DATASET SHIFT IN LAND-USE CLASSIFICATION FOR OPTICAL REMOTE SENSING

by

Francois Pierre Sarel Luus

Submitted in partial fulfillment of the requirements for the degree
Philosophiae Doctor (Electronic Engineering)

in the

Department of Electrical, Electronic and Computer Engineering
Faculty of Engineering, Built Environment and Information Technology

UNIVERSITY OF PRETORIA

April 2016

SUMMARY

DATASET SHIFT IN LAND-USE CLASSIFICATION FOR OPTICAL REMOTE SENSING

by

Francois Pierre Sarel Luus

Promoter:	Prof B. T. J. Maharaj
Co-promoter:	Dr F. van den Bergh
Department:	Electrical, Electronic and Computer Engineering
University:	University of Pretoria
Degree:	Philosophiae Doctor (Electronic Engineering)
Keywords:	Remote sensing, classification, clustering, manifold alignment, texture features, internal validation, geometric similarity, feature learning, neural network applications, neural network architecture

Multimodal dataset shifts consisting of both concept and covariate shifts are addressed in this study to improve texture-based land-use classification accuracy for optical panchromatic and multispectral remote sensing. Multitemporal and multisensor variances between train and test data are caused by atmospheric, phenological, sensor, illumination and viewing geometry differences, which cause supervised classification inaccuracies. The first dataset shift reduction strategy involves input modification through shadow removal before feature extraction with gray-level co-occurrence matrix and local binary pattern features.

Components of a Rayleigh quotient-based manifold alignment framework is investigated to reduce multimodal dataset shift at the input level of the classifier through unsupervised classification, followed by manifold matching to transfer classification labels by finding across-domain cluster correspondences. The ability of weighted hierarchical agglomerative clustering to partition poorly separated feature spaces is explored and weight-generalized internal validation is used for unsupervised cardinality determination. Manifold matching solves the Hungarian algorithm with a cost matrix featuring geometric similarity measurements that assume the preservation of intrinsic structure across the dataset shift. Local neighborhood geometric co-occurrence frequency information is recovered and a novel integration thereof is shown to improve matching accuracy.

A final strategy for addressing multimodal dataset shift is multiscale feature learning, which is used within a convolutional neural network to obtain optimal hierarchical feature representations instead of engineered texture features that may be sub-optimal. Feature learning is shown to produce features that are robust against multimodal acquisition differences in a benchmark land-use classification dataset. A novel multiscale input strategy is proposed for an optimized convolutional neural network that improves classification accuracy to a competitive level for the UC Merced benchmark dataset and outperforms single-scale input methods. All the proposed strategies for addressing multimodal dataset

shift in land-use image classification have resulted in significant accuracy improvements for various multitemporal and multimodal datasets.

OPSOMMING

DATASTELSKUIF IN LANDSGEBRUIKKLASSIFIKASIE VIR OPTIESE AFSTANDSWAARNEMING

deur

Francois Pierre Sarel Luus

Promotor:	Prof B. T. J. Maharaj
Kopromotor:	Dr F. van den Bergh
Departement:	Elektriese, Elektroniese en Rekenaar-Ingenieurswese
Universiteit:	Universiteit van Pretoria
Graad:	Philosophiae Doctor (Elektroniese Ingenieurswese)
Sleutelwoorde:	Afstandswaarneming, klassifikasie, groepering, menigvoudbelyning, tekstuurkenmerke, interne validering, geometriese ooreenkoms, neurale netwerktoepassings, neurale netwerkargitekture

Multimodale datastelverskuiwings wat bestaan uit beide konsep- en kenmerkverskuiwings word in hierdie studie bestudeer om tekstuurgebaseerde landsgebruikklassifikasie-akkuraatheid te verbeter vir optiese panchromatiese and multispektrale afstandswaardneming. Multitemporale en multisensorvariansies tussen afrig- en toetsdata word veroorsaak deur atmosferiese, fenologiese, sensor-, verwerking, illuminasie- en besigtigingsgeometrieverskille wat gekontroleerde klassifikasie-onakkuraathede veroorsaak. Die eerste datastelverskuiwingverminderingstrategie behels insetverandering deur beide beeldpunt- en voorwerpgebaseerde skaduweeverwydering voor kenmerkbepaling met grysvlak-samevallingsmatriks- en lokale binêre patroonkenmerke.

Komponente van ’n Rayleigh kwosiëntgebaseerde struktuurbelyningsraamwerk word gebruik om multimodale datastelverskuiwing te verminder by die insetvlak van die klassifiseerder deur ongekontroleerde klassifikasie gevolg deur struktuurpassing om klassifikasie-etikette oor te dra. Die vermoë van geweegde hiërargiese agglomeratiewe groupering om swak geskeide kenmerksruimtes te verdeel word ondersoek en gewigsveralgemeende interne validasie word benut vir kardinaliteitsbepaling sonder toesig. Menigvoudspassing los die Hongaarse algoritme op wat ’n koste-matriks met geometriese ooreenkomsmetings gebruik onder die aanname van die onderhouding van intrinsieke struktuur in die datastelverskuiwing. Lokale-omgewing- geometriese samevallingsfrekwensie-informasie word teruggewin en verbeterde klassifikasie-akkuraatheid word aangetoon met ’n nuwe integrasie van hierdie informasie.

’n Finale strategie vir die hantering van multimodale datastelverskuiwing is menigskaalkenmerkleer, wat gebruik word binne ’n konvolusionale neurale netwerk om optimale hiërargiese kenmerkevertonings te verkry in plaas van ontwerpde tekstuureienskappe wat suboptimaal kan wees. Kenmerkleer produseer eienskappe wat gewys word om robuust te wees teen multimodale verskille in ’n maatstaf-grondgebruikklassifikasiedatastel. ’n Nuwe menigskaalinsetstrategie word voorgestel vir ’n geoptimeerde konvolusionale neurale netwerk wat klassifikasie-akkuraatheid verbeter tot ’n mededingende vlak vir die UC Merced-datastel en verbeter op enkelskaal-insetmetodes. Al die voorgestelde strategië vir datastelverskuiwingvermindering het gelei tot beduidende verbeterings in akkuraatheid vir verskeie multitemporale en multimodale datastelle.

***"Yesterday I was clever, so I wanted to change the world.
Today I am wise, so I am changing myself." – Rumi***

ACKNOWLEDGEMENTS

I would like to acknowledge and express gratitude toward the following:

- My Creator, for every breath.
- I dedicate this work to my grandmother, who taught me the true meaning of selflessness.
- My grandparents, parents and brothers, for their continuous support and encouragement.
- My study leader, Sunil Maharaj, for his sage advice and support throughout the course of my studies.
- My advisor, Frans van den Bergh, for the constant guidance and for all the wisdom shared.
- Hans Grobler, for his diligent administration of the Advance Computing Cluster at the University of Pretoria.
- Rabindranath Bhattacharya, Alexey Khokhlov, Arkady Shadursky, Damo Mitchell, Zeev Kolman, and Brian Salmon.
- Credit goes to the University of Pretoria (UP, South Africa), Remote Sensing Research Unit (Meraka, CSIR, South Africa), and the National Research Foundation (NRF, South Africa) for the financial sponsorship of my Ph.D. degree.
- This work was supported by the NRF. The opinions expressed and the conclusions arrived at are those of the author and are not necessarily to be attributed to the NRF.

LIST OF ABBREVIATIONS

CNN	Convolutional neural network
DCNN	Deep convolutional neural network
FHS	Felzenszwalb and Huttenlocher's segmentation
FS	Formal settlements
FSB	Formal settlements with backyard shacks
GDI	Generalized Dunn's indices
GE1	GeoEye-1
GLCM	Gray-level co-occurrence matrix features
GPU	Graphics processing unit
HSV	Hue-saturation-value
ICA	Independent component analysis
IK2	Ikonos-2
Isomap	isometric mapping
KNN	*k*-nearest neighbor
LAT	Local adaptive thresholding
LBP	Local binary pattern
LBPA	LBP-All
MDS	Multi-dimensional scaling
MST	Minimal spanning tree
NBU	Non-builtup
NN	Nearest neighbor
OIS	Ordered informal settlements
OBIA	Object-based image analysis
PBM	Pakhira-Bandyopadhyay-Maulik
PCA	Principal component analysis
PMF	Probability mass function
QB2	QuickBird-2
RBF	Radial basis function
ReLU	Rectified linear units
RGB	Red-green-blue
SIFT	Scale-invariant feature transform
SVM	Support vector machine
TSVM	Transductive support vector machine
Umax	Unweighted maximum select
UPGMA	Unweighted pair group method with averaging
UPGMC	Unweighted pair-group method using centroids
Urnd	Unweighted random select

VLAT	Vector of Locally Aggregated Tensors
WCentroid	Weighted centroid
Wmax	Weighted maximum select
WPGMA	Weighted pair group method with averaging
WPGMC	Weighted pair-group method using centroids
Wrnd	Weighted random select
WV2	WorldView-2
WWard	Weighted Ward

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION

1.1 PROBLEM STATEMENT

The accuracy of trained classifiers across domains acquired under varying measurement modes is of cardinal importance in pattern recognition and classification. This study aims to demonstrate methods and systems engineered to produce more accurate land-use classifications under multimodal dataset shifts and to contribute general methods and algorithms that could be used in artificial intelligence areas outside remote sensing.

The proposed research theme is feature classification under measurement mode variances, with specific application to the remote sensing scenario of across-date and across-sensor settlement type and land-use classification. The aim is to develop general methods that can improve land-use classification accuracy if there are differences in illumination and viewing geometries between the train and test images, due to different acquisition times, or sensor differences due to different satellites being used.

1.1.1 Context of the problem

In the advent of electro-optical high spatial resolution digital satellites new avenues have been opened for detailed information extraction, especially in the area of land-use thematic mapping. Careful management of the earth's resources is becoming more critical, and an essential part of earth observation is environmental and land-use monitoring. As the global population is expected to rise to nine billion by 2050, and as the era of abundant cheap resources is drawing to a close, the need for intelligent resource management becomes evident.

Environmental conservation, natural resources management, land-use enforcement and the mapping of urban sprawl all benefit from information extracted from remotely sensed images. Settlement type and land-use classification, as well as settlement type change detection, can assist city planners in monitoring the expansion of informal settlements, for example, and help improve service delivery to those areas. Assisted classification of the earliest image can be used to train a basic classifier (shown in Figure 1.1) that can then automatically classify settlements on images acquired at later dates. An important theme in the proposed research is that of pattern classification with varying features, as it extends to a wide range of general classification problems.

The higher resolution images provided by current satellites require updated techniques and methodology to make full use of the underlying information. Although plenty of attention has been given to

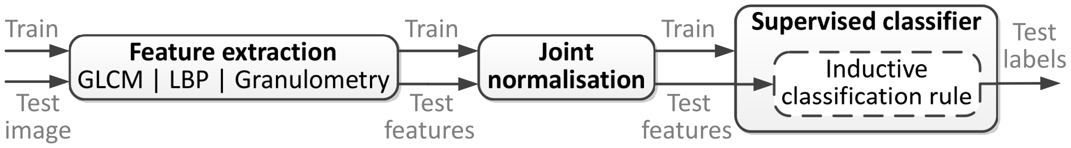


Figure 1.1. Basic supervised classification system.

multispectral high-resolution remote sensing, there is still insufficient research specifically into panchromatic-only information extraction. It is important to address this omission, as there are panchromatic-only satellites such as WorldView-1, a commercial earth observation satellite owned by DigitalGlobe, which provides high-resolution imagery with wide and frequent earth coverage that is sensitive to light of a colors in a given spectral band.

In addition, while most earth observation satellites provide both panchromatic and multispectral imagery, there is a distinct cost premium on multispectral data that makes panchromatic-only products the only affordable choice for many institutions. However, the panchromatic sensor typically provides imagery that is four times denser than multispectral, so there is a strong incentive to develop techniques specifically for panchromatic data.

1.2 DATASET SHIFT

Classification of multimodal remote sensing imagery is complicated by dataset shift, which is a change that happens between the training and testing environment of the classifier that causes inaccurate classification. Understanding dataset shift is critical to effectively addressing multimodal remote sensing classification, since dataset shift presents strongly in the remote sensing acquisition differences. In this section the subject of dataset shift is reviewed and its strong relation to the study in this thesis is emphasized.

1.2.1 Overview

Given a multidimensional feature or covariate x and a target or class variable y, the joint probability distribution $P(y,x)$ can be written as $P(y|x)P(x)$ in settings where class labels are causally determined by covariates. The prototypical classifier with parameters ω assigns a label $f(x,\omega) \in [1,\ldots,K]$ with one of K values corresponding to a particular class to an input covariate x. A loss function that detects classification errors when $f(x,\omega)$ does not equal the groundtruth label y is defined as

$$L(y,f(x,\omega)) = \begin{cases} 0 & \text{if } f(x,\omega) = y \\ 1 & \text{if } f(x,\omega) \neq y. \end{cases} \tag{1.1}$$

The risk functional that quantifies the overall classification error with classifier parameters ω is then given by

$$R(\omega) = \int\int L(y,f(x,\omega))p(y,x)\,\mathrm{d}x\,\mathrm{d}y. \tag{1.2}$$

The study objective is to minimize the risk $R(\omega)$ by discovering the optimal classifier design parameters ω_0 so that $\omega_0 = \mathrm{argmin}_\omega R(\omega)$. Ideally the parameters for two classifiers trained on the training data and test data, respectively, should be equivalent so that $\omega_{\mathrm{tr}} \approx \omega_{\mathrm{tst}}$.

The assumption in supervised discriminative classification is that the input prior probability density function (PDF) $P(x)$, class prior PDF $P(y)$ and conditional class PDF $P(y|x)$ remain unchanged between the training and testing scenarios [1]. In real world applications achieving this equivalence is complicated by variations between training and testing environments. Dataset shifts appear when the joint distribution $P(y,x)$ of the input and class variables (x and y) differ between the training and testing datasets (X_{tr} and X_{tst}). The concept of dataset shift is illustrated in Figure 1.2.

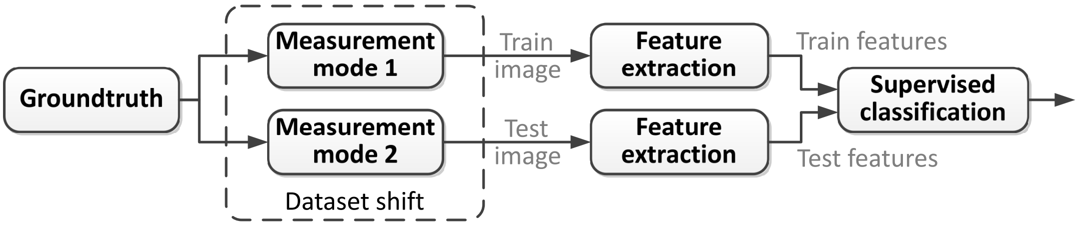


Figure 1.2. Supervised classification system affected by dataset shift.

Such dataset shifts may be caused by a non-stationary classification environment or domain shift, which is characterized by a change in the measurement system, or method of description [2]. This is prevalent in remote sensing, where training and testing dataset mismatches may appear in response to seasonal changes or a terrain distribution difference between the datasets.

1.2.2 Types of dataset shift

In general there are four types of dataset shifts [2], which vary in terms of prevalence and correction difficulty:

1. **Covariate shift** occurs when $P_{\mathrm{tr}}(x) \neq P_{\mathrm{tst}}(x)$ and $P_{\mathrm{tr}}(y|x) = P_{\mathrm{tst}}(y|x)$.
2. **Concept shift** occurs when $P_{\mathrm{tr}}(y|x) \neq P_{\mathrm{tst}}(y|x)$ and $P_{\mathrm{tr}}(x) = P_{\mathrm{tst}}(x)$.
3. **Prior probability shift** occurs when $P_{\mathrm{tr}}(y) \neq P_{\mathrm{tst}}(y)$ and $P_{\mathrm{tr}}(x|y) = P_{\mathrm{tst}}(x|y)$.
4. **Dataset shift** occurs when $P_{\mathrm{tr}}(x) \neq P_{\mathrm{tst}}(x)$ and $P_{\mathrm{tr}}(y|x) \neq P_{\mathrm{tst}}(y|x)$.

The covariate shift is well covered in the literature and there are several examples in different applications such as off-policy reinforcement learning [3], spam filtering [4], bioinformatics [5], brain-computer interfacing [6] and sample selection bias in economics [7]. Prior probability shifts are relevant only to classification scenarios where the class label y causally determines the values of the covariates x [2], but in the discriminative classification considered in this thesis the reverse is true, namely that the class label is causally determined by the covariate values x.

Dataset shifts that cause a change in both the conditional class probability distribution $P(y|x)$ (concept shift), conditioned on the input variable x, and the prior input probability distribution (covariate shift)

$P(x)$ are rarely discussed in the literature and are considered impossible to solve in the absence of assisting assumptions [2]. This fourth type of dataset shift is the one that is considered in this thesis, since multimodal remote sensing image variations can cause extracted texture features of the same class to change ($P(x)$) while also simultaneously changing the class definitions ($P(y|x)$).

1.2.3 Causes of dataset shift

The two main causes of dataset shift are sample selection bias, which is a form of covariate shift, and non-stationary measurement environments. Non-stationary measurement scenarios include adversarial classification problems, such as spam filtering and network intrusion detection. A set of benchmark datasets has been compiled by Moreno-Torres et al. for testing methods that deal with dataset shifts [8], but most of the datasets in the repository are synthetic or are derived from real-life datasets through artificial shifts.

Real-life reasons for dataset shift according to Storkey [9], [10] include:

1. **Simple covariate shift** is where the probability distributions of covariates x change and everything else stays the same.

2. **Prior probability shift** is where the probability distribution of y changes and everything else stays the same.

3. **Sample selection bias** is where training and test distributions differ as a result of an unknown instance rejection process.

4. **Imbalanced data** are a deliberate dataset shift for modeling or computational convenience.

5. **Domain shift** involves changes in measurement between the training and test samples.

6. **Source component shift** involves changes in the strength of contributing components. Three types are specified as follows:

   - Mixture component shift is where samples of (x,y) can come from a number of different sources s, but where the source priors can change, i.e. $P_{\text{tr}}(s) \neq P_{\text{tst}}(s)$ and $P_{\text{tr}}(x,y|s) = P_{\text{tst}}(x,y|s)$.

   - Mixing component shift is an aggregate of mixture component shifts that are sampled independently and identically distributed so that observations x are the average of observations drawn from each of the mixture component samples.

   - Factor component shift is where the data are dependent on a number of different factors, each of which can be decomposed into a form and strength. The forms of the factors remain the same, but the strengths can vary between the training and test scenario.

1.2.4 Solutions

1.2.4.1 Dataset shift characterization

The existence and characteristics of dataset shift between datasets have been determined in the literature through the following methods:

1. **Correspondence tracing** is where the effects of dataset shift are explored through the comparison of rule-based classifiers trained on both datasets, which uncovers the classification characteristics that qualitatively describe the nature of the dataset shift [11].

2. **Conceptual equivalence** discovers discrepancies between datasets as a method of contrast mining [12].

3. **Statistical analysis** has been used as a framework to analyze the changes between the training and test probability distributions [13].

1.2.4.2 Solutions for covariate shift

Sample selection bias is an important case of covariate shift where a training instance is drawn from the test distribution, then selected into the training sample with some probability or discarded otherwise [14]. The training and test probability distribution is reflected in the training and test samples, and can be estimated using kernel density estimation [15]. The estimated density ratio can subsequently be used to either weight or resample the training instances. The estimation of the density ratio is model-based and a classifier derived from the adjusted training sample is dependent on the adjustment process and is consequently generally non-optimal.

The case of training samples that are only biased with respect to the class ratio, i.e. a sample selection bias that depends only on the class label, has also been investigated [16]. Resampling weights for the training instances have been estimated by minimizing the Kullback-Leibler divergence between the test sample and weighted training sample [17].

A list of important solutions to the more general case of covariate shift is as follows:

1. Weighting the log-likelihood function: The loss on the test data distribution is minimized by weighting the loss on the training distributions with an instance-specific factor [10], based on knowledge of the training and test data distributions [18].

2. Importance weighted cross-validation: Under a covariate shift the standard model selection techniques such as cross-validation do not work, so an adjustment called importance weighted cross-validation is presented by Sugiyama et al. [19].

3. Integrated optimization problem: Bickel et al. derived a purely discriminative solution that is expressed as an integrated optimization problem, which leads to kernel logistic regression and

an exponential model classifier for covariate shift [14].

4. Kernel mean matching finds a weighting for the training instances so that the first statistical moments of the test sample and weighted training sample in the reproducing kernel Hilbert space (i.e. a high dimensional feature space) are close [20].

5. Sub-class re-estimation: Unlabeled test data can be used to adapt classifier outputs when class-conditional probability densities change because of changes in prior subclass probabilities [21].

6. Genetic programming has been used to address the dataset shift problem in a cancer diagnosis case study [22]. Genetic operators such as selection, mutation and crossover are applied to the test dataset to optimize the test accuracy, a computationally expensive approach that also requires a large proportion of test labels.

1.2.4.3 Solutions for concept shift

Klinkenberg addresses concept shift in information filtering, which involves the adaptive classification of documents with respect to a particular user interest where both user interest and document content can change over time [23]. The filtering system uses adaptive time windows over the training data, representative training sample selection and example weighting.

1.2.4.4 Solutions for prior probability shift

There are adaptive and robust approaches to prior probability shifts, which can adjust the classifier when the class prior PDFs of the training and test datasets differ. Robust approaches base classifier selection on measurements that are ideally transparent to changes in class distribution, such as receiver operator curve analysis [24], [25], [26].

Adaptive approaches first train the classifier and then adapt the classifier parameters by using the test data, which are usually unlabeled. Adjustments can be brought about by the end-user, as in the detection of oil spills in remote sensing data where class imbalance is addressed [27], or through the dynamic updating of classification rules [28]. Automatic adjustment is another type of adaptive approach and an example is the iterative procedure of expectation-maximization presented by Saerens et al. for adjusting the outputs of the trained classifier with respect to changed priors, without having to refit the model or without having to know the priors in advance [29].

1.2.4.5 Manifold alignment

Manifold alignment refers to the alignment of two feature space manifolds so that similar across-domain classes are distance-wise close and dissimilar across-domain classes are distance-wise far after alignment. Manifold is a term used here to refer to the intrinsic structure of an associate feature space or dataset, and is used in a general sense with no implied relation to a strict mathematical manifold. A typical example of a manifold is the feature space connected by a k-nearest neighbor graph; the basic structure of manifold alignment is shown in Figure 1.3.

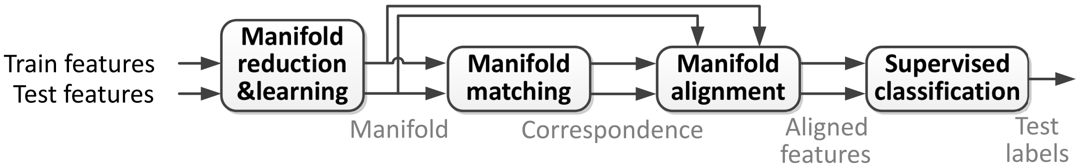


Figure 1.3. Manifold alignment for improving classification accuracy.

Manifold learning is a class of non-linear dimensionality reduction that attempts to describe the functional underlying structure or manifold of a dataset with a minimal number of features. Manifold learning methods include isometric mapping (Isomap) [30], locally linear embedding [31], Laplacian eigenmaps [32], Hessian eigenmapping [33], local tangent space alignment [34], and multi-dimensional scaling [35]. Isometric mapping seeks such a lower-dimensional embedding while preserving geodesic distances between manifold points. Locally linear embedding also uses partial eigenvalue decomposition, but instead preserves distances within local neighbourhoods. Spectral embedding or Laplacian eigenmaps find non-linear embedding via a graph Laplacian transformation of an adjacency graph before partial eigenvalue decomposition.

Manifold reduction as used in this thesis is a term that describes the unsupervised classification process of separating a dataset into distinctive classes or sub-classes for the purpose of establishing across-domain correspondences between the train and test manifolds in less computational time during manifold matching. Manifold alignment is a joint manifold learning method, since it finds a joint lower-dimensional embedding via generalized eigenvalue decomposition, but with the added function of preserving across-domain correspondences. Manifold reduction as used in this thesis is an initial component of effective joint manifold learning.

Manifold matching is strongly associated with the classification problem, since it aims to find correspondences between across-domain classes defined in the train and test manifolds. Non-injective or non-surjective manifold matchings incur severe computational complexity penalties, but design assumptions such as bijection and affine transformations are used to deliver a feasible demonstration of manifold matching.

1.2.5 Research gap

Land-use classification is one of the main tasks of remote sensing, and settlement type classification in particular is becoming more important in earth observation. A primary assumption in pattern classification is that of relative feature constancy, but in remote sensing especially, this feature invariance is not guaranteed because of the varying nature of satellite-borne image acquisition. This is a major problem that requires intense research efforts to introduce artificial invariance into the remote sensing classification system.

1.2.5.1 Multimodal domain adaptation

The primary research gap is the need for improved classification strategies for multimodal land-use classification, since only a few studies have investigated dataset shift correction or domain adaptation methods that address such dataset shifts. A Bayesian classifier that adapts class statistical parameters to match those of the testing dataset was used by Bahirat et al. [36] to update multitemporal land-cover maps.

Other domain adaptation methods have been applied in remote sensing, including those relying on cluster-distance metrics [37] and semi-supervised methods [38]. The deliberate transformation of a feature space to match a target space can be investigated, producing either a modified training or testing sample that may be used with any classifier, unlike the aforementioned classifiers that effectively adapt specific classifiers instead of datasets.

1.2.5.2 Panchromatic shadow detection

Accurate shadow detection is a necessary prerequisite for the introduction of input modification in Figure 1.4, but most of the available literature depends on color and multispectral distinction for detection. Shadowing is an inevitable acquisition artifact and the proposed research will endeavour to advance shadow detection techniques for panchromatic images especially. In particular, a watershed segment-merging algorithm has to be developed that can consistently produce quality panchromatic segmentation regardless of the specific image contrast [39]. The effect of shadow removal on settlement classification accuracy also has to be investigated for across-date imagery.

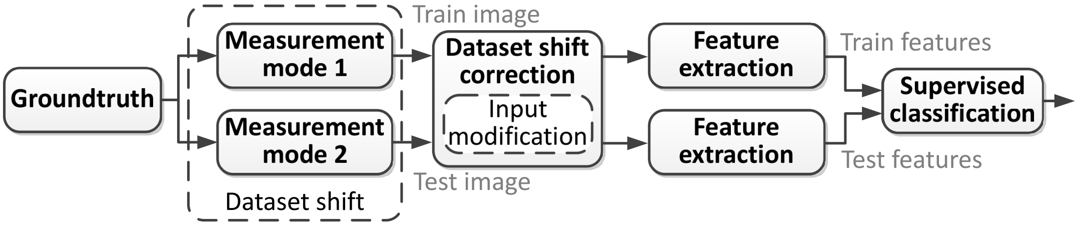


Figure 1.4. Supervised classification system with input dataset shift correction.

1.2.5.3 Post-feature dataset shift correction

Unlike shadowing, the viewing geometry requires augmented data such as digital surface models and relative satellite positioning to remove differences accurately in multitemporal images. This additional complexity may be avoided while maintaining a favorable result, by exploring semi-supervised and transductive classification or domain adaptation techniques such as the manifold alignment in Figure 1.5. These strategies can also reduce the dataset shift stemming from multisensor and phenological differences.

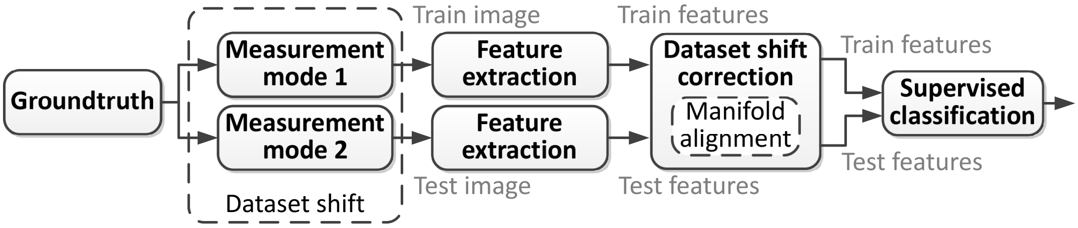


Figure 1.5. Supervised classification with manifold alignment dataset shift correction.

1.2.5.4 Weighted clustering for manifold reduction

Manifold alignment computational time can be managed by reducing given domain representations from the full set of samples to a representative description such as through statistical descriptors of salient clusters. Multiclass specimens appear in real world scenarios, especially in remote sensing, and there is an opportunity to demonstrate how weighted clustering can create artificial separation in feature spaces where classes cannot be properly distinguished.

The research gap can be filled by introducing sample weighting based on target properties and then properly investigating the effect of weighting on agglomerative clustering accuracy and internal validation. Weighted generalizations for select internal validation indices have been defined by Studer [40], which includes point-biserial correlation, Hubert's Gamma, Hubert's D, Hubert's C, Silhouette, Calinski-Harabasz and Pseudo R^2. A further research gap is the weighted generalization of many other known internal validation indices, such as those in the comprehensive compendium collected by Desgraupes [41].

1.2.5.5 Geometric similarity for manifold matching

The manifold alignment framework of Tuia et al. [42], reviewed in Appendix B, demands direct correspondence between across-domain clusters or points. Wang et al. [43] have demonstrated that geometric similarity measures can be used for manifold alignment and there is an opportunity to experiment with geometric similarity in texture feature spaces.

Geometric similarity calculations produce optimal neighborhood permutations, and there is a prospect for reusing using this information, which is normally discarded during manifold matching in the manifold alignment process of Wang et al. [43]. The research gap is to formulate a manifold matching measure suitable for features such as gray-level cooccurrence matrix (GLCM) texture features, incorporating geometric similarity. This requires a novel contribution, since the exact manifold matching measure depends on optimization for the specific features and classification scenario.

1.2.5.6 Multiscale feature learning

The main requirement for multimodal feature learning is that the training dataset should contain multimodal examples that display a large extent of the expected dataset shift. During classification the test examples should fall within the learned range of datasets with high probability, i.e. it should not cause a dataset shift outside of the previously observed dataset shifts. It is shown in this thesis that texture features such as GLCM and local binary patterns (LBP) are affected by the dataset shift in multimodal remote sensing, so there is a requirement for more robust features. A feature extractor/classifier should be learned that obtains features that are ideally unaffected by the type of dataset shifts that have been witnessed for the classification problem.

More specifically, there is a need for multiscale deep learning that can handle sample characteristics with varying sizes of representation, such as storage tanks in land-use imagery. Feature learning discovers the optimal features that optimize the multimodal classification objective and these features will consequently be minimally affected by multimodal dataset shift.

Competitively good feature representations for spatially organized forms of data such as images can be learned through methods such as convolutional neural networks (CNN) [44], which use the concept of receptive fields or filter definition field originating from Hubel and Wiesel's study of the feline striate cortex [45]. By using a sufficiently large number of convolutional layers a hierarchical feature representation can be obtained, which is a method of deep learning [46].

1.3 RESEARCH OBJECTIVES AND QUESTIONS

1.3.1 Dataset shift correction

The primary research objective is to improve land-use classification accuracy with optical remote sensing imagery under multimodal dataset shifts by exploring modifications and additions to the classification system that can minimize the detrimental effect of dataset shift. Dataset shift components or causes that can be isolated and addressed at each stage of classification must be determined. An outline of the main dataset shift correction methods is shown in Figure 1.6.

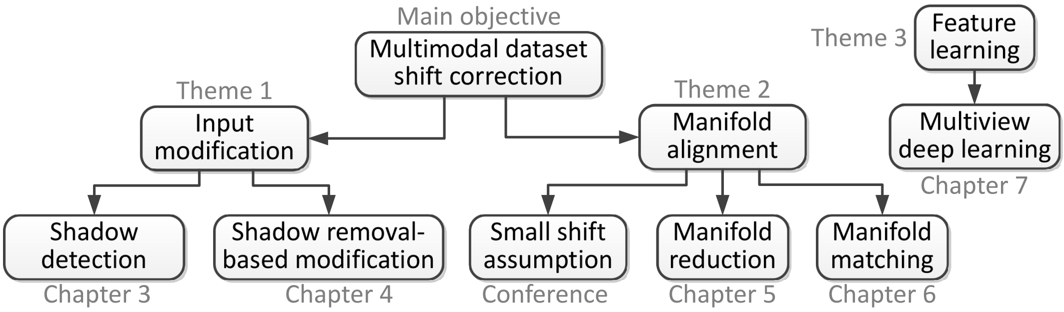


Figure 1.6. An outline of the thesis in terms of the main dataset shift correction objective.

1.3.2 Input modification

The aim is to engineer an input modification that can reduce dataset shift before feature extraction and to devise feature space manipulations that reduce dataset shift between train and test datasets before or during supervised classification, as shown in Figure 1.7.

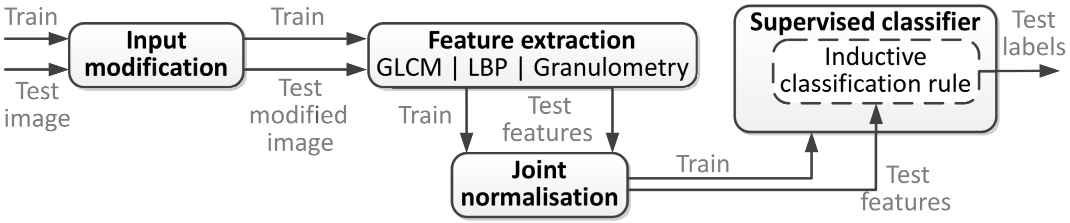


Figure 1.7. Supervised classification system with input modification.

1.3.3 Manifold matching and reduction

A main objective is to instantiate various components of the manifold alignment in Figure 1.8 to achieve a dataset shift correction. A dataset shift correction at the classification layer relies on the knowledge of across-domain correspondences, which is a task of the manifold matching component of manifold alignment. The manifold reduction component simplifies and imparts structure to the feature space to produce a manifold representation for computationally feasible manifold matching.

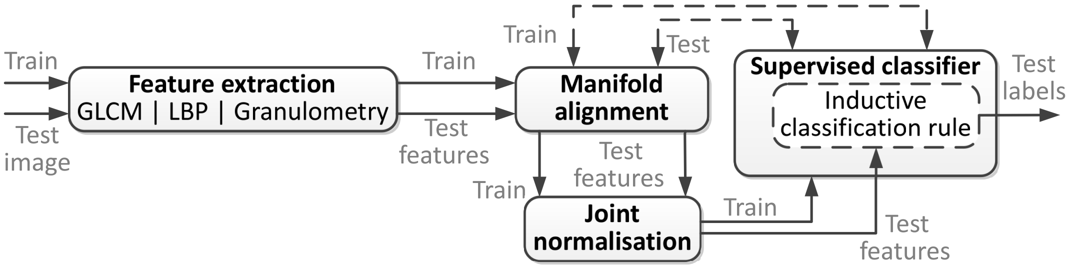


Figure 1.8. Supervised classification system with manifold alignment.

Unsupervised classification, shown in Figure 1.9, must be investigated by clustering a relatively large feature space into clusters that strongly relate to a target classification, and the challenge is specifically to undertake difficult clustering scenarios that require artificial feature space separation. Such complex clustering scenarios are prevalent because of the use of area texture-based features and land-use areas exhibiting features from multiple classes, which densifies interclass regions in the feature space and consequently deteriorates class separability.

Minimum-supervision or unsupervised manifold matching must be attempted by solving the classification problem under the assumption of the preservation of manifold geometry across the dataset shift. This will address the specific aim of correcting larger dataset shifts through manifold

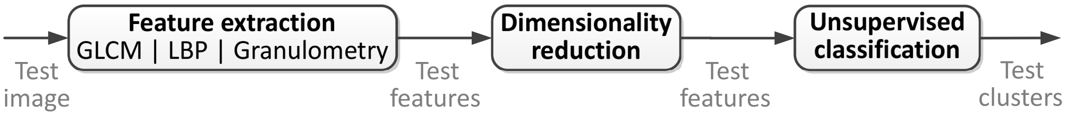


Figure 1.9. Basic unsupervised classification system.

reduction and manifold matching. For manifold alignment under a small dataset shift assumption the objective is to exploit the assumption or knowledge of a relatively small dataset shift, such as a multitemporal same-sensor scenario, to design a dataset shift correction measure.

1.3.4 Multiscale feature learning

The purpose of feature learning through CNN is to address dataset shift by discovering a hierarchical feature extraction that accurately characterizes the different classes despite the variations caused by dataset shift. The aim is to show how feature learning through CNN can produce competitive multimodal land-use classification accuracy when compared to current methods published in the literature for a benchmark remote sensing problem based on the UC Merced land-use dataset.

1.3.5 Research questions

1.3.5.1 Major research questions

1. How can the detrimental effect of multimodal dataset shift on remote sensing land-use classification be reduced to improve classification accuracy?
2. At which stages of pattern classification can dataset shift factors be removed or minimized?
3. What effect do multitemporal and multisensor dataset shifts have on texture-based land-use classification accuracy?
4. How can the manifold reduction and manifold matching stages of manifold alignment be implemented to address small and large dataset shifts?
5. How well can feature learning optimize a classifier for multimodal land-use classification in remote sensing images with within-class multiscale characteristics?

1.3.5.2 Chapter 3

1. How do the different threshold-based segmentations from the thresholding algorithm taxonomy of Sezgin and Sankur [47] compare in terms of panchromatic shadow detection accuracy?
2. How does global thresholding compare to locally adaptive thresholding in terms of panchromatic shadow detection accuracy?

1.3.5.3 In "The effects of segmentation-based shadow removal on across-date settlement type classification of panchromatic QuickBird images" [39]

1. How is shadow detection accuracy influenced when using segmentation or object-based shadow detection instead of fixed threshold shadow detection?
2. What is the relationship between change in shadow detection accuracy and change in settlement classification accuracy?

1.3.5.4 Chapter 4

1. Which strategy can be employed at the input level of feature extraction to deal with dataset shift in optical remote sensing?
2. What effects do multitemporal (same satellite) shadow profile differences have on texture-based settlement classification accuracy?
3. How does one achieve effective panchromatic shadow removal?
4. How much does adaptive threshold shadow detection improve classification accuracy compared to global threshold detection?

1.3.5.5 In "Mean translation of GLCM texture features for across-date settlement type classification of QuickBird images" [48]

1. What strategy can be employed after the output level of feature extraction to deal with dataset shift?
2. What strategy can be used to improve classification accuracy under a relatively small dataset shift?
3. How can manifold landmark points be obtained in the test data, under a relatively small dataset shift assumption, to enable the search for manifold matching?

1.3.5.6 Chapter 5

1. What manifold reduction strategy can be employed to create clustering separation in a poorly separated feature space?
2. How can a relevant sample weighting be obtained for texture-based land-use classification in remote sensing images?
3. What approach should be followed to obtain a scale-selective feature space when dimensionality reduction is used?
4. Which agglomerative clustering linkage is best for weighted clustering?
5. How can the optimal number of clusters be found, given a weighted feature space and hierarchical dendrogram?

6. Which internal validation indices perform best in a weighted clustering setting, and what role do sample weightings play in cardinality fitness?

1.3.5.7 Chapter 6

1. How does one perform unsupervised manifold (perfect) matching for relatively larger dataset shifts?
2. How can information derived during the optimal neighborhood permutation search be used to improve geometric similarity matching accuracy?
3. How should geometric similarity be employed, and which other correspondence measures should be applied to perform manifold matching accurately?

1.3.5.8 Chapter 7

1. How can features be learned that are optimal for minimizing the classification loss function under multimodal image variances?
2. How should a deep convolutional neural network (DCNN) be harnessed to improve classification where there are multiscale presentations of certain class characteristics, such as storage tanks that can vary in size depending on the sample?
3. How can a basic DCNN implementation be improved upon in order to increase classification accuracy?
4. What is the optimal DCNN architecture and configuration for the UC Merced dataset?

1.4 APPROACH AND HYPOTHESES

The basic supervised classification system is modified with the specific intent of reducing the detrimental effect of multimodal dataset shift on land-use classification accuracy, thereby producing a better classification system serving the primary aim of this study. The three predominant themes are input modification, manifold alignment and feature learning, which are approached separately with eventual demonstration of classification accuracy improvement. Two alternative approaches are discussed, namely generalized canonical correlation analysis and semi-supervised learning.

1.4.1 Input modification

The structured nature of viewing and illumination geometry variances in across-date imagery suggests that input correction measures are indeed plausible and can be developed. Shadowing is an example of one of the illumination effects that presents with more adverse variance in multitemporal imagery, but it is that well defined presentation that makes it possible to do shadow detection and removal, as in Figure 1.10

The hypothesis is that input modification should reduce feature dataset shift by removing shadows entirely from both train and test images, since existing shadow profile differences are a potential cause

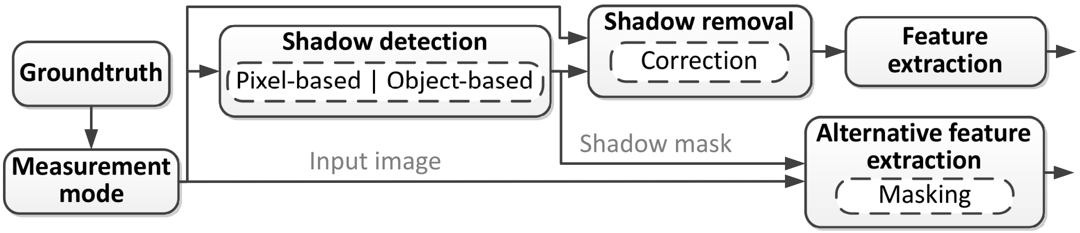


Figure 1.10. Feature extraction with shadow invariance input modification.

of dataset shift. Pixel-based thresholding and object-based segmentation are two different methods tested for shadow detection, and shadow removal is achieved either by producing a new image with the shadows corrected/lifted or by interfacing with feature extraction via a shadow masking process.

1.4.2 Manifold reduction and matching

To correct viewing geometry variation, phenological and other acquisition differences in the input is difficult in the absence of digital surface models and other augmenting information. Such complex variations may, however, be accounted for after feature extraction through a strategy such as manifold alignment. In the case of smaller dataset shifts due to the train and test inputs having been acquired from the same imaging vehicle, a small dataset shift assumption can alleviate the design burden of a normally unknown dataset shift.

In this situation supervised classification can be used to define an initial test manifold, as shown in Figure 1.11, and simplified divergence minimization can estimate a manifold match. Modal feature space translation correction can then be applied to ensure train and test manifold statistical moments coincide to reduce the associated dataset shift. This small dataset shift correction approach has been omitted from the thesis, but can be found in Luus et al. [48].

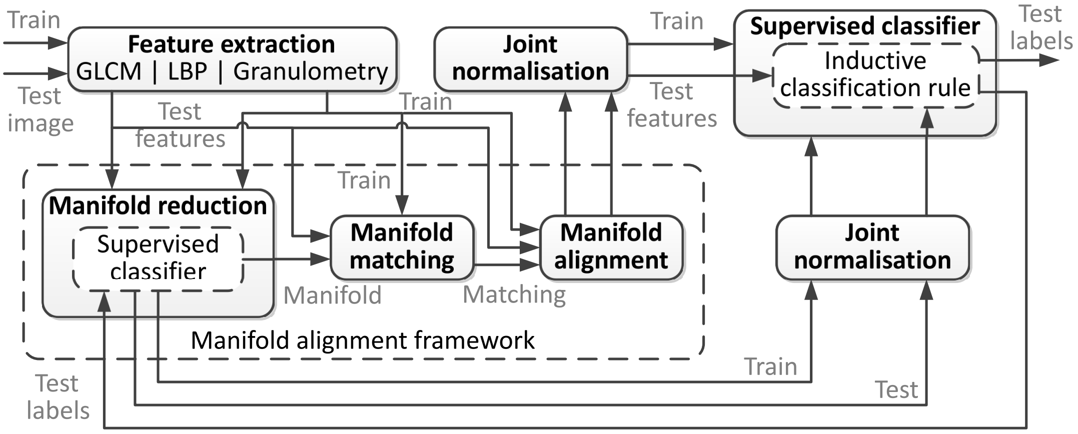


Figure 1.11. Supervised classification system with manifold alignment framework for small dataset shifts.

For the manifold alignment approach the assumption is made that the intrinsic structure or manifold of the train feature space undergoes an easily reversible transformation, e.g. an affine transformation, to produce the test feature space. This allows for a solution that discovers the transformation by attempting to preserve the manifold structure through the process shown in Figure 1.12, with the additionally required manifold learning, which is not indicated. Manifold learning informs the manifold alignment process of how the manifold geometry has to be preserved during alignment.

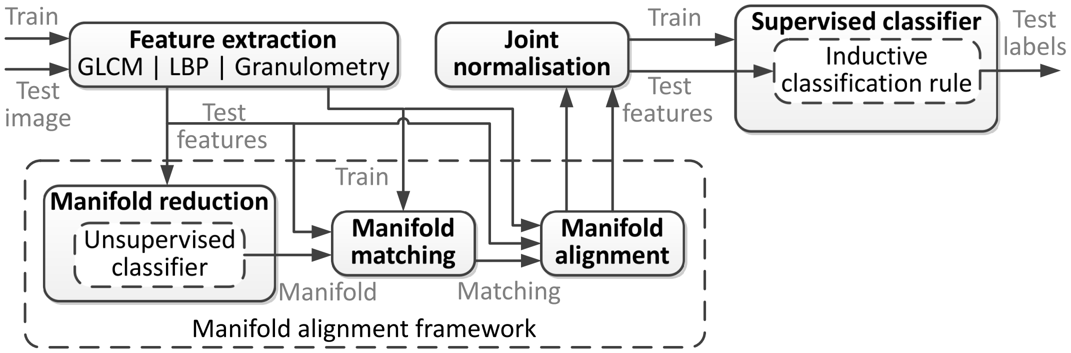


Figure 1.12. Supervised classification system with manifold alignment framework.

Manifold reduction is applied through unsupervised classification where target class properties enhance test feature space separability that enables weighted clustering to produce a target classification fit for the intended application. The target feature property of texture regularity is used to augment the feature space, producing a weighted feature space after scale-selective feature composition and kernel weight smoothing. Manifold matching is the most difficult component of manifold alignment, since nothing is known about the severity of the dataset shift, thus a small shift assumption or affine transformation assumption needs to be made to approach a solution.

Furthermore, in order to demonstrate minimum-supervision manifold matching in this thesis, one-to-one bijective correspondence (as indicated in Figure 1.13) between train and test classes is assumed, since correspondences that are either non-injective or non-surjective significantly escalate computational complexity.

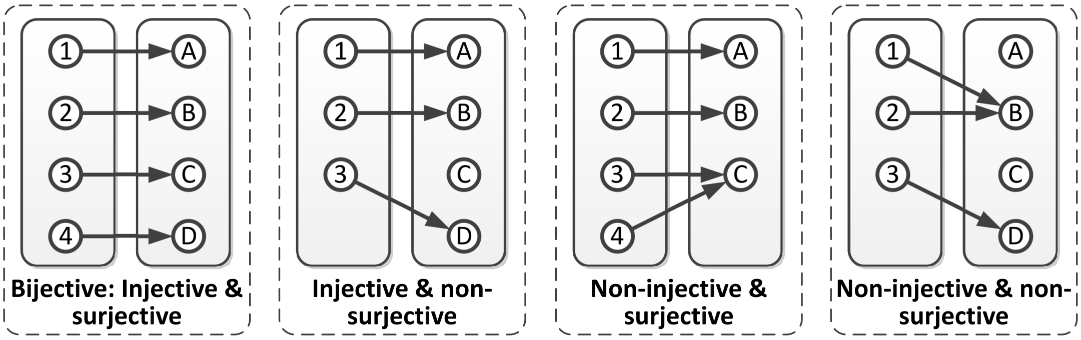

1
2
3
4
A
B
C
D
Bijective: Injective & surjective
Injective & non-surjective
Non-injective & surjective
Non-injective & non-surjective

Figure 1.13. Bijective, injective and surjective function combinations.

1.4.3 Multiscale feature learning

Input modification requires domain-specific expert knowledge of the factors that cause dataset shift and manifold alignment is constrained by assumptions and the low confidence of manifold matching. The genericity of manifold alignment can be retained with the approach of feature learning, without the need for expert knowledge of input modification. The feature learning approach to classification with dataset shift is illustrated in Figure 1.14.

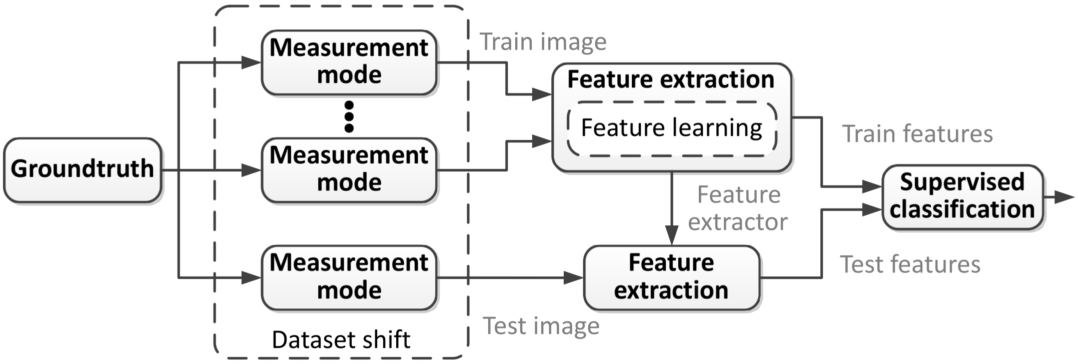


Figure 1.14. Supervised classification system with multimodal feature learning.

The end-to-end learning capability of deep learning allows for optimal features to be obtained that minimize the classification risk, which is very appropriate for addressing multimodal dataset shift when the training dataset contains sufficient multimodal variation so that robust features can be learned. This preemptive approach contrasts with the dataset shift correction strategies of shadow removal and manifold alignment, since it is expected to largely remove the need for these corrective measures.

The filter banks in the convolutional neural network (CNN) are iteratively modified to minimize the system loss, but the normal CNN structure first proposed by LeCun et al. [44] uses fixed scale features per convolutional layer. This may limit its expressiveness so strategies such as multiple input views [49], hybrid CNN [50] with variable receptive receptive field sizes and the Inception architecture of Szegedy et al. [51], which uses computationally efficient multiscale filter modules, have been proposed.

Deep convolutional neural networks can attain an average classification accuracy on a dataset with a rich set of different classes, such as the UC Merced land-use dataset [52] which displays multimodal variations. However, in order to achieve above-average classification accuracy that is competitive with the best methods available, a new strategy for harnessing the neural network has to be devised. The multiview strategy [49] is explored in Chapter 7 with the added dynamic of multiscale windows that can exploit the hypothesis of sub-sample redundancy in remotely sensed land-use images.

1.4.4 Semi-supervised learning

Instead of introducing illumination invariance into the texture features with shadow masking, as discussed previously, there is also the possibility of implementing effective invariance in the classifier itself. This will rely on a semi-supervised or transductive learning approach, and could have the capability of regulating complex acquisition variance.

Most natural learning occurs in a semi-supervised regime, where unlabeled data form part of the learning information. Intra-class variance shifts may be corrected with transductive classification and unlabeled clustering. A standard transductive classifier is the transductive support vector machine (TSVM), which uses unlabeled data to enhance a normal support vector (SVM) using the cluster assumption, which states that a decision boundary should lie in a low density region [53].

A TSVM was applied for pixel classification in remote sensing images [54], and also for semi-supervised learning in general remote sensing problems [55]. Semi-supervised SVMs outperform expectation maximization clustering, semi-supervised fuzzy c-means and indirect maximum likelihood with multivariate Gaussian distributions in a land-cover classification problem [56].

The SVM classifier uses the kernel trick, which performs a fixed mapping of feature vectors into a higher dimension in order to discover the nonlinear structure of the data. A data-independent kernel is used in that instance, such as a Gaussian or polynomial kernel, but it may not be consistent with the intrinsic manifold structure, geodesic distance, curvature or homology of the data [57].

An enhanced methodology involves warping the structure of the reproducing kernel Hilbert space to reflect the underlying geometry of the data. The local geometry may be captured by a nearest neighbor graph; the graph Laplacian can then be incorporated into the manifold adaptive kernel space and active learning can be performed [58].

1.4.5 Hypotheses and deductions

Conjectures are proposed in this subsection to explain observations and measurements acquired during the characterization phase of the study. Some immediate deductions are also made based on the associated hypotheses in order to initiate the engineering approach to the scientific method.

1.4.5.1 Major hypotheses

1. If dataset shift aspects between the train and test inputs to the feature extraction layer are corrected or equalized, then the dataset shift at the classification layer will also be reduced owing to the resulting features having smaller dataset shift.

2. Manifold alignment can be used to partially correct dataset shifts between the train and test feature spaces, because there are weakly supervised or unsupervised methods of manifold matching, which find the manifold correspondences and align them.

1.4.5.2 Chapter 3

1. Threshold-based shadow detection can relatively accurately delineate shadows because of the low intensity property of shadows.

2. Locally adaptive thresholds detect shadows below a threshold relative to local image intensity, which should produce more accurate shadows than with a global fixed threshold, since relatively low intensity admits greater sensitivity in images with contrast variation than globally low intensity.

1.4.5.3 In “The effects of segmentation-based shadow removal on across-date settlement type classification of panchromatic QuickBird images” [39]

A segmentation or object-based shadow detection approach can delineate shadow boundaries better than a threshold-based approach, since local features and contrast are taken into account.

1.4.5.4 Chapter 4

1. If dataset shift aspects between the classifier train and test inputs to the feature extraction layer are corrected or equalized, then the dataset shift at the classification layer will also be reduced because of the resulting features having smaller dataset shift.

2. Shadow profile differences between the classifier train and test images cause dataset shift at the classifier, and the removal of shadows in order to remove the shadow profile differences as well will reduce the corresponding dataset shift component and improve classification accuracy because of the resulting features having smaller dataset shift.

3. The more extreme a dataset shift becomes because of shadow profile differences, the more settlement classification accuracy will improve for an improvement in shadow removal accuracy, since more of the input dataset shift will be corrected with more accurate shadows.

1.4.5.5 In “Mean translation of GLCM texture features for across-date settlement type classification of QuickBird images” [48]

1. Definition can be created by induction in the test dataset to obtain landmark points in the test manifold as the first moment of each test class, since classification accuracy will be reasonable under a small dataset shift.

2. Manifold matching can be achieved under a relatively small dataset shift assumption by finding a perfect match between the two sets of manifold landmarks that minimises basic divergence, since there should be reasonable coincidence between train and test classes.

3. Feature space correction can improve classification accuracy by retraining the classifier after translating the training classes to have their first moments coincide with those of the test classes,

or by translating the test classes and reclassifying, since the corresponding dataset shift will also be reduced.

1.4.5.6 Chapter 5

1. Weighted clustering can attract cluster centroids toward classes with certain target properties, since agglomeration centroids gravitate toward higher weight regions.

2. Textural regularity as a target property can attract clusters toward more salient classes, since the target classification promotes classes with greater textural regularity.

3. Multiscale dimensionality reduction can be obtained with the principal components of only one particular scale, since the same groundtruth underlies the textures and the expectation is that sample importance will correspond well over the different scales.

4. Clustering linkages that incorporate sample weightings in the agglomerated cluster centroid calculation, but also the effective pairwise cluster dissimilarities, will provide more accurate clusterings, since the sample weightings have a greater impact on agglomeration.

5. Maximal weight input selection involves samples in the internal index calculation that improve cardinality decision accuracy compared to random selection, because the samples possess target characteristics and a greater affinity to the groundtruth classification.

1.4.5.7 Chapter 6

1. Local texture feature geometry is preserved across a multimodal dataset shift, since the feature relationships between classes are maintained and good features separate classes based on relative dissimilarity.

2. Across-domain classes that are more frequently matched together in optimal local neighborhood matchings are more likely to be matched in the across-domain matching, because such across-domain class pairs demonstrate a higher local geometry similarity.

3. Global translation and a basic divergence minimization objective can improve matching accuracy, since it corrects global domain differences and attempts to find the dataset shift with fewest assumptions, as stated by Occam's razor.

4. Relative class variances are possibly maintained under a dataset shift, since certain classes will usually have more , such as informal settlements, and other classes will have less variance, such as the non-builtup class.

1.4.5.8 Chapter 7

1. Basic texture features will distinguish poorly between distinct land-use classes where there are both multimodal and semantic within-class variations, because the low-level features may simultaneously be common across different classes, which causes excessive confusion.

2. The negative impact of multimodal image variances on classifier accuracy can be reduced with deep learning, since features are learned that are optimal for minimizing the classifier cost function.

3. A single DCNN with multiscale multiviews can improve composition-based inference of classes containing size-varying objects compared to single-scale multiview, since the size-varying objects have a greater probability of being featured at the right scale.

4. Increasing the number of different view scales can improve classification accuracy further, since a wider variety of object scales can then be accommodated.

1.5 RESEARCH GOALS

The primary intention is to develop processes and systems that can be used in more general problem scenarios, so that the contribution of this work can definitely be applied in problem domains other than texture-based land-use classification and remote sensing. The main research goal is to address multimodal dataset shift in texture-based land-use classification and to show that specifically engineered dataset shift reduction processes can increase classification accuracy.

1.5.1 Input modification

A specific goal is to remove across-domain differences in the input images, such as differing shadow profiles, so that the related dataset shift component is effectively reduced in the resulting features. The associated engineering task involves the design of a shadow detector and shadow removal method, and the aim is to explore threshold-based shadow detection but also more advanced object-based detection to improve shadow detection accuracy. A related target is to indicate experimentally that there is a strong correlation between shadow detection accuracy change and land-use classification accuracy change during shadow removal.

1.5.2 Manifold reduction and matching

Manifold alignment is a very appropriate framework for correcting dataset shifts, and the intention is to instantiate its critical parts, including manifold reduction and manifold matching. The instantiation of manifold reduction involves unsupervised classification and the target is to contribute a methodology specifically for the land-use classification problem, but also very general associated method novelties such as low-complexity weighted agglomerative clustering and weighted internal validation. The methods must show good use of the weighting information and result in a definite improvement in unsupervised classification accuracy.

Using manifold matching the desired contribution is a minimum-supervision or unsupervised method that can produce relatively accurate across-domain correspondence for multiple multimodal experiments. An impactful contribution would be the improvement of geometric similarity, which is a standard inclusion in manifold alignment without correspondence.

The intention is to achieve accurate classification under large dataset shift purely through the fact that test classes are established with manifold reduction and labels are transferred through correspondence derived from manifold alignment. A full conventional manifold alignment is thus not required if its components are established as stated.

1.5.3 Multiscale feature learning

The challenge of implementing a competitive deep learning solution for the UC Merced dataset involves the optimization of the CNN architecture and usage configuration. In addition, the need for a novel CNN usage strategy is emphasized in order to obtain competitive classification performance. The goal is to produce evidence for the hypothesis that label-preserving sub-sample redundancy in the UC Merced land-use image samples can improve classification accuracy.

1.6 RESEARCH CONTRIBUTIONS

Several methods, systems and study contributions have been made and a number of peer-reviewed articles have been produced based on this work. This student is responsible for 90% of the concepts and methodologies that are stated as being novel, and this student is also responsible for the implementation and execution of the research except where stated otherwise.

The social and economic impact of the contributions relate to the value of automation and integration of artificial intelligence and improving machine learning technologies can lead to increased value and subsequent impact. The benefit to academia and industry pertain to increased efficacy and improved functionality of a key technology for the future, namely machine learning.

1.6.1 Main contributions

A detailed list of the main contributions arising from all completed doctoral study activities are shown below.

1.6.1.1 Method contributions

1. Locally adaptive thresholding for panchromatic shadow detection.
2. Local contrast-robust segment-merging segmentation for panchromatic shadow detection [39].
3. Classifier-generic global and modal translation correction of feature spaces [48].

4. Scale-optimized texture feature space composition.
5. Weighted internal index generalisations.
6. Extremum-interpreted internal index knee-point accentuating filter.
7. Disruption-interpreted internal index suppression derivative.
8. Geometric similarity co-occurrence frequency.
9. Geometric similarity integrations of co-occurrence frequency.
10. Joint translation, divergence and geometric similarity manifold matching cost matrix.
11. Multiscale, multiview deep learning for multispectral land-use classification.

1.6.1.2 System contributions

1. Supervised and unsupervised object-based shadow detectors [39].
2. Texture-based land-use classification with input modification dataset shift correction.
3. Texture-based land-use classification with modal translation for small dataset shift correction [48].
4. Texture-based land-use classification with manifold matching based manifold alignment.
5. A CNN implementation with multiscale input views and classification probability averaging that produces a competitive classification accuracy for the UC Merced land-use dataset.

1.6.1.3 Study contributions

1. Threshold-based shadow detection comparison.
2. Shadow removal effect on multitemporal land-use classification accuracy.
3. Small dataset shift correction analysis [48].
4. Weighted agglomerative clustering linkage comparison.
5. Input-truncated internal index weighting analysis.
6. Manifold matching accuracy comparisons with instantiations of geometric similarity.
7. A CNN implementation with a heuristically optimized architecture that produces above-average accuracy on the UC Merced land-use dataset.

1.6.2 Research publications

Two internationally peer-reviewed conference papers, three internationally peer-reviewed ISI rated journal articles and another submission to an ISI rated journal have been produced based on the work presented in this thesis. The doctoral candidate was the lead author and researcher for all publications and was responsible for the concepts, methodology and execution. The list of generated publications is shown below.

1.6.2.1 Peer-reviewed conference publications

1. F.P.S. Luus, F. van den Bergh, and B.T.J. Maharaj, “The effects of shadow removal on multitemporal settlement type classification,” in IEEE Geoscience and Remote Sensing Symposium (IGARSS 2012), pp. 6196-6199, July 2012.

2. F.P.S. Luus, F. van den Bergh, and B.T.J. Maharaj, “Mean translation of GLCM texture features for across-date settlement type classification of QuickBird images,” in IEEE Geoscience and Remote Sensing Symposium (IGARSS 2013), July 2013.

1.6.2.2 Peer-reviewed journal publications

1. F.P.S. Luus, F. van den Bergh, and B.T.J. Maharaj, “The effects of segmentation-based shadow removal on across-date settlement type classification of panchromatic QuickBird images,” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 6, no. 3, pp. 1274-1285, 2013.

2. F.P.S. Luus, F. van den Bergh, and B.T.J. Maharaj, “Adaptive threshold-based shadow masking for across-date settlement classification of panchromatic QuickBird images,” IEEE Geoscience and Remote Sensing Letters, vol. 11, no. 6, pp. 1153-1157, 2014.

3. F.P.S. Luus, B.P. Salmon, F. van den Bergh, and B.T.J. Maharaj, “Multiview deep learning for land-use classification,” IEEE Geoscience and Remote Sensing Letters, vol. 12, no. 12, pp. 2448-2452, Dec. 2015.

1.6.2.3 Submitted journal publications

1. F.P.S. Luus, F. van den Bergh, and B.T.J. Maharaj, “Weighted agglomerative clustering for multimodal high-resolution multispectral land-use segmentation,” IEEE Transactions on Geoscience and Remote Sensing, Submitted, Dec. 2014.

1.7 OVERVIEW OF STUDY

This thesis is organized into chapters pertaining to each of the separate study goals as shown in Figure 1.15. Each chapter defines its own particular research objective and presents a methodology and experimental results, thus allowing each chapter to be considered in relative separation. A just-in-time approach is used for literature study and technique and method definitions to minimize cross-reference, with the exception of Chapter 2, which includes longer definitions that would obstruct the readability if included in the relevant chapter.

Four main datasets are used in this work, as shown in Figure 1.16, each exhibiting the desired multitemporal or multimodal properties that would properly demonstrate the efficacy of dataset shift correction measures and feature learning. Because of the multimodal nature of the datasets, chapter

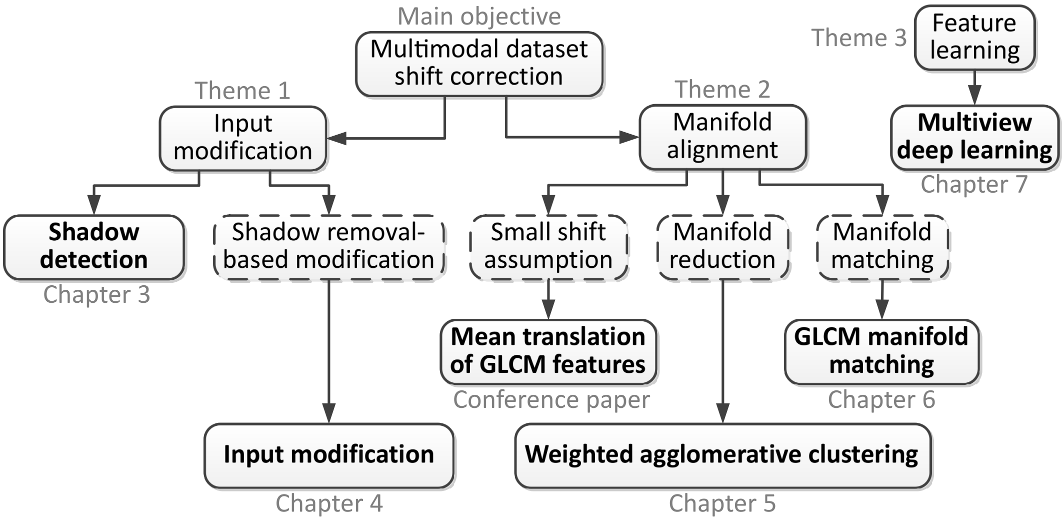


Figure 1.15. An outline of the thesis chapters.

results are only shown for particular relevant datasets. These datasets have been acquired in either panchromatic or pansharpened multispectral form over the United States of America, Rio de Janeiro (Brazil), Soweto and Johannesburg South (South Africa) as specified. The exact details of each dataset are shared in each associated study chapter.

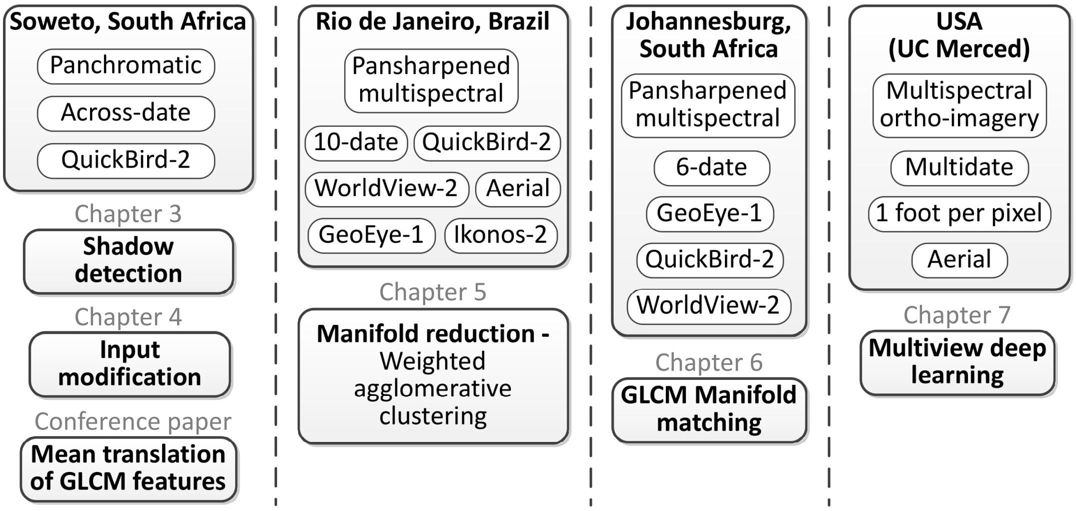


Figure 1.16. An overview of the datasets used by each study.

A literature study for the shadow detection and shadow removal methods used in input modification is performed in Chapter 2. Threshold-based shadow detection implementations are then explored in Chapter 3, followed by Chapter 4, which investigates dataset shift reduction based on input modification. Unsupervised classification based on weighted agglomerative hierarchical clustering is performed in Chapter 5 in order to reduce land-use feature manifold representations. Manifold alignment under a perfect match scenario is investigated with a reduced manifold in Chapter 6, which improves a geometric similarity measure that produces more accurate domain matching.

In Chapter 7 the role of feature learning in addressing classification problems with intra-training dataset shift is explored, and a multiscale input strategy is presented that provides competitive multimodal land-use classification performance. The thesis concludes in Chapter 8 with an overview of the approach and the results achieved, as well as a discussion of possible future research.

CHAPTER 2 SHADOW REMOVAL

2.1 CHAPTER OVERVIEW

A literature study of shadow removal is presented in this chapter with a focus on the causes of shadow variances, a radiometric modeling of the remote sensing (RS) system, as well as exploring the main components of shadow removal. Shadow removal firstly requires shadow detection, which is the first necessary preprocessing step of shadow removal, followed by de-shadowing or the actual shadow removal phase [59]. The purpose of this chapter is to provide an overview of all the relevant methods of shadow detection and shadow removal so that method selection can be performed for use in subsequent chapters. Synthesis beyond taxonomical integration would transport the study scope too far outside the focus on dataset shift in classification scenarios, so extensive incision into the technical merits/demerits is performed only in subsequent experimental analyses.

Three surveys have largely informed the content of this chapter and provided the elements of the shadow detection and shadow removal taxonomies presented here. The three important literature surveys on shadow detection and shadow removal have been contributed by Dare [60], Adeline et al. [59] and Shahtahmassebi et al. [61], as follows:

1. **Dare** [60] - “Shadow analysis in high-resolution satellite imagery of urban areas.”
2. **Adeline et al.** [59] - “Shadow detection in very high spatial resolution aerial images: A comparative study.”
3. **Shahtahmassebi et al.** [61] - “Review of shadow detection and de-shadowing methods in remote sensing.”

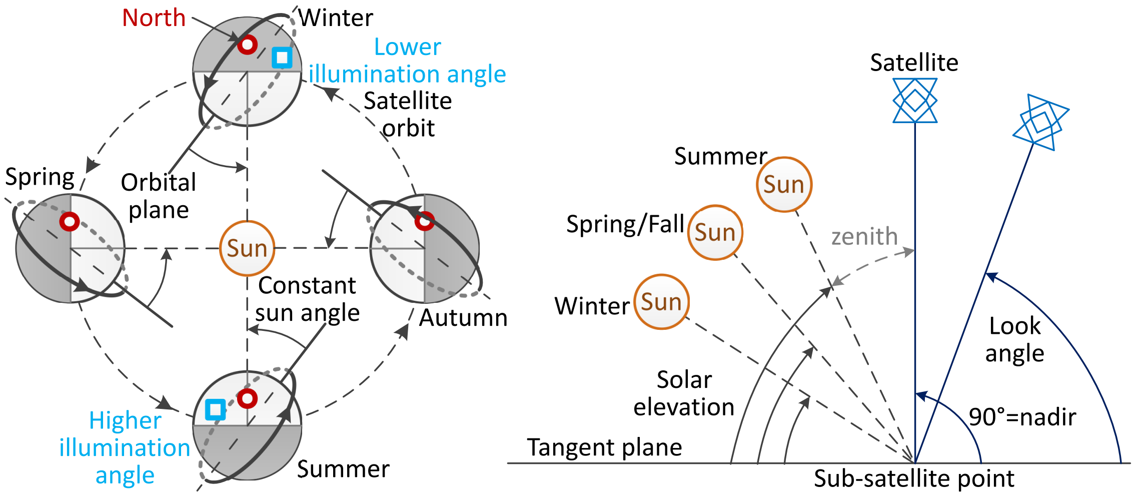
The goal of input modification in this RS study is to address dataset shift, which can hypothetically be achieved by removing a source of variance, namely shadows. By removing shadows a major component of illumination variance can be addressed. While shadows are regarded as a nuisance factor in this study, they do contain valuable information, which has been exploited for purposes such as 3D building reconstruction [62], building height estimation [63] and estimating solar incidence characteristics [64].

2.2 ACQUISITION-DEPENDENT SHADOW VARIATION

Remote sensing imaging introduces errors that are geometric in nature owing to the topography of scenes and atmospheric effects [61]. Shadows are a widely analyzed type of geometric effect, partly because of being a nuisance factor obscuring object details in scenarios such as land-use classification [65]. The acquisition-dependent variation of shadow profiles in temporal multimodal image analysis negatively influences applications such as change detection [66] and supervised classification [60].

2.2.1 Sun-synchronous satellites

Visible and infrared wavelength RS satellites normally follow sun-synchronous orbits, because of the consistent lighting and illumination angles. Seasonal implications for sun-synchronous satellites such as Ikonos involve changes in solar elevation angles, specifically producing a low elevation in winter (longer shadows) and a high elevation in summer (shorter shadows). An illustration of the seasonal influence on sun-synchronous satellites is shown in Figures 2.1(a) and 2.1(b).



(a) Sun-synchronous satellite orbit illustration

(b) Solar elevation angle and satellite look angle

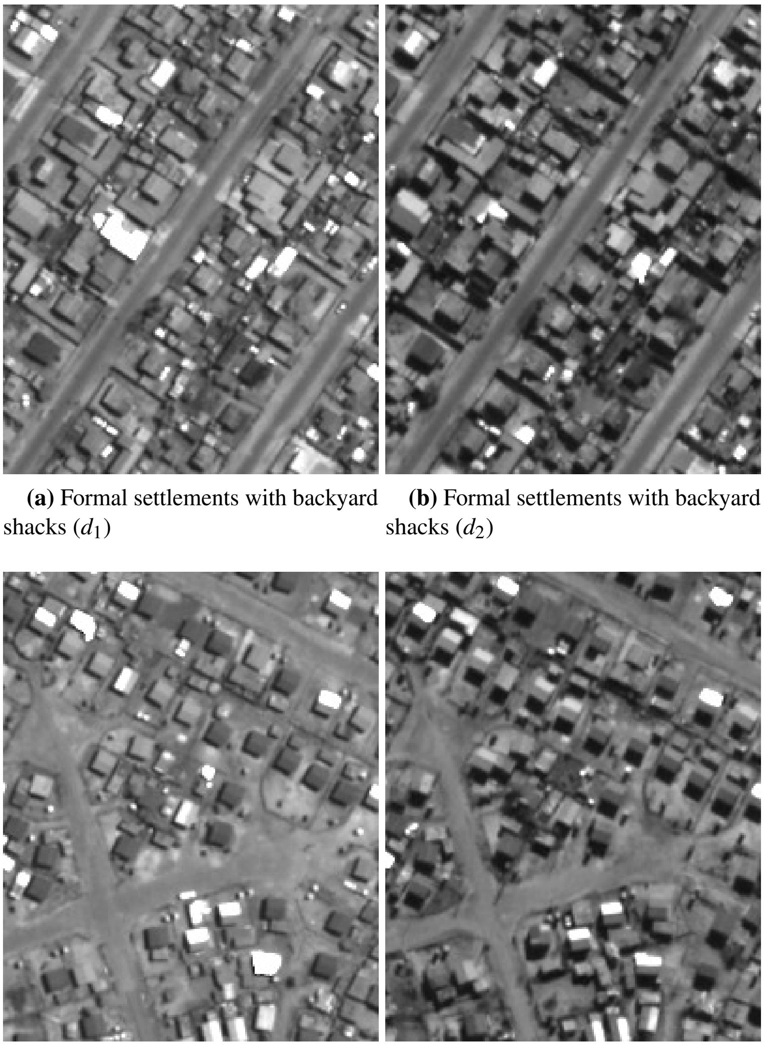
Figure 2.1. Sun-synchronous satellite orbit, solar elevation and the seasonal effect.

Sun-synchronous satellites acquire images at a fixed time every day, such as Ikonos with an equatorial crossing time of 10:30, so the cause of shadow profile variation is predominantly the seasonal solar elevation changes [60]. The earlier crossing time is chosen instead of 12:00, since the generally clearer atmosphere at that time of day is more important than the small gain in solar elevation. The look angle of the satellite can be increased to point to an off-nadir ground area with a later local solar time, which will reduce the apparent shadowing in the acquired image at the cost of increased feature occlusion because of the off-nadir attitude [60].

2.2.2 Urban shadow variation

The assumption of a flat scene geometry is usually made in the case of shadow correction for clouds, which is a popular topic in the literature for medium-resolution satellite images [67]. However, in this thesis the case studies consider mainly urban/suburban land-uses and non-builtup land-use to test builtup/non-builtup distinction.

Urban areas pose a greater challenge for shadow correction algorithms, because of higher proportions of shadow and higher object density, but also because of the high-resolution imagery, which reveals more shadows along with the scene detail [68]. An across-date example of panchromatic QuickBird imagery is shown in Figure 2.2 for urban/suburban land-uses, where the shadow profile differences due to the seasonal solar elevation differences are apparent.



(a) Formal settlements with backyard shacks (d_1)

(b) Formal settlements with backyard shacks (d_2)

(c) Formal settlements (d_1)

(d) Formal settlements (d_2)

Figure 2.2. Shadow difference examples in urban land-use images for two different dates d_1 (early summer) and d_2 (early winter). Panchromatic QuickBird images courtesy of DigitalGlobe™.

2.3 SHADOW MODELING

2.3.1 Radiometric modeling

To understand shadows in the context of RS it is necessary to formulate its properties in terms of a radiometric framework, which decomposes irradiance as the source of light in the scene and radiance as the collection of light at the satellite sensor. Shadows negatively affect land-use analysis in RS by causing radiometric distortions, such as radiative impact on the estimation of the reflective properties of surface materials [69].

2.3.1.1 Radiometry definitions

The radiometric framework is defined in terms such as radient flux, radiance, irradiance, spectral radiant flux, spectral radience and spectral irradiance. These terms are generally defined as follows:

1. **Radiant flux** is the radiant energy emitted, transmitted, reflected or received per unit time.
2. **Spectral radiant flux** is the radiant flux per unit wavelength.
3. **Radiance**, or historically called intensity, is the radiant flux emitted, transmitted, reflected or received by a surface, per unit solid angle per unit projected area, measured in the SI unit of watt per steradian per square meter.
4. **Spectral radiance**, or specific intensity, is the radiance of a surface per unit wavelength, measured in watt per steradian per square meter per nanometer.
5. **Irradiance** is the radiant flux received by a surface per unit area, measured in watt per square meter.
6. **Spectral irradiance** is the irradiance of a surface per unit wavelength, measured in watt per square meter per nanometer.

2.3.1.2 Ground irradiance

Adeline et al. define the total irradiance at ground level as the sum of the direct solar irradiance, the downwelling atmospheric irradiance due to light scattering by the atmosphere, the irradiance due to multiple scattering between the atmosphere and the ground, as well as the irradiance due to light reflection from surrounding surfaces [59]. The formulation for total ground level irradiance is given by

$$I_{\text{total}} = I_{\text{direct}} + I_{\text{diffused}} + I_{\text{coupling}} + I_{\text{reflected}}. \tag{2.1}$$

1. I_{total}: ground level total irradiance.
2. I_{direct}: direct solar irradiance.
3. I_{diffused}: downwelling atmospheric irradiance due to light scattering by the atmosphere.
4. I_{coupling}: irradiance due to multiple scattering between the atmosphere and ground.
5. $I_{\text{reflected}}$: irradiance due to light reflection from surrounding surfaces.

2.3.1.3 Sensor radiance

The radiance incident on the sensor surface is decomposed by Adeline et al. into the summation of direct radiance from the target scene to the sensor, the upwelling atmospheric radiance, as well as the scattered radiance light reflected from surrounding targets by the atmosphere in the field of view of the satellite sensor [59]. The composition of sensor level radiance is defined as

$$R_{\text{sensor}} = R_{\text{direct}} + R_{\text{environment}} + R_{\text{atmospheric}}. \tag{2.2}$$

1. R_{sensor}: total incident radiance on sensor's surface.
2. R_{direct}: direct radiance from target to sensor.
3. $R_{\text{atmospheric}}$: upwelling atmospheric radiance.
4. $R_{\text{environment}}$: scattered radiance light reflected from the surrounding targets and scattered by the atmosphere in the field of view of the sensor.

2.3.2 Radiometric properties

Adeline et al. analyzed the radiometric properties of shadow and identified the radiometric characteristics that distinguish shadow regions from sunlit regions [59]. Four unique shadow properties were identified through a simulation of a synthetic urban scene containing surfaces with Lambertian reflectance profiles.

These shadow properties are listed as follows based on the work of Adeline et al. [59]:

1. Shadow has much lower sensor radiance than sunlit counterparts over the whole reflective spectrum.
2. In constrained environments such as urban scenes the reflection effects due to 3D surroundings may not be negligible in shadow regions.
3. The sensor radiance component from shadowed regions decreases from short to long wavelengths because of scattering, so near-infrared channels can be used for better shadow detection.
4. The sensor radiance component from shadowed regions is material-dependent and material property retrieval can possibly be performed as in sunlit areas.

2.3.2.1 Radiance characterization

The most important observation regarding the radiative nature of shadow regions is that they receive less total irradiance, which is primarily due to significantly less direct irradiance. Irradiance of shadow regions that is caused by reflectance from surrounding surfaces, i.e. $I_{\text{reflected}}$, is another major source of shadow irradiance, which can reach between 10%-50% of the total irradiance [59].

A common false negative error in the use of radiative properties for shadow detection is the misclassification of high reflectance materials in shadow regions, since these bright surfaces are consequently much brighter than their lower reflectance counterparts in shadow regions. False positive errors are also possible when low reflectance objects such as dark cars in sunlit areas are misclassified as shadow because of the lower radiance from the object compared to sunlit counterpart surfaces.

2.3.2.2 Chromatic characterization

The chromatic characterization of shadows was defined in terms of the color properties of shadows, which are described by Tsai [70] as follows:

1. Lower luminance/intensity, because of most direct irradiance being blocked.
2. Higher saturation in blue-violet wavelengths, because of atmospheric Rayleigh scattering [71].
3. Increased hue values, i.e. the blue-magenta region towards the end of the hue-saturation-value (HSV) hue range.
4. The change of intensity of an area when shadowed or sunlit is positive proportional to the wavelength [72], or in other words the intensity reduction in shadow is larger for longer wavelengths, which is the reason blue-violet wavelengths have higher saturation in shadow.

The radiative impact of the atmosphere is significant in shadow regions, since I_{diffused} is the dominant component of shadow irradiance and directly involves the atmosphere as a radiative factor. However, scattering effects significantly decrease for longer wavelengths, which is seen as a wavelength-dependent reduction in I_{diffused} and I_{coupling} for shadow areas [59].

At shorter wavelengths $I_{\text{diffused}} > I_{\text{reflected}}$, but as wavelength increases the scattering effects in the atmosphere decrease, which makes reflected irradiance the dominant component, i.e. $I_{\text{diffused}} < I_{\text{reflected}}$. The reduction in scattering effects for longer wavelengths is also seen in the composition of the radiance that is incident on the satellite sensor, where $R_{\text{atmospheric}}$, $R_{\text{environment}}$ and R_{direct} drop with increase in the wavelengths.

2.3.3 Shadow component modeling

2.3.3.1 Shadow definitions

Shadow detection aims to separate an RS image into two regions, namely sunlit and shadow regions, with the goal of removing shadows, which potentially cause notable variances in multitemporal imagery. Shadow regions are formed when a non-zero fraction of direct irradiance from an illumination source is blocked.

Arévalo et al. categorized shadows into two classes [73], namely:

1. **Self-shadows**: the portion of the object surface that is not illuminated by direct irradiance.
2. **Cast shadows**: the shadow projected by the object in the direction of the light rays of the direct irradiance.

Shadows in RS imagery are generally cast shadows, because of the satellite look angle at acquisition time, which orients the rays of the sensor radiance in the same direction as the rays of the irradiance of the sun. In other words, the satellite looks in the same direction as the sunlight so that self-shadows will generally not be visible, as they are on the opposite side of the objects in the scene.

It is interesting to note that cast shadows are usually darker than self-shadow, because self-shadows receive more reflected radiance from surrounding surfaces [60]. Cast shadows also cause a significant reduction in spectral variation, which causes correlation failure when using stereo-autocorrelation, a technique that requires adequate spectral variation to enable correlation measures [74].

Cast shadows consist of two distinct components or regions [73], defined as follows:

1. **Umbra**: the part of the shadow where the direct irradiance is fully obscured.
2. **Penumbra**: the part of the shadow where the direct irradiance is only partly obscured.

2.3.3.2 Penumbra

The penumbra of the shadow is normally the transition between the umbra and sunlit regions in the image, and the penumbra width becomes more important for high-resolution RS, as it directly affects the shadow detection and removal approach. The penumbra width is strongly related to the solar angular width, which is related to the cross-sectional width of the sun as witnessed from the surface of the earth.

Dare uses trigonometry to calculate the penumbra width in terms of the solar elevation, solar angular width and object height, which is summarized as follows:

1. Penumbra width: w
2. Solar elevation: e
3. Solar angular width: ε
4. Object height: H

The trigonometry-based formulation for penumbra width is given by

$$w = H\left(\frac{1}{\tan(e-\varepsilon/2)} - \frac{1}{\tan(e+\varepsilon/2)}\right) \tag{2.3}$$

and a depiction of the problem scenario is shown in Figure 2.3. To put the penumbra width into perspective for high-resolution RS, consider the penumbra for a building height of $H = 25$ m, a solar elevation of $e = 38°$ and a solar angular width of $\varepsilon = 0.266°$. Firstly, the general shadow length of the building is given by $H/\tan(e)$, which is 32 m in this case. The penumbra width result is 0.61 m, which covers one pixel in a QuickBird panchromatic high-resolution image.

The penumbra is problematic in shadow detection and removal, since it contains ambiguity that requires a non-binary approach to handle correctly. However, the penumbra usually only occupies a small percentage of the cast shadow area, so many shadow removal approaches use only a binary shadow-sunlit classification, instead of a multi-level classification that can properly place the partial

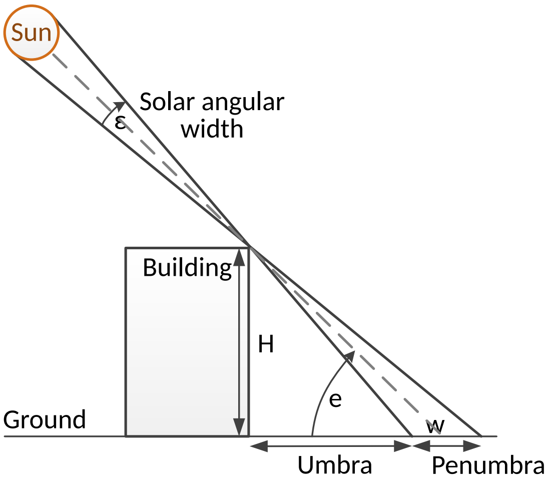


Figure 2.3. Penumbra width determination.

shadow characteristics of the penumbra region in relation to absolute shadow and absolute sunlit areas.

Dare suggests a threshold level in threshold-based shadow detection that does not classify the penumbra as shadow, since a subsequent shadow removal based on radiometric enhancement will generally create a bright border around corrected shadow areas [60]. Alternatively, shadow region processing can be used that excludes penumbra pixels from a detected shadow mask through a technique such as morphological erosion with a minimal structuring element. Shu and Freeman suggests defining three regions, namely sunlit, penumbra and umbra, instead of just sunlit and shadow, and then adjusting the brightnesses independently during shadow removal [75].

2.4 SHADOW DETECTION

2.4.1 Taxonomy

2.4.1.1 Categorization by Arévalo et al. and Adeline et al.

Arévalo et al. characterize shadow detection approaches into two classes, namely property-based and model-based methods [73]. Adeline et al. extend this shadow detection taxonomy by including physics-based and machine learning methods [59].

The general characteristics of the four classes of shadow detection approaches can be summarized as follows:

1. **Property-based**: uses shadow properties generally directly deduced from image data, such as radiometric attributes and spectral features.
2. **Model-based**: relies on augmenting information, such as the 3D geometry of the scene and atmospheric illumination conditions.

3. **Physics-based**: uses the physical properties of materials and knowledge of illumination conditions.
4. **Machine learning**: unsupervised an supervised classification generally based on shadow properties.

2.4.1.2 Categorization by Dare

In the shadow detection survey by Dare four general categories are specified [60], which are listed as follows, together with a categorization placing them in the combined taxonomy of Adeline et al. above:

1. **Thresholding**: Property-based, a large subcategory based on image data.
2. **Classification**: Machine learning.
3. **Region growing segmentation**: Property-based, object-based image analysis.
4. **Three-dimensional modeling**: Model-based, geometry-based shadow detection.

2.4.1.3 Categorization by Shahtahmassebi et al.

Shahtahmassebi et al. divide shadow detection methods into four categories and they introduce a new category called shade relief, which produces a categorization listed as follows:

1. **Thresholding**: Property-based, using color band ratios and spectral values.
2. **Invariant color model**: Property-based, using spectral properties of shadows.
3. **Modeling**: Model-based, three-dimensional modeling.
4. **Shaded relief**: Model-based, using solar zenith, solar elevation and digital elevation models to identify self-shadows.

2.4.1.4 Derived taxonomy

A shadow detection taxonomy was compiled from the aforementioned categorizations and classifications, and an overview is shown in Table 2.1.

2.4.2 Property-based shadow detection

2.4.2.1 Property-based characterization

Property-based shadow detection methods use the unique properties of shadows that can generally directly be deduced from the image information, so no augmenting data such as digital elevation models, 3D models or atmospheric information are required [59]. The types of shadow properties that are generally used for property-based shadow detection include the following [59]:

1. Radiometric attributes.

Table 2.1. Shadow detection taxonomy used in this study, composed from the classifications of Arévalo et al. [73] and Adeline et al. [59].

Property-based shadow detection	Arévalo et al. [73]
Thresholding	Adeline et al. [59]
Invariant color models	Adeline et al. [59]
Object-based algorithms	Adeline et al. [59]
Machine learning	Adeline et al. [59]
Supervised learning	Adeline et al. [59]
Unsupervised learning	Adeline et al. [59]
Model-based shadow detection	Arévalo et al. [73]
Geometry-based	Adeline et al. [59]
Physics-based	Adeline et al. [59]

2. Spectral features.
3. Textural attributes.
4. Spatial features.

One of the most important shadow properties is the radiometric property of shadows, which has been explored in subsection 2.3.2 and can be summarized as follows:

1. **Radiance property** of low intensity.
2. **Chromatic property** of high hue and high blue-violet saturation.

The subcategories that property-based shadow detection methods are divided into are listed as follows:

1. Thresholding.
2. Invariant color models.
3. Object-based algorithms.
4. Machine learning.

The list of examples found in the literature for each of the property-based shadow detection methods is shown in Table 2.2.

2.4.2.2 Thresholding

The shadow property of low intensity can be used to good effect for shadow detection [94], where histogram thresholding is a predominantly panchromatic detection method [76]. Histogram thresholding is property-based and is popular because of its speed and simplicity, since the assumption is made that there is a clear separation between shadow and sunlit histogram levels to facilitate the separation of the shadow and sunlit classes [66]. Object class variability can produce histogram shapes that are different from the assumed bimodal case, which increases the difficulty of defining an optimum threshold [59].

Table 2.2. Property-based shadow detection methods: Histogram thresholding, invariant color models, object-based algorithms and machine learning.

Property-based shadow detection	Arévalo et al. [73]
Thresholding	Adeline et al. [59]
Bimodal histogram splitting	Dare [60], Wei et al. [76]
Gaussian mixture model	Otsu [77]
Number of peaks and valleys	Chen et al. [78]
First valley detection	Liu and Yamazaki [66]
First peak classification	Wei et al. [76]
Visual inspection	Yamazaki et al. [79]
Improved object-based thresholding	Liu and Yamazaki [66]
Invariant color models	Adeline et al. [59]
Spectral ratio thresholding	Tsai [70]
Successive thresholding scheme	Chung et al. [80]
NIR exploitation	Teke et al. [81]
RGB+NIR	Fredembach [82], Nagao et al. [83]
Retinex theory	Wang and Wang [84], Aytekın et al. [85]
Object-based algorithms	Adeline et al. [59]
Region growing segmentation	Dare [60]
Heuristic segmentation	Liu and Yamazaki [66]
Rule-based object classification	Zhou et al. [86]
Morphological shadow index	Huang and Zhang [87]
Machine learning	Adeline et al. [59]
Supervised learning	Adeline et al. [59]
SVM	Levine and Bhattacharyya [88]
Pulse-coupled neural networks	Huang et al. [89]
Intrinsic image discrimination	Tappen et al. [90]
Wavelet features	Lorenzi et al. [91]
Unsupervised learning	Adeline et al. [59]
Intensity clustering	Yamazaki et al. [79]
Hyperspectral clustering	Ashton et al. [92]
Gaussian mixture models	Martel-Brisson [93]
Outlier detection	Shahtahmassebi et al. [61]

The property-based thresholding approaches found in the literature are summarized below:

1. **Bimodal histogram splitting**: Described by Dare as the most robust threshold selection for the pixel-level classification of shadow and sunlit areas [60], which involves setting the threshold as the mean of the two peaks in the bimodal histogram. Bimodal histogram splitting is suitable for large shadow detection, but finding a suitable threshold for the smaller shadows of urban structures is non-trivial [60].

2. **Gaussian mixture model**: Also known as Otsu's threshold, which is a robust binarization threshold [77].

3. **Number of peaks and valleys**: Chen et al. recognizes that a bimodal histogram with two classes featured as approximate Gaussian distributions would produce two different combined histograms, one with a single mode when the distribution means are close and another with two modes if the distribution means are sufficiently separated [78]. In the first case of a single combined mode the best threshold is at the single peak, but in the case of the bimodal histogram the best threshold is at the valley between the peaks.

4. **First valley detection**: A special case of a bimodal histogram corresponding to the decision process of Chen et al. [78] for the two Gaussian distributions being sufficiently separated to form a detectable valley between the two modes. The best threshold is then selected as the minimum location between the two Gaussian peaks.

5. **First peak classification**: Effective intensity thresholding was used by Wei et al. by choosing the lowest intensity class from unsupervised clustering by first peak selection in the histogram [76].

6. **Visual inspection**: Yamazaki et al. use a binarization threshold to distinguish between shadow and sunlit regions, and optimize the threshold based on visual inspection of the result [79].

7. **Improved object-based thresholding**: Liu and Yamazaki use a combined object-based and threshold approach by first performing object segmentation and then bimodal first valley threshold selection for thresholding the objects with [66].

Thresholding tends to perform worse for lower resolution imagery, since spectral mixing occurs within single pixels, thus obscuring the unique shadow properties. A common problem with thresholding is the confusion between shadows and water bodies, but several approaches have been developed to address the misclassification:

1. The misclassification of sunlit dark objects (water as shadows) or shadowed bright objects as sunlit, can be corrected afterward using texture features, edge features and other spectral information.
2. Dare uses region filtering where a variance threshold is set by visual inspection, based on the observation that shadow regions have higher variance than water regions [60].
3. Chen et al. use a spectral shape index to distinguish shadow from water regions [78].

2.4.2.3 Invariant color models

Adeline et al. divide invariant color model methods into two categories, namely red-green-blue (RGB) combinations and shadow invariant images [59]. The standard image colorspace RGB integrates radiance and chromaticity information, but the radiance and chromaticity can be separated by performing a color-space conversion to HSV, for example. The chromaticity information can be used separately to obtain lightening and shadow invariant derivations, or the true color at each image pixel can be retrieved as if shadows were absent, but color constancy is the main challenge here [73].

Invariant color models use the chromatic properties of shadows to perform shadow detection in multispectral imagery, rather than simpler histogram thresholding methods that only use intensity information [66]. Note that Dare compared shadow detection methods on panchromatic and pansharpened imagery, but achieved similar accuracy, indicating the strong radiometric effect of shadows over their chromatic character [60].

The unique chromatic properties of shadows include the following:

1. Higher saturation in the blue-violet wavelengths.
2. Increased hue values.
3. Greater intensity reduction for longer wavelengths.
4. Low intensity in NIR band.

Generally invariant color models assume that diffused irradiance dominates for the entire visible spectrum, i.e. $I_{\text{diffused}} \gg I_{\text{reflectance}}$. However, simulation by Adeline et al. shows that in shadow regions this is not normally the case, since $I_{\text{diffused}} < I_{\text{reflectance}}$ for greater wavelengths [59]. This means that the dominance of reflective irradiance at higher wavelengths can compromise the integrity of wavelength-dependent shadow properties, such as lower intensity at longer wavelengths.

The shadow detection approaches falling under the invariant color model category are listed as follows:

1. **Spectral ratio thresholding**: Tsai [70] presented an algorithm that uses the ratio value of the hue over the intensity to construct the ratio map for shadow detection in color aerial images. To accomplish this the RGB input image is first converted into a color invariant model, such as hue-saturation-intensity (HSI), and the ratio map is then calculated as $R = H/I$, which is then thresholded with Otsu's thresholding method.

2. **Successive thresholding scheme**: The detection accuracy was improved with a successive thresholding scheme used by Chung et al. [80]. Chung et al. stretch the gap between shadow and sunlit pixels in the ratio map (the one first proposed by Tsai [70]) by using an exponential function, and then perform global thresholding followed by local thresholding to refine candidate shadow regions.

3. **NIR exploitation**: Teke et al. primarily use the NIR band and the low NIR intensity property of shadows to perform shadow detection [81].

4. **RGB+NIR**: Relying only on the NIR property can produce confusion between water bodies and shadows, so Fredembach generates a combined space with both RGB and NIR channels where water and black objects have lower spectral responses than shadow, thus enabling the separation of shadow from water [82]. Nagao et al. created a linear combination of RGB and NIR channels, which can also improve shadow discrimination [83].

5. **Retinex theory**: The shadow property of maintained relative color but reduced intensity is embodied in the retinex theory, which has been exploited in several shadow detection approaches [84], [85], [91].

2.4.2.4 Object-based algorithms

Object-based or object-oriented shadow detection methods use the following enriched information, which has been shown by Chen et al. to improve shadow classification accuracy [95]:

1. Context
2. Texture
3. Spatial information
4. Radiometric pixel features
5. Spectral features of pixels

The object-based image analysis (OBIA) approach often involves segmentation as a first step [96], since it is an efficient technique of incorporting local information, such as texture and context [97]. Huang and Zhang have shown that, in contrast to pixel-based methods that cannot use spatial and contextual information, object-based methods can avoid false positives such as dark vehicles in sunlit areas being classified as shadow [87].

Pixel-based shadow detection methods generally produce more small shadows and more false negatives in the case of high reflectivity materials, such as bright roofs in shadow, but using neighborhood relationships the object-based shadow detection methods have a higher probability of classifying such surfaces as shadow [66].

1. **Region growing segmentation**: A popular category of region-based, bottom-up segmentation is region growing segmentation, which provides comparatively more control in the level of segmentation that is achieved. For segmentation with shadow detection as main purpose, it makes sense to choose starting or seed points corresponding to low-intensity pixels.

   Dare relates that a common merging predicate is the spectral distance and mean intensity of the neighborhood, since it can exclude radiometrically dissimilar pixels from merging with a segment [60]. Dare states that the optimal parameter for pixel agglomeration may be found in the histogram of the input image.

2. **Heuristic segmentation**: Liu and Yamazaki employed a heuristic segmentation algorithm that uses scale, color, smoothness and compactness information to optimize segment spectral homogeneity and spatial complexity [66], [98]. The scale parameter determines the maximum allowable heterogeneity in a segment, where a higher scale generally produces larger objects, and the smoothness and compactness parameters influence the properties of the eventual segment borders. Segment filtering can be used as a post-processing step in order to remove or regularize smaller objects.

3. **Rule-based object classification**: Zhou et al. use fractal net evolution for bottom-up region merging as the first step of segmentation [86]. The region merging segmentation is initialized at pixel level and a similarity-based merging predicate is then used, which is based on segment properties such as scale, color and shape. Rule-based classification is then performed with membership functions and a class hierarchy for eight land cover classes, using brightness thresholding for shadow/sunlit distinction based on visual interpretation with internally homogenous segments [86].

4. **Morphological shadow index**: Huang and Zhang proposed a morphological shadow index that is based on the spectral-structural characteristics of shadows, in order to indicate the presence of shadows in high-resolution imagery automatically through the local extraction of dark structures within a range of sizes in different scales and directions [87].

2.4.2.5 Machine learning

Shadow detection methods based on machine learning often involve classification approaches, in order to divide an image into shadow and sunlit regions. The machine learning category was first specified in the shadow detection taxonomy by Adeline et al. [59], but it should be noted that machine learning methods such as unsupervised classification do occur in other categories, such as thresholding. The categoric separation of machine learning from other categories in which it may appear, is motivated when the machine learning method is central to a shadow detection approach.

Li et al. divide shadow detection methods into supervised and unsupervised techniques [99]. Supervised machine learning methods for shadow detection require training samples from a groundtruth in order to train a classifier, such as the shadow examples manually determined by Zhan et al. [97].

Some important supervised machine learning methods for shadow detection are shown as follows:

1. **SVM**: Levine and Bhattacharyya use an SVM to classify shadow boundaries after the initial process of segmentation [88].
2. **Pulse-coupled neural networks**: Huang et al. employ a pulse-coupled neural network in a supervised setting to differentiate between shadow and sunlit regions [89].
3. **Intrinsic image discrimination**: Tappen et al. decompose a given image into two intrinsic images, namely a shading image that gives the illumination, and a reflectance image that gives the albedo or diffuse reflectivity under a Lambertian surface assumption [90]. The goal of the separation of intrinsic images is to remove the effects of shading from the reflectance information, which in effect removes shadows. Since there are two intrinsic images a dual purpose is served, namely shadow detection by using the shading image, and shadow removal by using the reflectance image.

   A supervised classifier is required that is trained with examples of shading and reflectance images, and the classifier can then be used to classify image derivatives as shading changes or reflectance changes. The classifier uses color information and it recognizes local patterns to detect shading, but Markov random field belief propagation is required afterward to propagate confident information to ambiguous regions.

4. **Wavelet features**: Wavelet features are calculated by Lorenzi et al. and used to conduct supervised multispectral shadow detection [91].

Unsupervised classification is the second category of machine learning methods for shadow detection and does not require training examples, since techniques such as clustering, Gaussian mixture models and outlier detection can naturally reveal classifications that can be used for shadow detection.

Important unsupervised classification approaches for shadow detection are given as follows:

1. **Intensity clustering**: Basic clustering methods such as k-means clustering can be used for shadow detection, since Yamazaki et al. have shown that the shadow class usually occurs in its own cluster [79]. The cluster corresponding to the lowest magnitude values can then be used to identify the shadow cluster.
2. **Hyperspectral clustering**: Ashton et al. have used k-means clustering on hyperspectral data in order to perform illumination suppression [92]. An issue with k-means clustering is the requirement of spherical clusters and an a priori cardinality, but the spectral variability of materials and the geometry in urban scenes can complicate the cardinality decision.
3. **Gaussian mixture models**: Martel-Brisson and Zaccarin have used GMMs to better fit multimodal distributions for moving cast shadow detection [93].
4. **Outlier detection**: Shahtahmassebi et al. suggested using outlier detection through clustering to perform shadow detection, instead of a simpler thresholding approach [61].

2.4.3 Model-based shadow detection

Shahtahmassebi et al. and Adeline et al. defined the model-based shadow detection category, which is defined as the group of shadow detection methods that is primarily based on a model [61], [59]. Where property-based shadow detectors use only the information present in the input imagery, the model-based shadow detection methods rely on extraneous and augmenting information, such as knowledge of the 3D geometry of the captured scene or atmospheric illumination conditions [59].

Adeline et al. categorized geometrical methods under the model-based category, but physics-based methods also depend on models, such as the blackbody radiator model, surface material models and reflectance correction models [59]. Physics-based shadow detection methods are consequently also included in the model-based category, which also addresses the situation of a singleton subcategory if only geometrical methods were included under the model-based category. The categorization of model-based shadow detection methods is shown in Table 2.3.

Table 2.3. Model-based and physics-based shadow detection approaches.

Model-based shadow detection	Arévalo et al. [73]
Geometrical methods	Adeline et al. [59]
Shade relief	Shahtahmassebi et al. [61]
Ray tracing	Thirion [100]
Line-of-sight analysis	Tolt et al. [101]
Terrain illumination correction	Wu et al. [102]
Physics-based approaches	Adeline et al. [59]
Linear unmixing	Boardman [103]
Backward radiance correction	Colby [104]
Blackbody radiator model	Makarau et al. [105]

2.4.3.1 Geometric methods

Geometric methods are the most common types of model-based shadow detection methods, since a number of options exist to obtain information about the topography and geometry of remotely sensed scenes. Sunlight models for the acquisition time can be combined with 3D information on the imaged scene, but the scene geometry and the sensor location and light source location relative to the scene is required, which may not always be available [106]. Working with such augmenting information is troublesome, as data availability is a prime concern and a further dimension of complexity is added to the analysis.

Moreover, the accuracy of geometric methods is limited by the accuracy of the 3D model accuracy. 3D models can be acquired through the following methods [59]:

1. Aerial photogrammetry (multiview, stereoscopic image pairs).
2. Satellite photogrammetry (multiview, stereoscopic image pairs).
3. Airborne laser scanning.
4. Terrestrial laser scanning.
5. Interferometric SAR.
6. Topographic data.

Photogrammetry performs poorly in a low-texture region and requires precise homologous point matching, while laser scanning can experience high multipath reflection and reduced accuracy in the presence of materials with strong absorption and reflection [59]. Rau et al. aim to generate true ortho-images of urban scenes through shadow removal that is based on geometric shadow detection with subsequent radiometric enhancement that has parameters determined by local histogram matching [107]. The actual elements on the ground that can cause shadows in RS are grouped into three categories by Shahtahmassebi et al. [61]:

1. Topography, such as shadows caused by mountains.

2. Urban objects, such as buildings and trees.
3. Clouds.

Important geometrical methods for shadow detection found in the literature are shown below:

1. **Shade relief**: Shahtahmassebi et al. listed shade relief as a primary shadow detection category, but it is rather a specific example of geometric model-based shadow detection methods [61]. Shade relief only identifies self-shadow components by using solar elevation, solar zenith and digital elevation models (DEMs).
2. **Ray tracing**: To perform shadow detection Thirion used DEMs and ray tracing, which is a method of calculating the paths of lightwaves from the sources of irradiance to the sensor [100].
3. **Line-of-sight analysis**: Shadow profiles have also been directly determined via a line-of-sight analysis, using information on the solar position and elevation in conjunction with surface models [101].
4. **Terrain illumination correction**: Topographic correction of surface reflectance was done by Li et al. [108] and terrain illumination correction was performed by Wu et al. [102] with shadow and occlusion detection using digital surface models.

2.4.3.2 Physics-based approaches

The second model-based shadow detection subcategory is physics-based, which describes a set of methods that primarily uses models of the physical properties of surface materials, illumination conditions and atmospheric influence on radiance and irradiance. Reflectance is essentially obtained, which is radiance converted by performing atmospheric correction that is influenced by the location of the scene, the solar elevation, solar zenith, viewing angle and aerosol profiles.

1. **Linear unmixing**: Linear unmixing determines the end-members or spectral source components, which should ideally match pure materials. The main assumption of linear unmixing is that the pixel reflectance is a mixture of linearly independent and fully illuminated end-member spectra. Shadows can form an end-member and can be identified as the end-member with the lowest radiance, but then there is the risk of confusion between shadows and low-reflectance non-shadow regions [103].

2. **Backward radiance correction**: Colby developed a backward radiance correction model using the Minnaert constant to minimize brightness differences for similar surface materials caused by topographic conditions, shadows or seasonal illumination changes [104].

3. **Blackbody radiator model**: The relationship between direct sunlight and scattered light may be modeled by a blackbody radiator model, as was done by Makarau et al. to perform multispectral shadow detection in a supervised setting [105].

2.4.4 Evaluation of shadow detection methods

2.4.4.1 Evaluation measures

The result of a shadow detector is a binary shadow mask, which indicates each pixel as either belonging to the shadow or sunlit classes. A groundtruth or producer's shadow mask represents the best possible result of shadow detection or the most accurate shadow mask. The shadow detector or user's shadow mask represents the shadow mask that needs to be evaluated for accuracy, which is done by performing a comparison with the producer's shadow mask.

An equivalence comparison of two binary images can produce four possible outcomes, which are defined as follows:

1. FP: False positive, producer pixel is shadow but corresponding user pixel is sunlit.
2. TP: True positive, both the producer pixel and corresponding shadow pixel are shadow.
3. FN: False negative, producer pixel is sunlit but corresponding user pixel is shadow.
4. TN: True negative, both the producer pixel and corresponding shadow pixel are sunlit.

The following four evaluation measures are subsequently defined as in Adeline et al. [59]:

1. **Producer shadow accuracy** (recall): $P_s = \frac{TP}{TP+FN}$.
2. **Producer non-shadow accuracy**: $P_n = \frac{TN}{TN+FP}$.
3. **User shadow** (precision): $U_s = \frac{TP}{TP+FP}$.
4. **User non-shadow**: $U_n = \frac{TN}{TN+FN}$.

Two combination measures of accuracy are also commonly used, namely overall accuracy and the F-score [109]. Congalton explains why overall accuracy is a good combination measure [110] (see [59]). These shadow detection evaluation measures are defined in terms of the above parameters as follows:

1. **Overall accuracy**: $\frac{TP+TN}{TP+TN+FP+FN}$.
2. **F-score**: $2\frac{P_sU_s}{P_s+U_s}$.

2.4.4.2 External validation indices

External validation indices measure how well a classification corresponds with a groundtruth classification. In the case of shadow detection, external validation indices measure how well a detected shadow mask corresponds with a groundtruth shadow mask. External validation indices are also composed of true positive, true negative, false positive and false negative terms.

The following external validation indices are used to measure shadow detection accuracy with [41]:

1. **Czekanowski-Dice index**: $\frac{2 \cdot TP}{2 \cdot TP+FN+FP}$.
2. **Jaccard index**: $\frac{TP}{TP+FN+FP}$.
3. **Rand index**: $\frac{TP+TN}{TP+TN+FN+FP}$.
4. **Rogers-Tanimoto index**: $\frac{TP+TN}{TP+TN+2 \cdot (FN+FP)}$.
5. **Sokal-Sneath index**: $\frac{TP}{TP+2 \cdot (FN+FP)}$.

The Rand index is equivalent to the overall accuracy, but the Jaccard index omits the true negative term TN. The F-score also omits TN and this results in a more appropriate accurate measure if the positive class has a low prior probability, which might be the case for the shadow class. Note that the F-score is equivalent to the Czekanowski-Dice index, since

$$2\frac{P_sU_s}{P_s+U_s} = 2\frac{TP}{TP+FN} \cdot \frac{TP}{TP+FP} \Big/ \Big(\frac{TP}{TP+FN} + \frac{TP}{TP+FP}\Big) \tag{2.4}$$

$$= 2\frac{TP}{TP+FN} \cdot \frac{TP}{TP+FP} \Big/ \Big(\frac{TP(TP+FP)+TP(TP+FN)}{(TP+FN)(TP+FP)}\Big) \tag{2.5}$$

$$= \frac{2 \cdot TP}{2 \cdot TP+FN+FP}. \tag{2.6}$$

2.4.4.3 Shadow detector performances

Adeline et al. ranked diversely different types of shadow detection methods in various high-resolution urban shadow detection scenarios, and found the following ranking in their experiments [59]:

1. **Histogram thresholding**, first valley detection with a modified intensity channel from Nagao et al. [83]
2. **Linear unmixing**, matched filter approach
3. **Supervised machine learning**, SVM
4. **Invariant color model**, YIQ (luma, in-phase and quadrature components)
5. **Unsupervised machine learning**, k-means

2.5 SHADOW REMOVAL

Shadow removal is the primary objective of this chapter and is performed after the initial step of shadow detection, which was discussed in the preceding section. The aim of shadow removal is to eliminate the varying factors that can cause dataset shift, which are shadows in this case as they vary mainly with seasonal acquisition differences. Several options exist for removing shadows, since the removal is actually intended to be part of a larger classification system and the shadow removal can consequently be performed at various stages of the classification system.

Three main shadow removal techniques given in the survey by Dare, which are listed as follows [60]:

1. Shadow masking.

2. Radiometric enhancement.
3. Multisource data fusion.

Only two primary categories for shadow removal are defined in this section, namely shadow masking and shadow restoration. Shadow restoration is fundamentally different from shadow masking, since shadow pixel values are not set to zero but are radiometrically enhanced to reveal details that were obscured by shadow. The result of shadow restoration often produces the effect as if objects did not obscure direct irradiance.

The multisource data fusion approach of Dare is categorized under shadow restoration in this section, since it involves the same type of image enhancement that characterizes shadow restoration, even if it involves multiple data sources [60].

Alternative names for shadow restoration have been defined by different authors as follows:

1. **Shadow restoration**, Zhou et al. [86].
2. **Image restoration**, Shahtahmassebi et al. [61].
3. **Radiometric enhancement**, Dare [60].
4. **Shadow compensation**, Li et al. [99].

Shadow masking and shadow restoration are the two main categories of shadow removal discussed in the remainder of this section.

2.5.1 Shadow masking

Shadow masking is significant in the context of the study of dataset shift, since it can allow a classifier to ignore selected image components that can vary across different acquisitions. Shadow pixels can be set to black to produce a modified image, but the modified image may not be appropriate for visual interpretation afterward or for producing a shadow restoration later on [60].

A classifier can receive the modified image and choose to ignore black pixels in the feature calculations, which is one method of achieving shadow masking in a land-use classifier. Another method is to input two images into the classification system, namely the original image and a binary shadow mask. The binary shadow mask can be used during feature extraction to decide whether to use a pixel or not. This approach incurs an increased implementation complexity in the classification system.

2.5.2 Shadow restoration

Shadow compensation, shadow restoration, image restoration and radiometric enhancement are grouped under one category of shadow removal, namely shadow restoration. There are minor differences between these terms, but the differences in nomenclature pertain mostly to the primary information types and methods used to achieve the same goal.

The strong commonality underlying these terms is that they intend to recover useful information in shadow regions [111]. Sarabandi et al. hypothesized that there is useful information in the weak radiance from shadow regions [112], and Nolè et al. noted that the sunlit regions neighboring shadow regions can be used in shadow restoration [113].

Shadow compensation refers to the adjustment of shadow pixels, in order to compensate for the unique properties of shadow pixels such as low intensity. Shadow restoration and image restoration are more general terms that relate more about the goal than the exact method to reach the goal. Radiometric enhancement is a more specific case of shadow compensation and shadow restoration, since it defines both the intent of enhancing the image and the type of information and method used.

Li et al. divided shadow compensation methods into two subcategories, namely intensity domain and gradient domain categories [99]. However, gradient domain shadow compensation methods have not been widely applied in shadow removal for RS, so the gradient domain subcategorization is not made. Gradient methods do feature in some intrinsic image methods, which were discussed in the supervised machine learning shadow detection, so gradient methods are subsumed here under an intrinsic domain category.

Li et al. specify two distinct modes of intensity domain shadow removal methods, which are characterized by the use of spatial similarity and by the modeling of shadowed images as products of shadow-free images and a shadow scale [99]. The product model of images strongly relates to intrinsic images, which are shadow and reflectance components of a given image. Several methods related to this approach have been applied in general shadow removal and are consequently listed as potential methods for shadow removal in RS. The subcategories of the shadow restoration category of shadow removal are listed in Table 2.4.

2.5.2.1 Intensity domain

Li et al. defined a popular category of shadow compensation, namely intensity domain shadow compensation [99]. An important mode of intensity domain methods are characterized by the use of spatial similarity to restore shadow regions, which implies the use of information in the surrounding sunlit regions to assist in the shadow compensation. Intensity domain methods primarily operate under the radiometric property of reduced pixel intensity of shadow regions, and generally resort to the use of local information of sunlit regions to perform shadow corrections.

Shu and Freeman suggested three methods that can be categorized under the intensity domain [75]:

1. Histogram equalization.
2. Algebraic grayscale transformation.
3. Mean and variance transformation.

The mean and variance transformation, or linear-correlation method as it is more commonly known,

Table 2.4. A categorization of shadow restoration methods.

Shadow restoration (Image restoration, radiometric enhancement, shadow compensation)	Zhou et al. [86] Shahtahmassebi et al. [61] Dare [60] Li et al. [99]
Intensity domain	Li et al. [99]
Histogram matching/equalization	Sarabandi [112]/Shu and Freeman [75]
Gamma correction	Nakajima et al. [68]
Spatial autocorrelation	Zhu et al. [114]
Linear-correlation	Nakajima et al. [68], Dare [60]
Piecewise linear-correlation	Zhan et al. [97]
Supervised chromatic correction	Makarau et al. [105]
Intrinsic domain	Li et al. [99]
Entropy minimization	Finlayson et al. [115]
Product model	Arbel and Hel-Or [116], Wu and Tang [117]
Poisson method	Xu et al. [118]
Multisource methods	
Multisource data fusion	Dare [60]

is a popular intensity domain method and has been found to perform relatively well [75]. There is similarity in the approaches of histogram matching and linear-correlation correction, since linear-correlation correction can be viewed as a special case of histogram matching, seeing that it uses only the primary statistical moments to approximate a histogram equalization. However, since linear-correlation correction is a popular intensity domain method, it is listed in a class of its own.

Multisource data fusion forms a separate category, as suggested by Dare [60], since the approach is unique in its use of local information across different sources. This approach is different from intensity domain shadow compensation approaches, since it pertains to substitution operations of shadow pixels with across-source sunlit counterparts rather than to a compensation transformation such as linear-correlation correction.

The listing of intensity domain methods is given below:

1. **Histogram matching/equalization**: Dare suggested radiometric enhancement that is similar to image balancing in orthomosaic generation, where radiometric differences across region boundaries are reduced by matching neighboring region histograms [60]. Histogram matching has also been proposed by Shu and Freeman as a main shadow removal method [75], and has been used by Sarabandi et al. for radiometric restoration by matching shadow region histograms with the histograms of sunlit regions of the same class [112]. Spatial radiometry variations are generally present in shaded images, so Dare states that it is best to perform histogram matching at a local level.

2. **Gamma correction**: Nakajima et al. describe a basic gamma correction method, which modifies shadow pixels exponentially so that the mean shadow pixel intensity μ_{shadow} and mean sunlit pixel intensity μ_{sunlit} match [68]. The gamma correction is more appropriate for higher resolution image restoration, according to Shahtahmassebi et al. [61].

   Gamma is determined as the ratio of the logarithms of the normalized shadow intensity mean and normalized sunlit intensity mean, which is then used to convert all shadow pixel intensities with the second equation

$$\gamma = \frac{\log\left(\mu_{\text{shadow}}/G\right)}{\log\left(\mu_{\text{sunlit}}/G\right)} \tag{2.7}$$

$$\mu_{\text{sunlit}} = G \times \left(\frac{\mu_{\text{shadow}}}{G}\right)^{1/\gamma}. \tag{2.8}$$

3. **Spatial autocorrelation**: Shahtahmassebi et al. suggest geo-statistical methods using the spatial autocorrelation technique of Zhu et al. [114], if the uniform variability assumption of interpolation techniques does not hold [61].

4. **Linear-correlation/mean and variance transform**: Linear-correlation is also called the mean and variance transformation by Dare [60], as well as Shu and Freeman [75]. Linear-correlation correction is sensitive to the local window size used to obtain sunlit region examples, since that will affect the sunlit statistics that are used by the method to perform compensation [112].

   The shadow and sunlit intensity means are given by μ_{shadow} and μ_{sunlit}, and the shadow and sunlit intensity standard deviations are given by σ_{shadow} and σ_{sunlit} in the following equation. This is the linear-correlation equation for adjusting shadow pixel intensity values through a simplified histogram matching where only the first and second statistical moments are employed, given by

$$y = \frac{\sigma_{\text{sunlit}}}{\sigma_{\text{shadow}}}\left(x - \mu_{\text{shadow}}\right) + \mu_{\text{sunlit}}. \tag{2.9}$$

5. **Piecewise linear-correlation**: A prominent concern of linear-correlation is that only a single linear equation is used, which may cause notable variation in the corrected shadow areas. If the penumbra width is significant or if there are distinct shadow intensities, then it is better to use a piecewise linear equation [119].

   Liu and Yamazaki divide shadow objects into dark (bottom 5% DN), medium and light-shadow (top 5% DN) classes, where a, b and c are the dark, medium and light shadow thresholds, respectively [66]. The piecewise division of shadows is given as follows by the modification coefficient Θ representing the darkness of a shadow, which has a range of $[-1,1]$:

$$\Theta = \begin{cases} \frac{x}{a} & \text{if} \quad 0 < x \leq a \\ 1 & \text{if} \quad a < x \leq b \\ \frac{b+c-2x}{c-b} & \text{if} \quad b < x < c. \end{cases} \tag{2.10}$$

The piecewise linear-correlation correction is then produced by the following equation, where r is the ratio of the radiances in the shadow and sunlit areas:

$$y = \Theta \cdot \frac{1}{r}(x - \mu_{\text{shadow}}) + \mu_{\text{sunlit}}. \tag{2.11}$$

6. **Supervised chromatic correction**: Makarau et al. used a supervised approach to compute chromaticity, using manual selection of shadow and sunlit pixels across a shadow edge as training examples [105].

2.5.2.2 Intrinsic domain

The intrinsic domain of shadow restoration methods refers to the decomposition of shaded images into two integral images, namely a shading image and a reflectance image. The dual purposes of intrinsic decomposition have been mentioned in the previous section on supervised shadow detection methods [90]. The first function of integral images is to provide a shadow mask, and the second function is to obtain a shadow-free reflectance image, which is the function explored in this set of methods.

Examples of shadow removal methods that belong in the intrinsic domain are given as follows:

1. **Entropy minimization**: Finlayson et al. [115] derive illumination invariant images through entropy minimization, without the need for calibration. They seek a projection through entropy minimization that results in a reflectance-information only image that is independent of lightning.

2. **Product model**: Li et al. describe a second mode of intensity domain shadow removal methods, where a shadowed image is defined as the product of a shadow-free image and a shadow scale [99]. A thin plate spline can be used to smooth the shadow scale, but the smoothness assumption does not hold in the case of compound shadows, which is prominent in dense urban areas [116]. This product model implicates intrinsic images and this method is consequently categorized as an intrinsic domain method.

3. **Poisson method**: Wu and Tang use an image model where the observed shadowed image is the shadowless image scaled by the shadow image [117]. They estimate the shadow and shadow-free PDFs with GMM from the corresponding histograms and solve the shadowed image via a Bayesian framework before estimating the shadowless image with a Poisson equation. This is also an example of a product model and intrinsic domain method, but the use of a Poisson method differentiates this example.

2.5.2.3 Multisource methods

Dare demonstrated multisource data fusion for high-resolution RS imagery [60], but multisource data fusion has also been done in a low-resolution setting for the removal of cloud shadows by Wang et al. [67]. Multisource data fusion uses co-registered images and replaces shadow pixels in one image with co-registered pixels in the other image, if these pixels are classified as being in sunlit regions.

The typical sun-synchronicity of optical high-resolution satellites means that it will be difficult to find sunlit counterparts for all shadow pixels. However, local summer acquisitions have higher solar elevation and shorter cast shadows than winter acquisitions, so winter shadows can be replaced by summer sunlit counterpart pixels.

Multisource data fusion that uses aerial and satellite imagery requires more sophisticated radiometric conversions, because of the sensing differences between aerial and satellite imaging vehicles [60]. Accurate registration is also a requirement, but topographic variation exacerbates poor registration. Zhou et al. showed that multisource fusion can perform relatively well as a method for shadow detection [86].

The following chapter on shadow detection evaluates a selection of methods reviewed in this chapter in addition to methods not used for shadow detection before.

CHAPTER 3 SHADOW DETECTION

3.1 CHAPTER OVERVIEW

In the previous chapter a literature study was performed for shadow removal and shadow detection. Shadow detection was shown to be an important first step in shadow removal, and it was indicated that there are two main approaches to shadow detection, namely pixel and object-based methods. Image segmentation and thresholding are two important components of pixel and object-based shadow detection, so this chapter explores different categories of segmentation and thresholding. This chapter studies threshold-based shadow detection methods that can subsequently be used as part of the input modification strategy to reduce dataset shift, so it is placed under the input modification theme to address dataset shift, as shown in Figure 3.1.

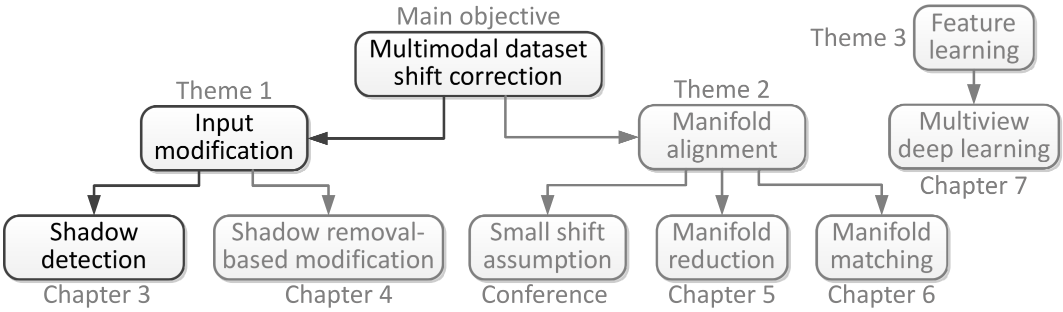


Figure 3.1. Indication of where this chapter fits into the thesis.

3.1.1 Contributions

1. Panchromatic shadow detection algorithms from the thresholding subcategory (Adeline et al. [59]) of property-based shadow detection (Arévalo et al. [73]) in Table 2.2 are reviewed on the Soweto panchromatic land-use dataset.

2. The shadow detection accuracy of unsupervised global thresholding methods are compared using the Czekanowski-Dice (F-score), Jaccard, Rand (overall), Rogers-Tanimoto and Sokal-Sneath external validation indices previously discussed in paragraph 2.4.4.2.

3. A qualitative and quantitative comparison is performed for 10 unsupervised global thresholding methods, but local adaptive thresholding (LAT) is also employed and its results are compared to those of global thresholding for a panchromatic land-use classification scenario.

4. LAT for panchromatic shadow detection is contributed to the known taxonomy of Adeline et al. [59] and shown to outperform global thresholding when threshold selection is supervised.

The problem statement follows in the next section where the hypotheses and research questions for this chapter are stated. Image segmentation is briefly reviewed and taxonomized in Section 3.3, where the general segmentation objective is described in terms of the gestalt laws of grouping and ideal segment properties. Image thresholding can typically be categorized as a region-based parallel type of image segmentation, but because of its importance in shadow detection a separate section is dedicated to image thresholding in Section 3.4 with a review of image thresholding taxonomy and descriptions of the main thresholding methods. Section 3.5 investigates global and LAT for panchromatic shadow detection with the Soweto dataset used to measure and compare shadow detection accuracies for various thresholding methods.

3.2 PROBLEM STATEMENT

Toward the goal of instantiating a method of input modification where shadows are removed before feature extraction, the various shadow detection options explored in the literature study of the previous chapter have to be evaluated.

3.2.1 Hypotheses

1. Threshold-based shadow detection can relatively accurately delineate shadows because of the low intensity property of shadows.
2. Locally adaptive thresholds detect shadows below a threshold relative to local image intensity, which should produce more accurate shadows than with a global fixed threshold, since relatively low intensity admits greater sensitivity in images with contrast variation than globally low intensity.

3.2.2 Research questions

1. How do the different threshold-based segmentations from the thresholding algorithm taxonomy of Sezgin and Sankur [47] compare in terms of panchromatic shadow detection accuracy?
2. How does global thresholding compare to locally adaptive thresholding in terms of panchromatic shadow detection accuracy?

3.3 IMAGE SEGMENTATION

Computer vision is generally a problem of inference, since it aims to determine the cause behind observed data [120]. One of the fundamental problems in computer vision is image segmentation, which is usually the first step of the process of image analysis. Image segmentation has been defined as "partitioning an image into several disjoint subsets such that each subset corresponds to a meaningful part of the image" [121]. Image segmentation is a key process in a wide range of applications in areas and problems such as medical image processing, remote sensing, recognition, object tracking and image reconstruction.

3.3.1 Segmentation objective

3.3.1.1 Gestalt laws of grouping

Two-dimensional image segmentation produces a number of spatially continuous regions or segments that normally display some pattern of coherence within. Wertheimer's *Gestalt theory* described a fundamental model of the perceptual clustering evident in the human perceptual system [122], which consisted of the following laws of grouping:

1. **Proximity**: Components that are closer together are perceived as a group.
2. **Similarity**: Components that share the same characteristics, such as color and texture, are perceived as a group.
3. **Closure**: A group with only a partial appearance can still be identified because of the tendency of perception to complete the appearance through a form of interpolation.
4. **Common fate**: Components that move together are perceived as a group.
5. **Good continuation**: Groups that overlap can be distinguished if there is a continuation of a characteristic (such as color) at the intersection.
6. **Good form**: Overlapping forms can be distinguished by differentiating them according to the characteristics of shape, pattern, color, etc.

3.3.1.2 Segment properties

Haralick and Shapiro [123] defined the following four properties generally desired for segments, which correspond partially to the concepts present in the *Gestalt laws of grouping*:

1. Uniform and homogeneous with respect to some characteristics
2. Simple interiors with strong regularity
3. Dissimilar adjacent regions
4. Simple and spatially accurate boundaries.

The specific application in which segmentation plays a role will influence the grouping objectives used. Therefore, Peng et al. [121] suggest that image segmentation should incorporate mid- and high-level knowledge of the application to obtain domain-specific segmentation. For object-based shadow detection the first step involves segmentation, and the unique properties of shadow, such as low intensity, can be used as a desired property as well.

3.3.2 Segmentation taxonomy

The categorization and classification of segmentation algorithms, methods and approaches are reviewed in this subsection. Various categorizations of segmentation methods can be found in the literature, such as the work of Tilton [124], who divided image segmentation approaches into three classes:

1. **Characteristic feature thresholding or clustering**, which does not usually exploit spatial information.
2. **Boundary detection**, which exploits spatial information but can suffer from incomplete edge detection on noisy images.
3. **Region extraction**, such as region growing, which can depend on the exact merging sequence followed.

Zhang [125] provides a general categorization of segmentation algorithms into four different categories, namely edge-based parallel, edge-based sequential, region-based parallel and region-based sequential. The methods are either sequential or parallel, as well as either edge-based or region-based, and specific examples of segmentation methods are given in Table 3.1 for each category. A third class of methods is noted in this thesis, namely graph-based segmentation, since graph-based approaches often incorporate both edge and region-based criteria.

3.3.2.1 Sequential and parallel

Gray-level image segmentation is generally based on the principles of similarity and discontinuity [126], and segmentation algorithms can be categorized into two types, namely sequential and parallel algorithms [127]. Sequential algorithms are characterized by the subsequent use of earlier information that is generated in the earlier stages of processing, which thus requires sequential steps where earlier information is used in later steps. Parallel algorithms base decisions on independent and simultaneous processing operations, such as histogram shape thresholding.

3.3.2.2 Edge-based and region-based

There is an additional type of categorization, namely whether a segmentation method is edge-based or region-based. The principle of similarity is used in region-based segmentation algorithms to form regions or segments where the constituent pixels or components have similar properties, such as intensity, hue or texture. Discontinuity can be considered the complement to similarity and is used in edge-based segmentation algorithms, which finds object contours explicitly to form segments [125].

3.4 IMAGE THRESHOLDING

One of the important segmentation methods that is frequently used in shadow detection is thresholding [60]. Image thresholding approaches are reviewed in this section with a specific focus

Table 3.1. Categories of segmentation algorithms, partly according to Zhang [125].

	Segmentation algorithm categorization	
	Sequential	**Parallel**
Edge-based	**Edge-based sequential**	**Edge-based parallel**
	Edge linking	Edge detection based
	Boundary following	Canny edge detection
	Dynamic programming for contouring	SUSAN operator
Region-based	**Region-based sequential**	**Region-based parallel**
	Multiresolution segmentation	Thresholding
	Region split and merge	Clustering
	Watershed segmentation	Histogram concavity analysis
	Region growing	Entropy minimization
	Statistical region merging	
Graph-based	**Graph-based sequential**	**Graph-based parallel**
	Felzenszwalb and Huttenlocher [128]	Wassenberg et al. [129]
	Graph-cut segmentation	Copty et al. [130]
	Urquhart [131]	Tilton [124]

on algorithms that will be subsequently tested in the shadow detection analysis performed in the next chapter. This section places thresholding in the context of shadow detection and then an overview of the categorization of thresholding algorithms is given. The main categories of thresholding are then reviewed in terms of their subcategories, with the main categories listed as:

1. Histogram shape thresholding
2. Clustering-based thresholding
3. Entropy-based thresholding
4. Attribute similarity thresholding
5. Spatial thresholding
6. Locally adaptive thresholding.

3.4.1 Thresholding for shadow detection

A segmentation produced through thresholding usually consists of only foreground and background segments, which is a type of binary segmentation. This segmentation does not differentiate between spatially separated segments of the same class, namely the foreground or background class, so there are essentially only two segments. A binary segmentation is thus a direct way of producing a shadow mask, which indicates all shadow regions in an image. However, the binary segmentation cannot be used for object-based approaches, since a more detailed segmentation is required so that individual objects can be characterized.

Thresholding requires a global threshold T or a local threshold $T(x,y)$, and an image binarization operation can then be performed for input image I as described below. For a global threshold T the pixels (x,y) of binary image I_b are defined as

$$I_b(x,y) = \begin{cases} 1 & \text{if } I(x,y) \leq T \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

For a local threshold $T(x,y)$ the pixel (x,y) of binary image I_b is defined as

$$I_b(x,y) = \begin{cases} 1 & \text{if } I(x,y) \leq T(x,y) \\ 0 & \text{otherwise.} \end{cases} \tag{3.2}$$

3.4.2 Thresholding taxonomy

Segmentation approaches that are often used are the region-based methods of thresholding and histogram concavity analysis, which are categorized in Table 3.1 in the previous section. The actual thresholding step itself can be considered as fully parallel, although the threshold decision usually involves more of a sequential type of processing. Histogram concavity analysis is a specific approach to determining which threshold to use, and it is a prominent approach in thresholding.

A number of thresholding based methods have been developed specifically for the separation of objects from the background, such as the separation of shadow from sunlit regions [132], [77]. In this section the foreground is shadow and the background is sunlit regions. Saha and Ray divided thresholding techniques into local and global types [133]. Sezgin and Sankur reviewed image thresholding techniques, which they characterized according to types of information used [47]. A categorization of thresholding is shown in Table 3.2, which is based on the taxonomy of Sezgin and Sankur [47].

3.4.3 Thresholding algorithm notation

In the following subsections the different categories of thresholding will be reviewed and some of the algorithm descriptions will entail the following basic notations: $p(g)$ is the normalized histogram or probability mass function (PMF) for an input image I with intensity range $g = g_{\min} \cdots g_{\max}$, where $g_{\max} \leq G$ is the maximum intensity value. The cumulative probability function associated with $p(g)$ is denoted by

$$P(g) = \sum_{i=0}^{g} p(i). \tag{3.3}$$

3.4.4 Histogram shape thresholding

In histogram shape thresholding the shape properties of the histogram are used in different forms to calculate an optimal threshold. The following types of histogram shape thresholding methods are considered in this subsection:

1. Convex hull thresholding
2. Peak-and-valley thresholding

Table 3.2. Image thresholding algorithm categorization, based on the taxonomy of Sezgin and Sankur [47].

Image thresholding categorization	
Histogram shape thresholding	
Convex hull thresholding	Rosenfeld's method [134]
Peak-and-valley thresholding	Sezan [135]
Shape-modeling thresholding	Ramesh et al. [136]
Bimodal mean thresholding	Prewitt and Mendelsohn [137]
First valley thresholding	Prewitt and Mendelsohn [137]
Clustering-based thresholding	
Iterative thresholding	Ridler and Calvard [138]
Clustering thresholding	Otsu's method [77]
Minimum error thresholding	Kittler and Illingworth [139]
Iterative minimum error thresholding	Kittler and Illingworth [139]
Fuzzy clustering thresholding	Jawahar et al. [140]
Entropy-based thresholding	
Entropic thresholding	Kapur et al. [141]
Cross-entropic thresholding	Li and Lee [142]
Fuzzy entropic thresholding	Shanbag [143]
Attribute similarity thresholding	
Moment preserving thresholding	Tsai [144]
Edge field matching thresholding	Hertz and Schafer [145]
Fuzzy similarity thresholding	Huang and Wang [146]
Topological stable-state thresholding	Pikaz and Averbuch [147]
Maximum information thresholding	Leung and Lam [148]
Enhancement of fuzzy compactness thresholding	Pal and Rosenfeld [149]
Spatial thresholding	
Co-occurrence thresholding methods	Pal and Pal [150]
Higher-order entropy thresholding	Abutaleb [151]
Thresholding based on random sets	Friel and Molchanov [152]
2D fuzzy partitioning	Cheng and Chen [153]
Locally adaptive thresholding	
Local variance methods	Niblack [154] Sauvola and Pietikäinen [155] Wellner [156]
Local contrast methods	White and Rohrer [157] Bernsen [158]
Center-surround schemes	Kamel and Zhao [159]
Surface-fitting thresholding	Yanowitz and Bruckstein [160]

3. Shape-modeling thresholding
4. Bimodal mean thresholding
5. First valley thresholding.

3.4.4.1 Convex hull thresholding

Rosenfeld's method selects as threshold one of the deepest concavity points on the histogram, by first calculating the convex hull of the histogram [134]. The optimal threshold T_{opt} according to Rosenfeld's method is given by

$$T_{\text{opt}} = \underset{g}{\operatorname{argmax}} \{\text{Hull}\,(p(g)) - p(g)\}\,. \tag{3.4}$$

Hull$(p(g))$ is the convex hull of the normalized histogram or probability mass function $p(g)$ for $g = g_{\min} \dots g_{\max}$, where $g_{\max} \leq G$ is the maximum intensity value. Hull$(p(g))$ can also be described as the smallest convex polygon that contains $p(g)$.

The goal of Rosenfeld's method is to find the concavities of H, which are the connected components of the set-theoretic difference Hull$(p(g)) - p(g)$. The PMF denoted by $p(g)$ is already bounded on the left, right and bottom by $\overline{(g_{\min},0)(g_{\min},p(g_{\min}))}$, $\overline{(g_{\min},0)(g_{\max},0)}$ and $\overline{(g_{\max},0)(g_{\max},p(g_{\max}))}$, respectively. In this notation an edge from point (a,b) to (c,d) on the (x,y)-plane is given by $\overline{(a,b)(c,d)}$.

To construct the top part of Hull$(p(g))$ we can use the Rutovitz algorithm for in-line generation of a convex cover [161]:

1. Starting at point $(k, p(k))$ for $k = g_{\min}$, compute the slopes $-90° < \Theta_i < 90°$ of line segments $\overline{(k, p(k))(i, p(i))}$ for $k+1 \leq i \leq g_{\max}$.

2. Find the rightmost point $k_1 = \text{argmax}_i\{\Theta_i\}$ having the slope $\max\{\Theta_i\}$ and let $\overline{(k, p(k))(k_1, p(k_1))}$ form a side of Hull$(p(g))$.

3. Repeat the process by replacing k with k_1 and finding the slopes of line segments $\overline{(k_1, p(k_1))(i, p(i))}$ for $k_1 + 1 \leq i \leq g_{\max}$, until reaching $k_j = \text{argmax}_i\{\Theta_i\} = g_{\max}$. The method thus produces j top boundary lines of $p(g)$, which are combined with the bottom and side edges to form Hull$(p(g))$.

3.4.4.2 Peak-and-valley thresholding

Sezan reduces the histogram to a two-lobe function through convolution with a smoothing and differencing kernel [135]. The threshold is selected between the first terminating and second initiating zero crossing.

3.4.4.3 Shape-modeling thresholding

Ramesh et al. minimize the sum of square error between a bilevel functional approximation of $p(g)$ through an iterative search for the optimal threshold, which defines the mean foreground and background levels of the approximation [136].

3.4.4.4 Bimodal mean thresholding

Prewitt and Mendelsohn smooth the histogram $p(g)$ through an iterative three-point mean filtering until the PMF is bimodal [137]. Specifically, starting with $p_1(g) = p(g)$, a three-point mean filtering $p_{i+1}(g) = p_i(g) \star [1,1,1]/3$ is performed with a convolutional filter $[1,1,1]$ until $p_{i+1}(g)$ is bimodal. The global threshold is then set as the mean $T_{\text{opt}} = (T_0 + T_1)/2$ of the two peaks, where the two peaks can be written as

$$\{T_0, T_1\} = \arg_g \left\{ \left(\frac{dp(g)}{dg} = 0 \right) \wedge \left(\frac{d^2p(g)}{d^2g} < 0 \right) \right\}. \tag{3.5}$$

3.4.4.5 First valley thresholding

Prewitt and Mendelsohn also perform iterative smoothing [137] to ensure bimodality, after which the threshold is set as the minimum between the first two peaks, i.e.

$$T_{\text{opt}} = \underset{g}{\operatorname{argmin}} \{T_0 < g < T_1\}. \tag{3.6}$$

3.4.5 Clustering-based thresholding

Clustering analysis can also be used to determine optimal thresholds, but the number of clusters is set to two to ensure a binary segmentation. The distinction can then be made between two clusters, namely shadow and sunlit regions. Examples of clustering-based thresholding that are considered in this subsection include the following:

1. Iterative thresholding
2. Clustering thresholding
3. Minimum error thresholding
4. Iterative minimum error thresholding
5. Fuzzy clustering thresholding.

3.4.5.1 Iterative thresholding

Ridler and Calvard use two-class Gaussian mixture models in an iterative scheme, where at iteration i a threshold T_i is set as the average of the foreground and background class means [138]. The termination condition is based on a sufficiently small $|T_i - T_{i+1}|$. The optimal threshold is then defined as

$$T_{\text{opt}} = \lim_{i \to \infty} \frac{m_f(T_i) + m_b(T_i)}{2} \tag{3.7}$$

with the foreground class mean given by

$$m_f(T_i) = \sum_{g=0}^{T_i} g \cdot p(g) \tag{3.8}$$

and the background class mean given by

$$m_b(T_i) = \sum_{g=T_i+1}^{G} g \cdot p(g). \tag{3.9}$$

3.4.5.2 Clustering thresholding

Otsu's method determines the threshold that minimizes the weighted sum of within-class variances of the foreground and background pixels, which is equivalent to maximizing the between-class scatter [77]. Otsu's method is one of the most popular global thresholding algorithms, but does not work well for images with significant overlap between the histograms of foreground objects and the background due to poor illumination [133].

Otsu's threshold can be defined in terms of the cumulative probability function $P(g)$, the foreground class mean $m_f(g)$ given in Equation 3.8, the background class mean $m_b(g)$ given in Equation 3.9 and the foreground and background region variances $\sigma_f^2(g)$ and $\sigma_g^2(g)$, as follows:

$$T_{\text{opt}} = \underset{g}{\text{argmax}} \left\{ \frac{P(g)(1-P(g))(m_f(g)-m_b(g))^2}{P(g)\sigma_f^2(g)+(1-P(g))\sigma_b^2(g)} \right\}. \tag{3.10}$$

The variance of the foreground region is given by

$$\sigma_f^2(T) = \sum_{g=0}^{T} (g - m_f(T))^2 p(g) \tag{3.11}$$

and the variance of the background region is given by

$$\sigma_b^2(T) = \sum_{g=T+1}^{G} (g - m_b(T))^2 p(g). \tag{3.12}$$

3.4.5.3 Minimum error thresholding

Kittler and Illingworth characterizes an image by a mixture distribution of foreground and background pixels [139]. They solve a minimum-error Gaussian density-fitting problem, without assuming equal variances, to produce the following threshold

$$\begin{aligned} T_{\text{opt}} = \underset{g}{\text{argmin}} \Big\{ & P(g)\log(\sigma_f(g)) + (1-P(g))\log(\sigma_b(g)) \\ & - P(g)\log(P(g)) - (1-P(g))\log(1-P(g)) \Big\}. \end{aligned} \tag{3.13}$$

3.4.5.4 Iterative minimum error thresholding

Kittler and Illingworth also proposed an iterative version of their algorithm in addition to the direct algorithm of minimum error thresholding [139]. The iterative implementation of minimum error thresholding defines a decision rule that requires a solution to the following quadratic equation:

$$g^2\left(\frac{1}{\sigma_f^2(T)}-\frac{1}{\sigma_b^2(T)}\right)-2g\left(\frac{\mu_f(T)}{\sigma_f^2(T)}-\frac{\mu_b(T)}{\sigma_b^2(T)}\right)+\left(\frac{\mu_f^2(T)}{\sigma_f^2(T)}-\frac{\mu_b^2(T)}{\sigma_b^2(T)}\right)$$
$$+2\left(\log(\sigma_f(T))-\log(\sigma_b(T))\right)-2\left(\log(P(T))-\log(1-P(T))\right)=0. \quad (3.14)$$

The iterative algorithm converges to the optimal threshold with the use of the following procedure:

1. Choose the initial threshold T as the average intensity value.
2. Compute μ_f, μ_b, σ_f, σ_b and $P(T)$ with the current T value.
3. Solve Equation 3.14 with the computed terms, and set the new T with the solution.
4. Repeat from step 2 until the new T value is unchanged.

3.4.5.5 Fuzzy clustering thresholding

Jawahar et al. assign fuzzy clustering memberships to pixels according to pixel value differences to the class means [140]. The threshold is established as the crossover point of the membership functions.

3.4.6 Entropy-based thresholding

In entropy-based thresholding a maximum information transfer to the thresholded image is achieved through maximization of its entropy. Alternatively, cross-entropy can be minimized between the input and output images to preserve information. The following entropy-based thresholding examples will be reviewed in this subsection:

1. Entropic thresholding
2. Cross-entropic thresholding
3. Fuzzy entropic thresholding.

3.4.6.1 Entropic thresholding

Kapur et al. [141] optimized the threshold by maximizing the sum of the foreground and background entropies, i.e.

$$T_{\text{opt}}=\underset{g}{\operatorname{argmax}}\left\{H_f(g)+H_b(g)\right\} \quad (3.15)$$

with the foreground and background entropies defined as

$$H_f(T) = -\sum_{g=0}^{T} \frac{p(g)}{P(T)} \log\left(\frac{p(g)}{P(T)}\right) \quad \text{and} \tag{3.16}$$

$$H_b(T) = -\sum_{g=T+1}^{G} \frac{p(g)}{1-P(T)} \log\left(\frac{p(g)}{1-P(T)}\right). \tag{3.17}$$

3.4.6.2 Cross-entropic thresholding

Li and Lee calculate the optimal threshold as the one that minimizes the information theoretic distance between the input and thresholded images, specifically by using the Kullback-Leibler distance [142].

3.4.6.3 Fuzzy entropic thresholding

Shanbag finds the optimum as the threshold that minimizes the sum of the fuzzy membership entropies [143].

3.4.7 Attribute similarity thresholding

In attribute similarity thresholding a threshold is chosen to preserve a similarity measure or attribute quality between the input image and thresholded image. The types of attribute similarity thresholding methods that will be reviewed in this subsection are:

1. Moment preserving thresholding
2. Edge field matching thresholding
3. Fuzzy similarity thresholding
4. Topological stable-state thresholding
5. Maximum information thresholding
6. Enhancement of fuzzy compactness thresholding.

3.4.7.1 Moment preserving thresholding

Tsai modeled the gray-level input image as the blurred version of an ideal binary image [144]. The optimal threshold should produce a binary image with its first three statistical moments equal to the first three input image moments. The search for this moment equality can be formulated as

$$\begin{aligned} T_{\text{opt}} = \underset{T}{\operatorname{argmin}} \Bigg\{ & \Big(\sum_{g=0}^{G} p(g)g - \sum_{g=0}^{T} p(g) \sum_{g=0}^{T} gp(g) - \sum_{g=T+1}^{G} p(g) \sum_{g=T+1}^{G} gp(g)\Big)^2 \\ & + \Big(\sum_{g=0}^{G} p(g)g^2 - \sum_{g=0}^{T} p(g) \sum_{g=0}^{T} g^2 p(g) - \sum_{g=T+1}^{G} p(g) \sum_{g=T+1}^{G} g^2 p(g)\Big)^2 \end{aligned}$$

$$+\Big(\sum_{g=0}^{G} p(g)g^3 - \sum_{g=0}^{T} p(g)\sum_{g=0}^{T} g^3 p(g) - \sum_{g=T+1}^{G} p(g)\sum_{g=T+1}^{G} g^3 p(g)\Big)^2\Bigg\}. \quad (3.18)$$

3.4.7.2 Edge field matching thresholding

Hertz and Schafer determine a global threshold as the value that maximizes the count of matching Sobel generated edges between the input grayscale image and thresholded image [145].

3.4.7.3 Fuzzy similarity thresholding

Huang and Wang find the threshold that minimizes the entropy of pixel fuzziness membership values with foreground and background classes [146].

3.4.7.4 Topological stable-state thresholding

Pikaz and Averbuch perform threshold selection based on the stability of object shapes and edges [147]. They measure the number of foreground objects or connected components with a minimum number of pixels and select the threshold at the point of maximum stability of this cardinality measure.

3.4.7.5 Maximum information thresholding

Leung and Lam established the optimum threshold as the one that minimizes the average residual pixel class uncertainty or maximizes the decrease in uncertainty after the thresholded image has been observed [148].

3.4.7.6 Enhancement of fuzzy compactness thresholding

Pal and Rosenfeld choose the threshold that maximizes the compactness of the connected components in the foreground set, where compactness is defined as the ratio of object area to the squared perimeter [149].

3.4.8 Spatial thresholding

Spatial thresholding generally uses the pixel neighborhood information for context and to calculate correlations and dependence models. The following main examples of spatial thresholding are considered in this subsection:

1. Co-occurrence thresholding methods
2. Higher-order entropy thresholding
3. Thresholding based on random sets
4. 2D fuzzy partitioning.

3.4.8.1 Co-occurrence thresholding methods

Pal and Pal defined the optimal threshold as the threshold producing the maximum number of background-to-foreground and foreground-to-background transitions, specifically by using the entropies of the co-occurrence probabilities of gray values in the foreground and background [150].

3.4.8.2 Higher-order entropy thresholding

Abutaleb selects as optimum threshold the specific threshold that minimizes the entropy of the co-occurrence histogram of pixel values and pixel neighborhood means [151].

3.4.8.3 Thresholding based on random sets

Friel and Molchanov produce a distribution of a random set from the input image and choose a threshold that generates a foreground with pixels possessing similarity in terms of their Chamfer distance [152], [162].

3.4.8.4 2D fuzzy partitioning

Cheng and Chen combine fuzzy entropy and local pixel co-occurrence histograms to find the threshold that maximizes the sum of foreground and background entropies at the crossover point of largest fuzziness [153].

3.4.9 Locally adaptive thresholding

Thresholds are calculated for each pixel using the local statistics like range and the variance of the pixel neighborhood. Local thresholding is superior to global thresholding for poorly and unevenly illuminated images [133]. The following types of locally adaptive thresholding are considered in this subsection:

1. Local variance methods
2. Local contrast methods
3. Center-surround schemes
4. Surface-fitting thresholding.

3.4.9.1 Local variance methods

Niblack defines a pixel threshold as the sum of the local pixel mean and a scaled version of the standard deviation for a given local window size [154]. For a given mean intensity $m(x,y)$ and standard deviation $\sigma(x,y)$ in a local window centered at (x,y), Niblack defines a threshold with a typical parameter $k = -0.2$ as

$$T(x,y) = m(x,y) + k \cdot \sigma(x,y). \tag{3.19}$$

The local window mean and standard deviation around pixel (x,y) are defined for a square window width w as

$$m(x,y) = \sum_{i=x-\frac{w-1}{2}}^{x+\frac{w-1}{2}} \sum_{j=y-\frac{w-1}{2}}^{y+\frac{w-1}{2}} \frac{I(i,j)}{w^2} \tag{3.20}$$

$$\sigma(x,y) = \sqrt{\frac{1}{N-1} \sum_{i=x-\frac{w-1}{2}}^{x+\frac{w-1}{2}} \sum_{j=y-\frac{w-1}{2}}^{y+\frac{w-1}{2}} (I(i,j) - m(x,y))^2}. \tag{3.21}$$

Sauvola and Pietikäinen improved on Niblack's threshold by adapting the scaling of the standard deviation [155]. Their threshold incorporates an extra parameter R with a typical value of $R = 128$, but parameter k now has a typical value of $k = 0.2$. The threshold is defined as

$$T(x,y) = m(x,y) + 1 + k\left(\sigma(x,y)/R - 1\right). \tag{3.22}$$

Wellner sets an adaptive threshold in terms of the local window mean [156], which is defined for a typical parameter $k = 15$ as

$$T(x,y) = m(x,y)(1 - k/100). \tag{3.23}$$

In a binarization algorithm comparison for the thresholding of 3D X-ray microtomographies of trabecular bone [163] it was shown that Wellner outperforms the local adaptive thresholds of Niblack [154] and Sauvola and Pietikäinen [155].

3.4.9.2 Local contrast methods

White and Rohrer set the threshold as a scaled version of the local neighborhood pixel mean [157]. The threshold is defined for a typical parameter $k = 1.1$ as follows

$$T(x,y) = m(x,y)/k. \tag{3.24}$$

Bernsen sets the local threshold as the mean of the minimum and maximum pixel values in the local neighborhood, provided that the difference between the extremes is large enough [158].

3.4.9.3 Center-surround schemes

Kamel and Zhao proposed a threshold calculation optimized for scenarios with well-defined objects such as text, since their method compares the average gray value in blocks proportional to the object with the surrounding neighborhood [159].

3.4.9.4 Surface-fitting thresholding

Yanowitz and Bruckstein combined edge and gray-level information to render a threshold surface, which can be used to obtain the optimal threshold [160].

3.5 SHADOW DETECTION

3.5.1 Input modification

The input modification strategy to address dataset shift involves the identification of isolatable input components that are related to the dataset shift. The reasoning behind input modification is that the presence of these input components, which are caused by varying measurement modes, can cause dataset shift in a classifier. So the hypothesis is that the removal of these varying components may reduce subsequent dataset shift, since a contributing factor to dataset shift is no longer present.

A major cause of dataset shift in remote sensing is the effect of seasonal changes on solar elevation, which introduces a subsequent variation in lighting geometry. A potentially strong factor of visible change in the remote sensing imagery is shadow, but its contrastive appearance may also allow for identification and isolation of affected image regions. There is a potentially wide variance of shadow profiles, especially in across-seasonal image pairs.

3.5.2 Shadow detection approaches

The identification and removal of shadows have the potential to improve subsequent feature constancy, which can improve land-use classification accuracy. The identification step initiates the process of shadow removal and it involves shadow detection. Panchromatic shadow detection algorithms are chosen from the thresholding subcategory (Adeline et al. [59]) of property-based shadow detection (Arévalo et al. [73]) in Table 2.2, for reasons of relatively low computational time requirements and since the other subcategories of property-based shadow detection are not as suitable, namely invariant color models, object-based algorithms and machine learning. Threshold-based shadow detection strategy is based on a binary classification of shadow or sunlit area, namely pixel-based shadow detection that classifies individual pixels as either shadow or sunlit.

Threshold-based segmentation and shadow detection explored in this chapter can be directly related to the bimodal histogram splitting used by Dare [60] and Wei et al. [76], the number of peaks and valleys used by Chen et al. [78], the first valley detection used by Liu and Yamazaki [66], and the first peak classification used by Wei et al. [76]. Several other important thresholding algorithms from Table 3.2 are also tested to expand the set of threshold-based shadow detectors found in the literature.

3.5.3 Data description

A 4.85×9.86 km^2 section of the subtropical highland of Soweto (Gauteng, South Africa) was selected as the study site, as shown in Figure 3.2. Panchromatic QuickBird imagery of the site was captured at a 0.6×0.6 m^2 pixel resolution, with a nominal 30° off-nadir wide accessible ground swath, on 18 October 2005 (d_1, early summer, rainy season) and 30 May 2006 (d_2, early winter). The across-date settlement classification is investigated for these two dates denoted by d_1 and d_2, of which QuickBird acquisitions are shown in Figure 3.3.

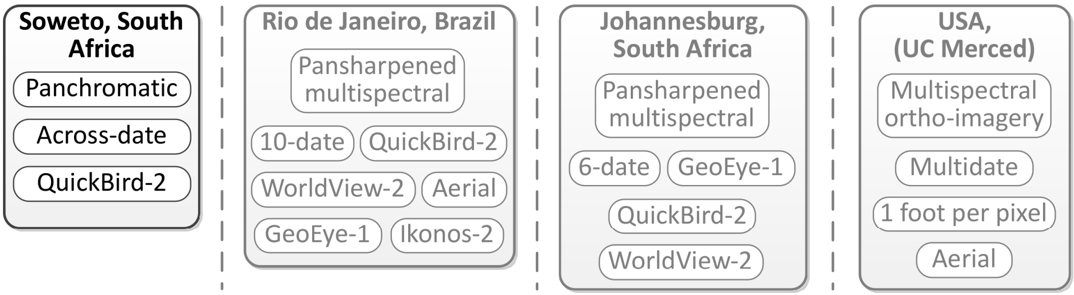


Figure 3.2. Soweto across-date dataset selection for this experiment.

The settlement classifier is evaluated in a study area with three main settlement types, as shown in Figure 3.3, namely formal settlements (FS), formal settlements with backyard shacks (FSB) and ordered informal settlements (OIS). NBU is the fourth class and includes natural vegetation; it is added to test classifier separability between builtup and non-builtup areas. FS are characterized by permanent residential structures that are positioned in a planned manner, while FSB have larger residential structures accompanied by smaller backyard shacks. An OIS is constituted when permanent and semi-permanent residential structures are arranged in a planned manner.

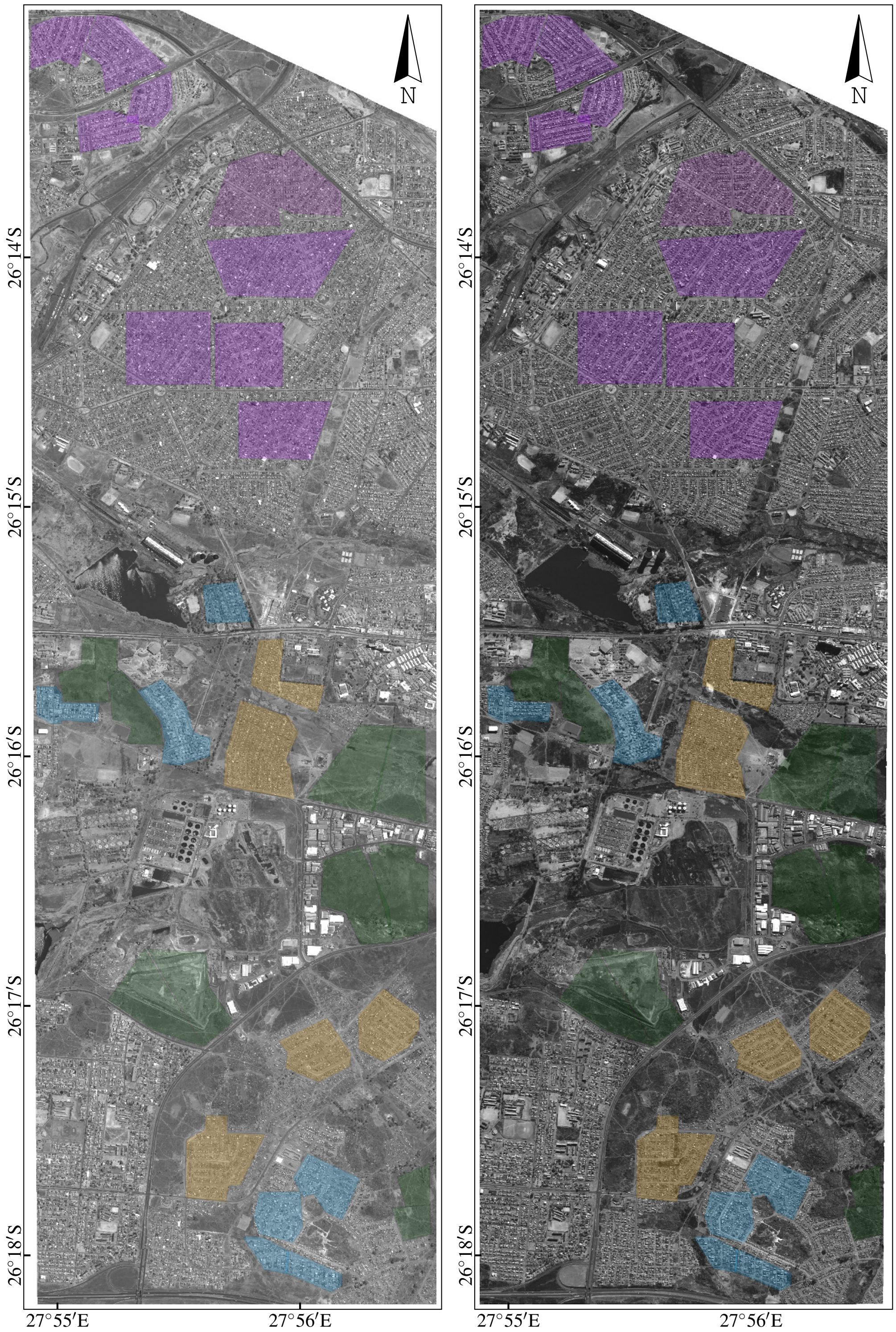
3.5.4 Global thresholding

3.5.4.1 Shadow detection groundtruth

Groundtruth shadow masks were created for the Soweto dataset, featuring representative coverage over FS, FSB and OIS polygons for both dates in the dataset. Appendix A contains the groundtruth shadow masks for the Soweto dataset, where shadows are indicated by red regions. In Figure 3.4 a groundtruth shadow mask is shown for a formal settlement polygon for the first date of the Soweto dataset.

3.5.4.2 Shadow detection accuracy

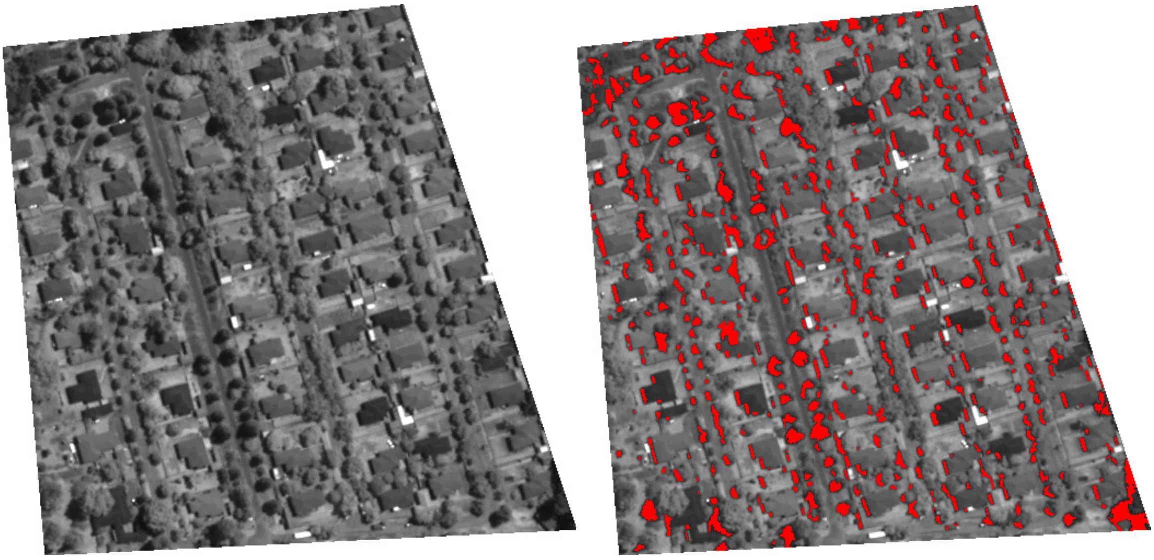
For global thresholding the same threshold is used to classify each pixel as shadow or sunlit, and the accuracy of the global thresholding for a given threshold can be determined with an external validation index and a groundtruth shadow mask. The global thresholding accuracy is illustrated in Figure 3.5, mainly in terms of the Czekanowski-Dice external validation index, which is equivalent to the F-score.



(a) QuickBird-2 (18 October 2005)

(b) QuickBird-2 (30 May 2006)

Figure 3.3. Across-date acquisitions of the study area of Soweto, with polygon selections of the various land-use classes. Class polygons outlined include (🟪) FSB, (🟦) FS, (🟫) OIS, and (🟩) NBU. Panchromatic background images courtesy of DigitalGlobe™.



(a) FS - Date 1 (Soweto).

(b) Groundtruth shadows in red (FS - Date 1).

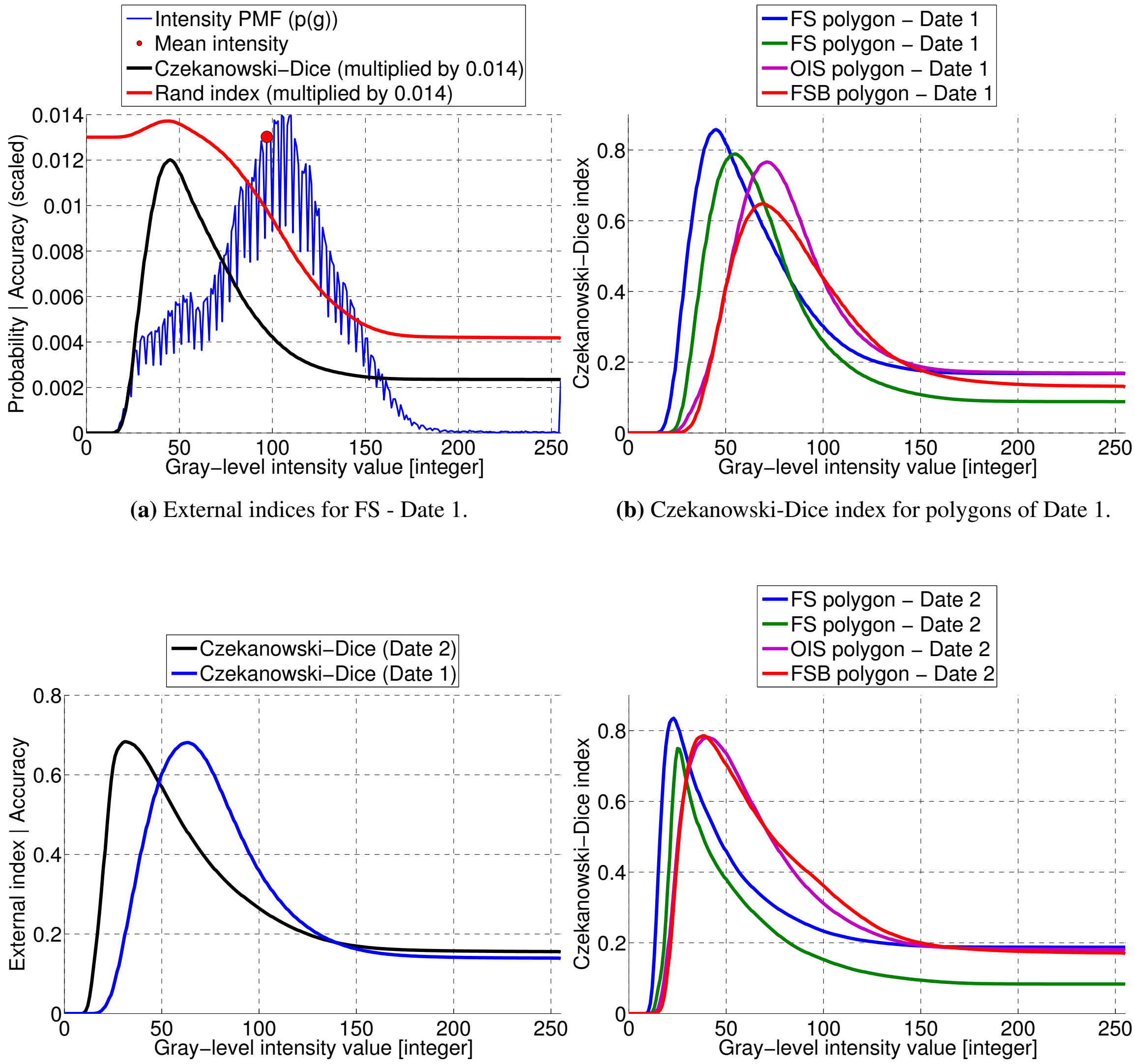
Figure 3.4. Groundtruth shadows for a co-registered formal settlement polygon over the two acquisitions of the Soweto dataset. Panchromatic background images courtesy of DigitalGlobe™.

Figures 3.5(b) and 3.5(d) show the Czekanowski-Dice index values of global thresholding for the representative groundtruth settlement types for both acquisitions of the Soweto dataset. Figure 3.5(c) is the averaged Czekanowski-Dice index values for both dates of the dataset, and it is observed that the shadows on date 2 have a lower intensity than on date 1. This is probably due to the reduced irradiance in winter for date 2 than in summer for date 1, because of the lower solar elevation during winter. The reduced irradiance also cause lower pixel intensities and reduced shadow intensities.

3.5.4.3 Rand index imbalance

Figure 3.5(a) compares the Czekanowski-Dice index with the Rand index, and it can be seen that the Czekanowski-Dice index penalizes very low shadow probabilities where the Rand index does not. The Rand index or overall accuracy is formulated as $(TP+TN)/(TP+TN+FP+FN)$, which becomes imbalanced for a small prior probability of shadow occurrence, since the proportion of true negative samples will be relatively large. A large value for TN can then be obtained even with an empty shadow mask that contains no shadow occurrences, which still results in a high overall accuracy.

The Rand index or overall accuracy is thus not ideal for measuring shadow detection accuracy, since it contains the true negative term in the numerator. The McNemar, Phi, Rogers-Tanimoto and Sokal-Sneath external indices also contain the true negative term in the numerator and are therefore also not the most suitable accuracy measures for shadow detection.



(a) External indices for FS - Date 1.

(b) Czekanowski-Dice index for polygons of Date 1.

(c) Mean Czekanowski-Dice indices for both dates.

(d) Czekanowski-Dice index for polygons of Date 2.

Figure 3.5. Global thresholding accuracy measure of shadow detection with groundtruth shadow masks for the Soweto dataset.

3.5.5 Unsupervised global thresholding

3.5.5.1 Shadow detection accuracy

The unsupervised global thresholding methods are differentiated only by how well they can choose binarization thresholds, since the shadow detection accuracy depends only on the global threshold. Two thresholding selection methods that choose the same threshold will thus have the same shadow detection accuracy for global thresholding. The goal of the unsupervised thresholding method is to separate the shadow density from the sunlit density with minimum error in an applicable intensity range.

The shadow detection accuracy of unsupervised global thresholding methods is compared in Table 3.3, using the Czekanowski-Dice (F-score), Jaccard, Rand (overall), Rogers-Tanimoto and Sokal-Sneath external validation indices previously discussed in paragraph 2.4.4.2. The best thresholds are found at the maximum external validation index values and so the shadow detection accuracy of the external validation indices are measured as a percentage of the maximum external validation index values with separate consideration of each external validation index.

Table 3.3. Unsupervised global thresholding accuracy.

Method	**Threshold**	**Czekanowski-**	**Jaccard**	**Rand**	**Rogers-**	**Sokal-**
		Dice			**Tanimoto**	**Sneath**
Maximum index	0 ±0	0.777±0.1	0.638±0.1	0.98 ±0.1	0.961±0.0	0.473±0.1
		Percentage of maximum index value				
Minimum error	254 ±0	19.0 ±4.9	12.6 ±3.5	47.4 ±13	32.4 ±12	8.95 ±2.6
Entropic	156 ±35	22.6 ±2.2	15.3 ±1.8	59.1 ±19	44.5 ±19	11.0 ±1.7
Bimodal mean	151 ±53	27.4 ±15	19.6 ±13	62.2 ±21	48.6 ±23	14.7 ±11
Moment pres.	113 ±17	30.1 ±5.9	20.9 ±4.4	71.7 ±7.3	56.8 ±8.5	15.3 ±3.5
First valley	187 ±94	33.9 ±27	26.8 ±26	63.4 ±25	51.5 ±30	22.0 ±25
Mean	104 ±18	35.0 ±8.4	24.9 ±6.7	77.0 ±5.7	63.4 ±7.4	18.6 ±5.4
Iterative	104 ±18	35.1 ±8.7	25.1 ±6.9	77.1 ±6.2	63.5 ±7.8	18.7 ±5.5
Clustering	104 ±18	35.2 ±8.9	25.1 ±7	77.1 ±6.3	63.6 ±7.9	18.7 ±5.5
Convex hull	68.8±22	69.8 ±25	61.9 ±29	93.2 ±5.8	87.9 ±9.8	55.5 ±32
Iter. min. error	59.3±29	80.1 ±27	74.9 ±30	95.8 ±6.7	92.6 ±11	70.2 ±31

The Czekanowski-Dice (F-score), Jaccard and Sokal-Sneath indices are more appropriate for shadow detection accuracy measurement, since these indices omit the true negative term that causes poor accuracy measures when the positive shadow class has a low prior probability. The following shadow detection accuracy observations can be made from Table 3.3 for the various unsupervised global thresholding methods:

- **Minimum error thresholding** [139] (paragraph 3.4.5.3) performs poorest, since the shadow and sunlit densities overlap too much for the intended minimum Bayes error to be detected.
- **Entropic thresholding** [141] (paragraph 3.4.6.1) performs poorly, since it tends to select the right-tail boundary, which is normally at a relatively high intensity.
- **Bimodal mean thresholding** [137] (paragraph 3.4.4.4) is not robust because of the occasional occurrence of the second mode at a high intensity, which causes misinterpretation when the shadow mode is not present.
- **Moment preservation** [144] (paragraph 3.4.7.1) finds a robust threshold, but does not consider the shadow class specifically so the thresholds are generally larger than the optimum.

- **First valley thresholding** [137] (paragraph 3.4.4.5) suffers from the same bimodal misrepresentation that misses the shadow mode, so the results are not robust and depend on the correct bimodal representation.

- **Mean**, **iterative** [138] (paragraph 3.4.5.1) and **clustering thresholding** [77] (paragraph 3.4.5.2) all find thresholds that are generally too large for shadow detection, since the shadow mode specifically is not regarded in these methods, as they aim to find general image binarization thresholds.

- **Convex hull thresholding** [134] (paragraph 3.4.4.1) use concavities to define peak and valley locations rather than requiring explicit bimodality, which results in a more robust determination of the valley between the shadow and sunlit modes.

- **Iterative minimum error thresholding** [139] (paragraph 3.4.5.4) outperforms minimum error thresholding, since it can avoid boundary minimums and focus on a relevant intensity range in the minimum error search.

3.5.6 Local adaptive thresholding

3.5.6.1 Wellner's thresholding

The local variance method of Wellner was reviewed in paragraph 3.4.9.1 and it was stated that it outperformed the local adaptive thresholds of Niblack [154] and Sauvola and Pietikäinen [155] in the thresholding of 3D X-ray microtomographies of trabecular bone [163]. Wellner sets the adaptive threshold in terms of the local window mean [156], which is defined with an offset parameter k as

$$T(x,y) = m(x,y) - k \cdot m(x,y)/100. \tag{3.25}$$

Wellner's threshold is thus a fixed percentage of the local intensity mean $m(x,y)$ around the pixel at coordinate (x,y), with the local mean calculated in a square window of width w. White and Rohrer's threshold is a percentage of the local intensity mean as well [157], but Niblack [154] and Sauvola and Pietikäinen [155] in addition incorporate local standard deviation. This incorporation of local standard deviation may be problematic, for instance with Niblack's threshold

$$T(x,y) = m(x,y) + k \cdot \sigma(x,y). \tag{3.26}$$

In this case, if the local standard deviation is small then it may produce a threshold that is too large for shadow detection. The LAT algorithm used in this subsection is Wellner's threshold, since it does not require the use of local variance measures and has been found in other studies to perform best. Sales ranked the different LAT algorithms as follows, based on the thresholding of 3D X-ray microtomographies of trabecular bone [163]:

1. Wellner

2. Bernsen
3. White and Rohrer
4. Sauvola and Pietikäinen
5. Niblack.

3.5.6.2 Shadow detection accuracy

Wellner's thresholding method was used for LAT shadow detection on the Soweto dataset for both dates of the dataset. Shadow detection accuracy is measured with the Czekanowski-Dice external validation index, which is the same index that was used for the evaluation of global thresholding. Wellner's thresholding has two parameters, namely the window size w and the percentage parameter k. A large k value results in a low threshold and shadow probability, while a small k value results in a threshold close to the window mean and a higher shadow probability.

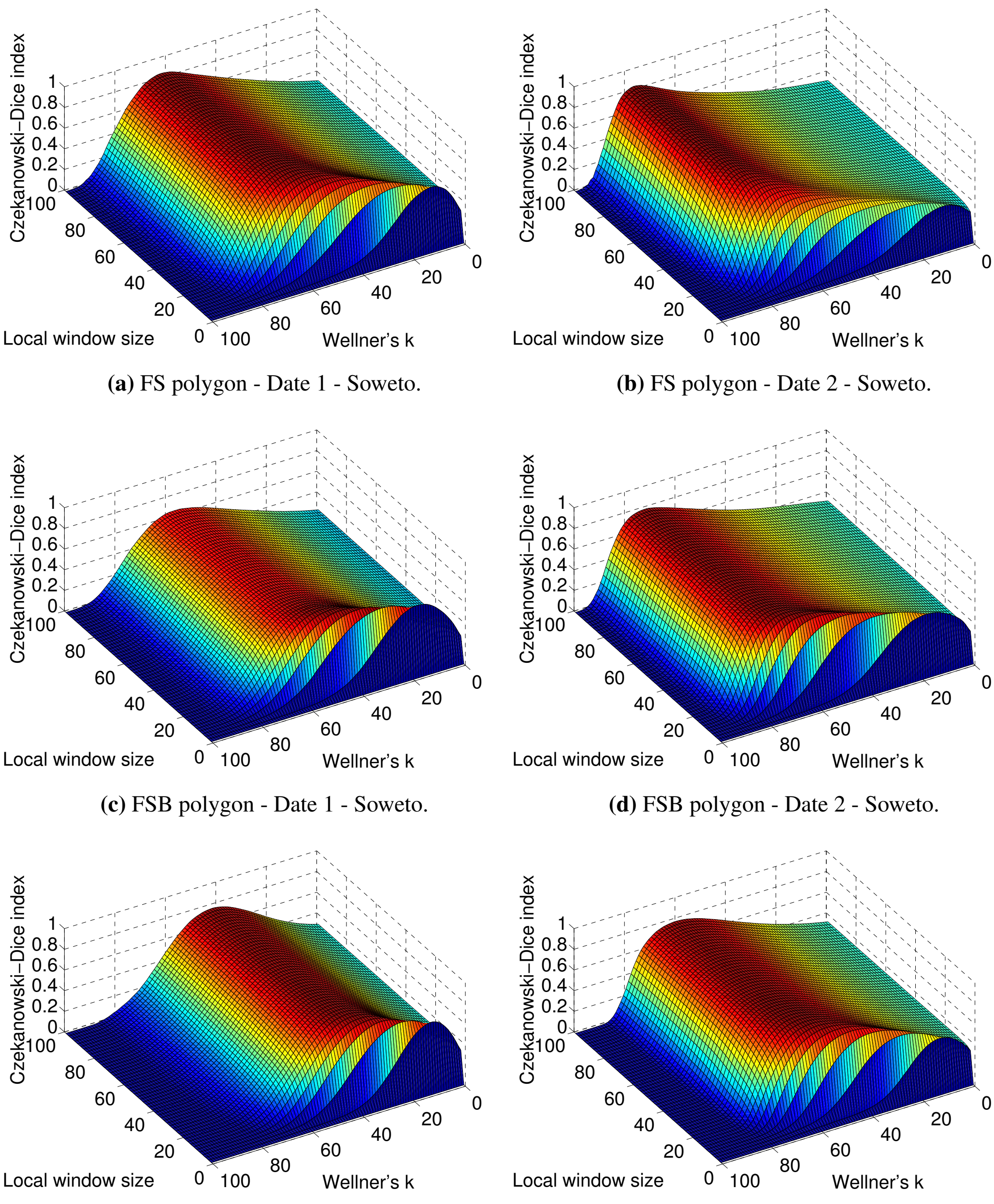
The results for Wellner's thresholding for the FS, FSB and OIS settlement types of both dates of the Soweto dataset are shown in Figure 3.6. The settlement type averaged external validation index values over the two dates of the Soweto dataset are given in Figure 3.7. It can be noted that Wellner's k parameter is larger for date 2 than date 1, which means that date 2 requires a lower shadow detection threshold than date 1. It can also be seen that the local window size has a greater effect in the one to 11 pixel width range, since the external validation index stabilizes above a window size of 11 pixels.

3.5.6.3 Unsupervised local adaptive thresholding

LAT algorithms generally have at least two parameters, namely the local window size and an offset parameter that defines a threshold in terms of the local window statistics. The local window size can be fixed for a type of imagery and the thresholding can perform well, provided that the local window size is not too small. Evidence for this statement can be seen in the Figure 3.7, where a local window size of 29 can be chosen and provide a mean stability in the high accuracy region of the two LAT parameters.

Using unsupervised global thresholding the input image histogram can be analyzed and a bimodal separation strategy can be used to find a good threshold. In the previous subsection it was shown that convex hull thresholding and iterative minimum error thresholding can provide relatively good global thresholds. These same methods can be used in LAT, since the only difference is that the histograms are determined over local square regions instead of the entire image.

Unsupervised threshold selection based on local histograms incurs a computational penalty, since the histograms have to be calculated around the local window for every pixel. The unsupervised threshold selection algorithm will also have to be run for each of the histograms. If the window is too small then the local histogram may not contain enough samples to define the shadow density, which will result in the unsupervised threshold selection algorithm not being able to function correctly.



(a) FS polygon - Date 1 - Soweto.

(b) FS polygon - Date 2 - Soweto.

(c) FSB polygon - Date 1 - Soweto.

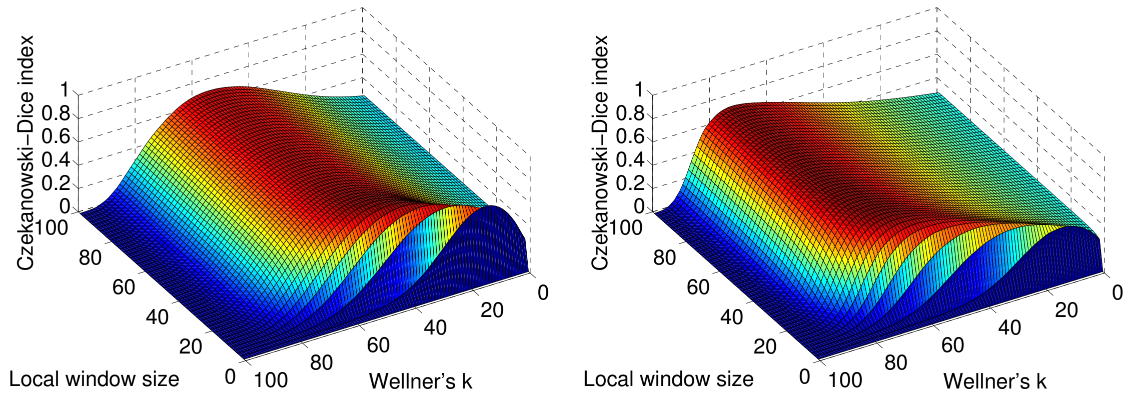
(d) FSB polygon - Date 2 - Soweto.

(e) OIS polygon - Date 1 - Soweto.

(f) OIS polygon - Date 2 - Soweto.

Figure 3.6. LAT accuracy of the different settlement types in terms of the Czekanowski-Dice index with the shadow groundtruth dataset.

For these reasons unsupervised LAT is not explored further in this thesis, but a supervised parameter decision is rather made on the shadow detection accuracy curves of the representative groundtruth shadow mask dataset, as shown in Figure 3.7. The local window size is set to a fixed size to eliminate further decisions for the window size parameter and to produce accuracy curves only in terms of Wellner's *k* value.



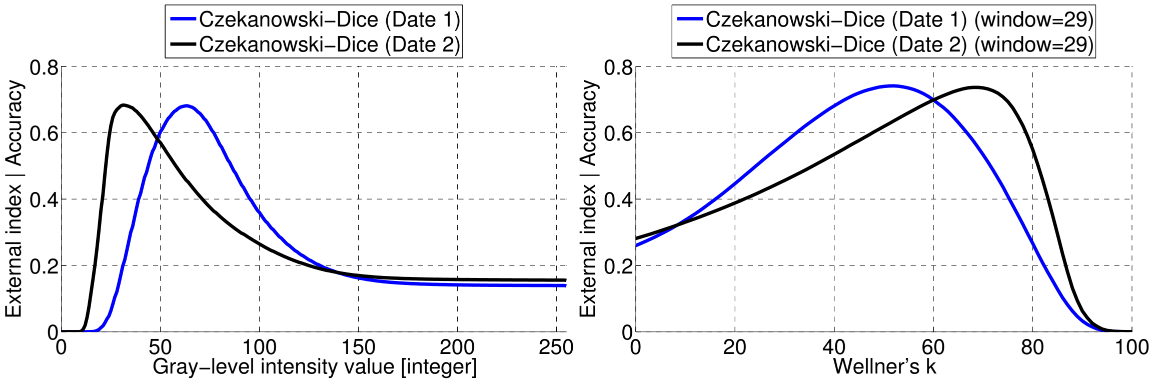
(a) Date 1 - Soweto. **(b)** Date 2 - Soweto.

Figure 3.7. Mean LAT accuracy over the three settlement types (FS, FSB and OIS) for both dates of the Soweto dataset in terms of the Czekanowski-Dice index with the shadow groundtruth dataset.

3.5.7 Comparison: Global thresholding and local adaptive thresholding

3.5.7.1 Shadow detection accuracy

The primary measurement of shadow detection accuracy is the Czekanowski-Dice external validation index, which is equivalent to the F-score as shown in paragraph 2.4.4.2. This is an appropriate evaluation measure that works especially well in cases with low shadow probability. Global thresholding is compared with LAT for the two dates of the Soweto dataset, in terms of the Czekanowski-Dice index in Figure 3.8.



(a) Global thresholding. **(b)** LAT.

Figure 3.8. Shadow detection accuracy comparison between global thresholding and LAT in terms of the mean LAT accuracy over the three settlement types (FS, FSB and OIS) for both dates of the Soweto dataset. Accuracy is measured in terms of the Czekanowski-Dice index with the shadow groundtruth dataset.

Note that the y-axis ranges in Figures 3.8(a) and 3.8(b) are the same so that the shadow detection accuracies can be compared directly. A summary of the shadow detection accuracy of global thresholding and local thresholding is given in Table 3.4 for good operating points specified in terms of the global threshold T, the local window width w and Wellner's parameter k.

Table 3.4. Comparison between global thresholding and LAT for both dates of the Soweto dataset in terms of the Czekanowski-Dice index calculated with the groundtruth shadows for two FS polygons and single polygons of the FSB and OIS settlement types.

Polygon	**Date 1 - Johannesburg**		**Date 2 - Johannesburg**	
	Global	Local adaptive	Global	Local adaptive
	(T=60)	(w=30, k=50)	(T=30)	(w=30, k=70)
FS 1	0.703	0.804	0.739	0.807
FS 2	0.771	0.746	0.671	0.632
FSB	0.582	0.678	0.647	0.801
OIS	0.635	0.722	0.643	0.707
Mean (μ	0.673	0.738	0.675	0.737
$\pm\sigma$)	$\pm$0.082	$\pm$0.053	$\pm$0.044	$\pm$0.084

For both dates of the Soweto dataset the LAT algorithm of Wellner outperforms global thresholding, although this requires a specific optimal window size derived from the mean accuracy graphs in Figure 3.7. A local window size of 29 pixels places the algorithm at the start of the stable high accuracy plateau in Figure 3.7, but the algorithm can still benefit from a relatively small window size.

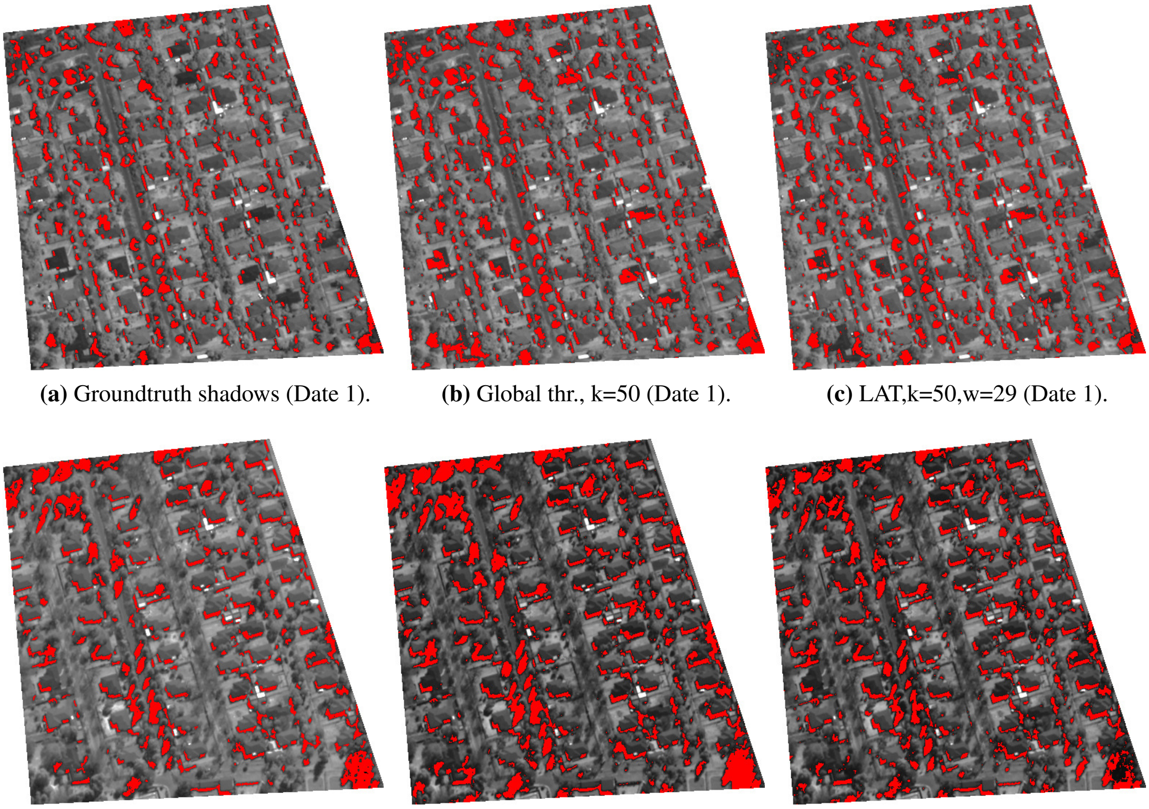
3.5.7.2 Shadow mask comparison

The groundtruth shadows for a formal settlement polygon are shown in Figures 3.9(a) and 3.9(d). The detected shadow masks of global thresholding are shown in Figures 3.9(b) and 3.9(e) and for LAT they are shown in Figures 3.9(c) and 3.9(f). The near-optimal thresholds for global thresholding and LAT are selected specifically for each date of the Soweto dataset, but the thresholds are not differentiated over different settlement types.

For some of the dark roofs in the formal settlement polygon of date 1 it can be seen that global thresholding selects more of the roof area as shadow than the LAT. The benefit of LAT is that the algorithm will take the local roof intensity into consideration in the decision on whether a region is shadow or not. In date 2 the difference in false positive overselection is less pronounced between global and LAT.

3.6 CONCLUSION

The objective is to instantiate shadow removal for input modification, so shadow detection solutions are explored in this chapter. The hypothesis that threshold-based shadow detection can relatively accurately delineate shadows because of the low-intensity property of shadows was investigated, as well as the hypothesis that LAT can produce more accurate shadows than global thresholding, since



(a) Groundtruth shadows (Date 1). **(b)** Global thr., k=50 (Date 1). **(c)** LAT,k=50,w=29 (Date 1).

(d) Groundtruth shadows (Date 2). **(e)** Global thr., k=25 (Date 2). **(f)** LAT,k=70, w=29 (Date 2).

Figure 3.9. Detected shadow masks for a co-registered formal settlement polygon over the two acquisitions of the Soweto dataset. Panchromatic background images courtesy of DigitalGlobe™.

relatively low intensity admits greater sensitivity in images with contrast variation than globally low intensity.

Panchromatic shadow detection algorithms from the thresholding subcategory (Adeline et al. [59], Table 2.2) of property-based shadow detection (Arévalo et al. [73]) are used on the Soweto panchromatic land-use dataset. Select thresholding algorithms from the taxonomy of Sezgin and Sankur [47] are also compared for shadow detection accuracy in terms of Czekanowski-Dice (F-score), Jaccard, Rand (overall), Rogers-Tanimoto and Sokal-Sneath external validation indices (paragraph 2.4.4.2). Locally adaptive thresholding is compared to global thresholding in terms of panchromatic shadow detection accuracy.

The minimum Bayes error is difficult to detect with minimum error thresholding because of extensive overlap of shadow and sunlit densities, so it achieves the lowest unsupervised global thresholding accuracy. Iterative minimum error thresholding avoids boundary minimums and focuses on a relevant intensity range so it achieves the highest shadow detection accuracies of the thresholding methods considered. Convex hull thresholding can robustly determine the threshold valley as no explicit bimodality is required, and this thresholding method attains the second highest accuracy for unsupervised global threshold detection.

Wellner's LAT is used for more accurate shadow detection and it is shown that the local window size parameter is robust despite multitemporal shadow profile differences. The potential shadow detection accuracy of global thresholding was compared to that of LAT and for both dates of the Soweto dataset LAT outperformed global thresholding. Global thresholding produced more false positives in the shadow mask, but LAT can take local intensity into account to reduce this and produce more accurate maps.

The following chapter on input modification uses the shadow detection methods evaluated in this chapter for dataset shift reduction.

CHAPTER 4 INPUT MODIFICATION

4.1 CHAPTER OVERVIEW

Multitemporal satellite-borne image acquisition introduces complex variances owing to a conflation of differences in viewing angles, illumination characteristics and environmental factors of the captured scenes. Multitemporal land-use analysis has to filter out these artificial differences to obtain an accurate account of actual land-cover changes. In the case of supervised land-use classification with limited groundtruth data, such calculated invariance can substantially improve across-date classification accuracy. The varying viewing and illumination geometry in multitemporal imagery must be accounted for, since those differences will normally become embedded in the classification features and compromise the supervised labeling.

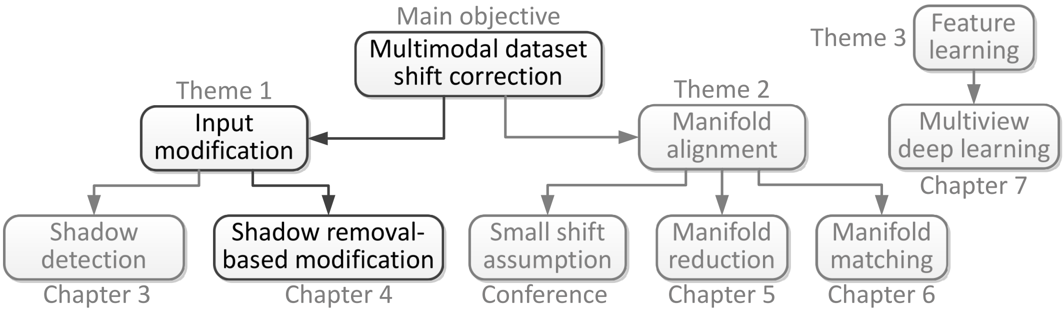


Figure 4.1. Indication of where this chapter fits into the thesis.

The focus of this study is on demonstrating the input modification strategy to reduce dataset shift for land-use classification with high-resolution panchromatic acquisitions, using texture features to distinguish between settlement classes. Texture features are used as they have been shown to be a good measure of settlement patterns (see [164]). The important multitemporal variance component of shadow is effectively removed before feature extraction, which allows for weakly supervised across-date classification. The effect of shadows on classification accuracy is specifically investigated in this study by performing shadow removal, before the calculation of texture features, which are then used as classifier inputs.

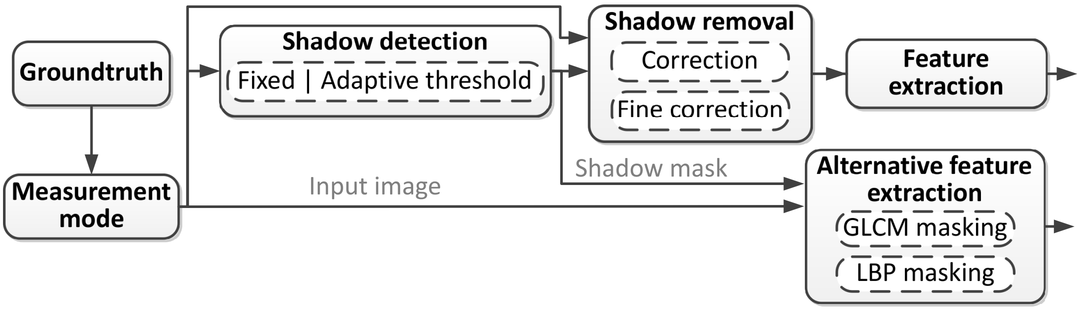


Figure 4.2. Basic input modification strategy.

4.1.1 Contributions

1. Dataset shift is reduced at the classifier level by removing the shadow component of illumination geometry at the input level of feature extraction.

2. It is shown that popular texture features are sensitive to differences in shadow profiles, and that across-date classification accuracy can be improved with shadow removal.

3. Shadow detection based on locally adaptive thresholding is employed and experimentally shown to outperform existing global threshold shadow detectors in increasing settlement classification accuracy.

4. The strong relationship between the best shadow detection threshold and the best settlement classification accuracy is indicated for the study data set.

5. An improved shadow correction algorithm that relies on region growing and localized histogram matching is contributed in this design, and demonstrated to be more effective than basic shadow correction methods.

6. It is shown that fine correction leads to more accurate classification across a wide range of shadow thresholds, compared to shadow correction that relies on global histogram matching.

7. Both same and across-date settlement accuracies are significantly improved with shadow masking during feature calculation.

8. A statistical study was performed and found to support the hypothesis that the increased accuracy is due to shadow masking specifically.

9. Alternative masking unrelated to shadows is applied to obtain evidence that it is shadow masking specifically that can improve land-use classification.

4.2 PROBLEM STATEMENT

The QuickBird satellite acquires scenery at potentially very different azimuth angles because of its maximum 45° off-nadir wide accessible ground swath. Viewing geometry variations aggravate the strong directional differences seen in urban surface features [165]. The sun elevation and solar illumination characteristics during acquisition may also have significant across-date differences, which produce considerable illumination variance, of which shadowing is a dominant factor. Changes in dynamic range of the scene intensity are another effect of illumination variance, but shadowing is more adverse since its presentation is locally coupled with surface features and thus more difficult to remove.

Apart from image brightness and reflection differences, which can be partially addressed with histogram matching, there can also be changes in cast shadow direction, length and area. It has been shown that texture features are sensitive to spurious differences in viewing and illumination geometry [164]. This study focuses specifically on the effect shadow differences have on land-use classification accuracy.

Shadow variances are among the more acute of across-date differences, and this is indicated in the comparison shown in Figure 4.3. Sample areas of the three main settlement types are shown for each of the study dates and the shadows detected with locally adaptive thresholding are shown in color. For a particular settlement type the same area is shown for both dates so that the shadow differences may be directly compared. Since the image for date 1 (d_1) was acquired during the summer season, and date 2 (d_2) during winter, the shadows of d_2 are longer because of the sun being in a more northerly position.

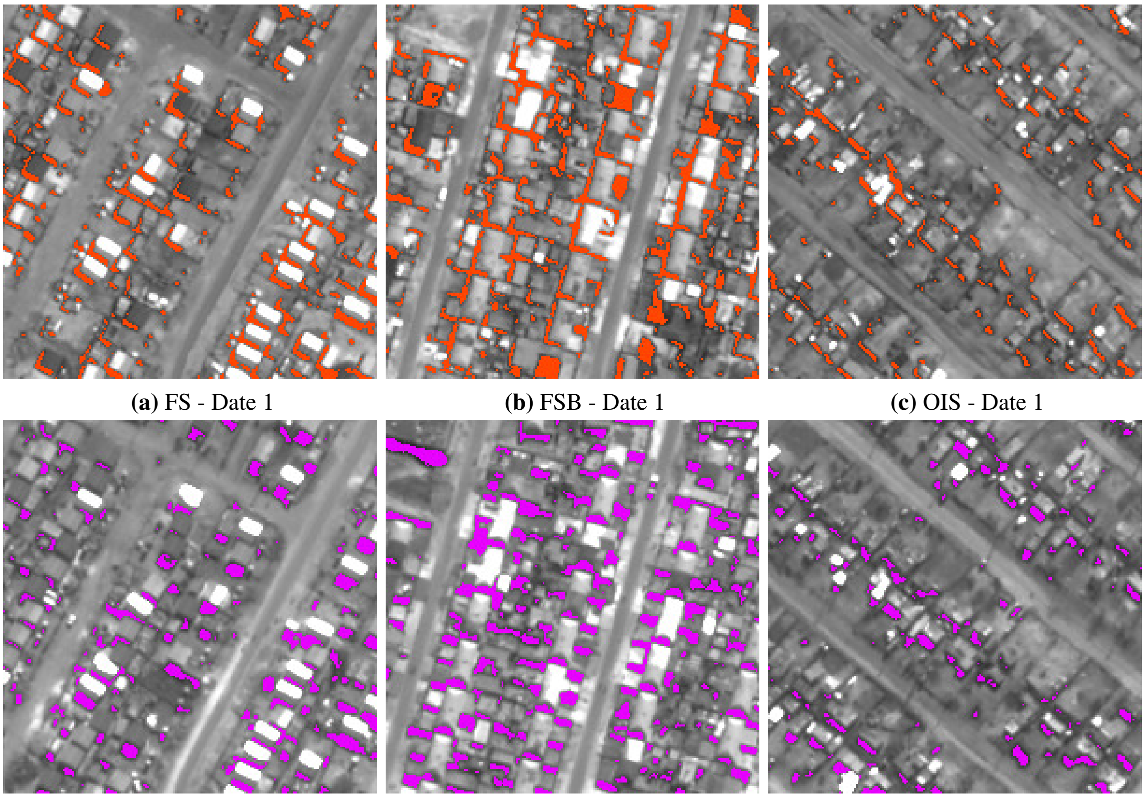
Pattern classification relies on differential measures to distinguish between different classes, but in the instance of across-date land-use classification the potentially severe shadow variances can cause intra-class feature variation. This leads to a higher incidence of class confusion and overlap, so it is the objective of shadow invariance to remove the significant illumination variance from the feature calculation.

4.2.1 Hypotheses

1. If dataset shift aspects between the classifier train and test inputs to the feature extraction layer are corrected or equalized, then the dataset shift at the classification layer will also be reduced because of the resulting features having smaller dataset shift.

2. Shadow profile differences between the classifier train and test images cause dataset shift at the classifier, and the removal of shadows in order to remove the shadow profile differences as well will reduce the corresponding dataset shift component and improve classification accuracy because of the resulting features having smaller dataset shift.

3. The more extreme a dataset shift becomes because of shadow profile differences, the more



(a) FS - Date 1 (b) FSB - Date 1 (c) OIS - Date 1

(d) FS - Date 2 (e) FSB - Date 2 (f) OIH - Date 2

Figure 4.3. Examples of the across-date shadow differences for three of the settlement classes found in Soweto. Panchromatic background images courtesy of DigitalGlobe™.

settlement classification accuracy will improve for an improvement in shadow removal accuracy, since more of the input dataset shift will be corrected with more accurate shadows.

4.2.2 Research questions

1. Which strategy can be employed at the input level of feature extraction to deal with dataset shift in optical remote sensing?

2. What effects do multitemporal (same satellite) shadow profile differences have on texture-based settlement classification accuracy?

3. How does one achieve effective panchromatic shadow removal?

4. How much does adaptive threshold shadow detection improve classification accuracy compared to global threshold detection?

4.3 TEXTURE FEATURES

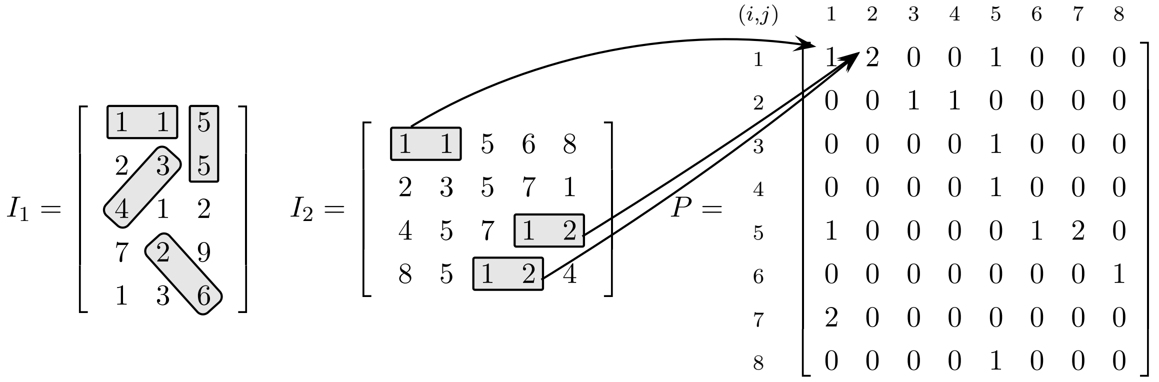
The fundamental characteristics that can be used in image classification are spectral, textural and contextual features. Spectral features are based on tonal or intensity variations in spectral bands, and textural features are based on the spatial distribution of intensity variations, whereas contextual features are derived from image objects, background and domain-specific knowledge thereof.

Texture is one of the most valuable discriminative features, since it is an intrinsic property of almost all image components. It contains information on the structural configuration of image elements and its relationship relative to the surroundings. Texture has an inextricable relationship with pixel intensity and this is exploited to derive the texture features presented in this section.

4.3.1 Gray-level co-occurrence matrix features

The calculation of GLCM features is reviewed in this subsection and is based on the seminal work by Haralick [166]. Given an X-by-Y two-dimensional digital image $I : X \times Y \to G$ quantised to the set of $G \in \{1, 2, \cdots, N_g\}$ possible pixel gray-level values, a gray-tone spatial-dependence matrix or GLCM can be obtained for a given horizontal or vertical spatial relationship.

The GLCM is a two-dimensional histogram $P : N_g \times N_g \to \mathbb{Z}$ that captures the frequency of two distinct tones co-occurring in a given spatial relationship where the tones are on the same horizontal, vertical or diagonal (45° or 135°) image grid line with a specified L_1 norm or Manhattan distance between them. These spatial relationships are indicated in image I_1 below for a horizontal and vertical Manhattan distance of 1 and a diagonal Manhattan distance of 2. An example of unnormalised GLCM is given by P for image I_2 below and it is shown how pixel value pairs (i, j) increment the corresponding matrix element in P.



$$I_1 = \begin{bmatrix} 1 & 1 & 5 \\ 2 & 3 & 5 \\ 4 & 1 & 2 \\ 7 & 2 & 9 \\ 1 & 3 & 6 \end{bmatrix} \quad I_2 = \begin{bmatrix} 1 & 1 & 5 & 6 & 8 \\ 2 & 3 & 5 & 7 & 1 \\ 4 & 5 & 7 & 1 & 2 \\ 8 & 5 & 1 & 2 & 4 \end{bmatrix} \quad P = \begin{array}{c|cccccccc} (i,j) & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ \hline 1 & 1 & 2 & 0 & 0 & 1 & 0 & 0 & 0 \\ 2 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 5 & 1 & 0 & 0 & 0 & 0 & 1 & 2 & 0 \\ 6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 7 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 8 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{array}$$

4.3.1.1 Notation

The following basic constructs are defined as they are subsequently used to calculate the statistical GLCM features [166]:

N_g, Number of distinct gray-levels.

$P,$	Unnormalised $N_g \times N_g$ GLCM.
$p,$	Normalized GLCM.
$p(i,j) = P(i,j)/\sum P,$	(i,j)th entry in the normalized GLCM.
$p_x(i) = \sum_{j=1}^{N_g} p(i,j),$	ith entry in the marginal-probability matrix.
$p_y(j) = \sum_{i=1}^{N_g} p(i,j),$	jth entry in the marginal-probability matrix.
$p_{x+y}(k) = \sum_{\substack{i=1 \\ i+j=k}}^{N_g} \sum_{j=1}^{N_g} p(i,j),$	$k = 2,3,\cdots,2N_g.$
$p_{x-y}(k) = \sum_{\substack{i=1 \\ \lvert i-j \rvert=k}}^{N_g} \sum_{j=1}^{N_g} p(i,j),$	$k = 0,1,\cdots,N_g - 1.$
$\mu_i = \sum_{j=1}^{N_g} ip(i,j),$	Mean row intensity level.
$\mu_j = \sum_{i=1}^{N_g} jp(i,j),$	Mean column intensity level.
$\mu_x = \sum_{i=1}^{N_g} p_x(i)/N_g,$	Mean of p_x.
$\mu_y = \sum_{i=1}^{N_g} p_y(i)/N_g,$	Mean of p_y.
$\sigma_x^2 = \sum_{i=1}^{N_g} (p_x(i) - \mu_x)^2/N_g,$	Variance of p_x.
$\sigma_y^2 = \sum_{i=1}^{N_g} (p_y(i) - \mu_y)^2/N_g,$	Variance of p_y.
$HX = -\sum_i p_x(i) \log\left(p_x(i)\right),$	Entropy of p_x.
$HY = -\sum_j p_y(j) \log\left(p_y(j)\right),$	Entropy of p_y.
$HXY1 = -\sum_i \sum_j p(i,j) \log\left(p_x(i)p_y(j)\right),$	Specific entropy for feature f_{12}.
$HXY2 = -\sum_i \sum_j p_x(i)p_y(j) \log\left(p_x(i)p_y(j)\right),$	Specific entropy for feature f_{13}.
$Q(i,j) = \sum_k \frac{p(i,k)p(j,k)}{p_x(i)p_y(k)},$	Required for feature f_{14}.

4.3.1.2 Textural features

For a given spatial relationship 14 GLCM features can be calculated, and the features of multiple spatial relationships can be combined to produce a single set of 14 feature means and 14 feature standard deviations that form the final features.

The GLCM is calculated for all possible patterns that occur in a specified window centered at a particular image location for a given spatial relationship, thus producing 28 features at each window center location. The window position is slid horizontally and vertically to cover the entire image and thus produce a GLCM feature image. The GLCM feature calculations are given below with the aid of the previously defined notations [166].

1) Angular second moment:
$$f_1 = \sum_i \sum_j \{p(i,j)\}^2 \tag{4.1}$$

2) Contrast:
$$f_2 = \sum_{n=0}^{N_g-1} n^2 \left\{ \sum_{\substack{i=1 \\ |i-j|=n}}^{N_g} \sum_{j=1}^{N_g} p(i,j) \right\} \tag{4.2}$$

3) Correlation:
$$f_3 = \frac{\sum_i \sum_j (ij)p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y} \tag{4.3}$$

4) Sum of squares:
$$f_4 = \sum_i \sum_j (i - \mu_i)(j - \mu_j)p(i,j) \tag{4.4}$$

5) Inverse difference moment:
$$f_5 = \sum_i \sum_j \frac{1}{1 + (i-j)^2} p(i,j) \tag{4.5}$$

6) Sum average:
$$f_6 = \sum_{i=2}^{2N_g} i p_{x+y}(i) \tag{4.6}$$

7) Sum variance:
$$f_7 = \sum_{i=2}^{2N_g} (i - f_8)^2 p_{x+y}(i) \tag{4.7}$$

8) Sum entropy:
$$f_8 = -\sum_{i=2}^{2N_g} p_{x+y}(i) \log\left(p_{x+y}(i)\right) \tag{4.8}$$

9) Entropy:
$$f_9 = -\sum_i \sum_j p(i,j) \log\left(p(i,j)\right) \tag{4.9}$$

10) Difference variance:
$$f_{10} = \sqrt{\frac{1}{N_g} \sum_{i=0}^{N_g-1} \left(p_{x-y}(i) - \sum_{j=0}^{N_g-1} \frac{p_{x-y}(j)}{N_g} \right)^2} \tag{4.10}$$

11) Difference entropy:
$$f_{11} = -\sum_{j=0}^{N_g-1} p_{x-y}(j) \log\left(p_{x-y}\right) \tag{4.11}$$

12) Inform. meas. of correlation 1:
$$f_{12} = \frac{f_9 - HXY1}{\max(HX, HY)} \tag{4.12}$$

13) Inform. meas. of correlation 2:
$$f_{13} = \sqrt{1 - \exp\left(-2(HXY2 - f_9)\right)} \tag{4.13}$$

14) Max correlation coefficient:
$$f_{14} = (\text{Second largest eigenvalue of } Q)^{1/2} \tag{4.14}$$

4.3.2 Local binary patterns

The second powerful texture feature used in this thesis is LBP, which capture both structural and statistical properties of the microstructures embodying the image textures. Gray-scale and rotation invariant uniform LBP are employed, and the literature is reviewed in the subsection according to [167].

4.3.2.1 Gray-scale invariant patterns

LBP achieves gray-scale invariance through a binary comparison between a center pixel of a pattern and the neighborhood members of the pattern placed according to the quantised angular space. Gray-scale invariance reduces the pattern information to binary relationships so that exact pixel value differences do not need to be stored, and in addition the remote sensing illumination variances do not affect the capturing of the same patterns as severely.

Gray-scale invariance is incorporated through a binary comparison between a center pixel and its neighborhood members. For a center pixel with value g_c and for a neighbor pixel value g_p with a P-point circularly symmetrical pattern of radius R the gray-scale invariant pattern can be coded as $LBP_{P,R}$ [167]

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p, \qquad s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}. \tag{4.15, 16}$$

4.3.2.2 Rotation invariant patterns

There is a possibility of the reoccurrence of a specific texture but at a different orientation, which would then require rotation invariant features to establish eventual texture similarity. Rotation invariance reduces all possible patterns to the smaller set of unique patterns so that texture equivalency is possible despite texture rotations.

The pattern set reduction is modeled so that equivalent patterns that are rotated are placed in the same pattern histogram bin. Rotation invariance (ri) can be accommodated in $LBP_{P,R}^{ri}$ by minimizing the circular bitwise right-shift $ROR(\cdot)$ of a pattern $LBP_{P,R}$, which then represents equal patterns with a rotation variance as exactly the same pattern value [167],

$$LBP_{P,R}^{ri} = \min\{ROR(LBP_{P,R}, i) \quad | \quad i = 0, 1, \cdots, P-1\}. \tag{4.17}$$

4.3.2.3 Uniform patterns

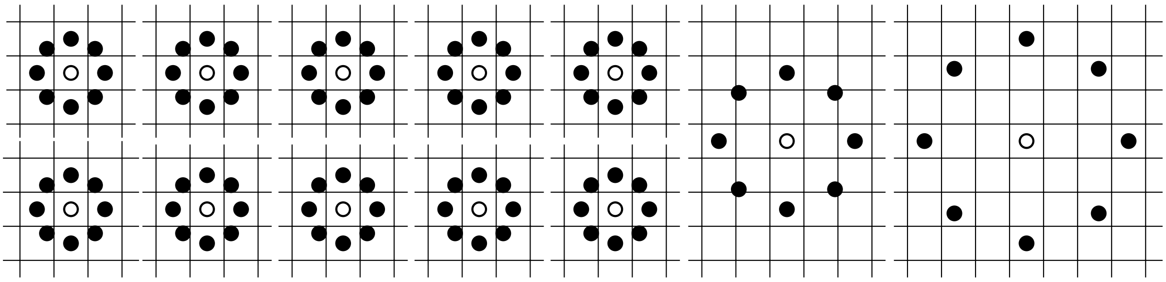
In practice nonuniform patterns with more than two transitions have a relatively small occurrence probability, so for the purpose of reducing the LBP features to the essential minimum, only uniform patterns are considered. For a neighborhood size of P a histogram with $P+2$ bins is constructed, with a bin for each of the $P+1$ uniform patterns and a final bin to contain all the nonuniform pattern counts.

The number of circular transitions $U(LBP_{P,R})$ of a given pattern $LBP_{P,R}$ is evaluated as [167]

$$U(LBP_{P,R}) = |s(g_{P-1} - g_c) - s(g_0 - g_c)| + \sum_{p=1}^{P-1} |s(g_p - g_c) - s(g_{p-1} - g_c)|. \tag{4.18}$$

A pattern is declared uniform if it has two or fewer transitions and the rotation invariant pattern is then allocated to the correct LBP histogram bin. The resulting histogram $LBP_{P,R}^{riu2}$ conceptualised in Figure 4.4(a) contains rotation invariant (*ri*) and uniform patterns with two or fewer transitions (*u*2), and is determined as [167]

$$LBP_{P,R}^{riu2} = \begin{cases} \sum_{p=0}^{P-1} s(g_p - g_c), & \text{if } U(LBP_{P,R}) \leq 2 \\ P+1, & \text{otherwise} \end{cases}. \tag{4.19}$$



(a) $LBP_{8,1}^{riu2}$ rotation invariant uniform patterns **(b)** $LBP_{8,2}^{riu2}$ **(c)** $LBP_{8,3}^{riu2}$

Figure 4.4. Local binary pattern examples for different pattern radii, namely $R = 1,2,3$.

4.3.2.4 LBP feature calculation

Each bin in the LBP histogram represents the unnormalised probability of encountering a given pattern at any orientation or gray-scale level, which constitutes the final features. The LBP histogram is calculated by fitting a given circular neighborhood template over every unique image position within a defined sampling window and incrementing the bin associated with the detected pattern.

The LBP features are then representative of the texture inside the sampling window at its center. An LBP image can then be formed through sampling window placements that achieve maximum coverage of a given image.

4.4 METHODOLOGY

4.4.1 Global threshold shadow detection

Shadow detection is the initial step, after which effective shadow removal is achieved through mainly shadow correction or masking. The shadow-effect mitigation method involves shadow detection and removal, as shown in Figure 4.5, where detection is the determination of a binary shadow mask indicating the perceived occlusion of sunlight. Global intensity thresholding can be applied to produce a rudimentary shadow mask [94] by declaring all pixels with an intensity less than the fixed threshold as shadow, but here the accuracy of the mask depends strongly on the specific threshold chosen. Shadow intensity may vary across different parts of the same scene, and a shadow area can then form a local gradient with intensities both less than and exceeding the fixed threshold.

4.4.2 Locally adaptive threshold shadow detection

Locally adaptive thresholding allows for more robust shadow detection, especially for gradient shadows, and its threshold parameters require less across-date modification than fixed thresholding. Locally adaptive thresholding is better suited as a panchromatic detector since the actual threshold is determined relative to the mean intensity in a local window. A pixel is declared as shadow in the mask if its intensity is less than the mean pixel intensity in the square window centered at that pixel, minus a given offset intensity value, which gives locally adaptive thresholding two main parameters, i.e. window size (number of edge pixels) and adaptive parameter. Shadow intensity perturbations in the same scene and across-date dynamic range differences may thus be better accommodated than with a fixed threshold.

To simplify analysis the shadow detection experiments are performed in terms of only one threshold, so multi-date images are initially histogram-matched to reduce dynamic range variances and then the threshold detector classifies every pixel with an intensity less than the threshold as shadow. Global intensity thresholding is used as a benchmark to test locally adaptive thresholding against, where Wellner's LAT method [156] (paragraph 3.4.9.1) is used with the main parameter *k* that is the percentage below the window mean at which to set the threshold and a supervised parameter *w* that is the local window size.

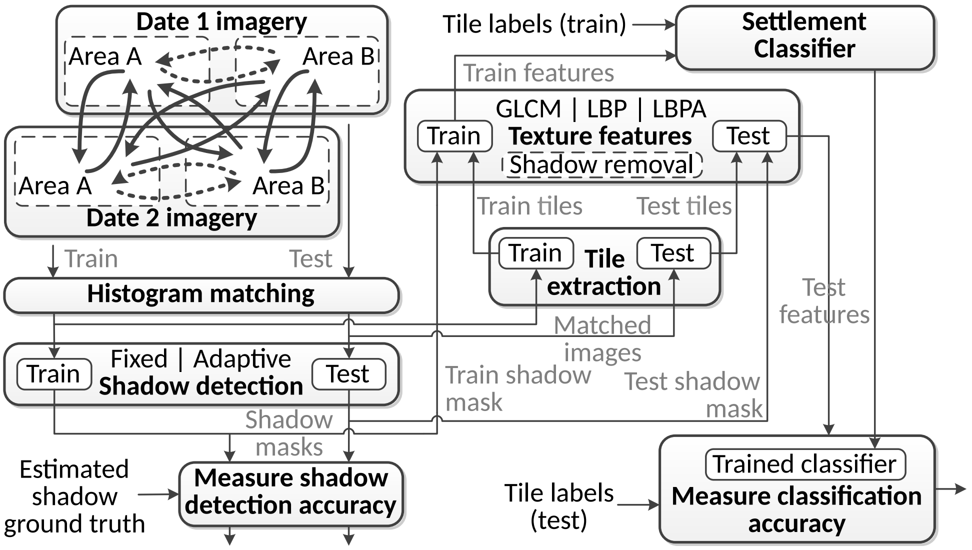


Figure 4.5. An outline of the experiment methodology, and the separation of same-date (- - -) and across-date (—) experiments (top left).

4.4.3 Texture feature calculations

GLCM [166] and LBP [167] texture features have been shown to perform well in settlement determination (see [164]) and are used in this study. However, texture features such as GLCM and LBP are sensitive to viewing and illumination geometry differences [164], so these features have the potential to benefit from calculated feature invariance.

4.4.3.1 GLCM features

Texture features are calculated per labeled tile, using either GLCM or LBP features. The first 13 of Haralick's GLCM features [166] are determined with the GLCM window having tile dimensions of 200×200 pixels. Using the first 13 GLCM features proved to be more accurate for the study area, compared to using a feature subset of size 6 based on information gain or correlation feature selection, or only the most relevant features (energy, contrast, variance, correlation, entropy and the inverse difference moment). GLCM pairs are used in all cardinal and ordinal directions with respective ℓ_1-norms of 1 and 2, and the features are averaged over the four spatial relationships.

4.4.3.2 LBP features

LBP features use the 10 basic patterns [167] in an eight-point circle with radii of one, four and eight pixels to render a total of 30 features. As shadow masking is performed the GLCM pairs or LBP patterns that fall within a shadow area are ignored during feature calculation. Both global and locally adaptive thresholding shadow detectors are tested and compared as part of the shadow masking process.

4.4.4 Shadow masking

The original image and its detected shadow mask are used to perform shadow masking, and in this study it has the purpose of effectively removing multitemporal shadow variances by removing the shadows. Histogram matching, gamma correction and linear-correlation correction [66] are popular methods of shadow correction, where shadows are lifted to have similar intensity to that of the surroundings. Corrected shadows suffer from posterization and the effect on classification accuracy is of concern. Alternatively this study will opt for shadow masking, where shadows are ignored during feature calculation.

Test data images are firstly histogram-matched to the train data image to address differences in image brightness. Image regions with gray-scale intensities smaller than a global shadow threshold, which is the same for both train and test images, are then classified as shadows. Shadow masking removes all shadow regions from texture feature calculations, whereas correction lifts shadow region intensities to match the outside.

The binary shadow mask provided by the detector is then directly used during texture feature calculation in order to mask out shadow areas. Every GLCM pixel pair is skipped if one of those pixels is located in shadow, which then effectively removes off-diagonal entries in the co-occurrence matrix that would have resulted with no masking. The LBP features are calculated in a similar manner so that a pattern is not placed when its central pixel forms part of a shadow area.

4.4.5 Shadow restoration

Histogram equalization, proposed by Shu and Freeman [75] and used by Sarabandi et al. [112], from the intensity domain (paragraph 2.5.2.1) of the shadow restoration taxonomy in Table 2.4 can be employed for shadow removal. A variant of histogram equalization is correction by adding to every shadow pixel the intensity difference between the polygon intensity mean and the corresponding shadow region mean. A two-pixel-wide transition edge outside each shadow region is then locally median-filtered in a five-pixel-wide window [168].

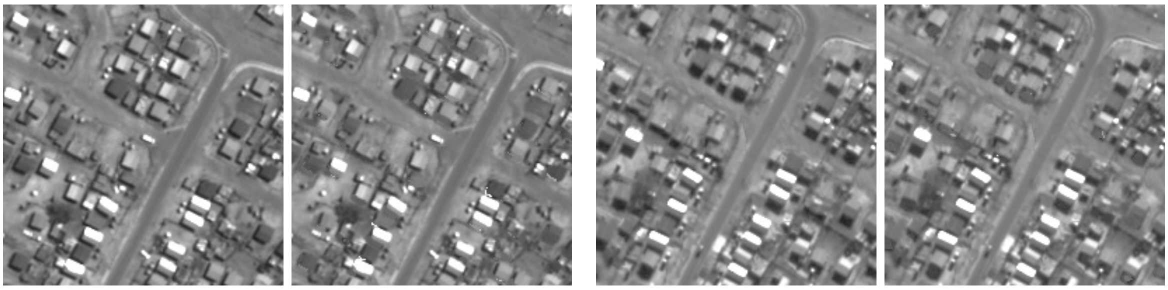
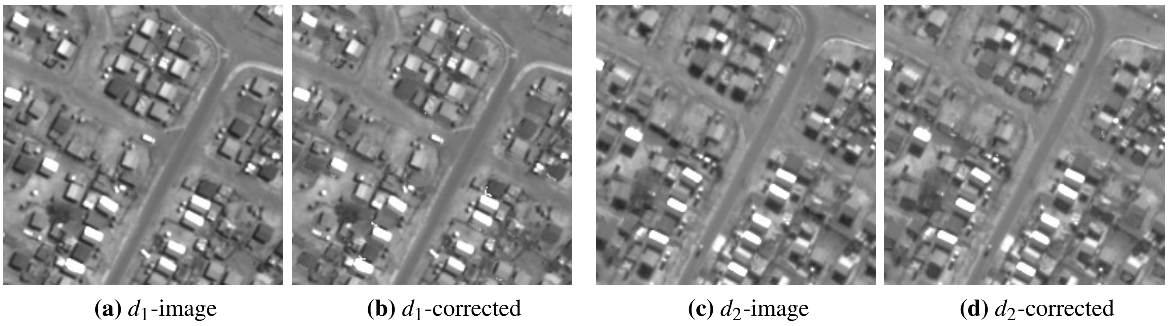

(a) Scene from d_1 **(b)** Shadow corrected (d_1) **(c)** Scene from d_2 **(d)** Shadow corrected (d_2)

Figure 4.6. Illustration of the differences in shadow profiles for the same area acquired on different dates, and the corresponding scenes with fine shadow correction. (a) and (c) courtesy of DigitalGlobe™.

Finer shadow correction is possible by taking as non-shadow mean the most frequently occurring intensity value in a two-pixel-wide edge outside the transition edge. This blends the lifted shadow better than with a global intensity mean, and addresses the situation where a shadow is surrounded by surfaces with distinctly different intensities.

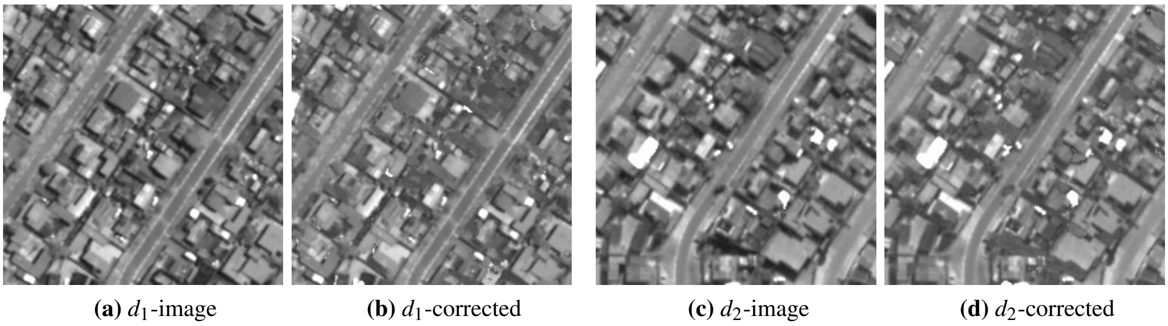
Fine shadow correction is displayed in Figure 4.6(b) and 4.6(d) for images with notably different shadow profiles in Figure 4.6(a) and 4.6(c). For a shadow threshold of 25% not all shadows are detected, but many of the larger shadow differences are corrected between the two dates.

The across-date differences in shadow profiles are depicted in Figures 4.7, 4.8 and 4.9 for different settlement types. The uncorrected images for the different acquisition dates d_1 and d_2 are denoted by d_1-image and d_2-image, and the shadow-corrected images are given by d_1-corrected and d_2-corrected. The shadow correction has been performed with the refined algorithm and shadow masks detected with a global threshold of $0.25Y_{\max}$ (maximum possible intensity is $Y_{\max}$).



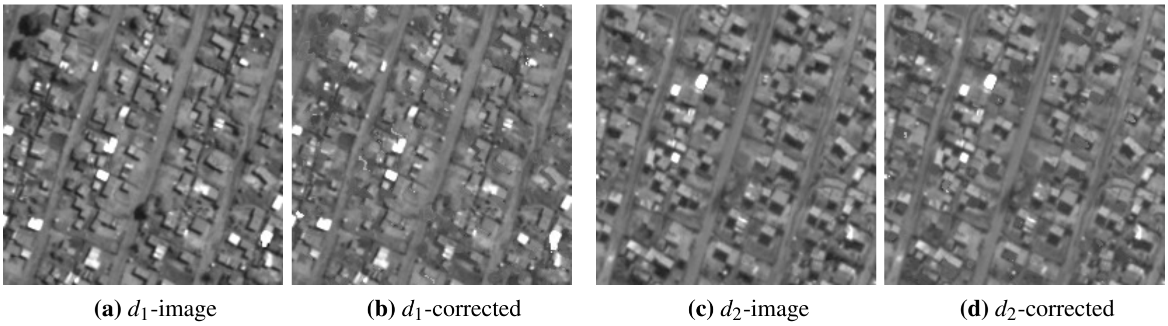
(a) d_1-image **(b)** d_1-corrected **(c)** d_2-image **(d)** d_2-corrected

Figure 4.7. FS and corresponding shadow-corrected images for different acquisition dates. (a) and (c) courtesy of DigitalGlobe™.



(a) d_1-image **(b)** d_1-corrected **(c)** d_2-image **(d)** d_2-corrected

Figure 4.8. FSB and corresponding shadow-corrected images for different acquisition dates. (a) and (c) courtesy of DigitalGlobe™.



(a) d_1-image **(b)** d_1-corrected **(c)** d_2-image **(d)** d_2-corrected

Figure 4.9. OIS and corresponding shadow-corrected images for different acquisition dates. (a) and (c) courtesy of DigitalGlobe™.

4.4.6 Feature classification

A multi-layer perceptron (learning rate of 0.3, momentum rate of 0.2, 500 training epochs, and unipolar sigmoid activation functions) is used as classifier with texture features as input and with four classes of FS, FSB, OIS and NBU areas as separate output units. The number of input units is equal to the number of texture features (13 for GLCM and 30 for LBP) and the number of units in the single hidden layer is the sum of the number of attributes and classes, divided by two. All features are numerical and are normalized in the range $[-1,1]$ simultaneously for both the train and test datasets.

The perceptron is trained with each one of the datasets (A_{d_1}, A_{d_2}, B_{d_1} or B_{d_2}) in turn, and tested on each of the remaining datasets. That means the classifier can be trained with the test set B_{d_1} and tested on the train set A_{d_2} to give an experiment denoted by $B_{d_1} \to A_{d_2}$. In this manner each experiment pairs different datasets and is performed for 10 repetitions, where the perceptron network weights are reinitialised with a different random seed each time.

4.5 DATA DESCRIPTION

The study area (selected in Figure 4.10) is a 2.7×9.3 km section of Soweto (Gauteng, South Africa), a subtropical highland of which two 0.6×0.6 m resolution panchromatic QuickBird images were captured on 18 October 2005 (d_1, early summer, rainy season) and 30 May 2006 (d_2, early winter). There are notable viewing and illumination geometry differences between the two images, especially in the shadow profiles where d_2 exhibits longer shadows because of the northern hemisphere being inclined towards the sun. Figure 4.11 shows the latest acquisition of the study area and the area selections of the land-use classes are indicated.

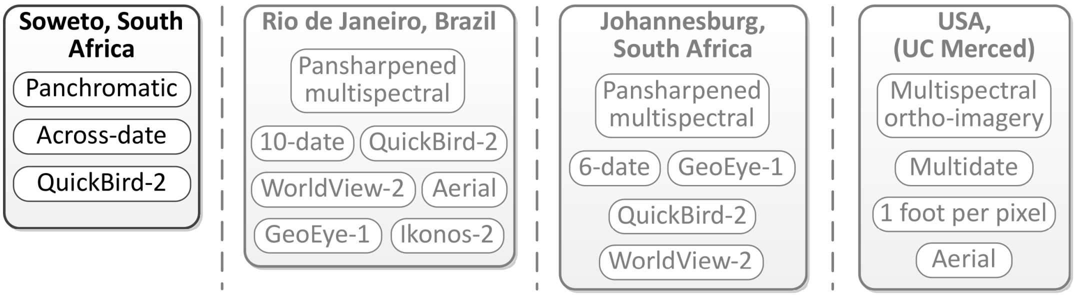


Figure 4.10. Soweto across-date dataset selection for this experiment.

Three prominent settlement types were considered, as shown in Figures 4.7, 4.8 and 4.9, namely FS, FSB and OIS. FS have permanent residential structures positioned in a planned manner, while FSB have residential structures accompanied by smaller shacks. OIS have permanent and semi-permanent residential structures ordered in a planned manner. In addition there is an NBU class that includes natural vegetation, to test the separability between non-builtup areas and urban land-use types.

Representative polygon pairs of each settlement type were extracted in an assisted manner for both dates, most of which form spatially adjacent selections of the same settlement class for the purpose of creating separate training and testing sets. Training and testing data sets are denoted by A and B, respectively, and the data sets are used interchangeably for either training or testing purposes.

The 11-bit panchromatic QuickBird scenes were converted to eight-bit images to enable visual inspection after each step and to simplify the texture feature calculations. The number of GLCM co-occurrence matrix entries is reduced by a factor of 2^6 with this change in bit-depth, and LBP features remain largely unaffected, since the intensity modification is a monotonic transformation. Across-date polygon pairs were then histogram-matched and square image tiles with representative dimensions of 120×120 m were obtained from every polygon.

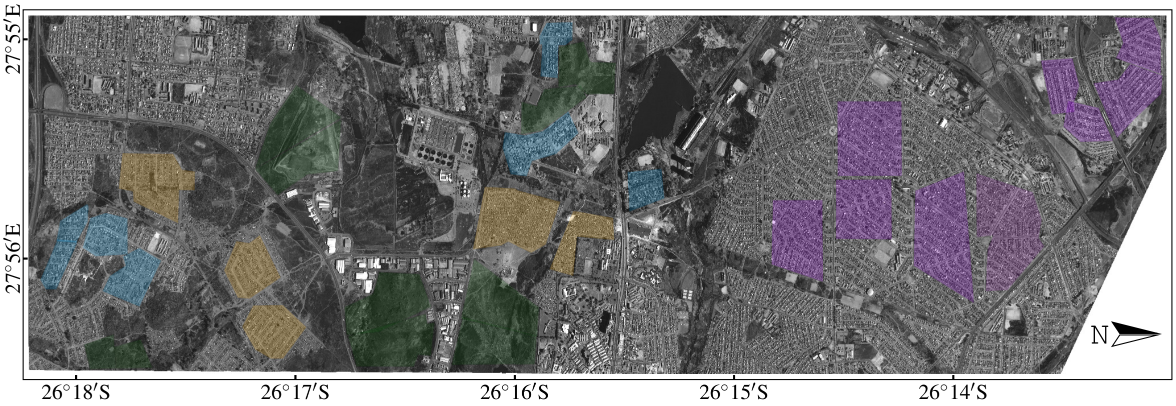


Figure 4.11. The study area of Soweto, acquired on 30 May 2006, with polygon selections of the various land-use classes. Class polygons outlined include (🟪) FSB, (🟦) FS, (🟫) OIS, and (🟩) NBU. Panchromatic background courtesy of DigitalGlobe™.

4.6 EXPERIMENTAL SETUP

The primary objective of the main experiment is to determine the efficacy of shadow removal in improving across-date land-use classification accuracy, and to investigate the relationship between shadow detection accuracy and land-use classification accuracy improvements. The purpose of the experiments in this study is to determine and compare the shadow detection accuracy of the various shadow detectors, and also to measure the change in land-use classification accuracy for the different shadow removal algorithms.

The experimental input includes at least two high-resolution panchromatic images of the same area without metadata, acquired on different dates with significant shadow profile differences, and a small set of ground truth shadow masks for the image of each date. The ground truth shadow masks cover part of each settlement type to give a fully representative shadow mask sample to the supervised segment-based shadow detector. For land-use classification two texture features are compared, namely LBP and GLCM features.

Shadow detection comparisons are performed between a global threshold pixel-based shadow detector and a locally adaptive thresholding pixel-based shadow detector. The shadow removal methods that are compared include shadow masking, basic shadow correction and fine shadow correction. Basic shadow correction using global histogram matching is compared in this study against fine correction that relies on region growing and localized histogram matching. Shadow masking removes shadow areas from the texture feature calculations without the need for image correction. The land-use classification performance with shadow masking under global thresholding detection and LAT is also investigated.

4.6.1 Settlement classification accuracy

A distinction is made between same-date and across-date experiments based on the hypothesis that shadow removal in the case of differing shadow profiles with across-date experiments will have a greater effect on land-use classification accuracy. For two-date experiments with area separation there are four possible same-date experiments and eight different across-date experiments, as shown in Figure 4.5.

The scenery of a specific date is divided into two parts, namely areas A and B, to provide separate training and testing data in order to reduce classifier overfitting. The area separation is performed in terms of whole polygon selections, which are depicted in Figure 4.11, with the aim of obtaining an approximately equal number of samples for both areas. The same area selections are used for images from other dates, so that an area is the identical set of polygon locations for all dates, but the actual image is then obtained from the specific date. Areas A and B are used interchangeably as either training or testing datasets to allow for more experiments to be performed.

Datasets consist of particular texture features and the corresponding groundtruth land-use type classification associated with every extracted tile. The training data sets (A_{d_1} and A_{d_2}) have 55, 148, 70, and 133 tiles respectively for the FS, FSB, OIS and NBU land-use types. The testing data sets (B_{d_1} and B_{d_2}) contain 50, 120, 67, and 113 tiles for the FS, FSB, OIS and NBU classes, respectively. The training and testing data sets are also used respectively as testing and training data sets, while area separation between training and testing data sets is always maintained for same-date experiments.

4.6.2 Shadow detection accuracy

The shadow detection accuracy is measured with the F-score or Czekanowski-Dice similarity coefficient using a representative sample of groundtruth shadow masks over all settlement types for both dates. This similarity index was used in the previous chapter on shadow detection analysis, since it emphasizes the true positive rate more than the true negative rate. This yields a more accurate measure for low positive probability shadow masks, since the true negative rate may still be high despite a poor true positive rate.

The set of groundtruth shadow masks used here are shared in Appendix A and consists of shadow masks for two FS, one FSB and one OIS polygons, which are repeated for acquisitions from both dates. These groundtruth masks have been determined with a human expert using a fuzzy selection tool locally to identify shadows. The fuzzy selection tool allows the user to expand and contract a closed two-dimensional selection based on increasing and decreasing the intensity variance in the selection.

The shadow detection accuracy is measured firstly to assess how similar a detected mask is to a human interpretation of the shadows via the groundtruth shadow masks. The inherent lack of object distinction in panchromatic imagery, and the gradient presentation of shadows can cause substantial subjectivity even in the human treatment of shadow selection. This ambiguity that complicates all panchromatic shadow detection methods, means that there is no absolute groundtruth, but the human-aided benchmark is set nonetheless to provide a point of reference. The detection accuracy is measured for every specific threshold value of the global threshold detector, as well as for the entire range of Wellner's k parameter for LAT detection.

4.6.3 Results and discussion

4.6.3.1 Land-use classification accuracy comparison

Input modification with basic shadow correction, fine shadow correction and global thresholding shadow masking is compared in Figures 4.12(a), 4.12(b), 4.12(d) and 4.12(e) in terms of the mean of Cohen's κ for ten-fold cross-validated land-use classification experiments. A comparison of Cohen's κ values are made between land-use classification with GLCM and LBP texture features for LAT shadow masking in Figures 4.12(c) and 4.12(f).

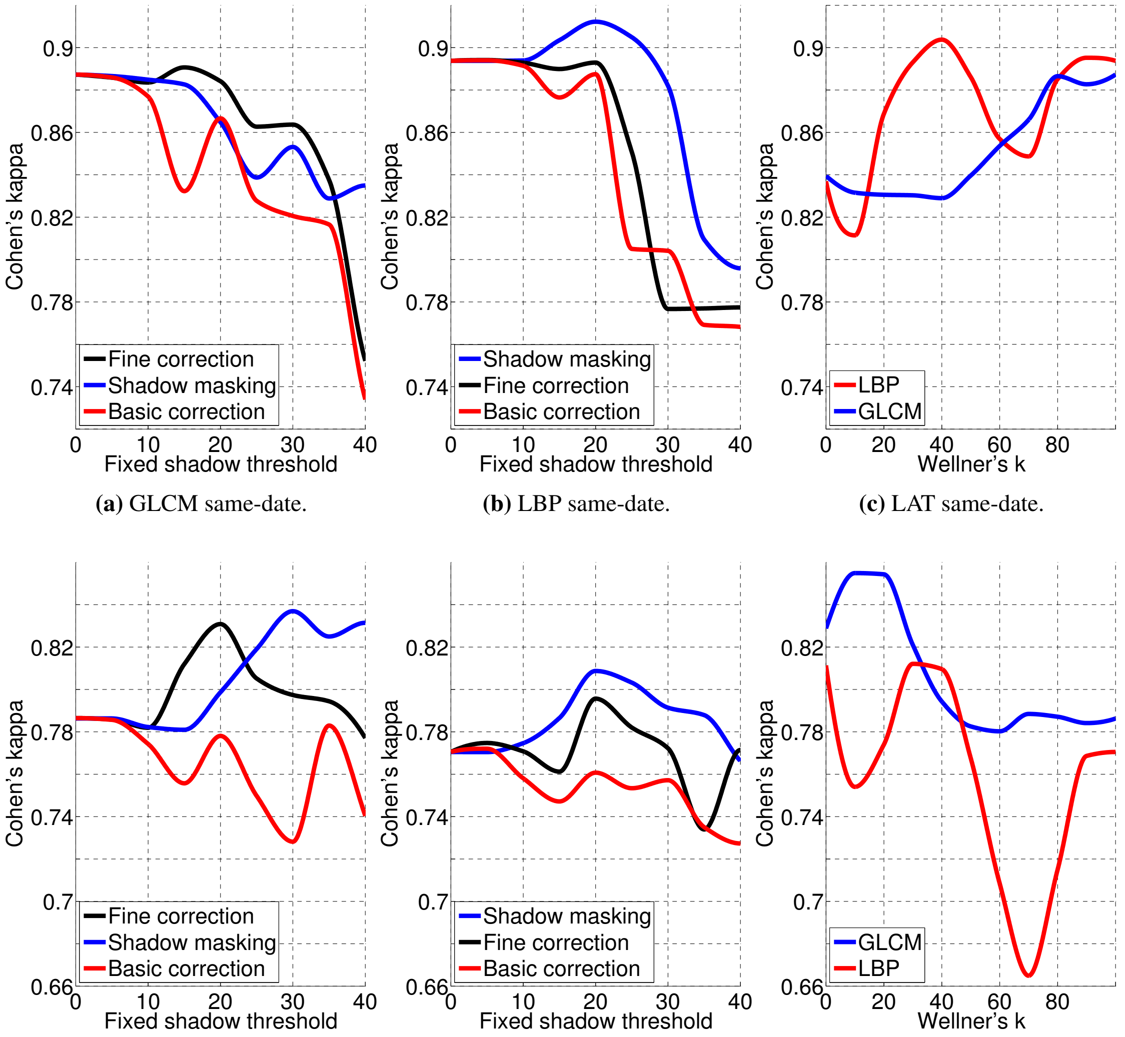
The land-use classification experiments are separated based on the texture features used and whether the training and testing datasets are from the same acquisition (same-date) or across different acquisitions (across-date). This produces the following four testing scenarios:

- GLCM same-date
- GLCM across-date
- LBP same-date
- LBP across-date.

For the shadow removal algorithms that use global thresholding as shadow detector the progression of increased shadow removal occurs from left to right on the x-axis in Figures 4.12(a), 4.12(b), 4.12(d) and 4.12(e). If the global threshold is 0, the shadow mask is empty and no shadow removal will take place. In contrast, the LAT shadow masking progresses from right to left on the x-axis in Figures 4.12(c) and 4.12(f) for Wellner's locally adaptive thresholding algorithm, since Wellner's k value indicates how far below the local window mean the threshold is set. If k =100 the threshold is 0 and no shadows will be detected.

For the same-date experiments the shadow profiles are expected to be similar for similar land-use classes, so in general shadow removal should not improve classification accuracy. However, for GLCM same-date experiments the fine shadow correction appears to improve accuracy slightly and for LBP same-date experiments the global shadow masking also appears to improve accuracy. These improvements are probably unrelated to the shadow component of dataset shift, since there should be no shadow profile differences in same-date imagery.

In shadow masking the GLCM and LBP histograms will probably have fewer entries, because of the omission of patterns incident on shadow pixels. The histograms are normalized, but there is the potential for same-date shadow masking to improve differentiation of land-use classes based on the shadow masking differentiating the class histograms. Some land-use classes have less shadow than



(a) GLCM same-date.

(b) LBP same-date.

(c) LAT same-date.

(d) GLCM across-date.

(e) LBP across-date.

(f) LAT across-date.

Figure 4.12. Land-use classification accuracy comparison in terms of Cohen's κ for GLCM and LBP texture features using global threshold shadow masking, LAT masking (Wellner, w=29), basic and fine shadow corrections with same and across-date distinctions. The fixed shadow threshold on the x-axis is the percentage of the maximum possible pixel value at which global thresholding is used to produce a shadow mask.

others, such as the NBU class that may have less shadow than the FS class. There can thus be a difference in how many pixels are ignored during feature calculation depending on the land-use class, which can cause class-specific deficiencies in the texture histograms that can help distinguish between classes.

The shadow detection parameters for optimal shadow removal are set as the global threshold and Wellner's k (w=29) where the maximum land-use classification accuracy is achieved. The shadow detection parameters and corresponding optimal land-use classification results are compared in Table 4.1 for basic correction, fine correction, global threshold masking and LAT masking. Cohen's

κ is the primary measure of classification accuracy, but the ratio of true positive classifications to all classified instances is also shown, as it is a more common accuracy measure. The null-hypothesis that κ and $\max(\kappa)$ have the same mean is tested with Welch's t-test for populations with potentially unequal variances. The value of interest here is the p-value, which indicates the probability of observing a statistic at least as extreme given that the null-hypothesis is true. If this probability is high then there is good evidence that shadow removal makes no statistically significant impact on the classification accuracy, but this is under the assumption that the same-date experiment accuracies and the across-date experiment accuracies are both normally distributed and i.i.d.

Table 4.1. Shadow removal comparison between no removal (T=0, k=100) and optimal removal ($\arg\max_T(\kappa)$) in terms of Cohen's κ and land-use classification accuracy. The p-value of Welch's t-test with a null-hypothesis stating that κ and $\max(\kappa)$ have the same mean is shown in the final column.

	Cohen's κ			**Classification accuracy**		
	No removal	**Optimal removal**		**No removal**	**Optimal**	
	κ ($\mu \pm \sigma$)	$\max(\kappa)$	$\arg\max_T(\kappa)$	Accuracy	Max acc.	p-val
	(T=0,k=100)	($\mu \pm \sigma$)		(T=0,k=100)		
GLCM						
Same-date						
Basic correction	0.887±0.025	0.887±0.025	T =0	0.920±0.017	0.920±0.017	1.000
Fine correction	0.887±0.025	0.891±0.027	T =15	0.920±0.017	0.922±0.018	0.860
Global masking	0.887±0.025	0.887±0.025	T =0	0.920±0.017	0.920±0.017	1.000
LAT masking	0.887±0.025	0.887±0.025	k =100	0.920±0.017	0.920±0.017	1.000
Across-date						
Basic correction	0.786±0.069	0.786±0.069	T =0	0.847±0.050	0.847±0.050	1.000
Fine correction	0.786±0.069	0.831±0.060	T =20	0.847±0.050	0.880±0.043	0.196
Global masking	0.786±0.069	0.837±0.063	T =30	0.847±0.050	0.884±0.044	0.151
LAT masking	0.786±0.069	0.855±0.075	k =10	0.879±0.050	0.897±0.052	0.081
LBP						
Same-date						
Basic correction	0.894±0.016	0.894±0.017	T =5	0.923±0.012	0.924±0.012	0.985
Fine correction	0.894±0.016	0.894±0.017	T =5	0.923±0.012	0.924±0.012	0.975
Global masking	0.894±0.016	0.912±0.020	T =20	0.923±0.012	0.937±0.015	0.224
LAT masking	0.894±0.016	0.904±0.059	k =40	0.884±0.012	0.930±0.043	0.785
Across-date						
Basic correction	0.770±0.061	0.772±0.061	T =5	0.837±0.043	0.838±0.043	0.961
Fine correction	0.770±0.061	0.796±0.083	T =20	0.837±0.043	0.856±0.057	0.502
Global masking	0.771±0.061	0.809±0.071	T =20	0.837±0.043	0.865±0.048	0.270
LAT masking	0.771±0.061	0.812±0.070	k =30	0.865±0.043	0.865±0.052	0.230

Since there are only four same-date experiments and eight across-date experiments the sample counts are low, which can exaggerate the p-value. However, if these conditions are met, there is good evidence in the last column of Table 4.1 that there is no real accuracy difference for most shadow removal techniques at a standard significance level of 0.05. Across-date experiments do have lower p-values than same-date experiments and both global masking and LAT masking with GLCM across-date experiments can reject the null-hypothesis at a significance level of 0.15.

Shadow differences are significantly aggravated in the case of taller buildings, and the shadow mask has the potential to occupy a relatively large portion of the image area. This might occur at lower solar elevations and shadow masking would then be a poor choice of shadow removal in view of the large area that would have to be masked. Alternatively, shadow correction would retain the entire image area for texture calculation, at the expense of the inaccuracies caused by posterization.

Differing viewing angles may also cause significant changes in the texture of scenes with medium- and high-rise buildings, and paired with the shadow differences a loss in settlement classification accuracy is expected with texture features. The amplified anisotropy of land surfaces with tall buildings requires a specialized approach, but for the relatively flat texture of the settlement types considered the texture features perform well.

4.6.3.2 Classified land-use map

An across-date classification of the study area as imaged on d_1 was performed with global shadow masking using $T = 30$ and GLCM features, where the classifier was trained with all of the d_2 data. A mean accuracy of 88.4% was achieved after 10 experiments, of which a classification instance is depicted in Figure 4.13, which was obtained with majority voting from redundant tile cover of each polygon. The classification results are relatively accurate for the FSB, OIS and NBU classes when compared to Figure 4.11, but there is confusion between the FS class and the FSB and OIS classes primarily owing to the small FS training data size. There is also underlying similarity between the FS and FSB classes that is hard to separate based on the texture features that were used.

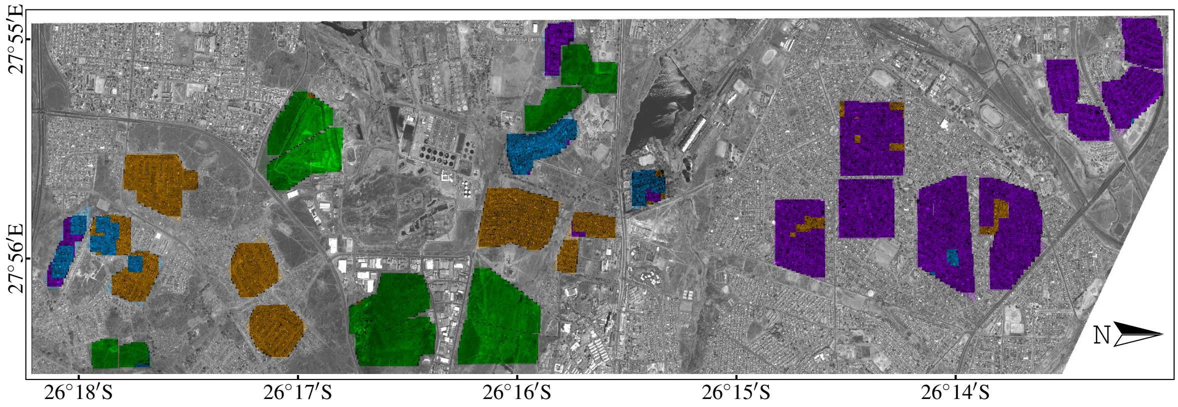


Figure 4.13. A second image of the study area, acquired on 18 October 2005, with thematic classification using training data from another date. Class polygons outlined include (🟪) FSB, (🟦) FS, (🟫) OIS, and (🟩) NBU. Panchromatic background courtesy of DigitalGlobe™.

4.6.3.3 Confusion analysis

There are four same-date experiments and eight across-date experiments involving the two acquisitions, each with dual area separation. The sum of confusion tables over all experiments pertaining to either same or across-date scenarios is given in Table 4.2 for GLCM land-use classification over the Soweto dataset using no shadow removal and global threshold masking. The objective of the confusion analysis is to determine the effect of shadow removal on classifier confusion, which is achieved through a comparison of shadow removal and global threshold masking.

Table 4.2. Confusion table without shadow removal and with shadow removal.

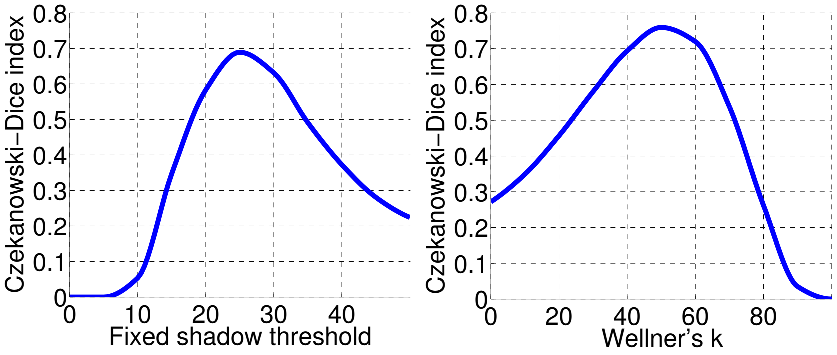
	Global threshold masking				**No removal**			
Same-date								
True label	**Predicted label**							
	FS	FSB	OIS	NBU	FS	FSB	OIS	NBU
FS	144	32	34	0	161	29	20	0
FSB	15	508	13	0	18	512	6	0
OIS	16	45	211	2	31	15	228	0
NBU	0	0	5	487	0	0	1	491
Across-date								
True label	**Predicted label**							
	FS	FSB	OIS	NBU	FS	FSB	OIS	NBU
FS	258	113	44	5	183	139	92	6
FSB	67	986	19	0	163	897	12	0
OIS	15	52	470	11	13	51	471	13
NBU	2	0	8	974	1	1	11	971

Higher classification accuracy generally corresponds to lower overall classifier confusion, so the expectation is that there will be reduced confusion where shadow removal is effective in improving classification accuracy. For the same-date experiments there is overall greater confusion with shadow removal, since there are no significant shadow profile differences that can benefit from correction. The largest confusion in this case is between FSB and OIS, probably because both land-use types include the presence of shacks.

For across-date experiments there is overall reduced confusion between classes if shadow removal is used, except for the OIS class. The OIS class is confused slightly more with the FS and FSB classes. The largest reduction in confusion with shadow removal is between the FS and FSB classes, which involve the largest structures and thus potentially the largest shadow profile differences.

4.6.3.4 Shadow detection accuracy

Two main shadow detectors are used in this chapter, namely global thresholding and LAT. Unsupervised threshold parameter optimization is not used here, but the full parameter range is rather investigated to properly detect land-use classification accuracy improvements and to determine the relationship between shadow detection accuracy and land-use accuracy improvements. Histogram matching is employed in the experiments of this chapter to enable the shadow detection accuracy evaluation in Figure 4.14 for both acquisitions under the same shadow detector parameters.



(a) Global threshold masking. **(b)** LAT masking.

Figure 4.14. Shadow detection accuracy in terms of the mean Czekanowski-Dice index using the shadow mask groundtruth data given in Appendix A.

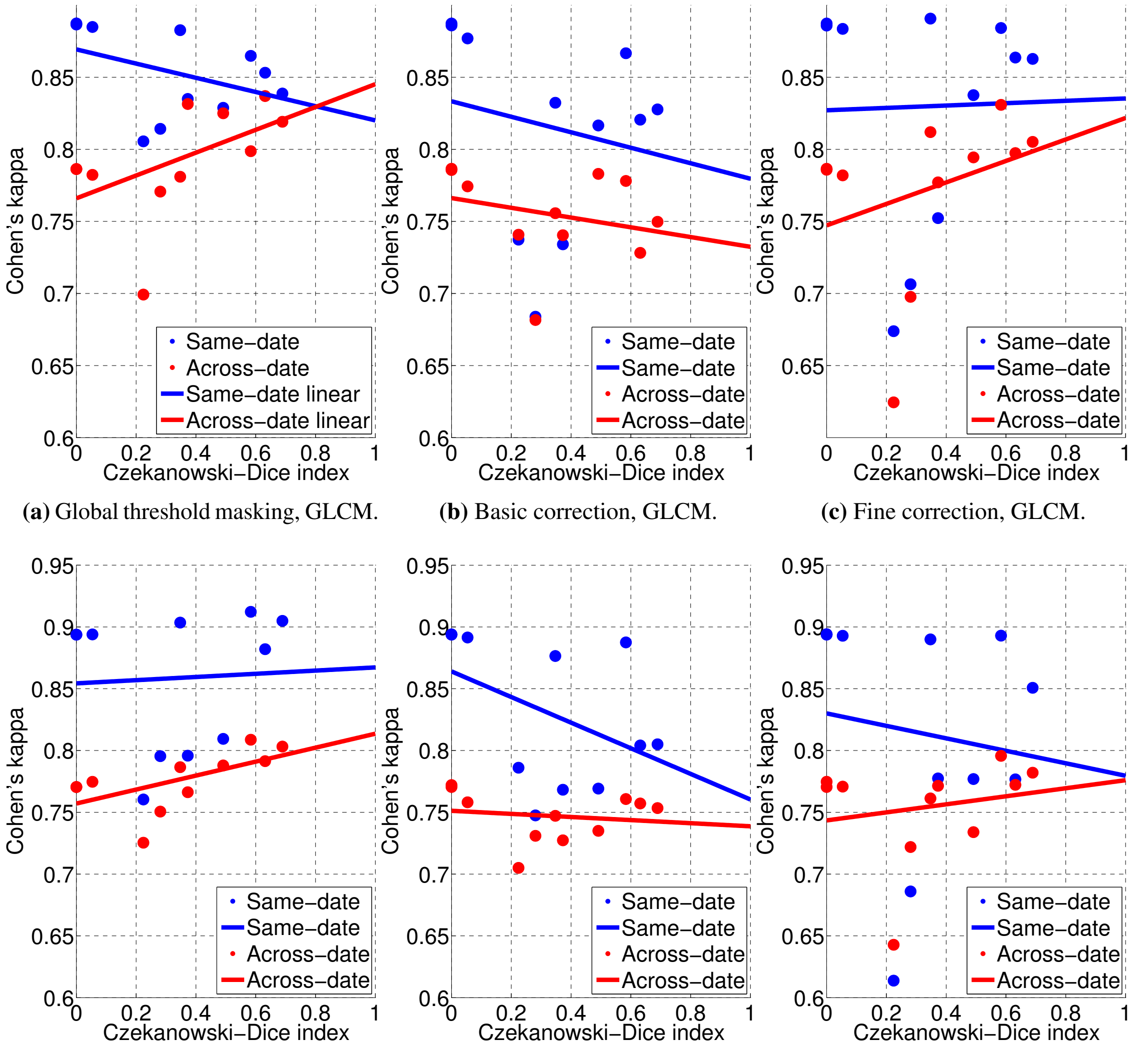
The histogram matching allows for the use of a single shadow detector parameter as opposed to dual parameters, while improving overall shadow detection accuracy when compared to not using histogram matching. Figure 4.14(a) indicates that the optimal global threshold is 25 for both histogram matched acquisitions, and Figure 4.14(b) shows that a k of 50 is optimal for LAT in terms of the shadow detection accuracy measured with the Czekanowski-Dice agreement index using a subset of groundtruth shadows.

4.6.3.5 Correlation analysis: land-use classification accuracy vs. shadow detection accuracy

It is instructive to evaluate the relationship between land-use classification accuracy and shadow detection accuracy in the case of shadow removal, since this relationship can help characterize the role of shadow removal in improving land-use classification accuracy through input modification for addressing dataset shift. Figure 4.15 indicates the correlation between land-use accuracy and shadow detection accuracy, respectively measured with Cohen's κ and the Czekanowski-Dice index.

Linear function fittings are also included for each set of experiments over the independent shadow detection parameter range, which provides a visual comparison of the correlation differences between same-date and across-date experiments. The distinction between same-date and across-date experiments is important in this relationship analysis, since the expectation is that shadow removal should be a lot more effective in the case of shadow profile differences, as is the case for across-date experiments.

The observation that the land-use classification accuracy and shadow detection accuracy correlation is positive for across-date experiments indicates that land-use classification could possibly improve as a result of more accurate shadow removal. If the correlation is negative, then more accurate shadow



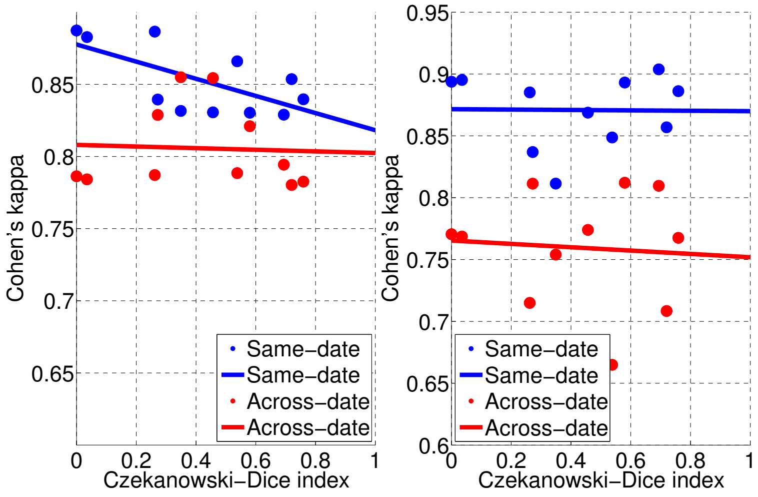
(a) Global threshold masking, GLCM. **(b)** Basic correction, GLCM. **(c)** Fine correction, GLCM.

(d) Global threshold masking, LBP. **(e)** Basic correction, LBP. **(f)** Fine correction, LBP.

Figure 4.15. Linear fittings for the correlation between land-use classification accuracy (Cohen's κ) and shadow detection accuracy (Czekanowski-Dice index) for global threshold masking, basic and fine corrections with same and across-date distinctions.

removal could be responsible for deteriorating class distinction. For all global thresholding shadow removal algorithms in Figure 4.15 it appears that the across-date correlations are more positive than the same-date correlations. In Figure 4.16 the correlations are given for LAT shadow masking.

The Pearson's correlation coefficients (denoted by ρ) of the same-date and across-date data points for fixed thresholding paired with the various shadow removal algorithms are shown in Table 4.3. A Pearson's correlation coefficient measures the linear dependence between two variables and is defined as the covariance of the two variables divided by the product of their standard deviations. The correlation coefficient ranges from -1 to 1, where a value of 1 implies that a linear equation can perfectly describe the relationship between the two variables. A linear relationship where one variable increases as another decreases is characterized by a correlation coefficient of -1. If the value is 0 then



(a) LAT masking, GLCM.

(b) LAT masking, LBP.

Figure 4.16. Linear fittings for the correlation between land-use classification accuracy (Cohen's κ) and shadow detection accuracy (Czekanowski-Dice index) for LAT shadow masking (w=29) with same and across-date experiment distinctions.

there is no discernible linear relationship between the two variables.

Table 4.3. Correlation coefficients and statistical significance.

	GLCM		LBP	
	Basic correction			
	Same	Across	Same	Across
Correlation ρ	-0.195	-0.264	-0.437	-0.153
Correlation p-val	0.565	0.433	0.179	0.654
Compare Δp-val	**0.357**		**0.108**	
	Fine correction			
	Same	Across	Same	Across
Correlation ρ	0.026	0.313	-0.130	0.191
Correlation p-val	0.941	0.349	0.703	0.574
Compare Δp-val	**0.046**		**0.046**	
	Global threshold masking			
	Same	Across	Same	Across
Correlation ρ	-0.404	0.513	0.057	0.591
Correlation p-val	0.218	0.107	0.867	0.056
Compare Δp-val	**0.018**		**0.011**	
	LAT masking			
	Same	Across	Same	Across
Correlation ρ	-0.654	-0.051	-0.015	-0.076
Correlation p-val	0.029	0.881	0.966	0.824
Compare Δp-val	**0.127**		**0.441**	

The statistical significance of each correlation was also determined in terms of a p-value calculated using the Student's t-distribution for a transformation of the correlation. A p-value is the probability of obtaining a test statistic at least as extreme as the one actually observed, assuming that the null-hypothesis is true. The null-hypothesis here is that there is no correlation between shadow detection accuracy and settlement classification accuracy, i.e. $\rho = 0$, formulated with the aim of rejecting the null-hypothesis to show that there is in fact correlation between the variables.

A p-value less than a low significance level of 0.05 would imply that the probability of obtaining a test statistic at least as observed given that the null-hypothesis is true is less than 5%, which would constitute strong evidence against a zero correlation. In that case it can be stated that the observed correlation is probably not zero with statistical significance at a significance level of 0.05. The most significant correlations occur for GLCM same-date experiments with LAT masking and for both GLCM and LBP across-date experiments with global threshold masking at a significance level of approximately 0.1.

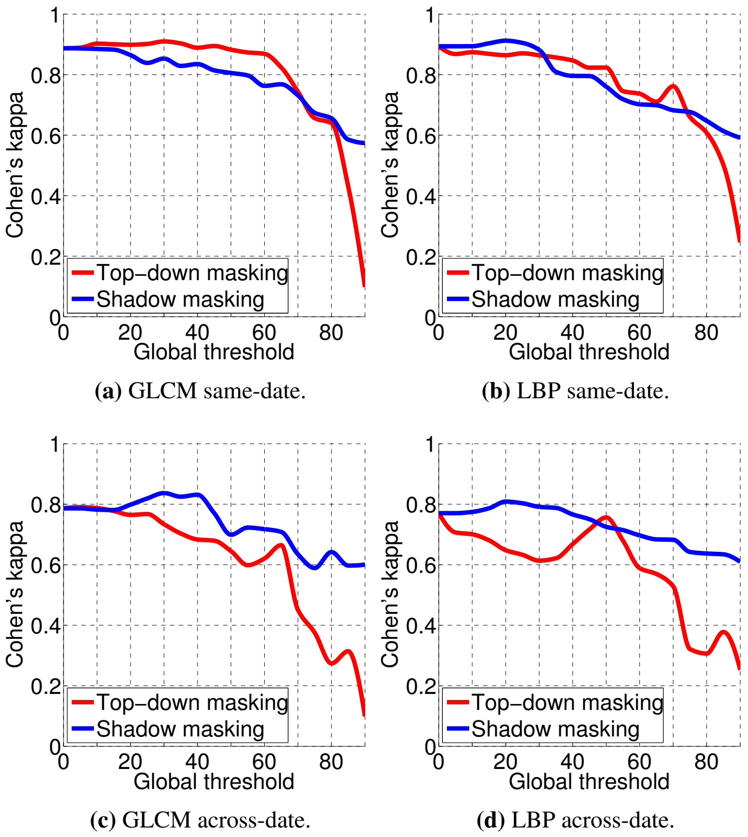
Emphasis is placed on the null-hypothesis test of equal correlation for the same-date and across-date samples of a specific shadow removal algorithm, and it is seen that there is a relatively small probability Δp-val of obtaining correlation differences at least as extreme as were observed, given that the null-hypothesis is true. The Δp-val is a notation used here to refer to the p-value of the test statistic that measures the correlation difference between same-date and across-date experiments. The Δp-val is calculated according to the correlated correlation coefficients case given by Meng et al. [169], where the Czekanowski-Dice similarity coefficient is the shared variable between the same-date and across-date samples.

At a significance level of 0.1 it can be seen in Table 4.3 that the global thresholding masking and fine correction shadow removal algorithms for both GLCM and LBP features produced a significant correlation difference between same-date and across-date experiments. This supports the hypothesis that more accurate shadow removal produces greater improvements in land-use classification accuracy in the case of across-date experiments with notable shadow profile differences between the training and testing datasets. At a significance level of 0.15 there is also evidence that there is a correlation difference for GLCM experiments with LAT shadow masking.

4.6.3.6 Shadow masking vs. top-down masking

Global threshold shadow masking effectively removes low-intensity pixels from the feature calculations, which may cause some class-specific feature alterations that can enhance distinction between classes. A comparison is therefore made between shadow masking and top-down masking, which is a masking unrelated to shadows and unrelated to the potential factor of dataset shift. The goal of this comparative experiment is to determine whether it is likely the shadow removal specifically that enhances class distinction, rather than simply intensity-based masking. This comparison is shown in Figure 4.17 with differentiation between GLCM and LBP texture features, as well as between same-date and across-date experiments.

It is the expectation that in across-date experiments the shadow masking will lead to greater land-use



(a) GLCM same-date.

(b) LBP same-date.

(c) GLCM across-date.

(d) LBP across-date.

Figure 4.17. Land-use classification accuracy.

classification accuracy improvements than with masking unrelated to across-date image differences. The alternative masking that is used in this experiment is top-down masking, which masks out pixels higher than a given global threshold. The top-down masking thus masks above a threshold, as opposed to shadow masking which masks below a threshold. The top-down masking can thus accentuate shadow profile differences and increase the effect of the dataset shift, thus potentially leading to a reduction in land-use classification accuracy. In Figure 4.17(c) and 4.17(d) it is observed that shadow masking improves accuracy at certain thresholds, whereas top-down masking does not manage to improve accuracy over the accuracy with no masking. This provides further evidence that it is shadow masking specifically that could be responsible for land-use classification accuracy improvements.

4.7 CONCLUSION

Effective feature variance occurs during across-date land-use type classification owing to differences in viewing and illumination geometry. The purpose of this study was to reduce dataset shift through input modification by detecting and removing shadow differences that can cause detrimental variation in texture features, and to test its efficacy in improving land-use type classification accuracy. A variant of histogram equalization, proposed by Shu and Freeman [75] and used by Sarabandi et al. [112], from the intensity domain (paragraph 2.5.2.1) of the shadow restoration taxonomy in Table 2.4, has been employed for shadow correction.

The relationship between shadow detection accuracy and increases in land-use type classification accuracy was investigated experimentally and statistically. It was observed that there is a definite stronger trend with across-date classification where more accurate shadow removal resulted in a typically larger improvement in the measured land-use accuracy compared to same-date experiments.

GLCM shadow correction and LBP shadow masking have improved settlement classification accuracy in same-date experiments, and both GLCM and LBP shadow correction and shadow masking can improve settlement classification accuracy in across-date experiments. The most statistically significant improvements in settlement classification accuracy were seen for GLCM across-date LAT masking, GLCM across-date global threshold masking, GLCM across-date fine shadow correction, LBP same-date global threshold masking, LBP across-date LAT masking and LBP across-date global threshold masking.

A confusion analysis revealed that the largest reduction in confusion was between the FS and FSB classes, probably because these classes typically involve the largest structures and thus the largest shadow profiles. Correlation between settlement type classification accuracy and shadow detection accuracy showed statistically significant differences between same-date and across-date experiments for both GLCM and LBP with fine shadow correction and global threshold masking.

Top-down masking was used as a control test to obtain further evidence that the land-use classification accuracy improvements are related to shadows in particular, which was seen in the results where top-down masking could not improve classification accuracy at all, whereas shadow masking could. These results support the theory that it is the shadow removal specifically that improves classification accuracy, and that while increases in same-date accuracies were witnessed, the main benefit lies in across-date classification situations.

The following chapter investigates the manifold reduction component of a manifold alignment framework for dataset shift reduction.

CHAPTER 5 WEIGHTED AGGLOMERATIVE CLUSTERING

5.1 CHAPTER OVERVIEW

The main goal of this chapter is to illustrate the ability of weighted clustering to invoke partitionings that more closely resemble a targeted groundtruth classification than with standard unweighted clustering. This weighted agglomerative clustering can be used as a form of manifold reduction, which is typically required for computationally feasible manifold matching. Texture feature sample weighting based on region salience is used in an unsupervised setting to influence clustering to produce clusters that favor prominent homogeneous land-use occurrences. Clustering accuracy is measured in absolute and relative terms for multiple multimodal datasets over a complex land-use configuration in terms of a groundtruth classification promoting regions with greater textural regularity.

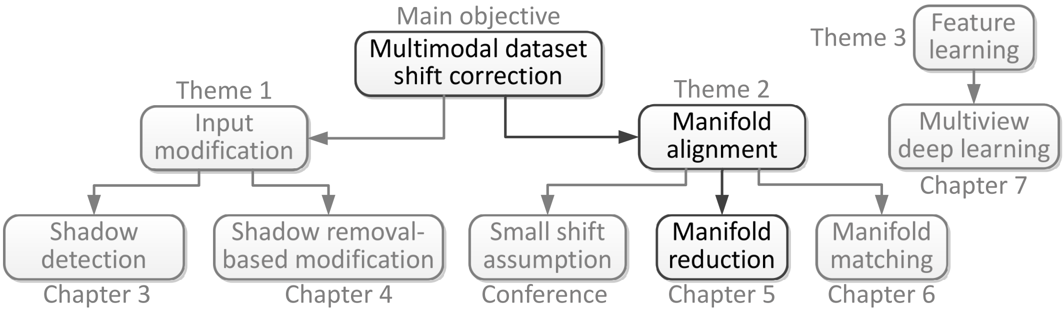


Figure 5.1. Indication of where this chapter fits into the thesis.

Multiscale, multichannel GLCM features are used with feature space compression via principal component analysis (PCA) and adaptive scale selection based on minimizing spatial feature variance is employed. Spatial feature variance at the optimal scale for each mapping unit is converted into a sample weighting later used for weighted clustering. Unweighted clustering is tested with standard linkages and compared against weighted clustering with weight-sensitive linkages.

A full dendrogram is constructed by the agglomerative clustering so that weakly supervised cardinality determination can be investigated. Weighted generalizations of numerous major internal validation indices are employed and a reduced computational time implementation with input sampling is proposed. The hypothesis is tested that maximal weight input selection involves samples in the internal index calculation that improve cardinality decision accuracy compared to random selection.

5.1.1 Contributions

1. It is demonstrated that weighted clustering leads to improved accuracy compared to unweighted clustering in terms of a target classification.

2. Internal validation algorithms with reduced computational time are used to allow for the processing of larger datasets.

3. Texture scale-selective sample composition that handles the multiscale presentation of land-use types is shown to improve clustering accuracy as well.

4. Input truncated implementations of weighted generalizations of multiple internal validation indices are deconstructed experimentally, and maximal weight input selection is shown to be responsible for improved clustering cardinality decisions.

5. A knee-point accentuating extremum filter and a suppressed first derivative alternative to disruption interpretation of internal validation criteria are further contributions and their efficacy is illustrated experimentally.

5.2 PROBLEM STATEMENT

Target clustering produces partitions desired for specific applications [170] and an unsupervised method of informing the clustering process via sample weighting is investigated in this study for complex land-use segmentation problems. Artificial separation can be achieved in a feature space by applying scalar real weights to each sample and thereby simulating varying density where high weight regions can attract cluster barycenters and distinguish between salient classes over regions of lower weight density.

Sample weighting enriches the clustering problem space in a manner that the addition of another feature dimension cannot replicate, at the cost of providing some engineered weight attribute. Dimensional expansion of a feature space accessed primarily through distance metrics requires the addition of a powerful new feature to separate classes and define class centers better. Even with the inclusion of such features the use of proper sample weighting still enhances separability and localization of target classes.

The hypothesis that weighted clustering can attract cluster centroids toward classes with certain target properties is investigated by translating the desired properties into an applied weighting of samples. Samples that more strongly possess target traits receive higher salience attribution. The clustering objective is the promotion of land-use types with greater textural regularity in order to provide target definition in an otherwise complex and poorly separated feature space. In scenarios with confusing and blending land-use occurrences a segmentation that emphasizes prominent homogeneous land-use classes is more useful than an arbitrary result producing no clear semantic separation.

A complex land-use clustering problem is considered with numerous land-use classes displaying irregular boundaries and many regions displaying confused class attribution. The clustering scenario is complicated by the blending of classes in regions bordering multiple different classes, which erodes the class separations in the feature space. The multimodal aspect of the study firstly enlarges the number of problem datasets by the inclusion of multiple high-resolution multispectral acquisitions from different imaging vehicles acquired under various acquisition conditions over a considerable timespan. This enables the proper demonstration of the ability of the clustering system to produce accurate results under a notable range of input characteristics, such as differences in image sharpness, texture presentation, color profiles, illumination and viewing geometry, atmospheric conditions, and phenological cycles.

5.2.1 Hypotheses

1. Weighted clustering can attract cluster centroids toward classes with certain target properties, since agglomeration centroids gravitate toward higher weight regions.

2. Textural regularity as a target property can attract clusters toward more salient classes, since the target classification promotes classes with greater textural regularity.

3. Multiscale dimensionality reduction can be obtained with the principal components of only one particular scale, since the same groundtruth underlies the textures and the expectation is that sample importance will correspond well over the different scales.

4. Clustering linkages that incorporate sample weightings in the agglomerated cluster centroid calculation, but also the effective pairwise cluster dissimilarities, will provide more accurate clusterings, since the sample weightings have a greater impact on agglomeration.

5. Maximal weight input selection involves samples in the internal index calculation that improve cardinality decision accuracy compared to random selection, because the samples possess target characteristics and a greater affinity to the groundtruth classification.

5.2.2 Research questions

1. What manifold reduction strategy can be employed to create clustering separation in a poorly separated feature space?

2. How can a relevant sample weighting be obtained for texture-based land-use classification in remote sensing images?

3. What approach should be followed to obtain a scale-selective feature space when dimensionality reduction is used?

4. Which agglomerative clustering linkage is best for weighted clustering?

5. How can the optimal number of clusters be found, given a weighted feature space and hierarchical dendrogram?
6. Which internal validation indices perform best in a weighted clustering setting, and what role do sample weightings play in cardinality fitness?

5.3 RELATED WORK

5.3.1 Weighted clustering

Clustering is a fundamental data analysis form and is popularly used especially in remote sensing applications, but the application of weighted clustering is less prevalent. Clustering algorithms are characterized as weight-sensitive, weight-robust or weight-considering in [171]. Weight-sensitive partitional clustering algorithms include *k*-means, *k*-medoids, *k*-median, and Min-sum, and weight-sensitive hierarchical algorithms include Ward's method and bisecting *k*-means.

Weighted model-based clustering is used for remote sensing imaging analysis in [172] where higher salience is granted to samples with relatively less noise. In other applications, such as brain magnetic resonance imaging segmentation [173], Gaussian smoothing of the feature space is used before fuzzy *c*-mean clustering to reduce noise and feature weighting is then used to bootstrap the clustering.

5.3.2 Multiscale features

Hierarchical segment merging has been used for high-resolution multiscale segmentation [174] and the fusion of samples from different feature scales based on combining the pixel means in all scales has been proposed [175] for urban change detection. The boost-classifier given in [176] is composed of weak classifiers, one for every segmentation level, which detects scale and feature sets best suited for a given dataset. Here classification results are combined in lieu of sample fusion to deal with the multiscale segmentation problem.

Sliuzas et al. specify two spatial scales for urban remote sensing analysis, namely a local scale concerned with the recognition of objects such as individual buildings and a strategic scale, which covers general land-use such as residential, commercial and industrial areas at city-block scale [177]. The increased availability of high-resolution imagery has produced an increased interest in local scale spatial classification and analysis that combine spatial and spectral information [178]. Land-use classification has been shown to benefit from higher spatial resolutions corresponding to sub-meter ground sampling distances [179].

OBIA is commonly used for accurate classification at local scale [180], but per-pixel interpretation of remotely sensed images is usually performed at a strategic scale because of the poor results of per-pixel-based methods at local scale [181]. However, pixel-based texture information can in some cases yield good accuracy in urban land-use classification, especially at the strategic scale [182]. Graphs of local variance have been used to anticipate the optimal scale parameter for forest classification in multispectral Ikonos images [183]. Multiscale features are determined in this study by selecting feature scales corresponding to the scale with minimum feature variance.

5.3.3 Land-use segmentation

Land-use maps are important for environmental monitoring such as pollution and deforestation analysis, but also for socio-economic purposes such as urban and transportation planning [184]. The three main methods of generating land-use or land-use land-cover maps from spatial, spectral, textural, contextual and relational information derived from remotely sensed images are pixel-based, object-based and field-based methods [185]. In general an image object is a group of pixels with similar textural or spectral properties [186] and size, shape, color, texture, pattern and shadow properties [187]. Homogeneous image objects are defined in OBIA primarily through segmentation and are then classified based on spatial, spectral, textural, contextual and relational properties [188].

Object-based classification consists of two main steps, namely image segmentation and image classification based on the derived spatially compact image objects [189]. OBIA is the best choice for efficient and rapid classification of medium-resolution optical remote sensing images [190], since representation in terms of discrete objects satisfy human understanding better [191] and because OBIA produces more visually consistent segmentations than pixel-based methods [192]. Urban features are typically defined by contextual arrangements, rather than individual pixel characteristics [193].

Wei et al. showed that OBIA can achieve a 94.4% classification accuracy for land-use, land-cover classification in medium-resolution imagery from forest-agriculture ecotones, which outperforms the 61.4% accuracy of a pixel-based approach [194]. Panchromatic pixel-based classification at local scale to classify roads, residential and commercial buildings has outperformed OBIA in one study when training and test samples were extracted from separate reference objects [179].

Segmentation algorithms are generally grouped into three types, namely point-based, edge-based and region-based techniques [195] of which region-based methods are widely used because of a guarantee of closed regions [196]. Numerous remote sensing segmentation algorithms have been proposed, but only a few are robust under operational settings [197]. The proprietary multi-resolution segmentation algorithm of eCognition Developer is a region-based technique that has shown strong performance in many remote sensing problems [198].

Multi-resolution segmentation can in general be controlled by shape, scale, smoothness and compactness parameters, of which the scale parameter is the most important, as it has a direct impact on the subsequent segmentation [199]. Qualitative visual inspection is commonly practiced in an iterative optimization approach [180], but because of its labor intensiveness and subjectivity of inspection the better approach involves supervised and unsupervised methods [200].

Measuring segmentation quality can be based on the number of closed regions in the segmentation, the number of pixels in each region and the color error of each region [201], or measures of region contrast and uniformity can be combined in a quality measurement function [202]. An unsupervised segmentation quality measure that combines a spatial autocorrelation indicator, which detects region separability, and a variance indicator that expresses region homogeneity has also been proposed for region-growing segmentation [192]. Classification reliability tends to be lower for urban settings than in rural areas because of the high spectral variability of urban materials and the occurrence of spectral signatures in multiple different urban classes [203].

5.4 AGGLOMERATIVE HIERARCHICAL CLUSTERING

5.4.1 Clustering linkages

Agglomerative hierarchical clustering with different linkages is generalized under the infinite family of algorithms defined by the recursive Lance-Williams function given by [204]

$$d(i \cup j,k) = \alpha_i d(i,k) + \alpha_j d(j,k) + \beta d(i,j) + \gamma |d(i,k) - d(j,k)|. \tag{5.1}$$

Single, complete, average, median, centroid and Ward linkages are all different measurements of dissimilarity between a pair of existing clusters, and the Lance-William update coefficients for each linkage are given in Table 5.1. The clustering algorithm has the objective of finding the pair of clusters with the smallest dissimilarity, according to a particular linkage, so that the pair of clusters can be merged into one new cluster [204].

Table 5.1. Agglomerative hierarchical clustering linkages described in terms of Lance-Williams coefficients [204]. The weight or number of points in a cluster i is given by $|i|$, which corresponds to the cluster weight W_i in the following subsection.

Linkage	**Lance-Williams coefficients**		
Single linkage	$\alpha_i, \alpha_j = \frac{1}{2}$	$\beta = 0$	$\gamma = -\frac{1}{2}$
Complete linkage	$\alpha_i, \alpha_j = \frac{1}{2}$	$\beta = 0$	$\gamma = \frac{1}{2}$
Average linkage (UPGMA)	$\alpha_i, \alpha_j = \frac{\lvert i\rvert}{\lvert i\rvert+\lvert j\rvert}$	$\beta = 0$	$\gamma = 0$
McQuitty's method (WPGMA)	$\alpha_i, \alpha_j = \frac{1}{2}$	$\beta = 0$	$\gamma = 0$
Median linkage (WPGMC)	$\alpha_i, \alpha_j = \frac{1}{2}$	$\beta = -\frac{1}{4}$	$\gamma = 0$
Centroid linkage (UPGMC)	$\alpha_i, \alpha_j = \frac{\lvert i\rvert}{\lvert i\rvert+\lvert j\rvert}$	$\beta = -\frac{\lvert i\rvert\lvert j\rvert}{(\lvert i\rvert+\lvert j\rvert)^2}$	$\gamma = 0$
Ward linkage	$\alpha_i, \alpha_j = \frac{\lvert i\rvert+\lvert k\rvert}{\lvert i\rvert+\lvert j\rvert+\lvert k\rvert}$	$\beta = -\frac{\lvert k\rvert}{\lvert i\rvert+\lvert j\rvert+\lvert k\rvert}$	$\gamma = 0$

5.4.2 Lance-Williams clustering algorithm

The Lance-Williams recurrence formula is used to update all the pairwise cluster distances that are effectively changed owing to the newest agglomeration, and the dissimilarity updates are performed after each agglomeration. Initially the input feature space consists of a set of feature samples or points that form the first singleton clusters at the start of the clustering algorithm. The pairwise cluster dissimilarities are normally initialized to the squared Euclidean distance between the two clusters in a cluster pair. At each agglomeration the new cluster center is determined, as shown in Table 5.2.

Table 5.2. Cluster center updates and cluster dissimilarity equivalents and initialisations [204]. A cluster center $\mathbf{g}_i$ is a d-dimensional vector and the Euclidean distance is given by $||\cdot||$.

Linkage	**Cluster center of $(i \cup j)$**	**Dissimilarity between cluster centers $\mathbf{g}_i$ and $\mathbf{g}_j$**
Median linkage	$\frac{\mathbf{g}_i+\mathbf{g}_j}{2}$	$\lVert\mathbf{g}_i - \mathbf{g}_j\rVert^2$
Centroid linkage	$\frac{\lvert i\rvert\mathbf{g}_i+\lvert j\rvert\mathbf{g}_j}{\lvert i\rvert+\lvert j\rvert}$	$\lVert\mathbf{g}_i - \mathbf{g}_j\rVert^2$
Ward linkage	$\frac{\lvert i\rvert\mathbf{g}_i+\lvert j\rvert\mathbf{g}_j}{\lvert i\rvert+\lvert j\rvert}$	$\frac{\lvert i\rvert\lvert j\rvert}{\lvert i\rvert+\lvert j\rvert}\lVert\mathbf{g}_i - \mathbf{g}_j\rVert^2$

5.5 CARDINALITY DETERMINATION

5.5.1 Internal validation indices

An internal validation index is a quantitative measure for evaluating the quality of a clustering, and are usually formulated in terms of the within-cluster and between-cluster dispersion and scattering. A number of different internal validation indices listed in Table 5.3 have been formulated in the literature, and a good review and comparison of these indices can be found in [41] and [205]. The asymptotic computational complexities are specified in Table 5.3, in addition to the references in the literature.

5.5.2 Weighted generalization

The internal validation index is generally used as a relative measure to find the optimal number of clusters or optimal cardinality, which is done by determining the index for a range of cardinalities and choosing the optimal cardinality. The hypothesis is that a more relevant clustering quality can be measured, to detect clusterings with improved salience balance over the clusters, if the internal indices directly incorporate sample weightings.

Weighted clustering and weighted generalizations for point-biserial correlation, Hubert's Gamma, Hubert's D, Hubert's C, Silhouette, Calinski-Harabasz and Pseudo R^2 internal validation indices are described in [40]. However, several more of the well-known indices [41] are generalized for weighted samples (included in Appendix C); this forms part of the contribution of this chapter.

5.5.3 Internal index interpretation

Hierarchical clustering produces a dendrogram that stores the clusterings for all cardinalities from N (the number of samples) down to one cluster (the dataset). The dendrogram structure thus allows for the computation of internal index interpretations at different cardinalities without having to recompute partitions. Usually a cardinality range $1 \leq k \leq K$ is chosen with a maximum considered cardinality K that is appropriate for the given scenario. The optimal cardinality according to a given internal validation index can normally be identified as the cardinality with a notable disruption or an extremum

Table 5.3. List of the internal validation indices considered in this chapter with citations and asymptotic computational complexities.

Internal index	Citation	Computational complexity
Baker-Hubert's Gamma	[206]	$O(dN^2 + N^4/k)$
Ball-Hall	[207]	$O(dN)$
Banfield-Raftery	[208]	$O(dN)$
C-index	[209]	$O(N^2(d + \log_2 N))$
Calinsky-Harabasz	[210]	$O(dN)$
Davies-Bouldin	[211]	$O(d(k^2 + N))$
Det_Ratio	[212]	$O(d^2N + d^3)$
G+	[213]	$O(dN^2 + N^4/k)$
GDI	[214]	$O(dN^2)$
Ksq_DetW	[215]	$O(d^2N + d^3)$
Log_Det_Ratio	[212]	$O(d^2N + d^3)$
Log_SS_Ratio	[216]	$O(d(k^2 + N))$
McClain-Rao	[217]	$O(dN^2)$
PBM	[218]	$O(d(k^2 + N))$
Point-Biserial	[219]	$O(dN^2)$
Ratkowsky-Lance	[220]	$O(d^2N)$
Ray-Turi	[221]	$O(dN)$
S_Dbw	[222]	$O(dNk^2)$
Scott-Symons	[212]	$O(d^2kN + d^3k)$
Silhouette	[223]	$O(dN^2)$
Tau	[224]	$O(dN^2 + N^4/k)$
Trace_W	[225]	$O(dN)$
Trace_WiB	[226]	$O(d^2N + d^3)$
Wemmert-Gançarski	[227]	$O(dNk)$

value.

5.5.3.1 Extremum interpretation

There are two main interpretations of internal validation indices, namely optimization-like criteria or an **extremum** interpretation and a difference-like criteria or a **disruption** interpretation [205]. The extremum interpretation finds a cardinality for an internal index $\mathcal{C} = \{c_k : 1 \leq k \leq K\}$ with either the global $\max(\mathcal{C})$ or $\min(\mathcal{C})$ in the signal, depending on the nature of the internal validation index, e.g. the Ball-Hall index measures the mean cluster dispersion over a partition, which should ideally be a minimum to ensure that the clusters are as compact as possible.

5.5.3.2 Disruption interpretation

The disruption interpretation can indicate more relevant cardinalities, where extremum interpretations sometimes tend toward global extrema outside a usable cardinality range. The disruption interpretation is performed as an extremum interpretation on a function $g(\mathcal{C})$ of the internal index $\mathcal{C}$, so a conversion is explicitly done to move derivative magnitude in an extremum interpreted domain. Vendramin et al. [205] instantiate Milligan and Cooper's difference between hierarchical dendrogram levels [228] with the following three possible realizations of the disruption conversion function $g(\mathcal{C})$:

$$g(\mathcal{C}) = \left|\frac{d\mathcal{C}}{dk}\right| \tag{5.2}$$

$$g(\mathcal{C}(k)) = \left|\frac{d\mathcal{C}(k)}{dk}\right| - \left|\frac{d\mathcal{C}(k+1)}{dk}\right| \tag{5.3}$$

$$g(\mathcal{C}(k)) = \left|\frac{d\mathcal{C}(k)}{dk} \Big/ \frac{d\mathcal{C}(k+1)}{dk}\right|. \tag{5.4}$$

The absolute values ($|\cdot|$) of the derivatives are proposed by Vendramin et al. [205] in Equations 5.2 and 5.3. The use of the absolute of the first derivative of the internal index in Equation 5.2 allows for the detection of the cardinality at the single greatest change in the index with an extremum interpretation. However, by using the absolute function there can be no differentiation between decremental or incremental changes, which could be an important distinction depending on the specific interal validation index used.

Similarly, in Equation 5.3 a difference is made between the absolute of two successive derivatives, but a peak formed by two line sections with derivatives of 1 and -1 will result in a zero due to the absolute function and can thus not be detected as a disruption. Equation 5.4 partially addresses this destructive superposition by allowing for a ratio comparison of successive derivative magnitudes, although the absolute function again removes discrimination between disruption polarities and the disruption measure is non-linear. These potential issues are avoided in this chapter by removing the absolute function and by using the proper second derivative to convert the internal index for extremum interpretation, according to

$$g(\mathcal{C}) = \frac{d^2\mathcal{C}}{dk^2}. \tag{5.5}$$

5.6 METHODOLOGY

The weakly supervised clustering system in Figure 5.2 receives a rectangular three-channel red-green-blue (RGB) image and produces internally a compressed salience-weighted scale-selective GLCM feature space. The weighted samples are clustered with a weighted hierarchical agglomerative clustering algorithm and an internal validation estimate of the optimum clustering cardinality is made. The output is a clustering with the number of clusters specified by the cardinality estimate.

5.6.1 Texture feature extraction

The first 13 of Haralick's GLCM features [166] are used to construct a multiscale feature set of six scales × three RGB channels × 13 features with six GLCM square window widths (s_1 to s_6) of 50, 100, 150, 200, 300 and 400 m all centered on 25 m-wide square mapping units for a total of $N = X \times Y$ mapping units. GLCM calculations are performed on the underlying high-resolution image at the full resolution provided by each respective imaging vehicle. GLCM pairs are used in all cardinal and ordinal directions with respective ℓ_1-norms of one and two, and the features are then averaged over

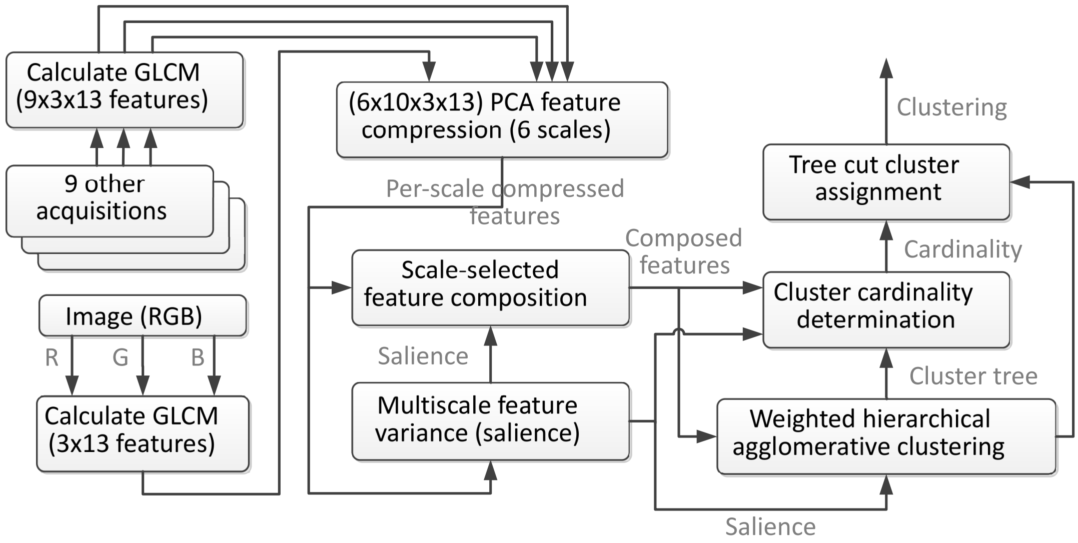

Calculate GLCM (9x3x13 features)
9 other acquisitions
Image (RGB)
R
G
B
Calculate GLCM (3x13 features)
(6x10x3x13) PCA feature compression (6 scales)
Per-scale compressed features
Scale-selected feature composition
Salience
Multiscale feature variance (salience)
Composed features
Clustering
Tree cut cluster assignment
Cardinality
Cluster cardinality determination
Cluster tree
Weighted hierarchical agglomerative clustering
Salience

Figure 5.2. Segmentation methodology based on salience-weighted clustering.

the four spatial relationships. The effect of different GLCM window sizes on features is shown in Figure 5.3.

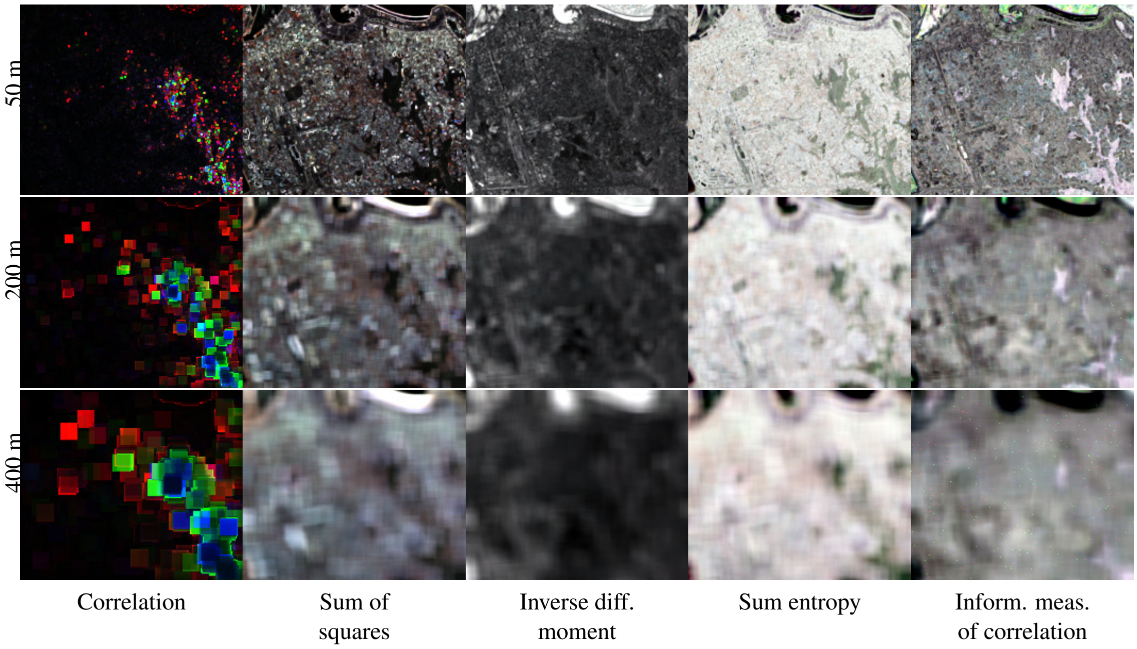


Figure 5.3. Selected multiscale GLCM features for 50, 200 and 400 m windows over the Rio de Janeiro area (D_8).

5.6.2 Dimensionality reduction

The multiscale feature sets $A^{(1)}$ to $A^{(6)}$ for a given acquisition are $N \times p$ matrices with $p = 3 \times 13$ feature dimensions. A PCA feature space compression is performed on the concatenation of all 10 dataset acquisitions at each of the six feature scales, meaning six scales × 10 acquisitions × N samples

$\times$ p features to produce a joint PCA over a $60N \times p$ feature space. The concatenated $60N \times p$ feature space is column-normalized and centered to standard $\mathcal{N}(0,1)$ distributions before performing the PCA.

The PCA produces a $60N \times p$ score that is split into the respective scales and acquisitions in order to calculate optimal sample feature scales and associate sample weightings. The multiscale feature sets $A^{(1)}$ to $A^{(6)}$ are thus compressed with a joint PCA to give $\mathcal{F}^{(1)}$ to $\mathcal{F}^{(6)}$. An Euclidean distance measure in the new space naturally incorporates the dimensional loadings reflecting feature importance.

5.6.3 Sample weighting

The framework in which weighted clustering is investigated requires a weight associated with each sample, which is modeled as local spatial feature variance. Spatial feature variance u for a given GLCM window size s_j is calculated as $u(s_j) = \|\text{std}(\{M\})\|_2$, namely the Euclidean norm of the standard deviation of the features $\{M \in \mathcal{F}^{(j-1)}\}$ of all the smaller scale GLCM windows that best fit inside and cover the single larger scale window s_j at a particular map location.

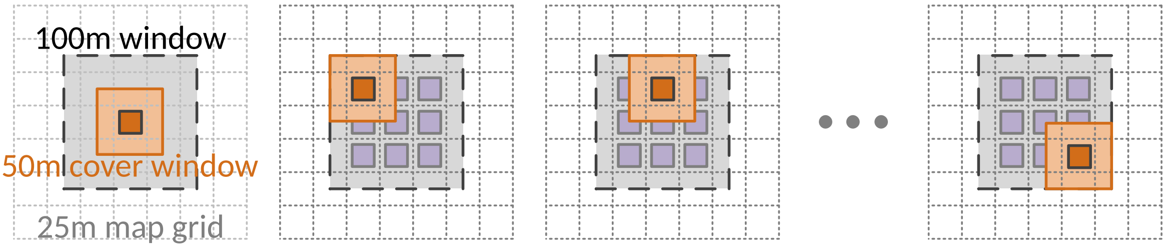


Figure 5.4. Conceptualization of sample salience for a 100 m scale calculated as the variance of the smaller 50 m scale GLCM window features, involving smaller windows that completely cover the larger scale window.

The process is illustrated in Figure 5.4 for a 100 m window size (s_j), where the smaller 50 m window (s_{j-1}) is used to find the variance. The window size $v_i = \text{argmin}_j(u(s_j))$ containing the smallest feature variance is selected for its map location i as the most appropriate feature scale and the spatial feature variance value $u_i = \min(u(s_j))$ at that scale as the associated textural regularity. $1 \times N$ vectors $U = \{u_i\}$ and $V = \{v_i\}$ respectively denote all the textural regularity values mapped to the range $[0,1]$ and the feature scales.

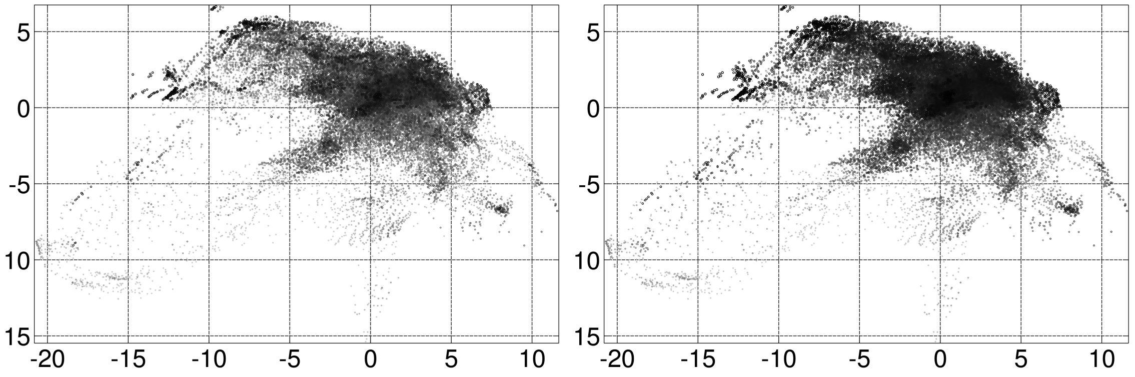
Sample textural regularity is changed into sample salience by $U \leftarrow \max(U) - U$. The probability density function of U is adjusted by moving the previous mean $\overline{U}$ to approximately 0.2 with $w_i = u_i^{\log 0.2/\log \overline{U}}$ forming the final salience or weight vector $\mathcal{W} = \{w_i\}$ when mapped to the range $[0.1,1]$. Reducing the mean weight allows fewer high-salience samples to influence the weighted clustering to a greater degree, as they will then possess relatively larger weights.

5.6.4 Multiscale feature sample composition

The multiscale texture interpretation of different land-use classes is addressed through multiscale feature sample composition based on the derived salience values. Given the optimal window size $v_i = \text{argmin}_j(u(s_j))$ for map location i, the corresponding feature in $\mathcal{F}^{(v_i)}$ is chosen to compose a new scale-selective feature space $\mathcal{F} = \{\mathcal{F}_i^{(v_i)} : \forall i\}$. The option to smooth the weightings $\mathcal{W}$ based on Gaussian kernel density estimation can be executed to produce a smoothed weighting $\mathcal{W}^{(S)} = \{w_i^{(S)} : \forall i\}$ according to

$$w_i^{(S)} = \frac{\sum_j \frac{w_j}{2\pi H_j} \exp\left(-\frac{\|\mathcal{F}_i - \mathcal{F}_j\|_2^2}{2H_j^2}\right)}{\sum_j \frac{1}{2\pi H_j} \exp\left(-\frac{\|\mathcal{F}_i - \mathcal{F}_j\|_2^2}{2H_j^2}\right)}. \tag{5.6}$$

Adaptive bandwidth $H_j = w_j \|\mathcal{F}_j - \text{knn}(\mathcal{F}_j, 100)\|_2$ can be used where $\text{knn}(\mathcal{F}_j, k)$ is the k-nearest neighbor of $\mathcal{F}_j$. The adaptive bandwidth is weighted to enlarge high-salience kernels relatively, and the k-nearest neighbor distance is used for bandwidth estimation. The smoothing result is shown in Figure 5.5 and the effect is that the mean adaptive neighborhood weighting is imparted to every sample.



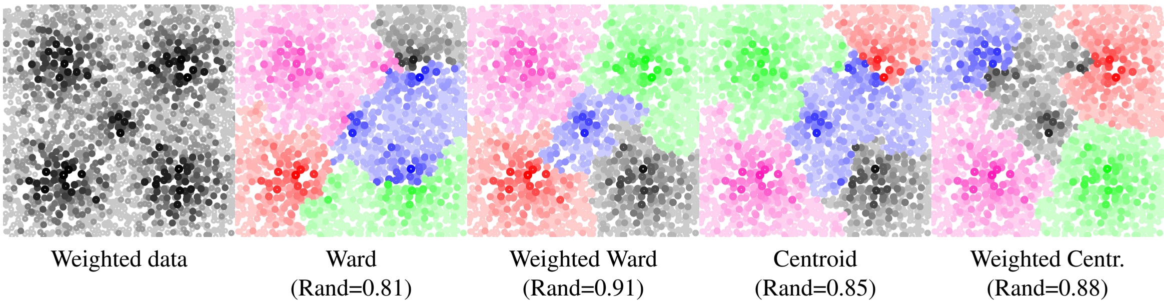
(a) Projection of $\mathcal{F}$ with $\mathcal{W}$

(b) Projection of $\mathcal{F}$ with $\mathcal{W}^{(S)}$

Figure 5.5. Effect of weight smoothing with kernel density estimation for an example dataset.

5.6.5 Weighted clustering

Hierarchical agglomerative clustering with weight-sensitive linkages is used to incorporate sample weightings into the agglomeration process. Single, complete, weighted (WPGMA) and median (WPGMC) linkages are weight-independent, while average (UPGMA), centroid (UPGMC) and Ward linkages are weight-sensitive. Weighted versions of centroid and Ward linkages are explored in this study, and the effect of sample weightings in moving cluster centroids toward higher weight regions is shown in Figure 5.6.



Weighted data

Ward (Rand=0.81)

Weighted Ward (Rand=0.91)

Centroid (Rand=0.85)

Weighted Centr. (Rand=0.88)

Figure 5.6. Toy problem as a weighted uniform distribution with partitionings (k=5) from both normal unweighted and weighted versions of Ward and centroid linkages. Clustering accuracy is shown in terms of the Rand index.

5.6.6 Cardinality determination

5.6.6.1 Weak supervision

Weighted generalizations $\overline{\mathcal{C}}$ of internal validation indices $\mathcal{C}$ are used to determine the optimal number of clusters in the weakly supervised system. It should be noted that in an otherwise unsupervised clustering system, the weak supervision is incorporated only in the cardinality determination. This weak supervision is required as a supervised selection of the top performing internal validation indices for a given clustering linkage, but the assumption is made that the top performing indices will in general always perform well.

5.6.6.2 Maximum weight input sampling

Computational time can be reduced by restricting the internal validation calculation to the largest weighing N_1/N fraction of samples per cluster, where there are a total of N dataset points and a sub-sample of size N_1. The computational time reduction is directly proportional to the given internal index complexity $O(f(N))$, such as a time of N^2 for the Silhouette index being reduced to an $(N_1/N)^2$ time for an $O(N^2)$ complexity.

5.6.6.3 Alternative internal index interpretation

A validation criterion $\mathcal{C}$ is optimized either by finding an extremum or disruption, and two alternative interpretations are proposed. For indices that tend to increase or decrease monotonically a disruption must be found to indicate a significant relative change in the criterion, usually with $\frac{d^2\mathcal{C}}{dk^2}$. The first derivative $\frac{d\mathcal{C}}{dk}$ has the potential to become relatively large, so $\frac{d}{dk}\arctan\left(\frac{d\mathcal{C}}{dk}\right)$ is proposed to produce fairer angle-based disruption comparisons.

The extremum of the disruption interpretation or of the direct criterion $\mathcal{C} = \{c_k\}$ can then be obtained by finding either the global $\max(\mathcal{C})$ or $\min(\mathcal{C})$ in the signal, depending on the nature of the criterion. If the index is optimal at a minimum then a maximum disruption $\max\left(\frac{d^2\mathcal{C}}{dk^2}\right)$ needs to be found, and vice versa. For direct extremum interpreted indices that still tend to be monotonic, a knee-point accentuating filter $f(\mathcal{C})$ is proposed to promote smaller usable cardinalities and is defined for maximum considered cardinality K by

$$f(c_i) = c_i + \sum_{j=i}^{K} c_j/(K-i+1) - \sum_{j=1}^{i} c_j/i. \tag{5.7}$$

5.7 DATA DESCRIPTION

The primary dataset characteristic requirement for this study is a compact area containing a complex arrangement of a wide variety of land-use classes that are predominantly without clearly defined boundaries. Such a site presents a difficult segmentation and clustering scenario where class/cluster boundaries become densified and disappear in the feature space because of intermediate points (areas) simultaneously containing characteristics of different classes.

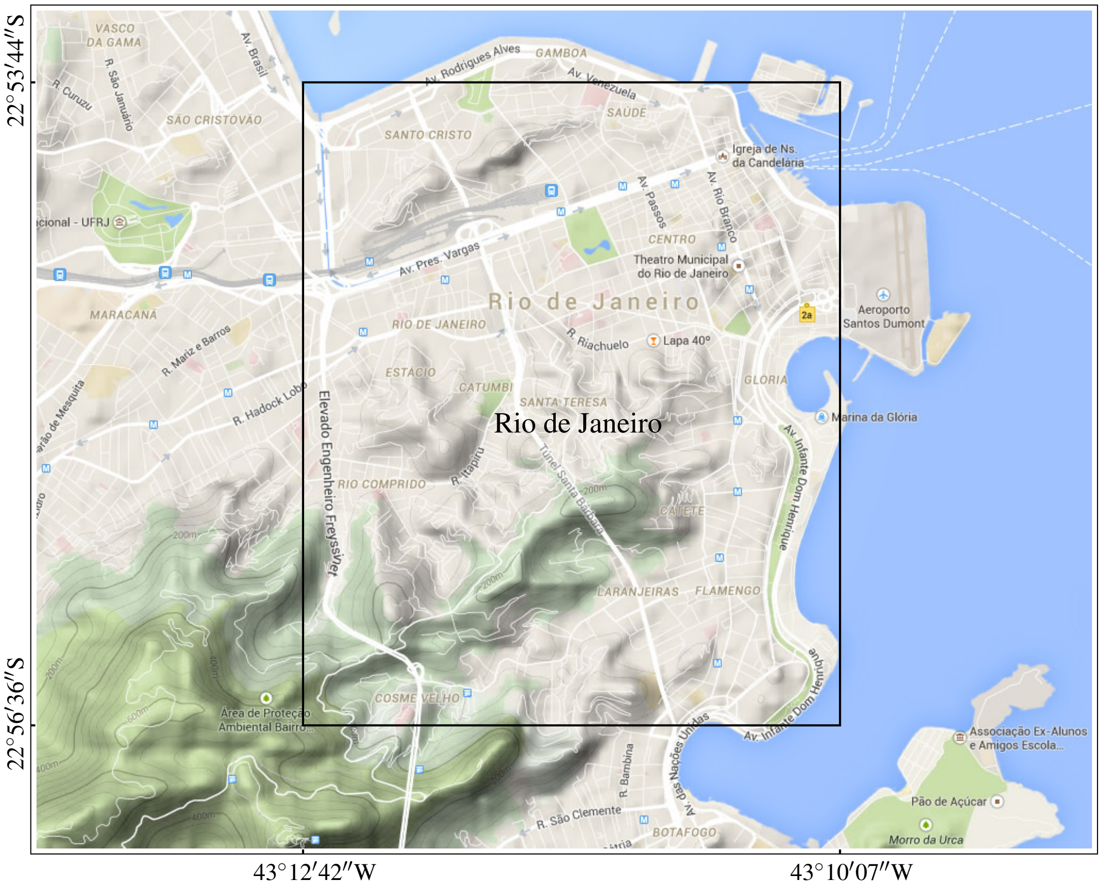


Figure 5.7. Overview map of the coregistered Rio de Janeiro image acquisitions (Courtesy of Google™ Maps).

A 5.25×4.475 km^2 rectangular area of the city proper of Rio de Janeiro was selected (see Figure 5.8) as shown in Figure 5.7, stretching from the coastline north of downtown Rio de Janeiro (22°53′44″S) to the south of Flamengo (22°56′36″S), and from the Canal do Mangue in the west (43°12′42″W) to the coastline in the east (43°10′07″W). Notable land-use types include formal and informal settlements

(19 separate favelas), industrial areas, low-rise building areas, skyscraper areas, as well as non-builtup classes featuring sea, forest, park and greenbelt areas.

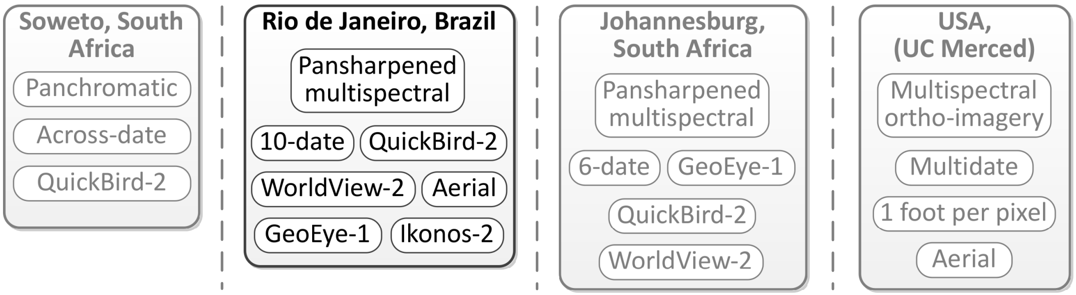


Figure 5.8. Rio de Janeiro multimodal dataset selection for this experiment.

A multimodal investigation of the study area is possible with a multitemporal and multisource selection of acquisitions, featuring three different pansharpened multispectral images for GeoEye-1 (GE1) and two images for each of the WorldView-2 (WV2), QuickBird-2 (QB2), and Ikonos-2 (IK2) satellites, as well as an additional aerial image obtained from Google™ Earth. Land-use class samples are shown for the different acquisitions in Table 5.5 and the acquisition characteristics are shown in Table 5.4. The multitemporal study period ranges from 2002 to 2013, but the isolated cases of land-use change mostly due to building upgrades have a negligible impact on the research objectives of larger-scale segmentation granularity with texture features.

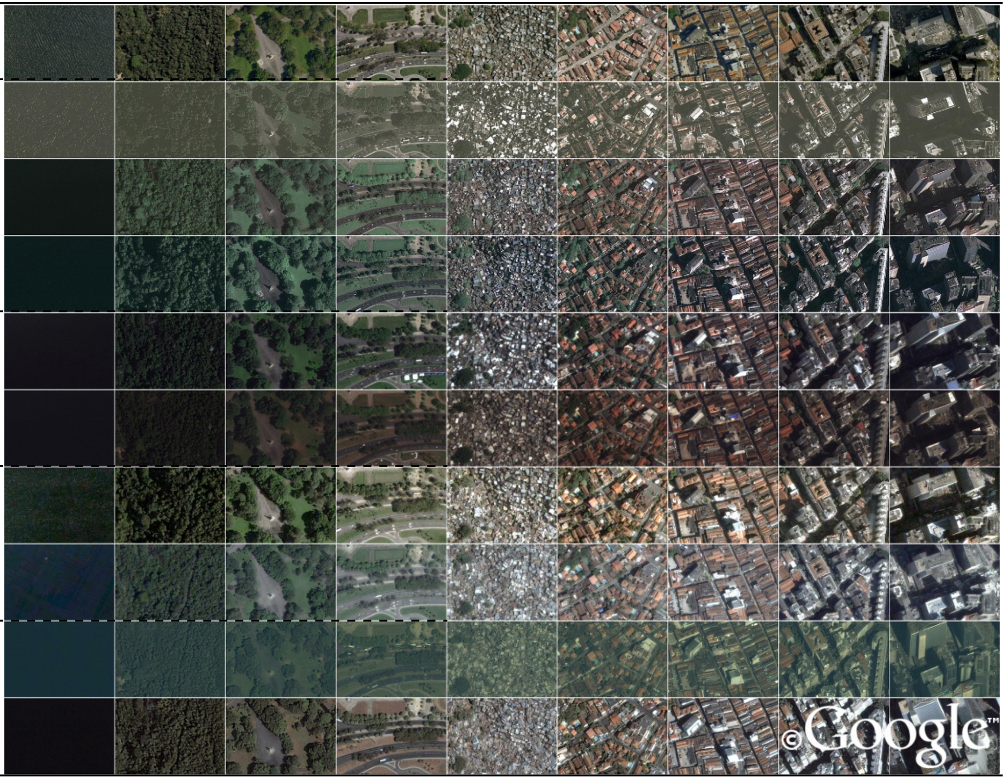
Table 5.4. Multi-satellite imagery class samples of 10 acquisitions from Aerial, GeoEye-1 (GE1), WorldView-2 (WV2), QuickBird-2 (QB2), and Ikonos-2 (IK2) satellites (Courtesy of Google™ Earth).

Acquisition	Date	Max Angle Off Nadir	Max GSD	Min Sun Elev.	Cloud Cover
D_6 (Aerial)	2009/06/25				
D_1 (GE1)	2013/05/28	11.22°	0.42 m	38.94°	0%
D_5 (GE1)	2009/09/18	25.89°	0.50 m	53.86°	5%
D_7 (GE1)	2009/05/23	18.27°	0.45 m	38.99°	0%
D_2 (WV2)	2013/04/12	20.38°	0.53 m	50.89°	0%
D_3 (WV2)	2012/09/15	12.07°	0.49 m	57.22°	1%
D_8 (QB2)	2006/05/17	12.96°	0.65 m	43.21°	5%
D_9 (QB2)	2005/09/30	10.41°	0.62 m	61.95°	6%
D_4 (IK2)	2011/10/06	20.60°	0.96 m	59.71°	0%
D_{10} (IK2)	2002/04/20	20.64°	0.92 m	47.21°	0%

5.8 EXPERIMENTAL SETUP

The objectives of the experimental analyses include quantitative and qualitative clustering accuracy comparisons between weighted and unweighted clustering, as well as determining the efficacy of internal index input truncation and different internal index interpretations. The following six main experiments are conducted to evaluate weighted clustering and weakly supervised cardinality determination:

Table 5.5. Multi-satellite imagery class samples of 10 acquisitions from Aerial, GeoEye-1 (GE1), WorldView-2 (WV2), QuickBird-2 (QB2), and Ikonos-2 (IK2) satellites (Courtesy of Google™ Earth).



Acquisition Date	Sea	Forest	Park	Green-belt	Informal	Formal	Industrial	Low-rise	Sky-scraper
D_6 (Aerial) 2009/06/25									
D_1 (GE1) 2013/05/28									
D_5 (GE1) 2009/09/18									
D_7 (GE1) 2009/05/23									
D_2 (WV2) 2013/04/12									
D_3 (WV2) 2012/09/15									
D_8 (QB2) 2006/05/17									
D_9 (QB2) 2005/09/30									
D_4 (IK2) 2011/10/06									
D_{10} (IK2) 2002/04/20									
Labeled samples	824	604	169	750	475	1155	610	2736	997

- Experiment 1: Clustering accuracy comparison
- Experiment 2: Clustering confusion evaluation
- Experiment 3: Visual clustering analysis
- Experiment 4: Cardinality decision accuracy
- Experiment 5: Internal index input truncation analysis
- Experiment 6: Internal index interpretation comparison

5.8.1 Clustering accuracy

The clustering objective is the maximization of partition similarity relative to groundtruth classes with penalization for larger cardinality decisions via $Y = \{1-k/K : 2 \leq k \leq K\}$. The groundtruth class selection prefers areas with greater texture regularity, but is not exhaustive in its labeling and only aims to provide an example of salient classes.

Table 5.6. External index definitions with concordant (*yy* and *nn*) and discordant (*yn* and *ny*) pairs and $N_T = yy + nn + yn + ny$.

External index	Formula
Kulczynski	$\frac{1}{2}\left(\frac{yy}{yy+ny}+\frac{yy}{yy+yn}\right)$
Czekanowski-Dice	$\frac{2yy}{2yy+yn+ny}$
Rogers-Tanimoto	$\frac{yy+nn}{yy+nn+2(yn+ny)}$
Folkes-Mallows	$\frac{yy}{\sqrt{(yy+yn)\times(yy+ny)}}$
McNemar	$\frac{nn-ny}{\sqrt{nn+ny}}$
Russel-Rao	$\frac{yy}{N_T}$
Hubert Gamma	$\frac{N_T\times yy-(yy+yn)(yy+ny)}{\sqrt{(yy+yn)(yy+ny)(nn+yn)(nn+ny)}}$
Jaccard	$\frac{yy}{(yy+yn+ny)}$
Sokal-Sneath-1	$\frac{yy}{yy+2(yn+ny)}$
Phi	$\frac{yy\times nn-yn\times ny}{(yy+yn)(yy+ny)(yn+nn)(ny+nn)}$
Rand	$\frac{yy+nn}{N_T}$
Sokal-Sneath-2	$\frac{yy+nn}{yy+nn+(yn+ny)/2}$

External validation indices $\mathcal{E}^{(i)}$ ($1 \leq i \leq 12$) in Table 5.6 are determined for partition cardinalities of 2 to $K = 50$ and are combined into a single objective

$$\mathcal{E} = \frac{1}{12}\sum_{i=1}^{12}\frac{\mathcal{E}^{(i)} \odot Y}{\max(\mathcal{E}^{(i)} \odot Y)}. \tag{5.8}$$

External indices of different clustering algorithms are pooled together to find $\max(\mathcal{E}^{(i)} \odot Y)$ in order to show relative accuracy in terms of the overall best index value for all algorithms compared. External indices are defined in terms of concordant (*yy* and *nn*) and discordant (*yn* and *ny*) pair counts using the groundtruth partition as reference. If a pair of samples belongs to the same cluster in the groundtruth as well as the clustering attempt then the *yy* count is incremented. If a pair belongs to different clusters in both partitions then *nn* is incremented, as it is a concordant pair. However, if a sample pair is in the same groundtruth cluster but in different clusters in the clustering attempt, then it increments the discordant pair count *yn* and vice versa for *ny*. The concordant *nn* count tends to grow with cardinality increase, which necessitates the inclusion of the cardinality penalty Y.

5.8.2 Experiment 1: Clustering accuracy comparison

Unweighted single, complete, average, median, weighted, unweighted centroid, weighted centroid, unweighted Ward and weighted Ward linkages are compared for the datasets D_1 to D_{10} in terms of $\mathcal{E}$. This accuracy measure is only relative to the best accuracy across all methods for each external index separately. Random cluster assignment is also tested for the range of considered cardinalities to serve as a baseline.

5.8.3 Experiment 2: Clustering confusion evaluation

Cross-tabulation of sample designations for the clustering and the groundtruth classes provides a view of the ability of the clustering to distinguish between groundtruth classes and to identify high-salience clusters. The interpretation of the resulting confusion table serves as a basis for qualitatively comparing the effect of weighted clustering to unweighted clustering in terms of groundtruth class distinction and high-salience clustering.

The cardinality of the clustering is set to $k = 9$ in each case, since there are nine groundtruth classes. In addition to the groundtruth samples there are also unlabeled samples, but no additional class is added for the unlabeled samples, since it is assumed that they must fit into the most similar groundtruth class. This assumption is not necessarily correct, as there might be other classes not accounted for in the groundtruth class set, but for the purpose of demonstrating the effect of sample weighting the assumption is reasonable.

The confusion table is filled column by column with each cluster distribution, where the rows account for a specific groundtruth class. Thus, considering a cluster in a column, the exact sample count breakdown of the cluster can be seen in terms of how the cluster of samples is distributed across the groundtruth classes. The objective of using the confusion table is to compare unweighted and weighted clustering in terms of cluster salience and groundtruth class separation.

5.8.4 Experiment 3: Visual clustering analysis

In addition to the previously discussed confusion evaluation, a visual analysis for evaluating clustering quality can also be conducted. The dimensionality reduction performed before clustering assists in this respect, since it compresses greater discrimination into two-dimensional representations just as it removes dimensional dependence for stronger distance measures. A PCA feature map will also be shown in addition to a salience map to visually qualify the discriminative power of the samples, based on color differentiation correlated to a groundtruth map.

The sample weightings will also be depicted for the viewer to ascertain the correspondence between groundtruth class feature space locations and high-weight pockets/regions. The main result planned here is the visual comparison between clusterings and the groundtruth classes in the dimensionally reduced feature space. The cluster locations on the geographical map are also shown, since the weight/volume of dense clusters can be hard to estimate visually in the feature space depiction.

A comparison is made between unweighted and weighted Ward clustering in terms of the clusterings as shown on the geographical maps. The comparison will be made in terms of visual clustering quality, with specific consideration of cluster spatial distribution and cluster weight/volume equality. The criteria for better visual clustering are based on a balanced spatial distribution of clusters with weight/volume concentrated at the cluster centroids and all clusters having appreciable salience.

5.8.5 Experiment 4: Cardinality decision accuracy

The 10 best internal indices with the highest clustering accuracy are determined for all the datasets combined for the best performing linkage. The cardinality mode for the selection of interpreted indices serves as the weakly supervised choice in partition size, estimating the optimal number of clusters. Each cardinality decision for the datasets is then scrutinized in terms of the clustering accuracy at that partition size. Both absolute and relative accuracy measures are provided, including Czekanowski-Dice, Jaccard and Rand criteria and the relative $\mathcal{E}$ clustering accuracy measure as main objective.

5.8.6 Experiment 5: Internal index input truncation analysis

The computational time of internal index calculation can be reduced by truncating the internal index sample input to restrict the input to a reduced size sample. The main aim of this experiment is to test the hypothesis that a maximal weight input selection will influence the validation criterion in a manner that more closely adheres to the salience-based groundtruth accuracy assessment.

This is tested by comparing the clustering accuracy $\mathcal{E}$ of maximal weight and random input selection for weighted Ward linkage. The accuracy measure is normalized relative to the maximum external indices of the clustering, since the cardinality determination only requires a performance measure relative to the best possible cardinality for a particular clustering. The effect of using weightings in the internal validation calculations is also analyzed by comparing unweighted and weighted internal criteria for both maximal weighting and random input selections.

Random input selection or random sampling is tested multiple times (≥ 10) and one-sample t-tests check whether accuracies compared to maximal weight and random input selection differ significantly for a given dataset. Maximal weight selection with weighted internal indices is then compared to other truncation configurations in terms of the percentage ratio of the number of datasets for which maximal weight selection with weighted internal indices is significantly more accurate or matches the accuracy of the other configuration, to the number of datasets in which the other configuration is significantly more accurate or matches the accuracy.

5.8.7 Experiment 6: Internal index interpretation comparison

The same relative clustering accuracy $\mathcal{E}$ described for the previous experiment is used to indicate how well an internal index interpretation reaches the maximum possible accuracy of the optimal cardinality for the given clustering dendrogram. This experiment compares extremum and disruption interpretations for all the internal validation indices considered, for the weighted Ward linkage. Extremum knee-point accentuating filtering and the alternative $\frac{d}{dk}\arctan\left(\frac{d\overline{\mathbb{C}}}{dk}\right)$ are also compared to indicate the indices for which these interpretations can improve performance.

5.9 RESULTS AND DISCUSSION

5.9.1 Experiment 1: Clustering accuracy comparison

The clustering accuracy comparison is shown in Table 5.7 in terms of the pooled joint external validation index accuracy measures per dataset, where the maximum possible accuracy is reported for every method at its optimal cardinality given a specific dataset. The purpose of pooling the external validation indices in this manner is to aggregate the different definitions of what a good clustering is and to represent the result as a scalar that could be interpreted as a comparative accuracy percentage measure.

Table 5.7. Clustering accuracy comparison of linkages for all considered Rio de Janeiro images. Standard linkages, centroid and Ward linkages are all considered.

	Rio de Janeiro acquisition										Mean
	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}	$\mu \pm \sigma$
Unweighted linkages											
Single	39.5	48.0	42.4	43.7	39.3	43.8	32.2	17.2	41.5	38.4	38.6±8.59
Random	54.8	59.6	56.8	54.9	50.5	53.0	47.6	56.8	54.5	50.0	53.9±3.63
Average	62.9	64.3	57.7	55.1	54.9	52.6	58.7	54.6	52.7	52.5	56.6±4.24
Centroid	62.8	62.5	58.6	55.1	57.0	53.4	60.6	53.7	53.9	53.3	57.1±3.80
Median	72.9	69.1	60.3	58.9	74.3	56.2	68.2	64.0	53.0	60.5	63.7±7.17
Weighted	75.5	75.6	64.7	59.0	74.8	61.9	75.3	65.8	59.0	62.2	67.4±7.14
Complete	84.0	75.1	57.8	60.8	68.9	68.0	73.3	69.4	65.4	84.3	70.7±8.77
Ward	81.7	88.1	93.2	91.6	91.9	90.3	90.5	97.3	93.4	87.9	90.6±4.15
Weighted linkages											
Centroid	62.8	63.0	58.0	55.9	55.8	54.3	59.0	53.8	52.2	54.0	56.9±3.75
Ward	99.0	94.4	91.4	98.9	99.8	85.0	92.5	99.8	93.4	99.4	95.4±4.92

Ward linkage consistently significantly outperforms the baseline of random cluster assignment (two-tailed p-val less than 0.0001 with Welch's unpaired t-test, with mean difference 95% confidence interval of 33.0 to 40.4), whereas other linkages do only slightly better. Complete linkage achieves the second highest mean clustering accuracy, followed by weighted linkage. Incorporating sample weights into the clustering with weighted clustering improves the mean clustering accuracy for both centroid and Ward linkages. In only two of the datasets, namely D_3 and D_6, the weighted Ward linkage does not outperform unweighted Ward linkages. At a significance level of 0.05, weighted Ward significantly outperforms unweighted Ward for a two-tailed p-val of 0.03 with Welch's unpaired t-test, with mean difference 95% confidence interval of 0.51 to 9.10.

5.9.2 Experiment 2: Clustering confusion evaluation

The cross-tabulation between the Ward clustering for D_5 and the groundtruth classes is shown in Table 5.8. The confusion table for weighted Ward clustering is given in Table 5.9 so that a qualitative discernment can be made about the effect of sample weighting on clustering confusion. The clustering cardinality is set at $k = 9$ and the distribution of each cluster is given per column, where each row is associated with a different groundtruth class or the remaining unlabeled samples class.

Table 5.8. Cross-tabulation of Ward cluster assignment (D_5) and groundtruth classes for $k = 9$.

	Ward cluster assignment								
Class label	1	2	3	4	5	6	7	8	9
Sea	404	213	202	0	3	0	0	2	0
Forest	0	0	0	387	0	30	187	0	0
Park	0	0	0	0	0	92	14	45	18
Greenbelt	0	0	0	0	35	657	29	0	29
Informal	0	0	0	0	0	9	48	418	0
Formal	0	0	0	0	0	0	259	896	0
Industrial	0	0	0	0	0	0	0	299	311
Low-rise	0	0	0	0	0	0	11	539	2186
Skyscraper	0	0	0	0	0	17	0	0	980
Unlabeled	138	501	83	266	873	3050	5512	11108	7739

It should be noted that the groundtruth classes may share commonalities and that the class distinction was made by a human operator for the purposes of obtaining basic class distinction in order to evaluate clustering accuracy. For example, forest, park and greenbelt have similar textural and spectral properties and clustering of samples may place these classes in the same cluster. For both normal and weighted Ward clustering the sea class is spread mainly over three clusters, because of the sea textures transforming to more extreme GLCM points that spread across the feature space extremities.

Table 5.9. Cross-tabulation of weighted Ward cluster assignment (D_5) and groundtruth classes for $k = 9$.

	Weighted Ward cluster assignment								
Class label	1	2	3	4	5	6	7	8	9
Sea	398	216	208	0	0	0	0	0	2
Forest	0	0	0	414	28	162	0	0	0
Park	0	0	0	0	92	14	1	0	62
Greenbelt	0	35	0	0	602	57	0	0	56
Informal	0	0	0	0	7	42	418	0	8
Formal	0	0	0	0	0	232	916	7	0
Industrial	0	0	0	0	0	0	161	424	25
Low-rise	0	0	0	0	0	11	283	2103	339
Skyscraper	0	0	0	0	6	0	0	741	250
Unlabeled	115	1391	122	245	2462	5519	8980	4004	6432

Weighted clustering concentrates the forest class more strongly into cluster 4 than unweighted clustering, and the park class is also concentrated more strongly with weighted clustering. The greenbelt class, however, has a higher concentration with unweighted clustering than with weighted clustering, but unweighted clustering assembles a low salience cluster (cluster 5) with a very low percentage of target class contribution. There is some confusion between forest and formal settlement classes for both clusterings, which is probably due to formal settlements frequently being surrounded by green space.

Weighted clustering produces higher saliency clusters, as evidenced by the lack of low-saliency clusters like cluster 5 for unweighted clustering. This permits weighted clustering to utilize cluster assignment better, which results in improved differentiation between industrial/low-rise/skyscraper collectives and low-rise/skyscraper collectives in clusters 8 and 9. There is stronger confusion between informal/formal settlement collectives and industrial/low-rise collectives with unweighted clustering, and industrial, low-rise and skyscraper classes do get confused in general for both unweighted and weighted clusterings.

5.9.3 Experiment 3: Visual clustering analysis

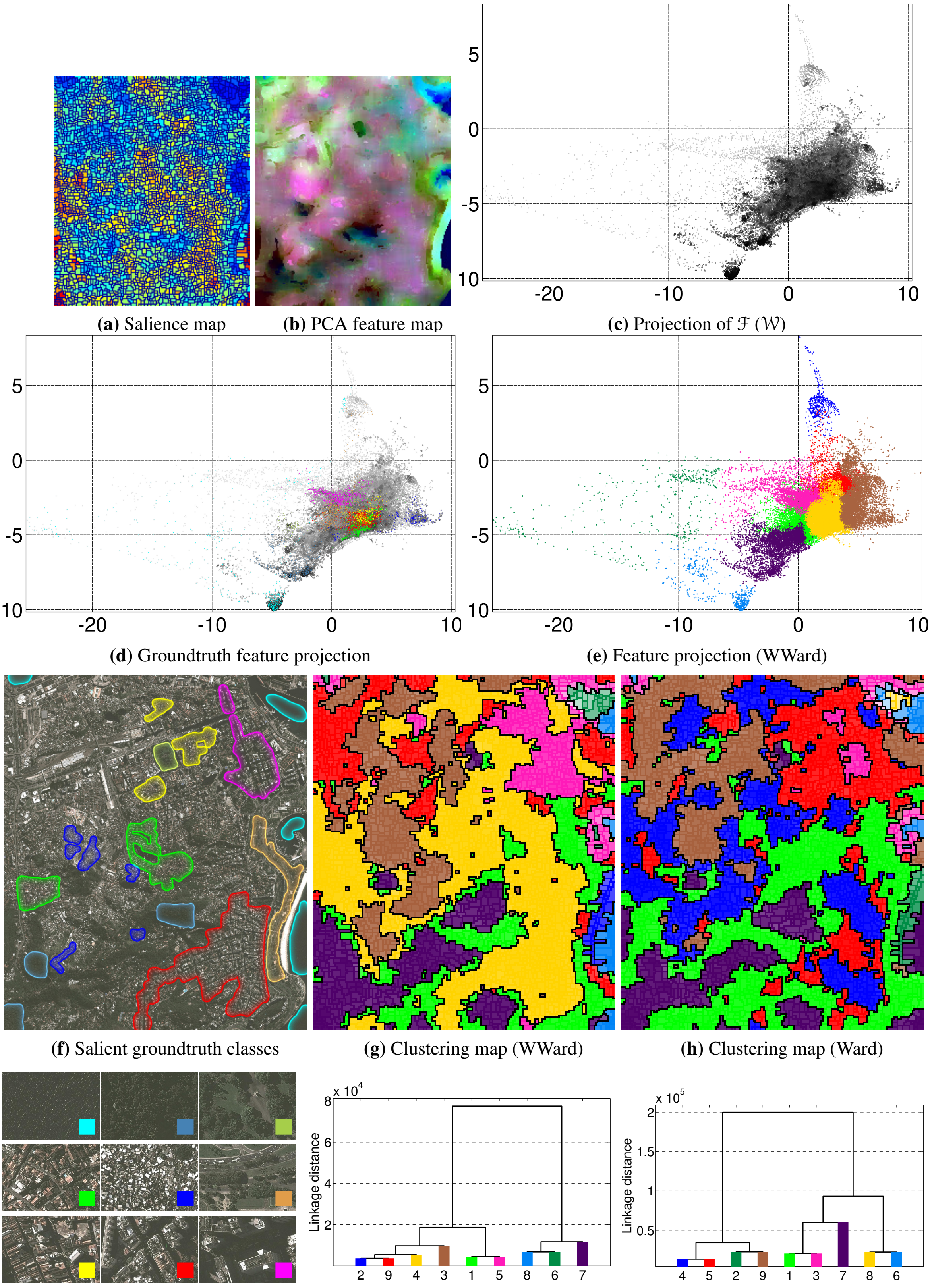
The ability of the clustering system to recreate the groundtruth classes, given the same cardinality of $k = 9$, is inspected visually in Figures 5.9, 5.10, 5.11, 5.12, 5.13 for D_1, D_3, D_7, D_9 and D_{10}. The salience maps show the adjusted salience distributions as used for the clustering, which promotes high-salience outliers to the upper end of the salience range. In some instances there are only a few high-salience outliers, such as with some sea and forest samples in Figure 5.11 for D_7.

The PCA feature maps generally show good distinction between builtup, non-builtup and water classes. The larger skyscraper class normally presents with a distinct color representation in the compressed feature display, although low-rise, industrial, formal and informal settlement classes appear less distinct. It should be noted that the color selection for the groundtruth feature projections (d) and salient groundtruth classes (f) does match, and that the color selection for the feature projection (e) matches with that of the clustering map (g). However, there is not necessarily a color scheme match between the groundtruth displays (d, f) and the clustering displays (e, g).

The core of the feature space generally has greater density and class occupancy, which can make it difficult to interpret the feature projections accurately and compare them to the groundtruth feature projection. The clustering maps allow for a more accurate assessment in this case, since they depict the clustering balance and distribution. The comparison of geographical clustering maps of clustering with weighted Ward linkage (g) and unweighted Ward linkage (h) generally reveals a more balanced cluster distribution for weighted Ward linkage.

5.9.4 Experiment 4: Cardinality decision accuracy

The performance of the internal validation index strategy for obtaining the optimal clustering cardinality is measured in Table 5.10 in percentage terms of the clustering accuracy relative to that at the real optimal cardinality for the given linkage and dataset. Absolute clustering accuracy measures are also given as Czekanowski-Dice, Jaccard and Rand criteria.



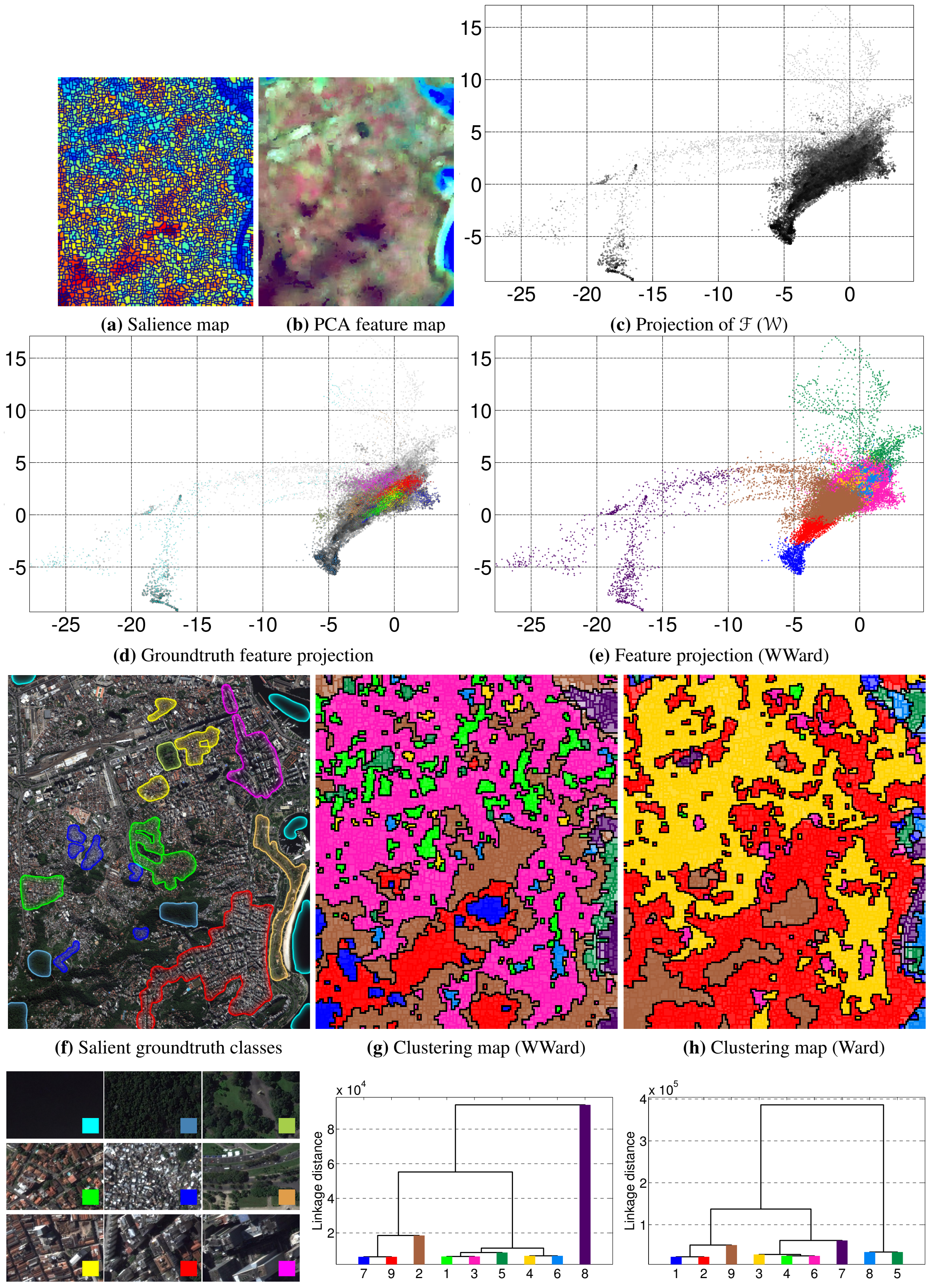
(a) Salience map **(b)** PCA feature map **(c)** Projection of $\mathcal{F}(\mathcal{W})$

(d) Groundtruth feature projection **(e)** Feature projection (WWard)

(f) Salient groundtruth classes **(g)** Clustering map (WWard) **(h)** Clustering map (Ward)

(i) Salient classes with color codes **(j)** Dendrogram (WWard) **(k)** Dendrogram (Ward)

Figure 5.9. Clustering results for Rio de Janeiro D_1 (GE1 - 2013/05/28), with batch weighted Ward for k=9 with the sample weighting $\mathcal{W}$. Multispectral images courtesy of Google™ Earth.



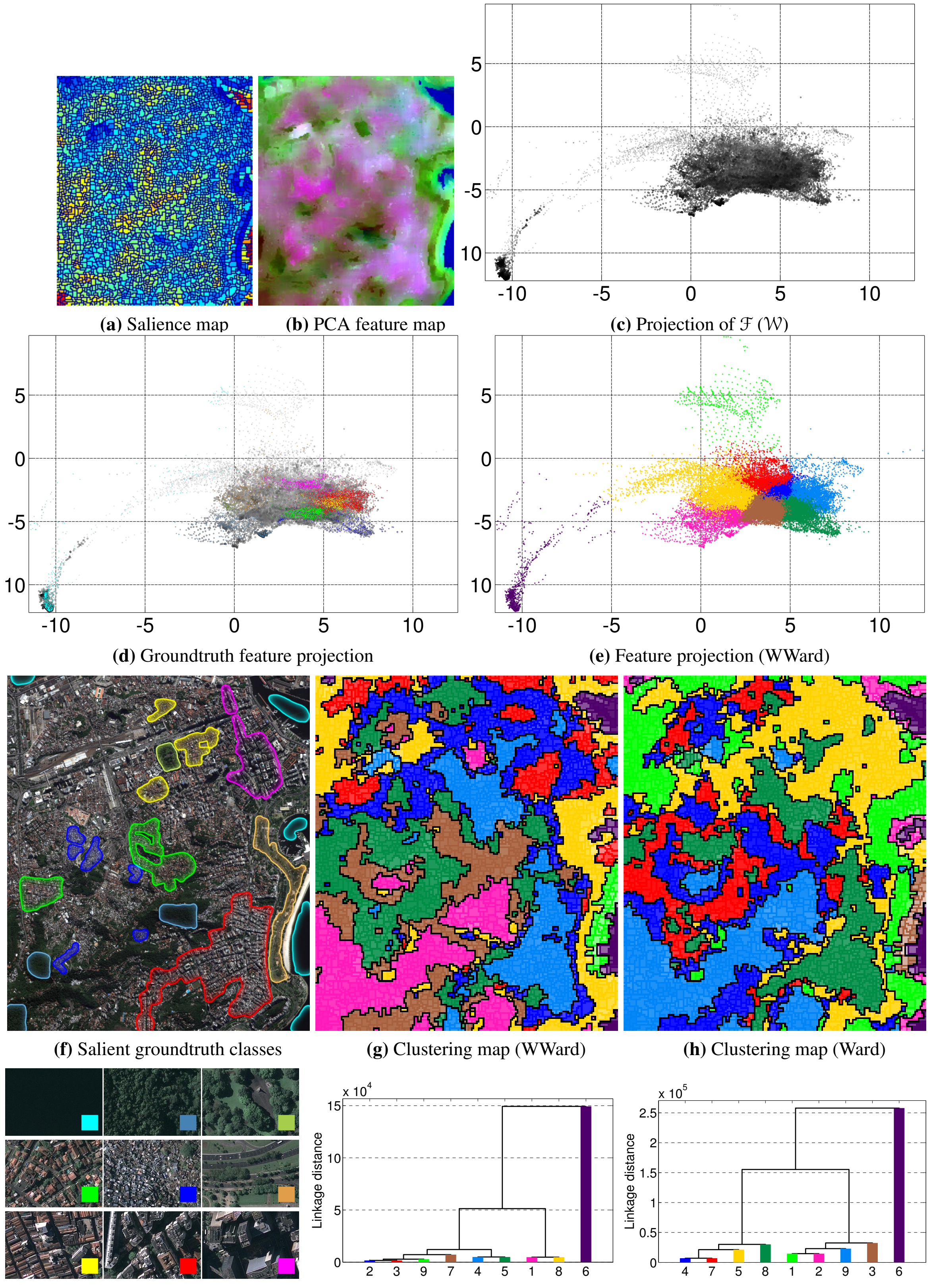
(a) Salience map **(b)** PCA feature map **(c)** Projection of $\mathcal{F}$ $(\mathcal{W})$

(d) Groundtruth feature projection **(e)** Feature projection (WWard)

(f) Salient groundtruth classes **(g)** Clustering map (WWard) **(h)** Clustering map (Ward)

(i) Salient classes with color codes **(j)** Dendrogram (WWard) **(k)** Dendrogram (Ward)

Figure 5.10. Clustering results for Rio de Janeiro D_3 (WV2 - 2012/09/15), with batch weighted Ward for k=9 with the sample weighting $\mathcal{W}$. Multispectral images courtesy of Google™ Earth.



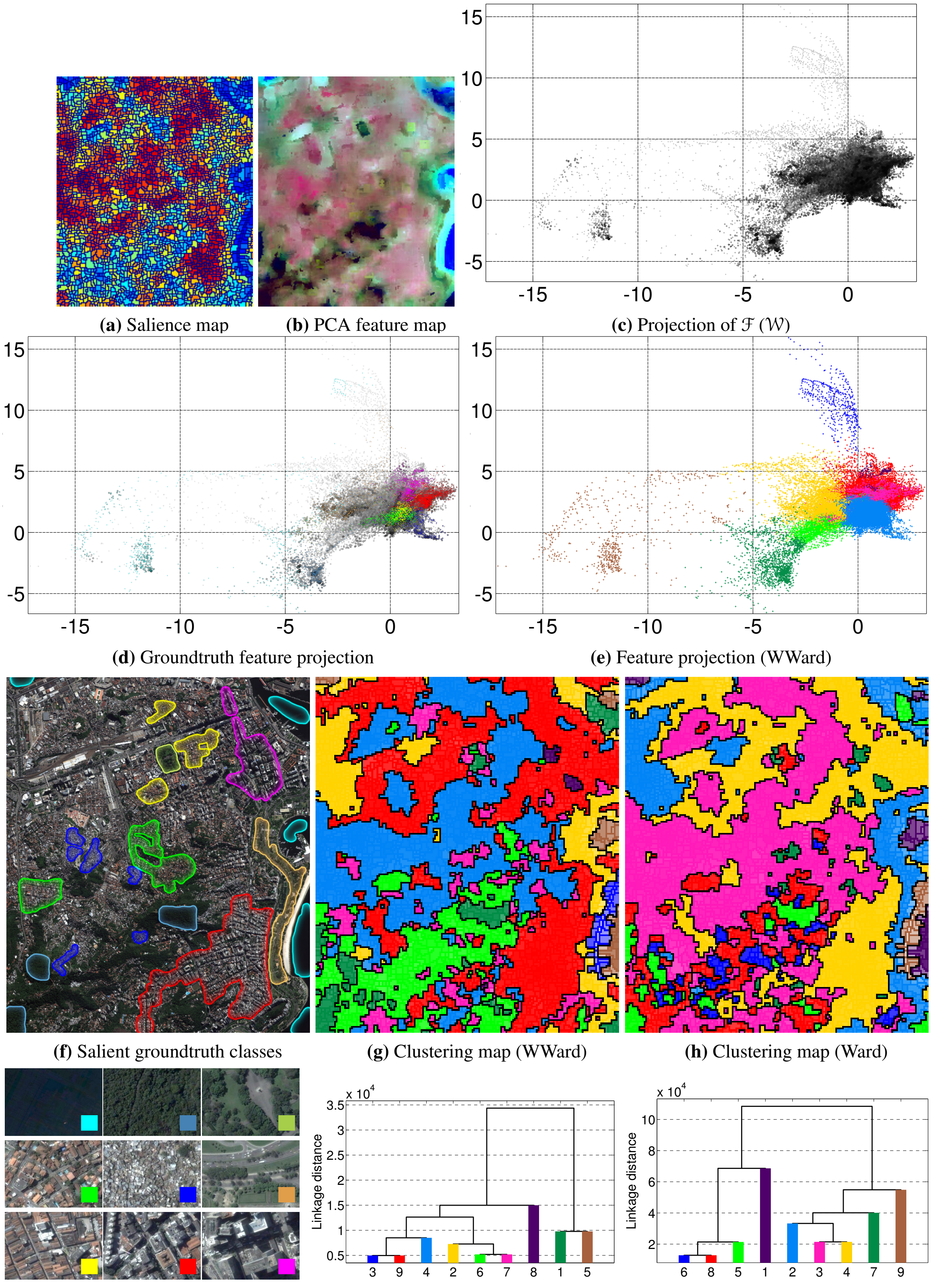
(a) Salience map **(b)** PCA feature map **(c)** Projection of $\mathcal{F}$ $(\mathcal{W})$

(d) Groundtruth feature projection **(e)** Feature projection (WWard)

(f) Salient groundtruth classes **(g)** Clustering map (WWard) **(h)** Clustering map (Ward)

(i) Salient classes with color codes **(j)** Dendrogram (WWard) **(k)** Dendrogram (Ward)

Figure 5.11. Clustering results for Rio de Janeiro D_7 (GE1 - 2009/05/23), with batch weighted Ward for k=9 with the sample weighting $\mathcal{W}$. Multispectral images courtesy of Google™ Earth.



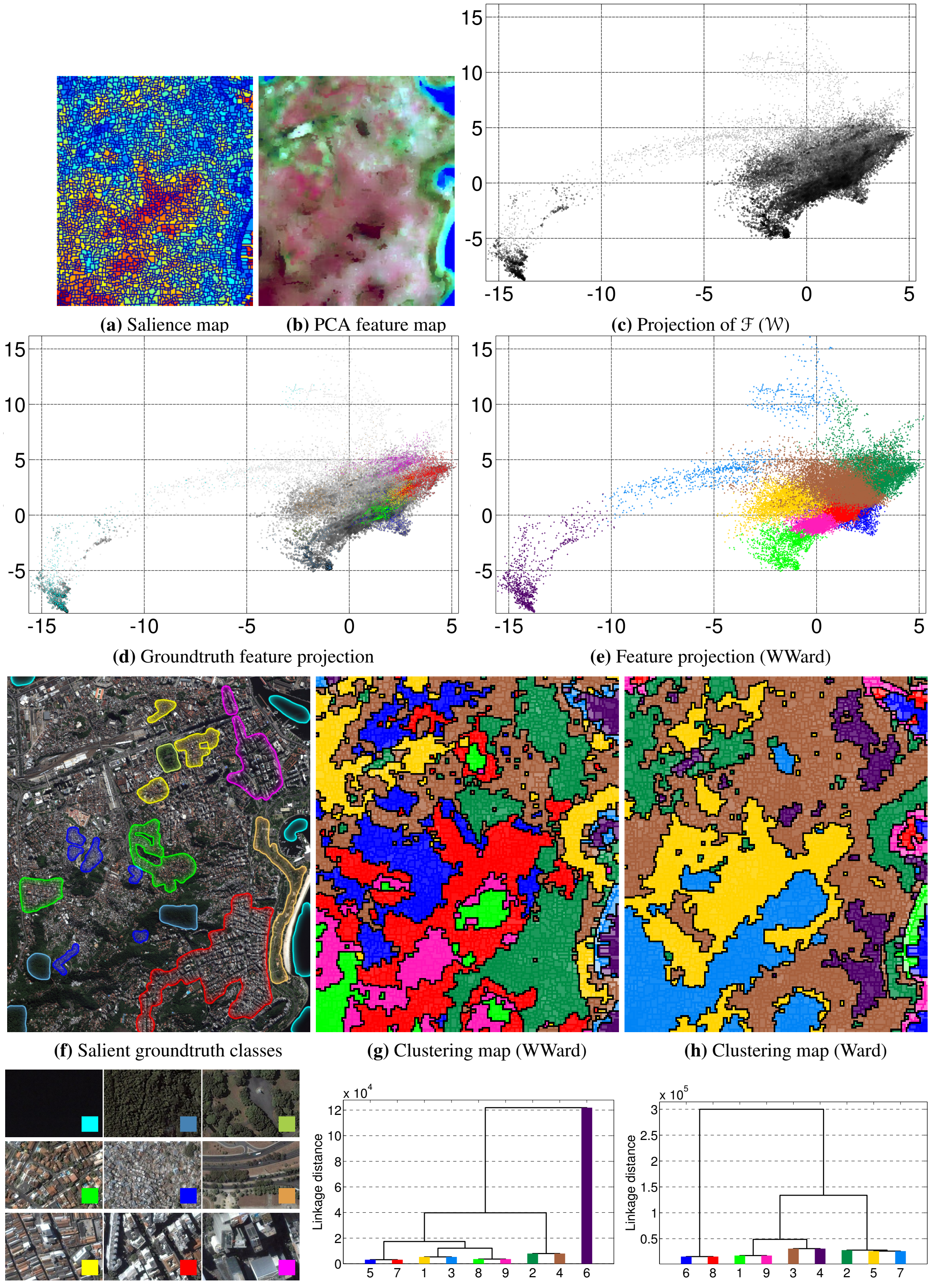
(a) Salience map **(b)** PCA feature map **(c)** Projection of $\mathcal{F}$ $(\mathcal{W})$

(d) Groundtruth feature projection **(e)** Feature projection (WWard)

(f) Salient groundtruth classes **(g)** Clustering map (WWard) **(h)** Clustering map (Ward)

(i) Salient classes with color codes **(j)** Dendrogram (WWard) **(k)** Dendrogram (Ward)

Figure 5.12. Clustering results for Rio de Janeiro D_9 (QB2 - 2005/09/30), with batch weighted Ward for k=9 with the sample weighting $\mathcal{W}$. Multispectral images courtesy of Google™ Earth.



(a) Salience map **(b)** PCA feature map **(c)** Projection of $\mathcal{F}(\mathcal{W})$

(d) Groundtruth feature projection **(e)** Feature projection (WWard)

(f) Salient groundtruth classes **(g)** Clustering map (WWard) **(h)** Clustering map (Ward)

(i) Salient classes with color codes **(j)** Dendrogram (WWard) **(k)** Dendrogram (Ward)

Figure 5.13. Clustering results for Rio de Janeiro D_{10} (IK2 - 2002/04/20), with batch weighted Ward for k=9 with the sample weighting $\mathcal{W}$. Multispectral images courtesy of Google™ Earth.

Table 5.10. Cardinality decisions per acquisition as the interpretation mode of the top 10 indices (fixed for all acquisitions) specifically for weighted Ward linkage. Clustering accuracies are shown at the chosen cardinalities in percentage terms of the objective, Czekanowski-Dice, Jaccard and Rand measures with the maximum possible cardinality decision accuracy.

	Rio de Janeiro acquisition										Mean
	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}	$\mu \pm \sigma$
Cardinality decision	4	5	3	6	5	4	7	7	8	4	
Objective	85.1	85.1	50.0	95.5	83.1	72.0	95.9	93.3	90.6	87.2	83.77±13.81
Objective opt.	95.3	91.7	92.4	96.1	95.3	89.5	95.9	93.3	92.5	97.3	93.92±2.43
Czekanowski-Dice	49.7	41.7	32.6	56.1	54.2	37.7	64.4	53.2	53.0	56.6	49.89±9.71
Czekanowski-Dice opt.	59.7	44.2	49.7	56.9	70.1	45.9	65.0	56.1	58.2	65.7	57.14±8.57
Jaccard	33.0	26.3	19.4	39.0	37.2	23.2	47.4	36.2	36.1	39.4	33.73±8.45
Jaccard opt.	42.5	28.4	33.1	39.8	54.0	29.8	48.1	39.0	41.0	48.9	40.45±8.39
Rand	69.4	67.1	30.9	78.7	79.0	63.3	87.4	79.4	79.8	76.4	71.14±15.83
Rand opt.	85.8	83.9	85.3	85.7	90.4	85.7	88.2	86.0	88.3	88.5	86.78±1.97

The cardinality decision involves obtaining the most frequent cardinality suggestion among the 10 best internal indices for a given clustering algorithm and linkage, obtained by evaluating the indices for all datasets combined. This weak supervision is motivated by providing evidence of good accuracy for most multimodal datasets, which is the case for weighted Ward linkage. The ability of the best internal index selection to perform well under multimodal dataset changes is measured to provide an indication of its generalization ability and thus support its fixed inclusion in the clustering system.

The objective opt. (optimum), Czekanowski-Dice opt., Jaccard opt. and Rand opt. rows in Table 5.10 give the best possible values that can be obtained at the optimal cardinality for the given clustering accuracy measure. The weakly supervised cardinality estimate achieves accuracies that are in general within 10-20% of the optimum.

5.9.5 Experiment 5: Internal index input truncation analysis

The main internal validation indices are analyzed in Table 5.11 under a truncated input scenario to determine whether maximal weight input selection and weighted index generalization improve cardinality decision accuracy. There are three main column sections representing the following internal index calculation sampling comparisons:

1. **Wmax:Urnd** compares weighted maximum salience selection with unweighted random selection.
2. **Wmax:Wrnd** compares weighted maximum salience selection with weighted random selection.
3. **Wmax:Umax** compares weighted maximum salience selection with unweighted maximum salience selection.

A selection is weighted if sample weights are used in a weighted internal index calculation, otherwise the selection is unweighted. A maximum weight/salience selection chooses the N points from the

Table 5.11. Weighted/unweighted and maximum/random selection performance comparisons for internal indices with sample sizes of N_1=1000, N_2=2000, and N_3=4000. Weighted maximum select is compared against unweighted random select (Wmax:Urnd), weighted random select (Wmax:Wrnd), and unweighted maximum select (Wmax:Umax) in terms of the ratio of experiments (shown as a percentage) in which weighted maximum select significantly outperforms each alternative.

	Weighted Ward								
	Wmax:Urnd			Wmax:Wrnd			Wmax:Umax		
Internal index	N_1	N_2	N_3	N_1	N_2	N_3	N_1	N_2	N_3
Point-Biserial	100	75	67	100	67	67	90	90	89
Baker-Hubert Gamma	100	100	100	90	90	100	100	100	100
Tau	100	100	100	100	100	100	100	100	100
G+	100	100	100	100	100	100	100	100	100
PBM	57	100	140	44	75	160	100	100	129
Davies-Bouldin	100	100	100	114	100	114	100	100	90
Det_Ratio	10	33	33	150	250	167	114	100	100
Ratkowsky-Lance	129	129	143	78	67	89	125	100	100
Wemmert-Gançarski	111	111	111	125	111	111	100	100	125
Silhouette	129	114	150	114	89	113	100	100	100
Trace_W	100	100	100	117	200	200	100	80	100
S_Dbw	89	133	140	100	225	114	100	111	100
Ball-Hall	114	167	200	100	129	167	100	90	111
Calinsky-Harabasz	100	175	175	129	114	150	100	143	100
C index	225	140	120	129	180	129	100	113	90
McClain-Rao	160	180	180	133	143	113	80	129	125
Banfield-Raftery	71	100	180	180	167	333	60	100	129
Ksq_DetW	88	67	86	160	225	160	200	114	250
GDI23	250	180	160	200	167	129	100	100	100
GDI33	250	150	200	167	180	150	100	100	90
Ray-Turi	63	75	117	450	250	160	129	129	129
GDI25	225	500	300	200	333	180	111	143	100
Log_SS_Ratio	175	100	120	225	300	1000	113	100	125
Log_Det_Ratio	180	129	100	1000	1000	333	111	111	129
GDI35	450	500	1000	200	250	333	125	143	100
Scott-Symons	1000	333	500	500	500	250	200	140	71
Trace_WiB	200	1000	1000	1000	1000	1000	100	70	90
Mean	169	178	209	213	226	215	109	107	106

dataset such that the sum of the sample weights is larger than or equivalent to any different selection of the same size. A selection is random when sample weights are not incorporated into the selection process and only random choices are made for constructing an N-sized sample. Each sampling comparison shown in a column section is divided into three subcolumns, which correspond to three different sample sizes, namely N_1=1000, N_2=2000, and N_3=4000. The values shown for each internal index for a given comparison and sample size is the percentage ratio of datasets for which weighted maximum salience selection outperforms the alternative selection with statistical significance.

The 10 internal index cardinality accuracy ($\mathcal{E}$) measures for the 10 acquisitions with a given sampling method A forms a vector $\{a_1, \cdots, a_{10}\}$ of accuracies, which are normally scalar values for deterministic samplings such as maximum weight sampling. However, accuracy distributions summarized by mean and standard deviations are used for random sampling. The accuracy for an alternative sampling scheme B is described similarly by vector $\{b_1, \cdots, b_{10}\}$, and accuracy comparisons are made for $(a_i, b_i), 1 \leq i \leq 10$, where a t-test can be used if standard deviations are available for multiple accuracy assessments in the case of non-deterministic sampling.

There are three possible outcomes to comparison (a_i, b_i) for a given dataset i:

c_1 : Sampling A assists in producing cardinalities with a significantly greater clustering accuracy than sampling B.

c_2 : There is no significant difference between the clustering accuracies produced.

c_3 : Sampling B assists in producing cardinalities with a significantly greater clustering accuracy than sampling A.

Consideration of all 10 datasets can provide the counts ($\sum c_j$) for the three conditions to construct the comparison ratio $(\sum c_1 + \sum c_2) : (\sum c_3 + \sum c_2)$ or $\frac{\sum c_1 + \sum c_2}{\sum c_3 + \sum c_2}$, which is expressed as a percentage in Table 5.11.

Maximum weight input selection clearly appears to produce better cardinalities than random input sampling, whereas the use of weights with weighted internal indices has a minimal effect overall. In the comparison with unweighted random sampling an increase in sample size on average produces a greater discrepancy in weighted maximum input selection. However, for the comparisons with weighted random sampling and unweighted maximum input selection there is on average no distinct comparative change with increases in sample size.

Random sampling consistently performs worse for the Wemmert-Gançarski, C index, McClain-Rao, GDI, Scott-Symons and Trace_WiB indices. Using weighted internal indices with maximum input selection manages to consistently outperform unweighted internal indices also with maximum input selection for the Ksq_DetW, Ray-Turi and Log_Det_Ratio indices. A more in-depth investigation is needed to determine why certain internal indices perform worse with random sampling or unweighted internal index calculations.

5.9.6 Experiment 6: Internal index interpretation comparison

Different internal index interpretation methods are compared in Table 5.12 for the major internal indices, and the best choice between extremum (denoted by "min" and "max") and disruption (denoted by "mind" and "maxd") intepretations is given in the final column. The comparison is performed in terms of the average percentage of the maximum joint clustering quality of the optimal cardinality for the given weighted Ward linkage. A greater average percentage means that the cardinality determination was more accurate.

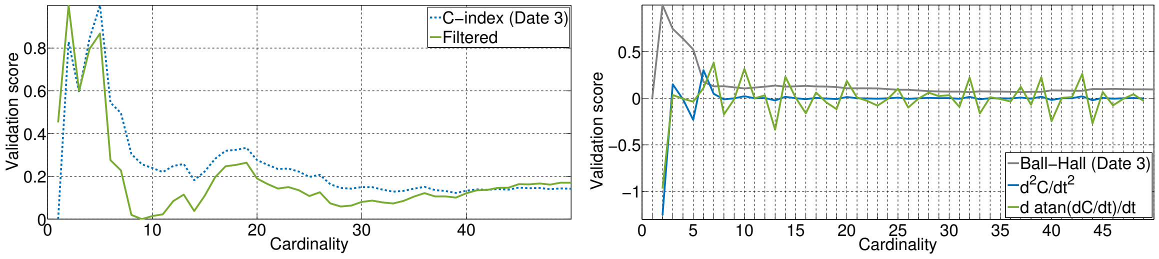
Table 5.12. Internal index performance comparison as average percentage of the maximum joint clustering quality of the optimum cardinality. Extremum (Ext.) and disruption ($\frac{d^2}{dk^2}$) interpretations are compared, including extremum filtering (Filt.) and the suppressed disruption derivative alternative (atan).

	Extremum		**Disruption**		
Index	Ext.	Filt.	$\frac{d^2}{dt^2}$	atan	Interpretation
Det_Ratio	0.00	0.00	7.05	10.73	mind
Log_Det_Ratio	0.79	8.32	69.40	56.01	mind
Banfield-Raftery	0.00	3.05	79.88	72.39	maxd
Log_SS_Ratio	0.00	5.09	79.68	75.60	mind
Trace_WiB	35.16	32.17	64.22	55.32	max
McClain-Rao	36.08	32.05	57.00	77.61	min
Ball-Hall	8.74	50.40	78.11	78.64	min
Point-Biserial	31.89	47.90	59.92	82.49	mind
Ray-Turi	54.65	53.58	57.30	64.06	min
Trace_W	0.00	74.25	77.28	85.98	min
S_Dbw	10.33	69.48	80.66	83.84	min
C index	45.00	49.99	76.26	80.20	min
G+	22.91	60.15	85.35	84.65	min
Silhouette	54.65	40.90	83.64	83.64	max
Davies-Bouldin	54.65	46.82	82.23	80.04	min
Scott-Symons	54.69	54.69	84.29	71.72	maxd
Ksq_DetW	45.35	71.42	82.61	72.64	maxd
GDI25	54.58	54.69	78.72	85.52	max
GDI35	54.58	54.69	78.96	85.52	max
GDI33	54.69	54.69	83.55	84.43	max
GDI23	54.69	54.69	83.74	85.50	max
PBM	68.51	68.56	75.30	74.53	mind
Ratkowsky-Lance	67.88	64.65	78.42	76.86	mind
Wemmert-Gançarski	73.13	72.64	72.64	70.25	max
Baker-Hubert Gamma	75.15	76.15	81.99	80.31	max
Calinsky-Harabasz	74.52	76.23	84.17	79.87	max
Tau	76.96	80.92	80.80	81.37	max
Mean	38.77	49.58	73.33	73.35	max

Knee-point accentuating filtering before extremum interpretation improves accuracy for 16 indices, where the more significant improvements are seen for Ball-Hall, Trace_W, S_Dbw and G+. The disruption interpreted indices do not really benefit from the knee-point accentuating filtering, since this filtering is designed to improve extremum interpretation. The alternative $\frac{d}{dk}\arctan\left(\frac{d\mathbb{C}}{dk}\right)$ disruption interpretation improves accuracy for 13 indices in the case of weighted Ward linkage clustering.

It appears from Table 5.12 that disruption interpretation of internal indices generally perform better than extremum interpretation, possibly because of the observed tendency of extremums located at higher cardinalities. Knee-point accentuating filtering clarifies the extremums for more accurate extremum interpretation, but disruption interpretation directly isolates the knee-points without the extremum requirement having to be fulfilled.

Examples of knee-point accentuating filtering and the alternative $\frac{d}{dk}\arctan\left(\frac{d\overline{C}}{dk}\right)$ disruption interpretation are shown in Figures 5.14(a) and 5.14(b).



(a) Knee-point accentuating filtering

(b) $\frac{d}{dk}\arctan\left(\frac{d\overline{C}}{dk}\right)$ disruption interpretation

Figure 5.14. Examples of internal index interpretations contributed in this chapter.

At every cardinality the extremum filter compares the signal mean for lower cardinalities with the mean for larger cardinalities and identifies major mean changes, which are then added to the original signal to press the knee-points out, thereby reducing the likelihood of large cardinality decisions in an otherwise generally monotonic signal.

5.10 CONCLUSION

The main objective of the study was to investigate whether clustering accuracy can be improved by imparting unsupervised context through sample salience that reveals the preferential utility embedded in the groundtruth reference. The preference for greater textural regularity present in the groundtruth is modeled as multiscale texture feature variance converted into a salience weighting for the feature space. The hypothesis that weighted agglomerative clustering differentiates between clusters through the tendency to move cluster centroids towards higher weight regions was tested indirectly by assuming that the salience weighting modeled the groundtruth preference for textural regularity reasonably well.

Weighting the feature space can create an artificial separation in a space where classes tend to blend owing to the presence of samples sharing traits from multiple classes simultaneously, such as a region geospatially bordering different classes. For the 10-date Rio de Janeiro multispectral dataset it was shown that weighted clustering with the Ward linkage achieves a greater mean clustering accuracy. Confusion analysis presented evidence that weighted clustering produces more salient clusters and differentiates better between certain groups of classes.

Weighted internal validation indices were used for weakly supervised clustering cardinality determination, and input truncated implementations were used to reduce computational time for large datasets. Sample weighting was used to good effect and it was experimentally illustrated that the main contributor to improved internal index performance was maximal weight input selection.

Improved internal index extremum and disruption interpretations were proposed and results indicated performance improvements for the majority of internal indices.

The following chapter investigates the manifold matching component of the manifold alignment framework, as a next step to the manifold reduction studied in this chapter.

CHAPTER 6 GLCM MANIFOLD MATCHING WITH GEOMETRIC SIMILARITY MEASURES

6.1 CHAPTER OVERVIEW

Across-satellite texture feature matching is considered in this chapter, with the purpose of optimizing domain matching with the use of geometric similarity. The preservation of GLCM feature space geometry across imagery from different high-resolution optical satellites is investigated by determining how the addition of geometric similarity to a graph matching cost function influences matching accuracy. The manifold matching of this chapter fits into the manifold alignment strategy for addressing dataset shift, where the previous chapter dealt with manifold reduction aspect required to render manifold matching computationally feasible.

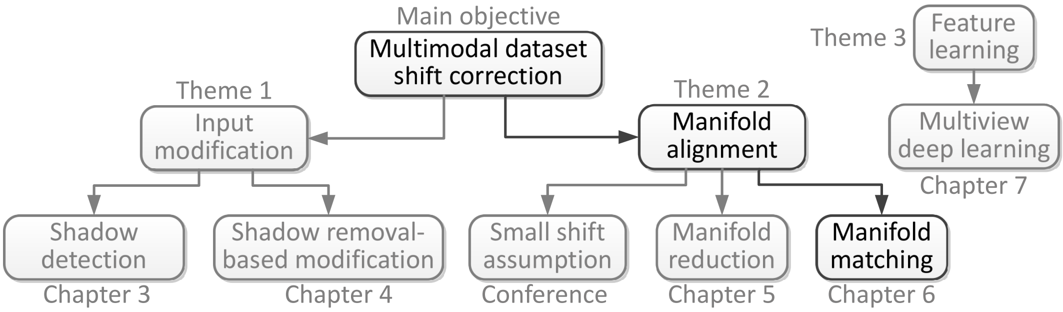


Figure 6.1. Indication of where this chapter fits into the thesis.

A generalized eigenvalue decomposition framework for manifold alignment [42] is reviewed in Appendix B and its requirement of correspondence knowledge or across-domain matching information is addressed in this chapter. The manifold matching problem is isolated to test directly whether a geometry-based matching performs well for GLCM feature spaces, so the Hungarian algorithm is used as solver and execution of manifold alignment is omitted. The focus is placed on engineering a matching cost function that simultaneously results in good matches for multiple land-use based matching problems.

A geometric similarity formulation is contributed that combines varying neighborhood sizes to eliminate the need to find an optimal neighborhood size, as is required by the geometric similarity from [43].

A local geometry matching co-occurrence matrix is introduced to access novel geometry pattern information, and it is shown how it can be best used to add novel information not included in basic geometric similarity. To facilitate the direct measurement of matching accuracy in this study, there is a strong assumption of perfect correspondence between across-domain areas through co-registration for all satellite images, so that all domain components can be matched for both domains and a perfect match can be found.

6.1.1 Contributions

1. Minimum-supervised manifold matching for relatively large dataset shifts and perfect correspondence is contributed.

2. A geometric similarity formulation is provided that eliminates the need for an optimal neighborhood size, yet maintains similar accuracy.

3. Local geometry matching co-occurrence is introduced as a new perspective on geometric similarity, and it is used in a manner that complements basic geometric similarity by contributing novel information.

4. Various cost function instantiations with geometric similarity components are tested on numerous domain-matching problems, and it is shown that the proposed modifications and additions to basic geometric similarity improve matching accuracy.

5. Supervised parameter learning for these cost functions is employed, but good generalization is shown for problems with the same domain cardinality.

The problem statement is given next in Section 6.2, followed by Section 6.3 on related work. The across-satellite imagery and dataset organization are described in Section 6.4 and then the different similarities, including geometric similarity and its variants, together with the combination thereof into the main cost function, are explained in Section 6.5. Key experiments are described and the main results of this chapter are given in Section 6.6 and the important observations and confirmations of hypotheses are discussed, before conclusions are drawn in Section 6.7.

6.2 PROBLEM STATEMENT

Feature matching in image registration is an important and challenging problem, especially in the case of severe dataset shifts [229]. A dataset shift appears when the joint distribution of the input and class variables differs between source and target datasets. Such dataset shifts may be caused by a non-stationary classification environment or domain shift, which is characterized by a change in the measurement system, or method of description. This is prevalent in remote sensing, where image mismatches may appear in response to seasonal and illumination changes, terrain distribution or sensor differences between the datasets.

A supervised across-satellite and across-date feature matching cost function, based on divergence, standard deviation similarity and geometric similarity [43], is developed in this chapter for GLCM features. The GLCM feature descriptors of areas with uniform land-use type are calculated from different high-resolution optical images acquired at different times and/or with different satellites. The objective is to optimize a graph matching cost function that can accurately find the correspondence between texture feature descriptors of two sets of areas with relatively similar land-use types and the same land-use prior probabilities.

The limitation to this very specific scenario, exemplified by a coregistration problem, is motivated partly by the difficulty of solving a non-perfect across-date cluster matching problem under dataset shift where some domain elements are unmatched or have multiple matchings. The problem scope is thus limited to perfect matching scenarios so that a feasible solution can be demonstrated.

6.2.1 Hypotheses

1. Local texture feature geometry is preserved across a multimodal dataset shift, since the feature relationships between classes are maintained and because good features separate classes based on relative dissimilarity.

2. Across-domain classes that are more frequently matched together in optimal local neighborhood matchings are more likely to be matched in the across-domain matching, because such across-domain class pairs demonstrate a higher local geometry similarity.

3. Global translation and a basic divergence minimization objective can improve matching accuracy, since it corrects global domain differences and attempts to find the dataset shift with the fewest assumptions as stated by Occam's razor.

4. Relative class variances are possibly maintained under a dataset shift, since certain classes will usually have more variance, such as informal settlements, and other classes will have less variance, such as the non-builtup class.

6.2.2 Research questions

1. How does one perform unsupervised manifold (perfect) matching for relatively larger dataset shifts?

2. How can information derived during the optimal neighborhood permutation search be used to improve geometric similarity matching accuracy?

3. How should geometric similarity be employed, and which other correspondence measures should be applied to perform manifold matching accurately?

6.3 RELATED WORK

Graph matching has recently been used in remote sensing to match multitemporal sequences of hyperspectral images by modeling the graphs for both domains as two observations of one common underlying hidden Markov random field, using similarity based on multivariate Gaussian probability density functions [230]. Pixel-level correspondences between across-date images are found through graph matching with simplified nearest neighbor graphs in [231].

The correspondence obtained between domains can be used for feature space manifold alignment to transfer labels between domains, for example, or as part of an image registration process. Feature point matching is used for image registration in [229], but spatial relationships are employed directly. Local spatial relationship and pattern similarity are exploited in [232] for feature point matching for the purpose of image registration.

The focus of this research is, however, on investigating the preservation of GLCM feature space geometry across imagery from different satellites by determining how a geometric similarity measure addition to the cost function influences domain-matching accuracy. This knowledge of the dataset shift behavior of texture features can be employed to improve feature point matching.

The manifold alignment framework of Tuia et al. [42] requires direct correspondence between across-domain clusters or points, which is basically classification of a test domain from a train domain. So the classification problem and the dataset shift problem essentially needs to be solved to obtain direct correspondence information. Manifold alignment then allows for previously unseen points to be mapped to a joint manifold to be classified, so it smooths the discrete probability function to a continuous function.

Note that the geometric similarity concept introduced by Wang et al. [43] is for a manifold alignment framework without correspondence, so no direct correspondence information is required, as neighborhood geometry relationships are used to establish across-domain relationships as part of the generalized eigenvalue decomposition. In this study the concept of geometric similarity is isolated from the manifold alignment without correspondence, since the extended study dataset lends itself to a manifold alignment with correspondence. So the task is to use the concept of geometric similarity and to evaluate its preservation for a texture feature space in a land-use classification remote sensing scenario.

The relationship of manifold alignment to the classification system is shown in Figure 6.2.

6.4 DATA DESCRIPTION

Twelve polygons, each of a uniform land-use type, were selected from the subtropical highland of Johannesburg (Gauteng, South Africa) (see Figure 6.3). Most land-use types are settlements, including informal, formal with backyard shacks, and formal housing. Non-builtup areas, open fields, a cemetery and golf course are included to define key diverse characteristics of the texture feature spaces. Each

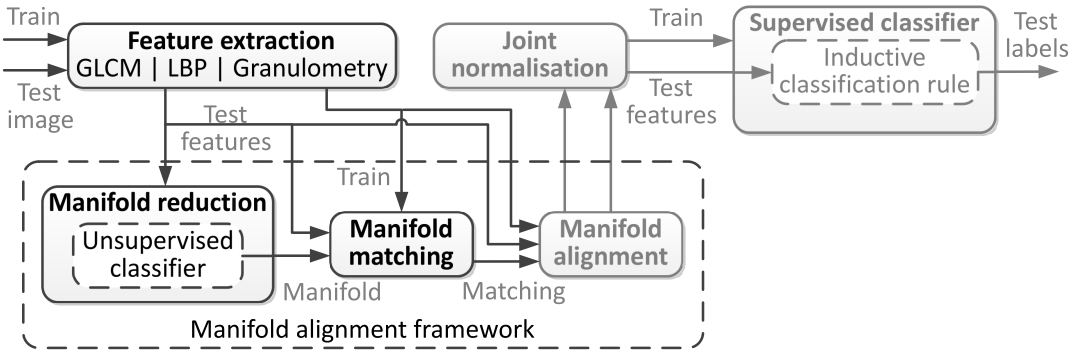


Figure 6.2. Supervised classification system with manifold alignment framework.

polygon is co-registered over six different across-date pansharpened satellite acquisitions, of which samples from the 6×12 polygons are shown in Table 6.1.

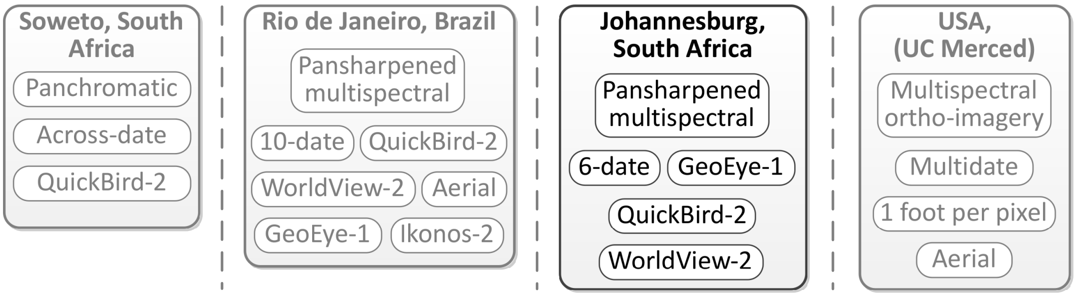


Figure 6.3. Johannesburg multimodal dataset selection for this experiment.

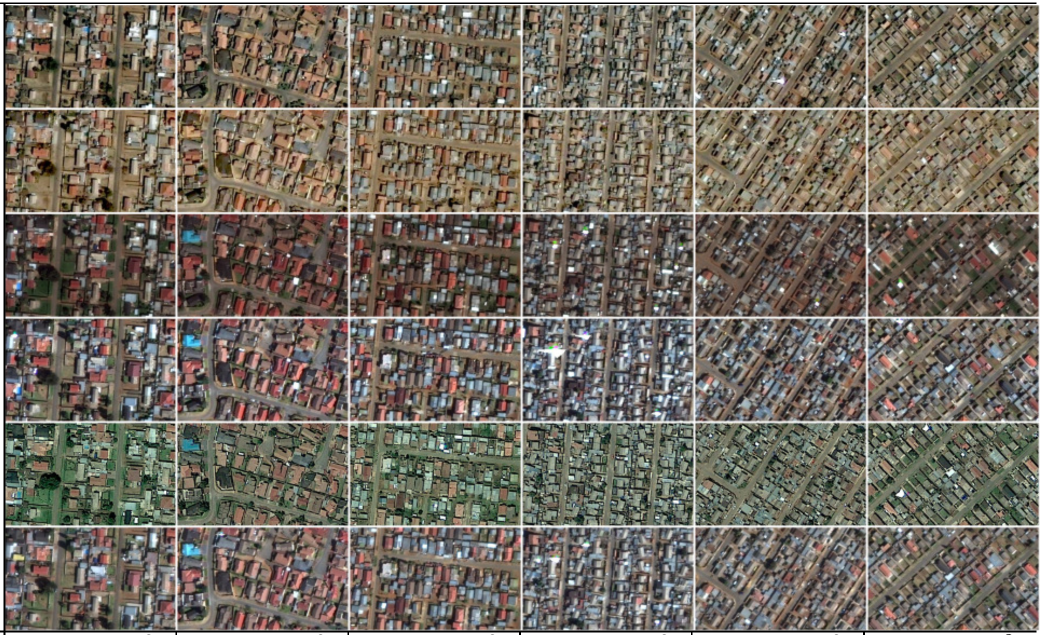
The QuickBird-2, WorldView-2 and GeoEye-1 pansharpened color images had initial respective resolutions of 0.6, 0.46 and 0.41 m/pixel-edge, which were then reduced to a common 0.6×0.6 m/pixel and converted to grayscale. Image tiles of dimensions 100.2×100.2 m (100.2 m = 167 pixels × 0.6 m/pixel) were then extracted from each polygon, with a mean coverage of 99.1% and an average coverage redundancy of 2.27 tiles/pixel. The first 13 of Haralick's GLCM features were determined for each tile (feature 14 omitted because of its computational complexity), with a 167×167 pixel window and GLCM pairs used in all cardinal and ordinal directions with respective ℓ_1-norms of one and two, and the features were averaged over the four spatial relationships.

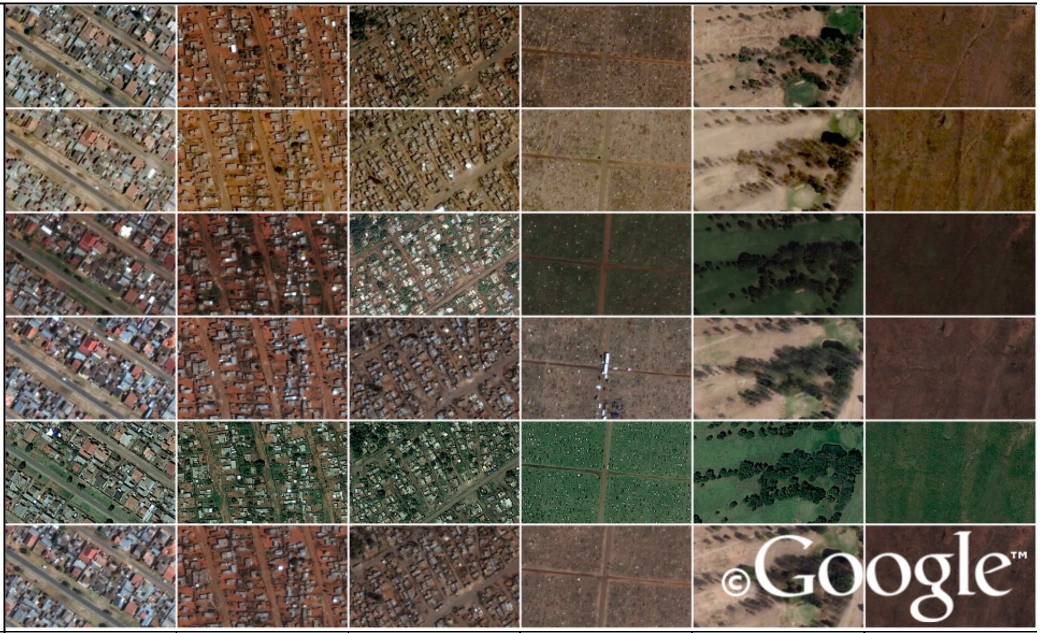
6.5 METHODOLOGY

6.5.1 Study objective

Domain $U = \{u_j\}_{j=1}^{|U|}$ is a $p \times |U|$ matrix (set cardinality denoted by $|\cdot|$) with co-domain $V = \{v_j\}_{j=1}^{|V|}$, where u_j and v_j are defined in a p-dimensional feature space ($p = 13$ for GLCM). A partitioning or labeling function $c_U(\cdot)$ partitions $U = \{U_i\}_{i=1}^{m}$ into $\eta(U) = m$ subsets $U_i \subseteq U$ so that $\{\forall i \in [1,m],\, \forall a,b \in U_i,\, \forall e \in U \setminus U_i \mid c_U(a) = c_U(b),\, c_U(a) \neq c_U(e)\}$, and $c_V(\cdot)$ partitions $V = \{V_i\}_{i=1}^{n}$ similarly.

Table 6.1. Multi-satellite imagery (S) of six acquisitions from QuickBird-2 (QB), WorldView-2 (WV) and GeoEye-1 (GE) satellites, coregistered over 12 different areas, each of a uniform land-use type. Imagery samples courtesy of Google™ Earth.

Satellite & Date	Eldorado Park East	Diepkloof East	Devland South	Diepkloof Zone 2	Chiawelo East	Molapo
D_1 (QB) 2007-9-18						
D_2 (QB) 2008-9-7						
D_3 (WV) 2011-3-31						
D_4 (WV) 2011-7-2						
D_5 (GE) 2012-1-1						
D_6 (WV) 2012-6-6						
Area	0.2863 km^2	0.7417 km^2	0.2179 km^2	0.2457 km^2	0.2819 km^2	0.5999 km^2
# Tiles	61×6	76×6	39×6	105×6	82×6	107×6

Satellite & Date	Orlando East	Olifantsvlei	Protea South	Avalon Cemetery	Soweto Golf Club	Klipriviers-oog South
D_1 (QB) 2007-9-18						
D_2 (QB) 2008-9-7						
D_3 (WV) 2011-3-31						
D_4 (WV) 2011-7-2						
D_5 (GE) 2012-1-1						
D_6 (WV) 2012-6-6						
Area	0.7902 km^2	0.4192 km^2	0.4061 km^2	0.6699 km^2	0.3788 km^2	0.5606 km^2
# Tiles	220×6	68×6	69×6	148×6	80×6	78×6

$\Gamma(C) = \Gamma(C(U,V)) = \arg_{f(C)} \min \sum_i C_{if(C)_i}$ is a bijection ($f(C)$) that one-to-one matches partitions in $\{U_i\}$ to those in $\{V_{\Gamma(C)_i}\}$ with all partitions matched and each partition matched to exactly one across-domain partition, which can be solved by the $O(n^3)$ Munkres or Hungarian algorithm [233]. The $\eta(U) \times \eta(V)$ cost matrix C (with $\eta(U) = \eta(V)$), with matrix indicing C_{ij} indicating row i and column j, captures the cost of matching across-domain subsets U_i and V_j. For perfect correspondence each subset U_i is uniquely matched to one subset V_j and vice versa ($m = n$), with the best possible match given by $1 \times \eta(U)$ vector Γ_{UV}.

For multiple source and target domain pairs $S = \{(D_{s(i)}, D_{t(i)})\}_{i=1}^{|S|}$, where source domains are denoted by $D_{s(i)}$ (a $p \times |D_{s(i)}|$ matrix) and target domains by $D_{t(i)}$ (a $p \times |D_{t(i)}|$ matrix), the study objective is to devise an overall cost matrix (function) $C^* = \arg_C \min \Theta_S(C)$ that minimizes the overall matching error. Equation 6.1 defines the normalized root-mean-square error $\Theta_S(C) \in [0,1]$ between the best match Γ_{UV} and $\Gamma(C)$ over all separately matched domain pairs, where $\delta(i,j)$ denotes the Kronecker delta function, which is 1 if i and j are equal and 0 otherwise, overall given by

$$\Theta_S(C) = \sqrt{\sum_{(U,V)\in S} \frac{\left(\sum_{i=1}^{\eta(U)} \delta(\Gamma_{UV_i}, \Gamma(C(U,V))_i)\right)^2}{\eta(U)^2 |S|}}. \tag{6.1}$$

Matching accuracy is noted in this chapter as $100(1 - \Theta_S(C))$ and is measured for S consisting of all possible pairs of acquisitions in the study dataset.

6.5.2 Data preprocessing

The data sets are normalized simultaneously to Student's t-statistic $T(U,V)_i = (U_i - \mu(U \cup V)) \cdot \sigma(U \cup V)^{-1}$, where $\mu(\cdot)$ and $\sigma(\cdot)$ are respectively the row-wise means and standard deviations. Standardized partition means $X = X(U,V) = \{x_i = \mu(T(U,V)_i)\}_{i=1}^m$ ($p \times m$ matrix) and $Y = Y(V,U) = \{y_i = \mu(T(V,U)_i)\}_{i=1}^n$, as well as partition standard deviations $X_\sigma = X_\sigma(U,V) = \{\sigma(T(U,V)_i)\}_{i=1}^m$ and $Y_\sigma = Y_\sigma(V,U) = \{\sigma(T(V,U)_i)\}_{i=1}^n$ describe the simplified domain and co-domain manifolds.

6.5.3 Divergence and standard deviation similarity

An assumption in this chapter is that there are GLCM features that undergo an overall constant shift or translation because of dataset shift resulting in a divergence unrelated to other types of geometry-preserving shifts. Divergence between domains resulting from an overall sample mean mismatch between domains can be reduced by using cost matrix C_Δ with specific matrix entries $C_\Delta(\delta,U,V)_{ij}$ given in Equation 6.2, representing the Euclidean norm ($\|\cdot\|_2$) between subsets U_i and V_j as

$$C_\Delta(\delta,U,V)_{ij} = \left\| X(U,V)_i - Y(V,U)_j + \delta\Big(\mu\big(X(U,V)\big) - \mu\big(Y(V,U)\big)\Big) \right\|_2. \tag{6.2}$$

Coincidence of domain and co-domain sample means is obtained through mean translation [48] with correction factor $\delta = 1$, as shown in Figure 6.4. Here the matching accuracy improves as $\delta \to 1$, which is indicative of possible evidence for the assumption of feature translation under dataset shift. An example of an across-satellite feature space transformation and domain matching based on corrected divergence is seen in Figure 6.5.

Another assumption is relative variance similarity between matched across-domain subset or cluster pairs, since the associated across-domain classes are expected to have a similar degree of feature variance relative to other classes. A smaller normalized difference between across-date cluster standard deviations indicates a higher likelihood of a match under this assumption. Normalized differences between subset standard deviations are used for component C_σ with scalar entries $C_\sigma(U,V)_{ij}$ as shown in Equation 6.3 given by

$$C_\sigma(U,V)_{ij} = \left\| \frac{X_\sigma(U,V)_i - Y_\sigma(V,U)_j}{X_\sigma(U,V)_i + Y_\sigma(V,U)_j} \right\|_2 . \tag{6.3}$$

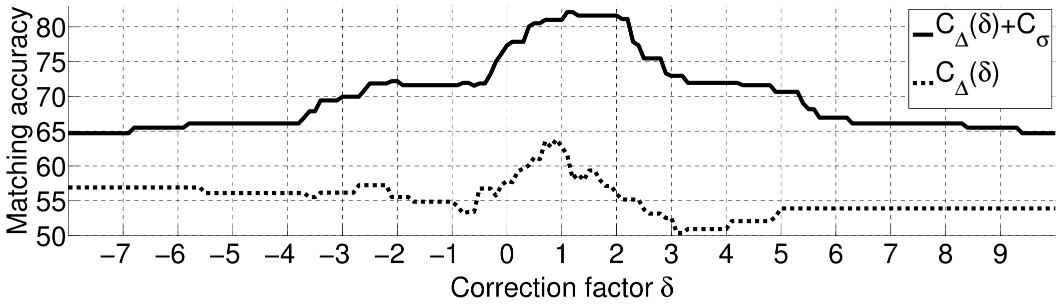


Figure 6.4. The influence of mean translation (Eq. 6.2) on domain-matching accuracy $(100(1-\Theta_S(C)))$ for divergence only $(C_\Delta(\delta))$ and divergence plus standard deviation similarity $(C_\Delta(\delta)+C_\sigma)$ cost matrices. The relative scaling between $C_\Delta(\delta)$ and C_σ is equal for this illustration, but is optimized in the remainder of this chapter.

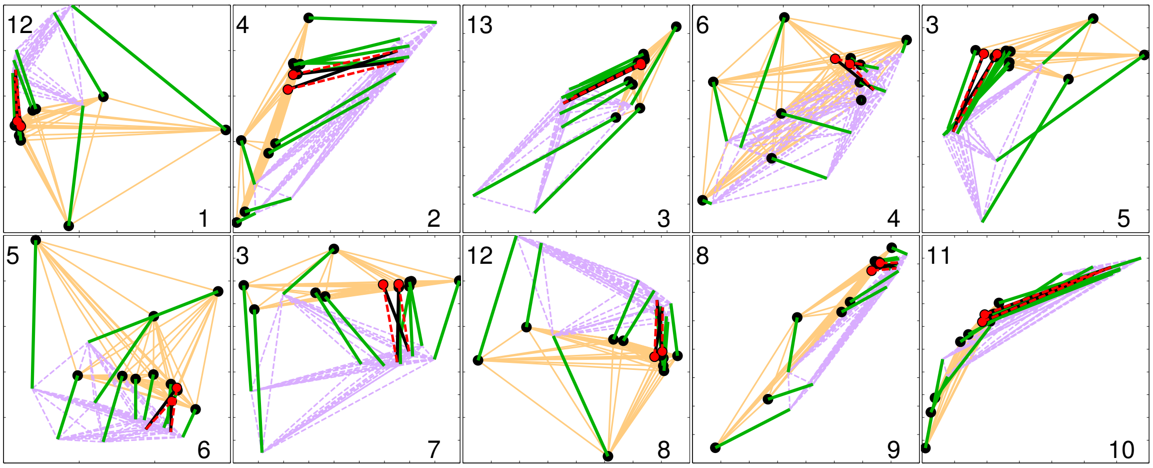


Figure 6.5. Domain-matching snapshot ($D_4 \leftrightarrow D_5$) with divergence ($C_\Delta(1)$). Each axis indicates the GLCM feature number. Correct point matches (—●) from D_4 (- - -) to D_5 (—) are indicated, also incorrect matches (- -●) with corrections (—●).

6.5.4 Geometric similarity calculation

The determination of geometric similarity outlined in [43] is reviewed in this subsection, before it is used in domain matching cost functions. The local geometry of a point x_i is represented by a $(k+1)\times(k+1)$ matrix R_{x_i}, which contains the pairwise Euclidean distances between the nearest neighbours of x_i. Similarly, R_{y_j} is a $(k+1)\times(k+1)$ matrix representing the local geometry of y_i. The distance between x_{z_a} and x_{z_b} is recorded as $R_{x_i}(a,b) = \| x_{z_a} - x_{z_b} \|_2$, where $z_1 = i$ and $\{z_2,\ldots,z_{k+1}\}$ are x_i's k nearest neighbor's indices. The k nearest neighbours of x_i are denoted as $N_k(x_i) = \{z_l\}_{l=1}^{k+1}$, and a permutation $\{N_k(x_i)\}_h$ fixes z_1 and permutes $\{z_2,\ldots,z_{k+1}\}$ according to permutation number $0 \geq h \geq k!$. $\{R_{x_i}\}_h$ is the associated matrix corresponding to the permutation $\{N_k(x_i)\}_h$.

A permutation-specific geometric similarity of the local contact patterns (size k) around points x and y is given by $d_k(x,y,h)$, for a specific y neighborhood permutation $\{N_k(y)\}_h$. Since x and y are from different domains, the neighborhood scales may differ, so either R_x or $\{R_y\}_h$ is scaled as in Equation 6.4 before calculating the difference between geometries, depending on which scaling minimizes the Frobenius norms of the difference matrices, scaling given by

$$d_k(x,y,h) = \min\left(\left\| \{R_y\}_h - \frac{\operatorname{tr}\left(R_x^T\{R_y\}_h\right)R_x}{\operatorname{tr}\left(R_x^T R_x\right)} \right\|_F , \left\| R_x - \frac{\operatorname{tr}\left(\{R_y\}_h^T R_x\right)\{R_y\}_h}{\operatorname{tr}\left(\{R_y\}_h^T\{R_y\}_h\right)} \right\|_F \right). \tag{6.4}$$

The k nearest neighbor geometric similarity between x and y is $d_k(x,y) = \min_{1\geq h\geq k!} d_k(x,y,h)$, for the local matching $\Gamma_{xy}^{(k)} = \arg_{\{N_k(y)\}_h} \min_{1\geq h\geq k!} d_k(x,y,h)$ between the neighbours of x and y that minimizes the geometric dissimilarity. The mean geometric similarity $\overline{d_{XY}}^{(k)}$ between simplified domains X and Y is shown in

$$\overline{d_{XY}}^{(k)} = \sum_{i=1}^{|X|}\sum_{j=1}^{|Y|} d_k(x_i,y_j) \Big/ (mn) \tag{6.5}$$

and the geometric similarity matrix $d_{XY}^{(k)}$ in

$$d_{XY}^{(k)} = \sum_{i=1}^{|X|}\sum_{j=1}^{|Y|} d_k(x_i,y_j) J(i,j) \Big/ \overline{d_{XY}}^{(k)} \tag{6.6}$$

gives the normalized geometric similarities of each possible across-domain point pair for neighborhood size k. $J(a,b)$ is a single-entry matrix where only $J_{ab} = 1$, with the rest of the elements being zero.

6.5.5 Basic geometric similarity cost function

The cost matrix C_1 in Equation 6.7 is based upon geometric similarity with growing neighborhood sizes, with a maximum size $K < \min(|X|,|Y|)$, where larger neighbourhoods contribute more information to the degree controlled by parameter λ, overall formulated as

$$C_1(\lambda,U,V) = \sum_{k=2}^{K} d_{XY}^{(k)} \exp\left(k/\lambda\right) \Big/ \sum_{k=2}^{K} \exp\left(k/\lambda\right) \tag{6.7}$$

The larger weight of larger neighborhoods is motivated by the fact that these include more information than smaller neighborhoods. The matrix is normalized to allow for integration into more complex cost matrices. The cost matrix C_1 constitutes a baseline formulation for domain matching with geometric similarity.

6.5.6 Geometric similarity and matching co-occurrence cost functions

The local geometry matching co-occurrence matrix $\Omega_A^{(k)}(X,Y)$ in Equation 6.8 records how often each possible across-domain pair is matched in optimal geometric similarity permutations for a neighborhood size of k, overall formulated as

$$\Omega_A^{(k)}(X,Y) = \sum_{x\in X}\sum_{y\in Y}\sum_{l=1}^{k+1} \frac{J\left(N_k(x)_l, \Gamma_{xy}^{(k)}(l)\right)\exp(1)}{(k+1)\exp\left(\frac{d_k(x,y)}{\overline{d_{XY}}^{(k)}}\right)}. \tag{6.8}$$

Each co-occurrence increment is conditioned by the normalized geometric similarity $d_k(x,y)/\overline{d_{XY}}^{(k)}$ of the base pair (x,y), so that the co-occurrence of matched neighbours $(x_{N_k(x)_l}, y_{\Gamma_{xy}^{(k)}(l)})$ is promoted when the geometries of the neighbourhoods $N_k(x)$ and $N_k(y)$ are more similar. $J(a,b)$ is a single-entry matrix where only $J_{ab} = 1$, with the rest of the elements being zero.

This co-occurrence matrix is incorporated into the denominator of cost matrix C_2 in Equation 6.9, since higher co-occurrence counts should translate to lower matching cost, overall given by

$$C_2(\lambda,\gamma,U,V) = \left(\sum_{k=2}^{K}\exp\left(\frac{k}{\lambda}-\frac{10}{\gamma}\right)\right)\oslash\left(\sum_{k=2}^{K}\exp\left(\frac{k}{\lambda}-10d_{XY}^{(k)}/\gamma\right)\odot\Omega_A^{(k)}(X,Y)\right). \tag{6.9}$$

The Hadamard (element-wise) product and division are respectively denoted by $\odot$ and $\oslash$. The geometric similarity matrix is moved to a heatmap formulation $\exp(-10d_{XY}^{(k)}/\gamma)$ controlled by factor γ, as this construct proved to perform best in conjunction with the co-occurrence matrix for the study data. γ is scaled by a factor of $1/10$ to get the parameter into a similar functional range as λ for aesthetic purposes; this scaling factor was determined based on the results in Figure 6.6.

An alternative local geometry matching co-occurrence matrix $\Omega_B^{(k)}$ is presented in Equation 6.10, where co-occurrences are suppressed for base pairs with greater geometric similarity, instead of promoting such co-occurrence increments as with $\Omega_A^{(k)}$, overall given by

$$\Omega_B^{(k)}(X,Y) = \sum_{x\in X}\sum_{y\in Y}\sum_{l}^{k+1} \frac{J\left(N_k(x)_l, \Gamma_{xy}^{(k)}(l)\right)}{(k+1)\overline{d_{XY}}^{(k)}d_k(x,y)^{-1}}. \tag{6.10}$$

Co-occurrence matrix $\Omega_B^{(k)}$ is used in cost function C_3 in Equation 6.11, which has the same form as C_2, overall given by

$$C_3(\lambda,\gamma,U,V) = \left(\sum_{k=2}^{K}\exp\left(\frac{k}{\lambda}-\frac{10}{\gamma}\right)\right)\oslash\left(\sum_{k=2}^{K}\exp\left(\frac{k}{\lambda}-10d_{XY}^{(k)}/\gamma\right)\odot\Omega_B^{(k)}(X,Y)\right). \tag{6.11}$$

The earlier co-occurrence matrix $\Omega_A^{(k)}$ could possibly repeat information already given by the geometric similarity matrix, because of the promotion of more geometrically similar base pairs. The suppression of information in $\Omega_B^{(k)}$ from across-domain pairs with a higher likelihood of being matched allows for $\Omega_B^{(k)}$ to contribute information not present in the geometric similarity $d_{XY}^{(k)}$. The interactive effect of parameters λ and δ on matching accuracy with C_3 is illustrated in Figure 6.6 (D described in

Section 6.6), where standard deviation similarity is added to resolve some of the remaining subset confusion.

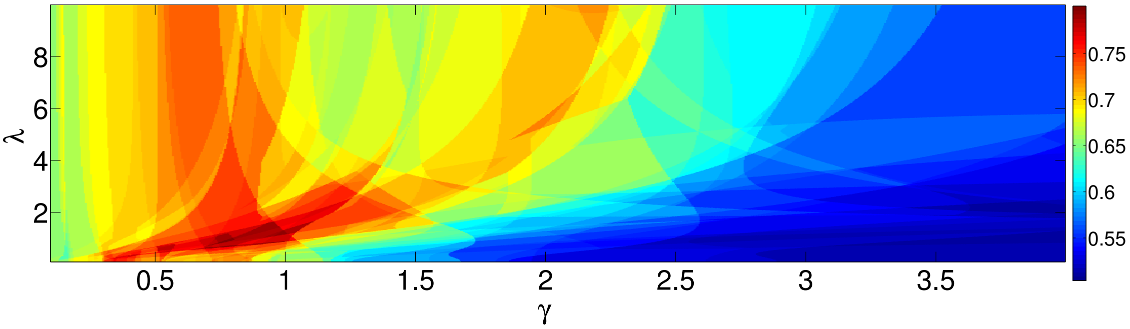


Figure 6.6. Domain matching accuracy $1-\Theta_S(C_3(\lambda,\gamma)+C_\sigma)$ as a colormap function of geometric similarity parameters λ and γ.

6.5.7 Combined cost function

Domain matching only with geometric similarity leads to significant confusion between close neighbours, as seen in the matching of Figure 6.7. Divergence, geometric and standard deviation similarities are thus combined in a cost function with optimal utility $\widehat{\Theta}_S(C_i)$ given in

$$\widehat{\Theta}_S(C_i) = \min_{\substack{\alpha,\beta,\delta,\\ \lambda,\gamma,\xi}} \Theta_S\big(\alpha C_\Delta(\delta)+\beta C_i(\lambda,\gamma)+\xi C_\sigma(\cdot)\big). \tag{6.12}$$

The contribution coefficients have property $\alpha+\beta+\xi=1$, so $\xi=1-\alpha-\beta$. Correction coefficient $\delta=1$ optimizes the divergence cost matrix, and $\gamma=1$ simplifies C_2 and C_3 without practical loss of generality, which leaves parameters α, β and λ to be learnt. The second cost term selects from geometric similarity measures C_1, C_2 or C_3. The matching accuracy is measured as a percentage-like $100(1-\widehat{\Theta}_S(C_i))$.

6.6 EXPERIMENTS AND RESULTS

An array S' of domain pairs is populated in $|D|!/|T|!$ different ways with every combination (D_i,D_j) of two images from every unique set of $|T|$ training images, selected from the study imagery of $|D|$ different acquisitions. At least one image from each different satellite is included in the training images T. Each training array S' has a corresponding test array $S\setminus S'$, where domain pair array S is populated from every two-image combination of all $|D|$ acquisitions.

The generalization performance of the combined cost function with geometric similarity $(K=11)$ was tested for different training sizes $|T|$, where parameters (α,β,λ) were optimized for the training array S' and the test matching accuracy measured with the optimized parameters on the previously unseen test array $S\setminus S'$. In the single case of $|T|=|D|$ the accuracy was tested on the training array only, corresponding to the last row in Table 6.2. Since there are multiple ways of instantiating S', the domain-matching accuracy means and standard deviations are calculated over the different S'.

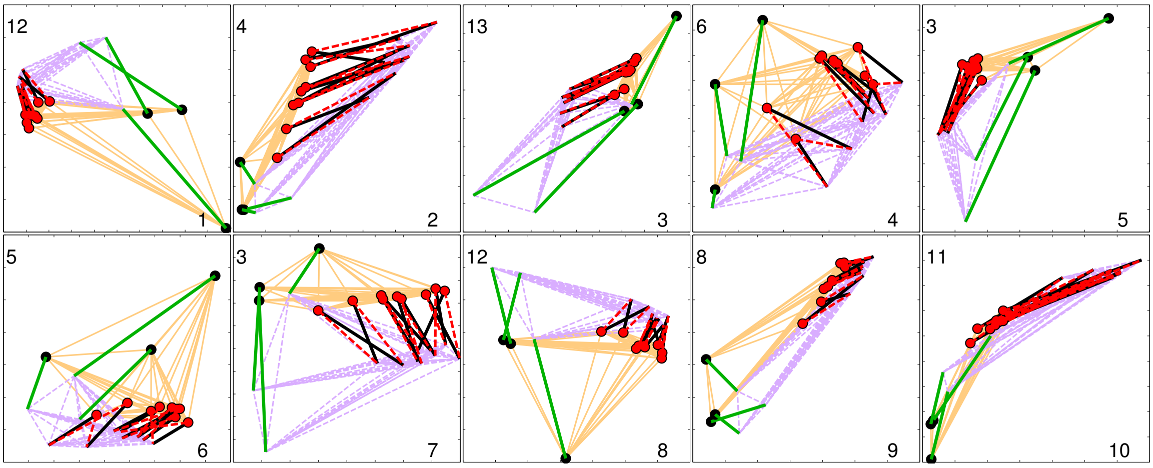


Figure 6.7. Domain-matching snapshot ($D_2 \leftrightarrow D_5$) with $C_1(6)$. Each axis indicates the GLCM feature number. Correct point matches (—●) from D_2 (- - -) to D_5 (—) are indicated, also incorrect matches (- - ●) with corrections (—●).

Table 6.2. Training accuracies $100(1-\widehat{\Theta}_{S'}(C_i))$ and test accuracies $100(1-\widehat{\Theta}_{S\setminus S'}(C_i))$ with optimal parameters $100(\alpha,\beta,\lambda)$ for different training image set sizes.

		Training accuracy (μ					$\pm\sigma$)	**Test acc.** (μ		$\pm\sigma$)
		C_1		C_2		C_3		C_1	C_2	C_3
Training set size $\|T\|$	3	87.6	±2.2	90.3	±10	**92.0**	±9.5	83.8	81.2	**84.3**
		(34, 35, 580)		(42, 34, 132)		(46, 28, 145)		±0.7	±8.9	±8.0
	4	88.9	±2.3	88.7	±3.8	**93.2**	±6.1	82.2	82.9	**88.7**
		(34, 35, 580)		(46, 28, 148)		(46, 28, 145)		±0.9	±3.9	±2.7
	5	87.3	±2.6	87.9	±1.5	**90.1**	±1.9	82.4	84.6	**89.0**
		(39, 32, 580)		(46, 28, 174)		(46, 28, 154)		±3.2	±7.0	±3.2
	6	85.9		87.8		**89.5**				
		(38, 32, 600)		(46, 28, 126)		(46, 28, 154)				

The domain-matching errors, out of a uniform maximum error of 12, for every domain pair are shown in Table 6.3 for different combined cost functions. The cost functions were optimized using all available domain pairs in array S. A comparison between the matching accuracy of different cost functions is shown in Table 6.4. For a training set size of 5 datasets there is a significant difference (with level of 0.1) between using C_1 and C_3 with two-tailed p-val of 0.09 using Welch's unpaired t-test, with a mean difference 95% confidence interval of -0.6 to 6.2. The variance appears to reduce with training set size and the mean difference appears to increase, which indicate the possibility of greater significant difference between using C_1 and C_3 with increase in training set size.

Table 6.3. Domain-matching errors for the different optimized geometric similarity cost functions, indicating only one off-diagonal triangular of each symmetric error matrix.

$0.38C_\Delta(1)+$ $0.32C_1(6) + 0.3C_\sigma$

$C_\Delta(1) + C_\sigma$

	D_1	D_2	D_3	D_4	D_5	D_6
D_1	•	0	3	0	0	0
D_2	0	•	2	0	2	0
D_3	0	2	•	3	0	3
D_4	0	0	3	•	2	0
D_5	2	2	0	4	•	2
D_6	0	0	4	0	4	•

$0.46C_\Delta(1)+$ $0.28C_3(1.54) + 0.26C_\sigma$

$0.46C_\Delta(1)+$ $0.28C_2(1.26) + 0.26C_\sigma$

	D_1	D_2	D_3	D_4	D_5	D_6
D_1	•	0	0	0	0	0
D_2	0	•	2	0	0	0
D_3	0	2	•	2	0	0
D_4	0	0	2	•	0	0
D_5	2	2	0	0	•	4
D_6	0	0	0	0	4	•

Table 6.4. Domain-matching accuracy ($100(1-\Theta_S(C))$) comparison for different cost functions C

Cost function	**Match accuracy**
$C_1(6)$	32.5
$C_2(1.26)$	39.3
$C_3(1.54)$	39.3
$C_\Delta(0)$	57.9
$C_\Delta(1)$	62.8
C_σ	64.7
$C_\Delta(1)+C_\sigma$	81.0
$0.49C_\Delta(1)+0.19d_{XY}^{(11)}+0.32C_\sigma$	83.8
$0.50C_\Delta(1)+0.08d_{XY}^{(3)}+0.42C_\sigma$	85.9
$0.38C_\Delta(1)+0.32C_1(6)+0.3C_\sigma$	85.9
$0.46C_\Delta(1)+0.28C_2(1.26)+0.26C_\sigma$	87.8
$0.46C_\Delta(1)+0.28C_3(1.54)+0.26C_\sigma$	**89.5**

6.7 DISCUSSION AND CONCLUSION

In this study the geometric similarity concept of Wang et al. [43] is isolated from the manifold alignment without correspondence, since the extended study dataset lends itself to a manifold alignment with correspondence. The objective was to use the concept of geometric similarity and to evaluate its preservation in a texture feature space for a land-use classification remote sensing scenario. The manifold matching problem can then be addressed and a solution for GLCM feature spaces can be demonstrated by using the Hungarian algorithm as solver and focusing on engineering a multicomponent cost matrix that performs generally well.

A multicomponent cost formulation for domain-matching with geometric similarity was presented, integrating divergence and standard deviation similarity to improve matching accuracy. There is confusion between nearby neighboring domain points, because of a relatively similar geometry with respect to the rest of the points. If divergence and standard deviation similarity is added to C_1 geometric similarity, the domain-matching accuracy improves from 32.5 to 85.9, out of a maximum 100, which would correspond to no errors being made for any matching between domain pairs in S. Geometric similarities for a range of different neighborhood sizes have been combined into a new geometric similarity formulation C_1, eliminating the need to find an optimal neighborhood size, which can vary significantly depending on the specific across-domain geometry. The combined formulation C_1 can achieve the same matching accuracy (85.9) as geometric similarity for an optimal neighborhood size of $k = 3$.

Local geometry matching co-occurrence was introduced to provide novel information on geometric similarity, and improved matching accuracy was achieved when combined with geometric similarity in cost function C_2, improving from 85.9 with C_1 to 87.8 with C_2. Local geometry matching co-occurrence matches the neighbourhoods of across-domain base pairs so that geometric dissimilarity is minimized. By suppressing across-domain point matches stemming from relatively geometrically similar base pairs, new information is probably contributed that is not given by geometric similarity alone. This reweighting of matching co-occurrences is implemented in C_3 and it increases matching accuracy from 87.8 for C_2 to 89.5 with C_3, for a dataset containing 15 different domain-matching problems. The parameters of the combined cost function with C_3 are also more stable and generalize better than with either C_1 or C_2, with the optimal α and β remaining unchanged and with λ varying relatively little.

The next chapter studies a feature learning strategy for reducing dataset shift in the case where there is no clear separation of dataset shift parts.

CHAPTER 7 MULTIVIEW DEEP LEARNING FOR LAND-USE CLASSIFICATION

7.1 CHAPTER OVERVIEW

In this chapter a multiscale input strategy for multiview deep learning is proposed for supervised multispectral land-use classification and it is validated on a well-known dataset. The hypothesis that simultaneous multiscale views can improve composition-based inference of classes containing size-varying objects compared to single-scale multiview is investigated. The end-to-end learning system learns a hierarchical feature representation with the aid of convolutional layers to shift the burden of feature determination from hand-engineering to a deep convolutional neural network (DCNN). This allows the classifier to obtain problem-specific features that are optimal for minimizing the multinomial logistic regression objective, as opposed to user-defined features, which trade optimality for generality.

A heuristic approach to the optimization of the DCNN hyper-parameters is used, based on empirical performance evidence. It is shown that a single DCNN can be trained simultaneously with multiscale views to improve prediction accuracy over multiple single-scale views. Competitive performance is achieved for the UC Merced dataset where the 93.48% accuracy of multiview deep learning outperforms the 85.37% accuracy of scale-invariant feature transform (SIFT) methods and the 90.26% accuracy of unsupervised feature learning.

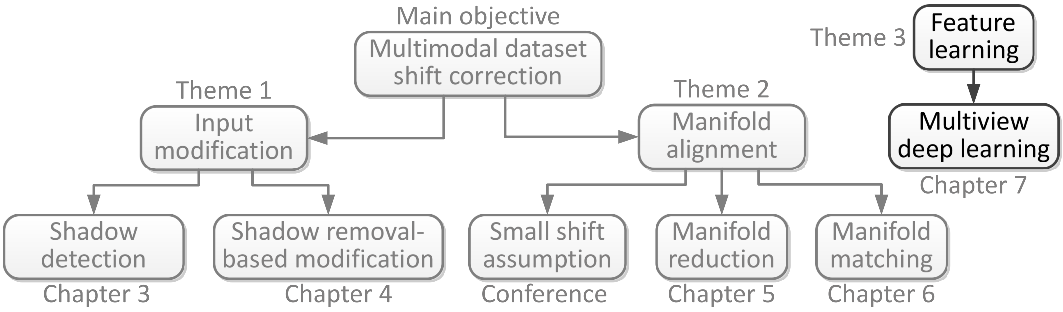


Figure 7.1. Indication of where this chapter fits into the thesis.

7.1.1 Contributions

1. Demonstration of pervasive and non-compartmentalized dataset shift when low-level features are used and the associated classification issues that arise from this dataset shift.

2. A basic DCNN with an architecture that produces above-average accuracy on the UC Merced dataset.

3. An improved DCNN with multiscale multiview input and classification probability averaging that results in a very competitive accuracy for the UC Merced dataset.

The problem statement is defined in terms of hypotheses and research questions in Section 7.2, whereafter related work is discussed in Section 7.3. An overview of the UC Merced dataset is given in Section 7.4 and the dataset shift problem is illustrated with low-level features in the form of GLCM texture features in Section 7.5. The deep learning design and methodology is discussed and the multiscale input method is described in Section 7.6, with the results presented in Section 7.7 where class confusion, convergence, visualization of the inner workings of the network, and comparison to the results of other published methods are addressed before a conclusion is reached in Section 7.8.

7.2 PROBLEM STATEMENT

Feature design has been a mainstay in classifier applications and much effort has been invested in hand-engineering specific features that are suitable only for select use-cases. The advent of graphics processing unit (GPU)-accelerated computational resources made feasible the implementation of multilayer convolutional neural network (CNN) approaches for classification. Deep learning discovers optimal features for the given problem in order to minimize the log loss cost function during classification. It is thus important to investigate the performance benefits of using the optimal problem-specific features learned by deep learning instead of using user-defined features that trades problem-specific optimality for general applicability.

The features discovered by deep learning are optimal in the sense that they minimize the multinomial logistic regression objective, and improved accuracy is expected compared to the use of more general user-defined features such as SIFT and Gabor features. The objective of this research was to design a DCNN for the UC Merced land-use dataset [52], a dataset compiled in 2010 and used as a benchmark in several land-use classification studies [52], [234], [235], [236], [237], [238]. The challenge is to optimize classification accuracy by finding a proper selection of DCNN hyper-parameters, which are defined as all the DCNN settings that exclude the learned neuron weights and biases, such as the architecture design, convolutional filter bank specifications, pooling layer specifications, and learning rate and momentum values.

While the hyper-parameter selection and the reduction of overfitting through data augmentation do have a significant impact on deep learning performance, an additional strategy is needed to achieve competitive performance. This requires moving beyond simple label-preserving transformations such as mirroring and rotation to augment the input dataset, while still adhering to the guiding principle of minimum intervention so that most of the feature learning burden can be delegated to the deep learning solution.

The approach contributed in this chapter is a generalization of the multiview strategy used by Krizhevsky et al. [49] that admits multiple view scales used to extract partial input sample patches. Classes with size-varying objects, such as storage tanks, can then potentially be recognized more accurately if consensus on multiscale views are used, a hypothesis tested in this research.

7.2.1 Hypotheses

1. Basic texture features will distinguish poorly between distinct land-use classes where there are both multimodal and semantic within-class variations, because the low level features may simultaneously be common across different classes, which causes excessive confusion.

2. The negative impact of multimodal image variances on classifier accuracy can be reduced with deep learning, since features are learned that are optimal for minimizing the classifier cost function.

3. A single DCNN with multiscale multiviews can improve composition-based inference of classes containing size-varying objects compared to single-scale multiview, since the size-varying objects have a greater probability of being featured at the right scale.

4. Increasing the number of different view scales can improve classification accuracy further, since a wider variety of object scales can then be accommodated.

7.2.2 Research questions

1. How can features be learned that are optimal for minimizing the classification loss function under multimodal image variances?

2. How should a DCNN be harnessed to improve classification where there are multiscale presentations of certain class characteristics, such as storage tanks that can vary in size, depending on the sample?

3. How can a basic DCNN implementation be improved upon in order to increase classification accuracy?

4. What are the optimal DCNN architecture and configuration for the UC Merced dataset?

7.3 RELATED WORK

Deep learning has been used previously in remote sensing for hierarchically extracting deep features with deep belief networks [239] or stacked auto-encoders in combination with PCA and logistic regression for hyperspectral data classification [240]. A hybrid DCNN was presented by Chen et al. [50] for improved vehicle detection in satellite images where variable-scale features are extracted through the use of multiple blocks of variable receptive field sizes or max-pool field sizes.

Remote sensing image fusion with deep neural networks (DNN) has been done by Huang et al. [241] using stacked modified sparse denoising auto-encoders for pretraining the hidden layers of the DNN to avoid the "diffusion of gradients" caused by random neuron initialization. Compressed-domain ship detection has also been performed with a DNN that provided high-level feature representation and classification in conjunction with an extreme learning machine that was used for feature pooling and decision making [242].

7.4 DATA DESCRIPTION

The UC Merced land-use dataset [52] is investigated, which is a set of aerial ortho-imagery with a 0.3048 m (1 foot) pixel resolution extracted from United States Geological Survey national maps. The UC Merced dataset has been used as a benchmark for land-use classifier evaluation in numerous publications [52], [234], [235], [236], [237], [238]. The dataset has been compiled from imagery over the US regions of Birmingham, Boston, Buffalo, Columbus, Dallas, Harrisburg, Houston, Jacksonville, Las Vegas, Los Angeles, Miami, Napa, New York, Reno, San Diego, Santa Barbara, Seattle, Tampa, Tucson and Ventura.

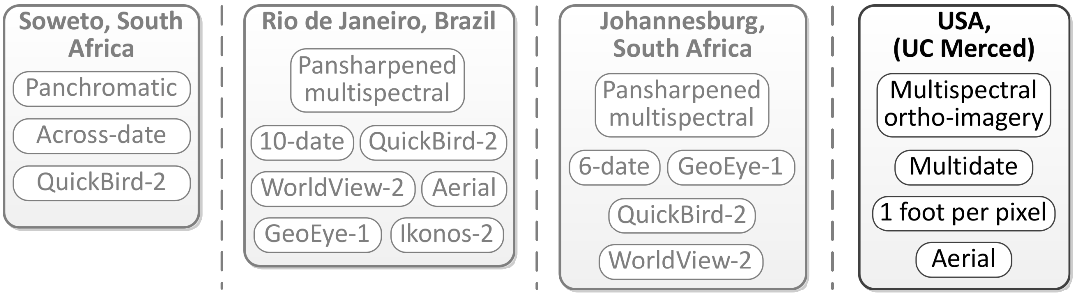


Figure 7.2. UC Merced dataset selection for this experiment.

The dataset consists of 21 land-use classes containing a variety of spatial patterns, some with texture and/or color homogeneity and others with heterogeneous presentation, as shown in Figure 7.3. The dataset was compiled from a manual selection of 100 images per class, each RGB image being approximately 256×256 pixels. The 21 land-use types include (a) agricultural, (b) airplane, (c) baseball diamond, (d) beach, (e) buildings, (f) chaparral, (g) dense residential, (h) forest, (i) freeway, (j) golf course, (k) harbor, (l) intersection, (m) medium density residential, (n) mobile home park, (o) overpass, (p) parking lot, (q) river, (r) runway, (s) sparse residential, (t) storage tanks, and (u) tennis court classes.

7.5 MULTISPECTRAL GLCM FEATURE CLASSIFICATION

7.5.1 Separability of basic features

Ordinarily with dataset shift there is a distinct change in measurement mode or difference between the training and testing data that can invalidate learnt class distinction rules. In the UC Merced dataset there is a pervasive dataset shift that makes it difficult to separate the samples into a train/test split that would display the conventional dataset shift scenario between the train and test datasets. This is because of the wide location variety of samples for any given class, where high-level semantic features are the predominant reason for class labeling. This stands in contrast to larger-area land-use classification problems, which are more accurately characterised by lower-level features such as texture features.

Classifications that depend on high-level features, such as the presence of certain objects and combinations of semantic elements, may display poorly separable low-level features. Differences between samples, such as different crops in agriculture and different sized storage tanks, can be high-level causes of dataset shift between the samples. Heterogeneous presentation of low-level features throughout a dataset presents a more severe form of dataset shift, where the solution to accurate supervised classification cannot exploit an explicit shift between the train and test data.

7.5.2 Multispectral GLCM features

To demonstrate that lower-level features become poorly separable in the case of higher-level semantic classification, multispectral GLCM texture features are investigated and their separability tested in supervised classification. The land-use image samples of approximately 256×256 pixels are converted into four GLCMs with pair offsets (row offset, column offset) of (2, 0), (0, 2), (-2, 2), and (2, 2).

The statistics calculated for each of the GLCMs include contrast, correlation, energy and homogeneity. The contrast measures local variations, correlation measures joint probability occurrence of pixel pairs, energy measures the sum of squared GLCM entries and homogeneity measures the closeness of the distribution of elements in the GLCM to the diagonal of the GLCM.

$$\text{Contrast:} \qquad f_1 = \sum_{i,j} |i-j|^2 p(i,j) \tag{7.1}$$

$$\text{Correlation:} \qquad f_2 = \frac{\sum_i \sum_j (ij)p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y} \tag{7.2}$$

$$\text{Energy:} \qquad f_3 = \sum_{i,j} p(i,j)^2 \tag{7.3}$$

$$\text{Homogeneity:} \qquad f_4 = \sum_i \sum_j \frac{1}{1+|i-j|} p(i,j) \tag{7.4}$$

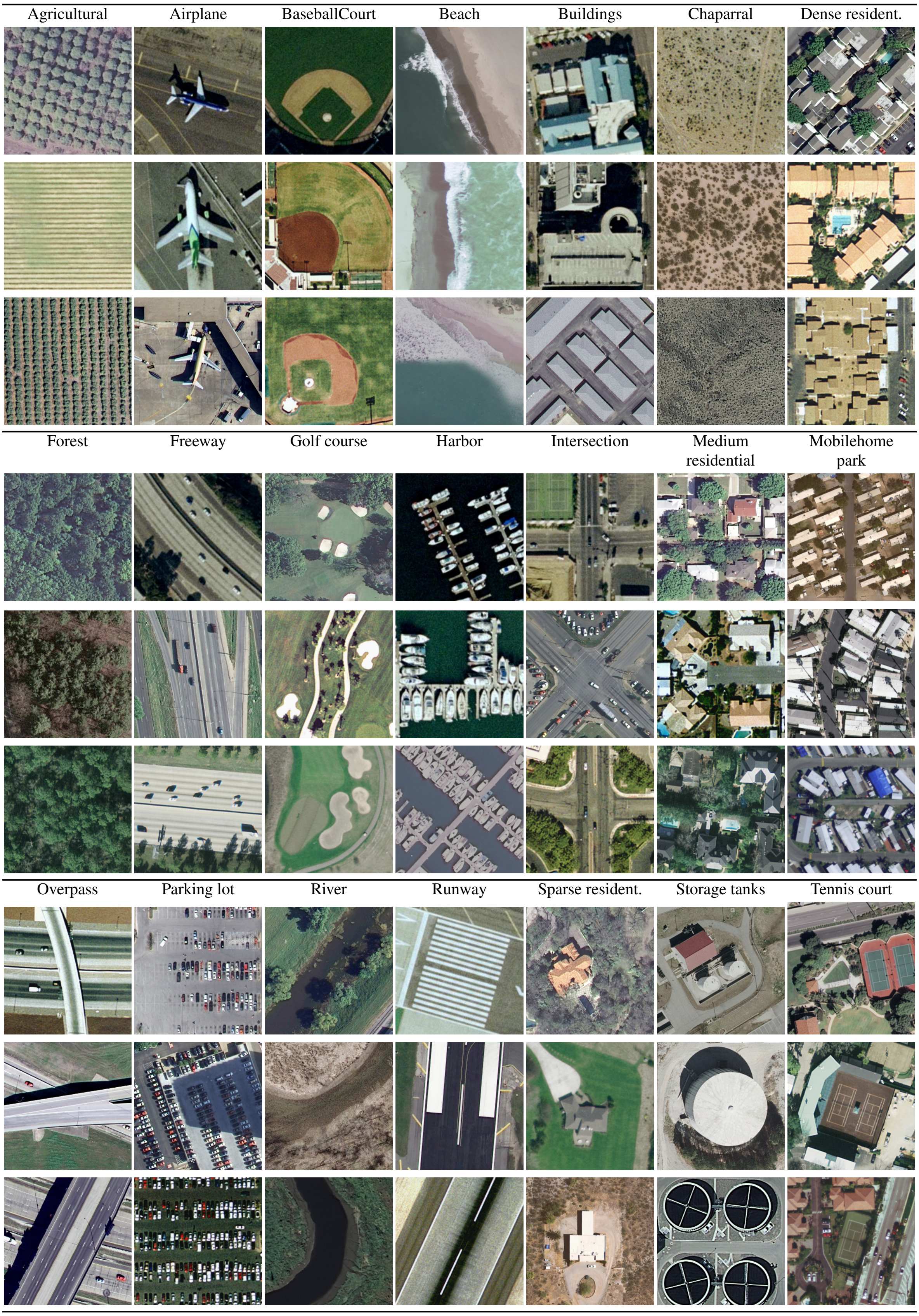


Figure 7.3. Image samples of the UC Merced land-use dataset classes, courtesy of United States Geological Survey National Maps.

The mean of each statistic f_i over the four GLCMs is used to obtain four features for a given color channel, which contribute to a total of 12 multispectral GLCM texture features (4×3 color channels) for a given image sample.

7.5.3 Principal component analysis

The fitness of dense multispectral texture features to distinguish between the higher-level semantic classes in the UC Merced dataset can be evaluated visually through PCA. PCA finds linear combinations of texture features that account for most of the variance observed in the feature samples, with orthogonality between different combinations. High fitness will manifest as clear separation between the primary principal component score clusters of different classes.

In Figure 7.4 the 21 different UC Merced class principal component score clusters are depicted for the two most significant principal components with color differentiation between classes. Each cluster is accompanied by a 1σ-ellipse centered at the class score mean. It is apparent that the multitude of classes are highly overlapped in the principal component space, with some exception for the *agricultural* (red) and *mobile home park* (blue) classes.

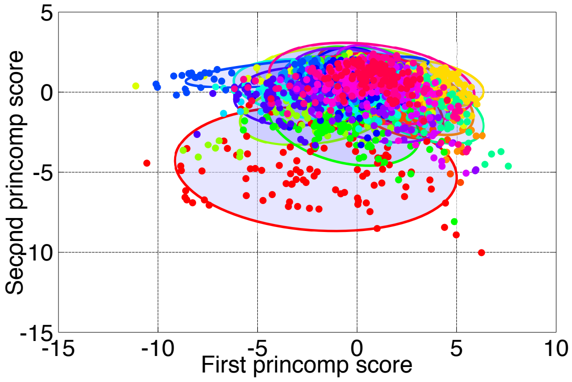
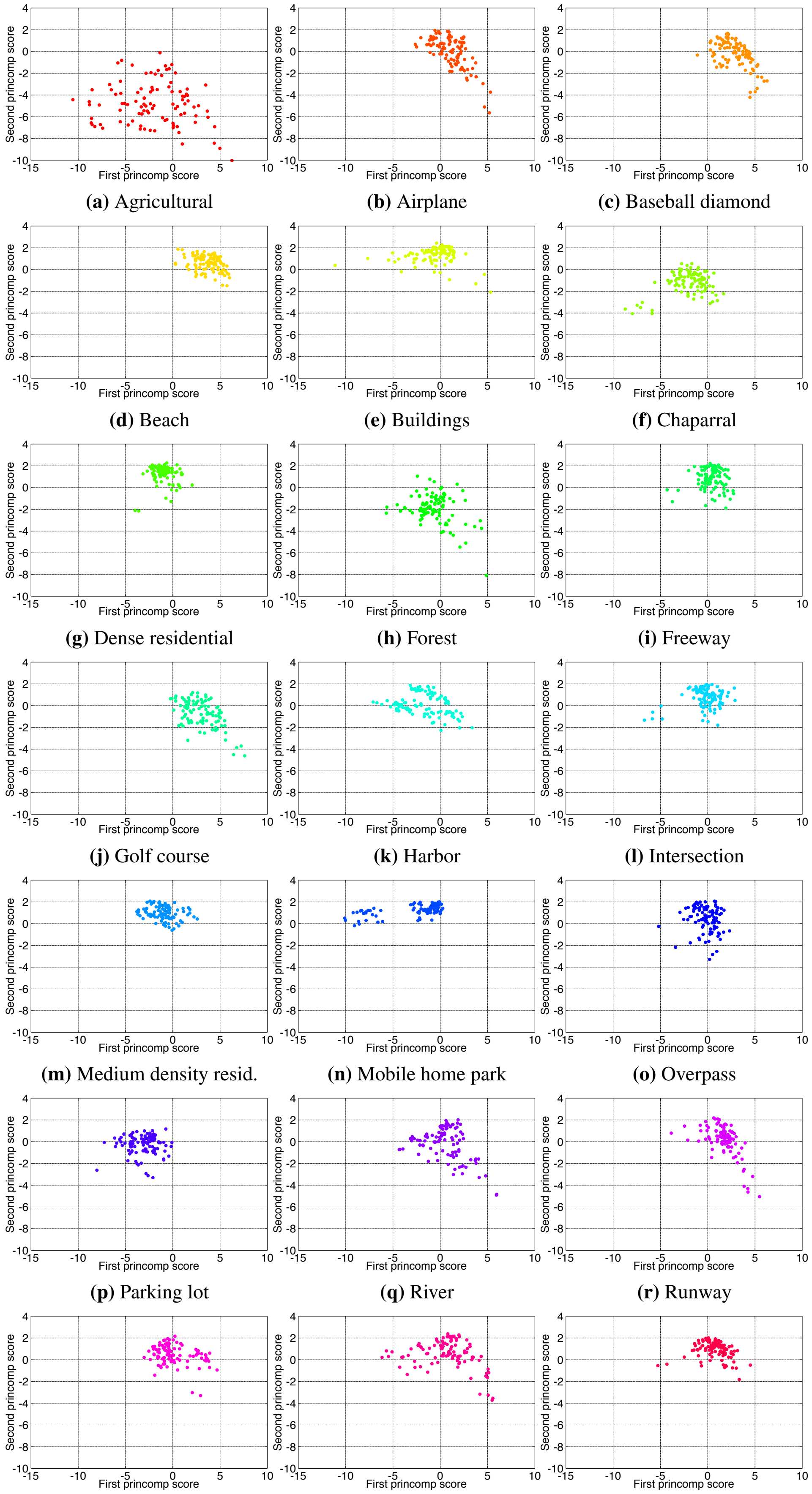


Figure 7.4. UC Merced class separation with multispectral GLCM features.

Consequently, the separability of the classes with multispectral texture features is likely to be poor, as depicted by the visual PCA in Figure 7.4. The 21 different land-use classes in UC Merced are also displayed separately in Figure 7.5, to show the details occluded in Figure 7.4 more clearly. Most classes appear to be unimodal with relatively low variance, with the notable exception of the *mobile home park* class, which presents with two modes, and the agricultural class, which has relatively higher variance.

7.5.4 Classification with neural network

Visual analysis in the principal component space in the previous subsection revealed high overlap of classes, so there is an expectation of a relatively low supervised classification accuracy. This expectation is tested by performing machine learning with the multispectral texture features using a neural network that is trained with standard backpropagation. The main objective is to measure the



(a) Agricultural **(b)** Airplane **(c)** Baseball diamond

(d) Beach **(e)** Buildings **(f)** Chaparral

(g) Dense residential **(h)** Forest **(i)** Freeway

(j) Golf course **(k)** Harbor **(l)** Intersection

(m) Medium density resid. **(n)** Mobile home park **(o)** Overpass

(p) Parking lot **(q)** River **(r)** Runway

(s) Sparse residential **(t)** Storage tanks **(u)** Tennis court

Figure 7.5. PCA of multispectral GLCM features of the UC Merced land-use classes.

Table 7.1. Multispectral GLCM NN confusion matrix for the lexicographically sorted UC Merced land-use classes for the full-sample training dataset.

		Predicted class label																				
		a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u
a	Agricultural	87	0	0	0	0	4	0	6	0	1	0	0	0	0	0	2	0	0	0	0	0
b	Airplane	0	68	6	7	0	0	4	0	2	2	2	0	1	0	1	3	0	1	1	2	0
c	Baseball diamond	0	4	70	6	0	0	1	0	1	6	0	0	1	0	1	0	0	1	6	2	1
d	Beach	0	0	17	78	0	0	0	0	4	0	0	0	0	0	0	0	0	0	1	0	0
e	Buildings	0	19	0	1	20	0	10	0	8	1	5	0	6	4	1	9	0	0	1	15	0
f	Chaparral	16	0	0	0	0	69	0	8	0	1	0	0	0	0	0	1	0	0	5	0	0
g	Dense residential	1	5	1	0	0	0	59	0	1	0	1	0	15	0	5	5	0	0	2	2	3
h	Forest	12	1	1	0	0	6	0	71	0	2	0	0	0	0	0	0	6	0	1	0	0
i	Freeway	0	5	8	4	1	0	1	1	42	1	0	0	6	0	6	3	2	1	13	5	1
j	Golf course	1	3	38	2	0	0	0	2	0	42	0	0	0	0	0	0	4	0	8	0	0
k	Harbor	0	11	0	0	0	0	0	0	0	0	82	0	0	0	0	7	0	0	0	0	0
l	Intersection	0	3	3	3	1	0	17	0	19	0	0	3	6	0	14	6	3	1	13	7	1
m	Medium residential	0	1	5	0	2	0	36	0	6	0	0	1	16	0	0	10	4	0	16	0	3
n	Mobile home park	0	3	0	0	4	0	38	0	0	0	0	0	8	30	0	7	0	0	8	2	0
o	Overpass	4	17	2	0	3	0	15	0	12	0	0	0	8	1	24	5	0	1	0	8	0
p	Parking lot	0	0	0	0	0	1	2	0	0	0	4	0	2	0	0	89	0	0	1	1	0
q	River	2	2	8	1	0	1	0	5	5	7	2	0	0	0	1	12	31	0	19	2	2
r	Runway	0	21	28	12	0	0	0	0	8	7	0	0	0	0	6	0	0	4	1	13	0
s	Sparse residential	0	0	17	3	0	5	5	2	3	5	0	0	9	1	0	0	6	1	41	0	2
t	Storage tanks	0	17	8	10	3	0	2	0	6	6	0	0	0	3	1	9	0	1	2	31	1
u	Tennis court	0	3	21	2	1	0	10	0	4	6	0	0	7	1	1	3	1	0	14	11	15

training accuracy, which can help to assess the fitness of the features to distinguish accurately between classes.

7.5.4.1 Neural network design

The 12 multispectral GLCM features are standardized before being input to a neural network with one hidden layer, with a number of units consisting of the geometric mean of the number of features and the number of classes or output nodes. The 16 hidden units have a sigmoid nonlinear function ($s(x) = (1+\exp(-x))^{-1}$) and the 21 output nodes are modeled by a softmax layer ($o(z)_j = \exp(z_j)/\sum_k \exp(z_k)$). The learning rate is set at 0.05 and the number of backpropagation iterations is 1000. The neural network is given adequate capacity to learn class distinctions, but the expectation is that poor features will not allow for the neural network to learn the necessary distinctions. This is tested by allowing the neural network to train on all the available data and then to measure the training accuracy and confusion, as shown in Table 7.1.

7.5.4.2 Classification accuracy

An instance training accuracy of 46.3% is observed when training the neural network with all available samples, which is a comparatively poor accuracy, since good features with adequate neural network capacity should normally obtain an accuracy close to 100%. An over-capacity neural network, i.e. a neural network with many more hidden units than the sum of input and output units, will usually overfit given a large enough number of training iterations and a suitable learning rate. Overfitting will increase the training accuracy even further, although the neural network is not likely to be able to generalize that performance. For the given machine learning task a significant increase in the number of hidden units did not increase training accuracy at any point.

7.5.4.3 Confusion analysis

According to the confusion table in Table 7.1 the best distinguishable classes include *parking lot*, *agriculture*, *harbor*, *beach* and *forest*. Some of these classes appear to have more homogeneous and sometimes dense textures that can be accurately characterized with GLCM features, which is a possible explanation for these classes being distinguishable with multispectral texture features alone. Some of the classes confused most often with texture feature classification are *intersection*, *runway*, *tennis court classes* and *buildings*, which tend to be classes with larger features defining the semantic character of the class, where the features are not likely to be characterized well with dense textures.

7.6 METHODOLOGY

In this section an overview of the training of a DCNN is first given, followed by a description of the important processing layers of a DCNN. The specific DCNN architecture instantiation developed for the UC Merced dataset is then defined and then methods of reducing training overfitting are given. A multiscale multiview input strategy is then described that utilizes the defined DCNN.

7.6.1 Deep learning

Deep learning is characterized as an end-to-end learning system typically consisting of more than five processing layers, which is usually supervised and produces a discriminative classification for a given input. The burden of feature determination is shifted to a DCNN, which learns the optimal features for the given problem in order to minimize a loss cost function. The features are learned in a hierarchical manner where higher-level features are learned in deeper convolutional layers as combinations of lower-level features determined in shallow layers.

7.6.1.1 Learning objective

Improved accuracy is expected by directly learning the features that minimize the multiclass log loss cost function $L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} y_{i,j}\log(p_{i,j})$ for a given dataset with N samples, compared to using predetermined features. The natural logarithm of the probability $p_{i,j}$ of sample i belonging to class j is counted by setting $y_{i,j} = 1$ only if i belongs to class j. Stochastic gradient descent can be used since the loss function is a sum of differentiable functions, and Nesterov's accelerated gradient in particular has been shown to be effective despite the use of noisy gradient estimates [243]. The update increment v_{t+1} and the updated network parameters w_{t+1} are calculated as

$$v_{t+1} = \mu \cdot v_t - \varepsilon \cdot \left\langle \left.\frac{\partial L}{\partial w}\right|_{w_i+\mu\cdot v_t} \right\rangle_{B_i} \tag{7.5}$$

$$w_{i+1} = w_i + v_{t+1} \tag{7.6}$$

with momentum μ and learning rate ε.

The loss gradient estimate $\frac{\partial L}{\partial w}$ is determined for the average loss over a smaller batch B_i of input samples for the DCNN parameters equal to $w_i + \mu \cdot v_t$.

7.6.1.2 Stopping condition

A validation split of the training data is omitted, since explicit measures are taken to reduce the likelihood of potential overfitting, such as data augmentation and dropout. So the expectation is that early stopping with validation set accuracy tracking is not required, as the test accuracy will stabilize because of very little overfitting with extended backpropagation learning. The final DCNN weight and bias parameters are based on the epoch registering the minimum value for the log loss cost function on the training data, which is a stopping condition that does not require a validation dataset.

7.6.2 Architecture definitions

7.6.2.1 Convolutional layers

A CNN consists of convolutional layers, each followed by optional sub-sampling and regularization layers, and ending in fully connected 1D hidden layers. A convolutional layer receives a 3D input and creates a 3D output that measures the filter responses at each input location, calculated as the sum of the element-wise incidence product between the filter and image window. This convolutional response encodes the input in terms of learned templates to systematically reduce input dimensionality as part of feature determination.

7.6.2.2 Activation functions

Each filter response becomes the input to a nonlinear activation function, which should be non-saturating in order to accelerate learning. Rectified linear units (ReLU) ($f(x) = \max(0,x)$) are used in lieu of saturating nonlinearities after every convolutional and fully connected layer, except for the final dense layer, which uses softmax activation ($f(x_j) = e^{x_j} / \sum_k e^{x_k}$) to maximize the multinomial logistic regression objective. Network implementation is simplified with the use of ReLU, as this activation function does not require input normalization to avoid saturation, although local normalization can promote improved generalization [49].

7.6.2.3 Sub-sampling layers

Sub-sampling layers normally proceed convolutional layers to further reduce feature dimensionality, but also to achieve translation invariance in the case of max-pool sub-sampling layers [49]. For example, a 2×2 max-pool layer divides the convolutional layer output into a set of non-overlapping 2×2 cells and only records the maximum activated filter response in each cell, thereby halving the input

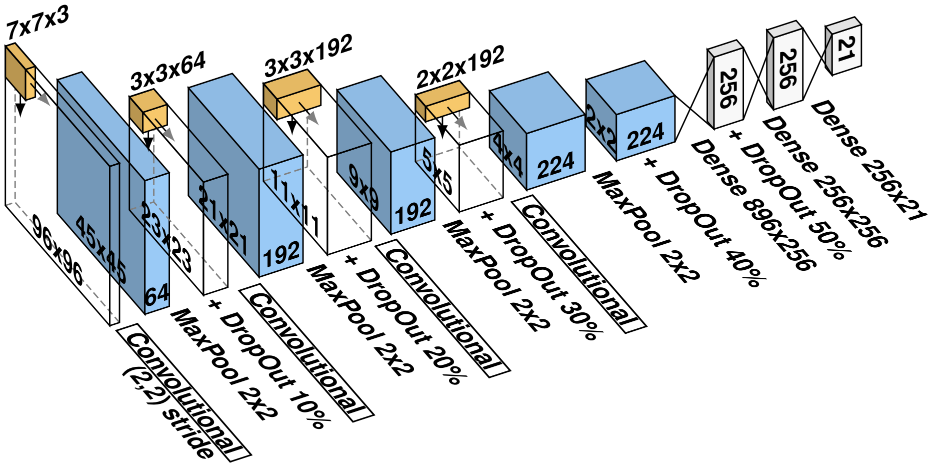

7x7x3
96x96
45x45
64
Convolutional (2,2) stride
3x3x64
23x23
21x21
MaxPool 2x2 + DropOut 10%
3x3x192
192
11x11
Convolutional
9x9
MaxPool 2x2 + DropOut 20%
2x2x192
192
5x5
Convolutional
4x4
224
MaxPool 2x2 + DropOut 30%
Convolutional
2x2
224
MaxPool 2x2 + DropOut 40%
256
Dense 896x256 + DropOut 50%
256
Dense 256x256
21
Dense 256x21

Figure 7.6. CNN architecture with four convolutional layers accepting $96 \times 96 \times 3$ inputs and resolving to a 21-class softmax output layer.

dimensions and producing features that are increasingly invariant to image object translations.

7.6.3 Architecture instantiation

The DCNN design given in this subsection was heuristically selected based on experimental investigation that adhered to the objective of layer dimension reduction, since it develops a strong hierarchical feature representation. The DCNN designed for the UC Merced dataset accepts a $96\times96\times3$ input, which can be converted from an RGB to HSV color model. The HSV color model can more directly concentrate chromaticity to single filter layers, which can potentially simplify features and allow for the reduction of network complexity.

The input is converted to $45\times45\times64$ neurons with the first convolutional layer using 64 filters of $7\times7\times3$ operating at a stride of $(2,2)$, before being sub-sampled with a 2×2 max-pool layer to obtain a $23\times23\times64$ output with 10% dropout. The second convolutional layer uses 192 filters of $3\times3\times64$ with a $(1,1)$ stride to produce a $21\times21\times192$ output, which is sub-sampled with a 2×2 max-pool to give a $11\times11\times192$ output with 20% dropout, as shown in Figure 7.6.

A third convolutional layer with 192 filters of $3\times3\times192$ and a stride of $(1,1)$ produces a $9\times9\times192$ output followed by a 2×2 max-pool layer, which outputs $5\times5\times192$ neurons with 30% dropout. The final convolutional layer has 224 filters of $2\times2\times192$ with a $(1,1)$ stride and gives a $4\times4\times224$ output, which is max-pooled with 2×2 cells to render a $2\times2\times224$ output with 40% dropout. A fully connected dense layer with 256 hidden units is used with ReLU activation and 50% dropout follows, after which another dense layer with 256 hidden units is used before resolving to 21 units in a softmax output layer. All neuron biases are set to 0 and network weights are initialized randomly according to normalized initialization $U\left[-\frac{\sqrt{6}}{\sqrt{n_j+n_{j+1}}}, \frac{\sqrt{6}}{\sqrt{n_j+n_{j+1}}}\right]$ given by Glorot et al. [244] where n_j and n_{j+1} are the number of neurons in layers j and $j+1$, respectively.

7.6.4 Reducing overfitting

7.6.4.1 Dropout

Convolutional and fully connected layers can be interconnected so that hidden neuron outputs are deactivated with probability p during training, with the remainder of the outputs multiplied by $\frac{1}{1-p}$. This strategy reduces the co-adaptation of neurons, since dropout forces neurons to provide more useful and robust contributions in combination with arbitrary active neuron combinations [49]. The set of dropped neurons changes randomly at every epoch, which changes the architecture and reduces overfitting at the cost of approximately $\frac{1}{1-p}$ times the convergence period compared to training without dropout.

7.6.4.2 Data augmentation

The original input dataset can be expanded with label-preserving transformations such as horizontal and vertical flips and rotation. This presents the network with an enlarged set of inputs, which may contain examples present in the test dataset but not in the original training dataset, thus improving classification accuracy. During training all views are flipped horizontally or vertically with probability of 0.5, but for testing the model averaging only considers the untransformed views. The classifier is trained with transformed views so that any untransformed view can be recognized during testing.

7.6.5 Multiview deep learning

Another form of data augmentation involves the use of multiple partial views of a given input sample to train with, and classifying test samples with the mean softmax output averaged over a predetermined set of classified patches or views, i.e. model averaging [49]. Some classes are distinguished by the presence of certain objects, such as airplanes and storage tanks, which only occupy a portion of a given sample. If these objects vary in size across different samples then multiscale views can potentially produce stronger activations with higher probability than single-scale views.

The main contribution proposed is that a single DCNN can be trained with multiscale views to obtain improved classification accuracy compared to using multiple views at one particular scale only. The UC Merced dataset samples are downsampled from 256×256 to 96×96 based on empirical evaluation of the optimal input size, and 10 multiscale views are extracted as described below. The first four augmenting views are acquired at the image corners at 75% input coverage, while the fifth view has 100% coverage. Views six to ten are obtained at the corners and center at 50% input coverage, and all extracted views are scaled to the input size of 96×96, as shown in Figure 7.7.

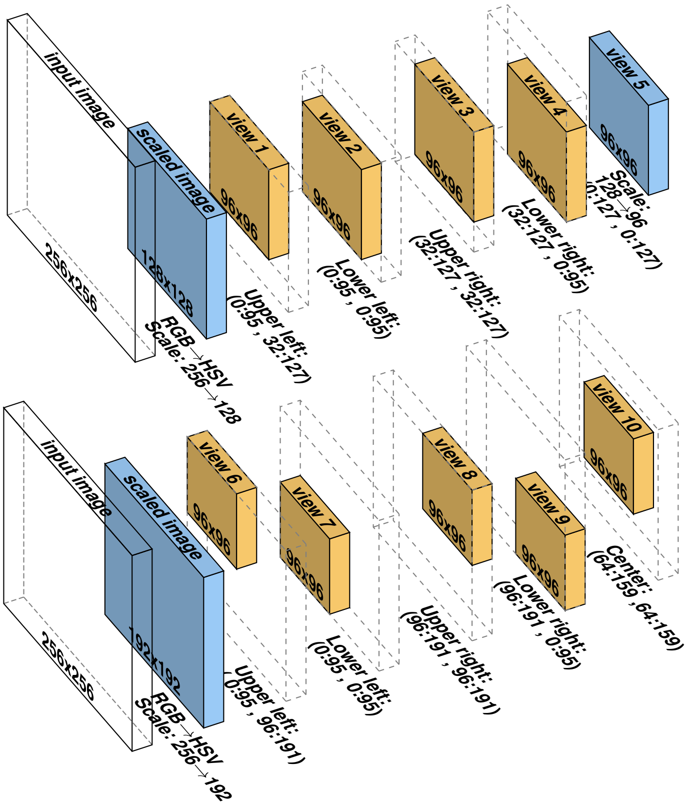


Figure 7.7. Partial view selection specifications for composing a multiview input dataset consisting of $10 \times 96 \times 96 \times 3$ inputs per sample.

7.7 RESULTS AND DISCUSSION

7.7.1 Experimental setup

The standard benchmark conditions for the UC Merced dataset first stipulated in [52] are followed to measure classification accuracy. Five-fold stratified cross-validation is used for all experiments, where four folds are used for training and model selection, and the remaining unseen fold of 20% of the dataset is classified to measure accuracy.

Initial empirical evaluation indicated that the salient hyper-parameters that influence accuracy most include the input size, first convolutional filter size and filter amount, and the network learning rate. Hyper-parameter range selections are based on values that resulted in high classification accuracy during an initial evaluation. Various architecture instantiations are evaluated empirically with the focus on optimizing the aforementioned hyper-parameters with full knowledge of the test accuracy.

The use of knowledge of test accuracy for hyper-parameter optimization is motivated as follows:

1. The hyper-parameter selection is highly constrained with relatively few values to be optimized.

2. Routine full-knowledge optimization is implicitly involved in the creation of machine learning architectures in the literature, where architectures or solutions with inferior performance would

Table 7.2. Five-fold cross-validation accuracy for various DCNN architectures. All instantiations use Nesterov Accelerated Gradient (linear momentum $\mu = 0.9 \rightarrow 0.999$, linear learning rate decrease to 0.0001, batch sizes $|B_i| = 128$)

Parameter	**Architecture**						
	#1	#2	#3	#4	#5	#6	#7
Input size	80×80	96×96					128×128
Filter 1 size	7×7	7×7	7×7	7×7	7×7	9×9	9×9
Learning rate	0.005	0.005	0.005	0.005	0.005	0.005	0.01
Max epochs	1000	1000	300	300	300	1000	1000
Multiview	1	1	5	5	10	1	1
Multiscale	×	×	×	✓	✓	×	×
HSV: Acc. (μ $\pm\sigma$)	86.76 ±1.74	88.00 ±2.88	90.53 ±1.87	91.18 ±1.62	92.34 ±1.25	87.10 ±1.98	83.29 ±2.83
RGB: Acc. (μ $\pm\sigma$)		87.14 ±3.77	91.10 ±0.80	92.76 ±1.46	93.48 ±0.82		

be discarded based on test accuracy.

3. Full-knowledge hyper-parameter optimization is used for both single-scale and multi-scale DCNN solutions, so performance differences of these multiview strategies can be highlighted.

4. The main learning task of the DCNN is still preserved with knowledge only of the training information.

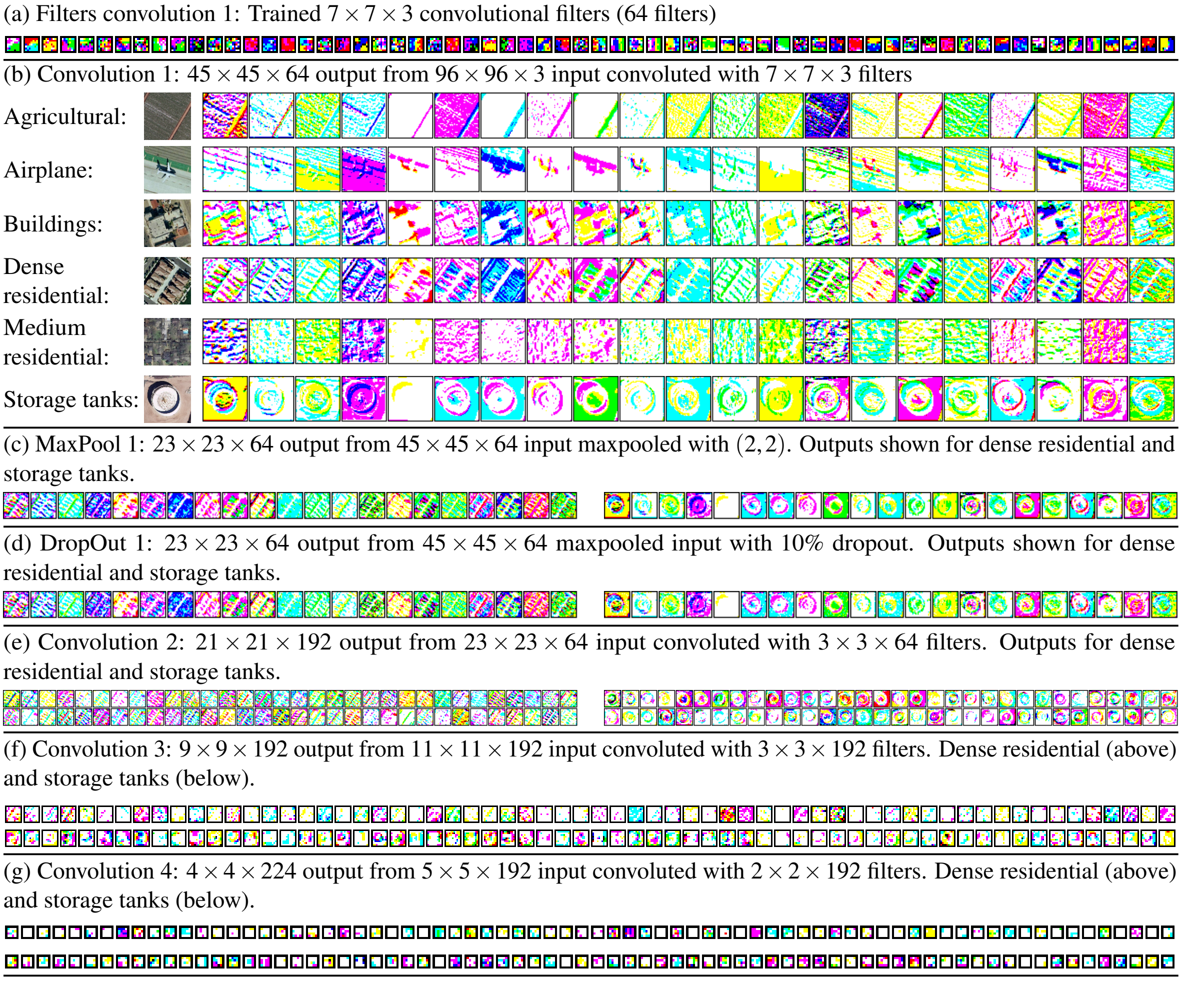
7.7.2 Architecture selection

Several architectures have been evaluated to obtain the best performing DCNN for the UC Merced dataset, and the results are shown in Table 7.2. The important design choices include the reduction in learning rate, using model averaging with an increasing number of multiple views, and finding the optimal input size of 96×96. The single-scale multiview input of Krizhevsky et al. [49] has been implemented in arch. #3, but its 91.1% accuracy is outperformed by the 92.75% of multiscale input (arch. #4). Using the first five views (arch. #4 in Table 7.2) specified in Figure 7.7 improved test accuracy from 87.14% to 92.76%, but using model averaging with all 10 views (arch. #5 in Table 7.2) resulted in an accuracy of 93.48% for RGB inputs.

Multiview (single-scale) input significantly (level of 0.05) improves over single-view given a two-tailed p-val of 0.05 using Welch's unpaired t-test, with a mean difference 95% confidence interval of -0.01 to 7.9. Multiscale significantly (level of 0.1) improves over single-scale given a two-tailed p-val of 0.07 using Welch's unpaired t-test, with a mean difference 95% confidence interval of -0.2 to 3.5.

7.7.3 DCNN weight visualization

Figure 7.8 displays a visualization of the trained single-view DCNN architecture #2 (see Table 7.2), showing the first convolutional filters and the convolutional responses for a selection of UC Merced classes. The first max-pool and dropout outputs are also shown to illustrate their functions of sub-sampling and omission noise. The second, third, and fourth convolutional filter banks are too large to display and are not included. The convolutional filters are the core features that are learnt by the DCNN and it is seen that the network reduces convolutional response dimensions to a final single-dimensional response appropriate for the use of softmax activation.


(a) Filters convolution 1: Trained 7 × 7 × 3 convolutional filters (64 filters)

(b) Convolution 1: 45 × 45 × 64 output from 96 × 96 × 3 input convoluted with 7 × 7 × 3 filters

(c) MaxPool 1: 23 × 23 × 64 output from 45 × 45 × 64 input maxpooled with $(2,2)$. Outputs shown for dense residential and storage tanks.

(d) DropOut 1: 23 × 23 × 64 output from 45 × 45 × 64 maxpooled input with 10% dropout. Outputs shown for dense residential and storage tanks.

(e) Convolution 2: 21 × 21 × 192 output from 23 × 23 × 64 input convoluted with 3 × 3 × 64 filters. Outputs for dense residential and storage tanks.

(f) Convolution 3: 9 × 9 × 192 output from 11 × 11 × 192 input convoluted with 3 × 3 × 192 filters. Dense residential (above) and storage tanks (below).

(g) Convolution 4: 4 × 4 × 224 output from 5 × 5 × 192 input convoluted with 2 × 2 × 192 filters. Dense residential (above) and storage tanks (below).

Figure 7.8. Filters and CNN layer outputs for single-view architecture #2 (see Table 7.2) and inputs from a selection of classes. Output visuals are mapped to full channel range and combined in some cases to occupy all RGB channels.

7.7.4 Convergence analysis

The convergence rate is illustrated for architecture #5 (see Table 7.2) in Figure 7.9, comparing the progression of training and testing accuracies in terms of training epochs. The single-view test

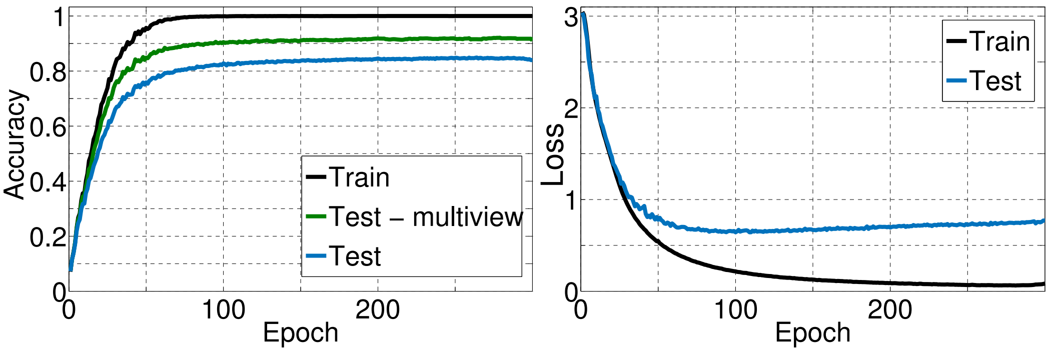


Figure 7.9. Averaged five-fold cross-validation accuracy graphs for multiview architecture #5 (see Table 7.2).

accuracy is also shown, which performs more poorly than with multiview model averaging. The best test accuracy loss is obtained around epoch 100 and degrades from that point, while training loss keeps improving. However, the measures employed to reduce overfitting allow for the test accuracy to keep improving by 1-2% even while the multinomial logistic regression score deteriorates.

7.7.5 Confusion analysis

A confusion matrix for a given DCNN instantiation is calculated in this section as the sum of the five confusion matrices on the test data of each fold in a five-fold cross-validation setup, so as to obtain full representation of all available data samples as test samples. The confusion matrix for single-view architecture #2 (see Table 7.2) is shown in Table 7.3, and the confusion matrix for 10-view architecture #5 is given in Table 7.4. The confusion matrix for architecture #3 (see Table 7.2) with five single-scale multiviews is shown in Table 7.5, where the predicted class label counts are given for each correct class row. The confusion matrix for architecture #4 with five multiscale multiviews is shown in Table 7.6, which is compared against the confusion matrix for architecture #3.

The largest class-specific accuracy increases with multiscale views over single-scale views (arch. #4 vs. #3) are seen in the *buildings*, *intersection*, *storage tanks*, *overpass*, *beach*, *dense residential*, *runway* and *tennis court* classes. The largest class-specific accuracy decreases with multiscale views over single-scale views are seen in the *sparse residential*, *river* and *baseball diamond* classes. Some of the largest clarifications made by multiscale views over single-scale views (arch. #4 vs. #3) are *sparse residential-forest*, *sparse residential-golf course*, *storage tanks-medium density residential*. The largest confusions introduced by multiscale views include *buildings-overpass*, *overpass-freeway*, *intersection-freeway*, *medium density residential-dense residential* and *beach-agriculture*.

Sparse residential samples often present characterizing buildings at widely different scales, whereas *beach* and *agriculture* classes often present with more homogeneous textures. Large characterizing structures such as *overpasses*, *intersections* and *freeways* may not always be captured coherently in one view as easily as smaller structures such as *storage tanks*. The result is that multiscale views can introduce confusion in the case of large features, but on the other hand it can present more consistent

Table 7.3. Confusion matrix with arch. #2 for a five-fold cross-validation sample result.

		Predicted class label																				
		a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u
a	Agricultural	96	0	0	3	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
b	Airplane	0	87	1	0	2	0	0	0	0	1	1	0	0	0	1	0	0	5	0	2	0
c	Baseball diamond	1	0	92	1	0	0	0	0	0	2	0	0	0	0	1	0	0	0	0	1	2
d	Beach	4	0	0	95	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
e	Buildings	0	2	0	0	63	0	11	0	2	0	0	1	1	1	2	3	0	0	0	6	8
f	Chaparral	2	0	0	0	0	97	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
g	Dense residential	0	0	0	0	4	1	70	1	1	0	0	0	16	3	0	1	0	0	0	0	3
h	Forest	1	0	0	0	0	2	0	96	0	0	0	0	0	0	0	0	1	0	0	0	0
i	Freeway	0	0	0	1	0	0	0	0	93	0	0	1	0	0	2	0	1	2	0	0	0
j	Golf course	2	0	1	4	0	0	0	0	0	89	0	0	0	0	0	0	2	1	1	0	0
k	Harbor	0	0	0	0	0	0	0	0	0	0	98	0	0	1	0	0	0	1	0	0	0
l	Intersection	0	0	0	0	0	0	0	0	5	0	0	86	2	0	3	1	2	0	0	0	1
m	Medium residential	0	0	0	0	0	0	6	0	0	0	0	1	87	3	0	1	1	0	0	0	1
n	Mobile home park	0	0	0	0	0	0	0	0	0	0	3	1	5	88	0	3	0	0	0	0	0
o	Overpass	0	0	0	1	2	0	0	0	3	0	0	5	0	0	86	0	1	1	0	0	1
p	Parking lot	0	0	0	0	0	2	0	0	1	0	1	1	1	0	0	94	0	0	0	0	0
q	River	2	0	0	0	0	0	0	6	2	4	0	0	0	0	0	0	82	4	0	0	0
r	Runway	0	0	0	2	0	0	0	0	3	1	0	0	0	0	0	0	0	94	0	0	0
s	Sparse residential	0	0	0	0	0	0	1	1	0	8	0	0	3	0	0	0	5	0	78	3	1
t	Storage tanks	0	4	2	1	8	0	0	0	0	3	1	5	0	1	0	0	0	0	10	64	1
u	Tennis court	0	0	3	1	3	0	1	3	1	0	0	1	5	0	1	1	5	0	2	1	72

Table 7.4. Confusion matrix with arch. #5 for a five-fold cross-validation sample result.

		Predicted class label																				
		a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u
a	Agricultural	98	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
b	Airplane	0	95	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	2	0	0	0
c	Baseball diamond	1	0	94	1	0	0	0	0	0	2	0	0	0	0	0	0	0	0	1	0	1
d	Beach	1	0	0	99	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
e	Buildings	0	0	1	0	80	0	4	0	0	0	0	0	0	2	0	0	1	0	2	10	0
f	Chaparral	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
g	Dense residential	0	0	0	0	1	0	81	0	0	0	0	3	12	1	0	0	0	0	1	1	0
h	Forest	0	0	0	0	0	0	0	97	0	0	0	0	0	0	0	0	2	0	1	0	0
i	Freeway	0	0	0	1	0	0	0	0	97	0	0	0	0	0	2	0	0	0	0	0	0
j	Golf course	1	0	1	0	0	0	0	0	0	95	0	0	0	0	0	0	0	0	3	0	0
k	Harbor	0	1	0	0	0	0	0	0	0	0	99	0	0	0	0	0	0	0	0	0	0
l	Intersection	0	0	0	0	2	0	3	0	2	0	0	87	1	1	2	0	1	0	0	1	0
m	Medium residential	0	0	0	0	0	0	5	1	0	0	0	1	86	1	0	0	1	0	4	1	0
n	Mobile home park	0	0	0	0	0	0	0	0	0	0	0	0	3	96	0	1	0	0	0	0	0
o	Overpass	0	0	0	0	1	0	0	0	6	0	0	1	0	0	91	0	0	0	0	1	0
p	Parking lot	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	98	0	0	0	0	0
q	River	0	0	0	2	0	0	0	2	3	1	0	1	0	0	0	0	91	0	0	0	0
r	Runway	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	99	0	0	0
s	Sparse residential	0	0	0	0	1	0	1	2	0	0	0	0	3	0	0	0	1	0	92	0	0
t	Storage tanks	0	0	1	0	4	0	1	1	0	0	0	4	1	0	0	0	2	0	3	83	0
u	Tennis court	0	0	0	0	2	0	1	0	0	0	0	2	4	0	1	0	1	0	3	3	83

Table 7.5. Confusion matrix with arch. #3 for a five-fold cross-validation sample result.

		Predicted class label																				
		a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u
a	Agricultural	94	0	0	2	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	1	0
b	Airplane	0	95	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0
c	Baseball diamond	0	1	94	0	0	0	0	0	0	4	0	0	0	0	0	0	1	0	0	0	0
d	Beach	5	0	0	94	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
e	Buildings	0	0	0	1	70	0	7	0	1	0	0	1	1	1	7	0	0	0	2	5	4
f	Chaparral	0	0	0	0	0	99	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
g	Dense residential	0	0	0	0	2	0	80	1	0	0	0	3	8	3	0	0	0	0	2	0	1
h	Forest	1	0	0	0	0	1	0	97	0	1	0	0	0	0	0	0	0	0	0	0	0
i	Freeway	0	0	0	1	0	0	0	0	97	0	0	2	0	0	0	0	0	0	0	0	0
j	Golf course	2	0	2	0	0	0	0	0	0	95	0	0	0	0	0	0	1	0	0	0	0
k	Harbor	0	1	0	0	0	0	0	0	0	0	99	0	0	0	0	0	0	0	0	0	0
l	Intersection	0	0	1	1	0	0	1	0	6	1	0	85	1	0	1	1	0	0	1	1	0
m	Medium residential	0	0	0	0	1	0	13	1	0	0	0	1	80	0	0	0	0	0	3	1	0
n	Mobile home park	0	0	0	0	0	0	0	0	0	0	0	1	4	94	0	1	0	0	0	0	0
o	Overpass	0	0	0	0	1	0	0	0	13	0	0	4	0	0	80	0	0	2	0	0	0
p	Parking lot	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	98	0	0	0	0	0
q	River	0	0	0	1	0	0	0	3	2	2	0	0	0	0	0	0	89	1	1	1	0
r	Runway	0	0	0	1	0	0	0	0	2	0	1	0	0	0	0	0	1	94	0	1	0
s	Sparse residential	0	0	1	0	0	0	1	2	0	2	0	0	4	0	0	0	0	0	87	0	3
t	Storage tanks	0	0	1	2	5	0	1	1	1	0	1	5	0	0	1	0	2	1	7	71	1
u	Tennis court	0	0	3	1	1	0	3	1	0	0	0	3	1	0	1	0	3	0	3	1	79

Table 7.6. Confusion matrix with arch. #4 for a five-fold cross-validation sample result.

		Predicted class label																				
		a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u
a	Agricultural	95	0	2	2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
b	Airplane	0	96	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0
c	Baseball diamond	0	0	92	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	2	1	1
d	Beach	1	0	0	99	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
e	Buildings	0	0	0	0	78	0	5	0	0	0	0	2	2	0	2	0	1	0	0	5	5
f	Chaparral	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
g	Dense residential	0	0	0	0	2	0	85	0	0	0	0	1	10	1	0	1	0	0	0	0	0
h	Forest	0	0	0	0	0	1	0	98	0	0	0	0	0	0	0	0	1	0	0	0	0
i	Freeway	0	0	0	0	0	0	0	0	96	0	0	3	0	0	1	0	0	0	0	0	0
j	Golf course	2	0	3	0	0	0	0	0	0	94	0	0	0	0	0	0	1	0	0	0	0
k	Harbor	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0
l	Intersection	0	0	0	0	2	0	0	0	2	0	0	93	1	0	1	0	1	0	0	0	0
m	Medium residential	0	0	0	0	0	0	9	0	0	0	0	1	82	1	0	0	1	0	4	1	1
n	Mobile home park	0	0	0	0	2	0	1	0	0	0	1	1	2	93	0	0	0	0	0	0	0
o	Overpass	0	0	0	0	0	0	0	0	8	0	0	3	0	0	87	0	0	2	0	0	0
p	Parking lot	0	0	0	0	1	0	0	0	0	0	0	1	0	1	0	97	0	0	0	0	0
q	River	0	0	0	3	0	0	0	5	2	1	0	0	0	0	0	0	86	2	0	1	0
r	Runway	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	99	0	0	0
s	Sparse residential	0	0	1	0	0	1	0	5	0	5	0	0	3	0	0	0	1	0	83	0	1
t	Storage tanks	0	0	1	0	5	0	0	0	0	0	0	3	3	0	0	0	0	0	8	79	1
u	Tennis court	0	0	1	1	0	0	2	0	0	1	0	0	3	0	0	0	2	0	3	3	84

Table 7.7. UC Merced accuracy comparison.

Date	Method	Accuracy (%)
2010	SPM [52]	74.00
2010	SPCK++ [52]	76.05
2015	Saliency-UFL [234]	82.72±1.18
2014	Bag-of-SIFT [235]	85.37±1.56
	Single-view Deep Learning	**88.00±2.88**
2014	SAL-LDA [236]	88.33
2015	Pyramid of Spatial Relatons [237]	89.1
2014	UFL [235]	90.26±1.51
	Multiview Deep Learning	**93.48±0.82**
2014	VLAT [238]	94.3

views in the case of smaller objects or features that occur at different scales.

7.7.6 Accuracy comparison

A five-fold stratified cross-validation comparison of all the important methods employed in the literature for the UC Merced dataset is shown in Table 7.7. The highest accuracies for the UC Merced dataset have been achieved with unsupervised feature learning (UFL) [235] and the vector of locally aggregated tensors (VLAT) method [238], which is an extension of visual dictionary approaches such as bag-of-words. Single-view DCNN is outperformed by these methods, but the 90.26% accuracy of UFL can be improved upon with a multiview DCNN, which achieves 93.48%. Multiview deep learning quite significantly improves over single-view given a two-tailed p-val of 0.01 using Welch's unpaired t-test, with a mean difference 95% confidence interval of 1.8 to 9.2.

The VLAT implementation of Negrel et al. [238] uses multispectral features in conjunction with a histogram of oriented gradient descriptors at four different scales. Descriptor clustering covariance matrices are used and PCA orthogonalization of the descriptor vector space forms highly discriminative aggregated descriptor tensors, which outperforms a singular DCNN that does not incorporate the benefit of clustering covariance.

7.7.7 Implementation details

Dataset augmenting through rotation did not improve performance in conjunction with horizontal and vertical flips, so it was omitted. Using a secondary neural network to combine the softmax arrays of multiple views for a given sample did not improve performance over model averaging. For the 10-view DCNN instantiation #5 (Table 7.2) a running time of 36.6 seconds per epoch was attained on an Amazon Elastic Compute Cloud g2.2xlarge instance with a GRID K520 GPU possessing 1536 CUDA cores and 4 GB video memory, of which 1 GB was used. A Python implementation was used based on Theano and Lasagne libraries [245], which provides a GPU-accelerated computational differentiation platform that automatically computes gradients for complex systems.

7.8 CONCLUSION

Dataset shift is normally defined in terms of two groups of samples where there is a distinct change in measurement mode between the two sample groups. The grouping itself needs to be provided as part of the problem definition and measurement metadata, or the grouping can be discovered with unsupervised or semi-supervised methods. The agglomerative clustering performed previously is an example of how the grouping can be discovered with unsupervised methods. Another type of dataset shift was demonstrated in this chapter, where there are no clearly defined groupings and thus no clearly defined shift to correct, yet there are many changes in measurement modes. The problem with this dataset shift was demonstrated for low-level features, specifically multispectral GLCM features.

An end-to-end learning system with hierarchical feature representation was designed in this chapter for complex land-use classification of high-resolution multispectral aerial imagery. DCNN architecture hyper-parameters were optimized in terms of cross-validation accuracy on the UC Merced land-use dataset, and it was shown that multiscale views can be used to train a single network and increase classification accuracy compared to using single-view samples. A specific comparison was made between the single-scale multiview strategy of Krizhevsky et al. [49] and the multiscale multiview strategy proposed in this chapter as a generalization of the single-scale DCNN input method.

The main hypothesis is that multiscale views allow for a greater range of object scales to be detected by the DCNN, and thus give a higher likelihood of class recognition. A confusion analysis was performed to ascertain the qualitative differences between five-view single-scale and multiscale input strategies, and it was observed that large structures such as overpasses and freeways can become more confused with multiscale views, although classes with typically smaller features such as storage tanks can benefit from multiscale views. In general, competitive performance was shown where multiview DCNN outperformed both SIFT-based methods and unsupervised feature learning on the UC Merced dataset.

CHAPTER 8 CONCLUSION

8.1 RESEARCH OVERVIEW

The study involves land-use classification for multimodal remote sensing where the challenge was to improve classification accuracy by addressing multimodal dataset shifts. Such dataset shifts occur when acquisition differences between the train and test datasets cause class features and class definitions to differ between the two datasets, rendering a trained classifier inaccurate. Dataset shift reduction is employed at the input-level of feature extraction and at the classifier layer.

The first strategy for reducing dataset shift involves isolating and removing an image component that differs consistently between the train and test datasets. A predominant isolatable component is shadowing, which was dealt with through a shadow removal system consisting of shadow detection and correction or feature masking. The ability of this input modification to increase classification accuracy was demonstrated using multiple different texture features and shadow removal methods.

The alternative approach of manifold alignment was identified with the intention of aligning corresponding train and test classes or directly transferring train labels via manifold matching after test class separation through unsupervised classification. An important contribution in this study is the investigation of the effect of weighting features for weighted clustering and weighted cardinality determination, including weighted generalization of internal validation indices. This unsupervised clustering investigation is the primary means of manifold reduction, which precedes manifold matching in order to reduce its computational time.

A third strategy namely, multiscale feature learning, is investigated for addressing dataset shift in land-use classification of remotely sensed images. This strategy takes into account an approximation of the full range of expected dataset shift in order to learn optimal feature representations that are robust for similar dataset shifts. The research overviews and important contributions of each specific chapter are given in the section below; these are then followed by suggestions for future research.

Appropriate baseline or control strategies and methods are used throughout to emphasize how the proposed methods are set apart, although there are not necessarily always an appropriate comparison of full dataset shift solutions to counterparts in the literature. Internal decomposition of methods are used together with extensive quantitative and qualitative analyses to contribute to validation of the techniques. The focus is thus on proposing and comparing functional techniques more at a component-level within the setting of remote sensing classification. Emphasis in the conclusion is placed on literature methods and how the proposed methods can introduce improvements.

8.2 SHADOW DETECTION

Shadow detection options are explored as part of instantiating shadow removal for the input modification strategy to dataset shift. The hypothesis that threshold-based shadow detection can relatively accurately delineate shadows because of the low-intensity property of shadows was investigated, as well as the hypothesis that local adaptive thresholding can produce more accurate shadows than global thresholding, since relatively low intensity admits greater sensitivity in images with contrast variation than globally low intensity.

Panchromatic shadow detection algorithms from the thresholding subcategory (Adeline et al. [59], Table 2.2) of property-based shadow detection (Arévalo et al. [73]) are used on the Soweto panchromatic land-use dataset. Select thresholding algorithms from the taxonomy of Sezgin and Sankur [47] are also compared for shadow detection accuracy in terms of Czekanowski-Dice (F-score), Jaccard, Rand (overall), Rogers-Tanimoto and Sokal-Sneath external validation indices (paragraph 2.4.4.2). The comparative analyses of shadow detection methods introduce thresholders not previously used for shadow detection.

The minimum Bayes error is difficult to detect with minimum error thresholding because of extensive overlap of shadow and sunlit densities, so it achieves the lowest unsupervised global thresholding accuracy. Iterative minimum error thresholding avoids boundary minima and focuses on a relevant intensity range so it achieves the highest shadow detection accuracies of the thresholding methods considered. Convex hull thresholding can robustly determine the threshold valley, as no explicit bimodality is required, and this thresholding method attains the second highest accuracy for unsupervised global threshold detection.

Wellner's local adaptive thresholding is used for more accurate shadow detection and it is shown that the local window size parameter is robust despite multitemporal shadow profile differences. The potential shadow detection accuracy of global thresholding was compared to that of local adaptive thresholding and for both dates of the Soweto dataset local adaptive thresholding outperformed global thresholding. Global thresholding produced more false positives in the shadow mask, but local adaptive thresholding can take local intensity into account to reduce this and produce more accurate maps.

8.3 INPUT MODIFICATION

An isolatable illumination and viewing geometry component of dataset shift was identified, namely shadowing, and a shadow removal strategy was implemented to reduce effective dataset shift through input modification. Panchromatic shadow removal was achieved through a system consisting of a pixel-based shadow detector and a shadow removal stage featuring either modified feature extraction with shadow pixel masking or shadow correction to produce a shadow-corrected input. A variant of histogram equalization, proposed by Shu and Freeman [75] and used by Sarabandi et al. [112], from the intensity domain (paragraph 2.5.2.1) of the shadow restoration taxonomy in Table 2.4 has been employed for shadow correction. Input modification addresses causal components of dataset shift particular to given data, so it constitutes a unique instantiation that is not compared to alternative strategies in the literature at a system level.

GLCM shadow correction and LBP shadow masking have improved settlement classification accuracy in same-date experiments, and both GLCM and LBP shadow correction and shadow masking can improve settlement classification accuracy in across-date experiments. The most statistically significant improvements in settlement classification accuracy were seen for GLCM across-date LAT masking, GLCM across-date global threshold masking, GLCM across-date fine shadow correction, LBP same-date global threshold masking, LBP across-date masking and LBP across-date global threshold masking.

A confusion analysis revealed that the largest reduction in confusion was between formal settlement and formal settlements with backyard shack classes, probably because these classes typically involve the largest structures and thus the largest shadow profiles. Correlation between settlement type classification accuracy and shadow detection accuracy showed statistically significant differences between same-date and across-date experiments for both GLCM and LBP with fine shadow correction and global threshold masking.

Top-down masking was used as a control test to obtain further evidence that land-use classification accuracy improvements are related to shadows in particular, which was seen in the results where top-down masking could not improve classification accuracy at all, whereas shadow masking could. These results support the theory that it is the shadow removal specifically that improves classification accuracy, and that while increases in same-date accuracies were witnessed, the main benefit lies in across-date classification situations.

8.4 WEIGHTED AGGLOMERATIVE CLUSTERING

The feature spaces generated from acquisitions of areas with complex land-use cases display poor separability because of samples exhibiting multiclass or unclassified traits. The notion that the separation of such a feature space can be obtained by weighting samples according to importance as defined by a target classification was explored in this chapter. The target property of texture regularity is exploited to create sample weights and scale-selective feature space composition is performed based on maximizing the component-wise saliences. This manifold reduction study focuses on the components of hierarchical agglomerative clustering and contrasting the benefit of sample weighting through the evaluation of how the proposed methods are set apart.

For the 10-date Rio de Janeiro multispectral dataset it was shown that weighted clustering with Ward linkage achieves greater mean clustering accuracy. Confusion analysis presented evidence that weighted clustering produces more salient clusters and differentiates better between certain groups of classes. An important demonstration shows how target properties can be derived to augment features with weight information, and how those weights can be used to improve unsupervised classification.

Random unweighted clustering formed a baseline unsupervised classification accuracy score of 53.9 for the multimodal Rio de Janeiro dataset. Hierarchical agglomerative clustering with an unweighted Ward linkage improved to a score of 90.6, and weighted Ward linkages achieved 95.4. The optimal number of clusters is searched for with internal validation index ensembles in truncated form where sample weights are used to improve performance. It is shown that sample weightings are best used for maximum weight selection of the truncated internal index input. An important contribution is weighted input-truncated internal validation indices designed to admit larger datasets through reduced computational complexity. The weighted generalization of a number of internal validation indices for cardinality determination has also been made in this work, including improved interpretations of these indices, to expand the weighted generalization defined by Studer [40] to those in the comprehensive compendium collected by Desgraupes [41].

Weighted internal validation indices were used for weakly supervised clustering cardinality determination, and input truncated implementations were used to reduce computational complexity for large datasets. Sample weighting was used to good effect and it was experimentally illustrated that the main contributor to improved internal index performance was maximal weight input selection. Unsupervised cardinality determination was implemented and an overall objective score of 83.77 was reached out of a maximum objective of 93.92 for the multidate Rio de Janeiro dataset.

Maximum weight input selection clearly appears to produce better cardinalities than random input sampling, whereas the use of weights with weighted internal indices has a minimal effect overall. In the comparison with unweighted random sampling an increase in sample size on average produces a greater discrepancy between weighted maximum input selection. However, for the comparisons with weighted random sampling and unweighted maximum input selection there is on average no distinct comparative change with increases in sample size.

Improved internal index extremum and disruption interpretations were proposed and results indicated performance improvements for the majority of internal indices. Knee-point accentuating filtering before extremum interpretation improved accuracy for 16 indices, where the more significant improvements are seen for Ball-Hall, Trace_W, S_Dbw and G+. The disruption interpreted indices do not really benefit from the knee-point accentuating filtering, since this filtering is designed to improve extremum interpretation. The alternative $\frac{d}{dk}\arctan\left(\frac{d\overline{c}}{dk}\right)$ disruption interpretation improves accuracy for 13 indices in the case of weighted Ward linkage clustering.

8.5 GLCM MANIFOLD MATCHING WITH GEOMETRIC SIMILARITY MEASURES

Larger dataset shifts due to both multitemporal and multisensor acquisition differences were addressed through partial manifold alignment consisting first of a manifold reduction stage covered in the previous chapter's unsupervised learning and then a manifold matching stage that works under the assumption of manifold preservation. Manifold matching is tested separately with the objective of minimizing correspondence cost through the Munkres algorithm based on a cost matrix that features geometric similarity, basic divergence and relative variance similarity.

The geometric similarity formulation of Wang et al. [43] is improved by incorporating the co-occurrence frequency information generated during optimal neighborhood permutation searches, which is normally discarded. The geometric similarity of Wang et al. is used in the isolated context of manifold matching, but it is integrated into multi-component cost matrices. The focus is on this integration and on improving information usage by geometric similarity, but comparisons at the system-level have no inclusion of further methods in the literature outside of the manifold alignment context.

Minimum-supervision manifold matching is contributed for bijective correspondence problems, with a manifold reduction prerequisite that presents class statistical moments as input. The matching is based on a cost function that depends on basic divergence, relative variance similarity, as well as geometric similarity. Geometric similarity calculations generate co-occurrence information that is normally discarded, but this information has been used to improve the geometric similarity cost function.

A matching cost function based only on basic divergence achieved an overall matching accuracy of 62.8% for the multimodal Johannesburg dataset, and with only variance similarity an accuracy of 64.7% was reached. Using textbook geometric similarity with divergence and variance similarity an accuracy of 85.9% was achieved, and a further improvement to 89.5% was possible when local geometry matching co-occurrence was incorporated into the geometric similarity cost.

8.6 MULTIVIEW DEEP LEARNING

If the full extent of dataset shift can be observed in a dataset for all classes, a classifier can be obtained that is accurate for test examples falling within the observed dataset shift range. Feature learning is an alternative approach investigated in this thesis, which overcomes the limitation of engineered features such as GLCM and LBP. Engineered features are probably sub-optimal since they have to be applicable in a wide range of problems.

Feature learning can discover the optimal features that minimize the classification objective loss function. The performance of feature learning in producing features and a classifier that can accurately classify remotely sensed land-use images is evaluated and compared against several methods for the UC Merced land-use dataset. Convolutional neural networks are used and optimized for the given problem and a CNN usage technique is designed to significantly improve classification accuracy.

An end-to-end learning system with hierarchical feature representation was designed for complex land-use classification of high-resolution multispectral aerial imagery. DCNN architecture hyper-parameters were optimized heuristically in terms of cross-validation accuracy on the UC Merced land-use dataset, and it was shown that multiscale views can be used to train a single network and increase classification accuracy compared to using single-view samples. A specific comparison was made between the single-scale multiview strategy of Krizhevsky et al. [49] and the multiscale multiview strategy proposed as a generalization of the single-scale DCNN input method.

The primary hypothesis is that multiscale views allows for a greater range of object scales to be detected by the DCNN, and thus gives a higher likelihood of class recognition. A confusion analysis was performed to ascertain the qualitative differences between five-view single-scale and multiscale input strategies, and it was observed that large structures such as overpasses and freeways can be confused more easily with multiscale views, although classes with typically smaller characteristics such as storage tanks can benefit from multiscale views. In general, competitive performance was shown where multiview DCNN outperformed both SIFT-based methods and unsupervised feature learning on the UC Merced dataset.

8.7 FUTURE RESEARCH

8.7.1 Weighted mean shift clustering

Unsupervised classification that incorporates sample weightings could alternatively be accomplished through density-based mean shift clustering, which is adapted to use the weightings. The weight adaptation could be achieved by factoring in the weightings into the adaptive bandwidth of mean shift to promote greater kernel sizes for more salient features. A potential problem that would need to be overcome is the issue of large variation in feature space density, since this may result in an uneven and unprioritised partitioning if the bandwidth adaptation is not sufficient. A further issue is that mean shift tends to find a natural partitioning cardinality, at least for well-behaved problem spaces, but in practice further steps need to be taken to ensure proper prioritization and refining to a partitioning of desired cardinality.

8.7.2 Non-bijective manifold matching

Bijective classification problems were the focus of this study, since non-bijective problems incur numerous added difficulties. The bijection assumption allows for solutions to be ventured upon, such as the use of an affine fit, but if this assumption cannot be made then options quickly become limited and confidence levels of solutions are reduced. Graph-based manifold matching methods could be applied to non-bijective problems where there is a definite overlap in train and test classes but not perfect correspondence. The challenge is in estimating the confidence level of the solution and basing the extent of the subsequent manifold alignment on that information or integrating the information into the manifold alignment framework.

8.7.3 Generalized eigenvalue decomposition manifold alignment

Manifold alignment based on the Rayleigh quotient framework involves an extra processing stage of generalized eigenvalue decomposition that is used after manifold matching to actually correct the dataset shift. Since classification was the only objective in this study, the solution only required manifold matching to transfer train labels to the test clusters. The opportunity exists to investigate whether actual dataset shift correction and normal supervised classification afterward can improve on simpler label transfer.

REFERENCES

[1] C. M. Bishop, *Neural Networks for Pattern Recognition*. New York, NY, USA: Oxford University Press, Inc., 1995.

[2] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," *Pattern Recognition*, vol. 45, no. 1, pp. 521–530, 2012.

[3] C. R. Shelton, "Importance sampling for reinforcement learning with multiple objectives," Ph.D. dissertation, Cambridge, MA, USA, 2001.

[4] S. Bickel and T. Scheffer, "Dirichlet-enhanced spam filtering based on biased samples," *Advances in Neural Information Processing Systems*, vol. 19, p. 161, 2007.

[5] S. Parsons, "Bioinformatics: The machine learning approach," *Knowl. Eng. Rev.*, vol. 19, no. 1, pp. 90–91, Mar. 2004. [Online]. Available: http://dx.doi.org/10.1017/S0269888904220161

[6] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain–computer interfaces for communication and control," *Clinical Neurophysiology*, vol. 113, no. 6, pp. 767–791, 2002.

[7] J. J. Heckman, "Sample selection bias as a specification error," *Econometrica: Journal of the Econometric Society*, pp. 153–161, 1979.

[8] J. G. Moreno-Torres, *Dataset shift in classification: Terminology, benchmarks and methods*. Editorial de la Universidad de Granada, 2013.

[9] A. Storkey, "When training and test sets are different: Characterizing learning transfer," *Dataset Shift in Machine Learning*, pp. 3–28, 2009.

[10] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset Shift in Machine Learning*. The MIT Press, 2009.

[11] K. Wang, S. Zhou, A. W.-C. Fu, J. X. Yu, F. Jeffrey, and X. Yu, "Mining changes of classification by correspondence tracing." in *SDM*. SIAM, 2003, pp. 95–106.

[12] Y. Yang, X. Wu, and X. Zhu, "Conceptual equivalence for contrast mining in classification learning," *Data & Knowledge Engineering*, vol. 67, no. 3, pp. 413–429, 2008.

[13] D. A. Cieslak and N. V. Chawla, “A framework for monitoring classifiers performance: When and why failure occurs?” *Knowledge and Information Systems*, vol. 18, no. 1, pp. 83–108, 2009.

[14] S. Bickel, M. Brückner, and T. Scheffer, “Discriminative learning under covariate shift,” *The Journal of Machine Learning Research*, vol. 10, pp. 2137–2155, 2009.

[15] M. Sugiyama and K.-R. Müller, “Input-dependent estimation of generalization error under covariate shift,” *Statistics & Decisions*, vol. 23, no. 4/2005, pp. 249–279, 2005.

[16] N. Japkowicz and S. Stephen, “The class imbalance problem: A systematic study,” *Intelligent Data Analysis*, vol. 6, no. 5, pp. 429–449, 2002.

[17] M. Sugiyama, S. Nakajima, H. Kashima, P. V. Buenau, and M. Kawanabe, “Direct importance estimation with model selection and its application to covariate shift adaptation,” in *Advances in Neural Information Processing Systems*, 2008, pp. 1433–1440.

[18] H. Shimodaira, “Improving predictive inference under covariate shift by weighting the log-likelihood function,” *Journal of Statistical Planning and Inference*, vol. 90, no. 2, pp. 227–244, 2000.

[19] M. Sugiyama, M. Krauledat, and K.-R. Müller, “Covariate shift adaptation by importance weighted cross validation,” *The Journal of Machine Learning Research*, vol. 8, pp. 985–1005, 2007.

[20] J. Huang, A. Gretton, K. M. Borgwardt, B. Schölkopf, and A. J. Smola, “Correcting sample selection bias by unlabeled data,” in *Advances in Neural Information Processing Systems*, 2006, pp. 601–608.

[21] R. Alaiz-Rodríguez, A. Guerrero-Curieses, and J. Cid-Sueiro, “Class and subclass probability re-estimation to adapt a classifier in the presence of concept drift,” *Neurocomputing*, vol. 74, no. 16, pp. 2614–2623, 2011.

[22] J. G. Moreno-Torres, X. Llorà, D. E. Goldberg, and R. Bhargava, “Repairing fractures between data using genetic programming-based feature extraction: A case study in cancer diagnosis,” *Information Sciences*, vol. 222, pp. 805–823, 2013.

[23] R. Klinkenberg, “Learning drifting concepts: Example selection vs. example weighting,” *Intelligent Data Analysis*, vol. 8, no. 3, pp. 281–300, 2004.

[24] N. M. Adams and D. J. Hand, “Comparing classifiers when the misallocation costs are uncertain,” *Pattern Recognition*, vol. 32, no. 7, pp. 1139–1147, 1999.

[25] F. Provost and T. Fawcett, “Robust classification for imprecise environments,” *Machine Learning*, vol. 42, no. 3, pp. 203–231, 2001.

[26] R. Alaiz-Rodríguez, A. Guerrero-Curieses, and J. Cid-Sueiro, “Minimax regret classifier for

imprecise class distributions," *The Journal of Machine Learning Research*, vol. 8, pp. 103–130, 2007.

[27] M. Kubat, R. C. Holte, and S. Matwin, "Machine learning for the detection of oil spills in satellite radar images," *Machine Learning*, vol. 30, no. 2-3, pp. 195–215, 1998.

[28] M. G. Kelly, D. J. Hand, and N. M. Adams, "The impact of changing populations on classifier performance," in *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1999, pp. 367–371.

[29] M. Saerens, P. Latinne, and C. Decaestecker, "Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure," *Neural Computation*, vol. 14, no. 1, pp. 21–41, 2002.

[30] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[31] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[32] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[33] D. L. Donoho and C. Grimes, "Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data," *Proceedings of the National Academy of Sciences*, vol. 100, no. 10, pp. 5591–5596, 2003.

[34] Z.-y. Zhang and H.-y. Zha, "Principal manifolds and nonlinear dimensionality reduction via tangent space alignment," *Journal of Shanghai University (English Edition)*, vol. 8, no. 4, pp. 406–424, 2004.

[35] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, no. 1, pp. 1–27, 1964.

[36] K. Bahirat, F. Bovolo, L. Bruzzone, and S. Chaudhuri, "A novel domain adaptation Bayesian classifier for updating land-cover maps with class differences in source and target domains," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 7, pp. 2810–2826, Jul. 2012.

[37] L. Gómez-Chova, G. Camps-Valls, L. Bruzzone, and J. Calpe-Maravilla, "Mean map kernel methods for semisupervised cloud classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 1, pp. 207–220, 2010.

[38] L. Bruzzone and M. Marconcini, "Toward the automatic updating of land-cover maps by a domain-adaptation SVM classifier and a circular validation strategy," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 4, pp. 1108 –1122, April 2009.

[39] F. P. S. Luus, F. van den Bergh, and B. T. J. Maharaj, "The effects of segmentation-based shadow removal on across-date settlement type classification of panchromatic QuickBird images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 6, no. 3, pp. 1274–1285, 2013.

[40] M. Studer, "WeightedCluster library manual: A practical guide to creating typologies of trajectories in the social sciences with R," LIVES Working Paper 24. Switzerland: NCCR LIVES, Tech. Rep., 2013.

[41] B. Desgraupes, "Clustering indices," *University Paris Quest Lab Modal'X*, Apr. 2013.

[42] D. Tuia, M. Volpi, M. Trolliet, and G. Camps-Valls, "Semisupervised manifold alignment of multimodal remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 12, pp. 7708–7720, Dec 2014.

[43] C. Wang and S. Mahadevan, "Manifold alignment without correspondence." in *Proc. 21st Int. Joint Conf. Artif. Intell.*, vol. 2, 2009, p. 3.

[44] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[45] D. H. Hubel and T. N. Wiesel, "Receptive fields of single neurones in the cat's striate cortex," *The Journal of Physiology*, vol. 148, no. 3, pp. 574–591, 1959.

[46] D. Ciresan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, 2012, pp. 3642–3649.

[47] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *Journal of Electronic Imaging*, vol. 13, no. 1, pp. 146–168, 2004.

[48] F. Luus, F. van den Bergh, and B. Maharaj, "Mean translation of GLCM texture features for across-date settlement type classification of QuickBird images," in *Proceedings of the 2013 IEEE Geoscience and Remote Sensing Symposium*, July 2013, pp. 1529–1532.

[49] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[50] X. Chen, S. Xiang, C.-L. Liu, and C.-H. Pan, "Vehicle detection in satellite images by hybrid deep convolutional neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 10, pp. 1797–1801, Oct 2014.

[51] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *arXiv preprint, arXiv:1409.4842*, 2014.

[52] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification,"

in *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 2010, pp. 270–279.

[53] S. K. Selvaraj, B. Bhar, S. Sellamanickam, and S. Shevade, "Semi-supervised SVMs for classification with unknown class proportions and a small labeled dataset," in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*. ACM, 2011, pp. 653–662.

[54] D. Chakraborty and U. Maulik, "Semisupervised pixel classification of remote sensing imagery using transductive SVM," in *Proceedings of the 2011 International Conference on Recent Trends in Information Systems (ReTIS)*, Dec. 2011, pp. 30–35.

[55] L. Bruzzone, C. Mingmin, and M. Marconcini, "A novel transductive SVM for semisupervised classification of remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 11, pp. 3363–3373, Nov. 2006.

[56] R. G. Negri, S. J. S. Sant'Anna, and L. V. Dutra, "Semi-supervised remote sensing image classification methods assessment," in *Proceedings of the 2011 IEEE Geoscience and Remote Sensing Symposium*. IEEE, 2011, pp. 2939–2942.

[57] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *The Journal of Machine Learning Research*, vol. 7, pp. 2399–2434, 2006.

[58] D. Cai and X. He, "Manifold adaptive experimental design for text categorization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 4, pp. 707–719, Apr. 2012.

[59] K. Adeline, M. Chen, X. Briottet, S. Pang, and N. Paparoditis, "Shadow detection in very high spatial resolution aerial images: A comparative study," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 80, pp. 21–38, 2013.

[60] P. M. Dare, "Shadow analysis in high-resolution satellite imagery of urban areas," *Photogrammetric Engineering & Remote Sensing*, vol. 71, no. 2, pp. 169–177, 2005.

[61] A. Shahtahmassebi, N. Yang, K. Wang, N. Moore, and Z. Shen, "Review of shadow detection and de-shadowing methods in remote sensing," *Chinese Geographical Science*, vol. 23, no. 4, pp. 403–420, 2013.

[62] Y.-T. Liow and T. Pavlidis, "Use of shadows for extracting buildings in aerial images," *Computer Vision, Graphics, and Image Processing*, vol. 49, no. 2, pp. 242–277, 1990.

[63] F. Cheng and K.-H. Thiel, "Delimiting the building heights in a city from the shadow in a panchromatic SPOT image: Part 1. Test of forty-two buildings," *Remote Sensing*, vol. 16, no. 3, pp. 409–415, 1995.

[64] G. Hai-tao, Z. Yan, L. Jun, and J. G.-w. Zhengzhou, "Research on the building shadow extraction and elimination method," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 37.

[65] A. Saha, M. Arora, E. Csaplovics, and R. Gupta, "Land cover classification using IRS LISS III image and DEM in a rugged terrain: A case study in Himalayas," *Geocarto International*, vol. 20, no. 2, pp. 33–40, 2005.

[66] W. Liu and F. Yamazaki, "Object-based shadow extraction and correction of high-resolution optical satellite images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. PP, no. 99, pp. 1–7, 2012.

[67] W. Bin, K. Muramatsu, and N. Fujiwara, "Automated detection and removal of clouds and their shadows from Landsat images," *IEICE Transactions on Information and Systems*, vol. 82, no. 2, pp. 453–460, 1999.

[68] T. Nakajima, G. Tao, and Y. Yasuoka, "Simulated recovery of information in shadow areas on IKONOS image by combing ALS data," in *Proc. Asian Conference on Remote Sensing*, 2002.

[69] S. Lachérade, C. Miesch, D. Boldo, X. Briottet, C. Valorge, and H. Le Men, "ICARE: A physically-based model to correct atmospheric and geometric effects from high spatial and spectral remote sensing images over 3D urban areas," *Meteorology and Atmospheric Physics*, vol. 102, no. 3-4, pp. 209–222, 2008.

[70] V. Tsai, "A comparative study on shadow compensation of color aerial images in invariant color models," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 6, pp. 1661–1671, June 2006.

[71] A. M. Polidorio, F. C. Flores, N. N. Imai, A. M. Tommaselli, and C. Franco, "Automatic shadow segmentation in aerial color images," in *Proceedings of the XVI Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI)*. IEEE, 2003, pp. 270–277.

[72] J. Huang, W. Xie, and L. Tang, "Detection of and compensation for shadows in colored urban aerial images," in *Proceedings of the Fifth World Congress on Intelligent Control and Automation (WCICA 2004)*, vol. 4. IEEE, 2004, pp. 3098–3100.

[73] V. Arévalo, J. González, and G. Ambrosio, "Shadow detection in colour high-resolution satellite images," *International Journal of Remote Sensing*, vol. 29, no. 7, pp. 1945–1963, 2008.

[74] M. P. Bishop, J. F. Shroder, and J. D. Colby, "Remote sensing and geomorphometry for studying relief production in high mountains," *Geomorphology*, vol. 55, no. 1, pp. 345–361, 2003.

[75] J. S.-P. Shu and H. Freeman, "Cloud shadow removal from aerial photographs," *Pattern Recognition*, vol. 23, no. 6, pp. 647–656, 1990.

[76] Y. Wei, Z. Zhao, and J. Song, "Urban building extraction from high-resolution satellite panchromatic image using clustering and edge detection," in *Proceedings of the 2004 IEEE Geoscience and Remote Sensing Symposium*, vol. 3, 2004, pp. 2008–2010.

[77] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 285-296, pp. 23–27, 1975.

[78] Y. Chen, D. Wen, L. Jing, and P. Shi, "Shadow information recovery in urban areas from very high resolution satellite imagery," *International Journal of Remote Sensing*, vol. 28, no. 15, pp. 3249–3254, 2007.

[79] F. Yamazaki, W. Liu, and M. Takasaki, "Characteristics of shadow and removal of its effects for remote sensing imagery," in *Proceedings of the 2009 IEEE Geoscience and Remote Sensing Symposium*, vol. 4. IEEE, 2009, pp. IV–426.

[80] K.-L. Chung, Y.-R. Lin, and Y.-H. Huang, "Efficient shadow detection of color aerial images based on successive thresholding scheme," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 2, pp. 671–682, 2009.

[81] M. Teke, E. Başeski, A. Ö. Ok, B. Yüksel, and Ç. Şenaras, "Multi-spectral false color shadow detection," in *Photogrammetric Image Analysis*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2011, vol. 6952, pp. 109–119.

[82] C. Fredembach and S. Süsstrunk, "Automatic and accurate shadow detection from (potentially) a single image using near-infrared information," EPFL Tech Report 165527, Tech. Rep., 2010.

[83] M. Nagao, T. Matsuyama, and Y. Ikeda, "Region extraction and shape analysis in aerial photographs," *Computer Graphics and Image Processing*, vol. 10, no. 3, pp. 195–223, 1979.

[84] S. Wang and Y. Wang, "Shadow detection and compensation in high resolution satellite image based on retinex," in *Proceedings of the Fifth International Conference on Image and Graphics (ICIG '09)*, Sept. 2009, pp. 209–212.

[85] Ö. Aytekın, A. Erener, İ. Ulusoy, and Ş. Düzgün, "Unsupervised building detection in complex urban environments from multispectral satellite imagery," *International Journal of Remote Sensing*, vol. 33, no. 7, pp. 2152–2177, 2012.

[86] W. Zhou, G. Huang, A. Troy, and M. Cadenasso, "Object-based land cover classification of shaded areas in high spatial resolution imagery of urban areas: A comparison study," *Remote Sensing of Environment*, vol. 113, no. 8, pp. 1769–1777, 2009.

[87] X. Huang and L. Zhang, "Morphological building/shadow index for building extraction from high-resolution imagery over urban areas," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 1, pp. 161–172, Feb. 2012.

[88] M. D. Levine and J. Bhattacharyya, “Removing shadows,” *Pattern Recognition Letters*, vol. 26, no. 3, pp. 251–265, 2005.

[89] W. Huang, Y. Xiao, and S. Lu, “Shadow detection of the high-resolution remote sensing image based on pulse coupled neural network,” in *Proceedings of the Seventh International Symposium on Multispectral Image Processing and Pattern Recognition (MIPPR2011).* International Society for Optics and Photonics, 2011, pp. 80 060Q–80 060Q.

[90] M. F. Tappen, W. T. Freeman, and E. H. Adelson, “Recovering intrinsic images from a single image,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 9, pp. 1459–1472, 2005.

[91] L. Lorenzi, F. Melgani, and G. Mercier, “A complete processing chain for shadow detection and reconstruction in VHR images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 9, pp. 3440–3452, Sept. 2012.

[92] E. A. Ashton, B. D. Wemett, R. A. Leathers, and T. V. Downes, “A novel method for illumination suppression in hyperspectral images,” in *SPIE Defense and Security Symposium.* International Society for Optics and Photonics, 2008, pp. 69 660C–69 660C.

[93] N. Martel-Brisson and A. Zaccarin, “Moving cast shadow detection from a Gaussian mixture shadow model,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2. IEEE, 2005, pp. 643–648.

[94] F. P. S. Luus, F. van den Bergh, and B. T. J. Maharaj, “The effects of shadow removal on across-date settlement type classification of QuickBird images,” in *Proceedings of the 2012 IEEE Geoscience and Remote Sensing Symposium*, July 2012.

[95] Y. Chen, W. Su, J. Li, and Z. Sun, “Hierarchical object oriented classification using very high resolution imagery and LIDAR data over urban areas,” *Advances in Space Research*, vol. 43, no. 7, pp. 1101–1110, 2009.

[96] A. Massalabi and D.-C. He, “Restitution of information under shadow in remote sensing high space resolution images: Application to IKONOS data of Sherbrooke city,” in *ISPRS, Commission, Istanbul.* Citeseer, 2004.

[97] Q. Zhan, W. Shi, and Y. Xiao, “Quantitative analysis of shadow effects in high-resolution images of urban areas,” *International Archives of Photogrammetry and Remote Sensing*, vol. 36, no. 8/W27, 2005.

[98] K. Navulur, *Multispectral Image Analysis Using the Object-Oriented Paradigm.* Boca Raton, FL, USA: CRC Press, Inc., 2006.

[99] H. Li, L. Zhang, and H. Shen, “An adaptive nonlocal regularized shadow removal method for aerial remote sensing images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52,

no. 1, pp. 106–120, 2014.

[100] J.-P. Thirion, "Realistic 3D simulation of shapes and shadows for image processing," *CVGIP: Graphical Models and Image Processing*, vol. 54, no. 1, pp. 82–90, 1992.

[101] G. Tolt, M. Shimoni, and J. Ahlberg, "A shadow detection method for remote sensing images using VHR hyperspectral and LIDAR data," in *Proceedings of the 2011 IEEE Geoscience and Remote Sensing Symposium*, July 2011, pp. 4423–4426.

[102] X. Wu, S. Collings, and P. Caccetta, "BRDF and illumination calibration for very high resolution imaging sensors," in *Proceedings of the 2010 IEEE Geoscience and Remote Sensing Symposium*. IEEE, 2010, pp. 3162–3165.

[103] J. Boardman, "Automated spectral unmixing of AVIRIS data using convex geometry concepts: In: Annual JPL Airborne Geosciences Workshop, 4, Pasadena, CA," *Summaries, JPL Publication*, pp. 93–26, 1993.

[104] J. D. Colby, "Topographic normalization in rugged terrain," *Photogrammetric Engineering and Remote Sensing*, vol. 57, no. 5, pp. 531–537, 1991.

[105] A. Makarau, R. Richter, R. Muller, and P. Reinartz, "Adaptive shadow detection using a blackbody radiator model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 6, pp. 2049–2059, June 2011.

[106] C. Fraser, E. Baltsavias, and A. Gruen, "Processing of IKONOS imagery for submetre 3D positioning and building extraction," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 56, no. 3, pp. 177–194, 2002.

[107] J. Rau, N.-Y. Chen, and L.-C. Chen, "True orthophoto generation of built-up areas using multi-view images," *Photogrammetric Engineering and Remote Sensing*, vol. 68, no. 6, pp. 581–588, 2002.

[108] F. Li, D. L. B. Jupp, and M. Thankappan, "Using high resolution DSM data to correct the terrain illumination effect in Landsat data," in *Proceedings of the 19th International Congress on Modelling and Simulation - Sustaining Our Future: Understanding and Living with Uncertainty (MODSIM)*, 2011, pp. 2402–2408.

[109] C. J. V. Rijsbergen, *Information Retrieval*, 2nd ed. Newton, MA, USA: Butterworth-Heinemann, 1979.

[110] R. G. Congalton, "A review of assessing the accuracy of classifications of remotely sensed data," *Remote Sensing of Environment*, vol. 37, no. 1, pp. 35–46, 1991.

[111] Q. J. Wang, Q. J. Tian, Q. Z. Lin, M. X. Li, and L. M. Wang, "An improved algorithm for shadow restoration of high spatial resolution imagery," in *Proceedings of the Remote Sensing*

of the Environment: 16th National Symposium on Remote Sensing of China. International Society for Optics and Photonics, 2008, pp. 71 230D–71 230D.

[112] P. Sarabandi, F. Yamazaki, M. Matsuoka, and A. Kiremidjian, "Shadow detection and radiometric restoration in satellite high resolution images." in *Proceedings of the 2004 IEEE Geoscience and Remote Sensing Symposium*, 2004, pp. 3744–3747.

[113] G. Nolè, M. Danese, B. Murgante, R. Lasaponara, and A. Lanorte, "Using spatial autocorrelation techniques and multi-temporal satellite data for analyzing urban sprawl," in *Computational Science and its Applications (ICCSA)*. Springer, 2012, pp. 512–527.

[114] X. Zhu, D. Liu, and J. Chen, "A new geostatistical approach for filling gaps in Landsat ETM+ SLC-off images," *Remote Sensing of Environment*, vol. 124, pp. 49–60, 2012.

[115] G. D. Finlayson, M. S. Drew, and C. Lu, "Entropy minimization for shadow removal," *International Journal of Computer Vision*, vol. 85, no. 1, pp. 35–57, 2009.

[116] E. Arbel and H. Hel-Or, "Shadow removal using intensity surfaces and texture anchor points," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 6, pp. 1202–1216, 2011.

[117] T.-P. Wu and C.-K. Tang, "A Bayesian approach for shadow extraction from a single image," in *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV)*, vol. 1. IEEE, 2005, pp. 480–487.

[118] L. Xu, F. Qi, and R. Jiang, "Shadow removal from a single image," in *Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications (ISDA'06)*, vol. 2. IEEE, 2006, pp. 1049–1054.

[119] V. Shettigara and G. Sumerling, "Height determination of extended objects using shadows in SPOT images," *Photogrammetric Engineering and Remote Sensing*, vol. 64, no. 1, pp. 35–43, 1998.

[120] D. A. Forsyth and J. Ponce, "A modern approach," *Computer Vision: A Modern Approach*, 2003.

[121] B. Peng, L. Zhang, and D. Zhang, "A survey of graph theoretical approaches to image segmentation," *Pattern Recognition*, vol. 46, no. 3, pp. 1020–1038, 2013.

[122] M. Wertheimer, "Laws of organization in perceptual forms." 1938.

[123] R. M. Haralick and L. G. Shapiro, "Image segmentation techniques," in *Proceedings of the 1985 Technical Symposium East*. International Society for Optics and Photonics, 1985, pp. 2–9.

[124] J. C. Tilton, "Image segmentation by iterative parallel region growing and splitting," 1989.

[125] Y.-J. Zhang, "An overview of image and video segmentation in the last 40 years," *Advances in Image and Video Segmentation*, pp. 1–15, 2006.

[126] R. C. Gonzalez and R. E. Woods, *Digital image processing*. Prentice Hall Upper Saddle River, NJ, 2002.

[127] A. Rosenfeld, "Image pattern recognition," in *IEEE Proceedings*, vol. 69, 1981, pp. 596–605.

[128] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004.

[129] J. Wassenberg, W. Middelmann, and P. Sanders, "An efficient parallel algorithm for graph-based image segmentation," in *Computer Analysis of Images and Patterns*. Springer, 2009, pp. 1003–1010.

[130] N. Copty, S. Ranka, G. Fox, and R. V. Shankar, "A data parallel algorithm for solving the region growing problem on the connection machine," *Journal of Parallel and Distributed Computing*, vol. 21, no. 1, pp. 160–168, 1994.

[131] R. Urquhart, "Graph theoretical clustering based on limited neighbourhood sets," *Pattern Recognition*, vol. 15, no. 3, pp. 173–187, 1982.

[132] J. Weszka and A. Rosenfeld, "Histogram modification for threshold selection," *NASA STI/Recon Technical Report N*, vol. 78, p. 15466, 1977.

[133] N. Ray and B. N. Saha, "Edge sensitive variational image thresholding," in *Proceedings of the 2007 IEEE International Conference on Image Processing (ICIP)*, vol. 6. IEEE, 2007, pp. VI–37.

[134] A. Rosenfeld and P. De La Torre, "Histogram concavity analysis as an aid in threshold selection," *IEEE Transactions on Systems, Man and Cybernetics*, no. 2, pp. 231–235, 1983.

[135] M. I. Sezan, "A peak detection algorithm and its application to histogram-based image data reduction," *Computer Vision, Graphics, and Image Processing*, vol. 49, no. 1, pp. 36–51, 1990.

[136] N. Ramesh, J.-H. Yoo, and I. Sethi, "Thresholding based on histogram approximation," *IEE Proceedings-Vision, Image and Signal Processing*, vol. 142, no. 5, pp. 271–279, 1995.

[137] J. Prewitt and M. L. Mendelsohn, "The analysis of cell images," *Annals of the New York Academy of Sciences*, vol. 128, no. 3, pp. 1035–1053, 1966.

[138] T. Ridler and S. Calvard, "Picture thresholding using an iterative selection method," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 8, no. 8, pp. 630–632, 1978.

[139] J. Kittler and J. Illingworth, "Minimum error thresholding," *Pattern Recognition*, vol. 19, no. 1, pp. 41–47, 1986.

[140] C. Jawahar, P. K. Biswas, and A. Ray, “Investigations on fuzzy thresholding based on fuzzy clustering,” *Pattern Recognition*, vol. 30, no. 10, pp. 1605–1613, 1997.

[141] J. N. Kapur, P. K. Sahoo, and A. K. Wong, “A new method for gray-level picture thresholding using the entropy of the histogram,” *Computer Vision, Graphics, and Image Processing*, vol. 29, no. 3, pp. 273–285, 1985.

[142] C. H. Li and C. Lee, “Minimum cross entropy thresholding,” *Pattern Recognition*, vol. 26, no. 4, pp. 617–625, 1993.

[143] A. G. Shanbhag, “Utilization of information measure as a means of image thresholding,” *CVGIP: Graphical Models and Image Processing*, vol. 56, no. 5, pp. 414–419, 1994.

[144] W.-H. Tsai, “Moment-preserving thresholding: A new approach,” in *Document Image Analysis*. IEEE Computer Society Press, 1995, pp. 44–60.

[145] L. Hertz and R. W. Schafer, “Multilevel thresholding using edge matching,” *Computer Vision, Graphics, and Image Processing*, vol. 44, no. 3, pp. 279–295, 1988.

[146] L.-K. Huang and M.-J. J. Wang, “Image thresholding by minimizing the measures of fuzziness,” *Pattern Recognition*, vol. 28, no. 1, pp. 41–51, 1995.

[147] A. Pikaz and A. Averbuch, “Digital image thresholding, based on topological stable-state,” *Pattern Recognition*, vol. 29, no. 5, pp. 829–843, 1996.

[148] C.-K. Leung and F. Lam, “Maximum segmented image information thresholding,” *Graphical Models and Image Processing*, vol. 60, no. 1, pp. 57–76, 1998.

[149] S. K. Pal and A. Rosenfeld, “Image enhancement and thresholding by optimization of fuzzy compactness,” *Pattern Recognition Letters*, vol. 7, no. 2, pp. 77–86, 1988.

[150] N. R. Pal and S. K. Pal, “Entropic thresholding,” *Signal Processing*, vol. 16, no. 2, pp. 97–108, 1989.

[151] A. S. Abutaleb, “Automatic thresholding of gray-level pictures using two-dimensional entropy,” *Computer Vision, Graphics, and Image Processing*, vol. 47, no. 1, pp. 22–32, 1989.

[152] N. Friel and I. S. Molchanov, “A new thresholding technique based on random sets,” *Pattern Recognition*, vol. 32, no. 9, pp. 1507–1517, 1999.

[153] H.-D. Cheng and Y.-H. Chen, “Fuzzy partition of two-dimensional histogram and its application to thresholding,” *Pattern recognition*, vol. 32, no. 5, pp. 825–843, 1999.

[154] W. Niblack, *An introduction to digital image processing*. Strandberg Publishing Company, 1985.

[155] J. Sauvola and M. Pietikäinen, “Adaptive document image binarization,” *Pattern Recognition*, vol. 33, no. 2, pp. 225–236, 2000.

[156] P. D. Wellner, “Adaptive thresholding for the DigitalDesk,” *Xerox, EPC1993-110*, 1993.

[157] J. M. White and G. D. Rohrer, “Image thresholding for optical character recognition and other applications requiring character image extraction,” *IBM Journal of Research and Development*, vol. 27, no. 4, pp. 400–411, 1983.

[158] J. Bernsen, “Dynamic thresholding of grey-level images,” in *Proceedings of the International Conference on Pattern Recognition*, 1986, pp. 1251–1255.

[159] M. Kamel and A. Zhao, “Extraction of binary character/graphics images from grayscale document images,” *CVGIP: Graphical Models and Image Processing*, vol. 55, no. 3, pp. 203–217, 1993.

[160] S. D. Yanowitz and A. M. Bruckstein, “A new method for image segmentation,” in *Proceedings of the 9th International Conference on Pattern Recognition.* IEEE, 1988, pp. 270–275.

[161] D. Rutovitz, “An algorithm for in-line generation of a convex cover,” *Computer Graphics and Image Processing*, vol. 4, no. 1, pp. 74–78, 1975.

[162] G. Borgefors, “Distance transformations in digital images,” *Computer Vision, Graphics, and Image Processing*, vol. 34, no. 3, pp. 344–371, 1986.

[163] E. Sales, W. Gomez, and W. C. Pereira, “Evaluation performance of local adaptive binarization algorithms for trabecular bone on simulated μct,” in *Proceedings of the 2011 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC).* IEEE, 2011, pp. 3084–3087.

[164] F. van den Bergh, “The effects of viewing- and illumination geometry on settlement type classification of QuickBird images,” in *Proceedings of the 2011 IEEE Geoscience and Remote Sensing Symposium*, Jul. 2011, pp. 1425–1428.

[165] J. A. Voogt and T. R. Oke, “Effects of urban surface geometry on remotely-sensed surface temperature,” *International Journal of Remote Sensing*, vol. 19, no. 5, pp. 895–920, 1998.

[166] R. M. Haralick, K. Shanmugam, and I. Dinstein, “Textural features for image classification,” *IEEE Transactions on Systems, Man and Cybernetics*, vol. 3, no. 6, pp. 610–621, Nov. 1973.

[167] T. Ojala, M. Pietikäinen, and T. Mäenpää, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, July 2002.

[168] Z. Zhang and F. Chen, “A shadow processing method of high spatial resolution remote sensing image,” in *Proceedings of the 3rd International Congress on Image and Signal Processing (CISP)*, vol. 2, Oct. 2010, pp. 816–820.

[169] X. Meng, R. Rosenthal, and D. Rubin, “Comparing correlated correlation coefficients,” *Psychological Bulletin*, vol. 111, no. 1, pp. 172–175, January 1994.

[170] M. Ackerman and S. Dasgupta, “Incremental clustering: The case for extra clusters,” in *Advances in Neural Information Processing Systems*, 2014, pp. 307–315.

[171] M. Ackerman, S. Ben-David, S. Branzei, and D. Loker, “Weighted clustering.” in *AAAI*, 2012, pp. 858–863.

[172] J. W. Richards, J. Hardin, and E. B. Grosfils, “Weighted model-based clustering for remote sensing image analysis,” *Computational Geosciences*, vol. 14, no. 1, pp. 125–136, 2010.

[173] K. Xiao, S. H. Ho, and A. Bargiela, “Automatic brain MRI segmentation scheme based on feature weighting factors selection on fuzzy c-means clustering algorithms with Gaussian smoothing,” *International Journal of Computational Intelligence in Bioinformatics and Systems Biology*, vol. 1, no. 3, pp. 316–331, 2010.

[174] L. Wang, T. Mei, and Q. Qin, “A region-based high spatial resolution remotely sensed imagery classification algorithm based on multiscale fusion and feature weighting,” in *Proceedings of the Sixth International Symposium on Multispectral Image Processing and Pattern Recognition*. International Society for Optics and Photonics, 2009, pp. 749–411.

[175] W. Wang, Z. Zhao, and H. Zhu, “Object-oriented change detection method based on multi-scale and multi-feature fusion,” in *Proceedings of the 2009 Joint Urban Remote Sensing Event*. IEEE, 2009, pp. 1–5.

[176] J. A. Dos Santos, P.-H. Gosselin, S. Philipp-Foliguet, R. d. S. Torres, and A. X. Falao, “Multiscale classification of remote sensing images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 10, pp. 3764–3775, 2012.

[177] R. Sliuzas, M. Kuffer, and I. Masser, “The spatial and temporal nature of urban objects,” in *Remote Sensing of Urban and Suburban Areas*. Springer, 2010, pp. 67–84.

[178] J. Gao, *Digital analysis of remotely sensed imagery*. McGraw-Hill Professional, 2008.

[179] Y. Wu, Y. Ke, H. Gong, B. Chen, and L. Zhu, “Comparison of object-based and pixel-based methods for urban land-use classification from WorldView-2 imagery,” in *Proceedings of the 3rd International Workshop on Earth Observation and Remote Sensing Applications (EORSA)*. IEEE, 2014, pp. 284–288.

[180] D. C. Duro, S. E. Franklin, and M. G. Dubé, “A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using spot-5 hrg imagery,” *Remote Sensing of Environment*, vol. 118, pp. 259–272, 2012.

[181] M. M. Nielsen, “Remote sensing for urban planning and management: The use of window-independent context segmentation to extract urban features in Stockholm,” *Computers, Environment and Urban Systems*, vol. 52, pp. 1–9, 2015.

[182] F. Pacifici, M. Chini, and W. J. Emery, “A neural network approach using multi-scale textural metrics from very high-resolution panchromatic imagery for urban land-use classification,” *Remote Sensing of Environment*, vol. 113, no. 6, pp. 1276–1292, 2009.

[183] M. Kim and M. Madden, “Determination of optimal scale parameter for alliance-level forest classification of multispectral IKONOS images,” *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, Salzburg*, 2006.

[184] D. Lu and Q. Weng, “A survey of image classification methods and techniques for improving classification performance,” *International Journal of Remote Sensing*, vol. 28, no. 5, pp. 823–870, 2007.

[185] M. Li, S. Zang, B. Zhang, S. Li, and C. Wu, “A review of remote sensing image classification techniques: The role of spatio-contextual information,” *European Journal of Remote Sensing*, vol. 47, pp. 389–411, 2014.

[186] L. Huang and L. Ni, “Object-oriented classification of high resolution satellite image for better accuracy,” *Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, pp. 211–218, 2008.

[187] Y. Gao, P. Marpu, I. Niemeyer, D. M. Runfola, N. M. Giner, T. Hamill, and R. G. Pontius Jr, “Object-based classification with features extracted by a semi-automatic feature extraction algorithm–seath,” *Geocarto International*, vol. 26, no. 3, pp. 211–226, 2011.

[188] G. P. Petropoulos, C. Kalaitzidis, and K. P. Vadrevu, “Support vector machines and object-based classification for obtaining land-use/cover cartography from Hyperion hyperspectral imagery,” *Computers & Geosciences*, vol. 41, pp. 99–107, 2012.

[189] R. A. Schowengerdt, *Remote sensing: Models and methods for image processing*. Academic press, 2006.

[190] M. S. Tehrany, B. Pradhan, and M. N. Jebuv, “A comparative assessment between object and pixel-based classification approaches for land use/land cover mapping using SPOT 5 imagery,” *Geocarto International*, vol. 29, no. 4, pp. 351–369, 2014.

[191] M. F. Goodchild, M. Yuan, and T. J. Cova, “Towards a general theory of geographic representation in GIS,” *International Journal of Geographical Information Science*, vol. 21, no. 3, pp. 239–260, 2007.

[192] G. Espindola, G. Câmara, I. Reis, L. Bins, and A. Monteiro, “Parameter selection for region-growing image segmentation algorithms using spatial autocorrelation,” *International Journal of*

Remote Sensing, vol. 27, no. 14, pp. 3035–3040, 2006.

[193] J.-P. Donnay, M. J. Barnsley, and P. A. Longley, *Remote Sensing and Urban Analysis: GISDATA 9*. CRC Press, 2003.

[194] S. Wei, Z. Chao, Y. Jianyu, W. Honggan, C. Minjie, Y. Anzhi, Z. Yingna, and S. Chongli, "Knowledge-based object oriented land cover classification using SPOT5 imagery in forest-agriculture ecotones," *Sensor Letters*, vol. 8, no. 1, pp. 22–31, 2010.

[195] M. Sonka, V. Hlavac, and R. Boyle, *Image processing, analysis, and machine vision*. Cengage Learning, 2014.

[196] J. C. Tilton and W. T. Lawrence, "Interactive analysis of hierarchical image segmentation," in *Proceedings of the 2000 IEEE Geoscience and Remote Sensing Symposium*, vol. 2. IEEE, 2000, pp. 733–735.

[197] C. Witharana and D. L. Civco, "Optimizing multi-resolution segmentation scale using empirical methods: Exploring the sensitivity of the supervised discrepancy measure Euclidean distance 2 (ED2)," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 87, pp. 108–121, 2014.

[198] M. Neubert, H. Herold, and G. Meinel, "Assessing image segmentation quality – Concepts, methods and application," in *Object-Based Image Analysis*. Springer, 2008, pp. 769–784.

[199] M. Kim, T. A. Warner, M. Madden, and D. S. Atkinson, "Multi-scale GEOBIA with very high spatial resolution digital aerial imagery: Scale, texture and image objects," *International Journal of Remote Sensing*, vol. 32, no. 10, pp. 2825–2850, 2011.

[200] H. Zhang, J. E. Fritts, and S. A. Goldman, "Image segmentation evaluation: A survey of unsupervised methods," *Computer Vision and Image Understanding*, vol. 110, no. 2, pp. 260–280, 2008.

[201] J. Liu and Y.-H. Yang, "Multiresolution color image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 7, pp. 689–700, 1994.

[202] M. D. Levine and A. M. Nazif, "Dynamic measurement of computer generated image segmentations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 2, pp. 155–164, 1985.

[203] S. Bhaskaran, S. Paramananda, and M. Ramnarayan, "Per-pixel and object-oriented classification methods for mapping urban features using IKONOS satellite data," *Applied Geography*, vol. 30, no. 4, pp. 650–665, 2010.

[204] F. Murtagh, G. J. Babu, J. G. Campbell, A. Heck, E. Feigelson, C. Fraley, S. Mukherjee, and A. Raftery, "Multivariate data analysis with Fortran, C and Java code," *Northern Ireland: Queens University Belfast, Astronomical Observatory Strasbourg*, p. 272, 2000.

[205] L. Vendramin, R. J. Campello, and E. R. Hruschka, "Relative clustering validity criteria: A comparative overview," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 3, no. 4, pp. 209–235, 2010.

[206] F. B. Baker and L. J. Hubert, "Measuring the power of hierarchical cluster analysis," *Journal of the American Statistical Association*, vol. 70, pp. 31–38, 1975.

[207] G. H. Ball and D. J. Hall, "ISODATA: A novel method of data analysis and pattern classification," *Menlo Park: Stanford Research Institute. (NTIS No. AD 699616)*, 1965.

[208] J. Banfield and A. Raftery, "Model-based Gaussian and non-Gaussian clustering," *Biometrics*, vol. 49, pp. 803–821, 1993.

[209] L. Hubert and J. Schultz, "Quadratic assignment as a general data-analysis strategy," *British Journal of Mathematical and Statistical Psychologie*, vol. 29, pp. 190–241, 1976.

[210] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974.

[211] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979.

[212] A. J. Scott and M. J. Symons, "Clustering methods based on likelihood ratio criteria," *Biometrics*, vol. 27, pp. 387–397, 1971.

[213] F. J. Rohlf, "Methods of comparing classifications," *Annual Review of Ecology and Systematics*, vol. 5, pp. 101–113, 1974.

[214] J. C. Bezdek and N. R. Pal, "Some new indexes of cluster validity," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 28, no. 3, pp. 301–315, 1998.

[215] F. H. B. Marriot, "Practical problems in a method of cluster analysis," *Biometrics*, vol. 27, pp. 456–460, 1975.

[216] J. A. Hartigan, "Clustering algorithms," *New York: Wiley*, 1975.

[217] J. O. McClain and V. R. Rao, "CLUSTISZ: A program to test for the quality of clustering of a set of objects," *Journal of Marketing Research*, vol. 12, pp. 456–460, 1975.

[218] B. S. Pakhira M. K. and M. U., "Validity index for crisp and fuzzy clusters," *Pattern Recognition*, vol. 37, pp. 487–501, 2004.

[219] G. W. Milligan, "A Monte Carlo study of thirty internal criterion measures for cluster analysis," *Psychometrika*, vol. 46, no. 2, pp. 187–199, 1981.

[220] D. A. Ratkowsky and G. N. Lance, "A criterion for determining the number of groups in a

classification," *Australian Computer Journal*, vol. 10, pp. 115–117, 1978.

[221] S. Ray and R. H. Turi, "Determination of number of clusters in k-means clustering and application in colour image segmentation," *4th International Conference on Advances in Pattern Recognition and Digital Techniques*, pp. 137–143, 1999.

[222] M. Halkidi and M. Vazirgiannis, "Clustering validity assessment: Finding the optimal partitioning of a data set," *Proceedings IEEE International Conference on Data Mining*, pp. 187–194, 2001.

[223] P. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[224] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, pp. 81–93, 1938.

[225] A. W. F. Edwards and L. Cavalli-Sforza, "A method for cluster analysis," *Biometrika*, vol. 56, pp. 362–375, 1965.

[226] H. P. Friedman and J. Rubin, "On some invariant criteria for grouping data," *Journal of the American Statistical Association*, vol. 62, no. 320, pp. 1159–1178, 1967.

[227] G. Forestier, P. Gançarski, and C. Wemmert, "Collaborative clustering with background knowledge," *Data & Knowledge Engineering*, vol. 69, no. 2, pp. 211–228, 2010.

[228] G. W. Milligan and M. C. Cooper, "An examination of procedures for determining the number of clusters in a data set," *Psychometrika*, vol. 50, no. 2, pp. 159–179, 1985.

[229] K. Zhang, X. Li, and J. Zhang, "A robust point-matching algorithm for remote sensing image registration," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 2, pp. 469–473, Feb 2014.

[230] J.-P. Jacobs, G. Thoonen, D. Tuia, G. Camps-Valls, B. Haest, and P. Scheunders, "Domain adaptation with Hidden Markov Random Fields," in *Proceedings of the 2013 IEEE Geoscience and Remote Sensing Symposium*, Jul. 2013, pp. 3112–3115.

[231] D. Tuia, J. Munoz-Mari, L. Gomez-Chova, and J. Malo, "Graph matching for adaptation in remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 1, pp. 329–341, Jan 2013.

[232] M. Zhao, B. An, Y. Wu, and C. Lin, "Bi-SOGC: A graph matching approach based on bilateral KNN spatial orders around geometric centers for remote sensing image registration," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 6, pp. 1429–1433, Nov 2013.

[233] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[234] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, pp. 2175–2184, Apr. 2015.

[235] F. Hu, G.-S. Xia, Z. Wang, L. Zhang, and H. Sun, "Unsupervised feature coding on local patch manifold for satellite image scene classification," in *Proceedings of the 2014 IEEE Geoscience and Remote Sensing Symposium*. IEEE, 2014, pp. 1273–1276.

[236] Q. Zhu, Y. Zhong, and L. Zhang, "Multi-feature probability topic scene classifier for high spatial resolution remote sensing imagery," in *Proceedings of the 2014 IEEE Geoscience and Remote Sensing Symposium*. IEEE, 2014, pp. 2854–2857.

[237] S. Chen and Y. Tian, "Pyramid of spatial relatons for scene-level land use classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 4, pp. 1947–1957, April 2015.

[238] R. Negrel, D. Picard, and P.-H. Gosselin, "Evaluation of second-order visual features for land-use classification," in *Proceedings of the 12th International Workshop on Content-Based Multimedia Indexing (CBMI)*. IEEE, 2014, pp. 1–5.

[239] Y. Chen, X. Zhao, and X. Jia, "Spectral-spatial classification of hyperspectral data based on deep belief network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. PP, no. 99, pp. 1–12, 2015.

[240] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 2094–2107, June 2014.

[241] W. Huang, L. Xiao, Z. Wei, H. Liu, and S. Tang, "A new pan-sharpening method with deep neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 5, pp. 1037–1041, May 2015.

[242] J. Tang, C. Deng, G.-B. Huang, and B. Zhao, "Compressed-domain ship detection on spaceborne optical image using deep neural network and extreme learning machine," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 3, pp. 1174–1185, March 2015.

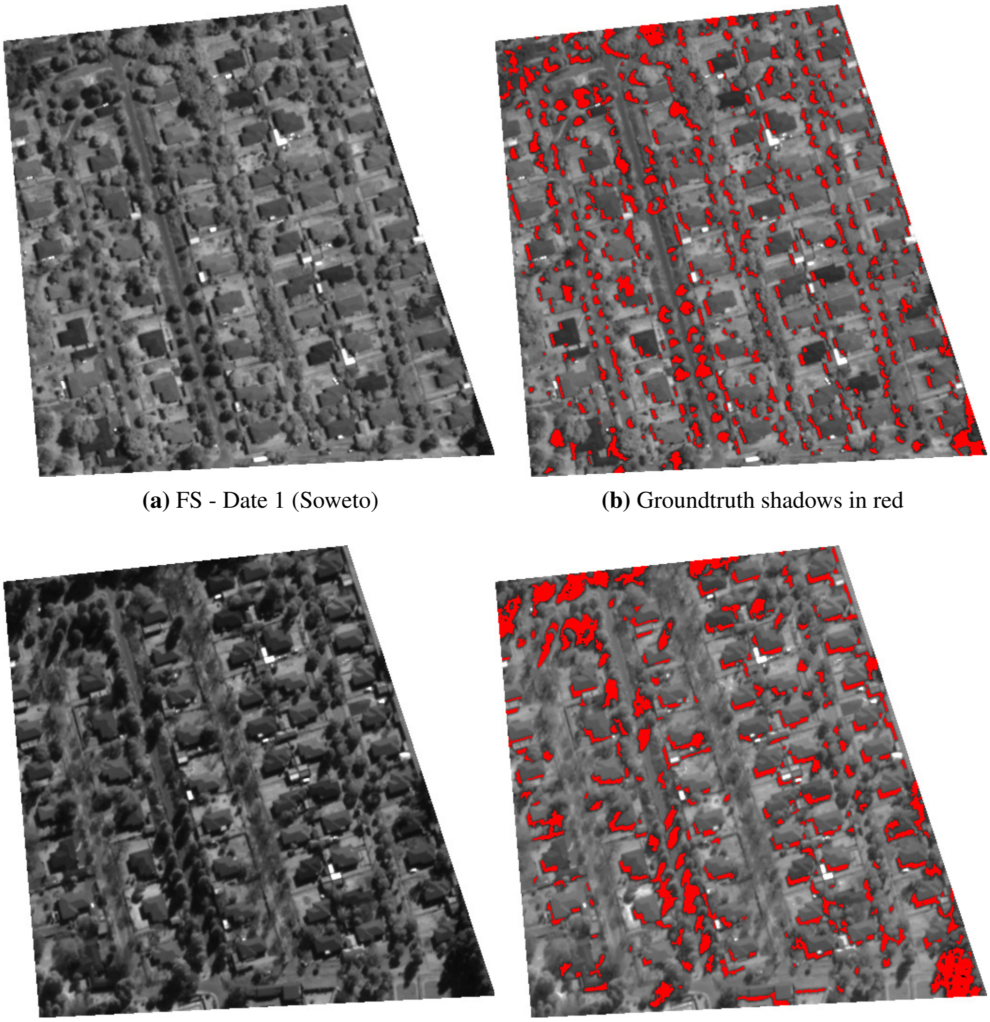
[243] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proceedings of the 30th International Conference on Machine Learning (ICML'13)*, 2013, pp. 1139–1147.

[244] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256.

[245] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. J. Goodfellow, A. Bergeron, N. Bouchard, and Y. Bengio, "Theano: New features and speed improvements," Deep Learning and Unsupervised

Feature Learning NIPS 2012 Workshop, 2012.

APPENDIX A SOWETO DATASET GROUNDTRUTH SHADOWS



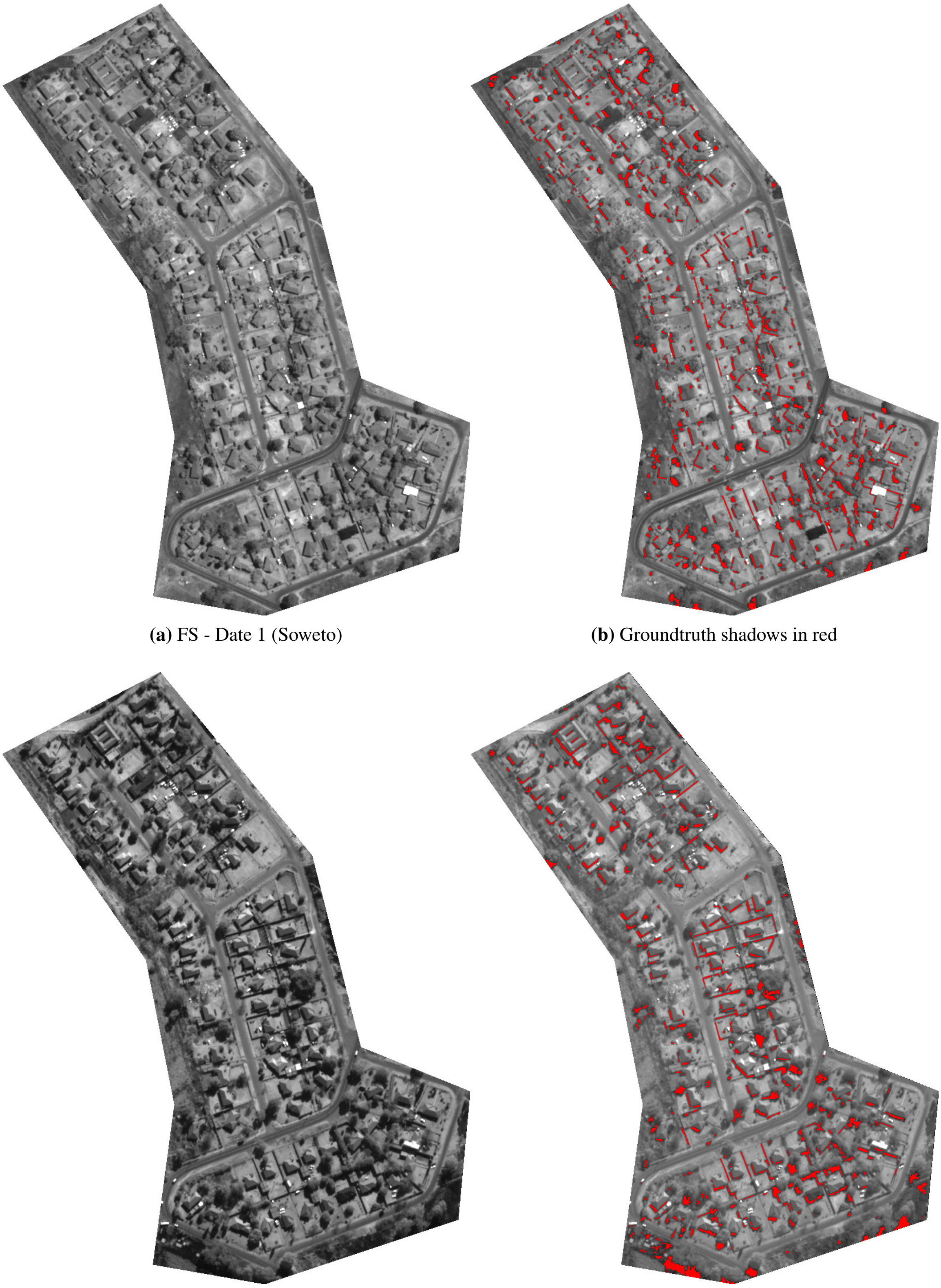
(a) FS - Date 1 (Soweto)

(b) Groundtruth shadows in red

(c) FS - Date 2 (Soweto)

(d) Groundtruth shadows in red

Figure A.1. Groundtruth shadows for a co-registered formal settlement polygon over the two acquisitions of the Soweto dataset. Panchromatic images courtesy of DigitalGlobe™.



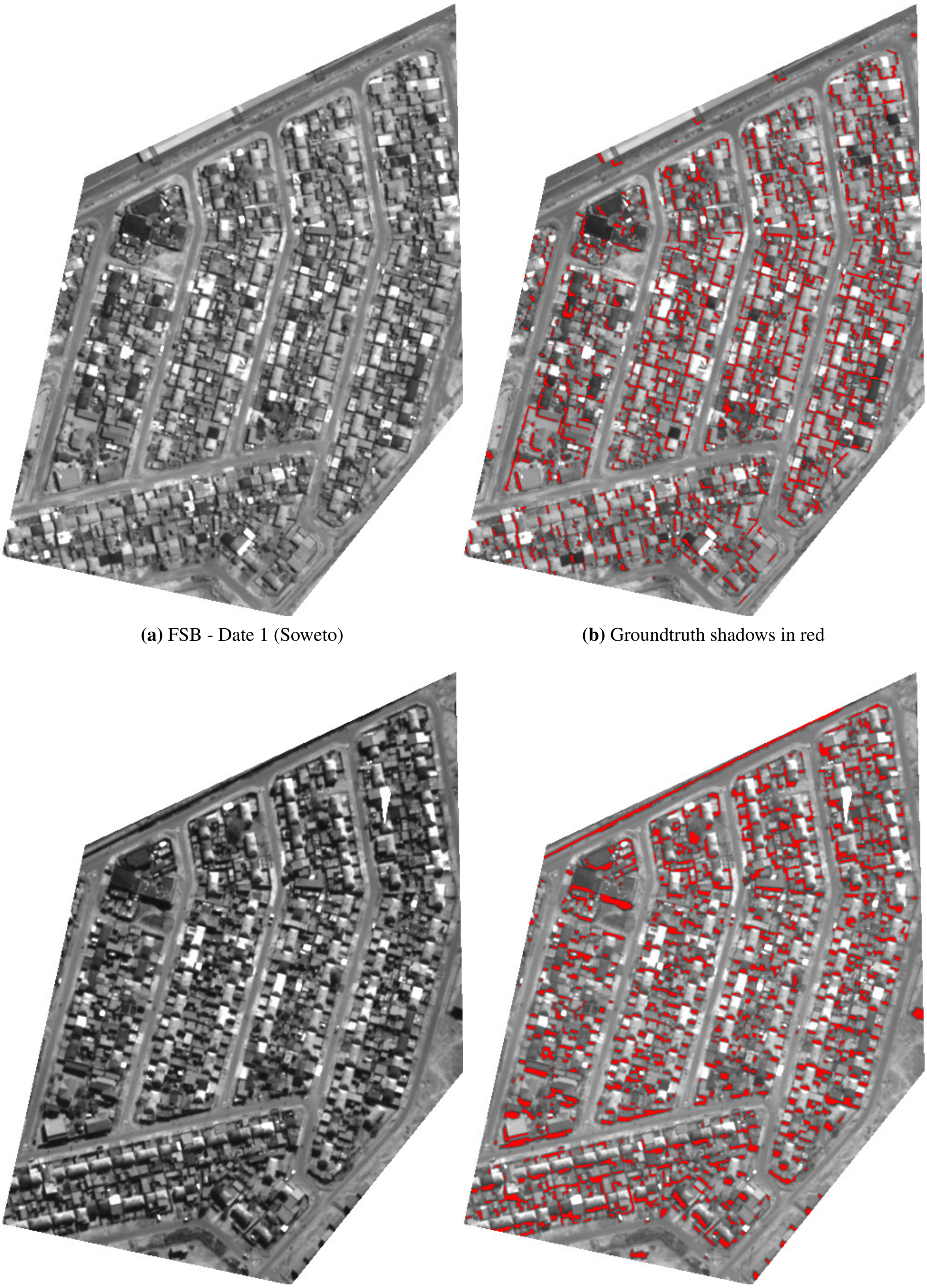
(a) FS - Date 1 (Soweto)

(b) Groundtruth shadows in red

(c) FS - Date 2 (Soweto)

(d) Groundtruth shadows in red

Figure A.2. Groundtruth shadows for a second co-registered formal settlement polygon over the two acquisitions of the Soweto dataset. Panchromatic images courtesy of DigitalGlobe™.



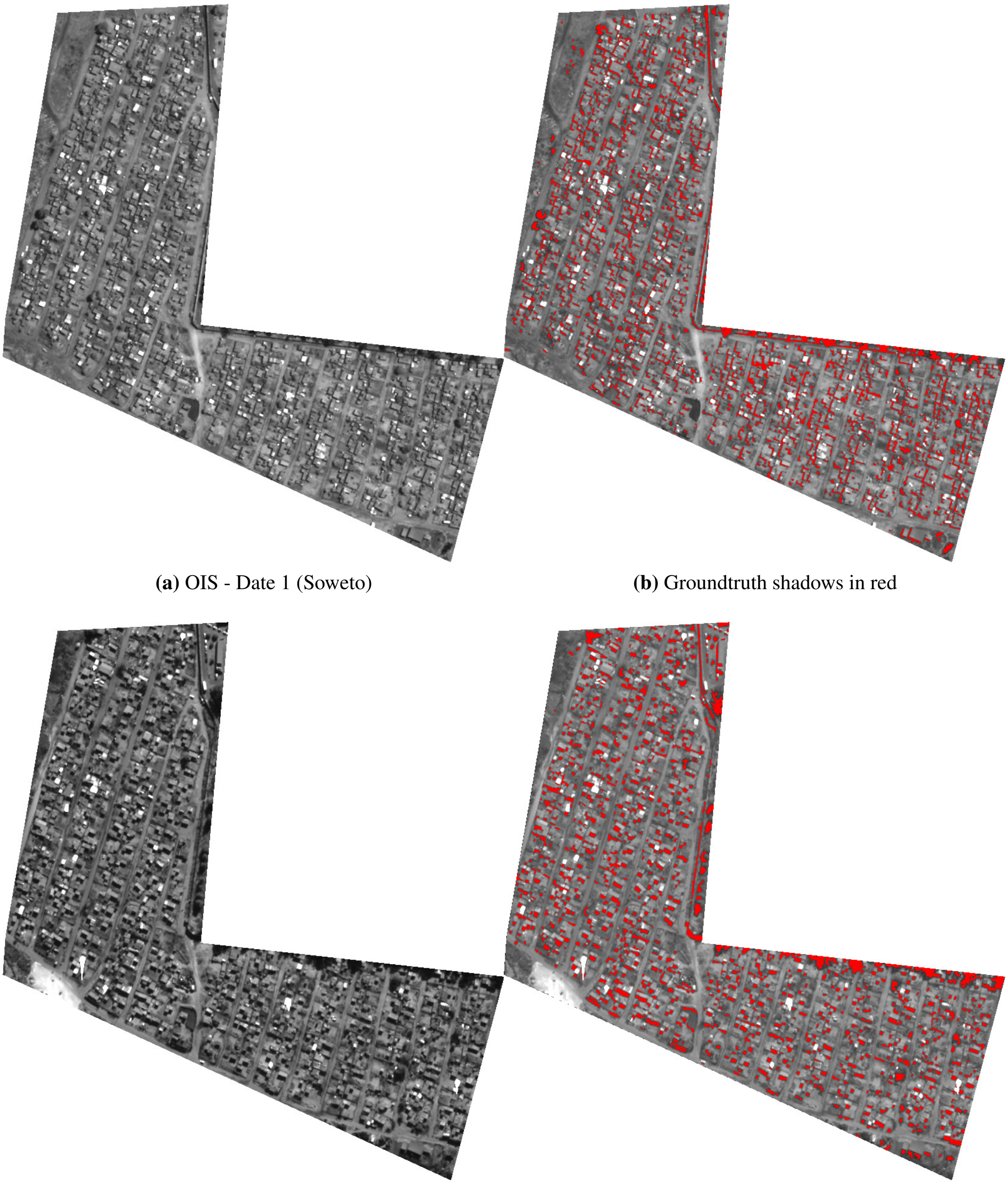
(a) FSB - Date 1 (Soweto)

(b) Groundtruth shadows in red

(c) FSB - Date 2 (Soweto)

(d) Groundtruth shadows in red

Figure A.3. Groundtruth shadows for a co-registered polygon of class formal settlement with backyard shacks over the two acquisitions of the Soweto dataset. Panchromatic images courtesy of DigitalGlobe™.



(a) OIS - Date 1 (Soweto)

(b) Groundtruth shadows in red

(c) OIS - Date 2 (Soweto)

(d) Groundtruth shadows in red

Figure A.4. Groundtruth shadows for a co-registered informal settlement polygon over the two acquisitions of the Soweto dataset. Panchromatic images courtesy of DigitalGlobe™.

APPENDIX B GENERALIZED EIGENVALUE DECOMPOSITION MANIFOLD ALIGNMENT

B.1 OVERVIEW

The concept of manifold alignment plays an important role in dataset shift reduction, since this framework describes the required components and how they interact to make supervised classification more accurate. Manifold alignment that is solved through generalized eigenvalue decomposition of the dual objective Rayleigh quotient formulation is effectively joint manifold learning with dataset shift reduction. This approach is attractive, since it preserves the structural relationships of the manifold and because it practically assures solvability through eigenvalue decomposition.

The generalized eigenvalue decomposition framework for manifold alignment is reviewed [42] in this thesis, since it is important to indicate how the study contributions of manifold reduction and manifold matchings can fit into this attractive solution. Manifold reduction produces an unsupervised classification of the target domain, which is used to form inter- and intra-domain similarity and dissimilarity objectives. Manifold matching establishes the across-domain class relationships, which is also required to formulate the relaxed similarity and dissimilarity constraints.

B.2 BASIC NOTATION

Given M feature domains derived from a multimodal array of image inputs, the associated data matrices are noted as $\mathbf{X}^m \in \mathbb{R}^{d_m \times n_m}$ for domains $m = 1, \dots, M$. The framework reviewed in this section allows for both labeled and unlabeled samples to be aligned, according to the algorithm provided in [42]. A data matrix $\mathbf{X}^m = \{\mathbf{x}_i^m \in \mathbb{R}^{d_m}\}_{i=1}^{n_m}$ can then consist of both unlabeled samples $\{\mathbf{x}_i^m \in \mathbb{R}^{d_m}\}_{i=1}^{u_m}$ as well as input-output labeled sample pairs $\{\mathbf{x}_j^m \in \mathbb{R}^{d_m}, y_j^m \in \mathbb{Z}\}_{j=1}^{l_m}$ for a total of $n_m = l_m + u_m$ samples for domain m.

Manifold alignment attempts to discover for each domain $m = 1, \dots, M$ a projection function $\mathbf{f}^m \in \mathbb{R}^{d_m \times d}$ onto a joint manifold $\mathcal{F}$ with dimensionality $\dim(\mathcal{F}) = d$, where $d = \sum_m d_m$. The number of samples per domain is not necessarily equal ($n_m \neq n_{m'}$) and domain dimensions may differ ($d_m \neq d_{m'}$). Joint matrix formulations are thus required, find the M mapping functions to a common latent space through generalized eigenvalue decomposition, and the joint block diagonal matrix $\mathbf{X} = \mathrm{diag}(\mathbf{X}^1, \dots, \mathbf{X}^M) \in$

$\mathbb{R}^{d\times N}$ can be conceptualised as in Equation B.1, with the total number of samples given by $N = \sum_m n_m$. The joint projection matrix $\mathbf{F} \in \mathbb{R}^{d\times d}$ contains the individual projection functions $\mathbf{f}_i^m \in \mathbb{R}^{d_m}$ for each domain and each new dimension, as given in Equation B.2.

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}^1 & 0 & 0 & 0 \\ 0 & \mathbf{X}^2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \mathbf{X}^M \end{bmatrix} \qquad \mathbf{F} = \begin{bmatrix} \mathbf{f}^1 \\ \mathbf{f}^2 \\ \vdots \\ \mathbf{f}^M \end{bmatrix} = \begin{bmatrix} \mathbf{f}_1^1 & \cdots & \mathbf{f}_d^1 \\ \mathbf{f}_1^2 & \cdots & \mathbf{f}_d^2 \\ \vdots & \vdots & \vdots \\ \mathbf{f}_1^M & \cdots & \mathbf{f}_d^M \end{bmatrix} \tag{B.1, 2}$$

B.3 RAYLEIGH QUOTIENT FORMULATION

The standard Rayleigh quotient or Rayleigh-Ritz ratio in Equation B.3 can be used to accommodate the generalized dual objectives of **A** and **B**, which can produce an optimized joint projection matrix $\mathbf{F}^{\text{opt}}$. The sample relationships that are to be minimized are quantified in **A**, and the relationships to be maximized are encoded into **B**.

$$\mathbf{F}^{\text{opt}} = \underset{\mathbf{F}}{\operatorname{argmin}} \left\{ \operatorname{tr}\left((\mathbf{F}^\top \mathbf{B}\mathbf{F})^{-1} \mathbf{F}^\top \mathbf{A}\mathbf{F} \right) \right\} \tag{B.3}$$

The minimized relationships of $\mathbf{A} = \mu\mathbf{G} + \mathbf{S}$ in this setting involve the minimization of the distances between samples in the aligned common latent space $\mathcal{F}$. These relationships include those between within-domain samples that are geometric neighbours ($\mathbf{G}$ with trade-off constant μ), as well as between inter- and intra-domain samples that belong to the same class ($\mathbf{S}$). The relationships to be maximized, namely $\mathbf{B} = \mathbf{D}$, involve the relative maximization of aligned space distances between inter- and intra-domain samples that belong to different classes ($\mathbf{D}$).

B.4 MANIFOLD GEOMETRY PRESERVATION

In Equation B.3 the minimization objective $\mathbf{A} = \mu\mathbf{G} + \mathbf{S}$ contains $\mathbf{G}$, which aims to preserve the manifold geometry in the transformation, but only the within-domain geometries. Standard graph methods such as KNN can be used to establish neighborhood relationships between within-domain samples, which are encoded in the local geometry matrices $\mathbf{W}_g^m \in \mathbb{R}^{n_m \times n_m}$ for each domain m as follows:

$$\mathbf{W}_g^m(i,j) = \begin{cases} 0, & \text{if } i \text{ and } j \text{ are not graph neighbours.} \\ 1, & \text{if } i \text{ and } j \text{ are graph neighbours.} \end{cases} \tag{B.4}$$

Squared Euclidean distances in the aligned common latent space $\mathcal{F}$ are used between samples, and only those neighboring sample pairs that need their distance relatively minimized are activated in the objective $\mathbf{W}_g^m$. The objective G associated with $\mathbf{G}$ first projects each d_m-dimensional sample $\mathbf{x}_i^m$ onto the d-dimensional latent space $\mathcal{F}$ via $\mathbf{f}^{m\top}\mathbf{x}_i^m$ ($\mathbb{R}^{d\times d_m} \times \mathbb{R}^{d_m \times 1} \to \mathbb{R}^d$).

$$G = \sum_{m=1}^{M} \sum_{i,j=1}^{n_m} W_g^m(i,j) \left\| \mathbf{f}^{m\top}\mathbf{x}_i^m - \mathbf{f}^{m\top}\mathbf{x}_j^m \right\|^2 = \operatorname{tr}(\mathbf{F}^\top \mathbf{X}\mathbf{L}_g \mathbf{X}^\top \mathbf{F}) \tag{B.5}$$

This objective formulation allows for re-expression using a joint graph Laplacian matrix $\mathbf{L}_g = \text{diag}(\mathbf{L}_g^1, \ldots, \mathbf{L}_g^M) \in \mathbb{R}^{N \times N}$, which consists of the domain-specific Laplacians $\mathbf{L}_g^m \in \mathbb{R}^{n_m \times n_m}$. Each Laplacian $\mathbf{L}_g^m$ is the difference between the degree matrix $\mathbf{U}_g^m(i,i)$ and the adjacency matrix $\mathbf{W}_g^m$.

$$\mathbf{U}_g^m(i,i) = \sum_j \mathbf{W}_g^m(i,j) \qquad\qquad \mathbf{L}_g^m = \mathbf{U}_g^m - \mathbf{W}_g^m \tag{B.6, 7}$$

B.5 CLASS SIMILARITY OBJECTIVE

The second minimization component of $\mathbf{A} = \mu\mathbf{G} + \mathbf{S}$ is the objective $\mathbf{S}$, which aims to minimize the distances between samples that belong to the same class both within and across all domains. Class similarity matrices $\mathbf{W}_s^{m,m'} \in \mathbb{R}^{n_m \times n_{m'}}$ are devised to include in the objective only those sample pairs that belong to the same class, as given below.

$$\mathbf{W}_s^{m,m'}(i,j) = \begin{cases} 0, & \text{if } y_i^m \neq y_j^{m'}, \text{ or when } y_i^m \text{ or } y_j^{m'} \text{ are unlabeled.} \\ 1, & \text{if } y_i^m = y_j^{m'}. \end{cases} \tag{B.8}$$

The joint graph Laplacian matrix $\mathbf{L}_s$ is associated with the joint class similarity matrix $\mathbf{W}_s = [\mathbf{W}_s^{m,m'}]$ and is determined as shown previously. The objective S associated with $\mathbf{S}$ can then be similarly expressed as with G, with the alternative formulation utilizing the Laplacian matrix.

$$S = \sum_{m,m'=1}^{M} \sum_{i,j=1}^{n_m,n_{m'}} W_s^{m,m'}(i,j) \left\| \mathbf{f}^{m\top}\mathbf{x}_i^m - \mathbf{f}^{m'\top}\mathbf{x}_j^{m'} \right\|^2 = \text{tr}(\mathbf{F}^\top \mathbf{X}\mathbf{L}_s\mathbf{X}^\top \mathbf{F}) \tag{B.9}$$

It should be noted that there are two requirements, namely a classification for each domain and class correspondence or domain matching between the different domains in order to establish class equivalency. The manifold reduction used in this thesis produces unsupervised classification for the test feature space and summarises the manifold, whereafter manifold matching is used to establish across-domain class correspondence.

B.6 CLASS DISSIMILARITY OBJECTIVE

In order to reduce the likelihood of a trivial solution to the eigenvalue decomposition, as when all points are mapped onto the same location, a maximization objective $\mathbf{B} = \mathbf{D}$ needs to be integrated to spread the projected samples onto a proper solution. Class dissimilarity matrices $\mathbf{W}_d^{m,m'} \in \mathbb{R}^{n_m \times n_{m'}}$ are constructed to distance sample pairs if they belong to different classes, which are given as follows.

$$\mathbf{W}_d^{m,m'}(i,j) = \begin{cases} 0, & \text{if } y_i^m = y_j^{m'}, \text{ or when } y_i^m \text{ or } y_j^{m'} \text{ are unlabeled.} \\ 1, & \text{if } y_i^m \neq y_j^{m'}. \end{cases} \tag{B.10}$$

The objective D that is associated with $\mathbf{D}$ in the original Rayleigh quotient can be expressed as in Equation B.11. The joint graph Laplacian matrix $\mathbf{L}_d$ is based on the joint class dissimilarity matrix

$\mathbf{W}_d = [\mathbf{W}_d^{m,m'}]$, and is used to permit integration into the Rayleigh quotient formulation.

$$D = \sum_{m,m'=1}^{M} \sum_{i,j=1}^{n_m,n_{m'}} W_d^{m,m'}(i,j) \left\| \mathbf{f}^{m\top}\mathbf{x}_i^m - \mathbf{f}^{m'\top}\mathbf{x}_j^{m'} \right\|^2 = \mathrm{tr}(\mathbf{F}^\top \mathbf{X}\mathbf{L}_d\mathbf{X}^\top \mathbf{F}) \tag{B.11}$$

B.7 ALIGNMENT PROJECTION

Using substitution, it can be shown that the generalized Rayleigh quotient formulation can be rewritten as in Equation B.12, where three Laplacian matrices $\mathbf{L}_g$, $\mathbf{L}_s$ and $\mathbf{L}_d$ encode the desired sample relationships respectively embedded in the objectives G, S and D.

$$\mathbf{F}^{\mathrm{opt}} \min_{\mathbf{F}} \left\{ \mathrm{tr}\left((\mathbf{F}^\top \mathbf{X}\mathbf{L}_d\mathbf{X}^\top \mathbf{F})^{-1} \mathbf{F}^\top \mathbf{X}(\mu\mathbf{L}_g + \mathbf{L}_s)\mathbf{X}^\top \mathbf{F} \right) \right\} \tag{B.12}$$

The overall minimization problem is solved through generalized eigenvalue decomposition where the eigenvectors φ_i with the smallest eigenvalues λ_i form the desired projections. The generalized eigenvalue decomposition is given as follows:

$$\mathbf{X}(\mu\mathbf{L}_g + \mathbf{L}_s)\mathbf{X}^\top \varphi = \lambda \mathbf{X}\mathbf{L}_d\mathbf{X}^\top \varphi \tag{B.13}$$

The optimal projection $\mathbf{F}^{\mathrm{opt}}$ to the common latent space $\mathcal{F}$ is then calculated from the eigenvectors and eigenvalues as shown in Equation B.14. The individual projections can then be ascertained as shown below, in order to project each individual domain onto the common space.

$$\mathbf{F}^{\mathrm{opt}} = \left[\sqrt{\lambda_1}\varphi_1 \middle| \cdots \middle| \sqrt{\lambda_d}\varphi_d \right] = \begin{bmatrix} \mathbf{f}_1^1 & \cdots & \mathbf{f}_d^1 \\ \mathbf{f}_1^2 & \cdots & \mathbf{f}_d^2 \\ \vdots & \vdots & \vdots \\ \mathbf{f}_1^M & \cdots & \mathbf{f}_d^M \end{bmatrix} \tag{B.14}$$

The projection $\mathcal{P}$ ($\mathbb{R}^{d\times d_m} \times \mathbb{R}^{d_m \times n_m} \rightarrow \mathbb{R}^{d\times n_m}$) of a dataset $\mathbf{X}_*^m$ is then the matrix multiplication of the associate projection function $\mathbf{f}^{m\top}$ with the dataset, as given below.

$$\mathcal{P}(\mathbf{X}_*^m) = \mathbf{f}^{m\top}\mathbf{X}_*^m \tag{B.15}$$

APPENDIX C WEIGHTED GENERALIZATION OF INTERNAL INDICES

Weighted clustering and weighted generalisations for point-biserial correlation, Hubert's Gamma, Hubert's D, Hubert's C, Silhouette, Calinski-Harabasz and Pseudo R^2 internal validation indices are described in [40]. Several more of the well-known indices [41] are generalized for weighted features in this appendix for use in the thesis.

C.1 PRELIMINARIES

The normal and weighted formulations of internal validation indices are denoted respectively by $\mathcal{C}$ and $\overline{\mathcal{C}}$. The overline notation is also used to refer to the weighted alternatives to each variable occurring in this generalization treatment. The generalization objective is to admit partial, fractionalised or otherwise real valued points that contribute variable representative amounts to a validation index calculation, so that points with greater salience affect the calculation to a greater degree. The generalized formulation guideline is adherence to the coincidence of normal and weighted formulations when all point weights $\mathbf{w} = \{w_i,\ 1 \leq i \leq N\}$ are one, as given by

$$\lim_{\substack{\mathbf{w}: w_i \to 1 \\ 1 \leq i \leq N}} \overline{\mathcal{C}}(\mathbf{w}) = \mathcal{C} \tag{C.1}$$

There are N observations denoted by M, with real scalar weights $\mathbf{w}$. The number of clusters for a given calculation is K, and the weight per cluster W_k and total weight W are given by

$$W_k = \sum_{i \in I_k} w_i \qquad\qquad W = \sum_{k=1}^{K} W_k \tag{C.2, 3}$$

The number N_W of unordered within-cluster pairs and the weight $\overline{N}_W$ of those pairs are

$$N_W = \sum_{k=1}^{K} \frac{n_k(n_k - 1)}{2} \qquad\qquad \overline{N}_W = \sum_{k=1}^{K} \sum_{\substack{i,j \in I_k \\ j > i}} w_i w_j \tag{C.4, 5}$$

Similarly, the total number N_T of unordered pairs and the associated weight $\overline{N}_T$ are

$$N_T = \frac{N(N-1)}{2} \qquad\qquad \overline{N}_T = \sum_{\substack{i,j=1 \\ j > i}}^{N} w_i w_j \tag{C.6, 7}$$

The number N_B of unordered between-cluster point pairs and their weight $\overline{N}_B$ are the difference between total and within-cluster values, such as

$$N_B = N_T - N_W \qquad \overline{N}_B = \overline{N}_T - \overline{N}_W \tag{C.8, 9}$$

The global centroid G and its weighted counterpart $\overline{G}$ are calculated as

$$G = \frac{1}{N} \sum_{1 \le i \le N} M_i \qquad \overline{G} = \frac{1}{W} \sum_{1 \le i \le N} w_i M_i \tag{C.10, 11}$$

The centroid G_k for cluster k and the weighted centroid $\overline{G}_k$ are given by

$$G_k = \frac{1}{n_k} \sum_{i \in I_k} M_i \qquad \overline{G}_k = \frac{1}{W_k} \sum_{i \in I_k} w_i M_i \tag{C.12, 13}$$

Between-group dispersion BS and the weighted version $\overline{BS}$ are calculated as

$$BS = \sum_{k=1}^{K} n_k d(G_k, G)^2 \qquad \overline{BS} = \sum_{k=1}^{K} W_k d(\overline{G}_k, \overline{G})^2 \tag{C.14, 15}$$

Normal and weighted within-cluster dispersions WS and $\overline{WS}$ are the sum of individual dispersions, so that

$$WS = \sum_{k=1}^{K} \sum_{i \in I_k} d(M_i, G_k)^2 \qquad \overline{WS} = \sum_{k=1}^{K} \sum_{i \in I_k} w_i d(M_i, \overline{G}_k)^2 \tag{C.16, 17}$$

The global scatter matrices T and $\overline{T}$ consist of the following entries T_{ij} and $\overline{T}_{ij}$ calculated in terms of coordinates $M_{li} = M_{l(i)}$, $\mu_i = G_{(i)}$ and $\overline{\mu}_i = \overline{G}_{(i)}$ for feature dimension i, so that

$$T_{ij} = \sum_{l=1}^{N} (M_{li} - \mu_i)(M_{lj} - \mu_j) \qquad \overline{T}_{ij} = \sum_{l=1}^{N} w_l (M_{li} - \overline{\mu}_i)(M_{lj} - \overline{\mu}_j) \tag{C.19}$$

Normal and weighted individual within-group scatter matrices $WG^{\{k\}}$ and $\overline{WG}^{\{k\}}$ are calculated in terms of $\mu_i^{\{k\}} = G_{k(i)}$ and $\overline{\mu}_i^{\{k\}} = \overline{G}_{k(i)}$ for dimension i, giving the matrix entries

$$WG_{ij}^{\{k\}} = \sum_{l \in I_k} (M_{li} - \mu_i^{\{k\}})(M_{lj} - \mu_j^{\{k\}}) \tag{C.21}$$

$$\overline{WG}_{ij}^{\{k\}} = \sum_{l \in I_k} w_l (M_{li} - \overline{\mu}_i^{\{k\}})(M_{lj} - \overline{\mu}_j^{\{k\}}) \tag{C.22}$$

Total within-group scatter matrices WG and $\overline{WG}$ are the sum of individual matrices producing the matrix entries

$$WG_{ij} = \sum_{k=1}^{K} WG_{ij}^{\{k\}} \qquad \overline{WG}_{ij} = \sum_{k=1}^{K} \overline{WG}_{ij}^{\{k\}} \tag{C.23, 24}$$

Between-group scatter matrices BG and $\overline{BG}$ are determined in terms of coordinate differences between cluster and global barycenters as follows:

$$BG_{ij} = \sum_{k=1}^{K} n_k (\mu_i^{\{k\}} - \mu_i)(\mu_j^{\{k\}} - \mu_j) \tag{C.25}$$

$$\overline{BG}_{ij} = \sum_{k=1}^{K} W_k (\overline{\mu}_i^{\{k\}} - \overline{\mu}_i)(\overline{\mu}_j^{\{k\}} - \overline{\mu}_j) \tag{C.26}$$

C.2 WEIGHTED GENERALIZATIONS

C.2.1 Ball-Hall index [207]

The mean $\mathcal{C}$ ($\overline{\mathcal{C}}$) of the individual mean cluster dispersion within each cluster k with barycenter G_k ($\overline{G}_k$) is

$$\mathcal{C} = \frac{1}{K}\sum_{k=1}^{K}\sum_{i\in I_k}\frac{d(M_i,G_k)^2}{n_k} \qquad \overline{\mathcal{C}} = \frac{1}{K}\sum_{k=1}^{K}\sum_{i\in I_k}\frac{w_i d(M_i,\overline{G}_k)^2}{W_k} \tag{C.27, 28}$$

C.2.2 Banfield-Raftery index [208]

The weighted sum $\mathcal{C}$ ($\overline{\mathcal{C}}$) of the logarithms of the individual mean cluster dispersions is

$$\mathcal{C} = \sum_{k=1}^{K} n_k \log\left(\frac{1}{n_k}\sum_{i\in I_k} d(M_i,G_k)^2\right) \tag{C.29}$$

$$\overline{\mathcal{C}} = \sum_{k=1}^{K} W_k \log\left(\frac{1}{W_k}\sum_{i\in I_k} w_i d(M_i,\overline{G}_k)^2\right) \tag{C.30}$$

C.2.3 C-index [209]

The sum S_W ($\overline{S}_W$) over all clusters of the sum of the N_W ($\overline{N}_W$) distances between all unordered pairs in each cluster is

$$S_W = \sum_{k=1}^{K}\sum_{\substack{i,j\in I_k\\ j>i}} d(M_i,M_j) \qquad \overline{S}_W = \sum_{k=1}^{K}\sum_{\substack{i,j\in I_k\\ j>i}} w_i w_j d(M_i,M_j) \tag{C.31, 32}$$

The subset $P_{\min}$ of the set P_A of all N_T unordered global point pairs is such that the sum of distances of the N_W pairs in $P_{\min}$ is minimum, according to

$$P_{\min} = \underset{\substack{P\subseteq P_A\\ |P|=N_W}}{\operatorname{argmin}}\left(\sum_{\{(i,j)\}\in P} d(M_i,M_j)\right) \tag{C.33}$$

The weighted version $\overline{P}_{\min}$ includes the point pairs with minimum distances such that their combined weight is $\overline{N}_W$, as shown by

$$\overline{P}_{\min} = \underset{\substack{P\subseteq P_A\\ \overline{N}_W\approx\sum_P w_i w_j}}{\operatorname{argmin}}\left(\sum_{\{(i,j)\}\in P} d(M_i,M_j)\right) \tag{C.34}$$

Similarly, the subset $P_{\max}$ of the set P_A of all N_T unordered global point pairs is such that the sum of distances of the N_W pairs in $P_{\max}$ is the maximum, according to

$$P_{\max} = \underset{\substack{P\subseteq P_A\\ |P|=N_W}}{\operatorname{argmax}}\left(\sum_{\{(i,j)\}\in P} d(M_i,M_j)\right) \tag{C.35}$$

The weighted version $\overline{P}_{\max}$ includes the point pairs with maximum distances such that their combined weight is $\overline{N}_W$, as shown by

$$\overline{P}_{\max} = \underset{\substack{P\subseteq P_A \\ \overline{N}_W \approx \sum_P w_i w_j}}{\operatorname{argmax}} \left(\sum_{\{(i,j)\}\in P} d(M_i,M_j) \right) \tag{C.36}$$

Consequently, the sum $S_{\min}$ ($\overline{S}_{\min}$) of the N_W ($\overline{N}_W$) smallest global distances is

$$S_{\min} = \sum_{(i,j)\in P_{\min}} d(M_i,M_j) \qquad \overline{S}_{\min} = \sum_{(i,j)\in P_{\min}} w_i w_j d(M_i,M_j) \tag{C.37, 38}$$

The sum $S_{\max}$ ($\overline{S}_{\max}$) of the N_W ($\overline{N}_W$) largest global distances is given by

$$S_{\max} = \sum_{(i,j)\in P_{\max}} d(M_i,M_j) \qquad \overline{S}_{\max} = \sum_{(i,j)\in P_{\max}} w_i w_j d(M_i,M_j) \tag{C.39, 40}$$

The C index thus measures the ratio of a type of within-cluster dispersion to the maximum possible dispersion for same-sized clusters, which is calculated as

$$\mathcal{C} = \frac{S_W - S_{\min}}{S_{\max} - S_{\min}} \qquad \overline{\mathcal{C}} = \frac{\overline{S}_W - \overline{S}_{\min}}{\overline{S}_{\max} - \overline{S}_{\min}} \tag{C.41, 42}$$

C.2.4 Calinski-Harabasz index [210]

The variance ratio criterion is given in terms of between-cluster variance BS ($\overline{BS}$) and within-cluster variance WS ($\overline{WS}$), so that

$$\mathcal{C} = \frac{BS/(K-1)}{WS/(N-K)} \qquad \overline{\mathcal{C}} = \frac{\overline{BS}(N-K)}{\overline{WS}(K-1)} \tag{C.43, 44}$$

C.2.5 Davies-Bouldin index [211]

The mean distance δ_k ($\overline{\delta}_k$) between points in a cluster k and the cluster barycenter is

$$\delta_k = \frac{1}{n_k} \sum_{i\in I_k} d(M_i,G_k) \qquad \overline{\delta}_k = \frac{1}{W_k} \sum_{i\in I_k} w_i d(M_i,\overline{G}_k) \tag{C.45, 46}$$

The distance $\Delta_{kk'}$ ($\overline{\Delta}_{kk'}$) between two cluster barycenters is

$$\Delta_{kk'} = d(G_{k'},G_k) \qquad \overline{\Delta}_{kk'} = d(\overline{G}_{k'},\overline{G}_k) \tag{C.47, 48}$$

The mean over all clusters of the maximum ratio for each cluster and a paired cluster of the sum of their cluster radii to intercluster distances forms the index given by

$$\mathcal{C} = \frac{1}{K} \sum_{k=1}^{K} \max_{k'\neq k} \frac{\delta_k + \delta_{k'}}{\Delta_{kk'}} \qquad \overline{\mathcal{C}} = \frac{1}{K} \sum_{k=1}^{K} \max_{k'\neq k} \frac{\overline{\delta}_k + \overline{\delta}_{k'}}{\overline{\Delta}_{kk'}} \tag{C.49, 50}$$

C.2.6 Det_Ratio index [212]

The ratio $\mathcal{C}$ ($\overline{\mathcal{C}}$) of the determinants of the total scatter matrix T ($\overline{T}$) and the within-group scatter matrix WG ($\overline{WG}$) is

$$\mathcal{C} = \frac{\det(T)}{\det(WG)} \qquad \overline{\mathcal{C}} = \frac{\det(\overline{T})}{\det(\overline{WG})} \tag{C.51, 52}$$

C.2.7 Generalized Dunn's indices (GDI) [214]

GDIij is the ratio $\mathcal{C}$ ($\overline{\mathcal{C}}$) of minimum between-cluster δ_i ($\overline{\delta}_i$) to maximum within-cluster distance Δ_j ($\overline{\Delta}_j$) for various instantiations of these distances. The ratio is given by

$$\mathcal{C} = \frac{\min_{k\neq k'} \delta_i(k,k')}{\max_k \Delta_j(k)} \qquad \overline{\mathcal{C}} = \frac{\min_{k\neq k'} \overline{\delta}_i(k,k')}{\max_k \overline{\Delta}_j(k)} \tag{C.53, 54}$$

A within-cluster distance $\Delta_1(k) = \max_{\substack{i,j\in I_k \\ i\neq j}} d(M_i,M_j)$ is omitted, since it is weight-robust and does not admit a weighted generalization. Within-cluster distance Δ_2 ($\overline{\Delta}_2$) is the mean distance between all ordered pairs of a cluster as defined by

$$\Delta_2(k) = \sum_{i\in I_k} \sum_{j\in I_k\setminus i} \frac{d(M_i,M_j)}{n_k(n_k-1)} \qquad \overline{\Delta}_2(k) = \sum_{i\in I_k} w_i \sum_{j\in I_k\setminus i} \frac{w_j d(M_i,M_j)}{W_k(W_k-w_i)} \tag{C.55, 56}$$

The within-cluster diameter Δ_3 ($\overline{\Delta}_3$) is calculated in terms of the cluster centroid so that

$$\Delta_3(k) = \sum_{i\in I_k} \frac{2d(M_i,G_k)}{n_k} \qquad \overline{\Delta}_3(k) = \sum_{i\in I_k} \frac{2w_i d(M_i,\overline{G}_k)}{W_k} \tag{C.57, 58}$$

Between-cluster distances δ_1 ($\overline{\delta}_1$) and δ_2 ($\overline{\delta}_2$) are weight-robust and are not generalized. These are defined as single-linkage and complete-linkage distances

$$\delta_1(k,k') = \min_{\substack{i\in I_k \\ j\in I_{k'}}} d(M_i,M_j) \qquad \delta_2(k,k') = \max_{\substack{i\in I_k \\ j\in I_{k'}}} d(M_i,M_j) \tag{C.59, 60}$$

The average linkage δ_3 ($\overline{\delta}_3$) between clusters k and k' is

$$\delta_3(k,k') = \sum_{i\in I_k} \sum_{j\in I_{k'}} \frac{d(M_i,M_j)}{n_k n_{k'}} \qquad \overline{\delta}_3(k,k') = \sum_{i\in I_k} w_i \sum_{j\in I_{k'}} \frac{w_j d(M_i,M_j)}{W_k W_{k'}} \tag{C.62}$$

The centroid linkage δ_4 ($\overline{\delta}_4$) is defined as the distance between cluster barycenters so that

$$\delta_4(k,k') = d(G_k,G_{k'}) \qquad \overline{\delta}_4(k,k') = d(\overline{G}_k,\overline{G}_{k'}) \tag{C.64, 65}$$

The weighted mean δ_5 ($\overline{\delta}_5$) of the mean distances between cluster points and their barycenter is formulated as

$$\delta_5(k,k') = \frac{\sum_{i\in I_k} d(M_i,G_k) + \sum_{j\in I_{k'}} d(M_j,G_{k'})}{n_k + n_{k'}} \tag{C.66}$$

$$\overline{\delta}_5(k,k') = \frac{\sum_{i\in I_k} w_i d(M_i,\overline{G}_k) + \sum_{j\in I_{k'}} w_j d(M_j,\overline{G}_{k'})}{W_k + W_{k'}} \tag{C.67}$$

The Hausdorff (D_H) distance δ_6 ($\overline{\delta}_6$) is not weight-sensitive and is defined as

$$\delta_6(k,k') = \max\left\{ \sup_{i\in I_k} \inf_{j\in I_{k'}} d(M_i,M_j), \sup_{j\in I_{k'}} \inf_{i\in I_k} d(M_i,M_j) \right\} \tag{C.68}$$

C.2.8 Baker-Hubert's Gamma [206]

The number of concordant quadruples (pair of unordered pairs) s^+ ($\overline{s}^+$) where a point pair from different clusters have a larger distance than a same-cluster pair is given by

$$s^+ \quad = \quad \overbrace{\sum_{\substack{k,k'=1\\k'>k}}^{K} \sum_{r\in I_k} \sum_{s\in I_{k'}}}^{\text{different-cluster pair}} \left(\overbrace{\sum_{l=1}^{K} \sum_{\substack{u,v\in I_l\\v>u}}}^{\text{same-cluster pair}} \frac{1}{2}\left(1 \quad + \quad \operatorname{sgn}\bigl(d(M_r,M_s) \quad - \quad d(M_u,M_v)\bigr)\right)\right) \tag{C.69}$$

$$\overline{s}^+ \quad = \quad \sum_{\substack{k,k'=1\\k'>k}}^{K} \sum_{r\in I_k} \sum_{s\in I_{k'}} w_r w_s \left(\sum_{l=1}^{K} \sum_{\substack{u,v\in I_l\\v>u}} \frac{w_u w_v}{2}\left(1 \quad + \quad \operatorname{sgn}\bigl(d(M_r,M_s) \quad - \quad d(M_u,M_v)\bigr)\right)\right) \tag{C.70}$$

Similarly, the number of discordant quadruples (pair of unordered pairs) s^- ($\overline{s}^-$) where a point pair from different clusters has a smaller distance than a same-cluster pair is given by

$$s^- \quad = \quad \sum_{\substack{k,k'=1\\k'>k}}^{K} \sum_{r\in I_k} \sum_{s\in I_{k'}} \left(\sum_{l=1}^{K} \sum_{\substack{u,v\in I_l\\v>u}} \frac{1}{2}\left(1 \quad - \quad \operatorname{sgn}\bigl(d(M_r,M_s) \quad - \quad d(M_u,M_v)\bigr)\right)\right) \tag{C.71}$$

$$\overline{s}^- \quad = \quad \sum_{\substack{k,k'=1\\k'>k}}^{K} \sum_{r\in I_k} \sum_{s\in I_{k'}} w_r w_s \left(\sum_{l=1}^{K} \sum_{\substack{u,v\in I_l\\v>u}} \frac{w_u w_v}{2}\left(1 \quad - \quad \operatorname{sgn}\bigl(d(M_r,M_s) \quad - \quad d(M_u,M_v)\bigr)\right)\right) \tag{C.72}$$

The Gamma index $\mathcal{C}$ ($\overline{\mathcal{C}}$) is then a measure of the ratio of concordant quadruples so that

$$\mathcal{C} = \frac{s^+ - s^-}{s^+ + s^-} \qquad\qquad \overline{\mathcal{C}} = \frac{\overline{s}^+ - \overline{s}^-}{\overline{s}^+ + \overline{s}^-} \tag{C.73, 74}$$

C.2.9 G+ index [213]

Using the number of discordant quadruples s^- ($\overline{s}^-$) from Baker-Hubert's Gamma index as a ratio $\mathcal{C}$ ($\overline{\mathcal{C}}$) to the total number of unordered quadruples renders the G+ index as

$$\mathcal{C} = \frac{s^-}{N_T(N_T-1)/2} \qquad\qquad \overline{\mathcal{C}} = \frac{2\overline{s}^-}{\overline{N}_T(\overline{N}_T-1)} \tag{C.75, 76}$$

C.2.10 Ksq_DetW [215]

This index is also known as $k^2|W|$ and is formulated in terms of the number of clusters K and the within-group scatter matrix WG ($\overline{WG}$) as

$$\mathcal{C} = K^2 \det(WG) \qquad \overline{\mathcal{C}} = K^2 \det(\overline{WG}) \tag{C.77, 78}$$

C.2.11 Log_Det Ratio [212]

The logarithmic version of the Det_Ratio is defined in terms of the total scatter T ($\overline{T}$) and the within-group scatter WG ($\overline{WG}$) as follows:

$$\mathcal{C} = N \log\left(\frac{\det(T)}{\det(WG)}\right) \qquad \overline{\mathcal{C}} = W \log\left(\frac{\det(\overline{T})}{\det(\overline{WG})}\right) \tag{C.79, 80}$$

C.2.12 Log_SS Ratio [216]

The logarithm $\mathcal{C}$ ($\overline{\mathcal{C}}$) of the between-cluster dispersion BS ($\overline{BS}$) and the total within-cluster dispersion WS ($\overline{WS}$) is

$$\mathcal{C} = \log\left(BS/WS\right) \qquad \overline{\mathcal{C}} = \log\left(\overline{BS}/\overline{WS}\right) \tag{C.81, 82}$$

C.2.13 McClain-Rao index [217]

The sum S_W ($\overline{S}_W$) over all clusters of the sum of the N_W ($\overline{N}_W$) distances between all unordered pairs is used as defined for the C index. The between-cluster distance sum S_B ($\overline{S}_B$) is similarly formulated as the sum of the N_B ($\overline{N}_B$) distances between all unordered between-cluster pairs, producing

$$S_B = \sum_{\substack{k,k'=1\\k'>k}}^{K} \sum_{i \in I_k} \sum_{j \in I_{k'}} d(M_i, M_j) \tag{C.83}$$

$$\overline{S}_B = \sum_{\substack{k,k'=1\\k'>k}}^{K} \sum_{i \in I_k} \sum_{j \in I_{k'}} w_i w_j d(M_i, M_j) \tag{C.84}$$

The relationship between the mean within-cluster distances and mean between-cluster distances is then shown by

$$\mathcal{C} = \frac{S_W/N_W}{S_B/N_B} \qquad \overline{\mathcal{C}} = \frac{\overline{S}_W \overline{N}_B}{\overline{N}_W \overline{S}_B} \tag{C.85, 86}$$

C.2.14 PBM index [218]

The maximum possible between-cluster distance D_B ($\overline{D}_B$) is calculated as

$$D_B = \max_{k<k'} d(G_k, G_{k'}) \qquad \overline{D}_B = \max_{k<k'} d(\overline{G}_k, \overline{G}_{k'}) \tag{C.87, 88}$$

The sum E_W ($\overline{E_W}$) of within-cluster distances is given by

$$E_W = \sum_{k=1}^{K} \sum_{i \in I_k} d(M_i, G_k) \qquad \overline{E}_W = \sum_{k=1}^{K} \sum_{i \in I_k} w_i d(M_i, \overline{G}_k) \tag{C.89, 90}$$

The total sum of distances E_T ($\overline{E_T}$) between observations and the global barycenter is

$$E_T = \sum_{i=1}^{N} d(M_i, G) \qquad \overline{E}_T = \sum_{i=1}^{N} w_i d(M_i, \overline{G}) \tag{C.91, 92}$$

The index by authors Pakhira, Bandyopadhyay and Maulik [218] is then defined as

$$\mathcal{C} = \left(\frac{1}{K} \cdot \frac{E_T}{E_W} \cdot D_B \right)^2 \qquad \overline{\mathcal{C}} = \left(\frac{\overline{E}_T \overline{D}_B}{K \cdot \overline{E}_W} \right)^2 \tag{C.93, 94}$$

C.2.15 Point-Biserial index [219]

Using the definitions S_W ($\overline{S}_W$) from the C index and S_B ($\overline{S}_B$) from Mclain-Rao the point-biserial correlation coefficient can be shown to be equivalent to

$$\mathcal{C} = \left(\frac{S_W}{N_W} - \frac{S_B}{N_B} \right) \frac{\sqrt{N_W N_B}}{N_T} \qquad \overline{\mathcal{C}} = \left(\frac{\overline{S}_W}{\overline{N}_W} - \frac{\overline{S}_B}{\overline{N}_B} \right) \frac{\sqrt{\overline{N}_W \overline{N}_B}}{\overline{N}_T} \tag{C.95, 96}$$

C.2.16 Ratkowsky-Lance index [220]

The square root of the sum over all feature dimensions of the quotient of the cluster dispersion and the total dispersion for a given feature dimension is

$$\mathcal{C} = \sqrt{ \frac{1}{pK} \sum_{j=1}^{p} \frac{ \sum_{k=1}^{K} n_k \left(\sum_{i \in I_k} M_{ij}/n_k - \sum_{i=1}^{N} M_{ij}/N \right)^2 }{ \sum_{i=1}^{N} \left(M_{ij} - \sum_{i=1}^{N} M_{ij}/N \right)^2 } } \tag{C.97}$$

$$\overline{\mathcal{C}} = \sqrt{ \frac{1}{pK} \sum_{j=1}^{p} \frac{ \sum_{k=1}^{K} W_k \left(\sum_{i \in I_k} \frac{w_i M_{ij}}{W_k} - \sum_{i=1}^{N} \frac{w_i M_{ij}}{W} \right)^2 }{ \sum_{i=1}^{N} w_i \left(M_{ij} - \sum_{i=1}^{N} w_i M_{ij}/W \right)^2 } } \tag{C.98}$$

C.2.17 Ray-Turi [221]

The ratio of the mean within-cluster dispersion to the smallest possible between-centroid squared distance is provided as

$$\mathcal{C} = \frac{WS/N}{\min\limits_{k<k'} d(G_k, G_{k'})^2} \qquad \overline{\mathcal{C}} = \frac{\overline{WS}/W}{\min\limits_{k<k'} d(\overline{G}_k, \overline{G}_{k'})^2} \qquad \text{(C.99, 100)}$$

C.2.18 Scott-Symons index [212]

The weighted sum $\mathcal{C}$ ($\overline{\mathcal{C}}$) of the logarithm of the determinant of each mean within-group scatter matrix is defined by

$$\mathcal{C} = \sum_{k=1}^{K} n_k \log\det\left(\frac{WG^{\{k\}}}{n_k}\right) \qquad \overline{\mathcal{C}} = \sum_{k=1}^{K} W_k \log\det\left(\frac{\overline{WG}^{\{k\}}}{W_k}\right) \qquad \text{(C.101, 102)}$$

C.2.19 S_Dbw index [222]

The global feature variance $\mathcal{V}_i$ ($\overline{\mathcal{V}}_i$) for feature dimension i is determined in terms of $\mu_i = G_{k(i)}$ ($\overline{\mu}_i = \overline{G}_{k(i)}$) as

$$\mathcal{V}_i = \frac{1}{N}\sum_{l=1}^{N}(M_{li} - \mu_i)^2 \qquad \overline{\mathcal{V}}_i = \frac{1}{W}\sum_{l=1}^{N} w_l (M_{li} - \overline{\mu}_i)^2 \qquad \text{(C.103, 104)}$$

The cluster feature variance $\mathcal{V}_i^{\{k\}}$ ($\overline{\mathcal{V}}_i^{\{k\}}$) for feature dimension i is determined in terms of $\mu_i^{\{k\}} = G_{k(i)}$ ($\overline{\mu}_i^{\{k\}} = \overline{G}_{k(i)}$) so that

$$\mathcal{V}_i^{\{k\}} = \frac{1}{n_k}\sum_{l\in I_k}(M_{li} - \mu_i^{\{k\}})^2 \qquad \overline{\mathcal{V}}_i^{\{k\}} = \frac{1}{W_k}\sum_{l\in I_k} w_l (M_{li} - \overline{\mu}_i^{\{k\}})^2 \qquad \text{(C.105, 106)}$$

The average scattering $\mathcal{S}$ ($\overline{\mathcal{S}}$) for clusters is then formulated as the sum of quotients of the norms of the cluster feature variance and global feature variance as shown by

$$\mathcal{S} = \frac{1}{K}\sum_{k=1}^{K}\frac{\sqrt{\sum_{i=1}^{p}\left(\mathcal{V}_i^{\{k\}}\right)^2}}{\sqrt{\sum_{i=1}^{p}\left(\mathcal{V}_i\right)^2}} \qquad \overline{\mathcal{S}} = \frac{1}{K}\sum_{k=1}^{K}\frac{\sqrt{\sum_{i=1}^{p}\left(\overline{\mathcal{V}}_i^{\{k\}}\right)^2}}{\sqrt{\sum_{i=1}^{p}\left(\overline{\mathcal{V}}_i\right)^2}} \qquad \text{(C.107, 108)}$$

Densities $\gamma_{kk'}(M)$ ($\overline{\gamma}_{kk'}(M)$) are determined as the number of points from both clusters k and k' in a ball with radius σ ($\overline{\sigma}$) centered at a point M. The radius is a measure of the norm of the standard deviation of clusters divided by the number of clusters, so that

$$\sigma = \frac{1}{K}\sqrt{\sum_{k=1}^{K}\sqrt{\sum_{i=1}^{p}\left(\mathcal{V}_i^{\{k\}}\right)^2}} \qquad \overline{\sigma} = \frac{1}{K}\sqrt{\sum_{k=1}^{K}\sqrt{\sum_{i=1}^{p}\left(\overline{\mathcal{V}}_i^{\{k\}}\right)^2}} \qquad \text{(C.109, 110)}$$

The density $\gamma_{kk'}(M)$ ($\overline{\gamma}_{kk'}(M)$) is then the number of points in the union of clusters k and k' that are within the σ-ball ($\overline{\sigma}$-ball) centered at point M, as given by

$$\gamma_{kk'}(M) = \sum_{i \in I_k \cup I_{k'}} \frac{\operatorname{sgn}\big(\sigma - d(M_i, M)\big) + 1}{2} \tag{C.111}$$

$$\overline{\gamma}_{kk'}(M) = \sum_{i \in I_k \cup I_{k'}} w_i \left(\frac{\operatorname{sgn}\big(\overline{\sigma} - d(M_i, M)\big) + 1}{2} \right) \tag{C.112}$$

The ratio $R_{kk'}$ ($\overline{R}_{kk'}$) of the density at the midpoint between clusters k and k' to the maximum of the cluster densities at their barycenters is defined as

$$R_{kk'} = \frac{\gamma_{kk'}\big((G_k + G_{k'})/2\big)}{\max\big(\gamma_{kk'}(G_k), \gamma_{kk'}(G_{k'})\big)} \tag{C.113}$$

$$\overline{R}_{kk'} = \frac{\overline{\gamma}_{kk'}\big((\overline{G}_k + \overline{G}_{k'})/2\big)}{\max\big(\overline{\gamma}_{kk'}(\overline{G}_k), \overline{\gamma}_{kk'}(\overline{G}_{k'})\big)} \tag{C.114}$$

The between-cluster density $\mathcal{G}$ ($\overline{\mathcal{G}}$) is the mean of the cluster densities between all unordered cluster pairs as given by

$$\mathcal{G} = \frac{2}{K(K-1)} \sum_{k<k'} R_{kk'} \qquad \overline{\mathcal{G}} = \frac{2}{K(K-1)} \sum_{k<k'} \overline{R}_{kk'} \tag{C.115, 116}$$

The final criterion is formulated as the sum of the average scattering and the average between-cluster density, as shown by

$$\mathcal{C} = \mathcal{S} + \mathcal{G} \qquad \overline{\mathcal{C}} = \overline{\mathcal{S}} + \overline{\mathcal{G}} \tag{C.117, 118}$$

C.2.20 Silhouette index [223]

The mean distance $a(i)$ ($\overline{a}(i)$) between a point i and the other points in its cluster k is

$$a(i) = \sum_{\substack{i' \in I_k \\ i' \neq i}} \frac{d(M_i, M_{i'})}{n_k - 1} \qquad \overline{a}(i) = \sum_{\substack{i' \in I_k \\ i' \neq i}} \frac{w_{i'} d(M_i, M_{i'})}{W_k - w_i} \tag{C.119, 120}$$

Similarly, the mean distance $b(i)$ ($\overline{b}(i)$) between a point i and the points of another cluster k', chosen so that the distance is the minimum, is given by

$$b(i) = \min_{k' \neq k} \left(\sum_{i' \in I_{k'}} \frac{d(M_i, M_{i'})}{n_{k'}} \right) \qquad \overline{b}(i) = \min_{k' \neq k} \left(\sum_{i' \in I_{k'}} \frac{w_{i'} d(M_i, M_{i'})}{W_{k'}} \right) \tag{C.121, 122}$$

The global silhouette index is then defined as the mean of the mean silhouette per cluster, which is a relatively good indication of how well points belong to their closest clusters, and is defined by

$$\mathcal{C} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i \in I_k} \frac{b(i) - a(i)}{\max\big(a(i), b(i)\big)} \tag{C.123}$$

$$\overline{\mathcal{C}} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{W_k} \sum_{i \in I_k} w_i \frac{\overline{b}(i) - \overline{a}(i)}{\max\big(\overline{a}(i), \overline{b}(i)\big)} \tag{C.124}$$

C.2.21 Tau index [224]

The τ index of Kendall is re-appropriated in terms of the difference between concordant and discordant cardinalities defined for Baker-Hubert's Gamma as a ratio to the number of unordered quadruples, formulated as

$$\mathcal{C} = \frac{s^{+} - s^{-}}{\sqrt{N_B N_W \left(\frac{N_T(N_T-1)}{2}\right)}} \qquad \overline{\mathcal{C}} = \frac{\overline{s}^{+} - \overline{s}^{-}}{\sqrt{\overline{N}_B \overline{N}_W \left(\frac{\overline{N}_T(\overline{N}_T-1)}{2}\right)}} \qquad \text{(C.125, 126)}$$

C.2.22 Trace-W index [225]

The trace of the within-group scatter matrix is equivalent to the within-cluster dispersion, which defines this index as

$$\mathcal{C} = \mathrm{Tr}(WG) = WS \qquad \overline{\mathcal{C}} = \overline{WS} \qquad \text{(C.127, 128)}$$

C.2.23 Trace-WiB index [226]

The trace of the quotient of the between-group scatter and within-group scatter matrices forms the Trace_$W^{-1}B$ index as

$$\mathcal{C} = \mathrm{Tr}(WG^{-1} \cdot BG) \qquad \overline{\mathcal{C}} = \mathrm{Tr}(\overline{WG}^{-1} \cdot \overline{BG}) \qquad \text{(C.129, 130)}$$

C.2.24 Wemmert-Gançarski index [227]

Using a less computationally complex alternative to the silhouette index is possible by measuring the ratio R_k ($\overline{R}_k$) of the distance of a point to the barycenter of its own cluster k to the distance of the closest barycenter of another cluster k', as formulated in

$$R_k(M) = \frac{||M - G_k||}{\min\limits_{k' \neq k} ||M - G_{k'}||} \qquad \overline{R}_k(M) = \frac{||M - \overline{G}_k||}{\min\limits_{k' \neq k} ||M - \overline{G}_{k'}||} \qquad \text{(C.131, 132)}$$

The index then sums over the clusters the difference between the number of cluster points and their values R_k ($\overline{R}_k$), if the difference is larger than 0, and finally divides by the total number of points so that

$$\mathcal{C} = \frac{1}{N} \sum_{k=1}^{K} \max\Big\{0, n_k - \sum_{i \in I_k} R_k(M_i)\Big\} \qquad \text{(C.133)}$$

$$\overline{\mathcal{C}} = \frac{1}{W} \sum_{k=1}^{K} \max\Big\{0, W_k - \sum_{i \in I_k} w_i \overline{R}_k(M_i)\Big\} \qquad \text{(C.134)}$$