CONTEXT BASED DETECTION OF URBAN LAND USE ZONES

By

Johan Louw
Supervisor: Dr George Sithole

February 2011

Submitted to the University of Cape Town in fulfilment of the requirements for a MSc Degree in Engineering

Department of Geomatics

Abstract

This dissertation proposes an automated land-use zoning system based on the context of an urban scene. Automated zoning is an important step toward improving object extraction in an urban scene. Object extraction from aerial imagery is important for numerous environmental and socio-economic decision making applications.

Traditional object classification systems rely solely on sensory measurements (colour, texture, and shape of pixels or pixel regions) to generate a description of a scene. These systems are limited by the available sensory content. Human visual perception relies to a large extent on an external contextual understanding. This understanding is used to interpret a scene based on image content.

If land-use (context) can be recognized in an urban scene, an external contextual land-use model can be used to improve object classification results. For instance, for a scene classified as industrial, typical industrial scene parameters can be used to constrain and improve the description of the scene.

In this dissertation urban context is characterized by analyzing training images of different land-use types in an aerial image dataset of the greater Cape Town region. A set of high-level features (e.g. *building density, road-building distance* etc) is measured off manually labeled building and road objects in these images. Through multivariate statistical visualization tests it is shown that different land-use scenes can be discriminated based solely on the high-level feature set. This demonstrates the effectiveness of features of this type in characterizing urban context. A feature selection routine is then used to obtain an optimum subset of high-level features that causes maximum discrimination of land-use classes. This feature subset can be regarded as a definition of context, or a 'scene descriptor'.

The proposed automated zoning routine works by segmenting and classifying land-use regions. Initially, bottom-up object classification is performed. Road block regions are then extracted from incomplete road data using a novel technique. These block regions are then classified to a land-use based on high-level features of the bottom-up data.

To test the effectiveness of the automated zoning routine, experiments were conducted on test scenes of an RGB aerial image dataset of the Cape

Town region. The block extraction results show a clear separation of land-use regions. This demonstrates the effectiveness of the block extraction routine in segmenting land-use regions. Blocks were classified to a land-use based on four high-level features. Classification accuracies of over 75% were achieved. The result shows that a recognition of context can be achieved through incomplete bottom-up results. This is an encouraging result within the framework of image understanding. These bottom-up results can then be potentially improved based on a contextual model such as the 'scene descriptor' mentioned above.

Plagiarism Declaration

1. **I know that plagiarism is wrong. Plagiarism is using another's work and to pretend that it is my own.**
2. **I have used the Harvard convention for citation and referencing. Each significant contribution to, and quotation in, this project, from the work, or works of other people has been attributed and has been cited and referenced.**
3. **This project is my own.**
4. **I have not allowed, and will not allow anyone to copy my work with the intention of passing it off as his or her own work.**

Signed:

Johan Louw
University of Cape Town
May 24, 2011

Acknowledgement

Firstly, I would like to extend my gratitude to my supervisor, Dr George Sithole, for offering me the privaledge of conducting this research, for overseeing the research, and contributing to technical problem solving. I wouldn't have been able to conduct this research without his continous guidance and skills.

I would like to thank the University of Cape Town for providing funds for this research.

A particular note of thanks is given to my Mother and Father, Patricia and Charles Louw, for their love, their continuous encouragement, and for financial support.

I would like to thank the City of Cape Town for providing the data used in this research.

I would like to thank the Department of Geomatics for their support and generous computing facilities.

I would like to thank my CUO collegues for their technical advice and for letting me share their office space.

Many thanks to my brother and sister, my digsmates, and my friends for their love and encouragement.

Lastly the biggest thanks goes to the LORD GOD my father for putting me on this planet and giving me the skills that I have, and giving me strength and guidance daily to complete this task.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Because of the increased pace of urbanization in recent years, the automated mapping of the urban environment has become a topical research subject. Numerous environmental and socio-economic decision making applications benefit from automated urban mapping, such as urban sprawl analysis, transportation infrastructure management, and architectural evaluation. The various output forms can include thematic maps, digital data files amenable to inclusion in a GIS, and more recently 3D city models. Remote sensing observation data in the form of aerial or space-borne imagery has traditionally been the main data source for automated urban mapping. Photographic imagery is attractive because it is rich in content, i.e., it contains geometric as well as colour and texture information about objects. The generation of an urban map from an aerial image requires a discrimination and extraction of relevant objects, e.g. buildings, trees and roads, in the image. This is referred to as 'image classification'.

Image classification is an extensively researched subject spanning numerous scientific fields. Classic image classification methods work by associating each image pixel with an object class based on the spectral properties of that pixel. With advances in sensor technology in the last two decades high spatial resolution imagery has become readily available, resulting in objects of interest being significantly larger than individual pixels. Researchers soon discovered that introducing a spatial awareness of neighbouring pixels into a classification system can greatly improve its performance. A modern trend is to classify groups of neighbouring homogeneous pixels, as opposed to individual pixels, to avoid mis-classification of individual pixels. This is known as Object Based Image Analysis (OBIA) (Benz et al., 2004; Aplin and Smith, 2008; Blaschke, 2009). In addition to spectral properties, shape and texture measurements of the pixel groups have been shown to be effective in improving classification results. This is especially evident in urban

scenes where typical man-made urban objects are often similar in colour but more robustly characterised by shape and texture. For example, roads and buildings my have a similar colour but roads in general have a longer *length-to-width* ratio.

1.2 Problem Statment

The performance of both the former 'pixel-based' and the later 'object-based' image classification techniques is inherently limited by the total amount of available sensory content, whether it be colour, shape or texture. Many different object types in urban scenes can be similar in shape and colour (e.g. road and pavement), and thus even with object-based techniques, misclassification of regions is possible.

Human visual perception relies to a large extent on a recognition of context. Sensory information is used to build gradually and select from a internal repertoire of 'perceptual hypotheses' (Gregory, 1970). When extracting buildings in a scene, for instance, a human interpreter will not solely look at the colour of individual pixels in an image, or even the shape and texture of individual buildings. His building detection decisions will be heavily influenced by a recognition that the scene is e.g. a middle-class residential urban scene in Cape Town, South Africa. The interpreter has a pre-built pattern of what this scene should look like before he interprets the image. He then utilizes this model to obtain the most likely description of the scene based on the image content. There has been research in emulating this contextual understanding to improve automatic image classification, especially for applications such as content-based image retrieval and multimedia applications, and to a lesser extent remote sensing.

This methodology has been widely referred to as 'image understanding'. In the case of an urban scene, urban context relates to the design of an urban scene. Contextual information may include planning specifications, spatial characteristics of the major urban objects in a particular land use area, building and road types, etc. It has been shown that the systematic inclusion of contextual information improves automated urban mapping results by providing contextual hypotheses and constraints in the mapping process (Porway et al., 2008; Liu et al., 2008).

1.3 Previous Work

Well established techniques exist for pixel based (see e.g. Richards and Jia, 2005) and region-based (see e.g. Blaschke, 2009) remote sensing image classification. Although research on remote remote sensing (aerial) Image Understanding (IU) systems has been scarce, the first aerial IU system can be seen as early as 1990 in (Matsuyama and Hwang, 1990). In this book

publication the authors present a general framework for aerial IU. The major concept of this framework is that a complete and idealized description of the input scene is constructed, even if it is partially depicted by the initial observed image features. This idealized description is induced by constraints of a contextual model. A contextual model is a template for what a scene should look like.

In more detail, a two step bottom-up, top-down analysis is performed. Standard image classification techniques are solely *bottom-up* (i.e. an unknown pixel / region is classified based on absolute sensory measurements of that pixel / region). An IU system uses *bottom-up* results (which are incomplete) to choose one of a predefined set of contextual models. This can be referred to as 'scene-matching'. In a top-down phase the chosen contextual model is used to improve the initial bottom-up results. This is done by matching / testing bottom-up results against the contextual model. Through the constraints / parameter ranges imposed by the contextual model, pruning and instantiation can take place. Pruning means to delete objects that are inconsistent with the model. Instantiation means to create a new object if it is required to fill a gap in the model. The goal is to reach a final scene description that causes as much consistency with the contextual model as possible. The contextual model complements missing but necessary information. Thus an IU system is designed to produce high quality final results albeit with poor quality bottom-up results.

More recently developed aerial IU systems are loosely based on this framework. Aksoy et al. (2003) presents a method for constructing contextual models for large scale scene prototypes (e.g. tree covered islands, residential areas with coastline etc). The intended application is content-based image retrieval but parallels can be drawn for urban scene analysis. Their contextual model consists of a set of spatial relations, some of them being fuzzy. The authors present a scene-matching routine, where bottom-up features are compared to scene prototype features in a Bayesian framework, and an appropriate prototype chosen.

Liu et al. (2008) similarly defines a set of spatial relations that are useful in constructing a contextual model. The authors present a case study where it is shown that the inclusion of a few of these spatial relation features improves the extraction of car objects from a high resolution image of an urban scene, using OBIA techniques.

Porway et al. (2008) defines a contextual model as a hierarchical model of an urban scene. The model starts at scene level (the entire scene). Scene level decomposes into groups of objects, such as blocks of buildings or rows of cars. Object groups decompose into single objects, such as a building, which decomposes further into parts and primitives. This hierarchical model helps capture objects, as well as the different characteristics of the scene, at varying scales. To learn this model, sample urban scene images were hand-labeled. Various contextual features, for example *road-building distance*, were mea-

sured off each image to train their contextual model. In a top-down analysis, the bottom-up results are improved in an Markov Random Field (see section 4.2.1) framework. Experimental results show a significant improvement in bottom-up results after implementation of the top-down phase.

An issue with the IU systems in (Liu et al., 2008; Porway et al., 2008) is that a contextual model for only one scene type is defined. Urban scenes while similar are significantly different due to land-use planning policies. There are significant structural differences between e.g. Central Business District (CBD), low cost housing development, industrial etc. A generalized contextual model requires a characterization of different land-use types. Before performing top-down analysis, appropriate land-use context must be chosen through scene-matching. In other words, automated land-use classification is required for a full urban aerial IU system.

A study related to South African urban context was done in (Busgeeth et al., 2008). A hierarchical, rule-based land-use classification typology is proposed. Quickbird imagery of the Soweto region in South Africa was analyzed. Informal settlements were distinguished from formal settlements based on the following high-level features: *average building size, building size variety, formalized / informalized street pattern,* and *tarred / gravel roads*. These features were, however, manually extracted from the imagery. For automation purposes, *formalized / informalized street pattern,* for instance, needs to be defined. Furthermore, only four features were used based on expert knowledge. Ideally, from all possible discriminatory features, an optimum subset needs to be established that causes greatest land-use separation.

In a second study in (Busgeeth et al., 2008), an automated land-use classification system is proposed. To train the classifier, sample scenes of image tiles 120m by 120m are generated from regions of known land-use. Local binary pattern features are automatically extracted from these regions and used to train a Support Vector Machine (SVM) (see section 3.3 for SVM theory). The classifier works by moving a 120m by 120m window over the input image to produce an overlapping set of tiles, where each tile is classified using the trained SVM. Through experimentation on the Soweto dataset, the classifier was able to separate built-up from non built-up areas, as well as formal from informal. This system is based on features directly extracted from raw imagery. Another option is to perform scene-matching using features extracted from classified image objects. This technique is considered closer to human visual perception, as it captures the important properties of individual objects (Jing et al., 2003). Content-based image retrieval studies such as (Aksoy et al., 2003; Ren et al., 2002; Rathi and Majumdar, 2002; Hernandez-Gracidas and Sucar, 2007) have shown effective recognition of scene types based on object properties as well as the spatial relationships between objects. For a system of this type applied to land-use classification, a segmentation and classification of land-use regions is required, based on object features (see figure 1.1).

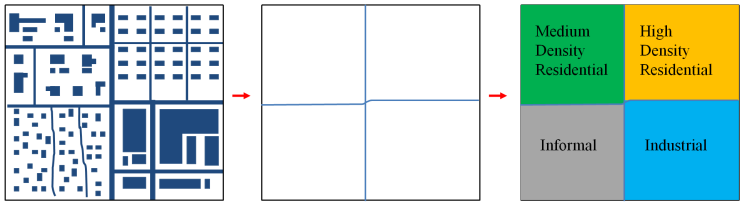


Figure 1.1: Land-use segmentation and classification based on image objects. 1) Bottom-up object classification is performed. 2) Land-use segmentation is performed based on object features. 3) Land-use classification is performed based on object features.

1.4 Research Objectives

Because of socio-economic, cultural, architectural and planning differences, urban scenes while similar are not the same. Consistencies do, however, exist in urban scenes of the same land-use type.

The objective of this thesis is to characterise urban land-use context for the purpose of automated zoning (land-use detection) from aerial imagery. 'Context' refers to the spatial relationships of predominant urban objects, or the general geometric structure of a scene. In this thesis the the term 'high-level features' refers to contextual scene measurements. Examples of high-level features are *average road-building distance, building density, average road width* etc.

Two core research questions are:

1. What is urban context? In other words, what are the high-level features that are most significant in discriminating different urban scene types?

2. Can automated land-use classification be performed based solely on high-level features extracted from bottom-up data? Bottom-up data refers to initial object classification results.

1.5 Scope of Research

1.5.1 Urban scenes to analyze

Urban scenes in the greater Cape Town region of South Africa were analyzed.

1.5.2 Urban objects to analyze

An urban scene contains many features, e.g., buildings, vegetation, roads, water bodies, rail tracks, pedestrians, motor vehicles, etc. This research will only concern itself with the classification and analysis of buildings and roads.

A visual analysis of South African urban scenes was conducted to show that these three objects are most predominant in characterizing the context of a scene.

1.6 Research Methods

In order to characterise the context of South African urban scenes, this research was subdivided into four parts as listed below:

1. Study of current image classification systems
2. Study of current aerial image understanding systems
3. Design and construction of a contextual model for different land-use types
4. Design and development of an automated zoning routine

1.6.1 Study of current image classification systems

The first part of this research (chapters 2 - 4) involves a study of the existing state-of-the-art image classification techniques, in particular applied to high resolution images of urban scenes. The aim of this study is to assess the potential of various techniques in extracting main urban objects, and to identify the general technique that produces the greatest object classification accuracy. The chosen technique will be adopted for a bottom-up image classification phase in this research.

1.6.2 Study of current aerial image understanding systems

The second part of this research (chapter 5) involves a study of existing aerial IU systems, with a focus on the construction of contextual models, and scene-matching strategies. Content-based image retrieval studies are reviewed, and techniques relevant to the objectives of this thesis drawn. A more in-depth review is made on the available urban aerial IU studies.

1.6.3 Design and construction of a contextual model for different scene types

The study of existing aerial IU systems in chapter 5 identifies a research gap in the area of land-use context characterization. This is required to build a generalized urban contextual model. In this dissertation a new contextual model is proposed based on hand-labeled sample scenes of various different land-use types in aerial imagery of the greater Cape Town region. Figure 1.2 illustrates the concept of the proposed contextual model. The method for constructing this model is based loosely on those presented in

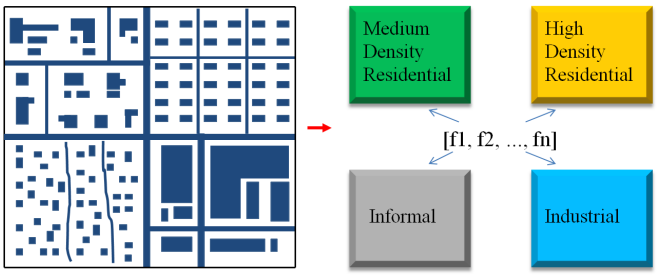


Figure 1.2: Characterizing urban context.
Our model distinguishes and isolates different land-use types. The [f1, f2,....fn] denotes a high-level feature set (*road-building distance, building density, average building size* etc). A unique feature space should exist for each land-use class.

the existing aerial IU studies. It is essentially a measurement of a set of high-level features (*road-building distance*, *building density* etc) in each sample scene. Multivariate statistical visualisation tools are then used to assess whether land-use scenes can be discriminated based on the high-level feature set. This tells us whether the features are useful for characterizing land-use context. A feature selection procedure is then used to establish a subset of high-level features that are most significant in discriminating land-use types. This answers research question 1 (section 1.4).

1.6.4 Design and development of an automated land-use classification routine

Automated land-use classification is a required step for a full urban aerial IU system. In this dissertation a novel automated land-use classification routine is proposed. Figure 1.3 illustrates the concept of the routine. The system works by initially performing bottom-up object classification. The bottom-up classification method is based on that method which was identified as most promising in the study of current image classification systems. Land-use regions are then segmented and classified based on these bottom-up results. Land-use segmentation is accomplished by extracting 'road block' regions from the bottom-up road data using a novel routine. A 'road block' is that region within a closed loop of road segments. Each block is classified to a land-use type based on a set of high-level features extracted from the bottom-up objects within that block.

A qualitative and quantitative assessment of block classification accuracy is carried out, in order to determine whether land-use context can be detected based solely on bottom-up data. This answers research question 2 (section 1.4).

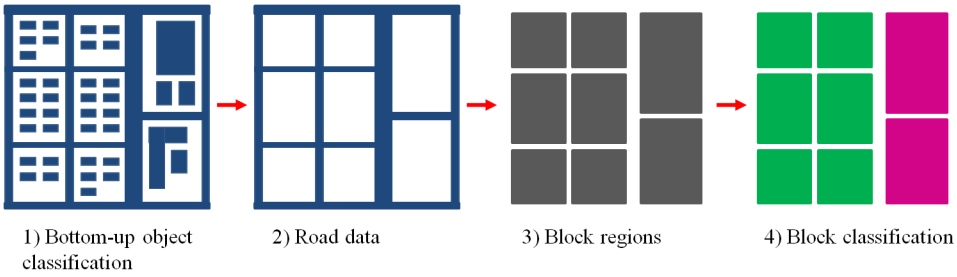


Figure 1.3: Automated land-use classification concept.
1) Bottom-up object classification is performed on a raw image. 2) The road classification results are considered for further analysis. 3) Block regions are extracted from road classification results. 4) Land-use classification is performed on the block regions based on high-level features of the object classification results in step 1).

1.7 Implications of Study

A land-use classification (zoning) tool such as the one proposed in this study can have the following uses:

- Automatic zoning can be performed of an urban area based solely on shape measurements of incomplete bottom-up urban objects. Automated zoning is useful for various land-use applications
- Land-use, and thus context, can be identified from bottom-up object classification results. In a top-down analysis, an appropriate contextual model can then be used to improve the bottom-up results. The idea is that the treatment of objects in an e.g. industrial scene will be different to that in an e.g. informal settlement. Certain 'industrial' parameters and constraints can be used to prune inconsistencies in bottom-up results, or generate hypotheses for missing objects.
- Knowledge based multisensory fusion tactics can be improved. Knowledge of scene context will provide additional cues to the identification of unknown objects present in different sensory data. For instance, an object present in a laser scan point cloud would be treated differently depending on whether the scene is industrial or informal. Object classification based on a combination of features, of a region present in multiple sensor data, will be improved with additional knowledge of the context in which that region falls.

1.8 Outline of Thesis

This thesis is structured as follows: In chapter 2 an overview of traditional image classification theory is presented. Traditional techniques are in general 'parametric' and 'pixel-based'. Chapter 3 provides a review of the currently more popular 'non-parametric' image classification techniques. In chapter 4 a review and analysis of modern region-based image classification methods is presented. In chapter 5 a general framework for aerial image understanding is presented. Content-based image retrieval and remote sensing case studies are then reviewed. A review on spatial relation theory is also given. In chapter 6 a detailed methodology is presented for the design and development of an urban contextual model, as well as a new automated land-use classification routine. Chapter 7 presents results obtained from applying the methodology to an experimental dataset. In chapter 8, conclusions are drawn and future research intentions are given.

Chapter 2

Traditional Image Classification

2.1 Introduction

Image classification refers to the association of an image pixel or region with a label representing a real world conceptual class (road, water body, agricultural field etc) (Mather, 2004). It is an extensively researched subject spanning numerous scientific fields. This chapter will deal with the standard approach to remote sensing (aerial) image classification, with an emphasis on urban scenes. Firstly, a brief overview of the properties of remote sensing imagery needs to be addressed.

2.2 Photographic Imagery

The two main types of remote sensing imagery are panchromatic images and multispectral images (Richards and Jia, 2005),. In panchromatic images, each pixel contains a single measure of the intensity of light reflected from an object. This measure of intensity is known as grey value, since pixels are displayed in shades of Gray. In multispectral images each pixel contains several measures of intensity, each from different wavelengths of light. The different wavelengths are known as spectral bands. Wavelengths for multispectral images typically include the visible as well as the Near Infrared (NIR) spectrum. The NIR band is useful for identifying vegetation regions. Hyperspectral sensors have been developed that can contain more han 200 spectral bands, providing additional information useful for scene classification. Airborne imagery has traditionally been the main source of data for urban mapping. More recently, space-borne imagery has become popular because of its high temporal resolution, large-scale coverage, and multispectral capabilities. In recent years the spatial resolutions available in satellite imagery have increased. This enables objects such as small houses to be

mapped. Since the only disadvantage of using space-borne imagery over airborne imagery (in the past and currently) is the lower resolution, and because resolutions of space-borne imagery have been increasing exponentially over the years, it can be expected that satellite imagery will be the primary mapping source in years to come.

2.3 Digital Image Classification

In automated digital image classification a pixel is allocated to a particular class based on its feature vector. The features are normally the spectral values in each band, e.g. blue, red and NIR brightness values. The feature vector takes the following form:

$$< blue, red, NIR >$$

This is collectively known as the spectral signature. A classification algorithm is employed to associate a certain class with a certain type of spectral signature. A classification algorithm can either be supervised or unsupervised. In supervised classification, the number and names of classes are known a priori. In unsupervised classification, pixels are allocated to a class without prior knowledge of the existence of those classes. .

2.4 Supervised Classification

With supervised classification a classifier is first 'trained' using representative data for each class (Richards and Jia, 2005). The representative data is selected by an analyst. For example, if a pixel-based classifier were to classify a scene into roads and buildings, an analyst would select representative samples of road pixels and building pixels in a training image, in order to 'teach' the classifier to recognise those classes. In more detail, for each building pixel that is selected, the values in each spectral band are observed. The feature vector is projected in a feature space, where the feature values determine the coordinates of its position (see figure 2.1). Where many features exist, as is the case with hyperspectral images, a feature selection procedure may be required in order to reduce the dimensionality of the feature space (see section 2.6 for feature selection strategies).

The goal in the training phase is to partition the feature space by finding optimum decision boundaries between classes. The partition and feature selection processes may be tested for accuracy through a cross validation procedure. This is accomplished by classifying subsets of the existing training data. An accuracy estimate can be deduced from the percentage of correctly classified training pixels.

With a partitioned feature space, unclassified image pixels can now be automatically classified. Each unclassified pixel is analyzed in the feature space

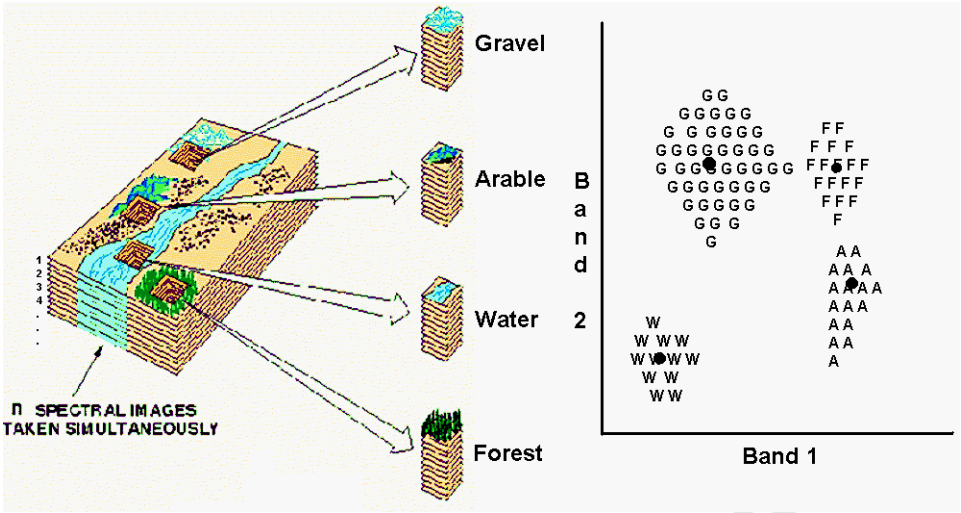


Figure 2.1: Training a remote sensing supervised classifier (UCL, 2011). Representative samples for *Gravel (G), Arable (A), Water (W)* and *Forest (F)* classes are selected in a training image. The spectral values of these samples are projected in a feature space in order to train a classifier, i.e., teach the classifier the spectral signatures for these classes.

where a decision is made as to which class it belongs to. Various classification algorithms can be adopted to make this decision. These classification algorithms can be further subdivided into parametric and non-parametric.

Parametric classifiers rely on the assumption that the feature distribution for each class is Gaussian distributed (Richards and Jia, 2005). Popular algorithms include the 'Minimum Distance' and 'Maximum Likelihood'. These statistical algorithms have been traditionally used in supervised remote sensing image classification. Non-parametric algorithms include machine learning methods, and will be dealt with in the next chapter.

2.4.1 Parametric methods

If the training data are assumed to have a particular distribution (usually the normal distribution), existing well established statistical theory can be employed to make inferences in the feature space. These are known as parametric methods, since parameters such as mean and variance are estimated from the distribution.

2.4.1.1 Minimum distance classifier

The minimum distance classifier works by computing the Euclidean distance between an unknown data point and each class cluster centroid (see figure 2.2). The class cluster centroids are the means of the normally distributed

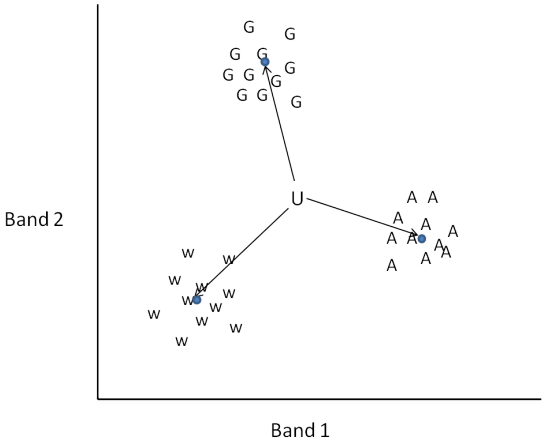


Figure 2.2: Minimum distance classification.
The unknown data point U is assigned to the class with the shortest distance to its cluster centroid.

training data. The unknown data point is then assigned to the class with the minimum distance. The advantage of the minimum distance classification is its simplicity and speed of computation. The disadvantage is that variations in the class distributions are not considered. Thus the classifier can be used when the number of training samples per pixel is limited, where variation will not be a significant influence (Richards and Jia, 2005).

2.4.1.2 Maximum likelihood classification

A popular technique is maximum likelihood classification. Here the mean (pertaining to the class cluster centroid) and the covariance matrix of the normally distributed training data is estimated. Using this information, for each pixel the relative likelihood, or probability, of that pixel belonging to a particular spectral class is computed. The pixel is then labeled according to the highest probability.

Technical overview

Maximum likelihood classification is based on Bayes rule (Richards and Jia, 2005). Let $c = (c_1, c_2, ..., c_i)$ denote a set of classes, where nc is the total number of classes. For a given pixel with feature vector x, the probability that x belongs to class c_i is $P(c_i|x)$, $i = 1, 2, ..., nc$. However, $P(c_i|x)$ is not known directly. Bayes theorem is thus used:

$$P(c_i|x) = \frac{P(x|c_i)P(c_i)}{P(x)} \tag{2.1}$$

where:

$P(c_i)$ is the *prior probability* that c_i occurs in the image.

$P(c_i|x)$ is the conditional probability of c_i given x, also referred to as the posterior probability.

$P(x|c_i)$ is the conditional probability of x given c_i.

$P(x)$ is the probability of x occurring in each class c_i.

$P(c_i)$ is normally assumed to be 1, but can be derived from additional knowledge about the image scene, such as context (see section4.2). In the classification stage we compare $P(c_1|x)$ with $P(c_2|x)$, and can thus cancel $P(x)$. Therefore, $P(x|c_i), i = 1, 2, ..., nc$ are the required conditional probabilities. These are computed by assuming the conditional probability density function (p.d.f.) for each class, derived from training samples, is normally distributed, and estimating mean and covariance parameters from that distribution.

With $P(c_i)$ and $P(x|c_i)$ known, the maximum likely hood classifier can classify an unknown feature vector x by computing the product $P(x|c_i)P(c_i)$ for each class and allocating it to the class with the highest product. This product is known as the Maximum A Posterior (MAP) solution.

The main disadvantage of the maximum likelihood classifier is that the normal distribution assumption limits its performance, since the assumption can be violated. This violation can often occur in the case of remote sensing data classification, especially when data is derived from complex landscapes (Lu and Weng, 2007). More elegant classifiers, such as those based on machine learning theory (Chapter 3), have been introduced into the field of remote sensing. These classifiers should draw our attention as they are distribution free and have shown significant levels of improvement over the more traditional statistical methods.

2.5 Unsupervised Classification

If the user has little idea of the number and names of distinct spectral classes, techniques like maximum likelihood are incompatible. Classification without prior class knowledge can be solved with clustering. Clustering is a general term for the grouping of data with homogeneous properties. In terms of image classification, it can be used to partition an image into unknown classes. In a second step these unknown classes are labeled based on knowledge of the scene or objects to be extracted. Clustering works by recognizing groups (clusters) of unknown pixels projected in a feature space. To recognise clusters in a feature space, a similarity measure has to be established. This is normally a simple distance measure such as the Euclidean distance or the L1 distance [1]. What will often happen is that several acceptable clusters will be recognizable in a certain area. Thus a quality measure is necessary

[1] L1 distance: The L1 distance between two points is the sum of the absolute differences of their coordinates.

to choose one cluster over others. A common quality measure is the Sum of Squared Errors (SSE). Popular clustering methods are ISODATA and K-means. The ISODATA method will know be discussed in order clarify the clustering concept.

ISODATA algorithm (Developed by Ball and Hall, 1965)

This algorithm operates as follows:

1. Points at arbitrary locations in feature space are selected to serve as candidate cluster centers.

2. The location of each pixel in the feature space is determined. Each pixel is allocated to the nearest candidate cluster based on a distance measurement to each cluster centroid.

3. The candidate cluster centers are re-computed to the mean positions of the allocated pixels.

Steps 2. and 3. are repeated until the SSE measure is reduced to a reasonable value. Once a clustering procedure such as the above is completed, post analyses of the encoded clusters will normally take place, such as the following:

1. Each cluster is checked to see whether it is too small to be statistically meaningful.

2. If clusters are too close together that they represent an unnecessary division of data (over-segmentation), they are merged.

3. Clusters that are too elongated (under-segmentation) are split.

2.6 Data Reduction

In many instances a large number of features are available for a classification task. An example is with hyperspectral imagery, where more than 200 bands can exist. It is impractical to use all the available features for classification, due to the computational requirement. Additionally, many of the features may be correlated in some way. A selection of the most significant and uncorrelated features will normally give a similar result, if not better, than the total number of available features (Guyon and Elisseeff, 2003). It is therefore desirable to use the smallest number of possible features that will cause maximum classification accuracy. A data reduction is required.

Two types of feature reduction techniques are established in the literature. The first approach is to project the original feature space into a

subspace of smaller dimensionality. A popular method is the Principle Component Analysis (PCA), where the principle components of the original feature space are extracted to create an artificial set of uncorrelated features (Jolliffe, 1986).

The second approach is based on an evaluation of class separability measures in the original feature space. A subset of features is selected from the original set that maximizes class separability. This is normally done by computing a class separability index (SI) for subset combinations of features, and choosing that subset with the highest SI. A popular separability index is the B-distance (Haralick and Fu, 1983).

To compute an SI for different feature combinations, a searching algorithm is required. One could use, for instance, the *optimum* search. This is an exhaustive search of all possible subsets, and can be computationally expensive. Another option is the *greedy* search, where the effect of removing one feature at a time is observed. If removing a feature reduces class separability, that feature is restored. A further option is the *random* search, where random subsets are extracted. This has the advantage of being inexpensive but important subsets may be missed. A search that can achieve similar results to the *optimum* but is less expensive is the *guided random search*, which is based on genetic algorithms (Holland, 1992).

Feature selection is often necessary for modern remote sensing supervised classification tasks, such as object-oriented classification (see section 4.3.4). It is also useful for the general purpose of identifying features that are significant in causing class separation. Such an analysis was used in the methodology section of this thesis (see e.g. section 6.3.8).

2.7 Conclusion

We have surveyed the traditional parametric supervised image classification methodologies, as well as non-supervised clustering techniques and data reduction techniques. As already mentioned, parametric methods are inherently limited by class distribution assumptions. In the next chapter we will review the more attractive non-parametric machine learning supervised classification methods, such as the Artificial Neural Network, that have been shown in remote sensing experiments to produce high quality classification results.

Chapter 3

Non-Parametric Image Classification Methods

3.1 Introduction

Since parametric methods rely on class distribution assumptions, they are limited to perform well only on those datasets that closely approximate these distributions. To deal with this issue, amongst others such as non-linear separable classes, the pattern recognition community was active in development of new intelligent systems in the 1970s and 1980s. Of particular interest was the Artificial Neural Network (ANN), and more recently the Support Vector Machine (SVM) and Decision Trees. These systems are non-parametric, for their ability to make decisions without making any statistical assumptions. The decision making is based rather on the fundamental assumption that an unknown pattern can be recognised through repeated examples of that pattern (Jin and Geman, 2006). The concept is that 'learning' or 'intelligence' can be obtained through these examples. They have thus been termed 'computational intelligence', otherwise known as 'machine learning' systems. Since the mid-90's there has been a strong interest in the remote sensing community in applying these algorithms to remote sensing data classification. Research has indicated that non-parametric classifiers may provide better classification results than parametric classifiers (Foody, 2002; Kavzoglu and Mather, 2003; Huang et al., 2002; Pal and Mather, 2005). A selection of popular machine learning techniques will now be discussed along with their application in remote sensing.

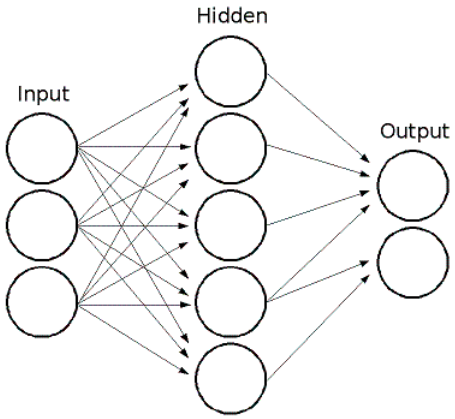


Figure 3.1: A typical multilayer perceptron neural network.
The multilayer perceptron consists of input layers of nodes (e.g. spectral values), one or more sets of hidden layers, and an output layer (e.g. classes in the classification system). A complex set of linkages exist between nodes. Neurons in the hidden layers consist of a set of functions.

3.2 Artificial Neural Networks

3.2.1 Introduction

An Artificial Neural Network (ANN) is a learning machine that attempts to imitate the workings of a human brain. ANNs do not require statistical assumptions about the distribution of data. They have the ability to develop their own input / output discriminant relations from training data (Pacifici et al., 2008). Thus, unlike the maximum likelihood classifier they are robust in recognizing patterns from non-Gaussian distributed data. ANNs have been well established in speech and handwriting recognition, as well image analysis (Tso and Mather, 2001). A detailed review of ANNs can be found in (Bishop, 1996).

3.2.2 The Multilayer Perceptron

One of the most widely used neural network models is the multilayer perceptron. It consists of input layers, one or more sets of hidden layers, and an output layer (see figure 3.1). The layers are made up of nodes, referred to as neurons in analogy with biological neurons. In remote sensing classification the input neurons could represent features such as spectral values. Each output node could represent a class in the classification system. A complex set of linkages exist between nodes, representing weights. Neurons in the hidden layer consist of a set of functions, normally sigmoids and / or summations. The number of hidden neurons is arbitrary.

3.2.3 Back-propagation

The most commonly used algorithm to train a multilayer perceptron is 'back-propagation'. The learning procedure is as follows. Input variables are passed through the system and evaluated for error against the output variables. The results of this evaluation are then passed back into the system, and the weights are adjusted accordingly in an attempt to reduce the error. This procedure iterates several times until the error is sufficiently reduced. The weights that are adjusted give significance to the input variables. In other words, in each iterative test against the ground truth output layer, the system is corrected, or taught, much the same way as a parent rebukes a baby every time it does a number of different things wrong, and eventually learns the correct way of living in general.

3.2.4 Practical use of the ANN

A disadvantage of the ANN is that the user has to determine its architecture, such as the number of hidden neurons as well as parameters such as learning rate. These parameters significantly affect the training time and performance of an ANN. Furthermore, no blueprint exists on how to define parameters, but rather rules of thumb. Another disadvantage is that the system is a black box. There is no easy way to trace back and pin point the cause of inaccurate results. Successful usage of an ANN relies on heuristic procedures.

3.2.5 ANNs applied to remote sensing data classification

In the past two decades ANNs have been increasingly used for remote sensing data classification, due to the belief that they can outperform statistical classifiers like the maximum likelihood. Benediktsson et al. (1990) provides a comparison between statistical classifiers and the ANN in classifying multisource remote sensing data. Results show that the two different approaches have unique advantages and disadvantages. Paola and Schowengerdt (1995) provides a literature survey of back propagation neural networks for multispectral image classification. They conclude that although a neural network has several unique characteristics, it will be a useful remote sensing tool only if made easier to use, faster, and more predictable. In (Kanellopoulos and Wilkinson, 1997) an experimental investigation of ANNs applied to the classification of satellite imagery is presented. Attempts are made to draw conclusions about 'best practice' techniques to optimize network training (e.g. choice of network architecture) and overall classification performance. They conclude that neural networks can now be used reliably and with much confidence for routine operational requirements in remote sensing. In (Kavzoglu and Mather, 2003) a comparison is made between ANNs and the maximum likelihood classifier for remote sensing land cover classification. Results show

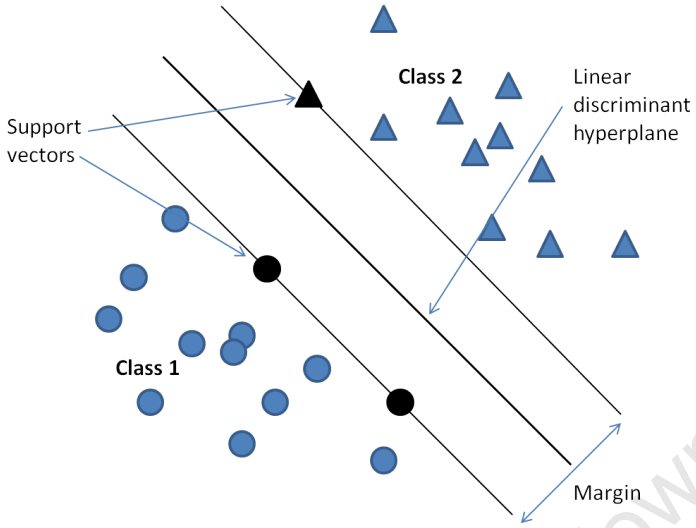


Figure 3.2: Support Vector Machine: The linearly separable case (Tso and Mather, 2009).
To determine the decision boundary in a feature space, a linear discriminant hyperplane is placed midway between class clusters. The discriminant is placed in the position where the maximum perpendicular distance (referred to as the margin) exists between the discriminant and a subset of points (the support vectors) on the edge of class clusters.

that the ANN produces higher classification accuracies than the maximum likelihood classifier when used with the recommended settings in the study.

3.3 Support Vector Machines

3.3.1 Introduction

Support Vector Machines (SVMs) offer a theoretically superior machine learning methodology (Tso and Mather, 2009). We can estimate their emergence in the late 70s (Vapnik, 1979), but they have not received significant attention until recent years. The success of the SVM is merited to their unique ability to minimize the so-called structural risk, or classification errors when solving the classification problem (Tso and Mather, 2009). With the maximum likelihood classifier, the minimization of classification errors is determined directly from the statistical distribution of training samples. The SVM, however, minimizes the probability of misclassifying previously unobserved data points drawn randomly from a fixed but unknown probability distribution (Vapnik, 1995).

3.3.2 Error minimization

The SVM structural risk minimization problem is solved with complex mathematics but a simple geometric interpretation (see figure 3.2). The Support Vector Machine determines the decision boundary between classes in a feature space by placing a linear discriminant hyperplane midway between the class clusters. The position of this discriminant requires the selection of a subset of training samples (the support vectors) that best describes the boundary between two classes. This subset is normally located along the edges of the class clusters. The discriminant is placed in the position where the maximum perpendicular distance (referred to as the margin) exists between the discriminant and these subset points. If classes are non-linearly separable, a kernel expansion is used, in which the feature space is projected onto a higher dimensional space where separation of classes becomes linear.

3.3.3 Advantages

A strong advantage of the SVM is that high accuracy can be achieved with a small number of training samples (see Foody and Mather, 2004). Indeed in the field of remote sensing this is often the case due to data availability and cost of data base production (Inglada, 2007). Another advantage is that SVMs can be used efficiently with high dimensional datasets (see Pal and Mather, 2005). Thus information loss through data reduction is prevented. A drawback of SVMs is that the original version was designed to solve binary classification. Most remote sensing scenarios deal with multiple classes. Multiclass algorithms have been proposed, including one-against-one, one-against-others, and directed acyclic graph (DAG) strategies (Tso and Mather, 2009). Another disadvantage is that required user-defined parameters have a strong influence on its performance.

3.3.4 SVMs applied to remote sensing data classification

The implementation of SVMs to perform remote sensing data classification is gradually expanding. In recent studies, SVMs were compared to other classification methods, such as ANNs, Nearest Neighbor (NN), Maximum Likelihood and Decision Tree classifiers for remote sensing imagery, and have performed as well, if not surpassed all of them in robustness and accuracy. Huang et al. (2002) compares the SVM to three other classifiers in a land cover classification experiment on a low resolution Thematic Mapper (TM) image. The three classifiers are the maximum likelihood, decision tree, and ANN. They conclude that SVMs provide higher overall accuracies than any other classifier. Pal and Mather (2005) compares SVMs with ANN and maximum likelihood methods for land cover classification of multispectral and hyperspectral images. Their results show that SVM achieves the highest accuracy. Results also show that an SVM can be used with high dimensional

data and small training sets. Foody and Mather (2004) show efficient and accurate SVM classification of multispectral satellite data with a small number of intelligently selected training samples. Tzotsos (2008) applies SVM classification to a segmentation of a Landsat TM image using the Definiens eCognition software (Benz et al., 2004). Spectral, shape and textural features were used. The results show that the SVM provides slightly higher classification accuracy than the standard NN classifier in eCognition (NN confusion matrix accuracy: 84.1%; SVM confusion matrix accuracy: 86%).

3.4 Fuzzy Set Theory

3.4.1 Introduction

The development of Fuzzy Set Theory was inspired by the fact that the decisions we encounter every day are often uncertain, or fuzzy, as opposed to deterministic (Tso and Mather, 2009). For instance, the concepts *cold*, *warm* and *hot* contain a level of subjectivity. One person's *hot* may overlap with another person's *warm*. They cannot be deterministically specified.

A similar problem occurs in remote sensing scenarios. In a coarse resolution image a pixel may lie on the border of water and land classes. Both classes are thus present in the pixel, and it is difficult to classify the pixel deterministically. This type of pixel has been termed *mixed*. The standard classification techniques in chapter 2 do not provide a good mechanism for coping with such uncertainty. Fuzzy Set Theory offers an approach that acknowledges the problem of uncertainty from the start in an ambiguous environment. The fuzzy concept has been adopted in numerous fields e.g. fuzzy logic control, process control, management and decision making, and operations research (Nedeljkovic, 2006). It has also been widely used for dealing with classification problems. Fuzzy-based classifiers are becoming increasingly popular in remote sensing (Tso and Mather, 2009).

3.4.2 Concept

With crisp sets, two choices are available: $[0, 1]$. For example, a particular day may be classified either as *cold*, *warm* or *hot*. If it is classified as *warm*, the set will read *cold*= 0, *warm*= 1, *hot*= 0. With fuzzy sets, we have the concept of partial membership in the form of a membership function m, where $0 \leq m \leq 1$. On the same day, the fuzzy set may read *cold*= 0.3, *warm*= 0.6, *hot*= 0.4. The three membership values need not sum to 1, which is partly how fuzzy theory differs from probability theory. While probability theory provides the probability of an event occurring, fuzzy theory provides the possibility of an event occurring. The fuzzy concept allows flexibility in decision making.

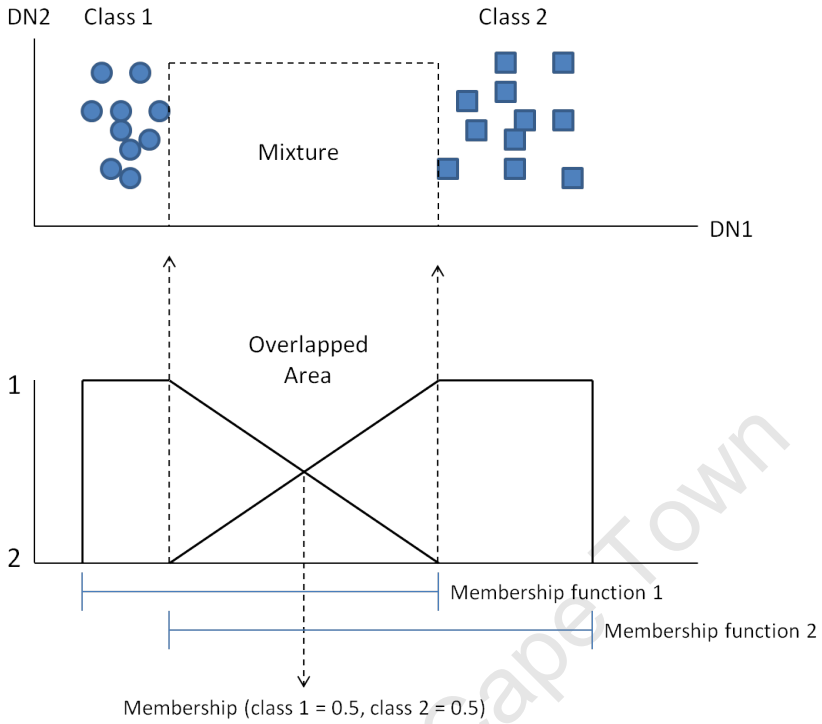


Figure 3.3: Fuzzy rule-base image classification (Tso and Mather, 2001). This illustration shows a fuzzy set for DN1 (Digital Number 1). 'Mixture' refers to a region in the feature space which is fuzzy. If an unknown pixel has features that fall within this fuzzy region, it is uncertain whether the pixel should be allocated to Class 1 or Class 2. Thus a fuzzy membership would be allocated to the unknown point, based on two membership functions, one for each class. If, for instance, the point falls at the intersection of the two overlapping membership functions, a membership of [Class 1 = 0.5, Class 2 = 0.5] would be allocated to the point.

3.4.3 Fuzzy rule-based remote sensing classification

The fuzzy logic concept is well suited to dealing with uncertainties in the image classification task, such as uncertainty in sensor measurements, vague (linguistic) class descriptions, and class mixtures due to limited resolution (Benz et al., 2004). The remote sensing community has been increasingly interested in exploiting fuzzy theory to improve image classification. A popular trend is to use fuzzy rules. The basic concept is as follows: For each unknown pixel p_i to be classified into one of n classes, a membership grade $\alpha_{ci,pi}$ is associated with each class c_i, such that an nc dimensional tuple of membership grades is established for each p_i, as follows: $f_{pi} = [\alpha_{c1,pi}, \alpha_{c2,pi}, ..., \alpha_{cn,pi}]$ (Benz et al., 2004). This tuple is useful as it contains insight into overall

reliability and class mixture. For instance, it can provide promising input to current and future remote sensing systems with multi-sensor sources and ancillary data. For a crisp classification problem, the membership grade with the highest value, $\alpha_{cj,pi}$, is considered, and the unknown pixel p_i is allocated to class c_j.

In more detail, a fuzzy partitioning of each feature p_i takes place during the training phase of a classification procedure. In other words, for each pixel feature a fuzzy set is established (see figure 3.3 which illustrates the fuzzy set of Digital Number 1 (DN1)). The size and type of a fuzzy set is controlled by a user defined membership function, normally in the shape of a trapezoid. The shape of the membership function determines the transition between full member and non-member, i.e. for a crisp set the membership function would be rectangular. The choice of the membership function is crucial and determines the performance of the classification system (Benz et al., 2004).

In order to classify an unknown pixel p_i, for each feature $\omega_{i,pi}$ measured in p_i, a membership degree $m_{ci,\omega i,pi}$, $0 \leq m_{ci,\omega i,pi} \leq 1$, is computed based on the fuzzy set for feature ω_i. Hence, instead of combining actual feature values to make a decision in a feature space, as is the case with crisp classification systems, fuzzy sets on these feature values are combined to make the decision. All further calculations in the classification system are based on the set of membership degrees $M_{ci,pi} = [m_{ci,\omega 1,pi}, m_{ci,\omega 2,pi}, ..., m_{ci,\omega \eta,pi}]$ where η is the total number of features. The inferencing stage works by constructing a set of *if-then* fuzzy rules (e.g. *if* feature $\omega_{i,pi}$ is a member of fuzzy set b_{cj} then pixel p_i is a member of class c_j).

An issue with fuzzy rule-base classifiers is that a large number of input features requires a higher number of fuzzy rules, which can become complex (Tso and Mather, 2001). This issue can potentially be solved with data reduction techniques (see section 2.6)

3.4.4 Fuzzy Set Theory applied to remote sensing data classification

Foody (1996) shows the importance of recognizing and accommodating for fuzziness of land cover classes. It is shown that fuzzy representations produce more accurate land cover classifications than crisp classifiers. In particular, results from a fuzzy ANN and statistical fuzzy c-means algorithm produced highest land cover classification accuracies.

Bardossy and Samaniego (2002) investigates the applicability of fuzzy rule-base modeling to classify a Landsat TM scene. It is shown that their proposed method with only 9 rules for four land cover classes produces slightly higher classification accuracies than the maximum likelihood classifier.

Shackelford and Davis (2003) investigates the use of a fuzzy theory in improving the urban land cover classification accuracy of high-resolution

satellite imagery of urban and suburban areas. Firstly, the imagery is classified using just spectral features with a standard maximum likelihood approach. Significant mis-classifications are observed at spectrally similar road and building classes. Secondly, texture features as well as a length-to-width shape features are included in the system in order to improve discrimination between spectrally similar classes. This produces higher urban land cover accuracies. Finally, a hierarchical fuzzy classification approach is investigated that makes use of both spectral and spatial features. This is shown to produce classification accuracies that are 8% to 11% higher than those from the standard maximum-likelihood approach.

Laha et al. (2006) proposes a contextual fuzzy rule-based classification system. For an unknown pixel to classify, information from fuzzy-generated possibilistic class labels of its neighbouring pixels is aggregated to make a final decision. The proposed method is tested with two Landsat-TM satellite images. Results are compared with the Markov Random Field (MRF) (see section 4.2.1) contextual classification method and found to perform consistently better.

3.5 Decision Trees

3.5.1 Introduction

A decision tree is a top-down hierarchical classification technique. The general purpose is to provide a comprehensive understanding of the relationships between objects at different scales of observation or levels of detail. The main attractiveness of a decision tree is that it can be viewed as a white box, in contrast to the ANN (Tso and Mather, 2009). It is easier to interpret and understand the relationships between inputs and outputs. A decision tree has the form of an upside down tree (figure 3.4). It is composed of a root node, interior nodes, and terminal nodes called leaf nodes. The root and interior nodes represent decision stages. The leaf nodes correspond to the final classification. The classification process works as follows: starting from the root node and ending at the terminal node, a decision is made at each non-leaf node. This decision determines the path to be followed.

An extension of the decision tree is the random forest, which consists of many decision trees. A random forest makes an optimal decision about class labels based on the classification output vote of individual trees (Breiman, 2001).

3.5.2 A decision tree's design

A decision tree's performance depends significantly on the nature of decisions and sequence of attributes in the tree (Tso and Mather, 2009). The design of a decision tree is thus important. One design approach is to rely solely

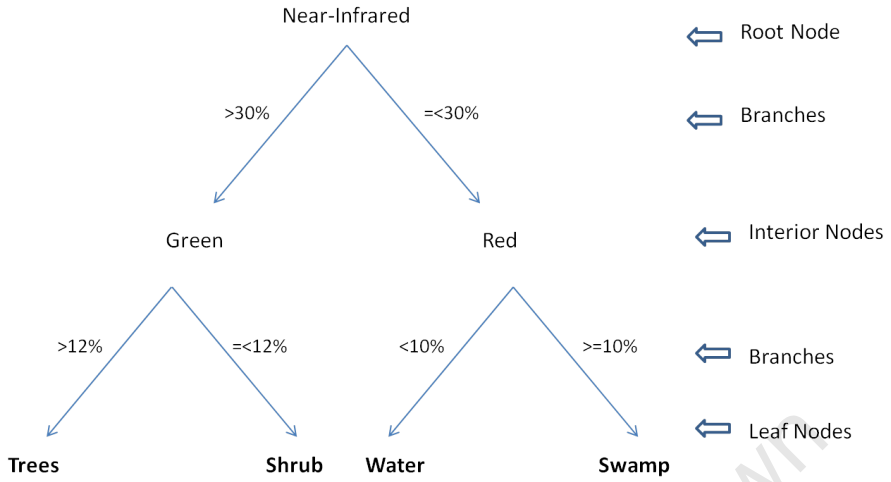


Figure 3.4: A simple hierarchical decision tree classifier (Tso and Mather, 2009).
A decision tree is composed of a root node ('Near-Infrared'), interior nodes ('Green', 'Red') and leaf nodes ('Trees', 'Water' 'Shrub', 'Swamp'). A decision is made at each non-leaf node to determine the path to be followed.

on an analyst's knowledge to manually separate classes in a hierarchical fashion. This requires statistics for all classes to be computed. Estimates of decision boundaries are then derived. This approach can be time consuming, and may not provide satisfactory results particularly for a large number of classes (Tso and Mather, 2009). A currently more popular approach is to develop automatic methods to arrange the structure of the tree and select features. These methods attempt to construct optimal designs by minimizing classification accuracy measures. They work by recursively splitting nodes into leaf nodes. At each node-splitting stage, the most effective feature is selected according to a measure of classification accuracy, or a simple statistical test. Automatically generated decision trees are often complex, having long and uneven paths. Pruning is thus often required, where a subtree is replaced with a leaf node.

3.5.3 Use of decision trees in remote sensing classification

The use of decision trees in remote sensing data classification has gained popularity in recent years, and has shown to be equal in classification performance, if not surpassed, state-of-the-art classifiers. Hansen et al. (1996) apply decision trees to land cover classification, and many authors have followed their lead (Muchoney et al., 2000; Friedl et al., 1999; Pal and Mather, 2003). German et al. (1999) compare the performance of the Minimum Distance, the Maximum Likelihood, and a multilayer perceptron ANN on land

cover classification of Landsat TM data. The performance is measured in terms of their learning ability, classification accuracy, and computational speed. They conclude that the decision tree classifier is the best all-round choice. Pal (2005) compares a random forest and a SVM for land cover classification of Landsat Enhanced Thematic Mapper Plus (ETM+) data with 7 classes. The two classifiers perform equally in terms of training time and classification accuracy, but the random forest requires less model parameters. Pal and Mather (2006) show that a decision tree can effectively be used for feature selection.

3.6 Conclusion

A selection of popular machine learning techniques has been reviewed, along with their usage in remote sensing. These studies have demonstrated the power of machine learning, by showing how they can out-perform traditional statistical methods in classifying data. This is largely due to their non-parametric nature, i.e., the ability to make predictions independent of class distribution assumptions. The Artificial Neural Network is a classical non-parametric machine, that has been popular in the remote sensing community especially in the 1990s, and has produced good results. More recently however, the theoretically superior Support Vector Machine and the hierarchical decision tree have been shown to produce even higher quality results (Huang et al., 2002; Pal and Mather, 2005; German et al., 1999). In this thesis the Support Vector Machine has be chosen to perform classification of urban scene images, due its theoretically superior nature, and because it outperforms the decision tree as shown in (Huang et al., 2002).

Chapter 4

Region-based Image Classification

4.1 Introduction

Pixel-based techniques are still in widespread use, but have drawbacks. In the last decade the spatial resolution of remote sensing imagery has increased significantly. For high resolution imagery, objects of interest are often significantly larger than pixel size. Intraclass spectral variation increases and interclass variation decreases (Aksoy and Akçay, 2005). This is especially evident in urban scenes where different man-made objects are made up of materials with similar spectral signatures (Aplin and Smith, 2008). This can lead to misclassified pixels, resulting in a 'salt and pepper' effect. A pixel in the middle of a lake, for example, might be misclassified as 'building' since a pixel-based classifier is based solely on the individual pixel's properties. There has thus been a strong move towards developing classification methods that are spatially aware of neighbourhood pixels. These are often termed 'contextual' classifiers in the literature. Indeed they introduce a degree of low-level context, but not the true high-level context that this dissertation seeks to address. Popular methods will be discussed in this chapter.

4.2 Exploiting Bayes rule

With maximum likelihood pixel-based techniques (see section 2.4.1.2) a pixel is assigned a probability of belonging to a certain class based solely on its own feature distribution. An option for introducing an awareness of neighbouring pixels is to exploit Bayes Rule to classify a pixel based on the influence of its own feature distribution as well as the feature distribution of neighbouring pixels (Tso and Mather, 2009). Recapping Equation 2.1, Bayes Rule is defined as follows:

$$P(c_i|x) = P(x|c_i)P(c_i) \tag{4.1}$$

where x = data, c = unknown classes, $P(c_i|x)$ is the conditional probability of c_i given x, $P(x|c_i)$ is the conditional probability of x given c_i, and $P(c_i)$ is the *prior probability* that c_i occurs in the image.

The maximum likelihood technique models just $P(x|c_i)$, the data distribution of a single pixel. What is required is $P(c_i)$, the prior probability of a pixel being a certain class. The $P(c_i)$ can be derived from the data distribution of neighbouring pixels. For example, a pixel in the middle of a lake would have a prior probability of being a 'lake' pixel based on its surrounding pixels.

4.2.1 Use of the Markov Random Field

A useful tool for characterizing $P(c_i)$ is the Markov Random Field (MRF). The MRF derives $P(c_i)$ from contextually dependent patterns as the sum of local contributions in the form of suitable potential functions chosen according to some useful statistical criteria (Jensen, 2005). We can use the Maximum A Posterior (MAP) for the statistical criteria, which is aimed at maximizing the posterior probability $P(c_i, x)$ over x. In (Tso and Olsen, 2005) the MRF is successfully exploited for high resolution image classification, and shows an improvement in results over pixel-wise classification.

4.3 Object Based Image Analysis

4.3.1 Introduction

A currently popular technique is to merge neighbouring and homogeneous pixels before classification begins. These homogeneous regions, as opposed to individual pixels, are then classified. This technique is formally known as ECHO (Extraction and Classification of Homogeneous Objects), its emergence probably credited to Kettig and Landgrebe (1976). A modern term is Object Based Image Analysis (OBIA), 'object' referring to a homogeneous region. The motivation behind OBIA is that the success of human interpretation is based on analysis of homogeneous regions of pixels, as opposed to individual pixels. To merge pixels, a segmentation of the image is normally performed.

4.3.2 Image segmentation

Image segmentation is the partitioning of an image into disjoint homogeneous regions (see figure 4.1). The basic concept is to merge image elements (pixels or pixel regions) according to homogeneity parameters (Schiewe, 2002). A popular strategy is bottom-up region growing where a seed region is selected

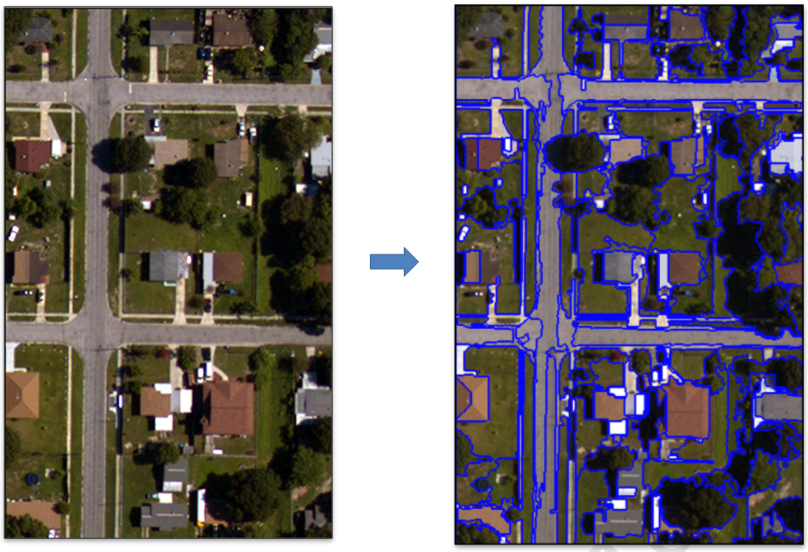


Figure 4.1: Image segmentation.
The figure shows a typical segmentation of an urban scene, performed in the eCognition software package (Benz et al., 2004). The input scene is on the left and the segmented scene is on the right.

and neighbouring pixels / pixel groups are tested against the seed region. These neighbouring pixels are added to the seed if similar. Another technique is region splitting (top-down), which starts with the entire scene, and splits regions until an appropriate segmentation is obtained.

A framework for testing heterogeneity between two image elements is described in (Schiewe, 2002) as follows:

Given two neighbouring elements A and B with features $f_{A,i}$ and $f_{B,i}$, $(i = 1, \ldots, n)$, we can derive a homogeneity measure by computing the Euclidean distance between $f_{A,i}$ and $f_{B,i}$. An option is to assign a weight g_i to each feature based on its importance in defining heterogeneity. The corresponding heterogeneity measure $\triangle h$ is given by:

$$\triangle h = \sqrt[+]{\sum_{i=1}^{n} g_i (f_{A,i} - f_{B,i})^2} \tag{4.2}$$

The measure $\triangle h$ is compared with a threshold in order to decide whether A and B should be merged or not. The threshold controls the size and number of segments in a final partition.

Further constraints concerning neighbourhood and similarity can be used:

1. In the simplest case, A accepts B if the homogeneity measure is below the given threshold.

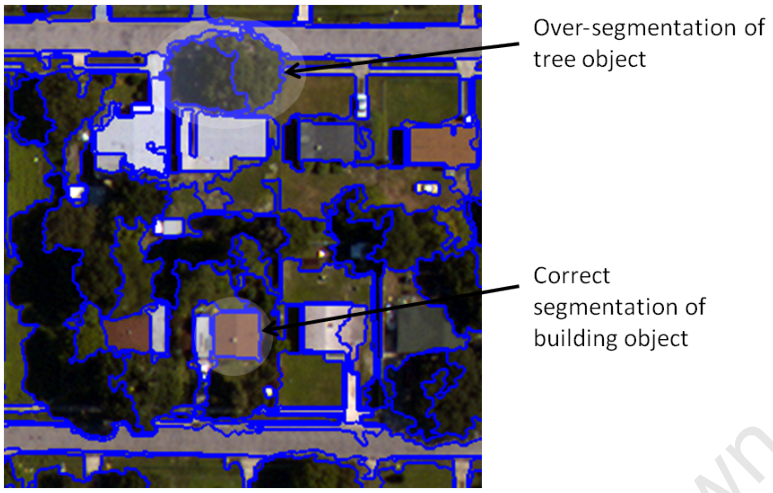


Figure 4.2: Object scale variation.

In this particular segmentation, tree objects are over-segmented whereas building objects are correctly segmented. Thus a different segmentation scale is required for different object types.

2. A may accept just that neighbouring element B which fulfills the homogeneity criterion best.

3. Alternatively, an element C is attached to A (which is attached to B) only if B and C, as well as A and C, are similar.

4.3.3 Object scale variation

An issue encountered in urban scene segmentation is that different objects require different scales of segmentation. A segmentation 'scale' is correlated with the size of the homogeneity threshold mentioned above. A segmentation at a certain scale may, for instance, properly define building regions, but improperly define tree regions (see figure 4.2). Each object has its own spectral homogeneity, and therefore its own threshold setting. If the threshold is smaller or larger, an over or under-segmentation respectively will be produced.

Bruzzone and Carlin (2006) performs a pixel-wise multi-scale classification of a high resolution image of an urban scene. For each pixel, context is considered in the following manner. Multiple scene segmentations are performed at different arbitrary scales. Thus an object will exist at multiple scales. The smallest scale is the pixel and the highest an object such as a building. Intermediate levels could be a roof face and a section of a roof face. At each scale, a selection of features is made. At the pixel level and small segment level, only spectral features can be considered. For larger segments that meaningfully represent objects, such as a roof face or roof,

geometric and relational features are selected, such as *area of object*, *number of sub objects*, and various shape features. By taking into consideration the local neighbourhood of each pixel, a multidimensional feature matrix is built, where each row could represent an object scale, and each column the list of features associated with that scale. An SVM (see section 3.3) is used to classify each pixel based on its feature matrix.

This is a promising approach for introducing multi-scale awareness of objects. However, it is heuristic in the sense that the system has no clear recognition of the correct object scale, but rather a guess via a combination of different scales with associative exhaustive feature sets.

In (Baatz and Schape, 2000) a segmentation routine is presented that was successfully implemented in a software package called eCognition (Benz et al., 2004), developed by Definiens. The eCognition environment is currently the most commonly used OBIA system (Lu and Weng, 2007). eCognition offers a multi-scale segmentation procedure, where several segmentations of the same image can exist at different scales. A drawback of the eCognition multi-scale segmentation approach is that scale parameters have to be manually defined by the user. Ideally what is required is a segmentation routine that chooses an appropriate threshold value for the objects of interest in an unsupervised fashion, based on the spectral / spatial characteristics of that object (automated scale selection).

Chang and Chen (2008) proposes an unsupervised scale selection segmentation routine for urban scene analysis. The method starts with an over-segmentation and then increments toward an under-segmentation. After each increment, for a particular object the change in segment edge density calculated. The scale where there is a maximum change in edge density is chosen for that object. The rationale behind this is that as soon as a segment area corresponds to a real world object, a sudden change in edge density is known to occur.

Taubenbock and Esch (2005) similarly starts with an over-segmentation and increments toward an under-segmentation. The shapes of segments are iteratively optimized according to a rule base until significant structures are realized.

Akcay and Aksoy (2008) produces segmentations at multiple scales for each individual image band based on the application of morphological opening and closing operations. Each segment at the different scales is evaluated as a candidate for a meaningful structure. This evaluation is based on a combined measure of spectral homogeneity (based on variances of spectral features) and neighbourhood connectivity (based on the sizes of connected components in the multi-scale hierarchy). Based on the observation that different urban objects appear more clearly in different spectral bands of an image, the segment candidates are grouped to find objects of interest.

4.3.4 Availability of new uncorrelated features

An ideal segmentation map contains segments of real world objects of interest. The availability of image objects provides a new paradigm in image interpretation. An existing classification algorithm can be employed to classify image objects, as opposed to image pixels. Image objects offer a large set of new uncorrelated features, such as the following:

1. Shape features: e.g. *area*, *compactness*, *length-to-width ratio*.
2. Textural features: e.g. Features based on analyzing pairs of pixels with similar grey values in a GLCM (Grey level Co-occurrence Matrix).
3. Spectral statistics (e.g. standard deviation - stdev - of all *brightness* values within an object region).

Shape can be a powerfully discriminating feature, especially in the case of an urban scene (Inglada, 2007). Urban features are often more robustly characterised by shape than by colour. Many different man-made object types are made of similar materials and thus have similar reflectance properties. For example, roads can have a similar colour to buildings but are normally different in shape. Shape features are independent of sensor characteristics and illumination conditions.

Thus the feature vector used to classify an image object may be significantly larger than a pixel-based feature vector, and a feature selection procedure is normally required (see section 2.6). Tzotsos (2008) points out that the Support Vector Machine (SVM) is a good choice for discriminating OBIA segments, since it performs well under a high dimensional feature space.

It is worth mentioning that the effectiveness of these object features will be heavily influenced by the quality of segmentation results. If a building, for example, is slightly over-segmented, and the segment contains spurious regions, it will no longer closely resemble the shape of a building. Thus segmentation quality is pivotal to the performance of OBIA .

4.3.5 Urban scene OBIA studies

From around the year 2000 there has been a sharp increase in the use of segmentation techniques to perform OBIA. The results in many studies have shown a significant improvement in image classification accuracy over pixel-based techniques. A comprehensive OBIA literature review can be found in (Blaschke, 2009). The following studies deal specifically with urban scene object extraction.

Shackelford and Davis (2003) investigates the use of a fuzzy theory in improving the urban land cover classification accuracy of high-resolution

satellite imagery of urban and suburban areas. The imagery is initially classified using just spectral features with a standard maximum likelihood approach. Significant mis-classifications are observed at spectrally similar road and building classes. Secondly, texture features as well as a *length-to-width* shape feature is included in the system in order to improve discrimination between spectrally similar classes. This produces higher urban land cover accuracies.

Mo et al. (2007) segments and classifies a high-resolution QuickBird multispectral image of an urban scene, also using a standard maximum likelihood approach. A feature selection was carried out on spectral, shape, and texture features. The classification results were found to be consistent with visual interpretation results. In addition, the results were compared with a pixel-based technique and found to be superior.

Su et al. (2008) shows how textural and local spatial features of segmented objects can be utilized to improve the classification of QuickBird imagery. Results based solely on spectral features are compared to those based on textural and spatial features, where a classification accuracy improvement of up to 7% is observed with the latter.

4.4 Conclusion

In this chapter we have discussed region-based classification techniques, including use of the MRF and the currently popular OBIA, which is based on image segmentation. These methods have been invented because of the current availability of high-resolution images. They essentially introduce an awareness of the spatial neighbourhood of pixels into the classification process, in order to emulate human visual perception. A human interpreter does not look at individual pixels when detecting objects in an image, but rather at groups of homogeneous pixels. Indeed the above studies have proved the effectiveness of object-based methods over pixel-based methods. An object-based approach is thus chosen in this research to perform bottom-up urban object classification.

In the last two chapters the state-of-the-art in image classification has been discussed. From chapters 2 and 3 we concluded that machine learning, and more specifically the Support Vector Machine (SVM), is the pattern recognition technique of choice. Thus a classification system assumed to be effective in urban object classification is SVM recognition of image objects, as proposed in (Tzotsos, 2008). This is the adopted bottom-up technique used in this thesis (see section 1.6.4).

Object-based image classification introduces a degree of low-level context by introducing an awareness of spatial neighbourhood. However, object-based techniques are based solely on sensory measurement and are thus inherently limited. This dissertation addresses the use of higher level context

in the form of external knowledge to produce high quality object classification results. This type of analysis can be found in image understanding theory, which is the subject of discussion in the next chapter.

Chapter 5

Image Understanding

5.1 Introduction

The performance of both 'pixel-based' (Chapter 2) and 'object-based' (Chapter 4) image classification techniques is inherently limited by the available sensory content (whether it be colour, shape or texture) each pixel or pixel region contains. Many different object types in an urban scene can be similar in shape and colour, for example road and pavement. Thus even with object-based techniques, mis-classification of regions is possible.

Human visual perception relies to a large extent on a recognition of context. In figure 5.1 (a) the blue square object could be a number of urban objects, including a building, a swimming pool, or a road segment. However, when viewed within its context (figure 5.1 (b)), it is more likely to be a building. Human visual perception works not only by looking at the colour, shape and texture of individual objects. Object detection decisions are heavily influenced by a recognition that a scene is e.g. a middle-class residential urban scene in Cape Town, South Africa. The interpreter has a pre-built pattern, or an understanding, of what a scene should look like before he interprets the image. He then utilizes this model to obtain the most likely description of the scene based on the image content.

5.1.1 An understanding through spatial relationships and semantics.

We will now take a more in-depth analysis of the nature of this 'understanding', in order to formalize it. Visual interpretation of any scene relies on inherent laws on the spatial arrangements or semantics of objects in a scene. These laws are different depending on the type of scene. Lets take a scene of a simple kitchen room, for instance. We hold intrinsic knowledge that coffee cups should be *on* tables, chairs *near* tables, table tops *parallel to* floor. If we saw a large object on the table that we were unsure of, we would unlikely classify it as a chair. This has similarities to the concept of semantics.

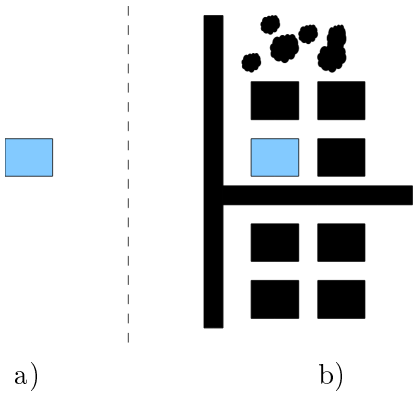


Figure 5.1: Classifying an object in an urban scene through context. The light blue object can be classified: a) without context: classification is based solely on low-level colour, texture and shape features. b) With context: classification is based on low-level features as well as higher level contextual features.

A house can decomposes into walls, walls into bricks, bricks into a certain colour or texture. The recognition of this house is not solely based on the colour and texture of an individual brick, but the combination of all elements - the mosaic of bricks together represent a wall, and the position and angles of walls together represent a house. We could then easily classify an object on the house as being a brick because it falls into the context of its surroundings. Likewise when looking at an urban scene, a city can decompose into residential area, CBD, heritage sites, a residential area into road networks, suburb blocks and parks, a suburb block into clusters of houses and trees, a cluster of houses into an individual house, and an individual house into roof planes, edges, walls. The recognition of a building will be largely influenced by an awareness of its semantic entities. In recent years there has been attempts to emulate this concept to improve image analysis, especially in applications such as content-based image retrieval and multimedia applications, and to a lesser extent urban remote sensing. This methodology has been widely referred to as 'image understanding'.

5.2 General Framework for Aerial Image Understanding

5.2.1 Concept

We will now look at a general framework for aerial Image Understanding (IU), adapted from (Matsuyama and Hwang, 1990). The fundamental concept in IU is that observed features are not simply labeled into a predefined set of classes, as is the case with the classical image classification task. A

complete and idealized description of the scene is constructed, even if it is only partially depicted by the initial observed features. The information for this idealized description is induced from what we will call a 'contextual model'. A contextual model is a template for what a scene should look like. It could consist of the various spatial relation laws / semantics mentioned above, and is often in the form of a hierarchical model, semantic network, or contextual feature space.

To recap the standard image classification routine in a theoretical manner, features in the form of object regions are extracted from an image by segmentation. These features are given more meaning by classification. The classification of a feature is normally based on the properties of that individual feature. This can be referred to as a *bottom-up* analysis. In IU, these extracted features, whether they are a segmentation or classification result, are then matched / tested against the contextual model. The goal is to reach a final scene description that matches the contextual model as closely as possible. The contextual model complements missing but necessary information. Thus an IU System is designed to produce good quality final results even with poor quality bottom-up results.

5.2.2 Scene-matching

Different contextual models are required for different scene types. Considering urban scenes, a contextual model could consist of a single generic model which is suitable for any urban scene anywhere in the world. Such a model is limited because urban scenes differ significantly according to socio-economic, cultural and land-use planning differences. Different contextual models ought to be defined that are tailored to a specific scene type (e.g. Southern African industrial, western city commercial, London CBD, Paris CBD). Thus the first important step in an image understanding procedure is to choose an appropriate contextual model. This choice can be made manually by prior knowledge of the scene type. If we knew that the scene to classify is an industrial area in Johannesberg, the analyst simply chooses the 'Southern African Industrial' model. What would happen if the scene to classify were made up of several different land-uses? In this case the classification system makes the choice for each target land-use area based on the initial bottom-up results. For this to happen, a predefined set of contextual models is required. For each land-use region, the land-use model that has greatest similarity to the bottom-up features is selected. This is known as 'scene-matching'. This is the same as land-use classification for the urban scene case.

5.2.3 Top-down analysis

With an appropriate contextual model, the top-down inferencing stage can begin. In this stage the bottom-up results are improved using the chosen

contextual model. Through the constraints / parameter range imposed by the contextual model, pruning and instantiation can take place. Pruning means to delete objects that are inconsistent with the model. Instantiation means to create a new object if it is required to fill a gap in the model. We will look at an example to clarify this concept. Lets say an agricultural field was detected near buildings in a scene during the bottom-up phase. In the top-down phase, the contextual model has a law that says, since this is a scene of a residential area, vegetated patches near buildings are likely to be 'lawn'. The detection of the agricultural field is thus inconsistent with the model, and it is deleted (pruning). A new 'lawn' object is created (instantiation) so that consistency with the model is reached.

In a top-down phase, reasoning has to be performed under incomplete data, since automated feature extraction results are never perfect. Many systems thus rely on the well established machine learning techniques that deal with reasoning under uncertainty, such as Bayesian statistics and Fuzzy logic.

Classification-based-segmentation

A fundamental issue with IU systems is that the initial bottom-up routine is imperfect. Indeed, a segmentation routine can never produce perfect results (Matsuyama and Hwang, 1990). The goal in IU is to first establish a global context or likely description based on an imperfect segmentation, and then to recover the imperfection. Benz et al. (2004) describes an iterative classification strategy for solving this problem, as follows: An initial, preliminary classification is performed. The classifier has now built up some knowledge about the scene, albeit with uncertainty. A top-down inference can then take place using a contextual model to improve the initial interpretation. A 3rd classification may then be performed, in fact as many as it takes to build up sufficient knowledge. Since the success of these classifications is completely reliant on a successful initial segmentation, a good technique is to perform a new and improved segmentation after every classification. Additional input to these intermediate segmentations is the newly discovered information from previous classification results. This has been termed 'classification-based-segmentation'. Thus a classifier learns higher level information over time.

Matsuyama and Hwang (1990) propose an aerial IU system that incorporates a classification-based-segmentation procedure. Initially, bottom-up features are extracted by a segmentation routine, and an appropriate contextual model chosen. During a top-down phase, hypotheses for instantiation are generated based on the model. Additional segmentations are then performed in order to satisfy these hypotheses. The segmentations are localized, and tuned to extract those specific features that the contextual model requires. For example, suppose the contextual model specifies that a driveway

is required at location D, near a particular house. If the driveway is already there in the initial segmentation, the relation is confirmed and stored. If the driveway is missing, another segmentation is performed only on the region around D, where parameters are tuned to extract driveway features. If it is detected, the hypothesis is confirmed. A reasoning agent convenes the global scene description, by checking consistency between bottom-up and top-down results, and choosing an appropriate routine based on the current scene description, with the goal of reaching an ideal description.

5.2.4 Philosophical point of view

Human visual perception consists of sensation and perception (Matsuyama and Hwang, 1990). Sensation can be defined as instantaneous absolute measurement. At sensation level, we are always looking at objects we have never seen before. Perception can be defined as persistent relative recognition. At perception level, we perceive familiar objects. If you look at a brick wall, sensation reveals a brownish color, a rough texture, rectangular shapes, and lines between the rectangles. This is a bottom-up analysis. Through perception, these features are grouped together and interpreted to be individual bricks that make up a wall. A more global awareness will group the wall with other walls, desks, chairs. Primarily through object types and the spatial relations of these objects, the scene will be interpreted to be a e.g. kitchen. Thus scene-matching has been performed. Now that we know we are looking at a kitchen, it would be easy to detect a chair, and this is the top-down phase. Sensory information need not contain complete information about a scene, since it is only used to trigger our recognition of the scene type. If we had to squint our eyes so that sensory information looses quality, we would probably still recognise the scene to be a kitchen. To quote Gregory (Gregory, 1970):

> "perception is not a matter of sensory information giving perception and guiding behavior directly, but rather the perceptual system is a 'look up' system in which sensory information is used to build gradually, and to select from, an internal repertoire of 'perceptual hypotheses' – which are the nearest we ever get to reality."

5.2.5 Defining a contextual model

We will now look at technical approaches to the formulation of a contextual model. Formulating the context of a scene type involves the generalization of a wide variety of complex spatial relations, clustering properties and patterns. The rest of this chapter will be devoted to a review of works that have attempted to solve this type of problem. Firstly, a review of spatial relation theory is required.

5.3 Spatial Relations

A formal description of spatial relations is a topic that has been important in geographical information sciences, robotics and Content Based Image Retrieval (CBIR), to answer questions about topology and to perform spatial queries (Liu et al., 2008). Four types of spatial relations can be identified in the literature (Hernandez-Gracidas and Sucar, 2007):

1. Topological relations (e.g. *meet*, *disjoint*): These are preserved under the topological transformations translation, rotation and scaling. There exists one and only one topological relation between any two objects.

2. Directional relations (e.g. *left*, *above*): These are preserved under scaling and translation, but variable under rotation. More than one directional relation may exist between any two objects.

3. Distance relations: These are based on distance measurements between the boundaries or centroids of objects. Preserved under rotation and translation, but variable under scaling.

4. Fuzzy relations: Measured in vague terms, e.g. *near*, *far*.

5.3.1 Topological relations

The first formalization of topological spatial relations between two entities was the 4 - Intersection Model (4-IM), developed by Egenhofer and Franzosa (1991). It is a simple model that is currently widely used and extended. The 4-IM determines the unique binary topological relation between two regions (without holes) in 2D (Rathi and Majumdar, 2002). This is determined by analyzing the intersections (or non-intersections) between their interiors and boundaries. The 4-IM is represented by a 2x2 matrix, as shown in equation 5.1 (adaption from Rathi and Majumdar, 2002):

$$F(A,B) = \begin{matrix} A_{int} \cap B_{int} & A_{int} \cap B_b \\ A_b \cap B_{int} & A_b \cap B_b \end{matrix} \tag{5.1}$$

where A_{int} is the interior of region A, and A_b is the boundary of region A. Thus between two regions, eight different topological relations may be defined (refer to figure 5.2):

- *disjoint* $\begin{matrix} 0 & 0 \\ 0 & 0 \end{matrix}$ *meet* $\begin{matrix} 0 & 0 \\ 0 & 1 \end{matrix}$
- *equal* $\begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$ *overlap* $\begin{matrix} 1 & 1 \\ 1 & 1 \end{matrix}$

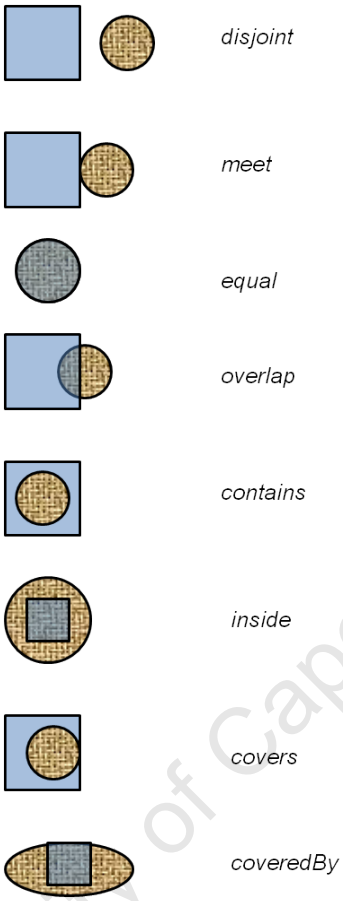


Figure 5.2: Topological relations: The 4 - Intersection Model (4-IM) (Egenhofer and Franzosa, 1991).
The 4-IM determines the unique binary topological relation between two regions A (blue square) and B (brown circle) in 2D.

- *contains* $\begin{matrix} 1 & 1 \\ 0 & 0 \end{matrix}$ *inside* $\begin{matrix} 1 & 0 \\ 1 & 0 \end{matrix}$
- *covers* $\begin{matrix} 1 & 1 \\ 0 & 1 \end{matrix}$ *covered by* $\begin{matrix} 1 & 0 \\ 1 & 1 \end{matrix}$

5.3.2 Directional relations

Directional spatial relations are normally employed to accompany and enhance topological relations in CBIR tasks. They can either be quantitative or qualitative. Quantitative relations refer to numerical values in degrees, minutes and seconds. This is normally computed between two objects by calculating the angle between a line joining the two centroids and a reference

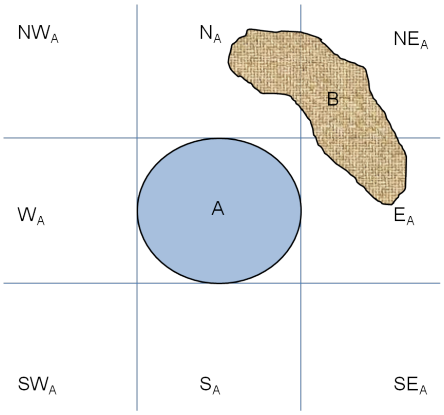


Figure 5.3: The direction-relation matrix model (Goyal and Egenhofer, 2000*a*).
The cardinal direction of object B with respect to object A is described by recording all tiles into which a part of B falls.

axis (Longley et al., 1990). Qualitative cardinal direction relations are based on a discrete set of symbols, e.g. {*left, right, above, below*}, or {N, S, E, W, NE, SE, SW, NW}. Cardinal direction relations are often employed in CBIR tasks (see section 5.4.1). Cardinal direction models such as the cone-based model and project-based model (Frank, 1991) have been proposed. However, these models are only suitable when the reference object is a point. They are inadequate to model the spatial relationships between objects found in a typical image segmentation, which are areal. The shape of areal objects will have an influence on cardinal direction (Goyal and Egenhofer, 2000*b*).

The direction-relation matrix model (Goyal and Egenhofer, 2000*a*) is a cardinal direction model designed for areal objects, where the influence of objects' shapes are considered. It works by considering the reference object within a reference frame (see figure 5.3). The reference frame has nine direction tiles: N, S, E, W, NE, SE, SW, NW. The cardinal direction of a target object with respect to the reference object is described by recording all tiles into which a part of the target object falls. This model is suitable for region-like objects, but fails to model point-like or line-like objects that are also typically found in an image segmentation. To resolve this issue, the direction-relation model was extended to the deep direction-relation matrix model (Goyal and Egenhofer, 2000*b*), which is able to produce consistent relations independent of object type (whether it be points, lines or polygons). This is accomplished by initially determining the object types of both reference and target objects and then establishing the appropriate relation.

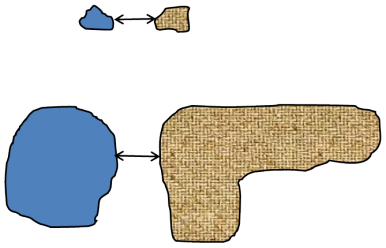


Figure 5.4: Quantitative distance between two groups of objects with different sizes (Liu et al., 2008)
The distance between the top two objects can be perceived to be smaller than the distance between the bottom two objects, due to the difference in object sizes.

5.3.3 Metric relations

Metric, or distance relations can be quantitative or qualitative. The quantitative distance between two areal objects A and B can be determined by computing the Euclidean distance between either the centroids or boundaries of A and B (Longley et al., 1990). However, as with directional relations, the shape and size of areal objects can have an influence on one's interpretation of a distance relation. Considering object size, the distance between A and B can be interpreted to be relatively larger if A and B were relatively large objects (see figure 5.4). Liu et al. (2008) propose a relative distance measure between two areal objects that takes the size of objects into account. It is defined as follows:

$$RelDistance(A,B) = \frac{|Distance(A,B)|}{\sqrt[4]{Area(C(A))Area(C(B))}}$$

where $Distance(A,B)$ is the Euclidean distance between either the centroids or the boundaries of A and B,$C(A)$ is the convex hull of A, and $C(B)$ is the convex hull of B.

The qualitative distance between two objects can be obtained by quantisizing the quantitative distance value to obtain a group of distinctions, such as *far* and *close.*

5.3.4 Fuzzy relations

Besides the the crisp topological, directional and metric relations mentioned above, fuzzy (see section 3.4 for fuzzy set theory) relations have been proposed in order to provide more robust qualitative decisions to vague quantitative relation measurements. For instance, a crisp set on the qualitative distance relations *near* and *far* would require a quantitative distance measurement that is neither *near* nor *far,* to be allocated to one of these cate-

gories. A fuzzy set on these categories, on the other hand, would introduce flexibility to subjectivity inherent in the definitions of these relations.

Liu et al. (2008) presents a framework of spatial relations for areal and non-overlapping objects such as those found in a typical image segmentation. The main objective is to employ these spatial relations to improve object classification in an OBIA (see section 4.3) environment. It is pointed out that the 4-IM is inadequate, since only two relations *(disjoint* and *meet)* occur (since objects do not overlap). They thus propose a new set of quasi-topological relations, some of them being fuzzy. They are termed 'quasi-topological' because they are based on the 4-IM model, but may not satisfy the constraint of topology, i.e. preservation under transformation.

Relations between two areal objects (Liu et al., 2008)

For relations between two areal objects, six relations are defined: *disjoint*, *surroundedBy*, *surround*, *invade*, *invadedBy*, and *s-meet* (refer to figure 5.5). Disjoint is the same as that in the 4-IM (see figure 5.2). The *surround* and *surroundedBy* relations are analogous to *contains* and *inside* of the 4-IM, and are determined by considering the boundaries and the convex hulls of objects. *Invade*, *invadedBy*, and *s-meet* are quantitative fuzzy relations. To determine whether an object B *invades* an object A, a fuzzy membership function is established based on the following formulation:

$$invades(A, B) = \frac{area(C(A) \cap C(B) \cap A)}{area(A)} \tag{5.2}$$

where *area* is a function that computes the area of a region, and $C(A)$ is the convex hull of A.

Relations between two line-like objects (Liu et al., 2008)

Relations between two line-like objects consists of the six quasi-topological relations by viewing them as areal objects, as well as additional relations by abstracting them to two-dimensional real lines. These additional relations are dependent on the objects shape, and include *along*, *connect*, *merge*, *mergedWith*, *l-meet*, and *crosses* (refer to figure 5.6). An object's shape is determined by computing a 'Relative Longness index' (RLI). This is the ratio of the length and the width of an object. Line Like Objects (LLO) have a large RLI. Point Like Objects (PLO) have a small RLI. A fuzzy definition of LLOs and PLOs is proposed, where a membership function is constructed based on the RLI. The relations are described as follows:

For two LLOs A and B:

- The *along* relation: A fuzzy set is defined over the following:

$$along(A, B) = \frac{length(b(B) \cap b(A))}{length(b(B))} \tag{5.3}$$

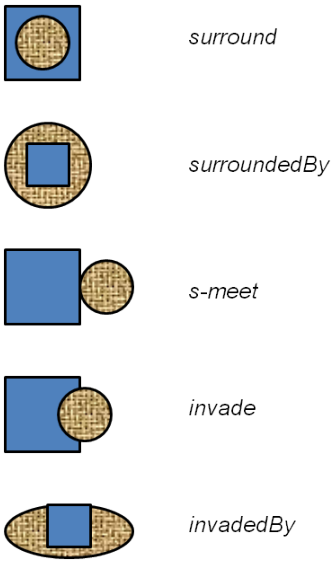


Figure 5.5: Quasi-topological relations between two areal objects (Liu et al., 2008).

where $length(b(B) \cap b(A))$ is the length of intersection of the boundaries of A and B, and $lengthb(B)$ is the length of the boundary of B.

- *Connect, merge, mergedWith*, and *l-meet*: For these relations, $length(b(B) \cap b(A))$ must be below a certain threshold. The relations are discriminated based on how A and B are adjacent.
- The *crosses* relation is determined when three LLOs are involved, due to the constraints of non-overlapping objects.

Aksoy et al. (2003) similarly proposes a set of fuzzy spatial relations for areal objects. The goal is to establish a visual grammar that will improve CBIR of remote sensing images. The following fuzzy spatial relations are defined: *disjoined*, *bordering*, *invaded by*, *surrounded by*, *near*, *far*, *right*, *left*, *above* and *below*. The computation of fuzzy membership functions are based on measurements such as perimeter, shape moments, and orientations of the objects concerned. The fuzzy relations are defined as follows:

For two objects A and B: Let

d_{AB} = Distance between A and B. This is based on the smallest Euclidean distance between the boundary pixels of A and B.

α_{AB} = Angle between A and B. This is the angle between the horizontal (column) axis and the line joining the centroids of A and B.

μAB = common perimeter between A and B / perimeter of A. The common perimeter between A and B is the length of that line segment that coincides with the borders of both A and B.

For the topological relations *disjoined*, *bordering*, and *invaded by*, a fuzzy set is defined on μAB. For the distance relations *near* and *far*, a fuzzy set is defined on d_{AB}. For the orientation relations *right*, *left*, *above* and *below*, a fuzzy set is defined on α_{AB} .

5.4 Image understanding studies

The question of how to reduce the gap between how humans perceive a scene and how current automated feature extraction systems describe a scene is largely unanswered. Studies that incorporate the IU principle can be found in several different research domains, the most prominent being Content Based Image Retrieval (CBIR), and to a lesser extent remote sensing. The majority of CBIR studies deal with ground-based images. Nevertheless, techniques that are relevant to aerial image analysis can be drawn from a selection of studies.

In the rest of this chapter a short review of CBIR theory will be presented, followed by relevant CBIR studies. This will be followed by a more in-depth review of aerial image understanding studies. The following terms will be used throughout the rest of this discussion. *Low-level* features are those spectral, shape and textural features that can be extracted from a single object obtained from a typical image segmentation. *High-level* features refer to contextual measurements in a scene, such as the spatial relationships between objects.

5.4.1 Content Based Image Retrieval

Image retrieval is the extraction of a subset of images from an image set according to predefined criteria (Smith and Chang, 1997). If these criteria are the visual contents of the image i.e. colour, texture, shape features, it is referred to as Content Based Image retrieval (CBIR). CBIR is an active research area in Computer Vision. It is a required tool for browsing, searching and retrieving images from the currently many large image databases in application domains such as fashion, crime prevention, publishing, medicine, and architecture, as well as remote sensing (Liu et al., 2007). A typical CBIR query could read "retrieve all outdoor scene images from an image database". Thus a characterization of prototype scenes is required, as well as a scene-matching methodology, which is relevant to the objectives of this thesis.

A CBIR system can work by extracting either global image features or local object features. In terms of global features, Oliva and Torralba (2001) for instance argue that a scene can be categorized without initial segmentation and classification of the image. A holistic representation of the structure of a real world scene is proposed, what they term a 'spatial envelope'. In other words a scene is seen as a single, unitary form rather than individual

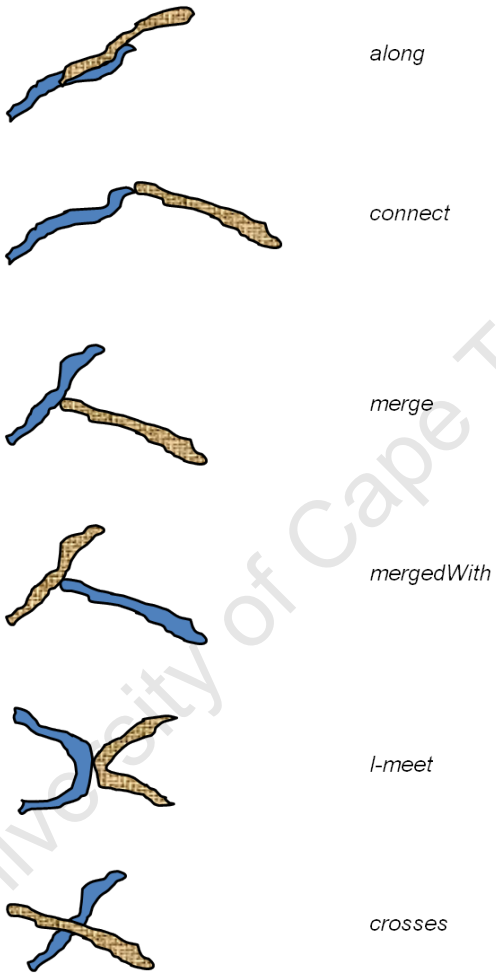


Figure 5.6: Quasi-topological relations between two line-like objects (Liu et al., 2008).

Line-like objects have a large Relative Longness Index (RLI). RLI is the ratio of the length and the width of an object.

objects. The spatial envelope consists of a set of perceptual dimensions (naturalness, openness, roughness, ruggedness and expansion) that represents the dominant spatial structure of a scene. These dimensions are reliably estimated using spectral and coarsely localized information. The model consists of a multidimensional space where scenes belonging to the same semantic categories (e.g. streets, coasts, trees) are projected closed together. Experimental results show that a holistic representation of a scene is sufficient to inform its likely semantic category.

Jing et al. (2003) argues that a single signature computed for the entire image cannot sufficiently capture the important properties of individual objects. Region-Based Image Retrieval (RBIR), on the other hand, represents images at object-level, usually through an initial segmentation of the image. RBIR is considered by Jing et al., (2003) to be closer to perception of the human visual system. It is currently an established and popular CBIR technique. We will prefer to use the term Object-Based Image Retrieval (OBIR) in keeping terminology consistent with the rest of the thesis.

Technical overview of OBIR:

A typical OBIR system is designed as follows (adaption from Liu et al. (2007)).

1. Perform segmentation of the entire scene. This is technically similar to the aerial image segmentation procedures covered in section 4.1.

2. Extract features from the image objects. These can include colour, texture, shape (low-level) and spatial relation (high-level) features.

3. Define similarity between two images. This is normally based on the distances between object features in two images.

OBIR based on spatial relations:

A standard OBIR system works by matching a training image A to a target image B by comparing the values of low-level features of objects in A and B.. Many authors have expressed the inability of such systems in reducing the *semantic gap*. The semantic gap is the difference between the way a human interprets a scene in an image, and that produced by low-level features. In an effort to reduce the semantic gap, the image understanding concept has been explored, mostly by modeling the spatial relations of image objects. For instance, sky and ocean may have similar colour, texture and shape properties, whereas sky is always *above* sea.

Ma and Manjunath (1997) segments an image and extracts colour, texture, and shape features from each image object. Spatial location information is appended to each object in order to improve retrieval performance. Spatial location is in the form of the coordinates of an object centroid as

well as its minimum bounding rectangle (MBR). Similarity between objects in two images is defined as the degree of overlap of their MBRs.

Smith and Chang (1997) measures similarity between two images based on colour, object sizes, and the absolute and relative spatial location of objects. The querying process is hierarchical: Candidate images are initially selected based just on colour and sizes. This selection is then pruned based on spatial location. The matching based on absolute spatial location works by first computing matching scores between individual objects in two images. These scores are based on a comparison of the coordinates of object centroids and their MBRs. Matching multiple objects in two images is then performed by intersecting the individual object matching scores. Thus a final match is obtained based on the absolute location of all objects in the images.

With these last two studies, in drawing from the IU framework in section 5.2, we can say that a contextual model is built based on the absolute spatial locations of individual objects in a scene. This can be useful for e.g. detecting inconsistencies in surveillance images, where the absolute position of objects is important. Intuitively, urban scene context is not well defined by the absolute position of urban objects, unless we wish to characterise e.g. 'scenes with schools in the upper left region'. What is more relevant is the relative spatial location of all objects in a scene, or the definition of a single model based on the spatial pattern of all objects.

Ren et al. (2002) defines six spatial relations for an object pair: They make the observation that directional as well as topological relations are required to provide a complete representation of the semantic / structural image content. The six relations are: *left*, *right*, *up*, *down* (directional), *meet*, and *front* (topological). These relations are modeled on segmented and classified objects in an image by comparing object centroids. For every object pair in an image, a spatial feature vector is constructed with the following format: <image name, object A name, object A index, object B name, object B index, *right, left, up, down, front, touch*>. The last six values are binary true or false. Thus for each image a set of spatial feature vectors exists, of equal length. Similarity between two images is measured by comparing their spatial feature vector sets. The overall similarity score is based on the number of identically classified objects, the proportion of identical spatial relations between the same two objects in both images, and the number of identical objects that share similar spatial regions. Thus in this case a contextual model of a scene consists of a list of topological / directional spatial relations between all object pairs in the scene.

Hernandez-Gracidas and Sucar (2007) improves object classification results in ground-based images by modeling segmented image objects with spatial relations within an MRF (see section 4.2.1) framework. The motivation is that improved object classification results will yield improved scene-matching (retrieval) performance. Their method operates as follows:

1. An image is segmented and classified based on low-level feature extraction.

2. The following topological and directional relations are then computed between objects: *meet*, *disjoint* (topological), *beside* (either *left* or *right*), *horizontally aligned*, *vertically aligned*, *above* and *below* (directional).

3. The object classification results are then improved by modeling the spatial relations with an MRF. This is done by representing the image content as a graph, where each image object is represented by a node in the graph and the spatial relation between two objects represented by an edge between the two nodes. Spatial relation probability laws are extracted from trained images and fused with expert knowledge. This is done by observing the frequency at which a spatial relation occurs between two object types. Then in the unknown query image, the probability of occurrence of a certain spatial relation between each pair of classified objects is obtained. This will generate a probability of an object belonging to a certain class based on its neighbourhood objects. The MRF works in such a way that the most probable classification configuration for the whole scene is generated.

The method was tested on a set of landscape images and shows an improvement of almost 9% compared to the initial bottom-up classification results.

Rathi and Majumdar (2002) propose an image retrieval system that includes spatial relations between objects, where the objective is to facilitate the effective searching of images on the world wide web. The motivation to include spatial relations is that users often search for images that contain specific objects with specific directional and topological relations between them. Their spatial relation model is described as follows: After objects in an image have been manually classified, spatial relations are computed between them. The 4-IM is used to describe the topological relation between two objects. The following directional relations are defined: *left-of*, *right-of*, *above*, *below*. As in (Hernandez-Gracidas and Sucar, 2007), an image is represented by a graph, where each object corresponds to a node in the graph. An edge between two nodes corresponds to the directional (if any) and topological relations between the two objects. An edge also contains the Euclidean distance between the centroids of the two objects. The problem of finding similarity between two images thus gets reduced to one of graph matching. A short overview of the graph matching procedure is as follows: The similarity between two images A and B, each containing two objects $O_{A,1}$, $O_{A,2}$ and $O_{B,1}$, $O_{B,2}$ respectively is defined as a combined measure of the following four components:

1. Object similarity: based on color similarity of the objects $O_{A,1}$, $O_{A,2}$ and $O_{B,1}$, $O_{B,2}$.

2. Directional Similarity: This determines to what extent the directional relations between $O_{A,1}$ and $O_{A,2}$ match the directional relations between $O_{B,1}$ and $O_{B,2}$.

3. Topological Similarity: Likewise, this determines to what extent the topological relations between $O_{A,1}$ and $O_{A,2}$ match those between $O_{B,1}$, $O_{B,2}$.

4. Distance Similarity: This is based on how well the Euclidean distance between the centroids of $O_{A,1}$ and $O_{A,2}$ match the Euclidean distance between the centroids of $O_{B,1}$ and $O_{B,2}$.

Thus from the above studies we draw two approaches to constructing a contextual model. One approach is to extract a list of high-level spatial relation features from the imagery. The other approach is to represent the image content as a graph, where an object corresponds to a node in the graph, and an edge between two nodes corresponds to the spatial relations between two objects. We will now move on to aerial image understanding studies.

5.4.2 Aerial image understanding studies

The following is a review of five studies that apply the IU principle to remote sensing (aerial) images, either to improve object classification or to perform scene-matching.

Durand (2007) presents a top-down methodology for improving the classification of objects in remote sensing images. A contextual model is constructed in the form of a semantic network using symbolic supervised machine learning tools presented in (Sheeren et al., 2006). A semantic network is made up of a set of concepts (e.g. building, building cluster, tree), their attributes, and their relations to each other) (see figure 5.7). In this study, concept attributes are identical to low-level features (colour, shape). In a bottom-up analysis, the authors segment an image and extract a set of spectral and shape features for each image object. To recap the standard object-oriented classification task, an individual image object is associated to a class based on a comparison of that object's features and an individual class' feature space. In this study, the authors propose a comparison of an object's features with their semantic network in order to gain an awareness of the context in which an object lies. This is done as follows. A local matching score is computed between an object and a concept in the semantic network. This matching score is based on the distance between the image object's features and the features of a concept in the network. If the matching is significant, a traverse down the semantic network takes place to compute a global matching score. The global matching score is a linear combination of local similarity measures between image object and concept, starting at the root of the semantic network and ending at the image object of interest.

The global matching score gives an assessment of how well an image object matches a certain concept within the hierarchy of concepts in the semantic network. Thus an image object is given semantic meaning. In other words, each object is initially classified based on the constraints of the contextual model. Thus image objects are not classified first and then evaluated against a contextual model. Rather, they are classified through an evaluation against a contextual model. The effectiveness of the proposed method is shown by experimentation with a high resolution multispectral Quickbird image of an urban district of Strasbourg, France.

Aksoy et al. (2003) formulates a set of spatial relations, which have already been discussed in section 5.3.4. The goal is to establish a visual grammar that will improve content based image retrieval of remote sensing images. In order to describe their procedure, lets define the term 'object' to mean e.g. residential area, lake, park, fields, CBD area, snow, tidal flats (large scale). A *prototype scene* consists of a group of particular objects (e.g. tree covered islands, residential areas with coastline, snow covered mountains). They define a *level 3* feature to be a combination of spatial relation features (all possible pairwise object combinations). For example, for the three small scale objects; road, building and car, six pairwise combinations exist. A car may be *surrounded* by road, car *right* of building, building *left* of car, building *above* road, road *below* building, road *surrounding* car.

A scene matching methodology is presented, where an unknown scene is allocated to one of a predetermined set of prototype scenes. The procedure operates as follows. For an unknown scene to be tested:

1. The scene is segmented to find object boundaries.
2. Objects are classified based on low-level features.
3. Spatial relation features are established.
4. A Bayesian framework is set up to match a scene based on its level 3 feature distribution.

In more detail on step 4., the Bayesian framework learns prototype scenes based on automatic selection of distinguishing (frequently occurring / rarely occurring) level 3 features. This concept will now be explained. To train the scene-matcher, the user selects a set of training images. Each training image contains example scenes for each prototype scene. For each training image:

1. Count the number of times each possible level 3 feature occurs, for each prototype scene. A level 3 feature of interest is that which frequently occurs in a particular prototype, but rarely in others. This is determined by selecting those level 3 features that cause the largest class separability. Thus for each prototype a contextual model is constructed that consists of a unique level 3 feature distribution. In other words, a

contextual model is learned for a prototype by observing which specific combinations of object spatial relationships are unique to that scene.

2. Use Bayes rule (see equation 4.1) to estimate a particular prototype for an unknown image scene based on its level 3 feature distribution.

In summary, this study defines different contextual models for different scene types. The models consist of a set of high-level spatial relation features unique to that scene type. Scene-matching is conducted by comparing the high-level feature space of each scene type to the high-level feature space of the input scene.

Liu et al. (2008) formulates a set of spatial relations, which are discussed in section 5.3.4. A case study is presented where it is shown that the use of these spatial relations can improve the extraction of car objects from a high resolution aerial image of an urban scene. Their method operates as follows:

1. Firstly the image is segmented with optimum parameters that enable delineation of roads and cars.

2. Roads are classified as being line-like, or having a number of lanes that are line-like, using a Relative Longness index (bottom-up phase).

3. Car objects are classified based on the following properties:

   (a) Non line-like.
   (b) *Surrounded* by or *invading* a road or lane.
   (c) *Neighbouring* road or lane.
   (d) Area of car objects between two thresholds.
   (e) If two car objects are close together, they are merged.

An accuracy assessment shows significant improvement in car detection results when using high-level features (as opposed to just low-level features), on a particular road for this particular dataset. This study provides promising contextual image interpretation in an OBIA environment. However, the interpretation essentially consists of specific spatial constraints for specific objects, optimally tuned to a specific dataset. An ideal contextual image interpreter should recognise the context of any type of scene, and interpret the scene using parameters suited to that scene type.

Porway et al. (2008) defines a contextual model in the form of a hierarchical, semantic network of an urban scene (refer figure 5.7). The objective of this study is to improve the extraction of urban objects in aerial images. The model starts at scene level, which is the entire scene. Scene level decomposes into groups of objects, such as blocks of buildings or rows of cars. Object groups decompose into single objects, such as a building, which decomposes

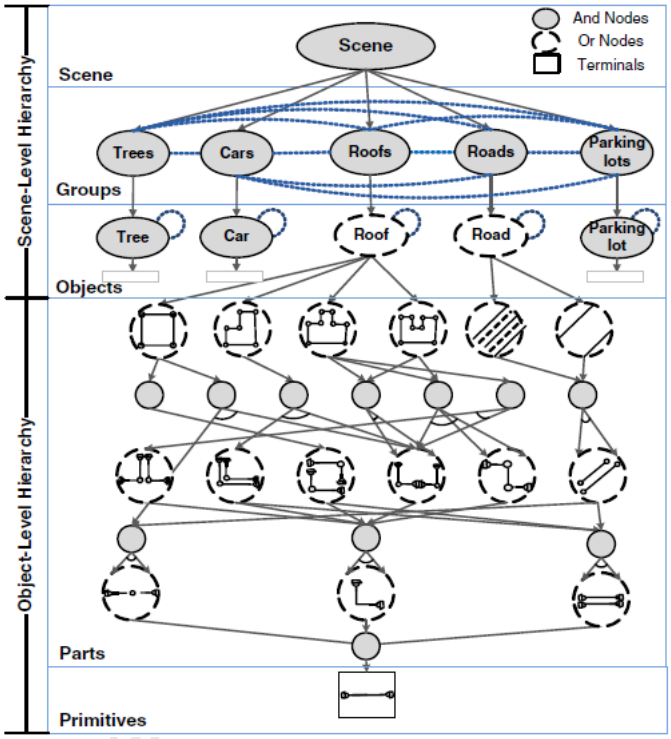


Figure 5.7: The hierarchical, semantic network of an urban scene used in (Porway et al., 2008).

Scene level decomposes into groups of objects, such as blocks of buildings or rows of cars. Object groups decompose into single objects, such as a building, which decomposes further into parts and primitives. This hierarchical model helps capture objects, as well as the different characteristics of the scene, at varying scales.

further into parts and primitives. This hierarchical model helps capture objects, as well as the different characteristics of the scene, at varying scales. Their approach to classifying objects in an urban scene is as follows:

1. Detect roofs, roads and other urban objects using Compositional Boosting (a method for finding image structures) and low-level features (bottom-up phase). This is designed to produce a high true positive rate, at the cost of many false positives.

2. Activate high-level features to resolve inconsistent false positives (top-down phase).

More detail on step 2 is as follows. In the training phase, high-level features such as *relative scale*, *relative position*, *relative orientation*, *percentage overlap* and *aspect ratio* are learned from a training set of manually labeled aerial images. For example, all 'buildings *near* roads' in a training image are identified, and the distance between them returned. These values are then modeled in a histogram, in order to learn a prior probability model to be used in an MRF framework. The classification proposals in step 1 that are inconsistent with the model are then pruned, with the goal of arriving at the most likely description of the scene. In an experiment, 196 multiresolution training images were taken from Google Earth and hand-labeled to learn their prior model. For their testing set, three Google Earth images were mosaicked together from many smaller high resolution images. This allowed their detectors to run at multiple scales for each image. Results show a drastic reduction in false positives after implementing the top-down phase.

As already mentioned, an ideal image understanding system should be able to work on any input scene. The contextual model should thus be constructed from a large amount of well distributed and representative sample scenes. The system proposed in (Porway et al., 2008) is promising because they used a large amount of training images to construct their contextual model. An issue with this system, however, is that only one contextual model is defined (no scene-matching methodology is implemented).

Urban scenes while similar are significantly different due to the planning policies for different land-use types. There are significant structural differences between (e.g. CBD, low cost housing development, industrial). The training images that were chosen in (Porway et al., 2008) either consisted of a combination of such scene types (which would yield soft peaked training histograms and thus less class discrimination), or more or less the same particular scene type (which would yield unwanted results if the scene that were to be tested is different in structure to that scene type). To build a more generic and robust classifier, contextual discrimination of scene types (such as that proposed in Aksoy et al., 2003) is required. Aksoy's system, however, is designed for large scale scene prototypes (e.g. tree covered islands), and

not necessarily urban land-use scenes. Land-use context needs to be characterised and formulated in order to discriminate between different land-use scenes.

This is the issue that this thesis focuses on. A characterization of land-use context is necessary for performing urban scene-matching (i.e. land-use classification) based on bottom-up classification results. With the knowledge of land-use for a scene to classify, top-down methods can be employed to produce an accurate final scene description, based on a land-use contextual model.

A land-use classification study related to South African urban context can be seen in (Busgeeth et al., 2008). The authors propose a hierarchical, rule-based land-use classification system designed to detect South African informal settlements. Three classification stages are carried out. The first distinguishes built-up from non built-up areas. The second differentiates formal from informal settlements. The third level refers to sub-classes. A formal settlement, for example, can decompose into high rise buildings / formal township / mining hostel. Informal settlements are distinguished from formal settlements based on the following features: *average building size, building size variety, formalized / informalized street pattern,* and *tarred / gravel roads.* Sub-classes are classified based on the following features: *material composition of buildings, homogeneous / heterogeneous roofs, building density, presence of engineering services,* and *presence of infrastructure.* Experiments were conducted on Quickbird imagery of the Soweto region in South Africa. Built-up areas were manually delineated. The above features used by the classifier were manually extracted from the imagery. The classification system was effective to the second level (in differentiating formal from informal).

In a second experiment in (Busgeeth et al., 2008), an automated land-use classification system is proposed. To train the classifier, sample scenes of image tiles 120m by 120m are generated from manually delineated regions of known land use. Local binary pattern features are automatically extracted from these regions and used to train a Support Vector Machine. The classifier works by moving a 120m by 120m window over the input image to produce an overlapping set of tiles. Each tile is classified using the trained Support Vector Machine. Experiments were conducted on the Soweto dataset. The classifier was able to separate built-up from non built-up areas, as well as formal from informal. No quantitative accuracies were given.

The first phase of this study shows that formal / informal land-use discrimination can be accomplished based on a set of high-level contextual features. These features give insight as to how land-use classes can be separated. The features were, however, manually extracted. The human eye might easily recognise *formalized / informalized street* pattern, but for automation objectives, this feature needs to be defined in order to be quantitatively extracted from imagery. Furthermore, only four features were used based on

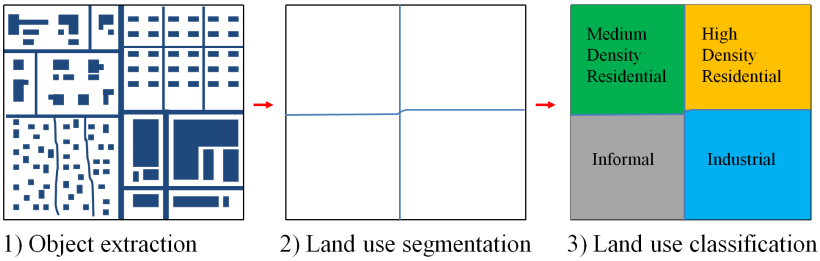


Figure 5.8: Land-use segmentation and classification based on image objects. 1) Bottom-up object classification is performed. 2) Land-use segmentation is performed based on object features. 3) Land-use classification is performed based on object features.

expert knowledge. Ideally, from all possible discriminatory features that can be automatically extracted from imagery, an optimum subset needs to be established that causes greatest land-use separation.

The second phase of their study is promising as it shows that land-use classification can be performed in an automated fashion. The system is based on global image features (features extracted from the raw imagery). The other option is to extract features from image objects, i.e., segmented regions. This technique is considered closer to human visual perception, as it captures the important properties of individual objects (Jing et al., 2003) (refer to section 5.4.1). A moving window strategy such as that proposed in (Busgeeth et al., 2008) will be ineffective in performing land-use classification based on the properties of objects, since a small window will not capture the properties of groups of objects. A larger window would not work since land-use borders wouldn't be captured. Thus a segmentation and classification of land use regions is required, based on bottom-up object features (see figure 5.8).

The next chapter will detail the design of a proposed urban aerial image understanding system. The two major contributions are in the contextual model definition and scene-matching phases. The approach to defining a contextual model is to establish a set of high-level features from training images, as used in most of the above studies. In more detail, urban land-use context is characterised by extracting a set of high-level features from sample images of different land-use types. An automated land-use classification system is also proposed, based on the above concept (figure 5.8).

Chapter 6

An Urban Aerial Image Understanding Approach

6.1 Introduction

In the previous chapter Image Understanding (IU) theory was presented along with various aerial and ground-based image approaches. This chapter will detail the design of an urban aerial image understanding system.

In the previous chapter in section 5.2 a general framework for aerial IU was presented. In summary, this framework consists of the following:

1. Bottom-up feature extraction
2. Contextual model definition
3. Scene-matching
4. Top-down inferencing

The major contribution of this thesis is in steps 2. and 3. The motivation is that whilst well established urban scene bottom-up techniques already exist, no formal definition of urban land-use context has been well established in the literature. 'Context' refers to the spatial patterns and shape features of predominant urban objects that characterise a scene type. The top-down phase is considered a future research direction.

6.2 Proposed Methodology and Rationale

The proposed approach consists of the following:

1. Construction of an urban contextual model.
2. Developement of an automated land-use classification routine.

For the rest of this discussion the above will be refered to as 'part 1' and 'part 2'. In part 1 of this research, manually labeled sample scenes of different land-use types are analyzed. A set of high-level features is extracted from the sample scenes in a similar manner to that in Porway et al. (2008). Multivariate statistical visualization tools are utilized to observe consistencies in feature measurements for samples of the same land-use type. Thus the separability of land-use classes in the high-level feature space is analyzed. This gives us an indication of the land-use discriminatory power of the set of high-level features. A feature selection procedure (see section 2.6) is used to compute an optimum subset of features that causes maximum class separability. This answers the question as to what is context (research question 1 in section 1.4). This high-level feature space may be regarded as a 'scene descriptor', and can be used in top-down object extraction analyses or land-use classification strategies.

Figure 6.1 illustrates a conceptual overview of our contextual model using just two land-use types, 'medium density residential' and 'industrial', as an example. Raw images containing these land-use scenes are manually digitized. The figure shows how a residential scene may appear different in structure to an industrial scene based on high-level features of building and road objects. For instance, the industrial scene has a larger variety of building sizes. Standard deviation (stdev) of building size may thus be a candidate high-level feature. Other high-level feature candidates may include *mean building area, mean / stdev road width, mean building-to-road distance, mean building-to-building distance, building density.*

A set of high-level features are measured off several sample scenes of each land-use type. For example, if 5 high-level features $(f_1, f_2, ..., f_5)$ were measured off 3 residential scenes (R_1, R_2, R_3) and 3 industrial scenes (I_1, I_2, I_3). A 6X5 observation matrix then exists, as shown in equation 6.1

$$\begin{array}{ccccc} R_{1,f1} & R_{1,f2} & R_{1,f3} & R_{1,f4} & R_{1,f5} \\ R_{2,f1} & R_{2,f2} & R_{2,f3} & R_{2,f4} & R_{2,f5} \\ R_{3,f1} & R_{3,f2} & R_{3,f3} & R_{3,f4} & R_{3,f5} \\ I_{1,f1} & I_{1,f2} & I_{1,f3} & I_{1,f4} & I_{1,f5} \\ I_{2,f1} & I_{2,f2} & I_{2,f3} & I_{2,f4} & I_{2,f5} \\ I_{3,f1} & I_{3,f2} & I_{3,f3} & I_{3,f4} & I_{3,f5} \end{array} \tag{6.1}$$

The multivariate statistical tools assess similarities within the residential feature vectors (first 3 rows), and likewise the industrial feature vectors (last 3 rows). They also assess dissimilarities between residential versus industrial feature vectors. Thus the separability of the residential and industrial classes is assessed, i.e., how well can these two classes be discriminated based solely on the high-level features? This assessment is made by visual analysis of the statistical plots. The plots also give an indication as to which features are significant in separating scenes. Feature significance is then robustly determined by applying a machine learning feature ranking routine to the matrix.

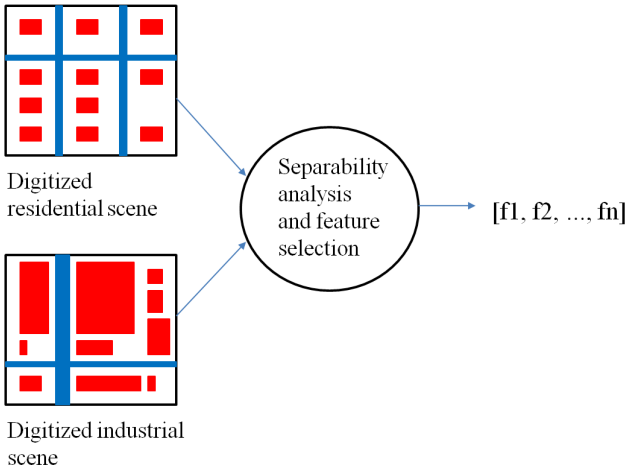


Figure 6.1: Constructing a contextual model. Manually digitized sample scenes of different land-use types are analysed. The [f1, f2,....fn] denotes a high-level feature set (*e.g. road-building distance, building density, average building size*). High-level features that cause discrimination of land-use types are selected.

This routine attaches an importance index to each $(f_1, f_2, ..., f_5)$ according to its significance in separating classes. Features with the highest ranking are then chosen to be our optimum subset of features, e.g. (f_2, f_4, f_5). This high-level feature space may be regarded as a definition of urban context.

In part 2 of this research an automated land-use classification routine is proposed. Figure 6.2 illustrates the concept. Land-use regions are segmented based on bottom-up object features. These segments are then classified to a land-use based on a set of high-level features of the bottom-up object classification results. These high-level features are the same as those established in part 1. The objective is to assess whether scene-matching of unknown scenes can be performed based on these features, and to quantitatively assess the scene-matching accuracies that can be obtained.

The high-level feature space, or scene descriptor, established in part 1, is a template for what a scene should ideally look like. A proposed future research direction would be to utilize this feature space as a constraint to improve bottom-up results (the top-down phase). For example, to improve object classification results of a scene that was classified as 'industrial', the 'industrial' high-level feature space is used to constrain the final description of the scene.

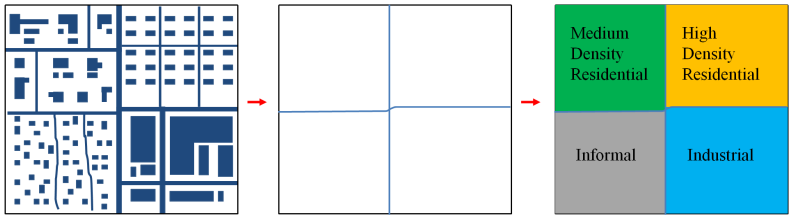


Figure 6.2: Land-use segmentation and classification based on image objects. 1) Bottom-up object classification is performed. 2) Land-use segmentation is performed based on object features. 3) Land-use classification is performed based on high-level features of bottom-up objects.

6.3 Construction of an Urban Contextual Model

From a theoretical point of view, the goal in the formulation of urban scene context is to emulate the human visual system. Let us visualize two scenes categorized as residential, in different parts of a city. Human vision would probably recognise that these two scenes are residential. In other words the context of the scenes is similar. Formulating these similarities, however, is difficult. It involves the generalization of a wide variety of complex spatial relations, clustering properties and patterns. The formulation process can begin by recognizing, by visual analysis, a discrete set of high-level features that have similar values for the same land-use type, and dissimilar values for different land-use types. This is the general approach used in the majority of image understanding studies (section 5.4). Useful urban scene features can include spatial relation distance features (see section 5.3.3), such as *building-building distance* and *road-building distance*, statistics of shape features such as *mean / stdev building area*, *mean / stdev road width*, and other contextual features such as *building density*. To build a 'residential scene signature', these high-level features need to be extracted from residential scene samples in order to gain a generalized model. This requires a formulation of each feature based on image content. The more scene samples and the more high-level features that are selected to learn a model, the more robust the model will be. The model will be in the form of a high-dimensional high-level feature space. The dimension is then reduced using data reduction techniques (section 2.6), in order to gain a more compact model.

The proposed approach to constructing an urban scene contextual model is as follows:

- Establish a set of land-use types (e.g. industrial, residential).
- Obtain a set of sample scenes for each land-use type.
- Manually label predominant urban objects in the sample scenes.

- Formulate a set of high-level features, and extract these features from each sample scene.
- Use multivariate statistical tools to test whether different land-use scenes can be discriminated (separated) based on these high-level features, and to observed which high-level features are significant in separating scenes.
- Use a machine learning tool to rank each feature according to its significance in separating land-use types. This ranking can be used to select an optimal subset of features.

6.3.1 Land-use selection

A set of urban land-use types should be chosen by observing the various dominant land-uses in a dataset. The goal in choosing land-use types is to include all possible types, so that when classifying a scene it is likely to fall into one of these categories. On the other hand too many types would lead to some with similar high-level feature spaces, which is redundant. In this research a limited number of land-use types were chosen due to time constraints. The following were chosen:

- Residential medium density
- Residential high density
- Informal residential
- Industrial
- Commercial

6.3.2 Sample scene selection

For each land-use type, sample scene datasets are required for training the model. The samples should be representative and diverse (obtained from a large study area). As already mentioned, the more samples for each land-use type the better, in order to generalize the formulation of context. In this study, due to time constraints and availability of data, about five samples were chosen for each land-use type from well distributed spatial locations and different suburbs in a 2007 high resolution RGB aerial image dataset of the greater Cape Town region. Each sample consists of 500 - 2000 buildings. Figures 6.3 - 6.7 show an example of a scene sample for each land-use type. The black border marks the scene extents.

Figure 6.3: Seapoint: an example of a medium density residential scene. The scene is characterised by invariant building sizes and road widths, and a regular street pattern.

Figure 6.4: Crossroads: an example of a high density residential scene. The scene is characterised by groups of compact buildings that are generally close together. The buildings are generally smaller than medium density residential buildings.

Figure 6.5: Boys Town: an example of an informal scene.
The scene has narrow roads, and small compact buildings that are irregularly dispersed.

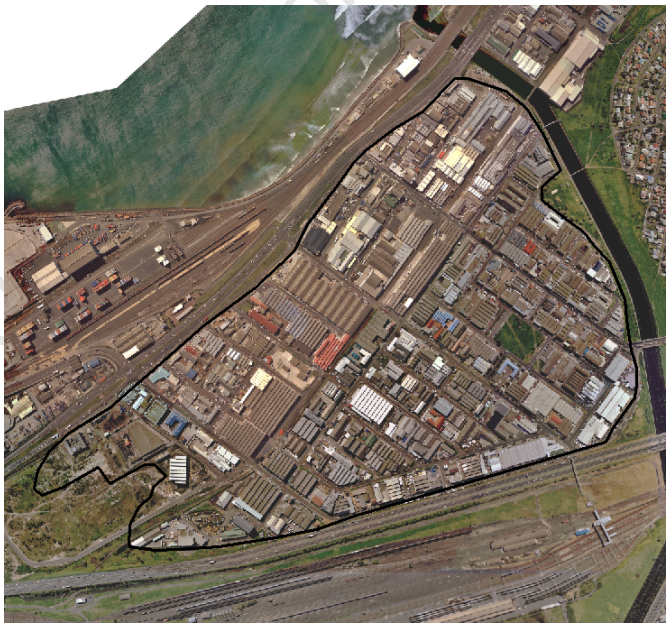


Figure 6.6: Paarden Eiland: an example of an industrial scene.
The scene is characterised by large buildings. There is also a large variety of building sizes. Buildings are close to one another, and close to roads.

Figure 6.7: Claremont: an example of a commercial scene.
A very large variety of building sizes is present, including one large building (a shopping mall). Like the industrial scene, buildings are close to one another and close to roads. There is also a variety of road widths.

6.3.3 Choice of scene objects

The definition of geometric context in a scene is based on one or more scene objects (e.g. tree, road, building, car) within that scene. Building and road objects are proposed, with the following rationale. The general aim is to emulate a human interpreter's success in discriminating scene types. A visual analysis was thus carried out on the scene samples to test the human eye in picking up predominant urban objects. In terms of geometry, the urban objects that were predominant in defining context at a large scale were buildings and roads.

These objects were manually digitized in each scene sample. Figure 6.8 shows the digitization of a medium density residential scene sample in the suburb of Seawinds. The digitization essentially produces a set of building and road polygons for each scene sample. These vector data are all that is needed to define the contextual model.

6.3.4 Choice of high-level features

An initial set of high-level features were formulated for further analysis based on expert knowledge. A visual analysis was carried out on the scene samples to discern which features might contribute to scene type discrimination.

Figure 6.8: Digitization of Seawinds sample scene.
Building (red) and road (dark blue) features are manually digitized in each sample scene.

Knowledge of urban planning specifications was also be used, e.g. different and discrete road width requirements according to land-use. The chosen set of features were extracted from the digitization of each scene sample through use of Esri ArcMap (Esri, 2010).

6.3.5 Technical description of features

The proposed set of features, along with a technical description and rationale, are listed in table 6.1.

6.3.6 Feature extraction methodology

The methodology for extracting these high-level features from the sample scenes is as follows. For a feature such as *building area*, the areas of all building polygons in a scene were measured (refer to figure 6.8 for an example scene of building and road polygons to be measured). Both the mean and the standard deviation of these *area* measurements were taken into account for further analysis. The rationale for standard deviation is that, when considering *building area* for instance, the variance of building areas may be discriminatory for e.g. industrial vs residential scenes. For other features such as *building density*, one measurement exists for each scene.

A total of 13 features then exists, as follows:

1. *Mean building area*

Table 6.1: Proposed initial set of high-level features

Feature	Description	Rationale
building area	The 2D area of a building footprint.	Buildings in e.g. an industrial area are generally larger than buildings in a residential area.
building compactness	Building compactness = $\frac{4\pi A_{building}}{P^2_{building}}$, where $A_{building}$= building area and $P_{building}$= building perimeter.	The compactness of commercial buildings, for instance, is generally higher than the compactness of medium density residential buildings.
road width	The shortest distance between two road edges (an edge is defined at a road curb). Sample measurements are made at road locations spatially dispersed within a scene.	The urban road width planning specifications are different for different land-uses.
road-building distance	The shortest distance between a building polygon and the nearest road edge. The distances between all buildings in a scene and their corresponding nearest road edge are measured.	This is a spatial relation feature (truly high-level). Buildings in commercial areas are generally closer to roads than in residential areas.
building-building distance	The distance between a building polygon and the nearest building polygon.	A high-level feature which models the building density of scene.
building density	Building density = $\frac{A_{buildings}}{A_{scene}}$, where $A_{buildings}$= sum of all building areas in a scene and A_{scene} = total scene area.	This should be a strong contextual feature due to terminology such as 'residential low / medium / high density'.
road-building distance to road width	Ratio of *road-building* distance to *road width*.	In a commercial / industrial scene, buildings are normally close to wide roads. In an informal scene, buildings can be far away from narrow roads.
building-building distance to building area	Ratio of *building-building* distance to *building area*.	In a commercial / industrial scene, large buildings normally exist close to one another, as opposed to residential scenes.

2. *Stdev building area*
3. *Mean building compactness*
4. *Stdev building compactness*
5. *Mean road-building distance*
6. *Stdev road-building distance*
7. *Mean road width*
8. *Stdev road width*
9. *Mean building-building distance*
10. *Stdev building-building distance*
11. *Building density*
12. *Mean road-building distance / mean road width*
13. *Mean building-building distance / mean building area*

The complete set of high-level features were extracted from each sample scene. Thus a high dimensional feature space exists. This can be represented as an observation matrix $X(i, j)$, where rows i correspond to observations (scene samples), and columns j to features.

6.3.7 Multivariate statistical visualization

Similarities / dissimilarities now need to be assessed in the scene observations of the same land-use. In other words we need to establish whether land-use classes can be separated in the high-level feature space. The features that are significant in separating classes also need to be established. The *Andrews plot* and the *glyph plot* are multivariate (high-dimensional) visualization tools useful for for this purpose. These plots, available in the Matlab statistical toolbox, were generated from the observation matrix X. The Andrews plot (Andrews, 1972) represents each observation by a function $f(t)$ of a continuous arbitrary variable t over the interval $[0, 1]$. $f(t)$ is defined for the *ith* observation in X as

$$f(t) = X(i, 1)/\sqrt{2} + X(i, 2)\sin(2\pi t) + X(i, 3)\cos(2\pi t) + ... \qquad (6.2)$$

An example of an Andrews plot is shown below in figure 6.9. The plot is based on 6 of the features listed in section 6.3.6 that were observed for the five land-use scenes listed in section 6.3.1. It is a plot of the function $f(t)$. $f(t)$ is on the y axis and t is on the x axis. As can be seen in the figure, plots of the

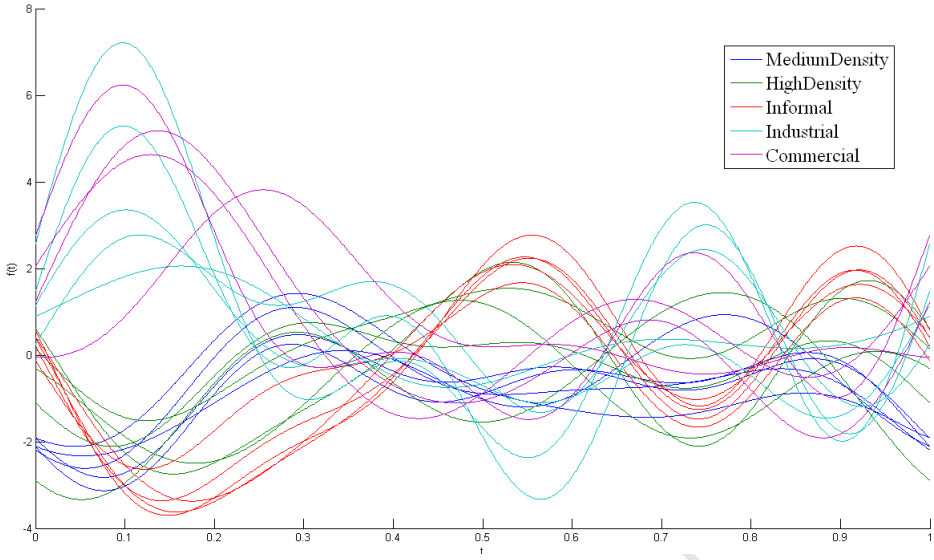


Figure 6.9: Example of an Andrews plot generated for 5 land-use classes and 6 features.
Plots of the same land-use follow a similar pathway. There is thus a degree of separation of land-use classes based on the 6 features.

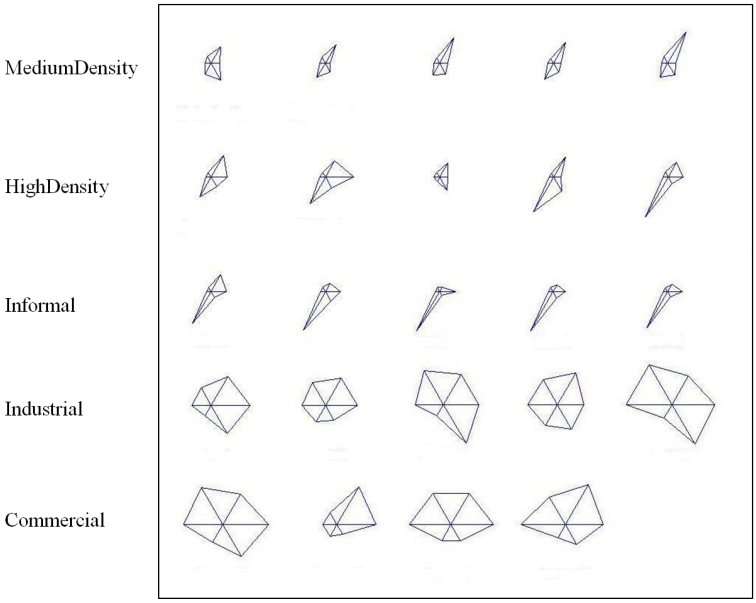


Figure 6.10: Example of a glyph plot generated for the same 5 land-use classes and 6 features.
Stars of the same land-use class have similar shapes, which means that classes can be separated based on the 6 features.

same land-use tend to follow a similar pathway, especially for the informal class in red. This tells us that there is a degree of discrimination among classes based on the 6 features. Thus the Andrews plot reveals information about the separation of classes.

The glyph plot represents each observation as a star, where the jth spoke in a star is proportional in length to the jth feature of that observation. Similarities in observation data can be established based on the shapes of the stars. Figure 6.10 shows an example of a glyph plot, generated for the same 5 land-use classes and 6 features. Each star corresponds to a single scene sample. Each star has six spokes, which correspond to the 6 features. The plot shows that stars of the same land-use class have similar shapes. This tells us that a unique signature exists for each land-use class based on the 6 features.

Thus like the Andrews plot, the glyph plot gives cues to the separation, or discrimination of classes. The glyph plot also tells us which features are significant in separating classes. This is done by observing which spokes in the star are prominent in characterizing classes. For instance the informal class is strongly characterised by stars with a long spoke in the 7 o'clock position (bottom left). This spoke would correspond to a particular feature, and this feature would thus be important in characterizing informal scenes. Further explanation and analysis of the Andrews and glyph plots are in the results section of this thesis (section 7.3.2), where the full 13 features listed in section 6.3.6 were analysed.

6.3.8 SVM feature selection

The 13 features described in section 6.3.6 are possibly redundant. It is desirable to establish a minimum subset of features that optimally separates land-use classes. This feature subset can be regarded as a 'scene descriptor', and provides a more meaningful definition of context, meeting the objectives of this dissertation. This feature space gives us a robust understanding of what characterises context. The feature space can be regarded as a contextual model within the image understanding framework. It can be considered general parameter constraints to the ideal description of particular scene types. Thus it can be utilized to constrain and improve bottom-up object classification results (top-down analysis). For example, to improve object classification results of a scene that was classified as 'industrial', the 'industrial' feature space should be used as a constraint, to produce a high-quality final scene description.

To establish an optimum subset, a feature selection technique (section 2.6) is required. A more meaningful analysis would be to rank each feature according to its significance, and then select a subset based on this ranking.

To perform feature ranking, an SVM feature ranking estimate based on a Random Forest (RF), proposed in (Chen and Lin, 2006), was carried out

using the LIBSVM (Chang and Lin, 2001) software package. The procedure works by computing the difference in RF accuracy measures after elimination of features. The procedure was applied to the observation matrix X. A technical overview of the procedure is as follows.

Given the observation matrix (training set) $X(i, j)$, where rows i correspond to observations (scene samples), and columns j to features, the procedure works by allocating an importance index to each feature based on an RF. An RF is a classification method (see section 3.5), but can also be used to provide feature importance. The RF feature importance allocator works as follows:

1. Split the training set into two parts.
2. Obtain an accuracy measure A by training the first and using an RF classifier to predict the second.
3. For a feature j, delete its values in the second set and obtain another accuracy measure B. The difference between the two accuracy measures $B - A$ gives an indication of the importance of feature j.

6.4 Development of an Automated Land-use Classification Routine

The concept of urban aerial image understanding is that higher quality object classification results will be produced if the semantics of a scene are recognised, i.e., a scene decomposes into land-use regions, which further decompose into urban objects (see Porway et al. (2008)). Instead of analyzing each object in an entire scene (as with standard image classifiers), groups of objects are analyzed within sub-scenes of the entire scene. With the knowledge of land-use type (context) for a sub-scene, top-down image understanding methods can be employed to produce a high quality scene description for that sub-scene. Each sub-scene will be treated differently and independently according to its land-use type.

A required stage in this system is automated land-use classification. Automated land-use classification refers to the production of a land-use classification map from a raw image of an urban scene without user intervention. This requires an automated segmentation of land-use regions, and then a scene-matching of those regions.

An automated segmentation of land-use regions is problematic, as there are no well-defined generic land-use discriminatory features. Our proposed approach is to segment 'block' regions in the original scene, by extracting those regions that fall within a closed loop of the bottom-up road data. The rationale for choosing block regions is not that each block region should consist of a different land-use, but that land-use regions will often be separated

by road networks. Within one land-use region there could be many blocks. The assumption with this approach is that an individual block falls within a closed loop of roads. Thus the approach will have limited effectiveness for scenes where this is not the case.

Thus the proposed scene hierarchy is a decomposition of a scene into road blocks, and then road blocks into urban objects. The objects within each block will be treated independently according to the land-use of that block. This model can easily be extended to produce land-use classification maps desirable for certain land-use applications, by merging blocks with the same land-use.

Figure 6.11 depicts a theoretical overview of the proposed urban image understanding system. The summation sign represents a combining / fusion of knowledge. Starting with a raw image, low-level features (colour, shape, texture) are extracted to produce an initial incomplete bottom-up object classification. Block regions and high-level features are then extracted from this bottom-up data. A land-use classification of block regions is then produced from these data. For each block region, this land-use knowledge (context) will be combined with the initial bottom-up features to produce a final complete scene description. This research focuses on producing a block classification map (solid lines). The dashed lines indicate future research intentions (the top-down phase).

Figure 6.12 offers a further conceptualization of the proposed automated land-use classification approach. It is essentially a 4 stage procedure:

1. Bottom-up object classification is performed
2. Road classification results are considered
3. Block regions are obtained from the road data
4. Each block is classified to a land-use

In more detail, the automated land use classification approach works as follows:

1. Segmentation and classification of building, road, vegetation and pavement objects (vegetation and pavement included to provide a more reliable overall scene classification) in a large scale (consisting of several different land-use types) scene based on low-level features. This is the bottom-up phase.
2. Extract block regions from the road extraction results using a novel technique.
3. Segment blocks in the original scene using these block region data. This is a classification-based-segmentation (see section 5.2.3).

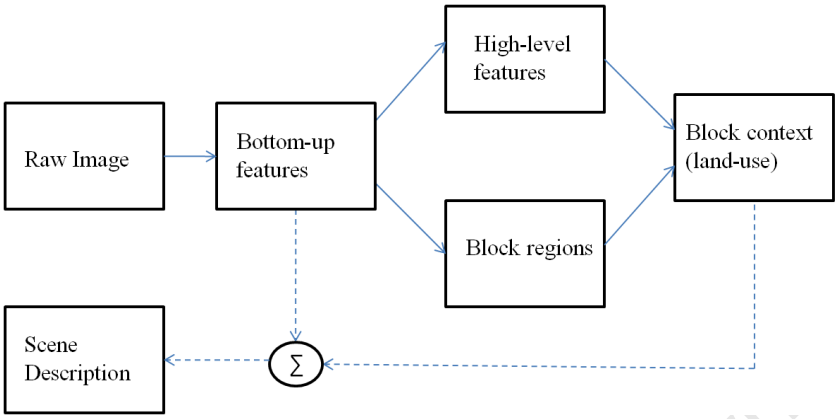


Figure 6.11: Theoretical overview of our urban image understanding system. The summation sign represents represents combining / fusion of knowledge. 'Bottom-up features' refer to the results of the initial object classification. 'High-level features' refers to spatial relation and contextual features based on the bottom-up objects (e.g. *road-building distance, building density*). 'Block context' refers to urban blocks with known land-use. Initially, bottom-up features are extracted from a raw image. Block regions are then obtained from these bottom-up features. Block regions are classified to a land-use based on high-level features. Future research intentions (dotted line) are to use this land-use contextual knowledge together with the initial bottom-up features to produce a final high quality scene description.

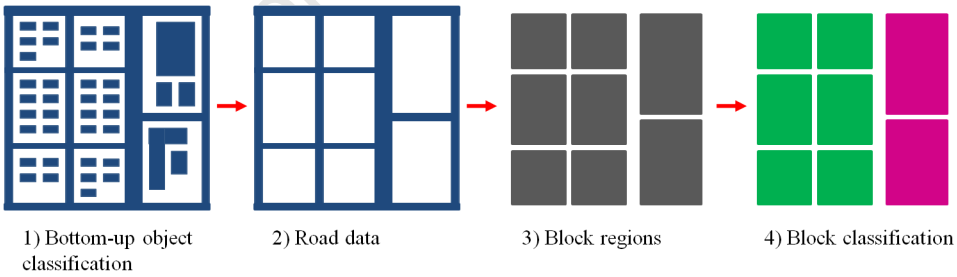


Figure 6.12: Automated land-use classification concept.
The figure offers further conceptualization of our automated land-use classification routine. 1) Bottom-up object classification is performed on a raw image, 2) the road classification results are considered for further analysis, 3) block regions are extracted from road classification results, 4) land-use classification is performed on the block regions based on high-level features of the object classification results in step 1).

4. Classify each block to one of the land-use types listed in section 6.3.1 based on high-level features extracted from the object classification results in step 1.

6.4.1 Technical overview of bottom-up phase

In chapters 2 - 4 a review of bottom-up classification literature is presented, with the intention of identifying the current state-of-the-art. It was concluded that Support Vector Machine (SVM) recognition of image objects, as proposed in (Tzotsos, 2008), is effective in urban object classification. This is the technique proposed for the bottom-up phase. A technical overview will now be detailed. For a raw image of an urban scene, objects are extracted in two stages:

1. Object segmentation
2. Object classification

Object Segmentation

Object segmentation was performed using a segmentation routine developed by Baatz and Schape (2000) and currently implemented in Definiens' eCognition software (Benz et al., 2004). It works by locally minimizing the average heterogeneity of image objects based on the colour of objects, the shape of objects, and the smoothness of borders of objects. The heterogeneity threshold is controlled by a user-defined scale value. An optimal scale value was chosen by heuristic visual analysis of different scale values. The goal during the visual analysis was to determine a scale value that causes the majority of objects to be properly segmented. Since different urban objects require different segmentation scale values, for a certain scale value some objects will be over-segmented whereas some will be under-segmented (see section 4.3.3). A scale value was chosen that causes an over-segmentation of objects, since proper segmentation of objects can then be recovered by merging segments after classification.

Object Classification

Classification of objects was performed as follows:

1. To train the classifier, sample objects were hand-labeled in eCognition. The samples were representative and spatially dispersed.
2. An initial, exhaustive set of colour, shape and textural (low-level) features were then selected from eCognition's built-in set of features.

3. Feature selection (refer to section 2.6) was then performed using eCognition's 'Feature Space Optimization' routine to reduce the dimensionality of the initial set of features. This routine works by choosing that subset of features that results in the average maximum separation distance (Euclidean) between the sample classes selected in step 1. For a maximum subset size s_{max}, the routine uses an exhaustive search to analyze all combinations from 1 to s_{max}. It is thus computationally expensive, but the advantage is that all combinations are taken into account.

4. Unknown objects were then classified using an SVM algorithm proposed in (Fan et al., 2005) and implemented in LIBSVM (Chang and Lin, 2001). This was accomplished using the eCognition Software Development Kit (SDK). In more detail:

   (a) Sample object feature values (training data) were exported from eCognition.

   (b) The SVM classifier was trained using these sample data, using the command *svm-train* in the LIBSVM package.

   (c) The feature values of unknown objects were exported from eCognition.

   (d) These unknown objects were then classified using the command *svm-predict* in the LIBSVM package. The classification results were then imported back into eCognition for display purposes.

Step 4. can also be performed using eCognition's built-in fuzzy Nearest Neighbour (NN) classification tool. This routine works in the same way as the KNN discussed in section **??**, except that a fuzzy set (see section 3.4), as opposed to a crisp set, is defined on the Euclidean distances between unknown data points and the nearest training points. After experimentation on a Cape Town aerial image dataset it was observed by visual analysis that this classifier produces classification results of similar quality to the SVM classifier. The eCognition fuzzy NN classifier was thus used in this research since it is simpler to use (a standard tool in eCognition).

6.4.2 Technical overview of block extraction

With the initial bottom-up object classification results, blocks can now be extracted from the road classification data. A 'block' is a region that falls within a closed loop of roads. The bottom-up road data are incomplete due to the limitations of automated classification, and thus it is not a trivial problem of obtaining those regions that fall within a closed loop of road regions.

A novel routine is proposed to extract block regions from incomplete road data. It works by extracting those regions from the road data that have a low 'road point' density. Road points are those points obtained from road classification results (the conversion of road classification polygons into points) (see figure 6.14 below). The assumption is that even with incomplete road data, block regions will be sparsely populated with road points.

An example of an urban scene in the Landsdowne area will be used to describe the routine. The scene is shown in figure 6.13. The input to the block extraction algorithm is road points obtained from an initial bottom-up road classification (see figure 6.14). The algorithm works as follows:

1. A Delaunay triangulation (Lattuada and Raper, 1996) is generated over the road points (see figure 6.15). A Delaunay triangulation imposes constraints that no point lies within the circum-circle of any generated triangle. This results in a maximization of the minimum angle of all triangles.

2. All 'short edges' in the triangulation are identified (figure 6.16). A 'short edge' is any edge of a triangle in the triangulation that has a length below a certain threshold. An appropriate threshold has to be chosen based on visual analysis of the results in figure 6.17.

3. All 'long edges' (edges with a length above the threshold value) are considered for further analysis. In a connected component analysis, connect triangles under the following criterion. Two triangles are connected if they are:

   (a) adjacent, and
   (b) share a 'long edge'.

   A set of regions then exists (the connected triangle components), which will refer to as 'block regions' (figure 6.17). Each block region is made up of a set of triangles. Thus each block region is essentially a set of points (the triangle vertices).

4. To obtain final block region estimates, generate a convex hull around the points of each block region (figure 6.18). The convex hull of a set of points P is the minimum convex set containing P (Brown, 1979). The convex hulls are the final block estimates.

The resulting block extraction consists of a set of polygons that estimate road block regions (figure 6.19). These data are used to segment the original scene again, i.e., to perform block segmentation. Figure 6.20 shows the segmentation of the original Landsdowne scene using the obtained block estimate data. This is a form of classification-based-segmentation (refer to section 5.2.3), i.e., the scene is segmented a second time based on information obtained from an initial object segmentation and classification.

Figure 6.13: Example of an urban scene in the Landsdowne area.
This scene is shown as an example to describe the block extraction routine.

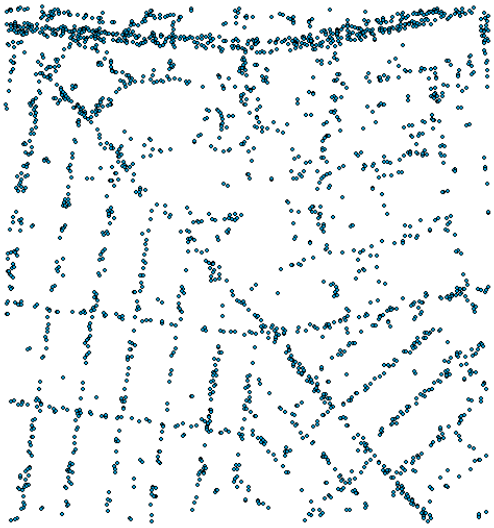


Figure 6.14: Road points.
'Road points' are those points obtained from bottom-up road classification data. The points in this figure were obtained by performing object classification on the above Landsdowne scene, and then converting the road classification results into points.

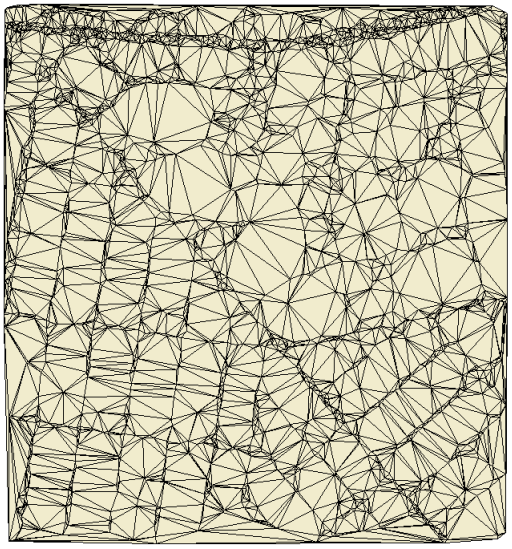


Figure 6.15: Delaunay triangulation generated over road points

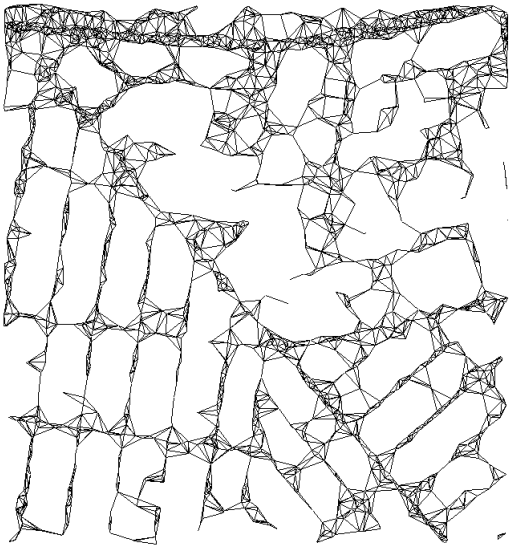


Figure 6.16: Short edges obtained from Delaunay triangulation.
The figure displays all 'short edges' in the triangulation. A 'short edge' is any edge of a triangle in the triangulation that has a length below a certain threshold. The large white spaces surrounded by these short edges are assumed to be road block regions.These block regions are identified in the next step (figure 6.17).

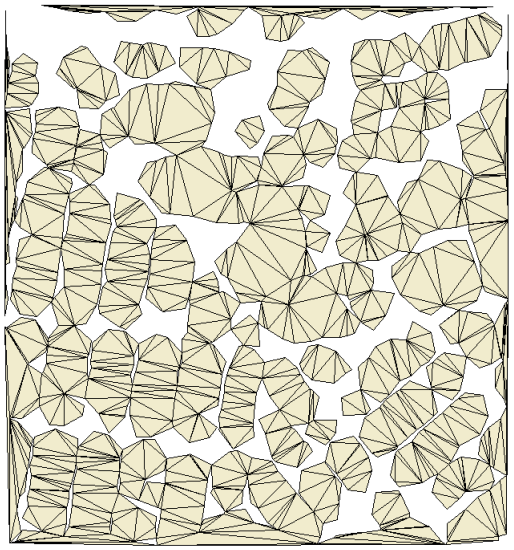


Figure 6.17: Block regions.
Block regions are those white space areas surrounded by short edges (figure 6.16). The regions are identified by connecting those triangles that are adjacent and share a common 'long edge'. A 'long edge' is any edge that was not labeled as 'short edge'.

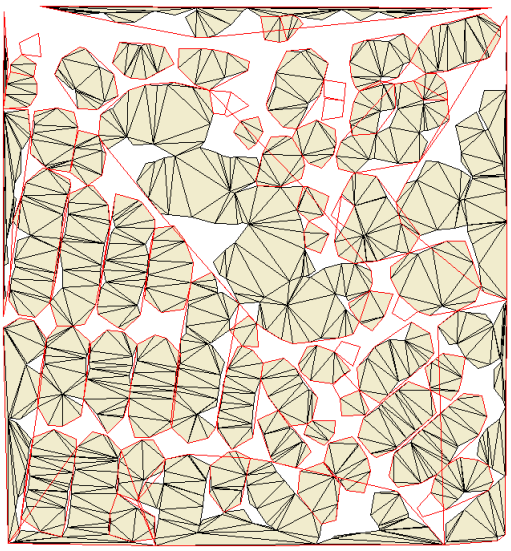


Figure 6.18: Convex hull generated over block regions.
Convex hulls (the red polygons outlining the block regions) are generated over the points of each block region. A convex hull can be visualized as an elastic band that has been stretched open to encompass a given set of points. These convex hulls are the final block estimates.

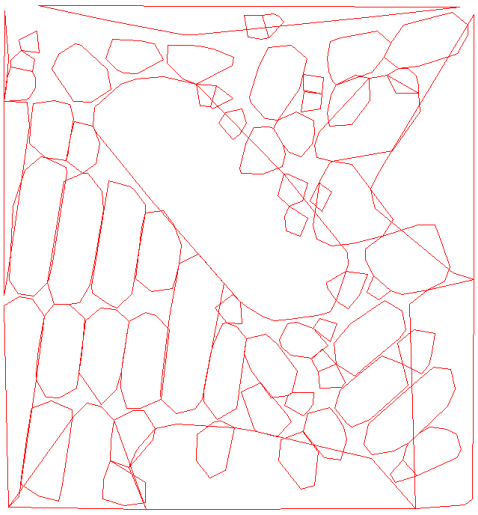


Figure 6.19: Resulting block extraction.

The final result consists of a set of polygons that estimate block regions in the input image.

6.4.3 Block classification

Once an image is segmented into block regions (figure 6.20), these block regions are classified to a land-use type based on high-level features extracted from the incomplete bottom-up object classification results. This is a form of scene-matching within the image understanding framework (section 5.2). The objective in this stage is simply to demonstrate that land-use scenes can be discriminated based on high-level features, and not to test the performance of a classification algorithm. Thus a complex machine learning classifier is not necessary.

High-level features similar to those in table 6.1 were formulated in eCognition. The high-level features were formulated in eCognition to be measured off a test image in a similar manner to that discussed in section 6.3.6. An optimum subset of high-level features was established using eCognition's Feature Space Optimization' tool with representative block samples. Blocks were classified using a simple rule-set formulated in eCognition based on the high-level feature subset.

A technical overview of the ruleset is as follows. For a block region b to classify to a land-use type l, based on a high-level feature subset $(f_1, f_2, ..., f_n)$, high-level feature values $b_{f1}, b_{f2}, ..., b_{fn}$ are measured off b. The rule-set has the following form:

IF $x_{f1} > b_{f1} > y_{f1}$ AND $x_{f1} > b_{f2} > y_{f2}$ .... AND $x_{f1} > b_{fn} > y_{fn}$ THEN $b = l$

where $x_{f1,f2,...,fn}$ and $y_{f1,f2,...,fn}$ are the decision border values of the

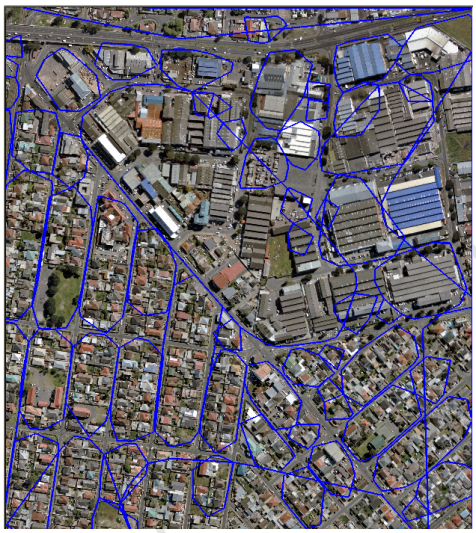


Figure 6.20: Block segmentation of Landsdowne scene.
The figure shows the final block extraction results overlaid on the original image. The blue lines indicate an estimate of block regions (not roads!). The block extraction routine is designed to segment block regions in an input scene (regions surrounded by a closed loop of roads), in order to obtain a partitioning of land-use regions. In this particular segmentation of the Landsdowne scene, land-use regions (industrial in the upper right and residential in the lower left) have been separated by the segmentation, even though some blocks have not been properly segmented. This is the desired result.

high-level feature space of sample blocks. Decision border values are determined for each land-use class by analysing various sample blocks of known land-use (training blocks). A feature value range is determined for a given high-level feature and a given class.

In simpler, non-mathematical language, this ruleset can be described in the following way: for a certain block to classify, measure high-level feature values off this block. If the values of each high-level feature fall between the decision border values of a certain land-use class, classify the block to that class.

6.5 Discussion

An approach to urban aerial image understanding has been presented, which comprises of a methodology for constructing a contextual model, and an automated land-use classification routine.

The urban contextual model proposed in section 6.3 is essentially a set of optimum high-level features chosen from a list of initial features (table 6.1). This list is not necessarily an exhaustive set. There are potentially many more features that might discriminate land-use scenes. The choice of features was partially influenced by time constraints and resources. However, if this set is able to discriminate scenes in an experimental test dataset, it would demonstrate the discriminatory power of features of this type. The significance of demonstrating the effectiveness of features of this type is that they can: 1) be used to perform land-use classification and 2) be used in top-down analysis. To elaborate on 2), if the feature set is shown to discriminate scene types, these features can be regarded as a contribution to a unique signature, or scene-descriptor, for each land-use type. They can thus be used to define the 'context' of a scene. The unique feature space for each land-use scene can be used as a contextual model to improve bottom-up object classification results.

The proposed land-use classification routine (section 6.4) assesses whether features of this type can be used to discriminate land-use scenes in an automated manner. The difference between this analysis and the former analysis is that the features are extracted from incomplete bottom-up object classification results, as opposed to manual (near complete) digitization results. Thus the potential of current bottom-up extractors can be tested in recognizing context.

A limitation of the proposed automated land-use classification routine is that it will only work upon success of the automated land-use segmentation. The proposed land-use segmentation routine is limited by an important assumption: land-use regions are separated by road blocks (closed loops of roads). If this is not the case for a given input scene, inaccurate land-use segmentation results may be generated. A correct land-use segmentation is

pivotal to the functioning of the rest of the system. Thus the biggest limitation with this system is the initial road block assumption. More research is required on the development of a robust land-use segmenter.

The block classifier is simple rule-based classification routine based on observed feature ranges in a small training dataset. The objective is to test the effectiveness of high-level features (automatically extracted from incomplete bottom-up data) in discriminating land-use classes. If these features are effective in causing reasonable accurate block classification results on experimental datasets, a more robust land-use classifier should be considered based on features of this type. A robust classifier should make use of a well-established machine learning algorithm, such as one of those presented in chapter 3 of this thesis. Furthermore, the classifier should be trained from a large and representative training set.

To test the effectiveness of the methodologies proposed in this chapter, experiments were conducted on various test datasets. The next chapter will detail experimental results.

Chapter 7

Results and Analysis

7.1 Introduction

In the previous chapter an urban land-use contextual model is described. The contextual model is constructed by extracting high-level features from manually labeled urban scenes. Statistical and machine learning tests are used to assess whether land-use scenes can be discriminated based on the high-level features, and to establish features that are significant in discriminating objects in a scene.

An automated land-use classification routine is then described, where an urban scene is segmented and classified into land-use zones. To perform the urban contextual analysis and test the effectiveness of the land-use classification routine, experiments were conducted on a test dataset.

In this chapter, experimental results will be presented, along with qualitative and quantitative analyses. Initially, the test dataset will be described. Results and analysis of contextual model tests will then be presented. Results obtained from applying the automated land-use classification routine to the dataset will then be presented.

7.2 Test Data

The dataset used for experimentation is 2007 ortho-rectified RGB aerial imagery of the greater Cape Town region, South Africa. The ground sample distance is 20cm. No image enhancement procedures were conducted on the dataset. Various scene samples were extracted from the dataset for experimentation.

7.3 Contextual Model Results

7.3.1 Manual high-level feature extraction from sample scenes

To construct the contextual model, the following land-use scene samples were extracted from spatially dispersed and representative regions in the test dataset. The suburb name of each scene sample will be listed for completeness. See figures 6.3 - 6.7 in the previous chapter for an example scene for each land-use.

- 5 x residential medium density samples: Newlands, Plumstead, Tokai, Seapoint, and Gardens.
- 5 x residential high density samples: Crossroads, Tafelsig, Seawinds, Bonteheuwel, and Mitchells Plain.
- 5 x informal settlement samples: Boys Town (near Cape Town International Airport), Nyanga, Seawinds, Charlesville and Gugulethu.
- 5 x industrial samples, Paarden Eiland, Parow, Retreat, Wetton and Hanover Park.
- 4 x commercial samples: Claremont, Victoria Street, Wynberg and Bellville.

In every scene sample, building and road objects were digitized through use of ESRI ArcMap's digitization tool (Esri, 2010). Figure 7.1 shows the digitization of the Seawinds sample scene. The 13 features described in section 6.3.6 were extracted from the building and road polygon database of each scene sample. Table 7.1 shows an excerpt of the observation matrix.

7.3.2 Multivariate visualization results

The observation data were then visualized using the Andrews and glyph plots discussed in section 6.3.7 to aid analysis. The results of the Andrews plot are shown in figure 7.2. Each plot represents a sample scene. The results show that there are consistencies in observations of the same land-use type, since there are no major deviations in the pathway of plots of the same land-use. There are major deviations, however, in the pathways of different land-use plots, which means that there is a degree of separation of classes. The result shows that land-use classes can be separated based solely on the high-level feature set.

The results of the glyph plot are shown in figure 7.3. The y axis denotes land-use type and the x axis denotes the various sample scenes that were observed. Thus each star in the plot corresponds to a single sample scene (about 5 sample scenes were observed for each land-use scene). In the

Figure 7.1: Digitization of Seawinds dataset
Building (red) and road (dark blue) features are manually digitized in each sample scene.

Table 7.1: Excerpt of sample scene observation matrix.
Just the residential scene samples and 3 high-level features are shown because of space constraints.

Scene type	*Mean building area* (m^2)	*Mean building compactness*	*Mean road width* (m)
Medium density	170.57	0.58	5.36
Medium density	159.43	0.60	5.14
Medium density	216.44	0.58	6.35
Medium density	168.49	0.61	6.93
Medium density	214.88	0.60	7.85
High density	71.62	0.76	5.06
High density	93.84	0.63	5.85
High density	122.55	0.68	5.29
High density	147.35	0.53	5.04
High density	68.46	0.73	5.58

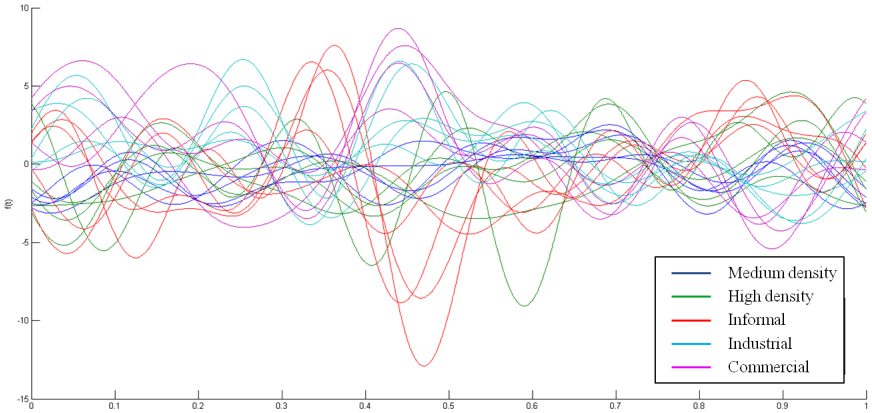


Figure 7.2: Andrews plot of high-level features observed from sample scenes. Plots of the same land-use follow similar pathways, which means that land-use scenes can be discriminated based on high-level features.

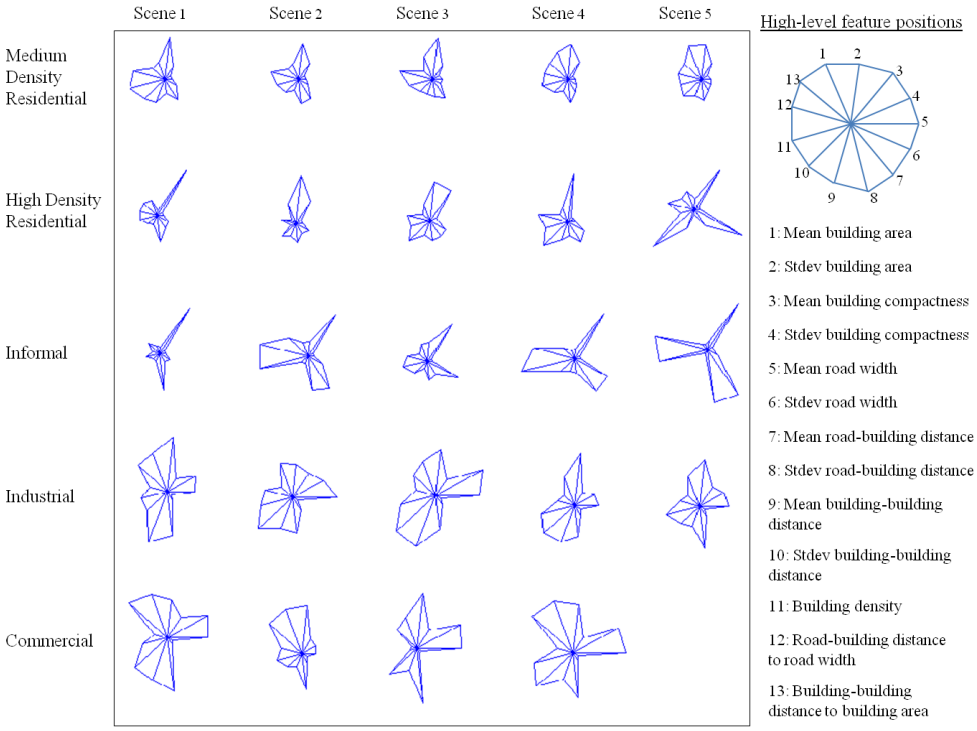


Figure 7.3: Glyph plot of sample scenes.
Each star represents a scene sample. Consistencies can be observed in the shapes of stars of the same land-use class, which means that a unique contextual signature can be recognised based on the high-level features.

figure we can see that scene samples of the same land-use exhibit a similarly shaped star. Thus a unique signature exists for a particular scene type, which confirms that the land-use scenes can be discriminated based solely on the high-level feature set. This demonstrates the potential effectiveness of features of this type in characterizing land-use context, and performing automated land-use detection.

The glyph plot shows which combinations of features are significant in separating land-use classes. Each star is made up of 13 spokes, which correspond to the 13 high-level features that were measured. On the right hand side of the plot a map shows the positions of the features on a star. The commercial plots for instance have a prominent fan at the 3 o'clock position. This tells us that the *stdev building compactness* and *mean road width* feature pair are prominent in characterizing commercial scenes. This same feature pair is prominent in industrial scenes, but not prominent in all the other scene types, which means that this feature pair can be used to separate industrial / commercial from the rest.

Informal settlements are strongly characterised by *mean building compactness*. The *stdev building area* feature is fairly consistent in characterizing scene types. This feature (spoke at the 1 o'clock position) is prominent especially for the medium density residential, industrial and commercial classes, but not prominent for the informal class. This is because informal buildings (shacks) are normally consistent in size.

Thus the glyph plot gives an indication as to what significantly characterises a particular land use type. This is useful to identify or single out a particular land use with respect to all the others, which can find application in land-use detection systems (e.g. informal settlement monitoring), or hierarchical land-use classification strategies.

7.3.3 High-level feature subset results: A definition of urban context

The SVM feature importance routine described in section 6.3.8 was then applied to the entire observation matrix in order to gain a quantitative understanding of what features are important in characterizing context, and to establish an optimum subset of features that causes greatest separation of land-use classes. To recap, the procedure works by computing the difference in RF accuracy measures after elimination of the feature concerned. Table 7.2 lists the feature importance results.

From these results a reasonable feature subset could be the first 6 features. The final feature subset is as follows:

$<$ *mean road width, stdev building area, mean building area, mean building compactness, stdev road width, stdev building-building distance* $>$

The results are consistent with the glyph plot results (*mean road width, stdev building area* and *mean building compactness* analyses described above).

Table 7.2: SVM feature importance results
The table displays a unitless ranking of each high-level feature according to its significance in discriminating land-use scenes.

Feature	**Feature importance**
Mean road width	4.748325
Stdev building area	4.090549
Mean building area	3.701236
Mean building compactness	2.161963
Stdev road width	2.098690
Stdev building-building distance	2.056538
Stdev building compactness	1.875651
Building density	1.860017
Mean building-building distance	1.031692
Mean road-building distance / mean road width	0.936287
Mean building-building distance / mean building area	0.797203
Stdev road-building distance	0.646435
Mean road-building distance	0.535549

Road width is the most highly ranked. This makes sense as there are particular urban planning road width specifications for different land use zones. This should result in consistent road width values for scenes of the same land use.

Mean building area is highly ranked probably because building type and thus building size is more or less consistent according to land use. Informal settlements buildings, for instance, are on average significantly smaller than buildings in most other zones. Residential buildings are normally slightly larger because of the income bracket difference. Industrial and commercial buildings can be large due to their functionality.

Land use functionality may also be correlated with *building compactness*, which causes its high ranking. Informal and high-density residential buildings are generally *compact* because of their size, low cost method of construction, and basic usage. Medium density residential buildings are less compact due to their more complex construction, to accommodate aesthetics and luxury.

Ironically, building density (ratio of total building area to total scene

area) is not highly ranked, which means that, contrary to common terminology ('medium density residential'), building density is not a major discriminating feature. This may be because high-density suburbs seem 'high density' due to the grid iron arrangement and close proximity of buildings. However, the average building size is also smaller so building density can be similar.

The subset feature space can be regarded as a scene descriptor, or a contextual model within the image understanding framework (see section 5.2.1). It can be considered general parameter constraints to the ideal description of particular scene types. Thus the feature space can be utilized to constrain and improve bottom-up object classification results (top-down analysis). For example, to improve object classification results of a scene that was classified as industrial, the industrial feature space would be used as a constraint to produce a higher-quality final scene description.

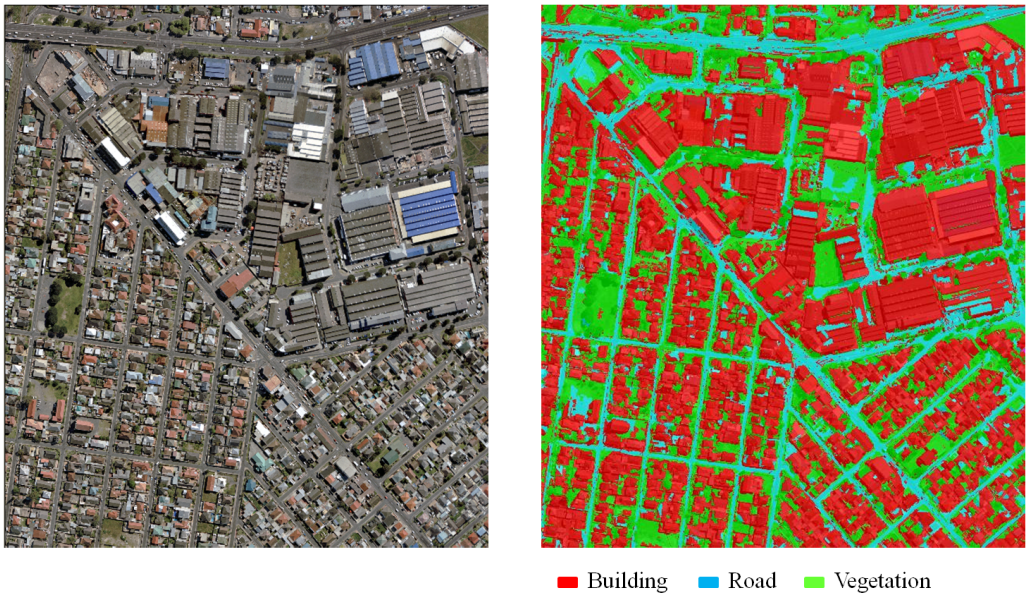
7.4 Automated Land-use Classification Results

The proposed automated land-use classification method described in section 6.4 was applied to 3 sample scenes from the same dataset described in section 7.2 to test the effectiveness of the approach. The 3 scenes are of the Landsdowne, N1 city and Eifendale suburbs. The Landsdowne scene consists of an industrial area and a medium density residential area. The N1 city and Elfindale scenes consist of a commercial area and a medium density residential area. It was difficult to find a scene with more than 3 land-use areas, or scenes containing mixed informal / high density residential areas, that were small enough to be processed in eCognition (version 7) software, since this version of eCognition has a limitation on the size of an image that can be processed..

7.4.1 Landsdowne dataset

The Lansdowne raw scene is shown below in figure 7.4 a). The scene consists of an industrial and a residential region. The industrial region is in the upper right, with characteristically larger buildings. It is apparent that the two land-use regions can be separated by the road network. It would thus make sense to segment the scene into road blocks in order to perform automated land-use classification.

Initial bottom-up object segmentation and classification was performed in eCognition as discussed in section 6.4.1. The scene was segmented using the eCognition multiresolution routine with a scale of 40. Segments were then classified using the eCognition fuzzy NN classifier based on 7 colour, texture and shape features obtained with eCognitions 'Feature Space Optimization' routine. See Appendix A for the results of the Feature Space Optimization.



a) b)

Figure 7.4: Landsdowne dataset: a) input scene; b) object classification results.
a) The scene consists of an industrial region in the upper right, and a residential region elsewhere. The two regions can be separated by the road network. b) Bottom-up object classification results: the scene is incomplete as there are misclassified segments.

a) b)

Figure 7.5: Landsdowne dataset: a) block segmentation results; b) block classification results.
a) Block segmentation results: block regions are more or less delineated. The industrial and residential regions are clearly separated by the segmentation, which is the desired result since we are interested in land-use classification. b) Block classification results: almost all residential blocks have been correctly classified, and a few industrial blocks have been misclassified as residential. A few regions have not been classified due to the incomplete block segmentation.

Object classification results are shown in figure 7.4 b). The figure shows that in general building and vegetation features are correctly extracted. There are, however, various misclassifications. Road segments have been misclassified as buildings (small red segments in the middle of the blue road areas). Building segments have been misclassified as road (blue segments within the red segments). Also, a lot of vegetation segments have been misclassified as building, especially in between buildings. This causes the building objects to be merged, which could prove problematic for certain building spatial relationship analyses. Thus the scene description is incomplete.

It is this limited competence of current automated object classifiers that has motivated the research in this thesis. If context (land-use type) can be recognised based on this incomplete data, top-down methods can be employed to improve the results, using a contextual model such as that described in this thesis (section 7.3.3), in order to produce a complete scene description.

The block extraction algorithm proposed in section 6.4.2 was used to extract block regions from the road point data. The original test scene was segmented again in eCognition based on these block regions. Figure 7.5 a) shows the block segmentation results. The block extraction consists of a set of oval-like polygons, that estimate block regions. Large gaps exist in between polygons. The gaps are caused by the block extraction algorithm's estimate of road regions. By visual analysis block regions are more or less delineated. Higher block segmentation quality exists in the residential area, where higher road classification results exist (figure 7.4 b)). This shows that the block extraction algorithm is only effective with high quality road classification input data. What is important (since we're interested in land-use classification) is that the industrial and residential regions are clearly separated by the segmentation. This demonstrates the effectiveness of the proposed algorithm in performing land-use segmentation.

7.4.2 N1 city dataset

The N1 city raw scene is shown in figure 7.6 a). The scene consists of a commercial area at the top and a medium density residential area at the bottom. A wide, main road separates the two land-use regions. What stands out is that the commercial area has much larger buildings, and a wide range of building sizes. The buildings are also much further apart from one another and the roads are wider.

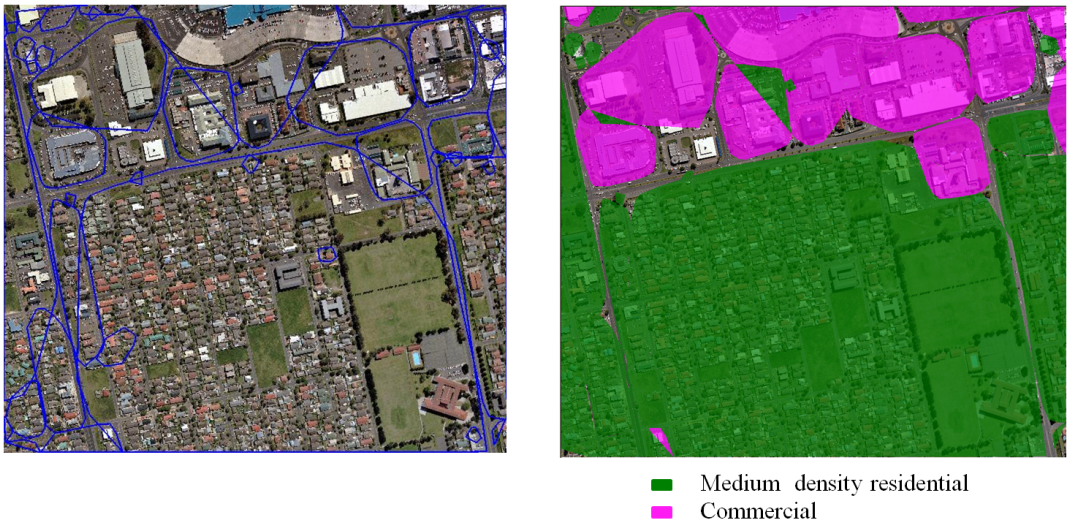
The same object segmentation and classification methodology as described in section 7.4.1 was applied to the dataset. Results are shown in figure 7.6 b). As with the Landsdowne dataset, the scene description is incomplete due to misclassified segments, especially commercial building segments classified as road, and road segments classified as building.

Block segmentation results are shown in figure 7.7 a). The results show

a) b)

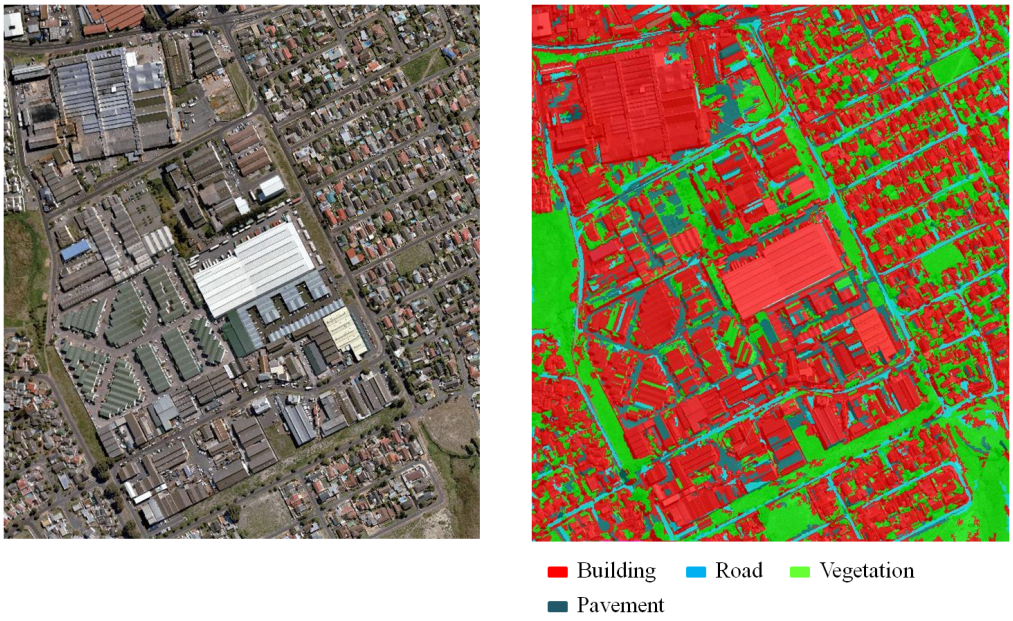
Figure 7.6: N1 city dataset: a) input scene; b) object classification results. a) The scene consists of a commercial area and a medium density residential area, separated by a road network. b) Object classification results: the scene is incomplete as there are misclassified segments.

that a large segment exists that contains many residential blocks. The commercial blocks have been more or less segmented. Some segments contain 2 or 3 blocks. This is not a concern since the main goal is for the segmentation to separate the two land-use regions. The correct separation is probably caused by the wide main road that marks the border between the two regions. Considering figure 7.6 b) the road classification data are prominent at this main road, but less prominent at the residential roads that form residential blocks. This suggests that the algorithm may be effective in separating land-use regions that are separated by main routes.



a) b)

Figure 7.7: N1 city dataset: a) block segmentation results; b) block classification results.
a) Block segmentation results. A large segment exists at the residential region that contains many blocks. This is not an issue since the main goal is to separate land-use regions. b) Block classification results: the unclassified areas are due to the incomplete block segmentation. A small residential segment has been misclassified as commercial. Small misclassifications such as these should be restored with an automated block merging algorithm.



a) b)

Figure 7.8: Elfindale dataset: a) input scene; b) object classification results. a) The scene consists of a commercial area and a medium density residential area, separated by a road network. b) Object classification results: the scene is incomplete as there are misclassified segments.

7.4.3 Elfindale dataset

The Eifindale scene is shown in figure 7.8 a). The scene consists of a commercial area and a residential area, separated by a road network. Object segmentation and classification of building, road, vegetation and pavement (included for a more reliable overall scene classification) objects was performed.

Object classification results are shown in figure 7.8 b). The results show that pavement objects could be useful for identifying commercial scenes, but only building and road objects were analyzed to stay consistent with the rest of the thesis.

Block segmentation results are shown in figure 7.9 a). Blocks are more or less delineated and the land-use regions have been separated. A large area to the upper right has not been segmented due to the poor road classification quality in that region.

a) b)

Figure 7.9: Elfindale dataset: a) block segmentation results; b) block classification results.
a) Block segmentation results: land-use regions have been separated. A large area to the upper right has not been segmented, which may be due to inadequate road classification results in that region. b) Block classification results: all block segments have been correctly classified.

7.4.4 Block classification

The block segments now need to be classified to a land-use type based on high-level features extracted from the incomplete bottom-up object classification results.

Since the concern is to operate only on the block polygon regions, the first step was to formulate a ruleset in eCognition to automatically eliminate the gaps in between the block regions. This ruleset classifies each segment as a 'block' if its compactness is above a certain threshold. See Appendix B for the full land-use classification ruleset. The following high-level features were then formulated in eCognition:

1. *mean building area*
2. *stdev building area*
3. *mean building compactness*
4. *stdev building compactness*
5. *mean building-road distance*
6. *stdev building-road distance*
7. *mean building-building distance*
8. *stdev building building distance*
9. *building density*

These features have the same definition as that in table 6.1 in the previous chapter. They were formulated for each block segment, i.e., to compute e.g. the *mean building area* feature for a given block, the average building area within that block region is computed. The high-level features were measured off representative block samples in all three datasets. Based on these measurements, eCognition's 'Feature Space Optimization' tool was used to compute a subset of features that causes maximum land use class separation.

The feature subset results are as follows.

<Mean building area, Stdev building area, Stdev road-building distance, Stdev building compactness, Stdev building-building distance>

These results are fairly similar to the feature importance results in table 7.2 (which were generated from manual object extraction), with the exception of *stdev road-building* distance. This could be because mis-classification of building objects often occur adjacent to roads, resulting in inconsistent road-building distances. Nevertheless, the feature subset similarities are a promising finding, as it means that current state-of-the-art automated urban object classifiers can produce results of high-enough quality to perform true semantic analysis.

The block classification rule-set discussed in 6.4.3 was then formulated based on the above four features and sample blocks in the three datasets. A commercial-residential classifier was formulated and applied to the N1 city and Elfindale datasets. An industrial-residential classifier was formulated for the Landsdowne dataset. See Appendix B for the full automated land-use classification ruleset. Results for the Landsdowne, N1 city and Elfindale datasets are shown in figures 7.5 b), 7.7 b) and 7.9 b), respectively.

Accuracy assessment

The block classification results for the three datasets show that there are large gaps where no classification exists. This is due to the incomplete block segmentation results. This is considered a limitation to a full aerial image understanding system, since when attempting to improve bottom-up object classification results, the objects within these unclassified regions will not be treated upon and improved. Nevertheless, a large portion of the imagery has been classified.

A potential solution to fill the gaps would be to merge block segments of the same land-use, and then classify those unclassified regions to the land-use of the adjacent segment with the largest percentage shared border. This would result in a final land-use classification map useful for top-down analysis or other land-use applications.

From a visual inspection of the land-use classification results, in the Landsdowne dataset (figure 7.5 b)) a few industrial segments have been misclassified as residential in the upper right. This is probably because these industrial blocks have similar features to the residential blocks. This is to be expected since within an industrial area certain sub-regions can be expected to be non-industrial like.

In the N1 city dataset (figure 7.7 b)) a few commercial segments have been misclassified as residential in the upper region, and a small residential segment has been misclassified as commercial in the lower left region. This particular segment is too small to be considered a block. Using the above-mentioned proposed block-merging approach, spurious segments such as these should be merged with their larger surrounding segments.

In the Elfindale dataset (figure 7.9 b)), a large area to the upper right has not been classified, due to the poor quality block segmentation in that area. This is due to the poor road classification quality in that region (figure 7.8 b). The available block regions have all been correctly classified.

Tables 7.3, 7.4 and 7.5 show the error matrices of block classification results for the Landsdowne, N1 city and Elfindale datasets respectively. In an error matrix $E_{i,j}$, the entry at the ith row and jth column is the number of blocks from the jth class that have been classified as the ith class. The user accuracy is defined as $e_{ii}/e_{i,sum}$, where e is an entry in the matrix, and $e_{i,sum}$ is the sum of row entries. The producer accuracy is defined as

Table 7.3: Error matrix for Landsdowne dataset block classification results.

Class	**industrial**	**residential**	**sum**	
industrial	22	1	23	
residential	17	62	79	
sum	39	63	102	
producer accuracy	0.56	0.98	**Overall**	**Kappa**
user accuracy	0.96	0.78	0.82	0.81

Table 7.4: Error matrix for N1 city dataset block classification results.

Class	**commercial**	**residential**	**sum**	
commercial	13	1	14	
residential	13	29	42	
sum	26	30	56	
producer accuracy	0.50	0.97	**Overall**	**Kappa**
user accuracy	0.93	0.69	0.75	0.73

Table 7.5: Error matrix for Elfindale dataset block classification results.

Class	**commercial**	**residential**	**sum**	
commercial	14	0	14	
residential	0	25	25	
sum	14	25	39	
producer accuracy	1.00	1.00	**Overall**	**Kappa**
user accuracy	1.00	1.00	1.00	1.00

$e_{ii}/e_{sum,i}$, where $e_{sum,i}$ is the sum of column entries. The overall accuracy Q is the percentage of blocks correctly classified, defined as $Q = \sum_i e_{ii}/N$, where N is the total number of objects. The Kappa Coefficient (K) is a statistical measure of the agreement between the classification map and the ground truthm, and is defined as $K = N \sum_i e_{ii} - \sum_i (e_{row} + e_{column})/N^2 - \sum_i (e_{row} + e_{column})$.

An overall block classification accuracy of 82%, 75% and 100%, and a Kappa coefficient of 0.81, 0.73, and 1.00 for the Landsdowne, N1 city and Elfindale datasets respectively, was achieved.

7.4.5 Discussion

The proposed routine was able to produce block classification accuracies of over 75% for the selected experimental test datasets. This demonstrates the potential to automatically detect land-use regions in an urban scene based solely on geometric measurements of incomplete building and road objects (high-level features).

Within the framework of an aerial image understanding system (section 5.2), the result shows that scene-matching, and thus a recognition of urban context, can be performed based on incomplete bottom-up data. This is a promising step toward improving the initial bottom-up results to produce a higher quality final scene description. This bottom-up improvement would be based on external expert knowledge in the form of a contextual model such as that proposed in this thesis (section 7.3). For instance, for those blocks that were classified as industrial in the Landsdowne dataset, the industrial feature space or 'scene descriptor' (section 7.3.3) would be employed to constrain and improve the bottom-up results within those block regions. Recapping the concept of image understanding, the idea is that a high-quality scene description should be produced albeit with low-quality bottom-up results. The bottom-up results need only be of high enough quality to enable reliable scene-matching, as was demonstrated by the automated land-use classification routine.

Block classification results of over 75% demonstrate the effectiveness of the four high-level features discussed in section 7.4.4 in discriminating land-use scenes. Classification was based on the simple rule-based classification routine presented in section 6.4.3. Note that the decision border values used by the classifier were determined based on sample blocks within the same datasets. A more robust and unbiased classifier should be trained by sample representative blocks in independent datasets. Due to time constraints and availability of data this could not be done, which is a limiting factor to the experimental routine. Nevertheless the results show that land-use scenes can be detected based on a small set of high-level features that have been automatically extracted from incomplete object classification results. The objective of this experiment was not necessarily to demonstrate the effec-

tiveness of the classification algorithm, but rather to test the discriminatory power of features of this type. A statistical or machine learning classifier such as those presented in chapters 2 and 3 should be considered for a more robust classification system.

As discussed in section 6.5, it must be noted that a limitation of this land-use classification routine is that it will only work if correct land-use segmentation is generated. The proposed automated land-use segmentation routine is based on the novel block extractor presented in section 6.4.2. The block extractor is limited by the assumption that land-use regions are separated by road blocks. The 3 test datasets used to test the routine all hold to this assumption. The land-use areas in these scenes can all be separated clearly by a road network, as can be seen in figures 7.4 a), 7.6 a) and 7.8 a). For input scenes where this is not the case, the entire system might be ineffective, since it pivots upon this initial assumption. Nevertheless, for scenes that do hold to this assumption, the experimental results have demonstrated the potential of the block extraction routine in providing a useable land-use segmentation (figures 7.5 a), 7.7 a) and 7.9 a)).

The results show that the block extraction algorithm will only work with road classification results of high enough quality. Considering the Landsdowne dataset, the road classification results in the residential region (bottom left of figure 7.4 b)) were proved to be of high enough quality to produce accurate block segmentation in that region (bottom left of figure 7.5 a)). It is evident in figure 7.4 b) that the road data are especially prominent in delineating residential blocks and less prominent in delineating industrial blocks. Thus the block segmentation results of the industrial region (top right of figure 7.5 a)) are incomplete, i.e., blocks are not clearly delineated.

Considering the N1 city residential road classification results in figure 7.6 b), the road classification data does not clearly delineate residential blocks (there is a lack of light blue road data in the residential area). Thus these residential blocks did not get segmented at all (figure 7.7 a)). The same can be said for the poor residential block segmentation results in the Elfindale dataset (figure 7.8 b)), especially in the upper right region. The road classification quality in this area (upper right of figure 7.8 b)) is poor, i.e., there is a lack of prominent light blue that delineates blocks.

In conclusion from the above analysis, the block extraction algorithm will only work with road classification results that are prominent in delineating and surrounding block regions.

Furthermore, consider the block classification errors in the Landsdowne and N1 city datasets, in figures 7.5 b) and 7.7 b) respectively. In the Landsdowne dataset, small industrial segments have been misclassified as residential. In the N1 city dataset, commercial segments have been misclassified as residential, and a small residential segment has been misclassified as commercial. It is important to note that almost all of these misclassified segments are small and spurious (they improperly estimated block regions). The results

indicate that classification accuracy is highly correlated with segmentation quality. In other words an accurate block segmentation is required for accurate overall land-use classification.

Thus the overall success of the system is highly dependent on the success of the initial land-use segmentation, which in turn is dependent on the quality of road classification results and the nature of the scene. This error propagation effect is considered a limitation to the system. Improvement is required in the area of land-use segmentation. Features other than geometric high-level features might be required for a more robust system. Contextual colour / texture scene features, as well as external ancillary data, might provide information needed for a more robust land-use segmentation routine.

Chapter 8

Conclusions

8.1 Conclusion

The objective of this dissertation is to characterise urban land-use context for the purpose of automated zoning from aerial imagery (refer to section 1.4). Automated zoning is useful for various land-use applications as well as top-down image understanding strategies. An automated zoning routine was proposed, that works by segmenting and classifying land-use regions based on bottom-up object classification data. Land-use regions are discriminated based on high-level features (e.g. *average road-building distance*, *average building size*) of the bottom-up objects. The routine was tested on experimental aerial image test datasets of the Cape Town region. Land-use classifications accuracies of over 75% were generated. This demonstrates the potential to perform land-use classification in an entirely automated fashion, based solely on geometric measurements of incomplete object classification data.

The motivation behind developing an automated zoning routine was to contribute to a full urban aerial image understanding system. The purpose of an aerial image understanding system is to improve object extraction in aerial images by exploiting context, in an attempt to emulate the human visual system.

A review of works that deal with image understanding was thus conducted. From the literature review a general framework for aerial image understanding was established. The concept of this framework is that a complete and idealized description of a scene is constructed, even if it is only partially depicted by image features. Thus unlike standard automated image interpretors that rely solely on sensation (absolute sensory measurements such as colour and texture), image understanding relies on sensation as well as perception, which is the way the human visual system operates.

The term 'contextual model' has been used to describe the formulation of perception. A contextual model is derived from expert knowledge and

training images, and can be regarded as a scene template or scene descriptor. Different scene types require a different contextual model. In the case of urban scenes, scene descriptions can be categorised according to land-use. The context of different land-use types thus need to be characterised.

A given urban scene may be made up of several different land-uses. A scene-matching methodology is thus required, where an appropriate contextual model is chosen based on initial bottom-up object extraction results. In a top-down analysis, the appropriate contextual model is used to improve bottom-up results.

In the literature a gap was identified in the area of defining a contextual model for different urban land-use types. In phase 1 of this research the question was thus asked "what is urban land-use context?". An urban contextual model was proposed, that consists of a set of high-level features (spatial relations and other contextual features). A set of 13 features were extracted from manually labelled sample scenes of different land-use types from a high resolution RGB aerial image dataset of the greater Cape Town region. Results of multivariate statistical visualization showed that a discrimination of land-use classes exists based on these features. This demonstrates the potential effectiveness of features of this type in characterizing land-use context, and performing automated land-use detection.

To answer the question of what is context, and gain a more robust definition of a contextual model, each of the 13 high-level features were ranked according to their significance in discriminating land-use types. This was accomplished with a Support Vector Machine (SVM) feature importance allocator. The results show that shape features such as *average road-width* and *average building area /compactness* are predominant in defining context.

In phase 2 of this research an automated land-use classification routine was developed, where automated segmentation and land-use classification of road block regions is carried out. Land-use classification is performed based on similar high-level features to those defined in phase l of the research. These features are automatically extracted from bottom-up object classification results. The idea is to assess whether scene-matching can be peformed based on the same contextual model proposed phase 1, but applied to automated feature extraction as opposed to manual feature extraction.

The land-use classification routine was tested on 3 urban scene datasets from the same RGB aerial imagery used in phase 1, to assess its effectiveness. Each scene is comprised of different land-use regions. Bottom-up object classification was performed with what is considered currently state-of-the-art; eCognition segmentation and classification based on colour, shape, and texture features. Results show a few mis-classifications of object segments, which means the scene description is incomplete.

Block regions were then extracted from the incomplete bottom-up road data. The proposed block extraction algorithm is a novel contribution. The original scenes were segmented again based on the resulting block region

estimates. Block segmentation results show a clear separation of land-use regions. This demonstrates the effectiveness of the proposed block extraction technique.

High-level features were then defined for each block region based on the bottom-up data. eCognition's feature selection tool was used to establish a subset of high-level features that causes maximum land-use class separation. The resulting subset is similar to that produced in phase 1, which was generated from manual image interpretation. This suggests that current state-of-the-art automated urban object classifiers can produce results of high-enough quality to perform true semantic analysis.

Blocks were classified to a land-use class based on the feature subset calculated in eCognition, yielding block classification accuracies of over 75% for the three datasets. This result shows that a recognition of context can be achieved from incomplete bottom-up results. This is an encouraging result within the framework of aerial image understanding. These bottom-up results can then be potentially improved based on external knowledge in the form of a contextual model such as the one proposed in phase 1.

The proposed automated zoning system has the following limitation. The system will only work if correct land-use segmentation is generated. Our proposed automated land-use segmentation routine is based on a block extraction algorithm, which is limited by the assumption that land-use regions are separated by road blocks. If this is not the case for a given input scene, inaccurate land-use segmentation results may be generated. Furthermore, the experimental results showed that accurate land-use classification accuracy is dependent on an accurate block segmentation. This is because almost all block mis-classifications occurred at inaccurate block segmentation locations. Experimental results also showed that the block extraction algorithm can only work well with accurate road classification data. Thus there is an inherent error propagation effect in the system, which may be a limitation. Improvement is required on the land-use segmentation algorithm.

8.2 Future Work

Scope for future research is considered in the following areas:

- In order to develop a more robust automated zoning routine, improvement is required in the area of land-use segmentation. Features other than geometric high-level features need to be tested for their effectiveness in segmenting land-use regions. A combination of geometric features, colour / textural scene features and external anxillary data might provide useful information for a more robust land-use segmentation and classification system.

- The 13 high-level features proposed in section 6.3.6 are an initial ex-

perimental set. There are potentially many more features that could be powerful in discriminating urban scene types. Other significant features could include *street pattern* (grid iron vs organic) features, as used in (Busgeeth et al., 2008), complex combinations of various spatial relation features, and features that take into account the semantics of a scene (e.g. land-use region decomposing into building clusters, which decompose into individual buildings). Furthermore, colour and texture features can be used, such the *tarred / gravel road* feature used in (Busgeeth et al., 2008). Objects other than buildings and roads could be considered, e.g. vegetation, cars. Vegetation density features might be useful in isolating residential scenes. The presence of a high density car object region indicates a commercial parking lot, which could accurately isolate commercial scenes. Various other urban planning land-use design specifications could be useful in discriminating scenes. The problem is essentially a reverse engineering of an urban planning problem, or procedural generation of urban scene models, in the computer science domain.

- Experimentation with more datasets is desirable to improve the automated zoning routine. Due to limitations of eCognition processing power, datasets of only a limited size could be experimented with in this study. Larger datasets consisting of more than two land-use regions are desirable to build a generalized contextual model and scene-matching methodology based on automated feature extraction results. A robust land-use classifier should be based on a machine learning algorithm such as those presented in chapter 3 of this thesis. The classifier needs to be trained from representitive land-use areas in training images. Datasets from cities in other parts of the world are attractive to achieve genericity. The same applies for constructing the contextual model from manually labeled samples. The more samples used from different and diverse cities / suburbs, the more robust the resulting contextual model will be.

- The most pertinent future work lies in the area of top-down analysis. Considering our block classification scheme, once a scene has been segmented and classified into land-use blocks, the bottom-up object extraction results within a block need to be improved based on the land-use of that block. This improvement will be induced by constraints of a contextual model such as the one proposed in this thesis. Our proposed contextual model is in the form of a high-level feature space (section 7.3.3). For example, to improve object classification results of a scene that was classified as industrial, the industrial feature space would be used as a constraint to produce a higher-quality final scene description. The contextual model needs to be constructed from

manually labeled (human interpreted) samples scenes in order to obtain a final scene description as close to human perception as possible. High-level features of bottom-up objects should be tested in this feature space. An analysis similar to that in (Matsuyama and Hwang, 1990) is proposed. If there are bottom-up spurious objects that cause inconsistencies in the model test, these objects are deleted. If objects are required in order to generate consistencies, they are instantiated. The goal is to obtain a final scene description that is consistent as possible with the contextual model. The Markov Random Field, as used in (Porway et al., 2008; Hernandez-Gracidas and Sucar, 2007), might be useful for this purpose.

Bibliography

Akcay, H. and Aksoy, S. (2008), 'Automatic detection of geospatial objects using multiple hierarchicl segmentations', *IEEE Transactions on Geoscience and Remote Sensing* **46**, 2097–2111.

Aksoy, S. and Akçay, H. (2005), 'Multi-resolution segmentation and shape analysis for remote sensing image classification', *Proc. 2nd Int. Conf. Recent Advances Space Technology, Istanbul, Turkey* .

Aksoy, S., Tusk, C., Koperski, K. and Marchisio, G. (2003), Scene modeling and image mining with a visual grammar, *in* C. Chen, ed., 'Frontiers of Remote Sensing Information Processing', World Scientific, Singapore, pp. 35–62.

Andrews, D. (1972), 'Plots of high-dimensional data', *International Biometric Society* **18**, 125–136.

Aplin, P. and Smith, G. (2008), 'Advances in object-based image classification', *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences,* **37**.

Baatz, M. and Schape, A. (2000), Multiresolution segmentation-an optimization approach for high quality multi-scale image segmentation, *in* J. Strobl, T. Blaschke and G. Griesebner, eds, 'Angewandte Geographische Informations-Verarbeitung XII', Wichmann Verlag,, Karlsruhe, pp. 12–23.

Ball, G. and Hall, D. (1965), *A novel method of data analysis and pattern classification*, Standford Research Institute, Menlo Park, California.

Bardossy, A. and Samaniego, L. (2002), 'Fuzzy rule based classification of remote sensing imagery', *IEEE Transactions on Geoscience and Remote Sensing* **40**, 362–374.

Benediktsson, J., Swain, P. H. and Ersoy, O. K. (1990), 'Neural network approaches versus statistical methods in classification of multisource remote sensing data', *IEEE Transactions on Geoscience and Remote Sensing* **28**, 540–552.

Benz, U., Hofmann, P., Willhauck, G., Lingenfelder, I. and Heynen, M. (2004), 'Multiresolution, object-orientated fuzzy analysis of remote sensing data for gis-ready information', *ISPRS Journal of Photogrammetry and Remote Sensing* **58**, 239–258.

Bishop, C. (1996), *Neural Networks for Pattern Recognition*, 1 edn, Oxford University Press, USA.
URL: *http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0198538642*

Blaschke, T. (2009), 'Object based image analysis for remote sensing', *International Society for Photogrammetry and Remote Sensing,* **65**, 2–6.

Breiman, L. (2001), 'Random forests', *Machine Learning* **45**, 5–32.

Brown, K. (1979), 'Voronoi diagrams from convex hulls', *Information processing letters,* **9**, 223–228.

Bruzzone, L. and Carlin, L. (2006), 'Multilevel context-based system for classification of very high spatial resolution images', *IEEE transactions on Geoscience and Remote Sensing* **44**, 4287–4308.

Busgeeth, K., Brits, A. and Whisken, J. (2008), Potential application of remote sensing in monitoring informal settlements in developing countries where complimentary data does not exist, *in* 'Planning Africa Conference', Johannesburg, South Africa, pp. 314–328.

Chang, C.-C. and Lin, C.-J. (2001), *LIBSVM: a library for support vector machines*. Date of access: 15 / 03 / 10).
URL: < *http://www.csie.ntu.edu.tw/ cjlin/libsvm*>

Chang, L.-Y. and Chen, C.-F. (2008), 'A multi-scale region growing segmentation for high resolution remotely sensed images', *Journal of Photogrammetry and Remote Sensing* **13**.

Chen, Y.-W. and Lin, C.-J. (2006), Combining svms with various feature selection strategies, *in* I. Guyon, M. Nikravesh, S. Gunn and L. Zadeh, eds, 'Feature extraction: Foundations and applications', Springer-Verlag, Berlin, pp. 315–324.

Durand, N., D. S. F. G. W. C. G. P. B. O. P. A. (2007), Ontology-based object recognition for remote sensing image interpretation, *in* 'IEEE International Conference on Tools with Artificial Intelligence', Patras, Greece, pp. 472–479.

Egenhofer, M. and Franzosa, R. (1991), 'Point-set topological spatial relations', *International Journal of Geographical Information Systems* **5**, 161–174.

Esri (2010), 'Esri arcmap user manual'. Online; accessed 20-October-2010. **URL:** *http://www.esri.com/software/arcgis/arcinfo/index.html*

Fan, R.-E., Chen, P.-H. and Lin, C.-J. (2005), 'Working set selection using second order information for training svm', *Journal of Machine Learning Research* **6**, 1889–1918.

Foody, G. (1996), 'Approaches for the production and evaluation of fuzzy land cover classification from remotely-sensed data', *InternationalJournal of Remote Sensing* **17**, 1317–1340.

Foody, G. (2002), 'Status of land cover classification accuracy assessment', *Remote Sensing of Environment* **80**, 185–201.

Foody, G. and Mather, A. (2004), 'Toward intelligent training of supervised image classifications: Directing training data acquisitions for svm classification', *Remote Sensing of Environment* **93**, 107–117.

Frank, A. (1991), Qualitative spatial reasoning about cardinal directions, *in* 'Proc. 7th Austrian Conference on Artificial Intelligence', Wien, Austria, pp. 157–167.

Friedl, M. A., Brodley, C. E. and Strahler, A. H. (1999), 'Maximizing land cover classification accuracies produced by decision trees at continental to global scales', *IEEE Transactions on Geoscience and Remote Sensing* **37**, 969–977.

German, G., West, G. and Gahegan, M. (1999), Statistical and ai techniques in gis classification: A comparison, *in* 'SIRC99 - The 11th Annual Colloquium of the Spatial Information Research Centre', University of Otago, Dunedin, New Zealand, pp. 2145–2152.

Goyal, R. and Egenhofer, M. (2000*a*), 'Cardinal directions between extended spatial objects', *IEEE Transactions on Knowledge and Data Engineering* **in press**.

Goyal, R. and Egenhofer, M. (2000*b*), Consistent queries over cardinal directions across different levels of detail, *in* 'IEEE 11th International Workshop on Database and Expert Systems Applications', Greenwich, UK, pp. 876–880.

Gregory, R. (1970), *The intelligent eye*, McGraw-Hill, New York.

Guyon, I. and Elisseeff, A. (2003), 'An introduction to variable and feature selection', *Journal of Machine Learning Research* **3**, 1157–1182.

Hansen, M., Dubayah, R. and Defries, R. (1996), 'Classification trees: An alternative to traditional land cover classifiers', *International Journal of Remote Sensing* **17**, 1075–1081.

Haralick, R. and Fu, K. (1983), Pattern recognition and classification, *in* R. Colwell, D. S. Simonett and F. T. Ulaby, eds, 'Manual of remote sensing, vol. 1 (2nd ed.)', American Society of Photogrammetry, Falls Church, VA, pp. 801–802.

Hernandez-Gracidas, C. and Sucar, L. (2007), Markov random fields and spatial information to improve automatic image annotation, *in* 'Proceedings of the Pacific-Rim Symposium on Image and Video Technology'.

Holland, J. (1992), 'Genetic algorithm', *Scientific American* **80**, 66–72.

Huang, C., Davis, L. and Townshend, J. (2002), 'An assessment of support vector machines for land cover classification', *International Journal of Remote Sensing* **23**, 725–749.

Inglada, J. (2007), 'Automatic recognition of man-made objects in high resolution optical remote sensing images by svm classification of geometric image features', *ISPRS Journal of Photogrammetry and Remote Sensing* **62**, 236–248.

Jensen, J. (2005), *Introductory digital image processing: A remote sensing perspective*, Pearson Prentice Hall.

Jin, Y. and Geman, S. (2006), Context and hierarchy in a probabilistic image model, *in* 'in CVPR', pp. 2145–2152.

Jing, F., Li, M., Zhang, L., Zhang, H. and Zhang, B. (2003), Learning in region-based image retrieval, *in* 'International Conference on Image and Video Retrieval'.

Jolliffe, I. (1986), *Principal Component Analysis*, Springer-Verlag, New York.

Kanellopoulos, I. and Wilkinson, G. (1997), 'Strategies and best practice for neural network image classification', *International Journal of Remote Sensing* **18**, 711–725.

Kavzoglu, T. and Mather, P. (2003), 'The use of backpropogating artificial neural networks in land cover classification', *International Journal of Remote Sensing* **24**, 4907–4938.

Kettig, R. and Landgrebe, D. (1976), 'Classification of multispectral image data by extraction and classification of homogeneous objects', *IEEE Transactions on Geoscience Electronics* **14**.

Laha, A., Pal, N. and Das, J. (2006), 'Land cover classification using fuzzy rules and aggregation of contextual information through evidence theory', *IEEE Transactions on Geoscience and Remote Sensing* **1**, 532–535.

Lattuada, R. and Raper, J. (1996), Applications of 3d delaunay triangulation algorithms in geoscientific modeling, *in* 'The Third International Conference/Workshop on Integrating GIS and Environmental Modeling', CD-ROM.

Liu, Y., Guo, Q. and Kelly, M. (2008), 'A framework of region-based spatial relations for non-overlapping features and its application in object based image analysis', *ISPRS Journal of Photogrammetry and Remote Sensing* **In Press, Corrected Proof**. <http://dx.doi.org/10.1016/j.isprsjprs.2008.01.007>.

Liu, Y., Zhang, D., Lu, G. and Ma, W.-Y. (2007), 'A survey of content-based image retrieval with high-level semantics', *Pattern Recognition* **40**, 262–282.

Longley, P., Goodchild, M., Maguire, D. and Rhind, D. (1990), *Geographic Information Systems and Science. Second Ed.*, John Wiley & Sons Inc., New York.

Lu, D. and Weng, Q. (2007), 'A survey of image classification methods and techniques for improving classification performance', *International Journal of Remote Sensing* **28**, 823–870.

Ma, W. and Manjunath, B. (1997), Netra: A toolbox for navigating large image databases, *in* 'Proceedings of the IEEE International Conference on Image Processing (ICIP)'.

Mather, P. (2004), *Computer Processing of Remotely-Sensed Images : An Introduction*, John Wiley & Sons.
URL: *http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0470849193*

Matsuyama, T. and Hwang, V.-S. (1990), *SIGMA: A knowledge-based aerial image understanding system*, Plenum Press, New York.

Mo, D.-K., Lin, H., Li, J., Sun, H. and Xiong, X.-J. (2007), 'Design and implementation of a high spatial resolution remote sensing image intelligent interpretation system', *Data Science Journal* **6**, 445–452.

Muchoney, D., Borak, J., Chi, H., Friedl, M., Gopal, S., Hodges, N., Morrow, N. and Strahler, A. (2000), 'Applications of the modis global supervised classification model to vegetation and land cover mapping of central america', *International Journal of Remote Sensing* **21**, 1115–1138.

Nedeljkovic, I. (2006), 'Image classification based on fuzzy logic', *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **34**, 1–6.

Oliva, A. and Torralba, A. (2001), 'Modeling the shape of the scene: A holistic representation of the spatial envelope', *IJCV* **42**, 145–175.

Pacifici, F., Del Frate, F., Emery, W., Gamba, P. and Chanussot, J. (2008), 'Urban mapping using coarse sar and optical data: Outcome of the 2007 grss data fusion contest', *IEEE Geoscience and Remote Sensing Letters* **5**, 331–335.

Pal, M. (2005), 'Random forest classifiers for remote sensing classification', *International Journal of Remote Sensing* **26**, 217–222.

Pal, M. and Mather, P. (2003), 'An assessment of the effectiveness of decision tree methods for land cover classification', *Remote Sensing of Environment* **86**, 554–565.

Pal, M. and Mather, P. (2005), 'Support vector machines for classification in remote sensing', *International Journal of Remote Sensing* **26**, 1007–1011.

Pal, M. and Mather, P. (2006), 'Some issues in the classification of dais hyperspectral data', *International Journal of Remote Sensing* **27**, 2895–2916.

Paola, J. D. and Schowengerdt, R. (1995), 'A review and analysis of back-propagation neural networks for classification of remotely sensed multi-spectral imagery', *International Journal of Remote Sensing* **16**, 3033–3058.

Porway, J., Wang, K., Yao, B. and Zhu, S. (2008), A hierarchical and contextual model for aerial image understanding, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition'.

Rathi, V. and Majumdar, A. (2002), Content based image search over the world wide web, *in* 'Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing'.

Ren, W., Singh, M. and Singh, C. (2002), Image retrieval using spatial context, *in* 'Ninth International Workshop on Systems, Signals and Image Processing', Manchester.

Richards, J. and Jia, X. (2005), *Remote sensing digital image analysis: An Introduction*, Springer, Berlin.
URL: *http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/3540251286*

Schiewe, J. (2002), Segmentation of high-resolution remotely sensed data: concepts, applications and problems, *in* 'Symposium on Geospatial Theory, Processing and Applications', Ottawa.

Shackelford, A. and Davis, C. (2003), 'A hierarchical fuzzy classification approach for high-resolution multispectral data over urban areas', *IEEE Transactions on Geoscience and Remote Sensing* **40**, 1920–1932.

Sheeren, D., Quirin, A., Puissant, A., Gancarski, P. and Weber, C. (2006), Discovering rules with genetic algorithms to classify urban remotely sensed data, *in* 'In Proc. IEEE International Geoscience and Remote Sensing Symposium', Vol. 97, pp. 127–136.

Smith, J. and Chang, S.-F. (1997), Visualseek: A fully automated content-based image query system, *in* 'Proceedings of the ACM International Conference on Multimedia'.

Su, W., Li, J., Chen, Y., Liu, Z., Zhang, J., Low, T., Suppiah, I. and Hashim, S. (2008), 'Textural and local spatial statistics for the object-oriented classification of urban areas using high resolution imagery', *International Journal of Remote Sensing* **29**, 3105–3117.

Taubenbock, H. and Esch, T. adn Roth, A. (2005), An urban classification approach based on an object-oriented analysis of high resolution satellite imagery for a spatial structuring within urban areas, *in* '1st EARSeL Workshop of the SIG Urban Remote Sensing', Berlin.

Tso, B. and Mather, P. (2001), *Classification methods for remotely sensed data. First Edition*, Taylor & Francis Group, Boca Raton, FL.

Tso, B. and Mather, P. (2009), *Classification methods for remotely sensed data. Second Edition*, Taylor & Francis Group, Boca Raton, FL.

Tso, B. and Olsen, R. (2005), 'A contextual classification scheme based on mrf model with improved parameter estimation and multiscale fuzzy line process', *Remote Sensing of Environment* **97**, 127–136.

Tzotsos, A. (2008), 'A support vector machine approach for object based image analysis'. Laboratory of Remote Sensing, School of Rural and Surveying Engineering.

UCL (2011), 'Ucl department of geography. supervised classification'. Online; accessed 07-February-2011.
URL: *http://www.geog.ucl.ac.uk/*

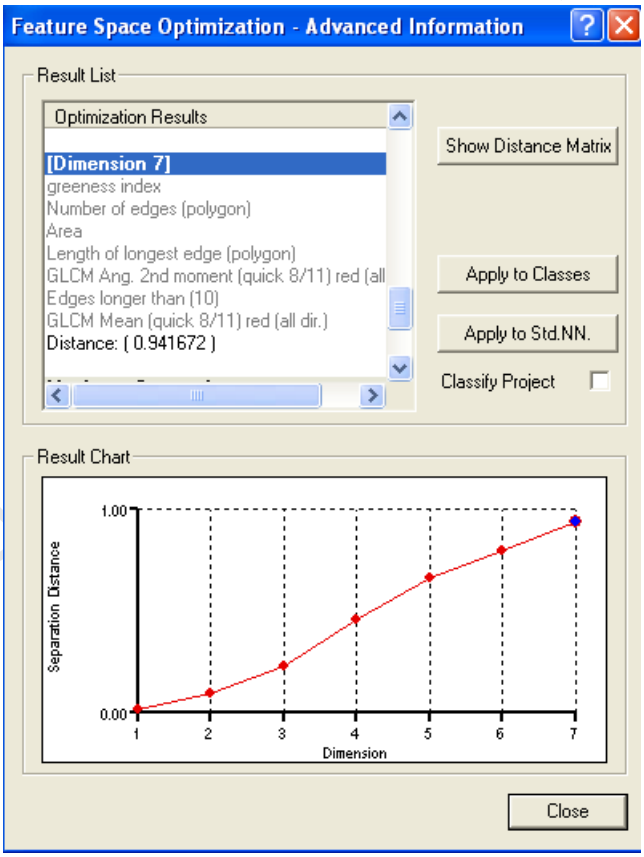
Vapnik, V. (1979), 'Estimation of dependences based on empirical data [in russian]'. English translation: Springer-Verlag, New York.

Vapnik, V. (1995), *The nature of statistical learning theory*, Springer-Verlag New York, Inc., New York, NY, USA.
URL: *http://www.amazon.com/gp/product/0387987800*

Appendices

Appendix A. Feature selection results for bottom-up object classification

The following shows the results of the optimum subset of features used to classify objects during the bottom-up phase of the automated land-use classification routine. The graph shows 'Separation Distance' vs 'Dimension'. Separation Distance refers to the best distance obtained between class clusters for a particular subset of features. The larger the Separation Distance, the better the classifier is likely to perform. Dimension refers to the number of features.



Appendix B. eCognition ruleset used to perform automated land-use classification