UNMANNED AERIAL VEHICLE MAPPING FOR SETTLEMENT UPGRADING

Caroline Margaux Gevaert

Graduation committee:

Chairman/Secretary

Prof.dr.ir. A. Veldkamp	

Supervisor(s)

Prof.dr.ir. M.G. Vosselman	University of Twente / ITC
Prof.dr. R.V. Sliuzas	University of Twente / ITC

Co-supervisor(s)

Dr. C. Persello	University of Twente / ITC

Members

Prof.dr. P.A.E. Brey	University of Twente / BMS
Prof.dr. K. Pfeffer	University of Twente / ITC
Prof.dr.-ing. M. Gerke	TU Braunsweig, Germany
Prof.dr. K. Schindler	ETH Zurich, Switzerland

ITC dissertation number 335
ITC, P.O. Box 217, 7500 AE Enschede, The Netherlands

ISBN 978-90-365-4635-5
DOI 10.3990/1.9789036546355

Cover designed by Job Duim
Printed by ITC Printing Department

UNIVERSITY OF TWENTE.

FACULTY OF GEO-INFORMATION SCIENCE AND EARTH OBSERVATION

UNMANNED AERIAL VEHICLE MAPPING FOR SETTLEMENT UPGRADING

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof.dr. T.T.M. Palstra,
on account of the decision of the graduation committee,
to be publicly defended
on Friday October 19, 2018 at 14.45 hrs

by

Caroline Margaux Gevaert

born on July 9, 1989

in Philadelphia, USA

This thesis has been approved by
Prof.dr.ir. M.G. Vosselman, supervisor
Prof.dr. R.V. Sliuzas, supervisor
Dr. C. Persello, co-supervisor

Acknowledgements

One thing I have noticed during my Ph.D. is that it is difficult to realize where an idea begins. Research, ideas and solutions seem to emerge through the many small communications and exchanges which interlace our daily life. The work behind this manuscript is the same. It is not the result of my ideas, but rather a showcase merging the ideas of the people surrounding me. Looking back, I see the ideas and influences of those who supported and guided me the past few years – and I am proud and extremely grateful for it. I'd like to make use of this opportunity to thank them.

Most prominently, you can see the ideas of my promotor and daily supervisors. George is the engineer with a pragmatic view, tough but fair and supportive. His knowledge of laser scanning and photogrammetry strongly influenced the involvement of the 3D aspect in my work. Claudio dives into the theory, examining the workings of the algorithms. He was extremely collaborative and supportive. Richard helped me focus on the context of my research – before my qualifier he asked me "but what can it do for the people"? I doubt I have answered that question, but perhaps we are now one step closer.

The ideas of many of the other department colleagues come back to various parts of this manuscript. Sander, one of the funniest people I know with the twinkle in his eye the first warning of an upcoming treat (whether a joke or a sweet). Francesco and Markus regarding the UAV aspects, Monika about informal settlements, and Yola's conversations regarding the ethical aspects. The others at ITC who have supported me: Watse in figuring out the Aibotix, Job and Benno for making the posters, Loes for helping with the Ph.D. formatting, the ITC travel unit, Theresa for all the support, kind Roelof, and the committee members who have dedicated their time and effort to review this manuscript.

To my office-mates Bashar, Mengmeng, Ye Lu, and Andrea – who never fails to remind me of what is truly important in life. To Anand and Kanmani, Biao, Shima, a great lady with similar passions for fencing and baking, Fashuai "wing-man" Li, Zhenchao, Sarah, Vera, Azar, and all my other Ph.D. colleagues.

All of this would never have been possible without the support of my friends in Rwanda when I decided to show up with a drone in a suitcase. City Engineer Dr. Alphonse Nkurunziza, Abias Philippe Mumuhire, Fatou Dieye, and Fred Mugisha. The World Bank and Ramani Huria / HOT teams for their support during the field work in Dar es Salaam. Also to UAV Agrimensura, for sharing their data of Uruguay. Fieldwork never failed to inspire me and brought me back home to Enschede full of energy.

Back home with so many friends who have come so close to my heart during the past years. Julia and Yiannis, for finding the tulips. Divyani, with her great smile and bottomless energy to uplift all of us. Mila and her daughters, for the painting and karaoke birthdays. Claudia and Martin, for the surprise gifts, family dinners, and introducing me to bouldering. The Fontainebleau (wine tasting?) group: Shayan, Ieva, Xavi, Laura, Juanri, and other friends from the Cube.

And looking towards the future, I'd like to thank some people who have given me great opportunities, even though they had little reason to do so. Jeroen for bringing me to Tanzania, and to Mark and Edward and the TURP team for bringing me back.

But no matter where I travel or end up, one thing which remains with me always is my family. To my mother, the strongest woman I know. To my sisters, Anouk for daring to do it differently and Charlotte for her fire. Friso a.k.a. "the man" – man, I really owe you one for putting up with me. To my father, for his perseverance.

To all of you who have supported me over the years, thank you.

Table of Contents

List of figures

List of tables

Chapter 1 - Introduction

1.1 Slum upgrading

Urbanization in developing countries is often paired with slum expansion, which is considered one of the main development challenges of our time. Target 11.1 of the Sustainable Development Goals (SDGs) is directed at ensuring "access for all to adequate, safe and affordable housing and basic services and upgrade slums" (United Nations, 2015). An estimated one-quarter of the world's urban population, 61.7% of the urban population in Africa, still live in slums (UN-Habitat, 2015). The true count may even be higher as official population estimations often depend on household surveys which do not take slums into account (Carr-Hill, 2013).

To establish an operational definition of slums, UN-Habitat defined a slum as having at least one of five characteristics: inadequate access to safe drinking water, inadequate access to sanitation, low quality of housing, overcrowding, and lack of tenure (UN-Habitat and Earthscan, 2003). The latter implies that all informal settlements – those lacking official tenure – are slums by definition, but not all slums are informal. The existence of this operational definition is valuable as it allows slums to be compared on a global level (Arimah, 2010). However, some critique this definition by stating that it is on household level without accounting for neighborhood characteristics (Jankowska, Weeks and Engstrom, 2012). Others indicate that it does not adequately capture the diversity of slums (Arimah, 2010), as even within a single city, slums may have differing characteristics (Sliuzas, Mboup and de Sherbinin, 2008; Jankowska, Weeks and Engstrom, 2012).

By whatever name it is called, improving the deprived conditions in these areas is at the top of various development agendas (AUC, 2015; United Nations, 2016; UN-Habitat III, 2017). Slum *eradication* is now considered to be ineffective as it treats the symptom rather than the underlying problems behind slum formation (Arimah, 2010). Instead, *in situ* slum upgrading projects which greatly reduce but do not eliminate the need to relocate inhabitants (UN-Habitat, 2012) are currently considered to be more appropriate (Abbott, 2002). These projects often focus on physical aspects such as: improving access to potable water and sanitation, provision of utilities such as electricity, and improving infrastructure such as roads and drainage (Turley *et al.*, 2013). Some strategies focus on improving streets to encourage the commercial development within the area, promote safety, and increase the identification of people with their neighborhood which would translate to increased household investments (UN-Habitat, 2012). Other studies argue that slum upgrading projects should focus on improving (access to) employment opportunities (Cohen, 2013; Pugalis, Giddings and Anyigor, 2014) rather than such physical interventions.

The 'best practice' for slum upgrading projects remains subject to debate. Part of the reason behind such diverse strategies is the lack of systematic evidence regarding their impact. One study analyzed more than 1000 publications and reports to find conclusive evidence regarding the socio-economic impacts of physical slum upgrading projects (Turley *et al.*, 2013). Improved water supply and sanitation improve public health in urban settings, but the results remain inconclusive. Due to the lack of concrete scientific evidence regarding the most effective interventions, and more importantly the great variety in slum characteristics and population needs, 'best practices' may focus on the methods rather than the specific goals. For example, a participatory approach to the upgrading process is strongly advocated (UN-Habitat and Earthscan, 2003) as including local stakeholders and slum residents helps to identify the actual needs of the local population and promotes a more sustainable change (Wekesa, Steyn and Otieno, 2011; Pugalis, Giddings and Anyigor, 2014).

1.2 Spatial information

An accurate overview of the current situation of the slum (existing housing and infrastructure, services, environmental conditions, hazards, etc.) is needed to identify key problems and plan the upgrading process. Therefore, spatial data is considered essential for informal settlement upgrading projects (Abbott, 2002; Kohli *et al.*, 2013; Taubenböck and Kraff, 2014). Informal settlements are often both literally and symbolically "empty spots on the map" (Paar and Rekittke, 2011; Pugalis, Giddings and Anyigor, 2014). Obtaining an accurate base map of these areas provides a sound basis for designing technical interventions (Paar and Rekittke, 2011; UN-Habitat, 2012), as well as improving the communication between stakeholders (Barry and Rüther, 2005), and empowering local authorities and communities (Abbott, 2003).

So how do we fill in these gaps on the map? Spatial data can be collected on the ground through field mapping exercises. A great benefit of this is the opportunity to involve the local residents in the mapping exercises. Another option is through the use of remotely sensed imagery, such as satellite or aerial imagery. This can speed up the mapping, collect information in areas with limited accessibility, allow experts off-location to be involved, and show evidence of the settlement at a certain timestamp. However, physical settlement conditions captured by remotely sensed imagery are not always representative of its current living conditions or other socio-economic aspects of the community (Taubenböck and Kraff, 2014). With this limitation in mind, satellite imagery supports informal settlement management through: identifying informal settlements, identifying changes in the boundaries of these settlements over time, generating surface data, classifying land use, identifying buildings and other objects for mapping purposes, and reconnaissance (Mason and Fraser, 1998). Remote sensing may play an

important role in providing information between censuses (Montgomery, 2008), and identify trends not visible through other data collection methods. For example, by identifying increases in backyard shacks which are not identified through official household surveys (Kakembo and van Niekerk, 2014). An overview of settlement characteristics which can be derived directly or indirectly from remotely sensed imagery is provided in Figure 1.1.

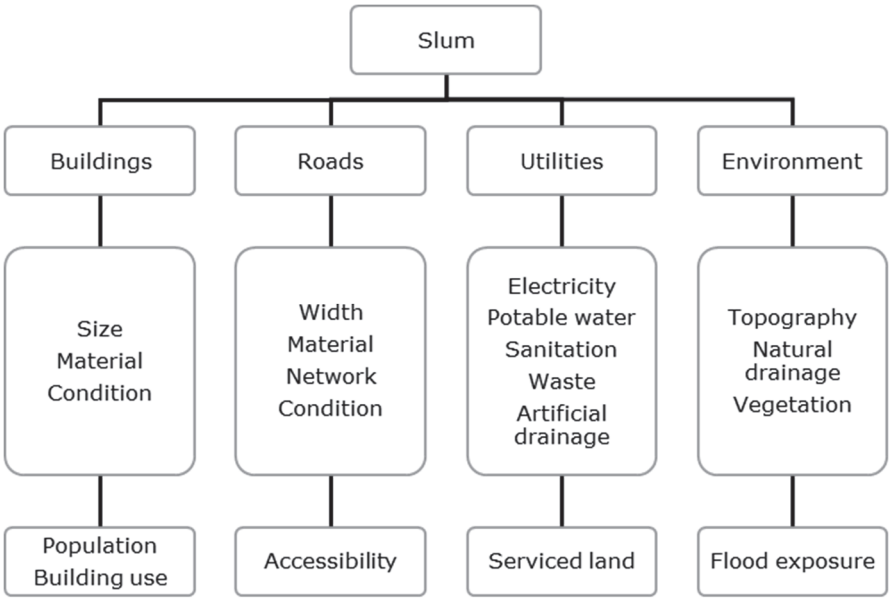


Figure 1.1: *Hierarchy of slum characteristics which can be identified through aerial imagery with sufficient spatial resolution. The general object types (2nd row) which make up a neighborhood, information which can potentially be derived from remotely sensed data (3rd row), and information which can be inferred with the support of auxiliary data (bottom row).*

There are a number of general characteristics of slums which make it especially difficult to extract geospatial information from remotely sensed imagery. Many studies characterize slums as having organic street patterns, high building densities, small building sizes, and a lack of open spaces (Baud *et al.*, 2010; Kohli *et al.*, 2012; Kit and Lüdeke, 2013; Kuffer, Pfeffer and Sliuzas, 2016). Continuous or even overlapping rooflines and heterogeneous roof materials also complicate the interpretation of satellite imagery (Owen and Wong, 2013). The advent of Very High Resolution (VHR) satellite imagery, has been an important development for this application. However, even having a spatial resolution of 50 cm is sometimes not enough for informal settlements (Kuffer, Pfeffer and Sliuzas, 2016). Aerial imagery is one option to obtain data with a higher spatial resolution, but costly. Especially for relatively small study areas, mobilizing an aircraft is impractical.

Elevation models are also important for upgrading projects. Overlapping aerial images of a slum can be used to obtain a Digital Surface Model (DSM). This provides the elevation as seen from overhead, i.e., the terrain height plus the height of the objects on top of it. Filtering out these elevated objects creates a Digital Terrain Model (DTM) which may be used for designing infrastructure and for identifying hazardous or flood-prone areas.

In summary, both imagery and derived elevation models can be very useful for informal settlement upgrading projects. However, input data with a higher spatial resolution and more advanced information extraction algorithms are required to provide useful spatial information in the challenging settings that typify slums.

1.3 Unmanned Aerial Vehicles (UAVs)

UAVs, also known as drones, Unmanned Aerial Systems (UAS) or Remotely Piloted Aircraft Systems (RPAS), are defined as small aircraft operated without an onboard pilot (Nex and Remondino, 2014). The widespread availability of cheap, off-the-shelf UAV systems coupled with developments in automatic image processing from the field of computer vision has led to a surge in UAV applications over the recent years (Colomina and Molina, 2014; Nex and Remondino, 2014).

For mapping applications, a UAV works in the same way as traditional aerial imagery. It flies a grid over an area, taking images at regular intervals. Photogrammetric software recognizes common points in each image, allowing the calculation of the interior and exterior camera parameters to calculate the relative position of each image and construct an initial 3D model of the area. The inclusion of Ground Control Points (GCPs) measured in the field allows for the positioning of this model in the real world. Dense matching can then be applied to obtain a detailed point cloud – i.e., a 3D model consisting of a much large number of points with X, Y, and Z coordinates as well as color information. A Digital Surface Model (DSM) can then be derived and an orthomosaic produced by stitching together parts of the original UAV images.

Like traditional aerial imagery, the orthomosaics obtained from UAVs may reach a spatial resolution on the scale of a few centimeters (Nebiker *et al.*, 2008). This depends on the UAV flight parameters such as flight height, camera type, and image acquisition angle. Images may be taken at oblique angles may also provide detailed façade information in urban settings (Xiao, 2013). Another benefit is the ability of UAVs to fly under clouds (although rain is still a problem), which is a recurring problem for optical satellite imagery. The DSMs obtained from UAVs may reach an accuracy level to rival that of field measurements with a DGPS (Haarbrink and Eisenbeiss, 2008; Harwin and

Lucieer, 2012), although this accuracy depends highly on the flight parameters, image quality, and GCPs.

1.4 Machine Learning

Information can be extracted from data through machine learning. For example, supervised classification methods can be used to recognize patterns in data from some labeled training samples, enabling a class label to be assigned to new data. The first step in supervised classification is usually to define relevant *features* which describe the data. For images, such features can be color and texture (Nichol, 2009). For point clouds, 3D features which describe the shape of neighboring points (Chehata, Guo and Mallet, 2009; Weinmann *et al.*, 2015), or height differences over larger areas can be used (Serna and Marcotegui, 2014). A set of *training samples* consisting of a class label and the corresponding features values is determined. It is important that the samples capture all the variations in one semantic class over the entire dataset. These training labels are then used to train the classification model which can later be applied to assign a class label to new data. Training samples can be costly to obtain as they often imply manual labeling.

A large variety of supervised classification models exist. The most suitable model for different classification tasks depends on elements such as required accuracy, number of features, availability of labeled training samples, and hardware capacity. *Random Forests* are made up of a large number individual classification trees which are each trained individually using random feature and training sample subsets (Breiman, 2001). These methods are therefore particularly robust to errors in the training labels (Frenay and Verleysen, 2014; Maas, Rottensteiner and Heipke, 2016), but require a large number of training samples.

Support Vector Machines (SVMs) are robust classifiers that are particularly suited to high dimensional feature spaces, have been proven to obtain high classification accuracies in remote sensing applications (Bruzzone and Persello, 2010), and can perform well with a limited number of training samples. SVMs map the training samples into a nonlinear feature space and construct class boundaries which maximize the margins between labels from different classes while minimizing the number of training errors. Kernels, such as the non-linear Radial Basis Function (RBF) kernel, are used to describe the distance between samples. Although often the same kernel is used for all features, Multiple Kernel Learning (MKL) defines many kernels with different parameters. Features are first divided into groups, and each group is assigned a kernel with different parameters, these kernels can then be combined into a single kernel to perform SVM. Using multiple kernels has been shown to outperform single-kernel strategies on some tasks (Gönen and Alpaydın, 2011) such as when

using features from LiDAR and multispectral satellite imagery for urban scene classification (Gu *et al.*, 2015).

More recently, *deep learning* has gained popularity due to unprecedentedly high classification accuracies on very difficult benchmark datasets in the computer vision community (Krizhevsky, Sutskever and Hinton, 2012; Simonyan and Zisserman, 2014; He *et al.*, 2016). For example, Convolutional Neural Networks (CNNs) use convolutional layers which apply a number of filters to an input patch to recognize patterns, nonlinear activation functions to learn complex representations, and pooling layers to generalize and prevent overfitting. By stacking these layers, deep networks can be constructed which are quite successful in image labeling tasks (Krizhevsky, Sutskever and Hinton, 2012; He *et al.*, 2016) and DTM extraction (Hu and Yuan, 2016). Fully Convolutional Networks (FCNs) are more suitable for pixel-wise classification tasks common to remote sensing, as they avoid redundant calculations and are more memory efficient (Shelhamer, Long and Darrell, 2017). Despite rapid developments in this field, limitations of this method include the considerable amount of training samples needed as well as substantial computing costs and associated hardware requirements.

1.5 Research Gap

The context of this work is two-fold. On the one hand, we see a clear need for high-quality, up-to-date information on slums to support upgrading projects. Understanding the present situation of the slum, identifying key problem areas, enabling stakeholders to visualize priorities and plan interventions together, the engineering of suitable upgrading measures – all steps require accurate spatial information. In this sense, slums are particularly challenging due to the lack of available data. Updating this information through remote sensing is also challenging due to typical slum characteristics: small buildings, narrow footpaths, irregular buildings, heterogeneous roof materials, and possibly even the environment such as the location on steep slopes. Geoinformatic methods for deriving information from imagery, such as classifying buildings and vegetation or the extraction of the underlying terrain, are typically developed on benchmark data from developed countries. For example, most DTM-extraction algorithms have been tested on relatively easy datasets (Tomljenovic *et al.*, 2015) and have difficulties in sloped urban environments and densely built-up areas characteristic of slums. In sum, not only is it important to acquire relevant spatial information to support upgrading projects, but the locations themselves challenge existing geoinformatic algorithms.

On the other hand, UAVs are booming. The global market may reach an estimated value of seven billion USD in 2020 (Thibault and Aoude, 2016). Their

agility as a data platform enable a user to quickly acquire images with a very high spatial and temporal resolution. The straightforward way to extract information from UAV data products would be to process the orthomosaic as you would a satellite image. However, we argue that one of the main opportunities of UAVs is the simultaneous acquisition of imagery and the 3D information. Identifying possible synergies between the 2D image-based information and the 3D geometric information should not be overlooked and is a recurring theme throughout the research presented in this dissertation.

The main focus of this research is on the use of machine-learning methods. Machine learning methods were flagged as an appropriate methodology for identifying slums from remotely sensed imagery (Kuffer, Pfeffer and Sliuzas, 2016). In the domain of computer vision, deep learning methods have been breaking records for a wide range of applications. Here, we consider the implications and required adaptations of successful machine learning methods to emerging data acquisition platforms (UAVs) for extracting information from challenging datasets (slum areas).

Finally, the importance of reflecting on the social and ethical aspects of scientific research is often forgotten (Flipse, van der Sanden and Osseweijer, 2013). Researchers' tendency to over-simplify the underlying social processes may deter the adoption of technological innovations (Pannell *et al.*, 2011). Regarding UAVs, some concerns have been voiced regarding the ethics of their usage (Haarsma, 2017) and the potential misuse of potentially sensitive information they capture (Culver, 2014). Specific concerns depend on the UAV operations (Finn and Wright, 2016) as well as the cultural context (Ordnance Survey, 2015) of the application in question. Potential benefits of geospatial information, such as urban governance and empowerment of deprived populations (Pfeffer *et al.*, 2013), obtained through the UAVs and negative externalities should be balanced. However, empirical research regarding the perceptions of the public towards UAV flights and the obtained geospatial information, as well as concrete investigations regarding how the obtained geospatial information can be used by local stakeholders is lacking.

1.6 Research Objectives

The main objective of the proposed research is to analyze the potential of UAVs to support informal settlement mapping projects. This is done through the following sub-objectives:

1) Identifying synergies between 2D and 3D information provided by UAVs
   To develop accurate classification models, the scene must be described by adequate features which are capable of distinguishing the different classes of interest. Informal settlements are often characterized by narrow

footpaths, irregular shapes, heterogeneous construction materials, and a large amount of clutter, which makes it difficult to distinguish these classes. For example, the color of a roof may be similar to the color of the ground. The simultaneous provision of highly detailed imagery and point clouds by the UAV enables users to benefit from advancements in both 2D image- and 3D scene-understanding. In this objective, we therefore compare which features are useful for describing buildings, vegetation, terrain, structures, and clutter in different informal settlements.

2) Adapting supervised classification methods to deal with heterogeneous data
The 2D and 3D feature sets correspond to different "views" of the same settlement. Therefore, the features are likely to have different statistical characteristics and should be considered differently by the classification model. Previous studies indicate that MKL indeed obtains better results than single kernel SVM for heterogeneous data. This objective investigates whether the same is true for the classification of UAV data. MKL literature describes different methods for combining kernels with different parameters. However, little attention is given to which features should be grouped and described by the same kernel.

3) Analyzing how reliable training labels can be obtained from existing geospatial data
The accuracy of a supervised classification model depends not only on the features and classification algorithm but also on the training samples used to train the model. The labeled samples must adequately describe the common characteristics and variations of the object in question. Unfortunately, it is generally costly and time-consuming to obtain such labels. Although many informal settlements remain unmapped, sometimes vector data is available from previous mapping efforts. In this case, there will be differences between the vector outlines and the newly acquired (UAV) imagery due to (1) changes in the scene itself such as building construction or demolition, and (2) misalignments due to digitization at a lower spatial resolution or other geo-referencing issues. This objective uses existing maps to provide training labels and then analyses how to automatically flag samples which are likely to be mislabeled and remove them from the training set.

4) Analyzing how to extract Digital Terrain Models in challenging settings
Informal settlement characteristics such as steep topography and a high building density are also challenging for DTM extraction algorithms. Deep learning could be utilized to learn these complex relations but must be adapted to the application of DTM extraction. In this context, we consider three specific research questions. The first challenge is how to acquire a large number of labeled samples to train the network in a fast and cheap

manner. Secondly, existing DTM algorithms often assume that ground samples are the lowest points within a local neighborhood. The size of this neighborhood must be larger than the largest elevated object in the scene. The size of objects such as buildings in the real world (in the order of meters) compared to the resolution of UAV imagery (centimeters) means that very large neighborhoods would need to be considered, which increases the computing costs of the deep learning algorithm. Therefore, avenues must be explored to increase the area under consideration by the algorithm while avoiding unnecessary increases in the computational costs. Thirdly, using only 3D information may not be enough to distinguish ground from non-ground in cases such as buildings on sloped terrain. Therefore, we again consider how interactions between 2D and 3D information may be exploited to improve DTM extraction.

5) Identifying opportunities of UAVs to support urban upgrading workflows
   Moving away from a machine learning approach of analyzing what information can be obtained from UAV imagery, it is important to consider how the images are actually used and perceived to be useful in a local context. To this end, we analyze how UAV imagery is used to support an upgrading project in Kigali, Rwanda and what the perceived utility is of the images for various stakeholders. Practical barriers towards the wide-scale utilization of UAVs at the time are also identified.

6) Analyzing the social impacts of using UAVs in the context of urban upgrading projects
   Apart from the perceived benefits of using UAVs to support urban upgrading projects, there are also widespread concerns regarding the ethical implications of acquiring such high-resolution images over urban settlements. In some cases, individuals may be recognized in the imagery as well as visualization of private spaces such as backyards. Therefore, issues such as privacy and possible misuse of the data may be a concern. The last objective is to consult the opinions of residents in informal areas regarding which information captured by the imagery they consider sensitive or private. This can then be used for further studies regarding how the use of UAVs aligns with other social values and 'best practices' advocated for upgrading projects.

The main study area for the current research was in Kigali, Rwanda. The researchers were provided with a very unique opportunity as the University of Twente / Faculty ITC funded the UAV data to be collected in 2015 at the same time and place as an urban upgrading project was being initiated by the City of Kigali – One Stop Centre in collaboration with the Rwanda Housing Authority and the World Bank. To examine the transferability of the methods and observations to other informal settlements, some chapters include UAV datasets from Tanzania, Uruguay, and Italy.

1.7 Outline

The framework of this dissertation can be seen as a set of concentric circles (Figure 1.2). We first analyze the implications of using UAVs for supervised classification tasks in a strictly algorithmic sense, then analyze how to practically obtain derived geospatial information products such as DTMs, and finally place the use of UAVs into the societal context of urban upgrading projects.

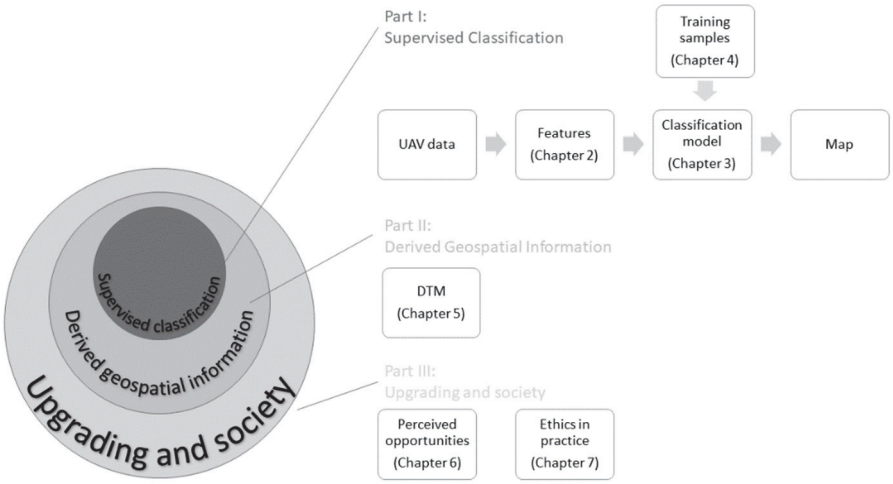


***Figure 1.2**: Organization of research topics in the dissertation.*

More specifically, the organization of the chapters is as follows:

Chapter 1 – introduces and motivates the work and describes the research objectives.

Chapter 2 – provides an overview of feature sets described in the scientific literature for urban classification using images (2D), DSMs (2.5D), and point clouds (3D). Various feature sets are combined to identify buildings, vegetation, terrain, structures, and clutter in two informal settlements (Kigali and Maldonado) using a SVM classifier. A detailed analysis of the results indicates which feature sets are especially useful for the identification of the different objects.

Chapter 3 – investigates how MKL can be optimized for the classification of UAV data. Using feature sets identified in Chapter 2, a data-driven MKL feature grouping strategy is developed which helps a user decide how to best employ MKL for their dataset. The proposed grouping strategy is compared with *a priori* and random feature grouping strategies through various MKL workflows on the

Kigali dataset. The results are also compared to standard (single-kernel) SVM and random forests.

Chapter 4 – presents an iterative technique to exploit existing base map data to provide labels for the newly acquired UAV imagery. An approach is proposed which utilizes global and local contextual cues to automatically remove unreliable samples from the training set and thereby develop an accurate classification model. The method is tested for the Kigali and Dar es Salaam datasets, and a sensitivity to the initial level of label noise is provided.

Chapter 5 – introduces the proposed methodology for DTM extraction. A review of existing DTM-extraction methods is provided as well as an overview of data and scene characteristics which are challenging for these algorithms. A new deep-learning based approach is proposed, which exploits simple rules to label training data – thus bypassing the costly process of manually labeling samples. The relatively shallow network is presented, and compared to both deeper deep learning networks and other reference DTM extraction approaches for three challenging datasets in Kigali, Dar es Salaam, and Lombardia.

Chapter 6 – considers the *observed* utility of the UAV imagery for upgrading projects. After distributing the UAV images to the upgrading project in Kigali, this chapter analyses how the images were used by various stakeholders and how they considered it to be useful. It also identifies some of the current constraints regarding the wide-spread usage of UAVs in these projects.

Chapter 7 – considers the ethics regarding the usage of UAVs as a geospatial collection tool to support urban upgrading projects. Stakeholder interviews in Kigali and Dar es Salaam describe their perceptions towards UAVs and identify which objects are considered to be private by the residents whose property is captured by the imagery. The ability of UAVs to contribute towards (or against) social values such as participation, empowerment, accountability, transparency, and equity are described.

Chapter 8 – synthesizes the results of the results of the individual chapters. Reflections on the work and future outlook are also provided.

It should be noted that chapters 2 through 7 are based on published scientific articles. There may therefore be some overlap in the introduction and motivation of the various chapters. However, this design enables each chapter to be considered individually, allowing a reader to focus on the areas which are of particular interest to him or her.

Chapter 2 – Classification Using Point-cloud and Image-based Features from UAV Data[1]

[1] This chapter is based on:

Gevaert, C.M., Persello, C., Sliuzas, R., and Vosselman, G. (2017) 'Informal Settlement Classification Using Point-cloud and Image-based Features form UAV Data', *ISPRS Journal of Photogrammetry and Remote Sensing*, 125, pp. 225-236. doi: 10.1016/j.isprsjprs.2017.01.017.

Abstract

Unmanned Aerial Vehicles (UAVs) are capable of providing very high resolution and up-to-date information to support informal settlement upgrading projects. To provide accurate basemaps, urban scene understanding through the identification and classification of buildings and terrain is imperative. However, common characteristics of informal settlements such as small, irregular buildings with heterogeneous roof material and large presence of clutter challenge state-of-the-art algorithms. Furthermore, it is of interest to analyze which fundamental attributes are suitable for describing these objects in different geographic locations. This work investigates how 2D radiometric and textural features, 2.5D topographic features, and 3D geometric features obtained from UAV imagery can be integrated to obtain a high classification accuracy in challenging classification problems for the analysis of informal settlements. UAV datasets from informal settlements in two different countries are compared to identify salient features for specific objects in heterogeneous urban environments. Findings show that the integration of 2D and 3D features leads to an overall accuracy of 91.6% and 95.2% respectively for informal settlements in Kigali, Rwanda and Maldonado, Uruguay.

2.1 Introduction

Informal settlements are a growing phenomenon in many developing countries, and the effort to promote the standard of living in these areas will be a key challenge for the urban planners of many cities in the 21st century (Barry and Rüther, 2005). These settlements refer to urban areas which lack legal tenure (Kuffer, Pfeffer and Sliuzas, 2016), and are often characterized by dense housing and sub-standard living conditions. The term is closely related to the term 'slums', referring to settlements which may lack legal tenure, lack access to water or sanitation, suffer from overcrowding and/or are characterized by non-durable housing (UN-Habitat, 2012). In the present study, we utilize the term informal settlement as it is more commonly used in the remote sensing community (Kuffer, Pfeffer and Sliuzas, 2016) and due to the possible negative connotations of the term 'slum' (Gilbert, 2007). The planning and execution of informal settlement upgrading projects with the purpose of ameliorating these conditions require up-to-date base maps which accurately describe the local situation (UN-Habitat, 2012). For example, the identification of buildings gives an indication of the population in the area, classifying terrain identifies footpaths for accessibility and utility planning or free space for the location of infrastructure. However, such basic information is often lacking at the outset of upgrading projects (Pugalis, Giddings and Anyigor, 2014), thus hindering the amelioration of the impoverished conditions in these areas. To create such base maps, satellite imagery is a powerful source of information regarding the physical characteristics of an informal settlement (Taubenböck and Kraff, 2013). However, as slums are often characterized by high building densities, small irregular buildings, and narrow footpaths, the spatial resolution provided by sub-meter satellite imagery is usually not sufficient (Kuffer, Barros and Sliuzas, 2014). Photogrammetric workflows can extract 2D orthomosaics, 2.5D Digital Surface Models (DSMs) and 3D point clouds from overlapping aerial imagery. Although this can be done from aerial or satellite imagery, UAVs have lower operational costs and allow for flexible and fast data acquisition (Nex and Remondino, 2014). This combination of flexible data acquisition and high spatial resolution of the acquired products motivate the use of UAVs to support urban planning in dense and dynamic areas such as informal settlements. Disadvantages of the use of UAVs include the limited spatial extent of UAV flights and the data processing requirements. Therefore, we consider them to be more adequate at a (settlement upgrading) project level where more detailed spatial information is required, rather than e.g. at a city level for the distinction between informal vs. formal settlements. The remaining question is then how to optimally integrate the information contained in the orthomosaic, DSM and point cloud in order to accurately classify these complex areas.

A well-known problem of classifying urban areas is the high within-class variability and low between-class variability of spectral signatures of the relevant classes. Also, when using very high-resolution (VHR) imagery, the objects to be classified are generally larger than the pixel size, which is problematic for purely pixel-based classification strategies (Blaschke, 2010). The classification of sub-decimeter orthomosaics in informal settlements can be expected to face similar problems. In the remote sensing community, a common strategy to address this issue is to include spatial-contextual features in the classification problem in addition to the spectral image attributes. Spatial-contextual information can also be incorporated through Object Based Image Analysis (OBIA), which is also currently the most common strategy for the classification of slum areas (Kuffer, Pfeffer and Sliuzas, 2016). Such approaches depend on adequate segmentation parameters, which may be difficult to transfer between study areas (Hofmann *et al.*, 2008) or even to represent different classes within the same study area (Myint *et al.*, 2011). Alternatively, a multilevel strategy to incorporate contextual features can be adopted by combining the radiometric characteristics at a pixel level with attributes of larger image segments and thus avoiding the need to define one set of optimal segmentation parameters (Bruzzone and Carlin, 2006). Their approach focusses on the spectral and spatial features at the different contextual levels, but could be extended to include texture features as these have proven to be an important supplement to spectral features in urban scene classification (Puissant, Hirsch and Weber, 2005; Tong, Xie and Weng, 2014).

Furthermore, the availability of 3D data are an important supplement to the orthomosaic as the inclusion of height information has been shown to greatly increase classification accuracy of urban scenes (Priestnall, Jaafar and Duncan, 2000; Hartfield, Landau and Leeuwen, 2011; Longbotham *et al.*, 2012). Especially the extraction of a normalized DSM (nDSM), which gives the elevation of pixels above the terrain, is useful for identifying elevated objects in urban scenes (Weidner and Förstner, 1995) and distinguishing between low vegetation and high vegetation (Huang *et al.*, 2008). A recent overview of building detection methods based on aerial imagery and LiDAR data indicates that state-of-the-art techniques which have access to both imagery and height information can identify large buildings with a very high correctness and completeness (Rottensteiner *et al.*, 2014). However, these building detection algorithms face difficulties when the buildings are relatively small (i.e. less than 50 m²), or when the height of the terrain is not uniform on all sides of the building due to sloped terrain. Unfortunately, informal settlements are often characterized by these challenging conditions, which emphasizes the need to investigate the synergies between 2D and 3D features to fully exploit the available UAV data and obtain a high classification accuracy.

Existing strategies regarding the combination of 2D and 3D features are often based on the integration of LiDAR with multispectral aerial imagery. (Yan, Shaker and El-Ashmawy, 2015) cite a number of studies where nDSM data derived from LiDAR was combined with vegetation indices from multispectral imagery to classify urban scenes (e.g. Hartfield et al., 2011). Other methods make use of elevation images which directly project the 3D points onto a horizontal plane without taking into account interpolation techniques which are typically applied for DSM extraction. Processing this summarized information in 2D space rather than the original 3D space can decrease computing costs (Serna and Marcotegui, 2014). In another example, (Weinmann *et al.*, 2015) describe a generic framework for 3D point cloud analysis which includes spatial binning features or accumulation maps, which are similar to elevation images. They define a horizontal 2D grid and calculate: the number of points within each bin, maximum height difference and standard deviation of height difference within each cell. (Serna and Marcotegui, 2014) use elevation maps to define the: minimum elevation, maximum elevation, elevation difference, and number of points per bin as a basis for detecting, segmenting and classifying urban objects. However, this method assumes the ground is planar. (Guo *et al.*, 2011) combined geometrical LiDAR features and multispectral features from imagery to analyze which features were most relevant to classify an urban scene into: building, vegetation, artificial ground, and natural ground. They use elevation images to include the inclination angle and residuals of a local plane, but found that the maximum height difference between a LiDAR point and all other points within a specified radius was the most relevant feature.

There are two main limitations of the previous methods. Firstly, most methods explicitly or inherently assume the terrain to be planar. Attributes such as the maximum absolute elevation or height above the minimum point within a horizontal radius, which are often considered to be the most relevant features (Guo *et al.*, 2011; Yan, Shaker and El-Ashmawy, 2015), will not serve to distinguish between buildings and terrain in a settlement located on a steep slope. Secondly, the methods generally focus on pixel-based features, or local neighborhood features. However, other research indicates that segment-based point cloud features provide important supplementary information to pixel-based attributes (Vosselman, 2013; Xu, Vosselman and Oude Elberink, 2014). Similarly, 2D object-based attributes significantly improve the classification of urban scenes from VHR satellite imagery (Myint *et al.*, 2011). Studies investigating the importance of features for urban scene classification should therefore consider segment-based features as well as point-based features.

The objective of this paper is to integrate the different information sources (i.e. UAV point cloud, DSM, and orthomosaic) and to analyze which 2D, 2.5D, and 3D feature sets are most useful for classifying informal settlements, a setting

which challenges the boundaries of existing building detection algorithms. In an effort address the challenge of identifying salient features in various conditions, UAV datasets over informal settlements in two different countries are compared. Feature sets describing 2D radiometrical and textural features from the orthomosaic, 2.5D topographical features from the DSM, and 3D features from the point cloud are selected from literature. Both pixel- or point-based features and segment-based features are included. The suitability of the feature sets for classifying informal settlements are tested through their application to two classification problems. The classification is performed using Support Vector Machines (SVMs), which have been shown to be very effective in solving nonlinear classification problems using multiple heterogeneous features. The first classification problem identifies major objects in the scene (i.e. buildings, vegetation, terrain, structures and clutter), whereas the second attempts to describe semantic attributes of these objects such as roof material, types of terrain, and specific structures such as lamp posts and walls. The results presented here are an extension of previous research regarding the suitability of various features sets for the classification of an informal settlement in Kigali, Rwanda (C. M. Gevaert *et al.*, 2016) in two significant ways. Firstly, the suitability of the feature sets in a different setting is analyzed through the application of the same framework to an informal settlement in Maldonado, Uruguay. Secondly, we provide an extensive analysis of the most suitable features per class, which supports other researchers in identifying which features could be most relevant for their specific classification problem.

2.2 Methodology

2.2.1 Data sets

Two UAV datasets of informal settlements were utilized in the current study. For each dataset, ten disjoint 1000 x 1000 pixel tiles were manually labelled into ten classes: three different types of roof material, high vegetation, low vegetation, bare surface, impervious surface, lamp posts, free-standing walls, and clutter (Table 2.1). The roof materials included a class for low-quality corrugated iron roofing, and two classes of high-quality material which depended on the dataset. The clutter class consists of temporary objects, such as cars, motorbikes, clothes lines with drying laundry, and other miscellaneous objects. These ten class labels were aggregated into a 5-class problem to identify the major objects in the informal settlement (buildings, vegetation, terrain, structures, and clutter) as indicated in Table 2.1. For pixels where the orthomosaic clearly indicated terrain, but the type of terrain was unknown (e.g. due to shadows), the pixels were labelled as terrain in the reference data of the 5-class but not included in the 10-class problem. The training data for the supervised classifier consisted of 200 samples for each of the ten classes,

randomly extracted from the labelled pixels. In the following sections, both study areas will be briefly described.

Table 2.1: *Classes defined in the 5-class and 10-class set-up.*

5-class	10-class
Building	Corrugated Iron roofs (low quality)
	High quality roof material I
	High quality roof material II
Vegetation	High vegetation
	Low vegetation
Terrain	Bare surface
	Impervious surface
Structures	Lamp posts
	Free-standing Walls
Clutter	Clutter

2.2.1.1 Kigali, Rwanda: A DJI Phantom 2 Vision+ UAV was utilized to obtain imagery over an unplanned settlement of 86 ha in Kigali, Rwanda in May, 2015. The characteristics of the settlement include small buildings (41% are smaller than 50 m²), often separated by narrow footpaths. Typical roof materials are corrugated iron sheets, and tile- or trapezoidal-shaped galvanized iron sheets that are often cluttered with objects such as stones. The area itself is located on a steep slope, and trees partially cover the roofs in many areas. The UAV was mounted with a 14 Megapixel RGB camera with a fish-eye lens (FOV = 110°). Each individual image has a resolution of 4608 x 3456 pixels, and they were acquired with approximately 90% forward- and 70% side-overlap. The images were processed with Pix4D software to obtain a point cloud with a density of up to 1014 points/m^2. A DSM and an 8-bit RGB orthomosaic with a spatial resolution of 3 cm were also obtained. The DSM was interpolated from the point cloud using the Inverse Distance Weighting (IDW) option in Pix4D. The main benefit of this technique is the preservation of smooth building outlines, avoiding a common speckled effect in areas where, for example, interpolation between points on overhanging roofs and the ground below may cause artefacts in the DSM and resulting orthomosaic. However, a disadvantage of this interpolation method is a slight rounding of roof corners in some areas.

2.2.1.2 Maldonado, Uruguay: This dataset was obtained by UAV Agrimensura, who utilized a microdrone md4-1000 quadcopter sporting a 24 Megapixel SONY Nex7 camera. Flights were planned at a height of 80 m with 80% forward- and 60% side-overlap. The flights were originally designed to cover 11.6 ha of the San Antonio settlement in Maldonado, Uruguay, to support urban projects by the Government of Maldonado. The settlement itself is also characterized by dense housing and low-quality corrugated iron roofs. The

terrain is flatter than the Kigali dataset, and there is more vegetation (both large trees overhanging the buildings and low vegetation in open areas). As with the Rwanda dataset, the images were processed with Pix4D to obtain the point cloud (123 points/m^2) and a 3 cm DSM (again using IDW) and orthomosaic. Details regarding the image acquisition can be found in (Birriel and González, 2015).

2.2.2 2D and 2.5D feature extraction from the orthomosaic and DSM

2D radiometric, textural, and segment-based features were extracted from the orthomosaic, and 2.5D topographical features were extracted from the DSM (see the overview in Table 2.2). Note that we refer to the DSM as being 2.5D as it is in the form of a raster and only assigns one value for the height information to each cell. This also underlines the distinction between the features extracted from the DSM and the 3D features extracted from the point cloud. The radiometric features consisted of the input R, G, and B color channels as well as the normalized values *r*, *g*, and *b* (calculated by dividing the color channel by the sum of all three channels per pixel). Vegetation indices (VIs) are typically used to identify vegetation, but are often dependent on the reflectance in the NIR spectrum which is not available in the UAV datasets employed in this study. (Torres-Sánchez *et al.*, 2014) compared a number of vegetation indices obtained from RGB UAV imagery and found that the excess green (ExG(2)) vegetation index (Woebbecke *et al.*, 1995) compared favorably to other indices for vegetation fraction mapping from UAV imagery. ExG(2) can be calculated using the normalized *r, g*, and *b* values as follows:

$$ExG(2) = 2g - r - b \tag{2-1}$$

In the absence of a Digital Terrain Model (DTM) which would allow calculating the height of objects above the ground, morphological filters can be applied to the DSM to identify how high a pixel is compared to its neighbors. More specifically, applying a top-hat mathematical morphological filter to a DSM will give the height of a pixel above the lowest point within the area delimited by a certain structuring element. Although there are more advanced methods for extracting a DTM, the combination of steep slopes and densely built structures in the Kigali dataset hindered an accurate extraction of the terrain. Furthermore, by using structuring elements of various sizes, we are able to identify the radius of the extracted object as well as the height above the surrounding area. This is because the size of the structuring element must be large enough to cover the entire object in question, but small enough to maintain the variation present in surface topography (Kilian, Haala and Englich, 1996). This size can be set in an automatic way based on granulometry to target a specific type of object such as buildings (Li *et al.*, 2014). However, as the present classification problem targets objects of varying sizes, multiple

circular top-hat filters are applied using structuring elements of varying radii: from 0.25 to 1.0 m at 0.25 m intervals, and from 1 to 10 m at 1 m intervals. Previous research has shown such an approach of using multi-scale DSM top-hat features to be successful in classifying urban scenes (Arefi and Hahn, 2005).

Table 2.2: *List of extracted features used in the classification problem. Dim. = dimension of input data, where 2D indicates the ortho-image, 2.5D indicates the DSM, and 3D indicates the point cloud. See the text for a details.*

Dim	Code	Features	Description
2D	R	Radiometric	Input RGB values, normalized color channels and vegetation index
	T	Textural	$LBP_{u,i}^{ri}$ and $VAR_{u,i}^{ri}$ features summarized over a local window
	2S	2D segment	Radiometric features averaged over mean-shift segments
2.5D	D	Topographic	Top-hat filters over DSM with various disk-shaped structuring elements
3D	3B	Spatial binning	Spatial binning to summarize 3D points in image grid
	3S	Planar segments	Planar segment features from point cloud
	3P	Point-based	3D Point-based features

Textural features from the orthomosaic are described by Local Binary Patterns (LBP) (Ojala, Pietikainen and Maenpaa, 2002). LBP texture features have compared favorably to, for example, texture features based on Gray-Level Co-Occurrence Matrices (GLCM) (Doshi and Schaefer, 2012). Furthermore, LBP features are rotationally invariant, which is important in aerial image applications as, for example, roof textures will not always be oriented in the same cardinal direction. The algorithm works as follows. It first analyses the N neighboring pixels at a radius R from the center pixel. Each neighbor is assigned the value of 1 if its grayscale value is higher than that of the center pixel and 0 if it is lower. This results in a binary code of N bits. Rotational invariance is achieved by applying a circular shift, or bitwise rotation, to the code to obtain the minimal binary value. For example, let's say we are considering the 8 directly neighboring pixels (R=1, N=8), so the texture will be described as a binary code where the first bit represents the neighbor directly above the pixel in question, and the remaining 7 bits represent the other neighbors in a clockwise direction. If for pixel A the neighbors in the first, third, and fourth positions have higher values than A and the others have lower values, it results in the binary code: 10110000. If pixel B has higher neighbors in the third, fifth and sixth slots, we obtain 00101100. By applying a bitwise rotation, we are essentially removing the zeros at the right. Thus 10110000

and 00101100 will both result in 00001011, indicating that pixels *A* and *B* both have the same rotationally-invariant texture. To reduce the number of unique codes, uniform patterns are defined as the codes containing a maximum of two binary 0/1 transitions. This allows for the definition of *N*+2 uniform, rotationally-invariant binary patterns, each of which can be assigned a unique integer. These LBP features are denoted as $LBP_{N,R}^{riu2}$ (where *'riu2'* refers to the rotationally invariant, uniform patterns). Due to the binary nature of these patterns, they fail to capture the contrast between the center pixel and its neighbors. Therefore, it is recommended to combine various $LBP_{N,R}^{riu2}$ operators with a variance measure $VAR_{N,R}$ (2-2), which compares the grayscale values of each neighbor (g_N) to the average grayscale value in the local neighborhood (μ) (Ojala, Pietikainen and Maenpaa, 2002).

$$VAR_{N,R} = \frac{1}{N}\sum_{N=0}^{N-1}(g_N - \mu)^2 \,, where\ \mu = \frac{1}{N}\sum_{N=0}^{N-1} g_N \qquad (2\text{-}2)$$

The application of LBP features to the orthomosaic involve converting the RGB image into a grayscale image, and calculating the $LBP_{N,R}^{riu2}$ pattern for each pixel. It should be noted that it is also possible to calculate such texture features separately per spectral band. However, in the present classification problems, the texture of the grayscale image provides sufficient discriminatory power. For example, the texture of corrugated iron roofs is caused by the shadows cast by the undulated shape of the roof and is therefore irrespective of the roofs' color. Next, a sliding window is applied to the orthomosaic to compute the normalized histogram. Thus, each $LBP_{N,R}^{riu2}$ feature results in *N*+2 features representing the relative presence of this pattern in the local neighborhood. Here, we apply two sliding windows, one of 3x3 pixels and the other 10x10 pixels. For this analysis, three LBP variations: $LBP_{8,1}^{riu2}$, $LBP_{16,2}^{riu2}$, $LBP_{24,3}^{riu2}$, and the corresponding VAR features were utilized.

The orthomosaic was segmented using the mean shift algorithm (Comaniciu and Meer, 2002) implemented in the EDISON toolbox. The algorithm transforms the RGB color values into L*u*v color space, and applies a kernel function to identify modes in the data. The algorithm requires two parameters to define the kernel bandwidths: h_s which represents the spatial dimension, and h_r which represents the spectral dimension. These parameters were set to 20 pixels and 5 grey levels respectively, based on experimental analysis. The segment features included in the classification consisted of the pixel-based radiometric features (i.e. R, G, B, *r*, *g*, *b* and ExG(2)) averaged over each segment.

2.2.3 3D feature extraction from the point cloud

Three types of 3D features were projected from the point cloud into 2D space: spatial binning features, planar segment features, and 3D-neighborhood based

features. Spatial binning features, similar to elevation images, can be used to project each 3D point into a horizontal 2D plane which is divided by a grid into equally sized bins. In this application, the geographical grid of orthomosaic is used to define the bins, so that each bin is exactly aligned to an image pixel. In this way, the 3D features can be directly combined with the 2D image-based features in the classification step. Three characteristics of the 3D points were calculated for each pixel: (i) the number of points per bin, (ii) the maximal height difference, and (iii) the standard deviation of the heights of all the points falling into the bin. To reduce the number of empty bins, the attributes of each point were assigned to the eight directly neighboring pixels as well as the pixel directly under it.

However, such spatial techniques greatly simplify important geometrical characteristics of objects in 3D space. Therefore, we also take into account features obtained from planar segments and the local neighborhood in 3D space. To integrate these features into a 2D space, the attributes of the highest 3D point for each pixel (or bin) was utilized. This was based on the premise that if there are multiple layers in a point cloud (terrain and an overhanging roof, for example), it is the highest point in the point cloud (i.e. corresponding to the roof) which will be visible in the orthomosaic.

Planar segments in point clouds have demonstrated their usefulness in the identification of building roofs and walls in urban scenes (Vosselman, 2013). Here, planar segments were obtained by applying a surface growing algorithm. The surface growing algorithm starts with a seed point within the point cloud, and selects the *k* nearest neighbors (10 in this case). If the group is sufficiently homogenous according to a defined attribute, then it is accepted as a segment. Here, planarity was utilized and the acceptance criterion was based on the sum of the squared residuals of the ten selected points to a local plane. The maximum residual error threshold was set to 0.30 m. If a group of points meets the requirements, it is accepted as a segment, and begins to 'grow'. Neighboring points within a defined radius (here 1.0 m) and maximum distance from the plane (again 0.30 m) are included in the segment and the planar coefficients of the segment are then recalculated. This process is repeated until no additional points are added to the segment, and then the same is done for the next seed point. A greedy approach is utilized in the segment growing process, which signifies that if a point is candidate for multiple segments that it is assigned to the segment for which it has the smallest residual error. After performing the surface growing algorithm, four segment features were calculated: (i) the number of points per segment, (ii) the average residual, (iii) the inclination angle of the plane, and (iv) the maximal height difference between the segment and directly neighboring points. To define the latter, the height difference map obtained for the spatial binning features was utilized. Pixels pertaining to the same segment were identified, and the maximal height

difference values of these pixels was assigned to all pixels pertaining to the same 3D segment. The four segment features were only calculated for segments which were identified at least once as being the highest segment in a bin, but all of the segment points pertaining to that segment were included in the calculation of the segment features.

The final step was to add more descriptive 3D-shape attributes from the point cloud into the image. Weinmann *et al.* (2015) present a set of 21 generic point cloud features. Point attributes are calculated by taking a local neighborhood into account. The size of this neighborhood will influence the characteristics of the calculated features. For example, in the context of informal settlements, a rock on a roof will display spherical characteristics if a very small neighborhood is taken into account, but planar characteristics if the neighborhood includes roof points over an extended area. This problem can be addressed by defining a different neighborhood size for each specific point (Demantké *et al.*, 2012; Weinmann *et al.*, 2015). This 'optimal neighborhood' is based on the normalized eigenvalues (e_1, e_2, and e_3, where $e_1 \geq e_2 \geq e_3$) of 3D matrix giving the X,Y,Z coordinates of the 3D points in the defined neighborhood. The ratio of these eigenvalues describe the general shape of the 3D points (Chehata, Guo and Mallet, 2009). For example, if the first eigenvalue is significantly larger than the others, the points are distributed linearly. If the first two are approximately equal, then the points describe a planar surface, and if all three are similarly sized than the points are scattered. The 'optimal neighborhood' size is defined by iteratively increasing the number of selected neighbors around a point through either a k-nn search (Weinmann *et al.*, 2015) or increasing radius size (Demantké *et al.*, 2012) and determining at which size the Shannon entropy (2-3) is minimized.

$$E_\lambda = -e_1 \ln(e_1) - e_2 \ln(e_2) - e_3 \ln(e_3) \tag{2-3}$$

It is suggested to restrict the optimal neighborhood search to a minimum of 10 and a maximum of 100 neighbors (Weinmann *et al.*, 2015). However, in the case of UAV point clouds such as those used in the current application, the density of the points implicates that even the 100 nearest neighbors may span less than 0.10 m^2 when point density is 1014 points per m^2. This could be insufficient to detect the geometry of larger objects, especially in the presence of clutter. Alternatively, utilizing a fixed radius to define the neighborhood may lead to computational difficulties due to the large number of points. Therefore, the 3D point features are calculated for the highest points per pixel in the dense point cloud, but the nearest neighbors are selected from a filtered point cloud which is subsampled to a 0.5 m 3D grid. Experimental results indicated that this greatly increased the computation speed while maintaining the features' discriminatory power.

After determining the optimal neighborhood for all the highest points per pixel, the 3D geometric, 3D shape, and 2D shape features were calculated based on the framework of 21 features (Weinmann *et al.*, 2015). The 3D geometric features consisted of the maximum altitude difference and standard deviation of the height values of neighboring points. The absolute maximum altitude feature was excluded, as the study area is sloped. From the 3D covariance matrix, combinations of the normalized eigenvectors are used to describe the linearity L_λ (2-4), planarity P_λ (2-5), scattering S_λ (2-6), omnivariance O_λ (2-7), anisotropy A_λ (2-8), eigenentropy E_λ (2-9), sum eigenvalues Σ_λ (2-10), and change of curvature C_λ (2-11).

$$L_\lambda = (e_1 - e_2)/e_1 \tag{2-4}$$

$$P_\lambda = (e_2 - e_3)/e_1 \tag{2-5}$$

$$S_\lambda = e_3/e_1 \tag{2-6}$$

$$O_\lambda = \sqrt[3]{e_1 e_2 e_3} \tag{2-7}$$

$$A_\lambda = (e_1 - e_3)/e_1 \tag{2-8}$$

$$E_\lambda = -\sum_{i=1}^{3} e_i - \ln(e_i) \tag{2-9}$$

$$\Sigma_\lambda = e_1 + e_2 + e_3 \tag{2-10}$$

$$C_\lambda = e_3/(e_1 + e_2 + e_3) \tag{2-11}$$

Elongated rectangular planes, such as fences along plot boundaries, normalization of the planarity P_λ by e_1 (2-5) may have low values (Vosselman, 2013). As an alternative, normalization using e_2 is proposed:

$$P2_\lambda = (e_2 - e_3)/e_2 \tag{2-12}$$

This alternative was included in the 3D feature set. Furthermore, the sum and ratio of eigenvalues in the 2D covariance matrix obtained by projecting the points in the neighborhood to a local plane were calculated. The spatial binning features described by the framework were not applied, as they are similar to those calculated previously.

2.2.4 Feature selection and classification

The feature sets are compared through supervised classification using a SVM classifier with a Radial Basis Function (RBF) kernel implemented in LibSVM (Chang and Lin, 2011). SVM classifiers maximize the margins between classes while minimizing training errors, a method which is proven to obtain high classification results and generalization capabilities even when few training samples are utilized (Bruzzone and Persello, 2010). To train the SVM

classifiers, all the features were normalized to a 0-1 interval. Then, a 5-fold cross-validation was used to optimize the values of the soft margin cost function C from 2^{-5} to 2^{15} and the spread of the Gaussian kernel γ from 2^{-12} to 2^{3} on the training set.

Five feature sets are compared (Table 2.3). The first feature set (RT) consists exclusively of 2D features, namely the radiometric and texture features calculated from the input orthomosaic. The second feature set (RD) simulates situations in which the DSM is also available, a 2.5D approach using radiometric and topographic features. The third set (R2ST) again consists of exclusively 2D features, but includes the mean shift segments, which are used to summarize the radiometric and textural features. The fourth set (R2ST3), includes the features which are directly obtained from the point cloud (spatial binning, planar segment, and point-based 3D features), again summarized over the mean shift segments.

To reduce the computational cost and prevent over-fitting the classifier, a feature selection method was applied to create the fifth feature set (FS). Feature selection methods consist of a *search strategy* and *criterion function*. In this case, the Sequential Forward Floating Search (SFFS) search strategy (Pudil, Novovičová and Kittler, 1994) was applied using the Hilbert-Schmidt Independence Criterion (HSIC) (Gretton *et al.*, 2005) for the criterion function. SFFS is based on Sequential Forward Search (SFS), a bottom-up feature selection method which starts with an empty set of selected features. The unselected features are iteratively evaluated using the criterion function. The feature which maximizes the criterion function is defined as the most significant feature and is added to the set of selected features. This process of evaluating the criterion function by iteratively evaluating the unselected features and adding the most significant feature is continued until the feature set reaches a pre-defined size. One main problem of SFS is the 'nesting' effect, which means that once a feature is selected, it can no longer be discarded from the final set of selected features. To avoid this issue, after the addition of each additional feature, SFFS backtracks to check if the removal of any features from the selected set increases the criterion function. More specifically, SFFS consists of three steps (Pudil, Novovičová and Kittler, 1994). The first step is *inclusion*, where SFS is applied to select the most significant candidate feature to add to the set. Secondly, *conditional exclusion* is applied by individually removing each feature and calculating the criterion function. If the removal of any feature except the one which was just added results in a higher criterion function, then this feature is removed from the set. The *continuation of conditional exclusion* iterates this procedure, iteratively searching for the least significant feature in the set and removing it if: (i) the feature set will still contain more than two features and (ii) the criterion function of the new subset is higher than the previously obtained subset containing the same number of features.

Regarding the criterion function, a number of methods such as Correlation-based Feature Selection (Hall, 1999) and Minimal-Redundancy-Maximal-Relevance (Hanchuan Peng, Fuhui Long and Ding, 2005) focus on maximizing the predictive power of the selected features towards the classification label while minimizing the similarity between the selected features. However, such methods rely on linear relations between features. As the current application utilizes a SVM classifier, it is important to select features based on non-linear relations. Therefore, the HSIC is utilized for the criterion function, which has successfully been used for feature selection methods in the transfer learning domain (Persello and Bruzzone, 2016). This measure evaluates the similarity between an input kernel $\boldsymbol{K}$ and a kernel $\boldsymbol{L}$ representing an ideal output kernel where samples adhering to similar class labels are assigned a value of 1 and samples adhering to different classes are assigned the value 0. It can be calculated as follows:

$$\widehat{HSIC}(X,Y) = \frac{1}{n^2} Tr(\boldsymbol{KHLH}) \tag{2-12}$$

where n is the total number of samples, and Tr indicates the trace. $\boldsymbol{H}$ is the centering matrix: $\boldsymbol{H}_{ij} = \delta_{ij} - (1/n)$, where δ_{ij} equals 1 when samples i and j adhere to the same class, and 0 otherwise. The SFFS feature selection strategy with HSIC criterion function was used to select the 60 most relevant features out of the entire 2D, 2.5D, and 3D feature set. The feature selection is applied separately for each of the three datasets and for both the 5-class and 10-class problem.

Supplemental experiments were conducted to identify which of the 3D features are most relevant for the different classes. To this end, the labelled samples of each of the ten classes were selected individually from the training dataset used in the classification. Each sample adhering to the selected class was set to a label of '1' and all other training samples were set to a value of '2'. Then, the SFFS feature selection was applied to select the three most relevant 3D features. This additional analysis allows to (i) identify which 3D features are most informative for each of the specific classes and (ii) to compare whether the same features are informative for similar objects in different informal settlements.

The classification results of the five different feature sets are compared using the Overall Accuracy (OA) of the prediction maps compared to the reference data. The reference data created manually by digitizing over the UAV orthomosaics. Furthermore, confusion matrices as well as the correctness (2-13) and completeness (2-14) are used to compare the relations between number of true positives (TP), false negatives (FN) and false positives (FP).

$$Correctness = \frac{TP}{TP+FN} \quad (2\text{-}13)$$

$$Completeness = \frac{TP}{TP+FP} \quad (2\text{-}14)$$

2.3 Results

The RT feature set has the worst performance for all experimental set-ups, and the highest performing feature set was always a combination of 2D and 3D features (RT2S3 and FS) according to the OA (Table 2.4), completeness, and correctness (Tables 2.5 and 2.6). Sample classification results consisting of one of the ten tiles for each dataset are presented in Figures 2.1 and 2.2. RT does achieve an OA of 80.8% in Maldonado, which is possibly due to the prominence of terrain in the form of low vegetation rather than reddish soils which facilitates the distinction between buildings and terrain based only exclusively spectral properties. The RT2S3 feature set achieves very high accuracies for both study areas from 91.6% to 95.2% in the 5-class problem and 86.1% to 92.2% in the 10-class problem. The high performance of the FS feature set is important to note, as it indicates that the improved performance of RT2S3 over RT2S is not solely the utilization of more features (117 instead of 60), but also due to the combination of 2D and 3D features.

Table 2.3: *Description of the feature sets used for the classification experiments. See Table 2.2 for a description of the feature set codes, FS indicates feature selection was applied. N indicates the number of features in the set.*

Feature set	**2D**			**2.5D**	**3D**	**N**
	R	T	2S	D	3	
RT	X	X				61
RD	X			X		20
R2ST	X	X	X			68
R2ST3	X	X	X		X	117
FS	X	X	X	X	X	60

Table 2.4: *Overall Accuracies (OA) achieved by the five feature sets for both study areas.*

	5-class OA (%)					**10-class OA (%)**				
	RT	RD	RT2S	RT2S3	FS	RT	RD	RT2S	RT2S3	FS
Kigali	74	87	90	**92**	91	52	79	85	**86**	84
Maldonado	81	91	95	**95**	95	57	83	91	91	**92**

Table 2.5: *Completeness and correctness of selected feature sets for the 5-class problems.*

Kigali

Completeness					
	Building	Vegetation	Terrain	Structures	Clutter
RT	0.693	0.941	0.797	0.460	0.239
RD	0.847	0.952	0.919	0.699	0.207
RT2S	0.901	0.949	0.909	0.826	0.718
RT2S3	**0.916**	**0.964**	**0.922**	0.769	0.676
FS	0.910	0.954	0.921	**0.846**	**0.721**
Correctness					
	Building	Vegetation	Terrain	Structures	Clutter
RT	0.911	0.892	0.726	0.032	0.153
RD	0.979	0.904	0.835	0.092	0.271
RT2S	0.986	0.939	**0.944**	0.123	0.315
RT2S3	**0.987**	**0.952**	0.930	**0.169**	**0.327**
FS	0.985	0.939	0.942	0.165	0.324

Maldonado

Completeness					
	Building	Vegetation	Terrain	Structures	Clutter
RT	0.699	0.901	0.787	0.710	0.584
RD	0.926	0.892	0.918	0.780	0.874
RT2S	0.907	0.965	0.955	**0.957**	**0.978**
RT2S3	**0.944**	0.958	0.951	0.941	0.960
FS	0.917	**0.968**	**0.959**	0.920	0.975
Correctness					
	Building	Vegetation	Terrain	Structures	Clutter
RT	0.889	0.973	0.749	0.026	0.053
RD	0.957	0.977	0.825	0.125	0.207
RT2S	**0.982**	0.990	0.907	0.163	0.169
RT2S3	0.974	**0.992**	0.910	0.182	**0.456**
FS	0.969	0.989	**0.924**	**0.232**	0.161

Furthermore, we see that the RT2S feature set which only utilizes features obtained from the 2D orthomosaic has a higher performance than the RD feature set which includes features from the DSM. For example in Kigali, the OA increases from 86.7% (RD) to 90.4% (RT2S). This is largely due to a decrease in the confusion of building and terrain, which increases the completeness of buildings from 0.847 to 0.901 and correctness of terrain from 0.835 to 0.944. For example, Figure 2.1 displays how a building (marked with a yellow box), located on a steep slope where the roof is almost equal to the

terrain above it, is captured by RTST but not RD. In Maldonado, the improved classification accuracy of RT2S as opposed to RD is mainly due to the improved classification of vegetation vs. terrain, where the completeness of vegetation improves from 0.892 to 0.965.

The 10-class problem focused on the ability to distinguish between different types of objects. Regarding buildings, there were difficulties in distinguishing between older corrugated iron sheets and new, grey iron sheets due to similar texture and coloring. However, the completeness of the different building roof types was still above 0.849 and correctness above 0.796 for both datasets using the RT2S3 feature set. The corrugated iron sheets also displayed confusion with the bare surface class. The ability to distinguish high- and low-vegetation also had the best results for RT2S3, showing gains in correctness and completeness for both classes as opposed to RT2S or RD. the most notable exception is the decrease in the correctness of low vegetation in the Uruguay dataset from 0.873 (RT2S) to 0.839 (RT2S3), although feature selection again improves the correctness to 0.895. Bare surfaces displayed a higher confusion with corrugated iron roofs than impervious surfaces, again likely due to the similar spectral properties. Regarding the terrain classes, results indicate that for all feature sets, there was much confusion between bare terrain and impervious surfaces. This is a common problem in remote sensing, as shadows from surrounding buildings and spectral similarity with pervious surfaces hinders the identification of impervious surfaces (Weng, 2012). The most difficult classes in the current setup are the structure classes of free-standing walls and lamp posts. Although the inclusion of point cloud features greatly improved the classification of walls and lamp posts, these classes were still over predicted, resulting in low correctness values.

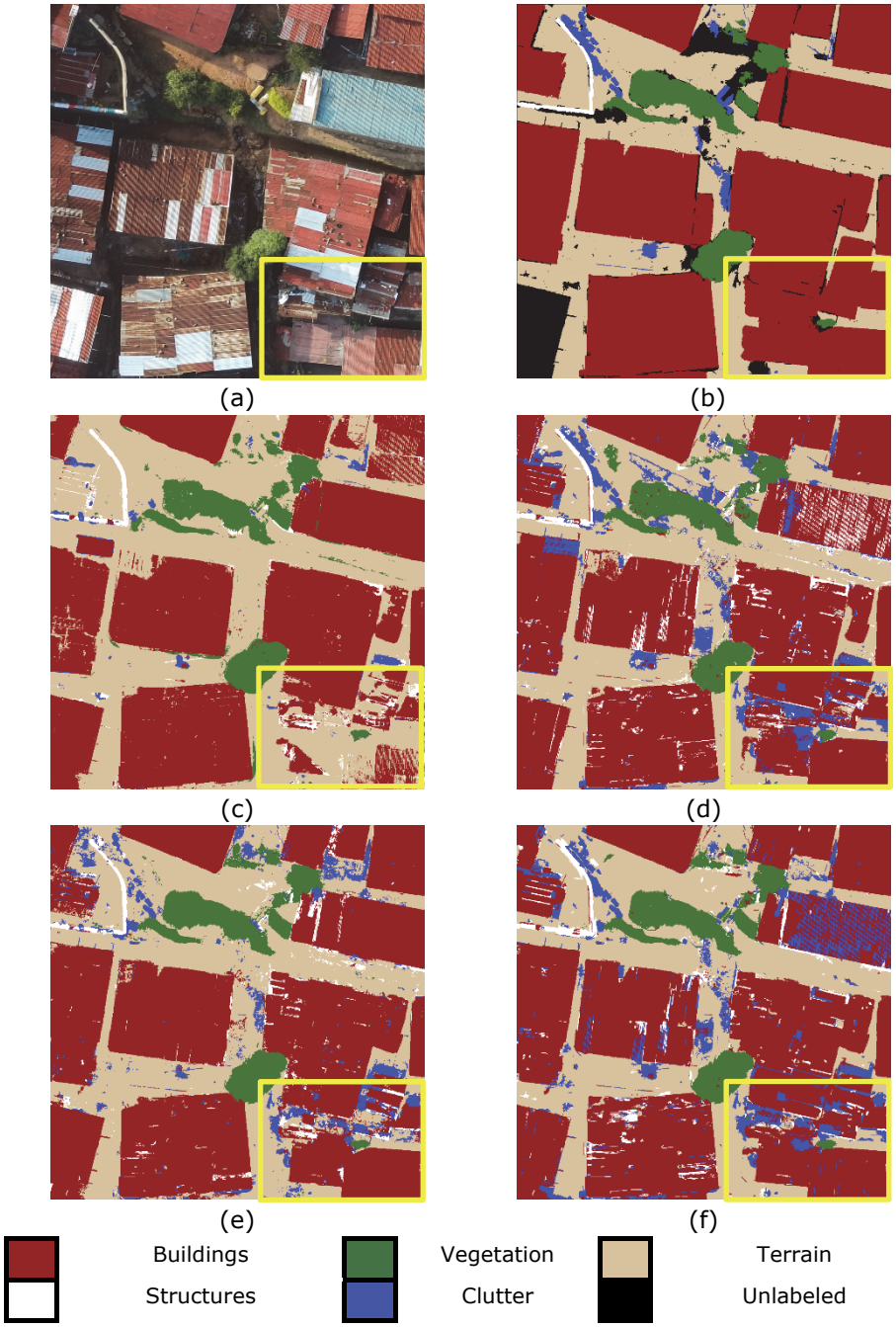


Figure 2.1: *Classification results (5-class) for one of the tiles in the Kigali dataset: input RGB image (a), reference data (b), RD prediction (c), R2ST prediction (d), R2ST3 prediction (e), and FS prediction (f). The yellow box indicates a building roof which is not captured in the RD feature set due to the steep slopes, but well captured in the RT2S, RT2S3, and FS sets.*

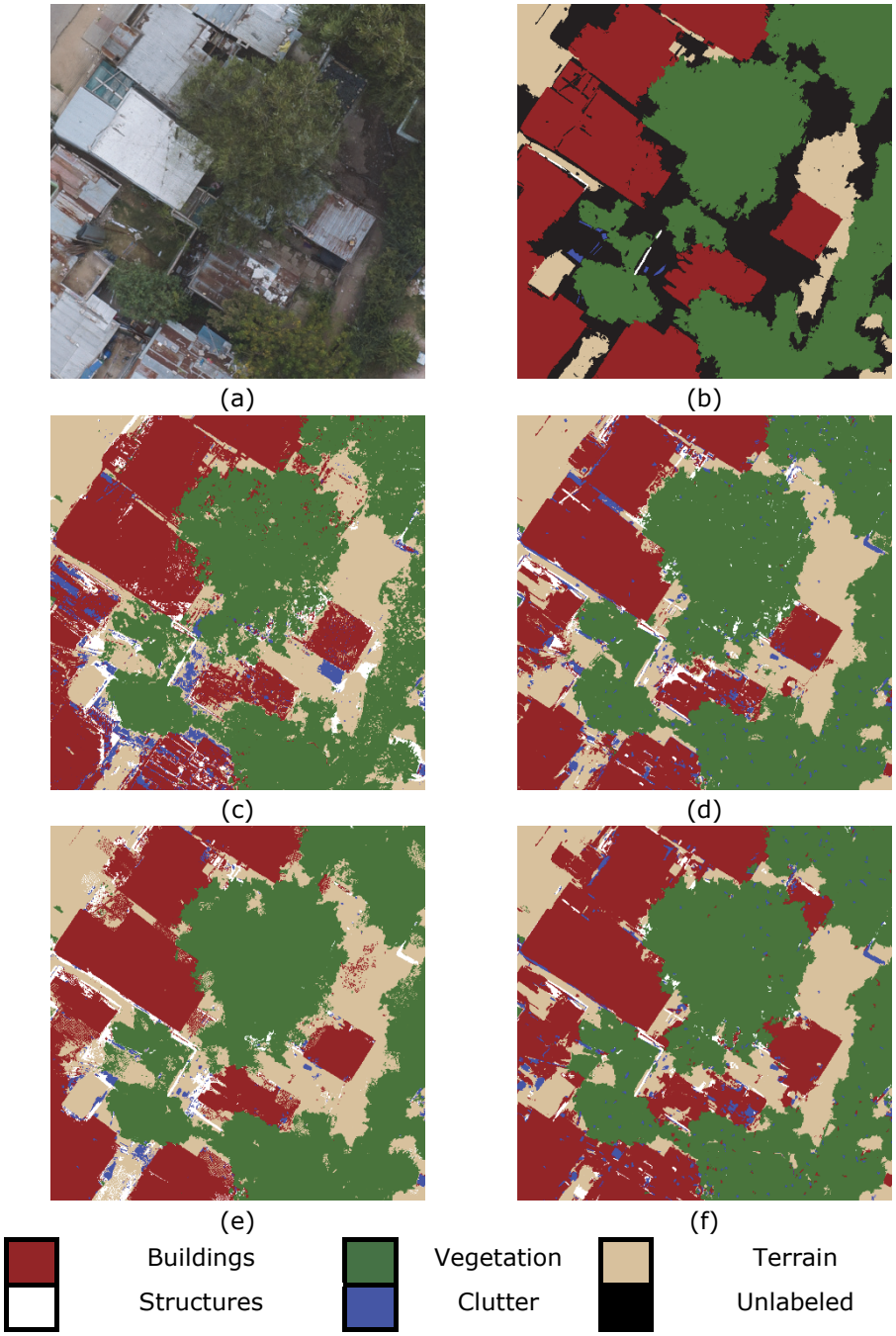


Figure 2.2: *Classification results (5-class) for one of the tiles in the Maldonado dataset: input RGB image (a), reference data (b), RD prediction (c), R2ST prediction (d), R2ST3 prediction (e), and FS prediction (f).*

Table 2.6: *Completeness and correctness of selected feature sets for the 10-class problem of the Kigali and Maldonado datasets. (CI = corrugated iron roof, Imperv. = impervious surface)*

		Buildings			Vegetation		Terrain		Structures		Clutter
		High I	CI	High II	High	Low	Bare	Imperv.	Walls	Lamps	Clutter
Kigali	*Completeness*										
	RT	0.847	0.487	0.943	0.568	0.553	0.627	0.397	0.474	0.942	0.412
	RD	0.946	0.829	0.964	0.795	0.880	0.759	0.612	0.800	0.993	0.527
	RT2S	0.962	0.858	0.972	0.825	0.909	**0.837**	**0.788**	0.832	**1.000**	**0.745**
	RT2S3	**0.979**	**0.890**	0.974	0.866	**0.951**	0.826	0.753	0.805	**1.000**	0.734
	FS	0.971	0.862	**0.979**	**0.883**	0.938	0.800	0.763	**0.878**	0.995	0.741
	Correctness										
	RT	0.195	0.910	0.874	0.766	0.226	0.609	0.484	0.037	0.002	0.134
	RD	0.584	0.979	0.925	0.889	0.466	0.739	0.705	0.122	0.018	0.217
	RT2S	0.770	**0.991**	**0.993**	0.920	0.620	**0.881**	**0.828**	0.097	0.036	**0.306**
	RT2S3	**0.796**	0.986	0.967	**0.931**	0.674	0.859	0.824	0.147	**0.254**	0.280
	FS	0.663	0.982	0.919	0.854	**0.682**	0.843	0.775	**0.194**	0.180	0.298
Maldonado	*Completeness*										
	RT	0.672	0.309	0.929	0.644	0.564	0.696	0.624	0.739	0.882	0.592
	RD	0.904	0.748	0.983	0.849	0.901	0.835	0.822	0.794	0.986	0.880
	RT2S	0.968	0.830	0.980	0.924	0.960	0.963	0.931	0.954	**1.000**	**0.980**
	RT2S3	0.962	**0.849**	0.980	0.923	0.952	0.945	0.912	0.934	**1.000**	0.972
	FS	**0.976**	0.846	**0.986**	**0.930**	**0.963**	**0.964**	**0.939**	**0.955**	0.995	0.972
	Correctness										
	RT	0.374	0.760	0.840	0.822	0.363	0.526	0.703	0.031	0.002	0.038
	RD	0.601	0.929	0.800	0.964	0.717	0.695	0.882	0.114	0.131	0.151
	RT2S	0.843	0.972	**0.956**	0.980	0.873	**0.898**	0.895	0.179	0.115	0.131
	RT2S3	0.864	0.972	0.899	0.982	0.839	0.818	**0.918**	0.152	**0.928**	**0.293**
	FS	**0.906**	**0.979**	0.929	**0.983**	**0.895**	0.881	0.896	**0.205**	0.193	0.132

As the results indicate the importance of including 3D features from the point cloud, it is interesting to analyze which 3D features (i.e. spatial binning, segment, or point-based features) are the most relevant for the various classes. Table 2.7 presents the results of the SFFS feature selection, identifying the most relevant 3D features for the ten different classes of both datasets. The most commonly selected 3D features are: the ratio of the 2D eigenvectors (selected 14 times), this ratio summarized over the mean shift segments (9 times), the standard deviation of the points per bin (8 times), the maximal height difference between a segment and the surrounding points (6 times), and the latter summarized over mean shift segments (6 times). Images displaying the values of these features are presented in Figure 2.3. It is interesting to note that the most common features represent point-based, spatial binning, and planar-segment based features respectively. With regard to buildings, the maximal height difference per planar segment, either pixel-based or averaged over mean shift segments, was selected as the most relevant feature five out of six times. The second most important feature was the ratio of 2D eigenvalues. Regarding the vegetation class, there was little consensus regarding the most important feature for high vegetation, whereas the maximal height difference per planar segment was selected as most important for low vegetation. The latter feature was also most important for identifying terrain, followed by the ratio of 2D eigenvalues and standard deviation per bin. The number of points per planar segment was important for both walls and street lights, whereas the features selected for the clutter class differed greatly between the both datasets.

2.4 Discussion

2.4.1 Importance of summarizing texture and 3D features over mean-shift segments

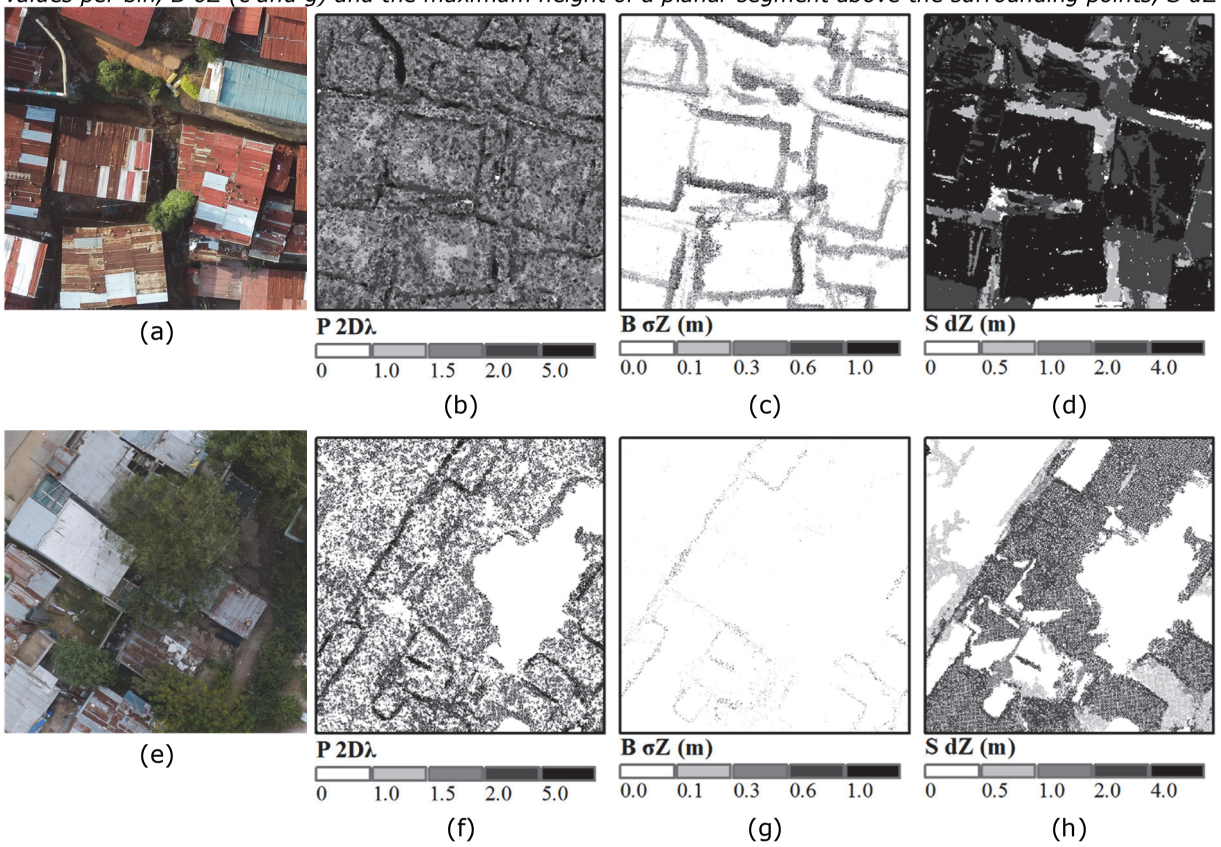
Out of all the feature sets, those which used a mean-shift segmentation to summarize texture or 3D features greatly increased the classification accuracy. As the extent of the moving window is fixed, it will summarize the textures of distinct classes at object borders, whereas this problem is avoided when using segments to summarize textural information. The discriminative power of the LBP texture features summarized per segment is possibly due to the high resolution of the UAV imagery, in which the texture of the corrugated iron roofs is clearly visible. Summarizing the 3D features over the segments also serves to decrease noise in the point cloud, such as outliers formed by dense matching errors. The R2ST3 feature set proves the utility of integrating both 2D and 3D features, especially in the context of the 10-class problem. In light of this observation, further research could, for example, summarize the 2D texture features over the planar 3D segments in order to reduce the misclassification of narrow corrugated iron sheets on roofs as walls.

2.4.2 Propagation of errors when using DSM features

Regarding the pixel-based methods, the suite of DSM top-hat filters allows for the distinction of objects of various sizes, which could be used to target elevated objects of uniform size. However, errors in the DSM are then propagated to the classification. For example, the Kigali and Maldonado datasets utilized an IDW interpolation incorporated to create the DSM, which causes the terrain next to or footpaths between buildings to be misclassified as building or vegetation since these pixels are falsely assigned a higher elevation value. This effect is more pronounced if training samples are not obtained from these locations. This hinders the suitability of the classification for upgrading projects, which requires the delineation of individual buildings or identification of footpaths to analyze accessibility in the settlement. As the building outlines are clearly visible in the orthophoto, the mean-shift segmentation improves the delineation of building outlines compared to relying on the DSM, especially when combined with 3D features.

This again emphasizes the advantage of 3D features summarized per image segment over the features extracted from a DSM. This advantage is accentuated if one takes into account that the manner in which the reference data was created, which was done manually over the orthophoto, favors features extracted from the imagery. Some objects, such as overhanging electricity wires or some poles, are not visible in the orthomosaic. They may therefore be labelled as 'terrain' in the reference data although they contain the 3D characteristics of a structure. Misalignment of the 3D point cloud and orthomosaic, although minimal in the current application as both products were derived from the same imagery, should be considered more carefully when utilizing the present methods to combine aerial imagery and LiDAR data.

Figure 2.3: *RGB images of a tile from the Kigali (a) and Maldonado (e) datasets, along with some of the most relevant features extracted from the corresponding point clouds: the ratio between the 2D eigenvectors, P 2D λ (b and f), standard deviation of height values per bin, B σZ (c and g) and the maximum height of a planar segment above the surrounding points, S dZ (d and h).*



2.4.3 Comparison of the three sets of 3D features

Three groups of features were derived from the point cloud to represent 3D attributes: spatial binning, planar segment features, and 3D point features. Results suggest that the three forms of 3D features are complimentary, as all three types are represented by the three most commonly selected features over all the classes. In the first place was the selection of the ratio of eigenvalues of the points within the optimal neighborhood projected into a 2D horizontal plane. Low values indicate that points are more evenly distributed on the horizontal plane (e.g. terrain and roofs), whereas higher values indicate a more linear structure (e.g. walls). The second feature was the standard deviation of points falling into the same bin. This feature indicates the different vertical layers of objects within the scene, as it will be lower at planar roofs or terrain, but higher for walls, lamp posts, and high vegetation. Regarding the third feature, in informal settlements, a single roof may consist of patches of materials displaying heterogeneous characteristics causing an over-segmentation of the mean shift based on radiometric features. However, these different materials may still generally lie on the same plane, allowing the 3D planar segments to summarize information over a larger area of the roof. The propagation of the maximal height difference with surrounding points over the entire planar segment is especially important for buildings and terrain. However, errors in the point cloud segmentation, such as building roofs and terrain being assigned to the same plane in sloped areas, caused visible artefacts in the classification. It is moreover interesting to observe that a number of 3D features averaged over mean shift segments were selected for the Kigali dataset, but not for the Maldonado dataset. It is unclear whether this would change if the mean shift segmentation parameters would be specifically tuned for each dataset, as the classification accuracies obtained with the current settings are still quite high.

In general, we can observe that some object classes such as buildings and terrain prioritize similar 3D features for both datasets. Others, such as clutter and vegetation display more variety. This is logical as different types of vegetation display different geometric characteristics and the lack of a NIR spectral band make it difficult to distinguish based on radiometric characteristics, though in this case the vegetation index and normalized green features provide adequate discriminative power. The clutter class is obviously more difficult to classify as it represents a wide range of objects with different spectral and geometrical properties.

2.4.4 Settlement heterogeneity and future applications

One of the main objectives of the current study is to analyze not only how to combine the 2D and 3D features, but also to investigate the transferability of these features to other settlements. The results indicate that despite the

different characteristics of the settlements, the classification accuracies using the different feature sets are comparable – that is to say that for both areas, the integration of 2D and 3D features achieves a high classification performance. Also, for all datasets the structure classes are over predicted, and there is difficulty in correctly classifying clutter, impervious surfaces, and bare surfaces.

Finally, it is important to consider the parameter tuning requirements. Many of the features require the definition of parameters: number of neighbors and radius for the texture features, structural element of the DSM top-hat features, planarity definitions in the surface growing for the planar segments, spatial and spectral bandwidths for the mean shift segmentation, etc. In the present analysis, these parameters were experimentally defined for the Kigali dataset, and the same values were utilized for the Maldonado dataset. This indicates that even without fine-tuning each of the parameters, high classification accuracies can be obtained. Applications to other study areas could use the values presented here as a starting point. SFFS could be applied to a training set to identify which of the features are most relevant for the new dataset, which could then be extracted to classify the entire study area. If the user would like to start fine-tuning the parameters, it is recommendable to give priority to the definition of the mean shift segmentation parameters, which must be lax enough to accommodate the spectral variability of, for example, the corrugated iron roofs, but fine enough to distinguish between building roofs and terrain.

2.5 Conclusions and Recommendations

This work illustrates the importance of integrating 2D radiometric, textural, and segment features, 2.5D topographical features, and 3D geometrical features for informal settlement classification. Through the integration of these features, a high classification accuracy can be obtained, despite the challenging characteristics of informal settlements, which often consist of small buildings with a mix of poor quality roof materials with clutter and possibly located on steep slopes. Various feature sets were applied to a 5-class problem: buildings, vegetation, terrain, structures (free-standing walls and lamp posts), and clutter (cars, laundry lines, miscellaneous objects on the ground); and a 10-class problem which distinguished roof material, high/low vegetation, pervious/impervious surfaces, and type of structure. Two informal settlements located in different settings were compared. Results indicate that using 2D radiometric features together with a series of top-hat morphological filters applied to the DSM had the highest accuracy of all pixel-based feature sets. However, inaccuracies in the DSM are propagated into the classification. Summarizing texture features over mean-shift segments obtains an improved classification even though it only requires the 2D RGB image as input. The high

spatial resolution of the UAV imagery allowed the texture features to capture typical corrugated iron roof patterns. However, the best results are obtained when integrating 3D features from the point cloud with image-based radiometric and texture features summarized over segments. The most relevant 3D features over the different datasets were: the ratio between the eigenvalues of the X,Y coordinates of a neighborhood of 3D points, the standard deviation in the height of points falling into the same bin of a defined grid, and the maximal height of a planar segment above neighboring points.

The observation that the highest classification accuracies were obtained by combining both 2D and 3D features for two datasets obtained from the same images demonstrates that both feature spaces contain complimentary information. As UAV imagery provides both a dense 3D point cloud and a high-resolution orthomosaic, both can be exploited to improve scene understanding. This is especially important in challenging scenes such as informal settlements, where many assumptions fundamental to building extraction algorithms (such as ground planarity and free-standing buildings) do not hold. Here, we demonstrate which feature sets can be combined to provide an accurate, up-to-date classification map of informal settlements, which is essential for upgrading projects. It also demonstrates the importance of using 3D features directly. Other studies can use the current findings to direct their attention to certain 3D features according to the target classes of their specific classification problem. Further research could focus on an analysis of how to fine-tune these features to enhance the recognition of various objects and materials in informal settlements. The application of this framework to multi-temporal settings could also be analyzed, as UAVs facilitate frequent image acquisition and may therefore be very useful for monitoring project implementation and impacts. Classification post-processing, which was considered outside the scope of the present study, could also reduce the presence of small pixel groups and improve the classification results.

Chapter 3 – Optimizing Multiple Kernel Learning for the Classification of UAV Data[2]

[2] This chapter is based on:

Gevaert, C.M., Persello, C., and Vosselman, G. (2016) 'Optimizing Multiple Kernel Learning for the Classification of UAV Data', *Remote Sensing*, 8, pp. 1025, doi:10.3390/rs8121025.

Abstract

Unmanned Aerial Vehicles (UAVs) are capable of providing high-quality orthoimagery and 3D information in the form of point clouds at a relatively low cost. Their increasing popularity stresses the necessity of understanding which algorithms are especially suited for processing the data obtained from UAVs. The features that are extracted from the point cloud and imagery have different statistical characteristics and can be considered as heterogeneous, which motivates the use of Multiple Kernel Learning (MKL) for classification problems. In this paper, we illustrate the utility of applying MKL for the classification of heterogeneous features obtained from UAV data through a case study of an informal settlement in Kigali, Rwanda. Results indicate that MKL can achieve a classification accuracy of 90.6%, a 5.2% increase over a standard single-kernel Support Vector Machine (SVM). A comparison of seven MKL methods indicates that linearly-weighted kernel combinations based on simple heuristics are competitive with respect to computationally-complex, non-linear kernel combination methods. We further underline the importance of utilizing appropriate feature grouping strategies for MKL, which has not been directly addressed in the literature, and we propose a novel, automated feature grouping method that achieves a high classification accuracy for various MKL methods.

3.1 Introduction

Unmanned Aerial Vehicles (UAVs) are gaining enormous popularity due to their ability of providing high-quality spatial information in a very flexible manner and at a relatively low cost. Another considerable advantage is the simultaneous acquisition of a photogrammetric point cloud (i.e., a 3D model consisting of a collection of points with X, Y, Z coordinates) and very high-resolution imagery. Due to these reasons, the use of UAVs for a wide range of applications is being analyzed, such as agriculture (Zhang and Kovacs, 2012; Gevaert *et al.*, 2015), forestry (Wallace *et al.*, 2012), geomorphology (Tarolli, 2014), cultural heritage (Remondino and Campana, 2014) and damage assessment (Anand Vetrivel *et al.*, 2015). Furthermore, the potential cost savings, improved safety and prospect of enhanced analytics they provide are being increasingly recognized as a competitive advantage from a business perspective (Thibault and Aoude, 2016).

Similar to traditional aerial photogrammetry, UAV imagery is processed to obtain a dense point cloud, Digital Surface Model (DSM) and orthomosaic. In (Nex and Remondino, 2014), the general workflow of utilizing UAVs for mapping applications is described. UAVs are generally mounted with a camera and fly over the study area to obtain individual overlapping images. Flights are planned according to the camera parameters, UAV platform characteristics and user-defined specifications regarding the desired ground sampling distance and image overlap. The acquired images are then processed using photogrammetric methods, for which semi-automatic workflows are currently implemented in various software (Sona *et al.*, 2014). Key tie points are identified in multiple images, and a bundle-block adjustment is applied to simultaneously identify the camera parameters of each image, as well as the location of these tie points in 3D space. Note that this step usually requires the inclusion of external ground control points for an accurate georeferencing. Dense matching algorithms, such as patch-based (Furukawa and Ponce, 2010) or semi-global (Hirschmüller, 2008) approaches, are then applied to obtain a more detailed point cloud. The point cloud is filtered and interpolated to obtain a DSM that provides the height information for the orthomosaic derived from the UAV images. Thus, geospatial applications making use of UAV imagery have access to the information in a point cloud, DSM and orthomosaic for subsequent classification tasks.

Much research regarding the classification of urban areas from aerial imagery still relies on features from either only the imagery or the imagery and DSM. For example, the orthomosaic can be divided into tiles and Linear Binary Pattern (LBP) texture features can be utilized to propose class labels that are present in that area (Moranduzzo *et al.*, 2015). Randomized Quasi-Exhaustive (RQE) feature banks can also be used to describe texture in UAV orthomosaics

for the purpose of classifying impervious surfaces (Tokarczyk *et al.*, 2015). Others use radiometric and Gray-Level Co-Occurrence Matrix (GLCM) texture features to identify inundated areas (Feng, Liu and Gong, 2015). The inclusion of elevation data greatly improves image classification results in urban areas (Hartfield, Landau and Leeuwen, 2011; Longbotham *et al.*, 2012). Indeed, a comparison of building extraction methods using aerial imagery indicates that the integration of image- and DSM-based features obtains high accuracies for (large) buildings (Rottensteiner *et al.*, 2014). However, combining the features derived from both the imagery and the point cloud directly (rather than the DSM) has been shown to prove beneficial for classification problems in the fields of damage assessment (A. Vetrivel *et al.*, 2015) and informal settlement mapping (Gevaert *et al.*, 2017).

Combining features from multiple sources or from different feature subsets pertains to the field of multi-view learning (Xu, Tao and Xu, 2013). For example, in this case, point-cloud-based and image-based features could be considered as different views of a study area. Although both are obtained from the same data source (UAV images), the point-cloud represents the geometrical properties of the objects in the scene, whereas the orthoimagery contains reflectance information (i.e., color). Three types of multi-view learning can be distinguished (Xu, Tao and Xu, 2013): co-training, sub-space learning and Multiple Kernel Learning (MKL). Co-training generally consists of training individual models on the different views and then enforcing the consistency of model predictions of unlabeled data. However, such methods require sufficiency, i.e., that each model is independently capable of recognizing all classes. This is not always the case, for example different roof types may be differentiated based on textural features from the imagery, but not geometrically distinguishable in the point cloud. Sub-space learning uses techniques, such as Canonical Correlation Analysis (CCA), to recover the latent subspace behind multiple views. The representation of samples in this subspace can be used for applications such as dimensionality reduction and clustering. MKL can be used in combination with kernel-based analysis and classification methods. Support Vector Machine (SVM) is a successful classification algorithm that utilizes a kernel function to map training sample feature vectors into a higher dimensional space in which the data are linearly separated. As a single mapping function may not be adequate to describe features with different statistical characteristics, MKL defines multiple mapping functions (either on different groups of features or the same group of features, but using different kernel parameters). A number of studies show that MKL achieves higher classification accuracies than single-kernel SVMs (Gönen and Alpaydın, 2011). For example, for the integration of heterogeneous features from LiDAR and multispectral satellite for an urban classification problem (Gu *et al.*, 2015). Although, as opposed to the integration of LiDAR and satellite imagery, the UAV point cloud and orthoimagery are obtained from a single set

of cameras and sensors, we could expect MKL to have a similar beneficial effect.

In this paper, we illustrate how the utilization of classification algorithms that are specifically tailored to the integration of heterogeneous features is more appropriate for exploiting the complementary 2D and 3D information captured by UAVs for challenging classification tasks. The objective of this paper is two-fold. Firstly, we demonstrate the importance of using classification algorithms, such as MKL, which support the integration of heterogeneous features for the classification of UAV data. Secondly, we describe various feature grouping strategies, including a novel automatic grouping strategy, and compare their performances using a number of state-of-the-art MKL algorithms. The methods are compared through a multi-class classification task using UAV imagery of an informal settlement in Kigali, Rwanda.

3.2 Background

Support Vector Machines (SVMs) are robust classifiers that are particularly suited to high dimensional feature spaces and have proven to obtain high classification accuracies in remote sensing applications (Bruzzone and Persello, 2010). These discriminative classifiers identify the linear discriminant function that separates a set of *n* training samples $\{\mathbf{x}_i, y_i\}_{i=1}^{n}$ representing two classes $y_i \in \{-1, +1\}$ based on their respective feature vectors $\mathbf{x}_i$ in a non-linear feature space obtained by a mapping function $\phi(\mathbf{x})$:

$$f(\mathbf{x}) = \langle \mathbf{w}, \phi(\mathbf{x}) \rangle + b, \tag{3-1}$$

where *b* is a bias term and **w** is the vector of weight coefficients, which can be obtained by solving a quadratic optimization problem defined as:

$$\min \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^{n} \xi_i, \tag{3-2}$$

$$with\ respect\ to: \mathbf{w} \in \mathbb{R}^q, \boldsymbol{\xi} \in \mathbb{R}_+^n, b \in \mathbb{R}$$
$$subject\ to: y_i(\langle \mathbf{w}, \Phi(\mathbf{x}_i) \rangle + b) \geq 1 - \xi_i \ \forall i = 1, \ldots, n.$$

where *C* is a regularization parameter representing the relative cost of misclassification, ξ_i represent the slack variables associated with training samples, and *q* is the dimensionality of the feature space obtained by $\phi(\mathbf{x})$. Rather than calculating the mapping function $\phi(\mathbf{x})$, the kernel trick can be employed to directly obtain a non-linear similarity measure between each pair of samples $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$. The optimization function is then solved using the Lagrangian dual formulation as follows:

$$\max \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j), \tag{3-3}$$

$$subject\ to\ \sum_{i=1}^{n} y_i \alpha_i = 0\ \ and\ 0 \le \alpha_i \le C\ \forall\ i = 1, \dots, n$$

where α_i are the Lagrangian multipliers. Various kernel functions are described in the literature, such as the common (Gaussian) Radial Basis Function (RBF) kernel:

$$k_{RBF}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\sigma^2}\right) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|_2^2\right). \quad (3\text{-}4)$$

The RBF kernel function has one parameter, σ (often replaced by $\gamma = 1/2\sigma^2$), which represents the bandwidth of the Gaussian function. The bandwidth parameter can be determined by heuristics, such as the median distance between samples (Gretton *et al.*, 2006) or cross-validation (Tuia *et al.*, 2010; Gu *et al.*, 2015).

Intuitively, one can understand that not all features may be best represented by the same kernel parameters. Instead, Multiple Kernel Learning (MKL) utilizes *P* independent input kernels, which allow nonlinear relations between training samples to be described by differing kernel parameters and/or differing input feature combinations. The calculation of the similarity between each pair of training samples using different kernel functions results in *P* different kernel matrices $\boldsymbol{K}_m$ that are then linearly or non-linearly combined into a single kernel $\boldsymbol{K}_\eta$ for the SVM classification:

$$\boldsymbol{K}_\eta(\mathbf{x}_i, \mathbf{x}_j) = f_\eta\left(\{\boldsymbol{K}_m(\mathbf{x}_i^m, \mathbf{x}_j^m)\}_{m=1}^{P} \middle| \boldsymbol{\eta}\right). \quad (3\text{-}5)$$

There are a number of advantages of MKL compared to standard SVM methods. Firstly, as it allows kernel parameters to be adapted towards specific feature groups, it may enhance the class separability. Secondly, the combined kernel $\boldsymbol{K}_\eta$ can be constructed by assigning various weights to the input kernels, thus emphasizing more relevant features. In extreme cases, certain feature kernels may be assigned a weight of zero, thus causing the MKL to act as a feature selection method. Due to these characteristics, MKL is an appropriate classification method for combining features from heterogeneous data sources.

Much of the research regarding MKL for classification focusses on the strategies that are used to combine the input kernels. For example, a fixed rule can be adopted, where each kernel is given an equal weight (Pavlidis *et al.*, 2001). Alternatively, the individual kernel weights could then be determined based on similarity measures between the combined kernel and, for example, an optimal kernel ($\boldsymbol{K}_y = \mathbf{y}\mathbf{y}^T$), which perfectly partitions the classes. Niazmardi et al. (Niazmardi *et al.*, 2016) refer to these as two-stage algorithms, as opposed to single-stage algorithms that optimize the kernel weighting and SVM parameters simultaneously. Although the latter group of methods, including SimpleMKL (Rakotomamonjy *et al.*, 2008) and Generalized MKL (Varma and

Babu, 2009), are more sophisticated and may potentially achieve higher classification accuracies, they often imply a higher computational complexity.

In fact, a review of MKL methods (Gönen and Alpaydın, 2011) suggests that although MKL leads to higher classification accuracies than single-kernel methods, more complex kernel combination strategies do not always lead to better results. Rather, simple linear combination strategies seem to work well for non-linear kernels, such as RBF kernels. Others reached similar conclusions, stating that "baseline methods", such as averaging or multiplying kernels, reach similar accuracies as more complex algorithms, but at a much lower computational complexity (Gehler and Nowozin, 2009).

Considering that one of the main motivations behind MKL is the ability to adapt kernel parameters for the various features, it is surprising that little work has been done regarding how to divide features into groups so they optimally benefit from the tailored feature mappings. Various MKL studies report different grouping strategies, but none of them seem to compare a wide range of grouping strategies and compare the influence of the grouping strategies on the classification accuracy. Intuitively, such an optimal feature grouping should: (i) group features that are optimally represented by the same kernel parameters; and (ii) group features in a way that allows less or non-relevant features to be suppressed by the kernel weighting strategy. The main difficulty is that measures used to determine the latter, i.e., feature relevance, through non-linear similarity measures often depend on the former, i.e., the chosen kernel parameters. At the same time, the optimal values for these kernel parameters depend on which features are included in the group.

In practice, some studies assign each feature to a unique kernel (Tuia *et al.*, 2010). This allows the optimal kernel parameters to be defined per feature and a feature selection to be introduced through MKL methods promoting sparsity. Alternatively, a multi-scale approach can be adopted (Yeh *et al.*, 2012; Gu *et al.*, 2015) which defines a range of r bandwidth parameters for each feature f out of a total of n_f input features to create a total of $r \cdot n_f$ kernels as the input for the MKL. However, such approaches may fail to describe the complex relations between features and suffer an increased computational complexity due to the presence of more kernels. Another grouping strategy depends on the origins of the image features. In this case, separate kernels are defined for spectral or spatial features, or multi-spectral and radar imagery, and have been shown to outperform assigning features to individual kernels (Tuia *et al.*, 2010). In an effort to define the groups automatically, one could consult the literature on view construction for co-training, the first multi-view learning technique. In general, co-training seems to work well when the different views provide complementary information (Xu, Tao and Xu, 2013). This is similar to the observations of Di and Crawford (Di and Crawford, 2012), who found that

using the "maximum disagreement" approach to spectral bands performed better than uniform or random sampling for hyperspectral image classification.

However, a direct comparison of different grouping strategies for MKL using heterogeneous features is still lacking. This paper addresses this issue and proposes an automatic algorithm that extracts potential kernel parameters from the training data and performs a backward feature-selection strategy to determine which features should be included in each kernel. It thus simultaneously functions as both a feature grouping and feature selection method.

3.3 Materials and Methods

Multiple kernel learning can be applied to UAV data through the workflow presented in Figure 3.1. Based on the input data, such as point clouds and imagery, the first step consists of extracting the relevant features from the input data. This may be supplemented by a feature selection strategy if desired. The next step is to divide the features into groups to define the input kernels, which are then combined into a single kernel that is used to define the SVM classifier. Depending on which MKL method is employed, the parameters for kernel weighting and SVM may be optimized jointly or separately. In the following section, we describe the heterogeneous features utilized in this study for the classification (Section 3.3.1), various feature grouping strategies (Section 3.3.2) and the multiple kernel learning algorithms utilized to classify the data (Section 3.3.3).

3.3.1 Feature Extraction from UAV Data

Four types of features were derived from the orthomosaic and point cloud: 14 image-based radiometric features, 54 image-based texture features, 22 3D features per pixel and 22 3D features averaged over image segments (Table 3.1). The image-based radiometric features consist of the original R, G, B color channels of the orthomosaic, their normalized values (r, g, b) and the ExG(2) vegetation index: $\mathrm{ExG}(2) = 2\mathrm{g} - \mathrm{r} - \mathrm{b}$ (Woebbecke *et al.*, 1995), at the pixel-level and averaged over image segments. Here, the segments were obtained through a mean shift segmentation (Comaniciu and Meer, 2002) with a spatial bandwidth of 20 pixels and a spectral bandwidth of five gray values.

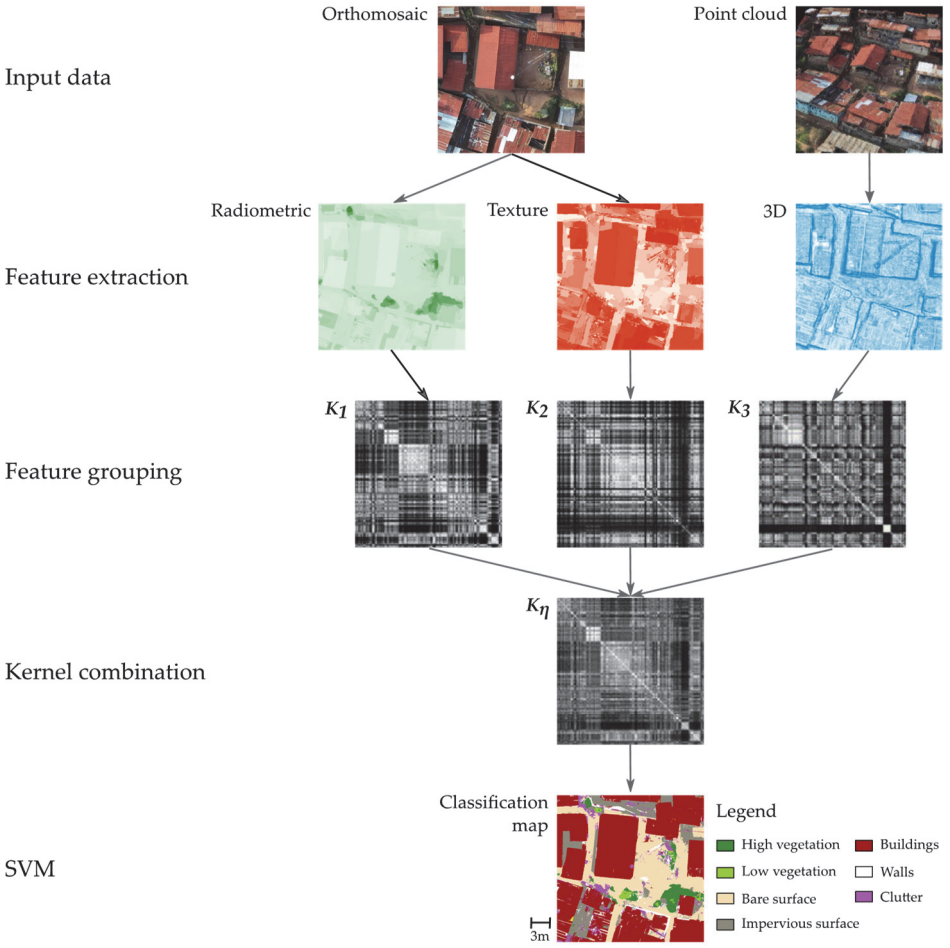


Figure 3.1: *An illustrative example of the multiple kernel learning workflow for UAVs: first, features must be extracted from the orthomosaic and the point cloud; then, the features are grouped, and the* $\boldsymbol{K}_m$ *input kernels are constructed. MKL techniques are used to combine the different input kernels into the combined kernel* $\boldsymbol{K}_\eta$*, which is used to construct the SVM and perform the classification.*

The image-based texture features are represented by Local Binary Pattern (LBP) features (Ojala, Pietikainen and Maenpaa, 2002). These rotationally-invariant texture features identify uniform patterns, such as edges and corners, based on a defined number of neighboring pixels (N) at a distance (R) from the center pixel. The relative presence of each $N + 2$ texture pattern in the local neighborhood can be summarized by constructing a normalized histogram for each mean shift segment, where the frequency of each bin is used as a feature.

Table 3.1: *A list of the features extracted from the point cloud and orthomosaic in the current study. N refers to the number of features in the group*

Type of Feature	*N*	Source	Description	
Radiometric	14	Image	Pixel-based	Color (R, G, B) Normalized color (r, g, b) Vegetation index (ExG(2))
			Segment-based	Color (R, G, B) Normalized color (r, g, b) Vegetation index (ExG(2))
Texture	54	Image	Local Binary Patterns	$LBP_{R=1,N=8}$ $LBP_{R=2,N=16}$ $LBP_{R=3,N=24}$
3D features	22	Point cloud	Spatial binning	Points per pixel Max. height difference Height standard deviation
			Planar segments	Number of points Average residual Inclination angle Max height difference
			Local neighborhood	Linearity, planarity, planarity (2), scattering, omnivariance, anisotropy, eigenentropy, sum of eigenvalues, curvature, maximum height, range of height values, standard deviation of height values, inclination angle, sum of 2D eigenvalues, ratio of 2D eigenvalues
3D features per image segment	22	Both	Same as point cloud features, but averaged over image segments	

The third type of features consist of 3D features extracted from the point cloud: spatial binning features, planar segment features and local neighborhood features. Spatial binning features describe the number of 3D points

corresponding to each 2D image pixel, as well as the maximal height difference and height standard deviation of these points. Planar segment features are obtained by applying a surface growing algorithm to the point cloud (Vosselman, 2012). The algorithm calculates the planarity of the 10 nearest neighbors for each seed point and adds points within a radius of 1.0 m, which are within a 0.30-m threshold from the detected plane. The latter threshold is relatively high compared to the spatial resolution as in the informal settlement, there are often objects, such as rocks or other clutter, on top of roofs. This could result in non-planar objects, such as low vegetation, being considered as planar. However, such class ambiguities may be rectified through the other features included in our feature set. From each planar segment, four features were extracted: the number of points per segment, average residual to the plane, inclination angle and maximal height difference to the surrounding points. The local neighborhood features are based on the observation that the ratio between the eigenvalues of the covariance matrix of the XYZ coordinates of a point's nearest neighbors can represent the shape of the local neighborhood (Demantké *et al.*, 2012). For example, the relative proportions between these eigenvalues may describe the local neighborhood as being planar, linear or scattered. More specifically, we consider an optimal neighborhood around each 3D point to define the covariance matrix and extract the 3D features described in the framework (Weinmann *et al.*, 2015). To assign 3D features calculated in the point cloud to 2D space, the attributes of the highest point for each pixel in the orthomosaic were assigned to the pixel in question. We thus obtain 3D features (spatial binning, planar segment and neighborhood shape) for each pixel.

The fourth and final type of features consist of averaging these pixel-based 3D features over the image segments. For a more detailed description of how the various features were extracted, the reader is referred to Gevaert et al. (Gevaert *et al.*, 2017). All feature values are normalized to a scale between 0 and 1 before feature grouping and classification.

3.3.2 Feature Grouping Strategies

3.3.2.1 Reference Feature Grouping Strategies: After calculating the input features, each feature f in the complete set of features S ($f \in S$), where n_f indicates the total number of features, must be assigned to a group G_m, m = 1,…, P, where each group will form an individual input kernel $\boldsymbol{K}_m$. Seven MKL grouping strategies are compared based on: (i) individual kernels; (ii) prior knowledge; (iii) random selection; (iv) feature similarity; (v) feature diversity; (vi) the kernel-based distance between samples; and (vii) a novel multi-scale Markov blanket selection scheme. In Case (i), each feature is assigned to an individual input kernel, so 112 features result in 112 input kernels $\boldsymbol{K}_m$. The prior knowledge strategy of Case (ii) consists of four kernel groups according to feature provenance: image-based radiometric features, image-based

texture features, 3D features per pixel and 3D features averaged over image segments (i.e., the four types of features listed in Table 3.1). In Case (iii), the random selection strategy divides the features arbitrarily into a user-defined number of features groups. For Case (iv) the similarity strategy is represented by a kernel k-means clustering (Schölkopf, Smola and Müller, 1998) over the feature vectors, thus grouping them into clusters that expose similar patterns in the input data. Di and Crawford (Di and Crawford, 2012) found that such an approach worked better than uniform or random feature grouping for multi-view active learning in hyperspectral image classification tasks.

For Case (v), diverse kernels are obtained by solving the Maximally-Diverse Grouping Problem (MDGP) through a greedy construction approach (Gallego *et al.*, 2013). The basic idea of the approach is to iteratively select one unassigned feature, calculate the value of a disparity function considering the assignation of this feature to each feature group G_m and appoint the feature the group to which its membership would maximize the disparity. To do this, the user first defines the desired number of groups P, as well as the minimum (a) and maximum (b) number of features per group. The population of each group G_m is started by randomly selecting one of the features in the feature set S. The remaining features in the set of variables not yet assigned to a group are iteratively assigned to one of the groups G_m. One feature f_i is selected at random, and the disparity function D_{f,G_m} (6) is calculated considering its inclusion into each group, which has not yet reached the minimal number of features (i.e., $|G_m| < a$). Once each group has reached the minimal number of features, each group that has not yet reached the maximum (i.e., $|G_m| < b$) is considered. Here, the disparity function describes the normalized sum of the distances between features:

$$D_{f_i,G_m} = \frac{\Sigma_{j \in G_m} d_{ij}}{|G_m|}, \tag{3-6}$$

where $|G_m|$ is the number of elements in group G_m and the distance d_{fi} is obtained from the Sample Distance Matrix (SDM). In this case, the SDM is an $n_f \times n_f$ matrix where the element $SDM_{i,j}$ gives the ℓ_2-norm of the difference between features f_i and f_j. In other words, the disparity function is defined as the sum of the Euclidean distance between all of the features within a group over all of the samples divided by the number of features within the group.

Allowing kernel parameters to be set differently for various groups of features has been mentioned as one of the benefits of MKL. Furthermore, the mean distance between training samples is sometimes used as a heuristic for the bandwidth parameter of an RBF kernel (Gretton *et al.*, 2006). Therefore, we also analyze the utility of grouping features based on the distance between samples. For Case (vi), we consider using the median (a) within-class vs. (b) between-class distances, as well as (c) a combination of both distances to

group the features. To implement this, an SDM is constructed for each single feature. Note that here, the SDM is a $n_s \, x \, n_s$ matrix representing the distance between the samples as opposed to the feature distance matrix described in the previous paragraph. The median within-class and between-class distances for each class is obtained by finding the median of the relevant SDM entries and using this median as a feature attribute. For example, the within-class distance of class *u* is the median of the SDM entries of all rows and columns representing samples belonging to class *u*. Similarly, the between-class distance of class *u* is the median of all entries corresponding to rows of samples labeled as *u* and columns of all samples not labelled as *u*. Note that the median is used instead of the mean to reduce the effect of possible outliers. A classification problem with *Q* classes will thus result in *Q* feature attributes representing within-class distances, and *Q* attributes representing between-class distances. These are simply concatenated to *Q* + *Q* attributes for the third approach (i.e., within- and between-class distances). These feature attributes are then used as the input for a kernel k-means clustering. Thus, features that have similar (within- or between-class) sample distances and that may thus be best represented by the same bandwidth parameter will be grouped together.

3.3.2.2 Proposed Feature Grouping Strategy: For the final method (vii), we propose an automatic grouping strategy. Remember that the benefits of applying MKL rather than single-kernel SVM models include feature weighting and the use of different kernel parameters for various feature groups. Regarding the former, MKL allows some feature groups to be given more emphasis; in some cases, it may even assign certain kernels a weight η_m of zero, thus suppressing noisy or irrelevant feature groups. However, MKL can only suppress certain input kernels and, therefore, can only function as a feature selector if each feature is indeed assigned to a unique kernel, as in Case (i) above. This may fail to account for non-linear relationships that would be identified if features are combined in the same kernel. The second potential benefit of MKL was to allow different kernel mappings for the different feature groups.

Gu et al. (Gu *et al.*, 2015) even recommended using different bandwidths for the same feature groups, allowing similarities between samples to be recognized along multiple scales. They used pre-defined bandwidth intervals from 0.05 to 2.0 in intervals of 0.05. Yeh et al. (Yeh *et al.*, 2012) also construct multiple input kernels for each feature by selecting different bandwidth parameters, defining a 'group' as the conjunction of different kernel mappings of a single feature. MKL is applied to ensure sparsity amongst features, thus functioning as a feature selection method. Unlike (Gu *et al.*, 2015), they select the bandwidth parameters in a data-driven manner based on the standard deviation of the distance between all training samples. Although both methods

enable a multi-scale approach to define optimal feature representations using multiple bandwidth parameters, both methods pre-define which features will be grouped together in the input kernels. This could potentially lose non-linear relations between different features.

A good feature grouping strategy should therefore remove irrelevant features before kernel construction and allow features to be grouped according to optimal kernel parameters. The novel feature grouping algorithm we propose here does this by first analyzing the dataset to identify candidate bandwidth parameters, performs a feature ranking for each candidate bandwidth and restricts the features within each group according to a pre-defined threshold. The latter can be based on either defining the number of features per kernel or by defining a limit to the cumulative feature relevance. This simultaneous feature grouping and feature selection workflow consists of three steps: (i) selecting candidate bandwidth parameters; (ii) ranking the features using each parameter; and (iii) defining the cut-off criterion that selects the number of features per group (Figure 3.2). An additional benefit of the method is that it provides a heuristic for choosing the bandwidth parameter for the RBF kernel.

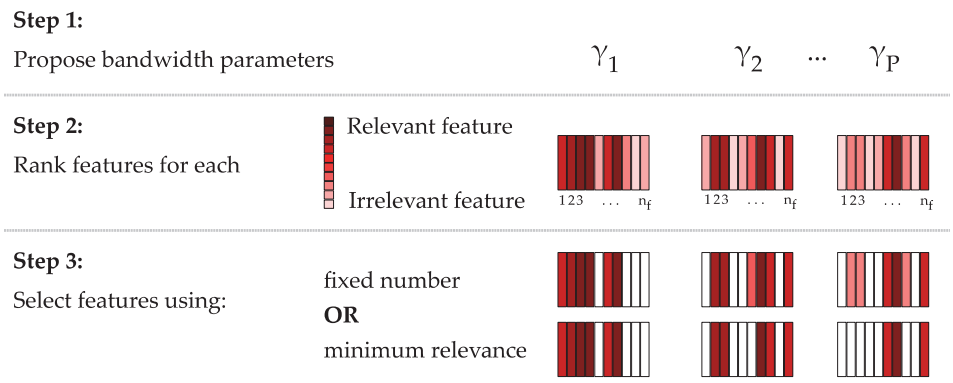
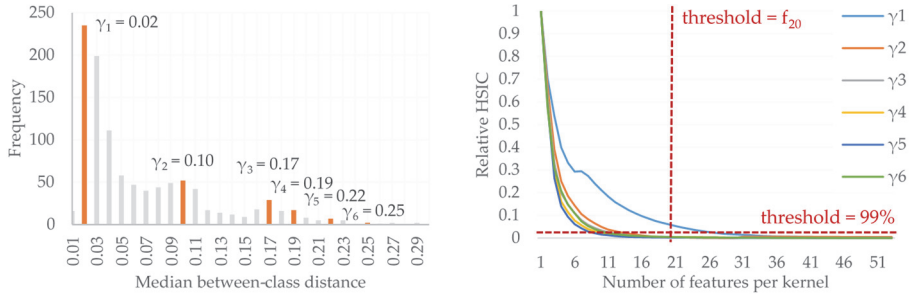


Figure 3.2: *A graphical illustration indicating how the automatic feature grouping strategy works. Step 1 consists of proposing a number of bandwidth parameters for the RBF kernel; in Step 2, a feature ranking is done using backwards-elimination and a kernel-class separability measure with the assigned γ to determine the relative relevance of each* n_f *features; in Step 3, a feature set is selected for each kernel based on (i) using a fixed number of features per kernel, e.g., six in the illustrated example, or (ii) a minimum cumulative feature relevance level, which may result in different numbers of features per kernel.*

In the first step, potential bandwidth parameters are identified by selecting the median between-class distances for each feature. These between-class distances are obtained by selecting all entries of the SDM that correspond to two samples from different classes. A histogram of these between-class distances is constructed, from which automatic methods can be used to select the potential bandwidth parameters (Figure 3.3a). Here, we simply select

histogram bins associated with local maxima, i.e., which have a higher frequency than the two neighboring bins. Each of the bins corresponds to a potential bandwidth parameter; thus, different bandwidths are used for the various kernel groups and capturing data patterns at multiple scales. If the histogram does not present local peaks, other strategies could be considered, such as taking regular intervals over the possible between-class distances.



(a) (b)

Figure 3.3: *The proposed feature selection method, first using peaks in the between-class distance histogram to identify candidate bandwidth parameters (a); and then using the feature ranking to determine which features to include in each group (b). The dashed red lines indicate the cut-off thresholds according to either a maximal number of features per kernel (f_{20}) or relative HSIC value (99%). Note that the graphs represented here do not reflect the exact data from the experiments, but have been slightly altered for illustrative purposes.*

In the second step, a feature selection method based on a kernel-based class separability measure and backwards-elimination (Strobl and Visweswaran, 2014) is employed to determine which features to include in each kernel. The idea is to use a supervised strategy to identify the Markov blanket of the class labels (Strobl and Visweswaran, 2014). That is to say, we attempt to identify which features are conditionally independent of the class labels given the remaining features. These conditionally independent features therefore do not influence the class labels and may be removed. By using kernel class separability measures, we can identify non-linear class dependencies in the reproducing kernel Hilbert space. This is implemented by constructing a kernel using all of the features and the candidate bandwidth in question and calculating the class separability measure for an ideal kernel. One by one, the features are removed, and the measure is calculated again. The feature whose removal results in the lowest decrease in class separability is considered to be the least relevant and is removed from the set. The process is repeated until all features are ranked from most to least relevant for each candidate bandwidth. The method would work with any kernel-based class separability measure.

Finally, the user must define the cut-off metric of which features to select in each kernel based on the provided feature ranking for each bandwidth. In this

case, the user can choose to either define the maximal number of features per kernel or to use a cumulative relevance metric, such as selecting the number of features that first obtain 99.9% of the maximum cumulative similarity measure provided by the feature ranking (Figure 3.3b). It should be noted that this feature grouping strategy also allows for a single feature to be included in various kernels. In theory, this could result in two groups containing identical features, but represented by different bandwidth parameters. The proposed methodology also potentially functions as a feature selection method, as irrelevant or redundant features are likely to be at the bottom of the feature ranking and may therefore not be included in any of the input kernels.

3.3.3 Kernel Weighting Strategies

3.3.3.1 Class Separability Measures and Ideal Kernel Definition: Various studies report that there are no large differences in different multiple kernel learning methods in terms of accuracy (Gönen and Alpaydın, 2011). Furthermore, two-stage algorithms that update the combination function parameters independently from the classifier have a lower computational complexity (Niazmardi *et al.*, 2016). Therefore, we hypothesize that the use of kernel class separability measures, or kernel alignment measures, between the individual kernels and an ideal kernel will provide an advantageous trade-off between computational complexity and classification accuracy. The ideal target kernel represents a case of perfect class separability, where kernel values for samples from the same class maximal and samples from different classes have minimal kernel values. Therefore, the similarity of an input kernel $\boldsymbol{K}_m$ to a target ideal kernel $\boldsymbol{K}_y$ provides an indication of the class separability given the input kernel. Such measures may be used to optimize kernel parameters or to define the proportional weights of the various feature kernels in the weighted summation. In this case, named class-separability-based MKL (CSMKSVM), the class separability measure $\mathcal{R}$ of each individual kernel $\boldsymbol{K}_m$ and an ideal kernel $\boldsymbol{K}_y$ is calculated, and then, a proportional weighting is applied as follows:

$$\eta_m = \frac{\mathcal{R}(\boldsymbol{K}_m, \boldsymbol{K}_y)}{\sum_{h=1}^{P} \mathcal{R}(\boldsymbol{K}_h, \boldsymbol{K}_y)} \; \forall m, \tag{3-7}$$

Qiu and Lane (Qiu and Lane, 2009) used a similar heuristic based on the kernel alignment measure (Cristianini *et al.*, 2002). Here, we compare four class separability measures found in the literature: the square Hilbert-Schmidt norm of the cross-covariance matrix (HSIC) (Gretton *et al.*, 2005; Persello and Bruzzone, 2016) (8); (ii) Kernel Alignment (KA) (Cristianini *et al.*, 2002) (9); (iii) Centered-Kernel Alignment (CKA) (Cortes, Mohri and Rostamizadeh, 2010) (10); and (iv) Kernel Class Separability (KCS) (Ramona, Richard and David, 2012) (11).

$$HSIC\left(\boldsymbol{K}_x, \boldsymbol{K}_y\right) = \frac{1}{n^2} Tr\left(\boldsymbol{K}_{\mathrm{x}} \boldsymbol{H} \boldsymbol{K}_y \boldsymbol{H}\right) \tag{3-8}$$

$$KA(\boldsymbol{K}_x, \boldsymbol{K}_y) = \frac{\langle \boldsymbol{K}_x, \boldsymbol{K}_y \rangle_F}{\sqrt{\langle \boldsymbol{K}_x, \boldsymbol{K}_x \rangle_F \langle \boldsymbol{K}_y, \boldsymbol{K}_y \rangle_F}} \tag{3-9}$$

$$CKA(\boldsymbol{K}_x, \boldsymbol{K}_y) = \frac{\langle \boldsymbol{K}_x^c, \boldsymbol{K}_y^c \rangle_F}{\sqrt{\langle \boldsymbol{K}_x^c, \boldsymbol{K}_x^c \rangle_F \langle \boldsymbol{K}_y^c, \boldsymbol{K}_y^c \rangle_F}} \tag{3-10}$$

where $\boldsymbol{K}^c = \boldsymbol{K} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\boldsymbol{K} - \frac{1}{n}\boldsymbol{K}\mathbf{1}\mathbf{1}^T + \frac{1}{n^2}(\mathbf{1}^T\boldsymbol{K}\mathbf{1})\mathbf{1}\mathbf{1}^T$

$$KCS(\boldsymbol{K}_x, \boldsymbol{K}_y) = \frac{\Sigma \boldsymbol{W} - \frac{1}{n}\Sigma \boldsymbol{K}_x}{tr(\boldsymbol{K}_x) - \Sigma \boldsymbol{W}} \tag{3-11}$$

where $\boldsymbol{W} = \frac{1}{n}\begin{pmatrix} \frac{1}{n_1}\boldsymbol{K}_{11} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{n_Q}\boldsymbol{K}_{QQ} \end{pmatrix}$

where $\boldsymbol{H}_{ij} = \delta_{ij} - (\frac{1}{n})$, δ_{ij} being the Kronecker delta and having a value of 1 if *i* and *j* adhere to the same class and 0 if they have different class labels, *Tr*(.) indicates the trace function, *n* is the total number of samples, n_1 is the number of samples in the first class, n_Q is the number of samples in class *Q*, $\langle \boldsymbol{K}_x, \boldsymbol{K}_y \rangle_F = \sum_{i,j=1}^{n_s} \boldsymbol{K}_x(\boldsymbol{x}_i, \boldsymbol{x}_j)\boldsymbol{K}_y(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and $\mathbf{1}$ is an *n* × 1 vector of ones. $\boldsymbol{K}_x$ is any input kernel and could therefore corresponds to either $\boldsymbol{K}_m$ or $\boldsymbol{K}_\eta$ depending on whether the class separability measure is being calculated for the input kernel or combined kernel, respectively.

3.3.3.2 Comparison to Other MKL Methods: Once the most adequate kernel class separability measure and ideal kernel definition have been selected, the following experiments serve to compare the proposed method to benchmark MKL methods and the influence of the various feature grouping strategies. Six benchmark Multiple Kernel SVM (MKSVM) methods are selected from the MATLAB code provided by Gönen and Alpaydin (Gönen and Alpaydın, 2011) (https://users.ics.aalto.fi/gonen/) and compared to the kernel Class-Separability method (CSMKSVM) described previously. They consist of methods using a Rule-Based linearly-weighted combination of kernels (RBMKSVM), Alignment-Based methods based on the similarity of the weighted summation kernel and an ideal kernel (ABMKSVM) and methods that initiate a linearly (Group Lasso-based MKL (GLMKSVM) and SimpleMKL) or nonlinearly (Generalized MKL (GMKSVM) and Non-Linear MKL (NLMKSVM)) combined kernel and use the dual formation parameters to iteratively update the weight. The first, RBMKSVM, is a fixed-rule method in which each kernel is given an equal weight $1/P$, and the resulting combined kernel is therefore simply the

mean of the input kernels. This is followed by CSMKSVM, where the weights are defined by the proportional class separability measure as described in the previous section. The second reference method, ABMKSVM, forgoes the use of a class separability measure, but rather optimizes the difference between the combined kernel and ideal kernel directly. This optimization problem can be solved as follows:

$$minimize\ \sum_{m=1}^{P}\sum_{h=1}^{P}\eta_m\eta_h\langle \boldsymbol{K}_m, \boldsymbol{K}_h\rangle_F - 2\sum_{m=1}^{P}\eta_m\langle \boldsymbol{K}_m, \boldsymbol{K}_y\rangle_F, \quad (3\text{-}12)$$

$$with\ respect\ to\ \eta \in \mathbb{R}_+^P \quad subject\ to\ \Sigma_{m=1}^{P}\eta_m = 1.$$

Other methods use the SVM cost term, rather than the distance to an ideal kernel, to update the weights. This can be done by initiating the kernel weights $\boldsymbol{\eta}$ to obtain a single combined kernel and performing the SVM on this kernel. The results of the SVM are then used to update the kernel weights. For example, recognizing the similarity between the MKL formulation and group lasso (Bach, 2007), Xu et al. (Xu *et al.*, 2010) update the kernel weights according to the ℓ_p-norm. For GLMKSVM, we use the ℓ_i-norm, which results in using (13) to update the kernel weights.

$$\eta_m = \frac{\|\mathbf{w}_m\|_2}{\Sigma_{h=1}^{P}\|\mathbf{w}_h\|_2} \quad (3\text{-}13)$$

$$\|\mathbf{w}\|_2^2 = \eta_m^2\Sigma_{i=1}^{n}\Sigma_{j=1}^{n}\alpha_i\alpha_j y_i y_j \boldsymbol{K}_m(\mathbf{x}_i, \mathbf{x}_j) \quad (3\text{-}14)$$

Similarly, SimpleMKL (Rakotomamonjy *et al.*, 2008) uses a gradient decent on the SVM objective value to iteratively update the kernel weights. The combined kernel is initiated as a linear summation where the weight of each kernel is defined as $1/P$. The dual formulation of the MKL SVM is solved (15), and the weights $\boldsymbol{\eta}$ are optimized using the gradient function provided in (16).

$$maximize\ J(n) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \boldsymbol{K}_\eta(\mathbf{x}_i, \mathbf{x}_j) \quad (3\text{-}15)$$

$$\frac{\delta J(\eta)}{\delta \eta_m} = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \boldsymbol{K}_{\boldsymbol{\eta}}(\mathbf{x}_i, \mathbf{x}_j)\ \forall m \quad (3\text{-}16)$$

Varma and Babu (Varma and Babu, 2009) use a similar gradient descent method for updating the weights (17), but perform a nonlinear combination of kernels (18), rather than a weighted summation of kernels, as in SimpleMKL.

$$\frac{\delta J(\eta)}{\delta \eta_m} = \frac{\delta r(\eta)}{\delta \eta_m} - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \frac{\delta \boldsymbol{K}_\eta(\mathbf{x}_i, \mathbf{x}_j)}{\delta \eta_m}\ \forall m \quad (3\text{-}17)$$

$$\boldsymbol{K}_{\eta}^{P}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(\sum_{m=1}^{D} -\eta_m\left(\mathbf{x}_i[m] - \mathbf{x}_j[m]\right)^2\right) \tag{3-18}$$

Here, the regularization function r(·) is defined as $1/2\left(\boldsymbol{\eta} - \frac{1}{P}\right)^T\left(\boldsymbol{\eta} - \frac{1}{P}\right)$. NLMKSVM also presents a non-linear combined kernel, namely the quadratic kernel presented in (19); where the weight optimization is defined as a min-max problem (Cortes, Mohri and Rostamizadeh, 2009) (20) and the weights defined as a ℓ_1-norm bounded set $\mathcal{M}$ (21) with η_0 = 0 and Λ = 1 in the present implementation.

$$\boldsymbol{K}_{\eta}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{m=1}^{P}\sum_{h=1}^{P}\eta_m\eta_h\boldsymbol{K}_m(\mathbf{x}_i^m, \mathbf{x}_j^m)\boldsymbol{K}_h(\mathbf{x}_i^h, \mathbf{x}_j^h) \tag{3-19}$$

$$\min_{\eta\in\mathcal{M}}\max_{\alpha\in\mathbb{R}^n} -\alpha^T\left(\boldsymbol{K}_{\eta} + \lambda\boldsymbol{I}\right)\alpha + 2\mathbf{y}^T\alpha \tag{3-20}$$

$$\mathcal{M} = \{\boldsymbol{\eta}: \boldsymbol{\eta} \in \mathbb{R}_{+}^{P}, \|\boldsymbol{\eta} - \eta_0\|_1 \leq \Lambda\} \tag{3-21}$$

3.3.4 Experimental Set-up

Remote sensing is a valuable tool for providing information regarding the physical status of informal settlements, or slums. Although many studies make use of satellite imagery, even sub-meter imagery may not be sufficient to distinguish between different objects, such as buildings, and to identify their attributes (Kuffer, Pfeffer and Sliuzas, 2016). This motivates the use of UAVs, which are capable of providing images at a higher spatial resolution, thus enabling improved detection and characterization of objects, as well as more detailed elevation information than is available from satellite imagery. The flexible acquisition capabilities also facilitate the acquisition of recurrent imagery to monitor project implementation, especially in the context of slum upgrading projects. These are some of the incentives that motivate the use of UAVs for informal settlement mapping.

The UAV dataset used for the experiments consists of a point-cloud and RGB orthomosaic of an informal settlement in Kigali, Rwanda, which was acquired using a DJI Phantom 2 Vision + quadcopter in 2015. The UAV acquired images with the standard 14 megapixel fish-eye camera (110° FOV) at an approximate 90% forward- and 70% side-lap. The images were processed using the commercial software Pix4D Mapper (Version 2.0.104). The point cloud densification was performed using the 'Low (Fast)' processing option, a matching window size of 7 × 7 pixels and only matching points that are present in at least four images. The DSM was constructed using the Inverse Distance Weighting (IDW) interpolation, with the noise filtering and surface smoothing options enabled. The resulting 8-bit orthomosaic has a spatial resolution of 3 cm. The average density of the utilized point clouds is 1031 points per m^2. This

density depends on the data processing parameters, as well as the characteristics of the land cover type. For this application, the point density ranges between 796 and 1843 points per m^2 according to the land cover (see Table 3.2).

The study area itself is characterized by small, irregular buildings, narrow footpaths and a steep topography. Ten thematic classes are defined for the classification problem: three different types of building roofs (corrugated iron sheets, galvanized iron with a tile pattern and galvanized iron with a trapezoidal pattern), high vegetation, low vegetation, bare surfaces, impervious surface, lamp posts, free-standing walls and clutter. The latter class may consist of, for example, laundry hung out to dry, the accumulation of solid waste on the streets, passing cars and pedestrians. Reference data were defined by visual interpretation and manually labelling pixels in the orthomosaic (based on the results of the over-segmentation and manually adjusting segment boundaries if necessary).

Table 3.2: *Number of labelled pixels and point cloud density for each thematic class.*

Semantic class	**Average Point Cloud Density (Points per m^2)**
Roof Type I (R1)	1843
Roof Type II (R2)	994
Roof Type III (R3)	796
High Vegetation (HV)	1367
Low Vegetation (LV)	951
Bare Surface (BS)	946
Impervious Surface (IS)	970
Walls (W)	1164
Lamp posts (L)	1561
Clutter (C)	1157
Total	1031

Ten sets of training data (n = 2000) were extracted from the Kigali dataset. Five sets followed an equal sampling strategy (n_c = 200), and five sets followed a stratified sampling strategy, which allows an analysis to be made regarding the sensitivity of kernel class separability measures to unequal class sizes. A set of 5000 samples was extracted for testing. The first set of experiments (Experiment I.A. and Experiment I.B.) compared the class separability measures and ideal kernel definitions using the prior knowledge (Case (ii)) feature grouping. The average Overall Accuracy (OA) for each of the folds, along with the standard deviation, is provided for the equal and stratified sampling training sets separately. In Experiment I.A. the class separability measures are used to define the optimal bandwidth parameter for each input RBF kernel $\boldsymbol{K_m}$. Experiment I.B., on the other hand, uses the class separability

measure both to optimize the bandwidth parameter of $\boldsymbol{K_m}$ and to perform the proportional kernel weighting in (7) to obtain the kernel weights $\boldsymbol{\eta}$. In both cases, the search space of the bandwidth parameter was defined by first defining the bandwidth parameter as the mean intra-class ℓ_2-norm and defining a range of γ as 2^{-5}- to 2^5-times this mean bandwidth. For these experiments, three different ideal kernel definitions are compared: assigning values of 1, $1/n_c$ and $1/n_c^2$ to samples belonging to the same class, where n_c represents the number of samples within that specific class.

The second set of experiments analyzed both the influence of feature grouping and MKL methodology. Regarding the feature grouping strategy, the random-, similarity-, diversity- and class-difference-based methods require the user to define the number of desired kernels. For these experiments, six kernels were defined, as this is the number of kernels identified by the automatic feature grouping method. For the novel feature grouping strategy, we use the results of Experiment I to select the best class separability measure (the HSIC). Furthermore, we report the results of using two different cut-off metrics to define how many features to include in each kernel: we report the results when defining a maximum of 45 features per kernel (HSIC-f_{45}) and when using the 99.9% cumulative relevance cut-off per kernel (HSIC-99.9%). These thresholds were selected based on the results of the feature ranking (e.g., Figure 3.3b). The minimum (a) and maximum (b) number of features per group using the diverse kernel strategy were set to a = 5 and b = 70 based on experimental analyses.

MKL was performed on these feature kernels using the seven algorithms described above. Note that the grouping strategy was applied separately for each fold, so the feature groups will not by definition be the same for each training set. Once the groups were identified, the same feature kernels were used as input for each MKL method.

The methods were again compared by computing the mean overall accuracy over each of the 10 folds with reference to the same 5000 sample test set. The error matrix of the CSMKSVM method using the HSIC-f_{45} feature grouping strategy is also presented, as well as the correctness (22) and completeness (23) for each of the 10 thematic classes.

$$Correctness = TP/(TP + FP) \tag{3-22}$$

$$Completeness = TP/(TP + FN) \tag{3-23}$$

where TP indicates the number of true positives per class, FP is the number of false positives and FN is the number of false negatives.

Furthermore, the MKL methods were compared to two baseline classifiers: a standard SVM classifier, implemented in LibSVM (Chang and Lin, 2011), where all features are combined in a single kernel, and a random forest classifier. For the SVM, RBF bandwidth parameter γ was defined as described previously, and the regularization parameter *C* was optimized through a 5-fold cross-validation between 2^{-5} and 2^{15}. Regarding the random forest classifier, the number of trees was optimized between 100 and 1500 in steps of 100.

In a final step, we provide classification maps of 30 × 30 m subsets of the Kigali dataset. Similar to the other experiments, 2000 labelled pixels were extracted from ten tiles representing the different characteristics of the study area through stratified sampling. These pixels were used to construct a single-kernel SVM, and CSMKSVM using the HSIC-f_{45} feature grouping strategy. Classification maps of three of the tiles are provided to illustrate the results.

3.4 Results and Discussion

3.4.1 Class Separability Measures and Ideal Kernel Definition

Kernel class separability measures require the definition of a target kernel. If each class has a similar number of samples, i.e., equal sampling, then the value assigned to the ideal kernel for samples adhering to the same class ($y_i = y_j$) does not have to be adjusted for the class size (Tables 3.3 and 3.4). However, this is not the case when there are large differences in the number of training samples per class, which may occur in stratified sampling, which is common to the processing of remotely-sensed images. The class separability measure used to define kernel weights will target the most common class, and therefore, results can be improved when the value is normalized by the number of samples per class (Table 3.3).

When the class separability measure is used both to define the bandwidth of the RBF kernel, as well as to define the relative kernel weights, this effect is minimized, and simply assigning a value of '1' to samples of the same class appears to be adequate (Table 3.4). Regarding the comparison between the various class separability measures, the HSIC outperformed KA, CKA and KCS through both a higher and more stable OA for the stratified samples, although KA performed slightly better in the case of equal sampling. Due to these observations, the subsequent analyses were carried out using the HSIC class separability measure and an ideal kernel where samples adhering to the same class are assigned a value of '1' and '0' otherwise, which is used both to optimize the kernel parameters and for the proportional kernel weighting. The HSIC is therefore also used as the kernel-based class separability measure for the proposed feature grouping strategy in the next experiments.

Table 3.3: *The overall accuracy obtained for Experiment I.A.: optimizing the bandwidth parameters γ_m for each input kernel $\boldsymbol{K_m}$ using various kernel class separability measures and ideal kernel definitions. n_c indicates the number of samples for a specified class. CKA, Centered-Kernel Alignment; KCS, Kernel Class Separability.*

Value $y_i = y_j$	HSIC	KA	CKA	KCS
Equal sampling (5 folds)				
1	**89.5 ± 0.46**	89.3 ± 0.43	89.1 ± 0.66	88.8 ± 1.13
$1/n_c$	**89.5 ± 0.46**	89.3 ± 0.43	89.1 ± 0.66	88.8 ± 1.13
$1/n_c^2$	**89.5 ± 0.46**	89.3 ± 0.43	89.1 ± 0.66	88.8 ± 1.13
Stratified sampling (5 folds)				
1	86.6 ± 0.59	86.8 ± 0.53	86.9 ± 0.45	86.8 ± 0.53
$1/n_c$	86.9 ± 0.38	86.8 ± 0.54	86.7 ± 0.42	86.8 ± 0.53
$1/n_c^2$	**87.2 ± 0.45**	86.9 ± 0.54	87.1 ± 0.44	86.8 ± 0.53

Table 3.4: *The overall accuracy obtained for Experiment I.B.: optimizing both the bandwidth parameters γ_m for each input kernel $\boldsymbol{K_m}$ and the relative kernel weights $\boldsymbol{\eta}$ using various kernel class separability measures and ideal kernel definitions. n_c indicates the number of samples for a specified class.*

Value $y_i = y_j$	HSIC	KA	CKA	KCS
Equal sampling (5 folds)				
1	90.3 ± 0.40	**90.6 ± 0.53**	89.2 ± 0.81	86.7 ± 0.67
$1/n_c$	90.3 ± 0.41	**90.6 ± 0.53**	89.2 ± 0.81	86.7 ± 0.67
$1/n_c^2$	90.3 ± 0.41	**90.6 ± 0.53**	89.2 ± 0.81	86.7 ± 0.67
Stratified sampling (5 folds)				
1	**87.2 ± 0.38**	82.7 ± 1.06	86.1 ± 0.48	80.8 ± 0.71
$1/n_c$	87.0 ± 0.81	84.1 ± 0.75	85.3 ± 0.52	80.8 ± 0.71
$1/n_c^2$	87.0 ± 0.57	86.0 ± 0.71	84.2 ± 0.86	80.8 ± 0.71

3.4.2 Comparison of Feature Grouping and Kernel Weighting Strategies

The first observation from the results of the various feature grouping and kernel weighting strategies (Table 3.5) is that almost all MKL methods perform better than a standard SVM where all features are described by a single kernel, which achieves an accuracy of 85.4%. Only some of the rule-based mean kernel weighting (ABMKSVM) classification results and GLMKSVM using individual features perform worse than the standard SVM. Furthermore, we see that most of the MKL implementations perform better than the random forest classifier, which has an average accuracy of 86.5%. A McNemar test with the continuity correction (Foody, 2004) indicates that the improved classification accuracy of the HSIC-f_{45} CSMKSVM method is significant compared to both

the results of the single-kernel SVM (*p*-value of 0.0021) and random forest classification (*p*-value of 0.0032).

Regarding feature grouping strategies, the HSIC grouping strategy proposed in this paper obtains the highest accuracy for most MKL algorithms, where all methods except ABMKSVM achieve an accuracy above 90%. Furthermore, the results suggest that the high accuracy is more stable than other grouping methods. Both stopping criteria (either by selecting a fixed number of features per kernel or thresholding the cumulative cost function) have a similar performance. The OA is also quite robust to the cut-off metrics. Selecting 45 features (HSIC-f_{45}) obtains the highest accuracy of 90.6% when combined with the CSMKSVM method, though using 30 or 60 features only lowered this accuracy by 0.1% and 0.2%, respectively. Similarly, the 90.5% accuracy achieved with HSIC-99.9% was only 0.2% higher than the accuracy obtained by HSIC-99.7%. Further analysis of the two methods indicated that a feature selection was indeed performed. HSIC-f_{45} selected an average of 78 features out of the 107 per fold, and HSIC-99.9% selected an average of 70 features per fold. This could advocate thresholding the cumulative feature relevance rather than fixing the number of features per kernel, as it is more suited in automatic workflows and uses a lower number of features while achieving a similar accuracy. The grouping strategy utilizing prior knowledge (i.e., feature provenance) also performs well for the CSMKSVM and NLMKSVM methods. The individual kernel grouping strategy works well for the NLMKSVM method. This is not entirely surprising, as NLMKSVM is a nonlinear kernel combination method and may therefore mimic the nonlinear similarity, which is achieved when various features are grouped into an input kernel through a non-linear mapping function.

Table 3.5: *The overall accuracy obtained for Experiment I.A.: optimizing the bandwidth parameters γ_m for each input kernel $\boldsymbol{K_m}$ using various kernel class separability measures and ideal kernel definitions. n_c indicates the number of samples for a specified class. CKA, Centered-Kernel Alignment; KCS, Kernel Class Separability.*

Feature Grouping Strategy	***p***	**MKL Method [1]**						
		AB	**CS**	**GL**	**G**	**NL**	**RB**	**S**
Reference MKL feature grouping strategies								
Individual	107	81.2±3.1	89.5±1.4	80.7±2.8	89.6±1.4	**92.0**±1.0	89.3±1.5	89.6±1.4
Prior knowledge	4	88.2±2.9	**90.2**±1.4	89.9±1.7	89.9±1.7	**90.7**±1.5	89.8±1.6	89.9±1.7
Random	6	83.6±2.6	87.6±2.0	87.3±2.4	87.3±2.5	87.5±2.0	87.3±2.2	87.3±2.5
Similarity	6	77.8±6.0	87.3±1.9	87.5±1.9	87.6±1.9	86.9±2.2	86.6±2.1	87.5±1.9
Diversity	6	82.2±4.0	86.9±2.0	86.9±1.9	86.9±1.9	87.1±2.0	86.9±1.9	86.9±1.9
Between-class sample distance	6	83.7±2.8	88.0±1.7	87.9±1.9	87.9±1.8	88.5±1.9	87.8±1.8	87.9±1.8
Within-class sample distance	6	81.5±6.9	87.9±2.2	87.5±2.3	87.5±2.3	88.1±2.1	87.5±2.2	87.4±2.3
Between + within-class distance	6	81.6±4.0	87.6±2.0	87.5±1.8	87.5±1.9	88.2±2.1	87.4±2.0	87.5±1.9
Proposed MKL grouping strategy								
HSIC-f_{30}	6	89.1±2.0	**90.5**±1.5	**90.3**±1.6	**90.3**±1.6	**90.3**±1.6	**90.5**±1.5	**90.3**±1.6
HSIC-f_{45}	6	89.5±1.6	**90.6**±1.5	**90.5**±1.6	**90.4**±1.7	**90.2**±1.7	**90.5**±1.5	**90.4**±1.7
HSIC-f_{60}	6	89.6±1.7	**90.4**±1.7	**90.4**±1.7	**90.4**±1.7	**90.1**±1.7	**90.4**±1.6	**90.4**±1.7
HSIC-99.7%	6	88.8±1.8	**90.3**±1.6	**90.0**±1.8	**90.1**±1.7	**90.1**±1.7	**90.1**±1.7	**90.2**±1.6
HSIC-99.9%	6	89.3±1.8	**90.5**±1.5	**90.3**±1.7	**90.3**±1.7	**90.3**±1.7	**90.5**±1.6	**90.3**±1.7
Reference classification strategies								
Single-kernel SVM	1	85.4±2.3						
Random forest	-	86.5±1.9						

[1] *AB = ABMKSVM, CS = CSMKSVM, GL = GLMKSVM, G = GMKSVM, NL = NLMKSVM, RB = RBMKSVM, S = SimpleMKL.*

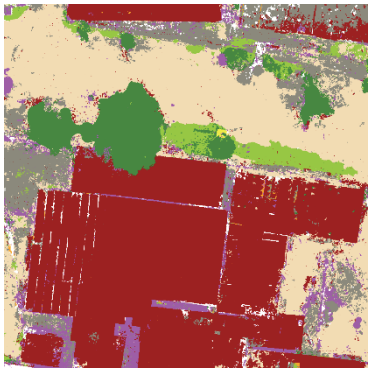
Similar to the results of previous studies (e.g., (Gönen and Alpaydın, 2011)), we observe a similar performance between the results obtained by the various MKL algorithms to combine the kernels. In this case, simply taking the mean of the input kernels (ABMKSVM) consistently performs worse than the other MKL algorithms. However, there is not one single algorithm that consistently outperforms the others. For the prior knowledge-based feature grouping strategy, the best OAs are achieved by the proposed proportional HSIC-weighting measure (CSMKSVM) with 90.2% and the nonlinear NLMKSVM method with 90.7%. Regarding the proposed feature grouping strategy, CSMKSVM also obtains slightly better results than the other MKL methods at 90.6% when utilizing 45 features per kernel.

The ability of the selected features to distinguish between the different land cover classes will also depend on the study area. For example, vegetation will be more difficult to distinguish for study areas in which it is not always green (for example, the leaf-off season in temperate climates, ripening agricultural crops or arid climates). It is possible that some of the 3D features will capture the geometric traits of vegetation in these situations. However, the extent to which this is possible will greatly depend on the characteristics of the study area. Furthermore, the UAV flight parameters and data processing options will influence the suitability of the 3D features. Low texture, pixel saturation or mismatch in the scale of objects in the UAV images may cause artefacts in the point cloud, such as irregular elevation values. This, in turn, influences the quality of the 3D features, such as planar segments. Similarly, the (textural) characteristics of different land cover types will influence the point cloud density and affect the suitability of 3D features. For a more detailed analysis of the interaction between different features from UAV images and why these features were selected, the reader is referred to (Gevaert *et al.*, 2017).

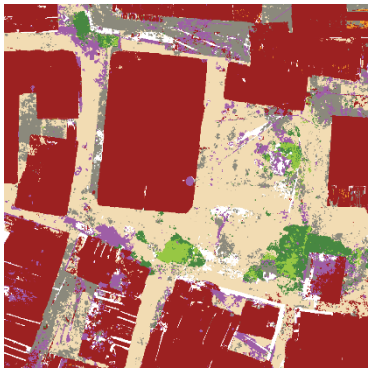
A visual analysis of the results is presented in Figure 3.4, which compares the classification maps of a standard SVM and the proposed HSIC-f_{45} CSMKSVM method. The results indicate the standard single-kernel SVM is noisier than the MKL methods. Although MKL performs better than the standard SVM method, there are still difficulties in distinguishing between bare versus impervious surfaces, surfaces versus clutter, building roofs versus clutter and building roofs versus walls (Table 3.6).

(a)

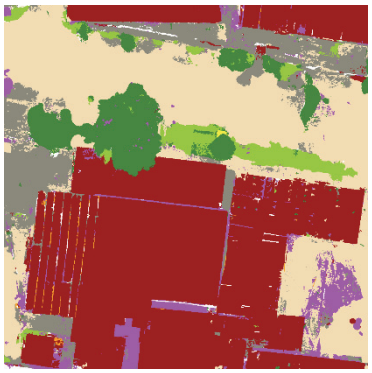
(b)



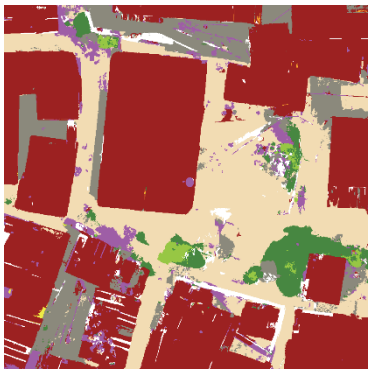
(c)



(d)



(e)



(f)

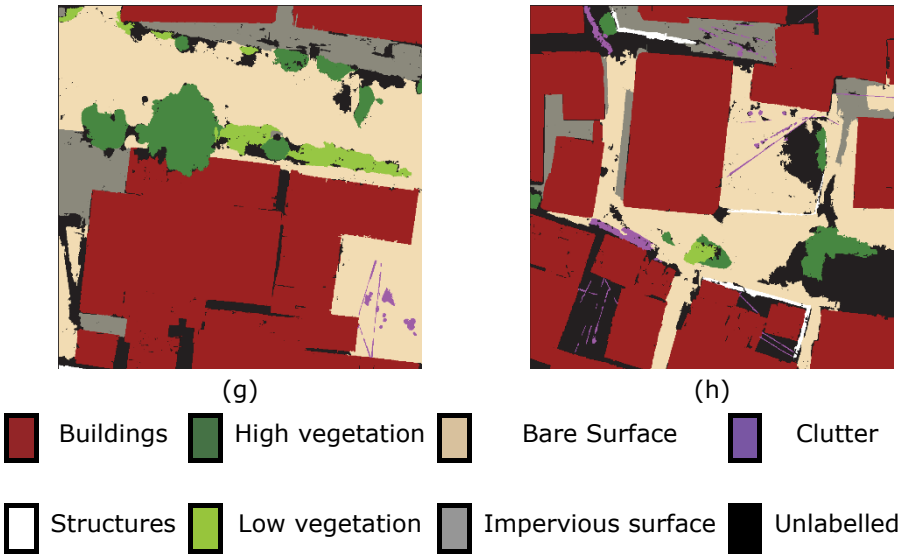


Figure 3.4: *Sample classification results of two image tiles, with the input RGB tile in the first row (a,b); followed by the classification results using a standard single-kernel SVM (c,d); the classification results using the proposed CSMKSVM measure and the HSIC-f_{45} feature grouping strategy (e,f); and the reference classification data (g,h).*

Table 3.6: *Error matrix of the HSIC-f_{45} CSMKSVM method; numbers indicate the total number of pixels over the 10 folds. The final column provides the completeness (Comp.) of each class, and the final row provides the correctness (Corr.). R1, R2 and R3 correspond to 3 types of roof materials; HV = high vegetation, LV = low vegetation, BS = bare surface, IS = impervious surface, W = wall structures, L = lamp posts, C = clutter.*

		Predicted Class Label										
		R1	**R2**	**R3**	**HV**	**LV**	**BS**	**IS**	**W**	**L**	**C**	**Comp. (%)**
Reference class label	**R1**	*2048*	55	0	0	0	0	0	3	0	24	96.2
	R2	215	*19968*	7	29	3	293	299	321	0	405	92.7
	R3	0	21	*1759*	9	0	5	10	9	0	7	96.7
	HV	2	10	0	*3351*	192	16	2	17	0	30	92.6
	LV	2	8	0	67	*1595*	7	0	0	0	1	94.9
	BS	0	69	10	66	19	*7346*	403	144	1	182	89.2
	IS	9	92	26	14	17	453	*5738*	228	15	198	84.5
	W	1	27	1	11	0	69	32	*1086*	0	73	83.5
	L	0	0	0	0	0	0	0	0	*1020*	0	100
	C	16	69	0	24	6	139	105	129	1	*1371*	73.7
Corr. (%)		89.3	98.3	97.6	93.8	87.1	88.2	87.1	56.1	98.4	59.8	OA = 90.6

3.5 Conclusions

In this paper, we demonstrate the suitability of MKL as a classification method for integrating heterogeneous features obtained from UAV data. Utilizing a novel feature grouping strategy and a simple heuristic for weighting the individual input kernels (CSMKSVM), we are able to obtain a classification accuracy of 90.6%, an increase of 5.2% over a standard SVM implementation and 4.1% over a random forest classification model. These improvements are statistically significant with a *p*-value <0.005, which indicates strong evidence as standard tests use confidence levels of 0.05 or 0.01 to indicate significant differences. A series of experiments reinforces observations by other researchers that complex kernel weighting strategies do not seem to perform significantly better than simple heuristics, such as a proportional weighting based on the HSIC class separability measure.

Furthermore, we observe that much of the literature on MKL classification has focused on ways to weigh the kernels, but not how to group the features appropriately. Experiments demonstrate the importance of the latter to effectively apply MKL. In this application, satisfactory results are obtained when grouping features based on their provenance (i.e., radiometric, texture or 3D features). A novel, automated grouping strategy is also proposed, which consistently obtains high classification accuracies for all seven MKL methods that were tested here. Furthermore, for most MKL methods, the proposed feature grouping strategy performed better than when using individual kernels for each feature. This underlines the importance of proper feature grouping, which not only produces a high and stable overall accuracy, but also reduces the number of input kernels for the MKL and, thus, reduces the computational complexity. These observations support a deeper understanding of MKL for classification tasks. Future applications of classification tasks with heterogeneous features are recommended to start by grouping features according to the proposed automated method and to use CSMKSVM to weight the input kernels for the SVM classification. Finally, this manuscript demonstrates that features extracted from point clouds and orthoimagery derived from UAVs are suitable for land cover classification. Additional research would be needed to analyze to what degree the features are sensitive to the type of UAV, flight parameters and algorithms utilized to produce the point clouds and orthoimagery.

Chapter 4 – Context-based Filtering of Noisy Labels for Automatic Basemap Updating from UAV Data[3]

[3] This chapter is based on:

Gevaert, C.M., Persello, C., Oude Elberink, S., Vosselman, G., and Sliuzas, R. (2017) 'Context-Based Filtering of Noisy Labels for Automatic Basemap Updating From UAV Data', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pp. 1-11. doi: 10.1109/JSTARS.2017.2762905.

Abstract

Unmanned Aerial Vehicles (UAVs) have the potential to obtain high-resolution aerial imagery at frequent intervals, making them a valuable tool for urban planners who require up-to-date basemaps. Supervised classification methods can be exploited to translate the UAV data into such basemaps. However, these methods require labeled training samples, the collection of which may be complex and time consuming. Existing spatial datasets can be exploited to provide the training labels, but these often contain errors due to differences in the date or resolution of the dataset from which these outdated labels were obtained. In this paper we propose an approach for updating basemaps using global and local contextual cues to automatically remove unreliable samples from the training set and thereby improve the classification accuracy. Using UAV datasets over Kigali, Rwanda and Dar es Salaam, Tanzania, we demonstrate how the amount of mislabeled training samples can be reduced by 44.1% and 35.5% respectively, leading to a classification accuracy of 92.1% in Kigali and 91.3% in Dar es Salaam. To achieve the same accuracy in Dar es Salaam, between 50000 and 60000 manually labeled image segments would be needed. This demonstrates that the proposed approach of using outdated spatial data to provide labels and iteratively removing unreliable samples is a viable method for obtaining high classification accuracies while reducing the costly step of acquiring labeled training samples.

4.1 Introduction

The utilization of geospatial information to support urban planning is becoming common practice worldwide. A fundamental building block is the basemap (or topographic map), which provides information regarding the location of elemental objects of the urban fabric. As a foundation for many urban planning activities, it is imperative that this basemap provides accurate and up-to-date information. This is not always the case, as changes in an urban setting may occur more rapidly than the updating of such basemaps, which traditionally occurs through the manual digitization of satellite or airborne imagery. This is particularly relevant for informal settlements, which tend to be very dynamic urban environments.

Recent technological developments regarding data acquisition platforms such as Unmanned Aerial Vehicles (UAVs), also known as Remotely Piloted Aerial Systems (RPAS), display potential for quickly delivering high-quality spatial data for geomatics applications (Nex and Remondino, 2014). UAVs are capable of bringing imagery with a spatial resolution of mere centimeters and accurate 3D information to urban planners at a low cost. However, the area which can be covered by a single flight is currently limited due to the technical characteristics of the type of UAVs commonly used for mapping activities (Nex and Remondino, 2014) and national legislation often limits the maximum area that can be covered by a single flight (Stöcker *et al.*, 2017). UAVs are therefore especially suited for mapping tasks which require multiple acquisitions over a limited study area, such as incremental map updating.

In order to exploit the information contained in remotely sensed imagery, the images are usually translated into vector-based semantic information such as the basemaps mentioned previously. In many situations, basemap updating through UAV imagery is performed through manual digitization of new features (Koeva *et al.*, 2016), possibly even involving stakeholders through a participatory mapping approach (Ramani Huria, 2016). An alternative strategy to digitization is the extraction of semantic information through supervised image classification methods.

Supervised classification methods make use of representative training samples to characterize the common characteristics and variability of objects pertaining to each semantic class. Based on the observed distributions of the samples in a defined feature space, a classification model is constructed. This model allows labels to be assigned to new, unlabeled samples. Supervised classification algorithms have been successfully applied in a number of studies to extract semantic information from UAV imagery of urban scenes. Some studies divide the orthoimagery into a grid of coarser resolution, and use a pre-trained library to propose which urban objects may be present within each grid cell

(Moranduzzo *et al.*, 2015). Other studies include 3D features from the point cloud or Digital Surface Model (DSM) to support the image-based features, e.g. (Chen *et al.*, 2016; Gevaert *et al.*, 2017). The application of morphological filters of adaptive sizes on the DSM result in useful features for identifying urban objects with differing scales and can complement the information contained in the imagery (Zhang *et al.*, 2015).

One of the main difficulties in classifying sub-decimeter resolution imagery obtained through UAV platforms is the spectral and spatial variability of urban objects (e.g. (Moranduzzo *et al.*, 2015; Zhang *et al.*, 2015)). It is possible to address the spectral variability by clustering all the pixels in an unsupervised manner and using a majority voting of the reference labels per cluster to label the pixels (Senthilnath *et al.*, 2017). Unfortunately, the collection of these reference labels is expensive, time consuming, and requires a relatively high level of knowledge to ensure that they are representative of the class distributions.

Rather than manually labelling training samples, it is also feasible to use existing spatial datasets to provide the labels. For example, vector data from existing basemaps or sources such as OpenStreetMap could be used (Mnih and Hinton, 2012; Chen and Zipf, 2017). However, there are likely to be changes in the scene if there is a time lapse between the collection of the vector data at t_0 and the newly acquired UAV imagery at t_1. Furthermore, the existing vector information may have been digitized over imagery of a lower spatial resolution, causing misalignments when superimposed over the UAV imagery. Therefore, if existing spatial data is utilized to provide the training samples, it must be taken into account that a number of the training labels are likely to be incorrect.

Various strategies have been developed to deal with such errors in the training sample labels, i.e. *label noise*. A recent overview of the effect of label noise on classification algorithms (Frenay and Verleysen, 2014) observed that there are three main strategies to address this issue: (i) utilizing noise-robust classification algorithms, (ii) data cleansing to remove potentially noisy labels from the training data, and (iii) explicitly modelling label noise. Other strategies have been developed to specifically combat label noise for remote sensing applications. For example, by modelling label noise by combining noise robust logistic regression and Conditional Random Fields (CRFs) for updating geospatial databases (Maas, Rottensteiner and Heipke, 2016); or by using the contextual information in a semi-supervised setting in order to assess the reliability of training samples and obtain a classification algorithm that is more robust to mislabeled training samples (Bruzzone and Persello, 2009).

In this paper, we utilize a data cleansing strategy which exploits context to identify samples which are likely to have an incorrect label. Research from the fields of computer vision (Galleguillos and Belongie, 2010) and the human visual system (ten Oever *et al.*, 2016) indicate that both global and local contextual cues are important for object recognition. Global context has to do with the statistics of the image as a whole. The underlying idea is that similar but disjoint objects in a single study area will have similar variations. So, by using existing labels over the entire scene, the classifier uses global statistics of the study area to identify the common variations of these objects in a feature space. Objects which are mislabeled are likely to fall outside of these common variations, causing the classifier to be uncertain about the output label. Local context has to do with the similarity of neighboring samples. Generally speaking, neighboring pixels or segments which have similar characteristics could be expected to belong to the same semantic class (Schindler, 2012). This forms the basis of the contrast-sensitive Potts model which is commonly used in image analysis techniques such as CRFs to ensure a smooth labeling (e.g. (Shotton *et al.*, 2009)).

Object- or segment-based labeling, as opposed to pixel-based labelling, gives a single label to a group of pixels. Such a contiguous group of pixels with similar characteristics, is known as an *image segment*. This technique forms the basis of Object Based Image Analysis (OBIA) (Blaschke, 2010), and could also be interpreted as a way to ensure smooth labels (Schindler, 2012). It has been advocated that OBIA is especially suitable for remote sensing applications where the object of interest is larger than the spatial resolution of the image (Blaschke, 2010). It has been one of the most common techniques for slum identification from high resolution satellite imagery, though parameter tuning is important to avoid over- or under-segmentation (Kuffer, Pfeffer and Sliuzas, 2016). In light of the difficulty of tuning image segmentation parameters, super-pixels could be used (Achanta *et al.*, 2012). These are in essence an over-segmentation of the image. Super-pixel based image analysis lowers the data redundancy and can speed up classification tasks compared to pixel-based strategies, while avoiding errors due to undetected object boundaries in cases of under-segmentation.

The main motivation behind this work is to combine both the *global contextual uncertainty* (i.e. class representation and object variability within the scene) with the *local contextual consistency* (i.e. similar neighbors having similar classes) to automatically identify and remove noisy labels from training data. By iteratively training supervised classifiers and removing potentially mislabeled samples after each iteration, the training sample set is iteratively cleaned and the accuracy of the classification model is improved. This allows existing vector outlines to be exploited as labels for newly-acquired UAV imagery, thereby reducing the need for the collection of costly training samples

and speeding up the basemap updating workflow. The adopted methodology combines various aspects of the state-of-the-art in remote sensing of urban areas and image processing, such as Object-Based Image Analysis, the integration of 2D and 3D features, and the inclusion of contextual information to improve classification accuracies.

The proposed technique is demonstrated through two applications. The first uses recently acquired UAV imagery and outdated building outlines of an informal settlement in Kigali, Rwanda. The second application demonstrates how the same method can be applied to improve the accuracy of crowd-sourced data. More specifically, to verify the building outlines of an informal settlement in Dar es Salaam, Tanzania which were digitized by community members using OpenStreetMap. Various experimental setups demonstrate the necessity of using both global and local contextual cues, the sensitivity of the proposed method to the proportion of training labels which are incorrect, and approximate the number of training samples which would need to be manually labelled in order to obtain the same classification accuracy as the automated workflow.

4.2 Proposed Method

The proposed method takes image segments with descriptive features from the newly acquired dataset and an initial class label obtained from the outdated basemap data as input. Then, it applies three steps to identify samples with unreliable labels and remove these from the training set. These three steps (steps 4 to 6 in Figure 4.1) are: pre-filtering the image segments based on the uniformity of noisy labels acquired at t_0 for each segment from the images at t_1, performing a supervised classification, and finally removing unreliable training samples. Here, uniformity refers to the percentage of pixels in an image segment which are assigned the label of the most prominent class within that segment. Label reliability is based on the *label consistency*, *local contextual consistency*, and *global contextual uncertainty*. The last two steps (classification and removing unreliable training samples) are repeated iteratively to improve the classification model. This improved classification model can then be used to assign a class label to each image segment and obtain a classified map.

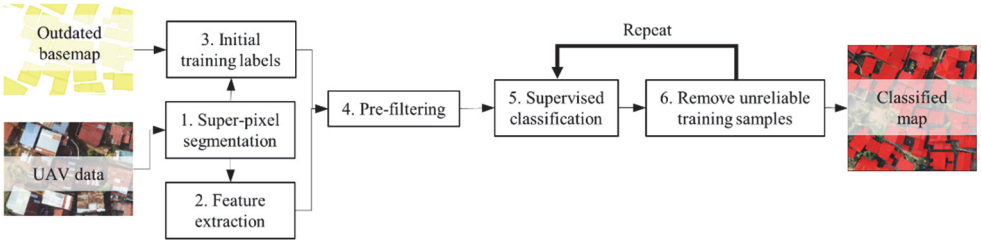


Figure 4.1: *Workflow of the proposed method for automatically identifying unreliable labels when using existing spatial data to provide training labels for the classification of UAV data.*

In the following section, we explain how the proposed method works. We use the notation $S = \{s_1, s_2, \dots s_n\}$ for the n segments in the image acquired at t_1 and $R \subseteq S$ to the set of segments which are used to train the classifier. Each image segment s_i has an area A_i, a feature vector $\mathbf{x}_i$ obtained from the image dataset at t_1, and a class label c_i^k where k refers to the iteration. For example, c_i^0 indicates the class label of s_i according noisy labels acquired at t_0, and c_i^1 refers to the label assigned to s_i after the first iteration of the algorithm. E_i refers to the set of all image segments adjacent to s_i and $l_{i,j}$ refers to the length of the shared border between s_i and $s_j \in E_i$. Pseudocode for the algorithm is provided in Algorithm 1. Please note that this section describes the general workflow of the proposed method whereas the exact implementation employed for our UAV datasets (including image segmentation and feature extraction) is described in Section III B. Experimental Analysis.

Algorithm 1: *iterRF-LG*

Inputs: $S = \{s_1, s_2, \dots, s_n\}$ image segments with features from t_1, $R \subseteq S$ subset of segments which are used as training samples, $C^k = \{c_1^k, c_2^k, \dots, c_n^k\}$ segment labels at iteration k, user-defined number of iterations k_{max}, local contextual consistency threshold $\psi_{\min}$, global contextual uncertainty threshold θ_{min}.

Procedure:

1. Set $R = S$ and initialize C^0 with noisy training labels from t_0
2. If uniformity(s_i) < minimum uniformity criterion, remove s_i from R

For $\boldsymbol{k = 1 : k_{max}}$

3. Train a random forest classifier using segments in R and labels from C^{k-1}
4. Apply the classification model to S and update C^k.
5. If $c_i^k \neq c_i^{k-1}$ OR $\psi_i < \psi_{\min}$ OR $\theta_i < \theta_{min}$, remove s_i from R

Outputs: improved segment labels $C^{k_{max}}$

The reasoning behind the pre-filtering step (i.e. step 4 in Figure 4.1) is that the building outlines at t_0 may not always align with the image segments obtained from the imagery at t_1, causing these image segments to contain

conflicting labels. Therefore, as an initial simple filtering mechanism, only "pure" segments where the percentage of labels from a single class meets a user-defined threshold are selected for the training set R used to develop the classification model. The segments which do not meet this purity criterion are only incorporated at the end of the workflow, when the final classification model is used to classify the entire image and obtain the final classification map.

For the supervised classification step, we propose to use random forests (Breiman, 2001) as they have been demonstrated to be more robust to label noise compared to other classification methods (Folleco *et al.*, 2008; Frenay and Verleysen, 2014) and they can easily deal with large numbers of training samples, which is useful as all the segments are labeled in this application of map updating. Furthermore, it is intuitive to derive a confidence measure for the prediction, which is needed for the global contextual uncertainty criterion.

Then, the *label consistency*, *local contextual consistency*, and *global contextual uncertainty* are used to remove unreliable training samples. *Label consistency* implies that the label of a training sample is consistent with the label assigned at the previous iteration, i.e. $c_i^k = c_i^{k-1}$. For example, if one segment represents a building at t_1, it may be non-building according to the outdated basemap labels at t_0. However, as the features of the segment are likely to be similar to other buildings in the area, it could feasibly be classified as building in the second iteration. Therefore, segments where a label is inconsistent causes it to be removed from the training set. This strategy has been previously employed for data cleansing techniques (Thongkam *et al.*, 2008; Jeatrakul, Wong and Fung, 2010), but may be dangerous when used on its own as it may also remove potentially informative samples (Matic *et al.*, 1992; Guyon, Matic and Vapnik, 1996).

The underlying idea of the second criterion, *local contextual consistency*, is that if there are misalignments in the object boundaries at t_0 and at t_1, then the correctly labeled parts of the object may be used to identify neighboring mislabeled segments (see the example in Figure 4.2a). This is implemented by comparing the labels of neighboring pixels or image segments, and introducing a penalty for neighbors which have different labels but similar feature vectors. We exploit the idea of edge potentials commonly adopted in CRFs, and define our contextual consistency criterion using a contrast-sensitive Potts model (Boykov and Jolly, 2001):

$$\phi\left(c_i^k, c_j^k, \mathbf{x}_\mathrm{i}, \mathbf{x}_\mathrm{j}\right) = \begin{cases} \exp[-\beta\left\|\mathbf{x}_\mathrm{i} - \mathbf{x}_\mathrm{j}\right\|^2] & if\ c_i^k \neq c_j^k \\ 1 & if\ c_i^k = c_j^k \end{cases}, \tag{4-1}$$

where c_i^k and c_j^k indicate the class labels and $\mathbf{x}_i$ and $\mathbf{x}_j$ the feature vectors of two neighboring image segments s_i and s_j. This assigns a value of 1 to edges between neighbors of the same class, and $\exp[-\beta\|\mathbf{x}_i - \mathbf{x}_j\|^2]$ to edges between neighbors adhering to different classes, where β equals the average square gradient between all neighboring segments as in (Shotton *et al.*, 2009). The local contextual consistency index of segment s_i (ψ_i) is the weighted sum of (1) for all neighboring segments:

$$\psi_i = \sum_{(i,j)\in \mathrm{E}_i} w_{i,j} \cdot \phi(c_i^k, c_j^k, \mathbf{x}_i, \mathbf{x}_j), \tag{4-2}$$

where $w_{i,j}$ is the relative weight of the neighboring segment s_j. The relative weight ($w_{i,j}$) of each neighbors' edge potential is normalized by border length and relative size of the neighbors (Gould, Fulton and Koller, 2009) as follows:

$$w_{i,j} = \frac{l_{i,j} \cdot A_j}{\sum_{(i,j)\in \mathrm{E}_i} l_{i,j} \cdot A_j}. \tag{4-3}$$

This increases the influence of neighboring segments which share a longer border and larger neighboring segments, as larger segments are presumed to provide more stable feature values. Note that ψ_i penalizes similar segments with different labels, but it doesn't indicate which of the two neighbors is likely to be correct and which is likely to have the noisy label. Therefore, we only remove the samples for which both the local contextual consistency and the classifier uncertainty fall below a defined threshold. Furthermore, it is important to note that (1) only looks at neighboring segments which have different labels. This is useful for updating building outlines, but not in detecting new objects which are isolated. For example, if a building appears at t_1 in the middle of an area which was entirely labelled as non-building at t_0, ψ_i will not be suitable for identifying mislabeled training samples (i.e. Figure 4.2e).

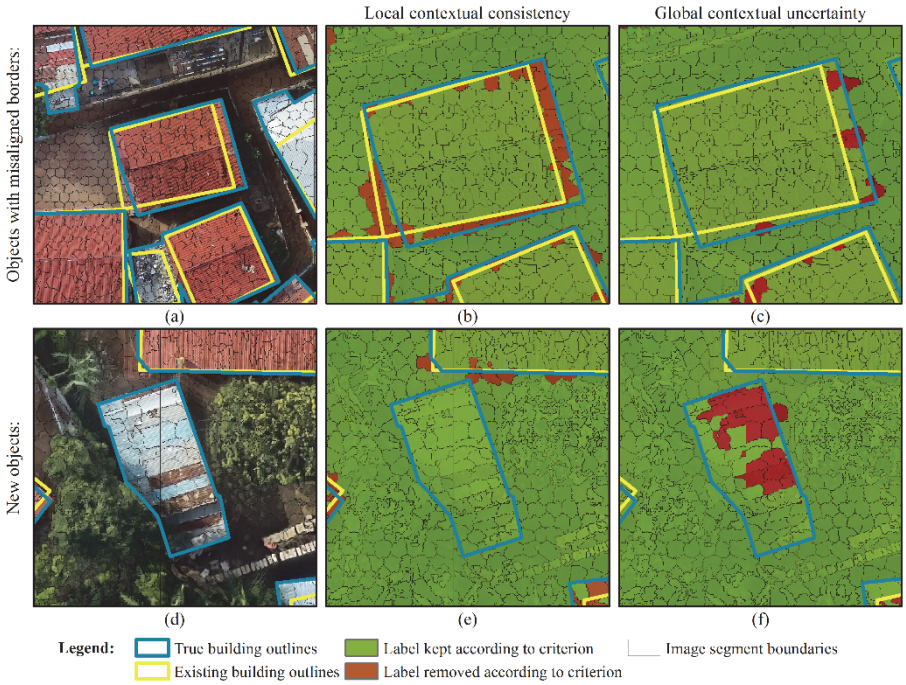


Figure 4.2: *Illustrative examples from the Kigali dataset showing the interplay between the local contextual consistency (b,e) and global contextual uncertainty criteria (c,f). The local contextual consistency is especially useful for updating object boundaries (a-c), whereas the global contextual uncertainty is required to capture new objects (d-f).*

In such situations, mislabeled training samples can be identified by taking into account the *global contextual uncertainty* of segment s_i (θ_i). That is to say that the classification model describes the statistical attributes of the objects outlined by the outdated vector data at t_0 in the feature space derived from the imagery at t_1. If a sample is mislabeled by the provided labels, then it is likely to lie closer to its true label in the feature space and may therefore cause the classifier to be uncertain about the assignation of the label. If a group of neighboring segments consistently have a high uncertainty according to the classification model, they are likely to be mislabeled. Therefore, we calculate the weighted average of the classifier uncertainty over all neighboring segments:

$$\theta_i = \sum_{(i,j)\in \mathrm{E}_j} w_{i,j} \cdot u_j^k \,, \tag{4-4}$$

where $w_{i,j}$ is again the relative weight between neighboring segments s_i and s_j as defined in (2) (i.e. the same weights are used for both the local contextual consistency and global contextual uncertainty) and u_j^k is the classifier uncertainty for segment s_j at iteration k. Note that although both ψ_i and θ_i

change at each iteration, we omit the superscript *k* in order to simplify the notation.

We propose utilize random forests for the supervised classification task. Random forests consist of a group of classification trees, where each tree is trained using a random subset of the training samples and each decision node depends on a random subset of the features (Breiman, 2001). In the testing stage, a sample passes through each tree and each tree casts a vote for the label of the most prominent class of training samples which ended up at that leaf. The final label of the sample is determined through a majority voting of the results of each tree in the forest. The uncertainty can easily be calculated as the fraction of reference samples of the leaf which have the most prominent label multiplied for each tree in the forest. For binary classification problems, u_j^k ranges from 0.5 to 1 with higher values representing more confident predictions.

In the proposed method, the steps of supervised classification and removing the unreliable training samples using the three consistency criteria are repeated iteratively. The number of iterations could be fixed by the user. Alternatively, the user could automatically stop the iterations by tracking the number of samples removed from the training set or the number of samples which have are assigned different labels compared to the previous iteration. The accepted classification model can then be used to classify the entire image.

4.3 Experimental Analysis

4.3.1 Data sets

4.3.1.1 Kigali, Rwanda: The first study area concerns an informal settlement in Kigali, Rwanda (Figure 4.3a). In 2015, a DJI Phantom 2 Vision+ UAV was flown over the study area. The images were processed with Pix4Dmapper which provided a Digital Surface Model (DSM) and a true-color orthomosaic with a spatial resolution of 3 cm and point cloud with a density of up to 1014 points/m². Further details regarding this dataset can be found in (Gevaert *et al.*, 2017). A subset of 150 m x 150 m was selected for the present analysis. The outdated building outlines were provided by the local government as vector data, which was initially digitized over a 2008 orthomosaic of 22 cm and partially updated using 2014 Pléiades satellite imagery resampled to 50 cm pixels (Bachofer, 2016). True reference data was obtained by manually digitizing the building outlines of the UAV orthomosaic. Around 11% of the segment labels provided by the existing outlines are incorrect according to this reference data.

4.3.1.2 Dar es Salaam, Tanzania: The second dataset was obtained by the Dar Ramani Huria project with the support of the World Bank[4]. This project mobilizes community members and university students to map flood-prone areas of the city. The results are used to support disaster response and are made available to the public through OpenStreetMap. To support these mapping activities, UAV flights were undertaken with a SenseFly eBee mounted with a 14MP Canon Powershot RGB camera. The images were again processed with Pix4Dmapper to obtain a point cloud with an average density of 50 points/m², and a 5 cm DSM and orthomosaic. A 300 m x 300 m subset located in the Tandale ward was used for the present analysis (Figure 4.3b).

For this dataset, the 'noisy' labels consist of the building outlines which were digitized over the 2015 imagery within the Ramani Huria project. Similar to the Kigali dataset, the true reference data was manually digitized over the subset for the purposes of the present study. Although most objects were correctly digitized by the Ramani Huria project, a label noise of about 10% is observed.

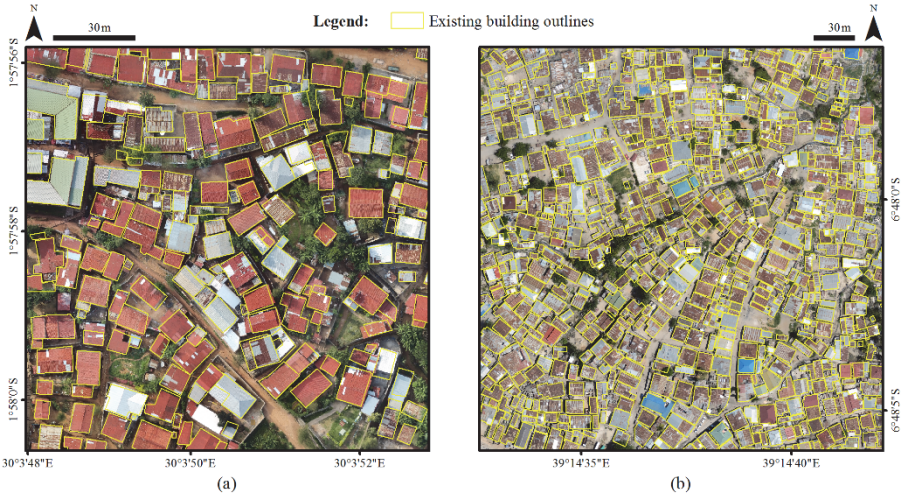


Figure 4.3: *The Kigali (a) and Dar es Salaam (b) datasets used in the present study. The building outlines (i.e. noisy labels) from t_0 are displayed in yellow over the images acquired at t_1.*

4.3.2 Experimental Set-up

4.3.2.1 Image Segmentation: To segment the orthomosaic, the SLIC super-pixel algorithm was used (Achanta *et al.*, 2012). SLIC first defines a regular grid over the image, where the grid interval is based on a user-defined target super-pixel size. These samples are used to initialize a k-means clustering, where each pixel in the image is assigned to the nearest cluster center. The

[4] http://ramanihuria.org/

proximity to cluster centers is calculated in a five-dimensional space which consists of spectral (L, a, b, values of the pixel in CIELAB colorspace) and spatial (the x,y image coordinates) components. After all pixels are assigned to a super-pixel, the cluster centers are updated by averaging the Labxy values of all pixels within the super-pixel, and the process is repeated. In our experimental analysis, the target super-pixel size was set to approximately 0.5 m² (i.e. 555 pixels for the Kigali dataset and 200 pixels for Dar es Salaam). However, it was noted that a number of very small segments were still present, which could unnecessarily slow down the image processing workflow. Therefore, all segments with an area less than 0.05 m² (i.e. 55 pixels for the Kigali dataset, 20 for Dar es Salaam) were merged with the most similar neighboring segment larger than 0.05 m². This segmentation strategy resulted in a total of 59812 image segments for the Kigali dataset and 103227 segments for Dar es Salaam.

4.3.2.2 Feature Extraction: For each segment, the average R, G, B color values, normalized r,g,b, and the ExG(2) vegetation index (Woebbecke *et al.*, 1995) were calculated, as well as a normalized histogram displaying the relative frequency of Local Binary Pattern (LBP) texture patterns (Ojala, Pietikainen and Maenpaa, 2002) within each segment. Features from the point cloud were also included in the classification: the number of points falling into the spatial extent of each pixel, as well as the range and standard deviation in the elevation values of these points. Planar segment and local neighborhood features of the highest point per pixel were also assigned to each image pixel. Planar segments were obtained through a surface-growing algorithm (Vosselman, 2012) and the number of points, average residual, inclination angle, and maximum height difference of the plane above neighboring points from different segments were included as features. The local neighborhood features were calculated according to the framework proposed by (Weinmann *et al.*, 2015). The values of each 3D feature per pixel was averaged over the image segments. These image-based and point-cloud based features together form the set of features used for the classification task. A more comprehensive overview of the utilized features is provided in (Gevaert, Persello and Vosselman, 2016).

4.3.2.3 Initial Training Labels: The vector data from the existing basemap are first rasterized using the same grid as the UAV orthomosaic. Next, a majority voting is used to assign a binary label (building vs. non-building) to each SLIC segment, which represents the outdated label at t_0. The true labels at t_1, which are used for the performance assessment, were obtained by manually digitizing the building outlines in the UAV orthomosaic, and assigning them to the image segments in the same way.

4.3.2.4 Pre-filtering: The previous processing steps yield: the input feature data consisting of segments which are described by radiometric, textural and geometric features derived from the UAV data, the outdated building vs. non-building labels at t_0, and true reference labels from the UAV data at t_1. In the pre-filtering step, only segments for which at least 60% of the segment had the same label were retained. Experimental results using a threshold of 0% (i.e. not applying the pre-filtering) and 100% (i.e. only including 'pure' segments in the classification) are also provided in the results section to indicate the importance of this pre-filtering step.

4.3.2.5 Iterative Supervised Classification: The next step performs the iterative classification with a random forest classifier. The number of trees is optimized through cross-validation, by randomly selecting 500 training samples, training forests with up to 200 trees, and selecting the number of trees with the lowest cross validation error. This optimal number of trees was then used to train the random forest classifier using all the training samples which was subsequently used classify the entire dataset.

4.3.2.6 Removal of Unreliable Training Samples: Four strategies are applied to illustrate the importance of combining both local and global contextual cues for identifying unreliable labels. The first method, *iterRF*, does not take any contextual criteria into account and simply uses the label consistency criterion. The second method, *iterRF-L*, removes samples for which the local contextual consistency index ψ_i is lower than the threshold value of 0.7. Whereas *iterRF-G* only employs the global contextual uncertainty index θ_i and the same threshold value. Finally, the proposed method *iterRF-LG* uses both local contextual consistency and global contextual uncertainty criteria. For all four methods, 15 iterations (of steps 5 and 6 in Figure 4.1) were performed.

4.3.2.7 Assessment: The four strategies are compared through the Overall Accuracy (OA) of all the segments, the OA of only the originally mislabeled segments, the number of false positives and false negatives in the training set, and the percentage of mislabeled samples in the training set. Note that the underlying idea of the proposed method is to eliminate the need of collecting labelled training data by exploiting existing geospatial information. Therefore, in order to compare the proposed method to the traditional method of manually labelling training samples for the classifier, we provide an experiment which indicates how many (correct) training sample labels would need to be collected in order to obtain the same classification accuracy. This is done by randomly selecting ten folds of a set number of reference samples at t_1, constructing a random forest classifier, and obtaining the OA. Finally, we also present the results of a sensitivity analysis. This was performed by taking the true labels of the Kigali dataset, and randomly changing training sample labels to induce a noise level of 0%, 5%, 10%, 20%, 30%, 40% or 50%. Using these labels,

iterRF-LG was again performed for 15 iterations. The average OA over three trials is reported for each iteration and noise level.

4.4 Results and Discussion

Table 4.1 provides the OA of the four different strategies for removing label noise from the training set after 15 iterations. The *iterRF* strategy, which only takes label consistency into account, does not improve the results significantly. The number of mislabeled and the classification accuracy is relatively stable after the first 15 iterations (Figure 4.4). The local contextual consistency (*iterRF-L*) and global contextual uncertainty (*iterRF-G*) achieve a similar accuracy for the Dar es Salaam dataset, correctly classifying about 90.4% of the image segments. Although a comparable number of noisy labels remain in the training set after 15 iterations (Figure 4.4), *iterRF-L* (91.2%) outperforms *iterRF-G* (90.1%) for the Kigali dataset. However, it is clear that the proposed method which combines all three criteria, *iterRF-LG*, obtains the best performance. For both datasets, a McNemar test with continuity correction (Foody, 2004) indicates that the results between *iterRF-LG* and the three other methods are statistically significant (p-value of < 0.001). The proposed method correctly classifies 92.1% of the segments for the Kigali dataset, corresponding to an improvement of 3.3% compared to using the initial, noisy training labels. This improvement was 1.7% for the Dar es Salaam dataset. The improvement is more visible when we consider only the segments which were mislabeled in the noisy training labels: the proposed method increased the accuracy of these segments from 6.6% to 47.9% in the Kigali dataset, and 8.5% to 41.1% in the Dar es Salaam dataset. Finally, the success of the method in removing unreliable labels is visible through the reduction of the fraction of mislabeled samples in the training data. This was effectively reduced from 11.1% to 6.2% in the Kigali dataset (effectively removing 44.1% of the mislabeled samples from the training set), and from 10.0% to 6.4% in the Dar es Salaam dataset (removing 36.0% of the mislabeled samples).

Table 4.1: *Accuracy measures of the proposed iterative strategies after 15 iterations.*

Data cleansing strategy	OA (%) All segments	OA (%) Mislabeled segments	False positives (%)	False negatives (%)	Percentage of mislabeled samples in training set
Kigali dataset:					
Using noisy labels	88.2	6.6	3.27	8.54	11.1
iterRF	88.9	7.3	2.94	8.15	11.0
iterRF-L	91.2	32.2	1.49	7.30	8.3
iterRF-G	90.1	23.6	2.14	7.79	9.1
iterRF-LG (proposed method)	*92.1*	*47.9*	*1.00*	*6.91*	*6.2*
iterRF-LG (no pre-filtering)	91.2	46.8	1.08	6.74	5.9
iterRF-LG (only uniform segments)	92.2	54.0	0.84	6.96	4.3
Dar es Salaam dataset:					
Using noisy labels	89.0	8.5	3.73	7.28	10.0
iterRF	89.6	8.6	3.45	6.92	10.0
iterRF-L	90.4	24.6	2.85	6.78	8.45
iterRF-G	90.4	24.7	3.33	6.31	8.23
iterRF-LG (proposed method)	*91.3*	*41.1*	*2.82*	*5.92*	*6.45*
iterRF-LG (no pre-filtering)	91.1	37.2	2.95	5.98	6.52
iterRF-LG (only uniform segments)	90.3	52.0	2.86	6.83	3.45

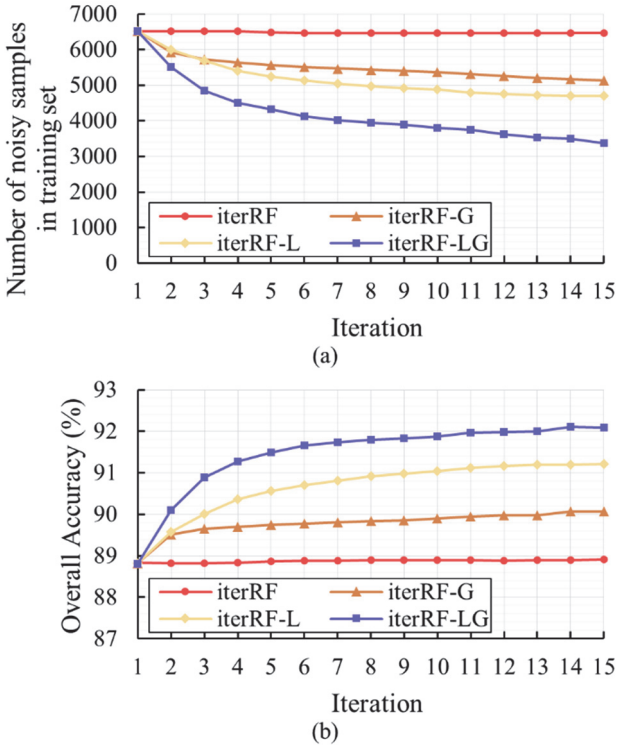


Figure 4.4: *The number of noisy training samples remaining in the set of samples used to train the classifier after each iteration (a) and the resulting Overall Accuracy for the Kigali dataset using the four different methods for filtering the training labels (b).*

The influence of the pre-filtering step on the results of *iterRF-LG* is also visible in Table 4.1. Increasing the uniformity criterion results in a decrease in the number of mislabeled segments in the training data and a more accurate classification of the mislabeled segments for both datasets. Using only pure segments (i.e. a uniformity of 99%) in the pre-filtering stage improves the OA of the entire Kigali dataset by 0.1%, but decreases the accuracy of the Dar es Salaam dataset by 1.0%. The results therefore suggest that increasing the uniformity criterion in the pre-filtering stage reduces the number of noisy labels in the training set and improves the classification accuracy of mislabeled samples. However, using a strict uniformity criterion may decrease the classification accuracy of the entire dataset, perhaps due to the exclusion of informative training samples.

The improved results of *iterRF-LG* compared to the other three methods is also visible in the output classification maps (Figure 4.5). For example, there is a notable reduction in false positives in the Kigali dataset. A building missed in the manual delineation of the buildings in Dar es Salaam (Figure 4.5c, top left), is correctly identified through the *iterRF-LG* method (Figure 4.5d). In the Kigali

dataset, a number of roofs are still not recognized by the classification model, remaining false negatives in the *iterRF-LG* method (Figure 4.5b). A visual analysis of the image indicates that many of these errors are in locations where the building extensions have been covered with a different roofing material than the (correctly labelled) adjacent construction. This causes a difference in the feature vectors of neighboring segments, and may therefore mislead the local contextual consistency criterion. If this is coupled to a consistent change in the representation of objects between t_0 and t_1 – for example if building extensions consist of a new type of roofing material which is not well-represented by the existing labels – then the global contextual uncertainty may also fail. Note that in (1), the all segment features are weighted equally when determining the similarity between neighboring segments. Further research could consider incorporating more advanced techniques to select or weight the different features as previous research indicated that considering 3D and 2D features separately may improve image classification results (Gevaert, Persello and Vosselman, 2016).

Another set of experiments compared the proposed workflow with a traditional workflow, where image segments must be labeled manually. Experimental analysis indicates that approximately 600 correctly labeled training samples would be needed in the Kigali dataset to obtain the same accuracy as *iterRF-LG* after 15 iterations (Figure 4.6a). For the Dar es Salaam dataset, this is much higher, and between 50,000 and 60,000 training samples would be needed (Figure 4.6b). This could be due to the spectral similarity of building roofs and ground in the Dar es Salaam dataset. Furthermore, the slightly lower spatial resolution of the Dar es Salaam dataset makes it difficult to capture the texture of the corrugated iron roofs, which proved to be an important distinguishing attribute for the Kigali dataset (Gevaert *et al.*, 2017).

The large number of training samples required for the Dar es Salaam dataset can easily be dealt with by a random forest classifier. Other supervised classification methods, such as SVM also achieve high accuracies in remote sensing applications (Bruzzone and Persello, 2010). Future investigations regarding the use of SVM instead of random forests for the proposed *iter-LG* method would require two adaptations. Firstly, the number of training samples would need to be reduced by sampling or using an SVM variant which is capable of dealing with large numbers of training samples such as DC-SVM (Hsieh, Si and Dhillon, 2014). Secondly, a classifier uncertainty measure (u_j^k) would need to be assigned to each training sample to calculate the global contextual uncertainty. SVM does not directly provide a probability for the classification output, although strategies exist which use proxies to indicate the classification certainty, e.g. (Demir, Persello and Bruzzone, 2011).

Figure 4.5: *Results of the classification using the noisy labels (a,c) and after the fifteenth iteration of iterRF-LG (b,d) for the Kigali (a,b) and Dar es Salaam (c,d) datasets.*

Finally, the results of the sensitivity analysis for the Kigali dataset are presented in Figure 4.7. The results indicate that after 15 iterations the classification accuracy is above 93% for noise levels of up to 30%. In these experiments, the noise is introduced by switching the labels of randomly selected training labels. It is possible that the label noise in practical applications is more systematic (e.g. new constructions make use of a different roofing material, or a concentration of adjacent mislabeled samples), which may have a more significant impact on the results of the proposed method.

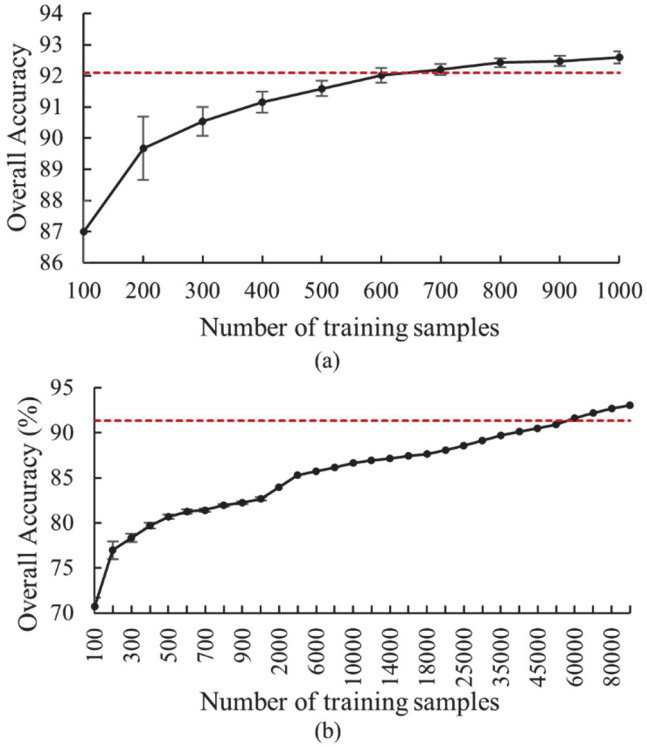


Figure 4.6: *A comparison between the Overall Accuracy achieved through iterRF-LG after 15 iterations (red dashed line) and the mean Overall Accuracy achieved by randomly selecting a set number of training samples with true labels (black line) for the Kigali (a) and Dar es Salaam (b) datasets.*

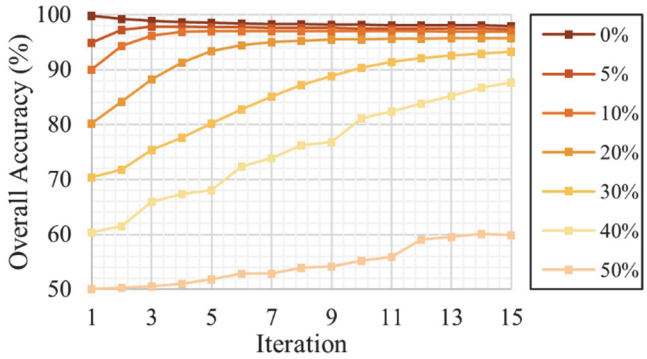


Figure 4.7: *Overall Accuracy of iterRF-LG for the Kigali dataset after 15 iterations with initial label noise levels ranging from 0% to 50%.*

4.5 Conclusions

In this paper, we utilize two datasets to demonstrate how existing spatial data may be exploited to obtain labeled training samples for the application of supervised classification algorithms to UAV data. Considering that a number of labels provided by this outdated spatial data will be erroneous, local and global image cues are used to filter out unreliable training samples. The local contextual criterion encourages neighboring image segments to have consistent labels. At the same time, a global contextual criterion uses the entire scene to capture the distribution of the semantic classes in the feature space, and is suitable for identifying isolated new objects. Sensitivity analyses show that classification accuracies of 93% or more are achieved, even in presence of up to 30% erroneous training samples. There are two main implications of these results. The first is that the proposed method may lead to a considerable speed-up in the implementation of supervised classification methods for base-map updating by reducing the need of manually labelling image segments to train the classifier. Secondly, the interaction between the local and global cues emphasizes that the inclusion of spatial contextual information is beneficial for data cleansing techniques in geomatics applications.

The proposed method may also be used for a number of other applications. For example, it could be used in a quality control application to verify the accuracy of volunteered geographical information such as OpenStreetMap. Furthermore, it could be used in a domain adaptation application, where the training labels are obtained from a classification model trained on a certain study area could be applied to a similar study area for which no data is available, rather than outdated spatial data. The main caveat of this method is that it assumes that the noisy data labels provided by the outdated spatial data cover all the representations of the semantic classes in the new UAV imagery. Therefore, if an entirely new variation of an object appears between t_0 and t_1, for example if an alternative type of roof material is only used in new constructions, the mislabeled segments will not be filtered by the proposed method. Further developments could explore active learning methods (Persello, 2013) to target such segments and potentially improve the classification accuracy, though this would require (limited) manual labelling.

Chapter 5 – A Deep Learning Approach to DTM Extraction from Imagery Using Rule-based Training Labels[5]

[5] This chapter is based on:

Gevaert, C.M. , Persello, C., Nex, F, and Vosselman, G. (2018) 'A deep learning approach to DTM extraction from imagery using rule-based training labels' *ISPRS Journal of Photogrammetry and Remote Sensing*, 142, pp.106-123. doi: 10.1016/j.isprsjprs.2018.06.001

Abstract

Existing algorithms for Digital Terrain Model (DTM) extraction still face difficulties due to data outliers and geometric ambiguities in the scene such as contiguous off-ground areas or sloped environments. We postulate that in such challenging cases, the radiometric information contained in aerial imagery may be leveraged to distinguish between ground and off-ground objects. We propose a method for DTM extraction from imagery which first applies morphological filters to the Digital Surface Model to obtain candidate ground and off-ground training samples. These samples are used to train a Fully Convolutional Network (FCN) in the second step, which can then be used to identify ground samples for the entire dataset. The proposed method harnesses the power of state-of-the-art deep learning methods, while showing how they can be adapted to the application of DTM extraction by (i) automatically selecting and labelling dataset-specific samples which can be used to train the network, and (ii) adapting the network architecture to consider a larger surface area without unnecessarily increasing the computational burden. The method is successfully tested on four datasets, indicating that the automatic labelling strategy can achieve an accuracy which is comparable to the use of manually labelled training samples. Furthermore, we demonstrate that the proposed method outperforms two reference DTM extraction algorithms in challenging areas.

5.1 Introduction

Airborne Laser Scanning (ALS), satellite imagery, and aerial or UAV imagery can provide a *Digital Surface Model* (DSM) which describes the elevation of the Earth's surface. This model describes the elevation of the top of objects, i.e. the elevation of the ground plus the height of objects such as buildings and vegetation which is on top of the surface. However, many applications actually require a model where these elevated objects are removed, i.e. a *Digital Terrain Model* (DTM), as depicted in Figure 1. The difference between the DSM and DTM is referred to as a normalized Digital Surface Model (nDSM), and gives the height of the elevated objects. The conversion of a DSM to a DTM is known in literature as DTM extraction, bare-ground extraction, or point cloud filtering. This process generally consists of two phases: first selecting pixels or points which represent the ground and then using these points to interpolate a surface model of the terrain.

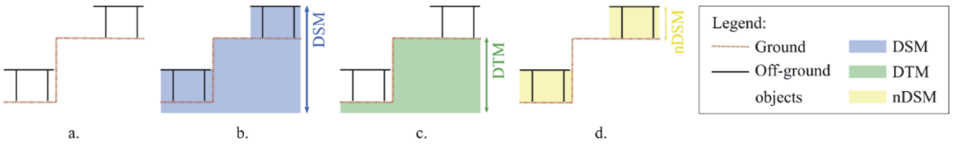


Figure 5.1: *Given a scene with the ground and objects such as buildings (a), the Digital Surface Model (DSM) provides the height of the ground plus any objects on top of it (b), the Digital Terrain Model (DTM) filters off-ground objects and therefore provides the elevation of only the ground surface (c), and the normalized Digital Surface Model (nDSM) represents the difference between the DSM and DTM, essentially giving the height of the objects on top of the terrain (d).*

Most DTM extraction algorithms have been tested on relatively easy datasets (Tomljenovic *et al.*, 2015). However, we can identify a number of specific scenarios which present difficulties for DTM extraction from point clouds of urban areas (Figure 5.2). A number of difficulties arise due to errors inherent in the data itself. For example shadows cause difficulties for dense matching algorithms, resulting in noise in the point cloud (Figure 5.2a). Also, lack of texture or unsatisfactory camera calibration may cause noise or outliers in the point cloud (Figure 5.2b). The DSM interpolation step may also cause errors, such as increasing the extent of elevated objects when using Inverse Distance Weighting (Figure 5.2c) or artefacts along overhanging objects when using Delaunay triangulation (Figure 5.2d). Other sources of difficulties for DTM extraction algorithms are due to the characteristics of the scene itself. For example, sloped surfaces may cause a step-like pattern where ground and off-ground cannot be distinguished (Figure 5.2e) or off-ground objects to be co-planar with the ground (Figure 5.2f). Finally, elevated objects which are significantly larger than the other objects in the scene (Figure 5.2g) or agglomerations of neighboring objects (Figure 5.2h) form contiguous off-ground areas which affects the size of the local neighborhood which must be considered to identify ground points as most algorithms somehow assume that

ground points will locally be the lowest point. In our approach, we demonstrate how complementary information from the imagery can be included to successfully extract a DTM in these challenging areas.

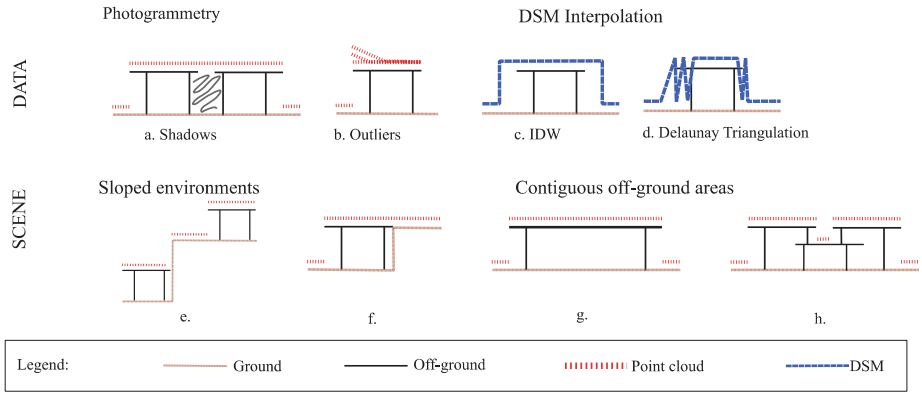


Figure 5.2: *An overview of sources of errors in DTM extraction algorithms. The data itself has errors, such as shadows (a) and outliers (b) which are byproducts of the photogrammetric workflow. Also, DSM interpolation methods such as Inverse Distance Weighting (IDW) (c) and Delaunay Triangulation (d) create artifacts in the DSM. Scene characteristics such as sloped environments (e and f) and contiguous off-ground areas due to exceptionally large buildings (g) or connected buildings (h) also cause difficulties.*

Existing algorithms for DTM extraction from DSMs or point clouds can be roughly divided into five groups: (i) morphological filtering, (ii) progressive densification, (iii) surface-based, (iv) segment-based, and (v) deep learning methods. In *morphological filtering*, the ground is defined as the lowest point within a specified neighborhood. Variations of this method include: making the threshold dependent on the distance to the center point (Sithole and Vosselman, 2005) or adapting the filter to the slope calculated from an existing DTM (Sithole and Vosselman, 2005; Debella-Gilo, 2016). Morphological methods are very sensitive to the size of the search neighborhood. For example, if the element is too small, it may cause elevated objects slightly lower than the surrounding objects to be mistakenly labelled as ground (e.g. Figure 5.2h). To avoid this, some approaches use neighborhoods of various sizes. For example, Kilian et al. (1996) use structuring elements of various sizes and then link the likelihood of a point being considered ground with the size of the structuring element for which the point is labelled as ground. Similarly, Mongus et al. (2014) use morphological profiles of various sizes, and record: the largest response, the size of the structuring element at the first response, and the cumulative sum of responses up to the largest response as three features. These are said to reflect the height of features compared to the direct surrounding, planimetric size of the elevated object, and estimation of the height of the object.

After the ground points are obtained through the filtering, an interpolation can be performed to obtain the DTM surface. For example, a Triangulated Irregular Network (TIN) represents the surface through a series of triangles where the vertices are the ground points. This can be done through a Delaunay triangulation, which constructs a TIN in such a way that no points are within the circumcircle of one of the surface triangles, and the vertices of all triangles are maximized (Lawson, 1972). This is a common method, and a wide range of adaptations have been developed to optimize it (Tsai, 1993). Another interpolation method is Inverse Distance Weighting (IDW), where all points within a given neighborhood are utilized as input for the surface, but nearer points are given more weight on the surface estimation than further points (Hohn, 1991). The performance of the interpolation algorithms depends on e.g. surface characteristics and dataset density (Chaplot *et al.*, 2006). However, in the current study we focus on the correct identification of ground points, and a further comparison of interpolation methods is not considered.

Progressive densification methods select a number of 'seed' points which are likely to represent the ground, and then successively add points to those classified as ground. For example, Axelsson (2000) used a grid to select the lowest points which are then used to construct an initial TIN model. This TIN is progressively densified by adding points which are less than a user-defined distance from an existing TIN face, and form an angle less than a user-defined threshold with the three vertices of this face. With a total error of 11.2% algorithm had comparatively good results on the ISPRS benchmark set (Sithole and Vosselman, 2004); though it is said to have difficulties in identifying cliffs and sharp ridges (e.g. Mongus et al., 2014; Zhang et al., 2016).

Surface-based or *interpolation* methods estimate a surface from all the input points and suppress the influence of off-ground points on the interpolation. For example, at the first iteration, a surface can be interpolated using all available points. One can then assume that points on the ground are likely to be below the interpolated surface. These lower points are then assigned a higher weight in the interpolation for the next iteration(Kraus and Pfeifer, 1998). Alternatively, an active shape method can be applied, which describes the surface as a rubber cloth and forms it to the laser points in a bottom up fashion (Elmqvist *et al.*, 2001). The surface is adjusted iteratively using an energy function which weighs the 'stiffness' of the interpolated surface (internal force) against the individual point observations (external force). Zhang et al. (2016) propose a similar approach based on 'cloth simulation filtering' to identify potential ground points. Surface-based methods experience difficulties in areas with steep slopes (Liu, 2008), which may require explicit post-processing (e.g. Zhang et al., 2016).

Segment-based methods generally consist of 3 steps: (1) the segmentation of a point cloud or DSM, (2) the classification of the segments as ground or non-ground, and (3) the interpolation of the DTM from ground segments (Beumier and Idrissa, 2016). Point cloud segmentation may use profiles (Sithole and Vosselman, 2005) or region-growing techniques. For the latter, local minima are often used to obtain seed points, which are densified through e.g. planar segmentation (Pérez-Garcia *et al.*, 2012), or similarity of normal vectors (Tóvári and Pfeifer, 2005). For 2D raster methods, slope is often used to define segmentation boundaries. For example, Hingee et al. (2016) calculate the slope of the DSM, which is used to segment the raster. Segments where majority of pixels are flowing 'in' are candidates for ground, then surface fitting is applied. Tomljenovic et al. (2016) also uses slope to delimit segments, the largest segment is considered to be the ground. Note that Mongus et al. (2014) mention that slope-based filtering doesn't work well in sloped study areas. Yan et al. (2012) use a locally lowest points to initiate the region growing segmentation, where slope is used to determine whether pixels are included in the segment or not. They define a segment as terrain or non-terrain based on the signed height differences between neighboring segments. Beumier and Idrissa, (2016) use a maximum height difference and two-step connected component algorithm to define the segments, and additionally define a minimum region size parameter. Segment size and its relative elevation to the neighboring segments are used to identify ground segments. Segment-based methods may speed up processing compared to pixel- or point-based methods and reduce sensitivity to noisy data, though the quality of the results is heavily dependent on the quality of the segmentation to begin with.

Finally, *deep learning* algorithms have recently been improving accuracies on a wide range of supervised classification tasks (e.g. He et al., 2016; Krizhevsky et al., 2012; Simonyan and Zisserman, 2014). In computer vision applications, Convolutional Neural Networks (CNNs) were used to give a single semantic label to an entire image patch. CNNs consist of a combination of: convolutional layers which apply a series of filters to the input image, nonlinear activation layers which allow complex representations to be learned, and pooling components to help prevent overfitting. CNNs have been very successful in classification tasks which require assigning a label to an image patch or scene (Krizhevsky, Sutskever and Hinton, 2012; He *et al.*, 2016). More recently, Fully Convolutional Networks (FCNs) have been developed for tasks which require assigning a label to each pixel within an image, i.e. semantic segmentation (Shelhamer, Long and Darrell, 2017). They are more efficient for semantic segmentation than conventional CNNs as they avoid redundant calculations, improve memory efficiency and incorporate more training data into the optimization of the weights. Furthermore, a significant benefit of FCNs is that, unlike patch-based CNNs, they can be easily applied to images with different dimensions.

Due to the convincing results achieved on computer vision benchmarks, CNNs (Hu *et al.*, 2015; Romero, Gatta and Camps-Valls, 2016; Zhang, Zhang and Du, 2016) and FCNs (Sherrah, 2016) are also increasingly being applied to classify satellite and aerial imagery in remote sensing applications. For example, some studies utilize networks trained on computer vision datasets (Audebert, Saux and Lefèvre, 2016) or synthetic multispectral imagery (Kemker and Kanan, 2017), and fine-tune the weights using real aerial imagery. Deep learning has also been applied for DTM extraction from point clouds, where Hu and Yuan (2016) recently achieved state-of-the-art results on the ISPRS benchmark dataset using a CNN. The authors first convert the point cloud into a 2D grid consisting of three attributes: the minimum, maximum and mean height per grid cell. More than 17 million pre-labelled training samples were then used to train a CNN capable of distinguishing ground vs. non-ground points. The method obtained accurate results, but required a large amount of labelled data.

We foresee that there are three main concerns which must be overcome in order to efficiently exploit the power of deep learning for DTM extraction. Firstly, the collection of a sufficient amount labelled training data for training the networks is costly and time-consuming. In some cases, such labeled data may be available due to extensive manual labor, but here we consider cases where such labeled data is not available. Secondly, previous DTM extraction algorithms indicate that it is important to consider elevation differences over a local neighborhood which exceeds the size of the largest off-ground object in the scene. However in the case of DTM extraction from extremely high resolution UAV data products, covering such an extensive area would require very large image patches as input for a FCN. The challenge is therefore how to consider the information over a large area while limiting the number of network parameters which must be tuned as well as the size of the input patch used to train the network. Thirdly, even if a network is correctly tuned, there are still cases in which using only the elevation information is not enough to distinguish ground from off-ground samples (e.g. Figure 5.2e,f).

In the field of large-scale urban scene reconstruction, researchers have shown how incorporating both 3D and 2D information is beneficial. For example, to jointly perform image segmentation and dense stereo reconstruction. This can be done by jointly optimizing the random field formulations of both problems (Ladický *et al.*, 2012). At an object level, learning the mean shape of an object from 3D scans can be combined with image-based cues of anchor points to improve the accuracy of multiview stereo workflows (Bao *et al.*, 2013). Probabilistic models (Ulusoy, Black and Geiger, 2017) or 3D deep learning strategies (Riegler *et al.*, 2017) can learn object shapes to support 3D reconstruction in occluded or texture-less areas. Classification problems on a larger scale also benefit from the integration of 2D and 3D information. For

example, a voxel's preference for a certain semantic class (image-based cues) can be supplemented by the likelihood of certain surface orientations (3D geometric cues) (Hane *et al.*, 2013). Smart strategies using hierarchical voxel schemes can be used to maintain a high classification accuracy while reducing the memory and computational times enormously (Blaha *et al.*, 2016). Other workflows combine an even wider range of data sources: from OpenStreetMap, LiDAR, aerial photography and semantic data for large-scale scene reconstruction (Cabezas, Straub and Fisher, 2015). Although scientific research displays great potential for this field of large-scale 3D scene reconstruction and semantic interpretation, we acknowledge that a number of applications simply require an input DTM. The purpose of this manuscript is therefore to exploit these observations of synergies between information contained in the visual and geometric information of a scene, but applied to a simpler task of DTM extraction with more conservative computational and data requirements.

More specifically, this paper proposes the use of deep learning in the form of a FCN for DTM extraction. DSMs derived from photogrammetric point clouds and the corresponding true-orthophotos are used as input. The utilization of both sources of information is one of the main points of our approach, and is key to DTM extraction in challenging areas. Our method uses a simple rule-based mechanism to automatically identify ground and off-ground samples which are then used to train the network, thus eliminating the need to collect large sets of manually labelled training data. Secondly, the network takes a large surface area into account by considering topographic features derived from DSM (which summarize the height of a pixel to the local topographical tendencies) and by applying dilated filters in the network architecture. Finally, difficult scenarios which may confuse existing DTM methods are solved by exploiting the RGB information obtained from the UAV imagery in conjunction with the DSM. In the following manuscript we describe the proposed method for DTM extraction and demonstrate its accuracy using three challenging datasets. The use of the three different datasets attests to the versatility of the method for VHR aerial imagery due to the dataset characteristics (i.e. two were acquired with a UAV and one through aerial imagery and all three have different spatial resolutions) and scene characteristics.

The proposed methodology is assessed by casting it as a binary classification problem (i.e. ground vs. off-ground). We illustrate the importance of combining image-based and DSM-based features by performing sets of FCN experiments using differing input features. Furthermore, we perform experiments using the ground-truth labels vs. the rule-based training labels to support the claim that simple morphological rules are a viable alternative to manually labelling the large number of training samples required to train deep networks for DTM extraction. Further experiments compare the proposed network architecture to

deeper networks, apply the algorithm to the ISPRS benchmark data, and consider the possibility of direct regression-based DTM prediction.

5.2 Proposed method

The proposed methodology consists of two steps (Figure 5.3). The first step is a rule-based selection of a set of candidate ground (S_g) and off-ground (S_o) training samples. The idea is that if a supervised classification will follow the initial rule-based method, it is not necessary to label *all* of the pixels as ground or off-ground. Rather, it suffices to have a large number of confident samples. In this case, simple morphological filters are applied to the DSM to select the training samples, as the filters can be executed quickly and the algorithm parameters are intuitive to the user (i.e. neighborhood search window clearly corresponds to the expected size of the object in the scene).

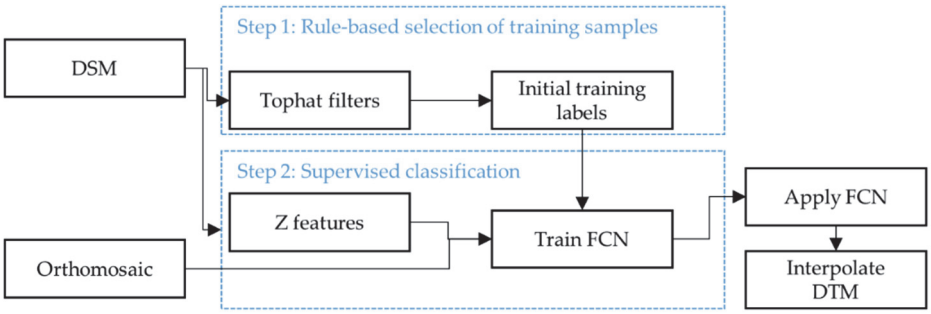


Figure 5.3: *Workflow of the proposed methodology. The first step consists of applying top-hat filters to the DSM to select and label initial training samples. The second step combines the RGB channels of the orthomosaic with features derived from the DSM together with the labeled samples from the first step to train a FCN. This FCN is then applied to the entire dataset to identify the ground samples, which can then be used to create a DTM through interpolation.*

The second step consists of a supervised classification combining radiometric features from the imagery with geometric features from the DSM to refine the initial labeling. Image-based classification in urban settings can be challenging due to high within-class variability (e.g. different roof materials and colors as well as the presence of clutter on roofs) and low between-class variability (e.g. especially when ground and roof pixels appear similar in true color). Contextual features such as texture can improve the separability of these two classes (Gevaert et al., 2017), but hand-crafting informative texture features can be challenging. Therefore, we use a FCN as a classifier. In addition to the recent success of deep learning methods for various image classification tasks, their ability to learn powerful contextual features from the data itself supports the development of automatic workflows. Once the pixels corresponding to ground samples have correctly been identified, the final DTM can be interpolated. In the following sections, we describe both steps in more detail.

5.2.1 Rule-based training sample selection using morphological filters

Let us define $\gamma_{w_s}(DSM)$ as a morphological top-hat filter on the DSM. This filter returns the height of the central point above the lowest point in a disk-shaped neighborhood w_s with a radius of s. However, rather than utilizing multiple scales (Mongus, Lukač and Žalik, 2014), we utilize only two neighborhoods: w_{small} and w_{big}. The first filter, w_{small}, is used to identify off-ground objects. The idea is that pixels which are higher than their direct neighbors provides a set of confident off-ground samples (i.e. the filter indicates that the pixel is higher than neighbors within a small neighborhood), and that these selected samples will be representative of the image-based characteristics of off-ground objects within the dataset. Problems with contiguous elevated objects (e.g. Figure 5.2g and h) are addressed as we assume that pixels along the edges of elevated objects (such as roofs in the figure) will have a similar appearance in the image as pixels in the interior of these roofs.

The set of off-ground training samples is defined as:

$$S_o = \{\gamma_{w_{small}}(DSM) > \tau_o\}, \tag{5-1}$$

where τ_o represents the threshold in meters which defines the minimum height difference between a DSM pixel and its neighbors to be considered as off-ground. Later experiments on the ISPRS benchmark indicated that datasets containing large buildings with flat roofs, unique roofing material and located on flat terrain (e.g. industrial areas) are not always captured by the rule in (1). These areas can therefore benefit from an additional criterion to select off-ground samples: $(DSM - \zeta_{w_{small}}(DSM)) > \tau_o$, where ζ is a morphological erosion filter.

Similarly, we can consider that ground pixels have a minimal response to $\gamma_{w_{big}}(DSM)$. That is to say, a ground point is likely to be lower than pixels within a larger neighborhood. As with other morphological methods, this search range should be large enough to extend over large objects in the scene, yet not too large as this will be problematic in sloped areas. The set of ground training samples S_g is then:

$$S_g = \{\gamma_{w_{big}}(DSM) < \tau_g\}. \tag{5-2}$$

In practice, we set $\tau_g = 0.5 \cdot \tau_o$ which reduces the number of parameters to be tuned by the user. Thus, ground and off-ground samples are selected and labeled automatically for each dataset through two simple rules which require the user to tune only three intuitive parameters: w_{small}, w_{big}, and τ_o.

5.2.2 Fully convolutional neural networks

The selected training samples are used to train a FCN. Detailed descriptions are available regarding the applications of CNN (Castelluccio *et al.*, 2015; Hu *et al.*, 2015), and FCNs (Sherrah, 2016; Persello and Stein, 2017) for image classification tasks in remote sensing. When applying a FCN to DTM extraction applications, especially when utilizing data with an extremely high spatial resolution such as those acquired with UAVs, one of the main concerns is how to consider a large spatial extent without increasing the computational costs of the network. Considering contextual information over a large spatial extent is important for DTM extraction algorithms. For example, the search neighborhood in morphological filtering methods should be larger than the largest off-ground object. Similarly, when using a FCN for DTM extraction, the receptive field of final layer should be large enough to capture relative elevation differences between off-ground pixels and the surrounding ground pixels. We do this in two ways: by adapting the network architecture and through the use of specialized feature inputs.

Both CNNs and FCNs can be defined as a sequence of layers which generally consist of convolutional, nonlinear activation, and pooling components. Using the same notation as (Volpi and Tuia, 2017), the convolutional layers consist of a set of K' filters with a size of $M \times M \times K$, where M is the width and height of the square filter and K corresponds to the number of input channels of the previous layer. For example, for an RGB image this K would have a value of three. Each filter is convolved over the input layer $\boldsymbol{x}$, producing a response $\boldsymbol{x}'_{\mathrm{ijk'}}$ for the k'^{th} filter at row i and column j of the output layer $\boldsymbol{x}'$ as follows:

$$\boldsymbol{x}'_{\mathrm{ijk'}} = \sum_{k=1}^{K} \sum_{q=1}^{M} \sum_{p=1}^{M} \boldsymbol{w}_{\boldsymbol{pqk}} \cdot \boldsymbol{x}_{pqk} + b\,, \tag{5-3}$$

where $\boldsymbol{w}_{\boldsymbol{pqk}}$ is the filter value of row p, column q, and channel k of the input layer and b are the bias parameters which are learned by training the network. One of the main advantages of the convolutional layers is that, once optimized in the training stage, the weights of the filter are fixed as it passes over the image in the testing stage. This not only decreases the number of parameters to be learned, but also introduces translation invariance. The dimensions of $\boldsymbol{x}'$ depend on the stride (s) and padding (z). The stride is the interval for which each convolution is calculated. A stride equal to one indicates that the convolution is calculated for each pixel of $\boldsymbol{x}$ whereas values higher than one indicate that pixels are skipped and $\boldsymbol{x}'$ will therefore be downsampled. The padding indicates the number of zeros added to the border of the input image to enable pixels along the edges of $\boldsymbol{x}$ to be processed. The receptive field of a filter refers to the area of the original input image which affects the filter response. This can be increased by applying multiple convolutional layers, increasing the size of the filters, or increasing the stride. Another way to

increase the receptive field without increasing the number of variables to be tuned is by inserting a defined number (d) of 0s between weights of $\boldsymbol{w}$. This technique is known as the atrous method (Chen *et al.*, 2015) or dilation (Yu and Koltun, 2015). For an input layer $\boldsymbol{x}$ with dimensions $W \times H \times K$, the dimensions of $\boldsymbol{x}'$ will then be: $\left(\frac{W+2z-M(d-1)-1}{s}\right) \times \left(\frac{H+2z-M(d-1)-1)}{s}\right) \times K'$.

Convolutions are generally followed by a nonlinear activation, which introduces non-linearity into the system thus allowing more complex representations to be learned. One of the most common methods currently used is the Rectified Linear Unit (ReLU), defined as $\boldsymbol{x}'_{k'} = \max(0, \boldsymbol{x}_{k'})$ (Nair and Hinton, 2010). This function is capable of efficient network training and it avoids the vanishing gradient problem (He *et al.*, 2015).

The third main component of FCNs are the pooling layers. The purpose of pooling layers is to summarize the filter responses and improve translation invariance (Krizhevsky, Sutskever and Hinton, 2012). A common strategy is max-pooling, which returns the highest response over a small window (generally 2 x 2 or 3 x 3). Pooling layers commonly utilize a stride set equal to the pooling size window. This returns a single value for each (e.g. 2 x 2) window and thus downsamples the image. In semantic segmentation applications, the final output layer should have the same dimensions as the input layer. Therefore, the network may make use of deconvolutional layers which again upsample the features at a later stage in the network (Shelhamer, Long and Darrell, 2017; Volpi and Tuia, 2017). Alternatively, it is possible avoid downsampling in the pooling layer by avoiding pooling layers altogether or by using pooling layers with a stride equal to one (Sherrah, 2016). Results of the ISPRS 2D semantic labelling contest[6] suggest that the latter strategy is competitive with more complex deconvolutional strategies (Volpi and Tuia, 2017).

5.2.3 Proposed network

In our proposed network, we therefore utilize a FCN with no downsampling to ensure that the output ground prediction map will automatically have identical dimensions as the input dataset. The network consists of three convolutional layers (Table 5.1). The first two are followed by ReLUs, and a max-pooling with no downsampling. As there is no downsampling, no deconvolutional layers are needed to ensure the output map has the same dimensions as the input map. This strategy has been previously used by FCN architectures for the classification of satellite imagery (Sherrah, 2016; Persello and Stein, 2017). The receptive field is increased greatly in the second convolutional layer through the use of dilated filters. The use of dilated filters also introduces a

[6] http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html

multi-scale effect, which is typically achieved through downsampling. However, the use of dilated filters as opposed to downsampling has the additional benefit that fewer parameters are required, thus speeding up training and reducing potential overfitting (Yu and Koltun, 2015). The architecture used here also reduces overfitting by introducing a batch normalization (Ioffe and Szegedy, 2015) after each convolution and dropout (Srivastava *et al.*, 2014) after the final convolution.

Table 5.1: *An overview of the FCN network architecture utilized for the DTM extraction.*

Layer	Filter size M (pixels)	Filter dilation d (pixels)	Number of filters K'	Padding z (pixels)	Receptive field size (pixels)
Convolutional1 Batch normalization ReLU	5 x 5	1	16	2	5 x 5
Pooling1	3 x 3	-	-	1	7 x 7
Convolutional2 Batch normalization ReLU	9 x 9	6	16	24	55 x 55
Pooling2	3 x 3	-	-	1	57 x 57
Convolutional3 Batch normalization Dropout	1 x 1	1	2	0	57 x 57

The second manner to increase the extent under consideration by the FCN is by incorporating DSM feature which describe the topography over a large area as input channels for the network. We choose to include these DSM features as they allow the network to consider topographical variations over an extended area without increasing the computational costs of training the network. I.e., while the utilization of dilated kernels may reduce the number of parameters to be tuned, it still requires a larger input patch for training the network and thereby increases the memory requirements. Remember that existing DTM extraction methods indicate that the relative height must be considered over an area larger than the largest elevated object (i.e. Figure 5.2 g,h). If a dataset has a spatial resolution of 3 cm, then it requires a receptive field of for example 667 x 667 pixels in order to take a 20 x 20 m area into account, which is considerably larger than the patch size of the current network. Therefore, we include features which describe the surface topography over a larger area as input for the FCN. These features are inspired by DTM extraction methods which take an existing DTM into account (Sithole and Vosselman, 2005; Debella-Gilo, 2016) and the surface-based or interpolation

methods (Kraus and Pfeifer, 1998). Namely, we define a grid over the DSM and select the point with the elevation which corresponds to the lowest 10% of the pixels in the cell. We avoid selecting the lowest point per cell as photogrammetric point clouds may contain many outliers which could negatively affect the interpolation (Nex and Gerke, 2014). A bicubic interpolation is then applied to these lowest points. In addition to the original DSM, we utilize two grids: one of 1 x 1 m to preserve local topographical details and the other of 20 x 20 m to describe the general surface topography. The combination of these three features representing the absolute height forms one feature set (Z). Another feature set simulates the height of objects above these surfaces and consists of the difference between the DSM and the two interpolated height features (nZ). An overview of these feature sets is given in Table 5.2.

Table 5.2: *Description of the different feature sets used to train the FCN.*

Feature set name	Data source(s)	Number of channels	Description
RGB	Image	3	Red, green, blue color channels
Z	DSM	3	DSM Local topography: interpolation of lowest elevation decile every 1 m General topography: interpolation of lowest elevation decile every 20 m
nZ	DSM	2	DSM – Local topography DSM – General topography
DTM	DSM	1	An approximated DTM formed by interpolating a surface from all pixels labeled as ground.
nDSM	DSM	1	An approximated normalized DSM formed by subtracting the DTM above from the input DSM.

5.3 Experimental analysis

5.3.1 Data sets

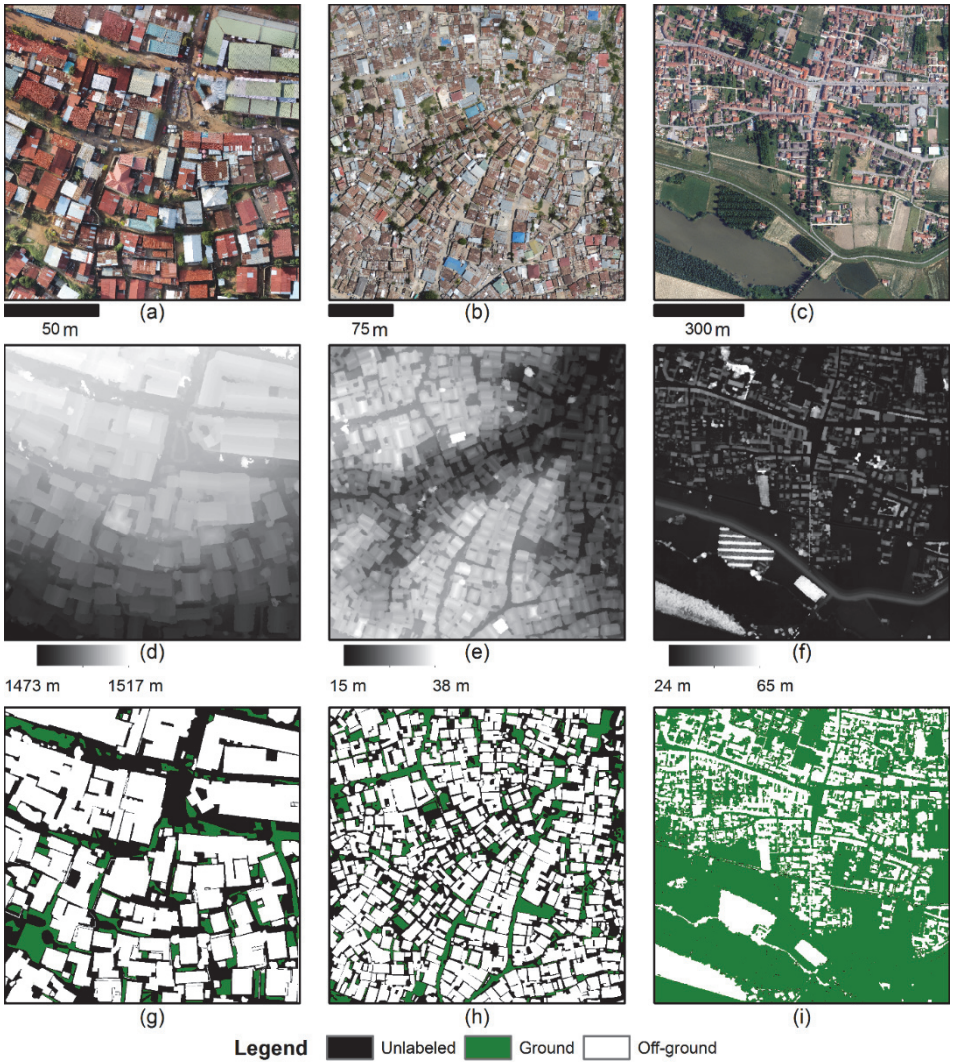


***Figure 5.4**: Images of the Kigali (a), Dar es Salaam (b), and Lombardia (c) datasets, and their respective DSMs (d-f) and manual reference data (g-i).*

5.3.1.1 Kigali, Rwanda: The first dataset consists of UAV imagery collected over an informal settlement in Kigali, Rwanda (Figure 5.4a,d). Images were collected with a DJI Phantom 2 Vision+ quadcopter and processed with Pix4Dmapper to obtain a DSM and true-color orthomosaic with a spatial resolution of 3 cm. A subset of 5000 x 5000 pixels (150 x 150 m) was selected which contains densely grouped buildings separated by narrow footpaths which

are often shadowed. The terrain of the lower part of the image contains steep slopes, making it a challenging scene for DTM extraction algorithms. More information regarding the UAV data collection and processing can be found in Gevaert et al., (2017). The reference data (Figure 5.4g) was manually created by visual interpretation.

5.3.1.2 Dar es Salaam, Tanzania: The second dataset consists of UAV imagery over Dar es Salaam, Tanzania (Figure 5.4b,e). The images were collected in 2015 with a SenseFly eBee mounted with a 14 MP Canon Powershot RGB camera in the context of a World Bank project (Dar Ramani Huria[7]). These images were processed with Pix4Dmapper to obtain a DSM and true-color orthomosaic with a spatial resolution of 5 cm. A subset of 6000 x 6000 pixels (300 x 300 m) was selected for the current analysis. The area again covers an informal settlement. Although the area is not as steeply sloped as in Kigali, the area also challenging due to the presence of contiguous off-ground areas and spectral similarity between the ground and off-ground objects. Reference data for the ground and off-ground object classes was again manually digitized over the orthomosaic (Figure 5.4h).

5.3.1.3 Lombardia, Italy: The third dataset was obtained over Lombardia, Italy with a Vexcel UltraCam Xp on May 29, 2015. The aerial images were processed to obtain an orthomosaic and DSM with a Ground Sampling Distance (GSD) of 20 cm. A subset of 5000 x 5000 pixels (1000 x 1000 m) was selected for the experimental analyses. The area consists of a residential area, river, dense forests, agricultural fields and a dike (Figure 5.4c,f). A DTM of this area was obtained by the Compagnia Generale Ripreseaeree (CGR S.p.A.) by manually editing the DSM. Therefore, the reference data for the classification part of the experimental analyses was determined by classifying all pixels where the difference between the DSM and DTM was greater than 50 cm as off-ground, and pixels where they were equal as ground. Pixels where the difference was between 0 and 0.5 m were left unlabeled (Figure 5.4i).

5.3.1.4 ISPRS Benchmark Dataset: The proposed method was also tested on the ISPRS 2D Semantic Labelling dataset of Vaihingen[8]. The dataset consists of 33 tiles, for which orthophotos and DSMs with a spatial resolution of 9 cm are provided. Sixteen tiles have reference labels corresponding to six semantic classes: impervious surfaces, buildings, low vegetation, trees, cars and clutter/background. In accordance with the benchmark, a 3x3 erosion filter was used on these reference data to remove border pixels from the quality analyses. It should be noted that while it is useful to test the algorithm using an existing benchmark, it is not the optimal dataset to demonstrate the utility

[7] http://ramanihuria.org/
[8] http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-vaihingen.html

of DTM-extraction techniques such as ours which targets two classes: ground and non-ground. We therefore consider the ISPRS class "impervious surfaces" to equate ground, whereas non-ground consists of the ISPRS classes: buildings, trees, and cars. The ISPRS classes low vegetation and clutter are not considered in our accuracy analysis due to inconsistencies. For example, the "low vegetation" class contains both shrubs (off-ground) and grass (ground) pixels.

5.3.2 Experimental set-up

5.3.2.1 Setting up the proposed network – feature sets, reference labels and dilation: For each of the three datasets, experiments were conducted which trained a FCN with randomly selected patches and utilizing either the true reference labels or the labels assigned through the proposed morphological rule-based method. Ideally, the classification accuracy of the training samples labelled through the proposed rule-based method should approximate the accuracies obtained when using the manually labelled reference data. Furthermore, we motivate the use of dilated filters in the network architecture by providing the results of a FCN in which no dilation is applied in the second convolutional layer.

The parameter values for the rule-based method were tuned on the Kigali dataset, and the same parameters ($w_{small} = 6m$, $w_{big} = 20m$, $\tau_o = 1.0m$, and $\tau_g = 0.5m$) are applied to the Dar es Salaam and Lombardia datasets. Experimental analyses indicated that slight variations in w_{small} (0.2 – 1m), w_{big} (10 – 20 m), and τ_o (0.4 – 1 m) did not significantly change the results for these three datasets. Given the feature sets described in Table 5.2, experiments were performed using FCNs exploiting only the imagery (RGB), only the DSM (Z and nZ), or both imagery and DSM (RGBZ, RGBnZ, RGBDTM, RGBnDSM). Note that the DTM feature sets, obtained by interpolating the elevation values of pixels labelled as 'ground', were calculated separately for both the true reference labels and the labels assigned through the rule-based method. All features were normalized according to the maximum and minimum values of the respective dataset.

For each of these combinations, three folds of 2000 randomly selected patches of 167 x 167 pixels were used to train a FCN using stochastic gradient descent (SGD) with momentum (Krizhevsky, Sutskever and Hinton, 2012) and a batch-size of 32. The networks were trained with a learning rate of 0.0001 for 30 epochs followed by another 10 epochs with a training rate of 0.00001. Weights for all convolutional layers were initialized using the improved Xavier initialization to $\sqrt{\frac{2}{M^2 \cdot K'}}\mathcal{N}(0,1)$ (He *et al.*, 2015). The dropout rate of the final

convolution as 0.5, and a batch size of 32 was used. The network was implemented in MatConvNet[9].

The accuracy assessment is conducted using the mean Producer's Accuracy (mPA) and mean User's Accuracy (mUA), providing the average and standard deviation across the three folds of randomly selected samples. The mPA (Eq 5-4.) and mUA (Eq. 5-5) are calculated using the true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) of the ground class.

$$mPA = \frac{\left(\left(\frac{P}{TP+FN}\right)+\left(\frac{TN}{TN+FP}\right)\right)}{2} \tag{5-4}$$

$$mUA = \frac{\left(\left(\frac{TP}{TP+FP}\right)+\left(\frac{TN}{TN+FN}\right)\right)}{2} \tag{5-5}$$

5.3.2.2 Comparison with deeper network architectures: The proposed method utilizes a much smaller network than those which are proposed for other deep learning tasks. This was a conscious choice, as deeper networks also have more parameters and therefore require more sophisticated hardware and longer training times than smaller networks. In this study, a preference was given to smaller networks, as the network is trained and tested separately for each dataset. However, we provide comparisons with deeper network architectures in order to justify this decision. For this purpose we select three FCNs with Dilated Kernels (DK) which were specifically designed for remote sensing applications (Persello and Stein, 2017). Table 5.3 displays the network architecture, where three networks with varying depths were tested: FCN-DK4, FCN-DK5, and FCN-DK6. These networks have 4, 5, and 6 convolutional layers respectively (e.g. FCN-DK5 contains all of the layers DK1 – DK5 in Table 5.4 but not DK6). Similar to the proposed network, the FCN-DK networks consist of modules of a convolutional layer followed by batch normalization, a non-linear activation function (in this case leaky Rectified Linear Units or lReLU) and a max-pooling layer with no downsampling. Dilated convolutions are used to increase the receptive field size while limiting the number of parameters. Experiments were run using the rule-based reference labels and the RGBnZ feature set and the same three folds of training samples. All networks were trained with 150 epochs at a learning rate of 10^{-6} followed by another 20 epochs with a learning rate of 10^{-7}. Shallower network architectures may require less epochs, but using the same hyper-parameters for all four network architectures enables a fairer comparison.

[9] http://www.vlfeat.org/matconvnet/

Table 5.3: *An overview of the layers for the three FCN network architectures FCN-DK4, FCN-DK5, FCN-DK6.* M *is the filter size in pixels,* d *is the filter dilation in pixels,* K' *is the number of filters,* z *is the padding in pixels, and* r *gives the dimensions of the receptive field in pixels.*

Name	Layer	M	d	K'	z	r
DK1	Convolution1 Batch normalization lReLU	5 x 5	1	16	2	5 x 5
	Pooling1	5 x 5	-	-	2	9 x 9
DK2	Convolution2 Batch normalization lReLU	5 x 5	2	32	4	17 x 17
	Pooling2	9 x 9	-	-	4	25 x 25
DK3	Convolution3 Batch normalization lReLU	5 x 5	3	32	6	37 x 37
	Pooling3	5 x 5	-	-	6	49 x 49
DK4	Convolution4 Batch normalization lReLU	5 x 5	4	32	8	65 x 65
	Pooling4	5 x 5	-	-	8	81 x 81
DK5	Convolution5 Batch normalization lReLU	5 x 5	5	32	10	101 x 101
	Pooling5	5 x 5	-	-	10	121 x 121
DK6	Convolution6 Batch normalization lReLU	5 x 5	6	32	12	145 x 145
	Pooling6	5 x 5	-	-	12	169 x 169
Classification	Convolution7 Dropout Loss	1 x 1	1	2	0	169 x 169

Table 5.4: *An overview of which layers are included in each of the three FCN network architectures FCN-DK4, FCN-DK5, FCN-DK6.*

Name	Included in FCN-		
	DK4	DK5	DK6
DK1	x	x	x
	x	x	x
DK2	x	x	x
	x	x	x
DK3	x	x	x
	x	x	x
DK4	x	x	x
	x	x	x
DK5		x	x
		x	x
DK6			x
			x
Classification	x	x	x

5.3.2.3 Comparison with existing DTM extraction methods: The proposed method is compared to two existing DTM extraction algorithms, namely LAStools[10] which implements a variation of progressive densification DTM extraction (Axelsson, 2000) and gLidar[11] which is based on differential morphological profiles (Mongus, Lukač and Žalik, 2014). There are three parameters for the LAStools implementation: the step size, bulge, and standard deviation. The step size indicates the dimensions of the grid used to select the initial ground samples. This parameter was optimized for each dataset by trying a step size of 5 m to 40 m at 5 m intervals. The bulge parameter refers to the height in meters that the TIN is allowed to go up during the refinement stage. Values from 0.3 to 1.8 m in steps of 0.3 were tested for each dataset. The final parameter refers to the maximal standard deviation for planar patches, values from 0 to 40 centimeters were tested in steps of 10 cm. All parameter combinations were tested for the three datasets, and the combination which maximized the mPA regarding the true reference data are reported. For the gLidar implementation, parameter settings described in the work by Mongus et al. (2014) were set to: $S = 50\,m$, $k = 0.01$, $n = 0.10$, and $b = 0.5$. A detailed description of the meaning of these parameters can be found in the original presentation of the algorithm (Mongus, Lukač and Žalik, 2014). Finally, we also compare the proposed method to the manually generated DTM for the Lombardia dataset. The pixels which were labelled as ground by the proposed method were selected and a bilinear interpolation was performed to construct a DTM. The cumulative error between the predicted and reference DTMs for the pixels labelled as ground are provided.

5.3.2.4 Performance on the ISPRS benchmark: The sixteen labelled tiles of the ISPRS benchmark were used to test the performance of our proposed algorithm. Although good results were obtained with the parameter settings used for the previous datasets ($w_{small} = 6m$, $w_{big} = 20m$, $\tau_o = 1.0m$, and $\tau_g = 0.5m$), the presence of larger buildings and a relatively flat terrain in the ISPRS benchmark dataset caused slightly better results to be obtained with $w_{small} = 3m$ and $w_{big} = 30m$. As the roofing material of these larger buildings was also different from the surrounding buildings, the additional criterion for off-ground samples mentioned in Section 2.1 was implemented. Conform to the benchmark, the User's Accuracy (precision), Producer's Accuracy (recall), and F1-scores are provided as quality metrics for the ground (impervious surfaces) and non-ground (buildings, trees, and cars) classes.

5.3.2.5 Regression-based DTM experiments: An interesting question is whether the proposed method can be altered to directly predict the DTM using regression-based deep learning rather than using first classifying the ground

[10] https://rapidlasso.com/lastools/
[11] https://gemma.feri.um.si/gLiDAR/index.html

pixels and then interpolating the DTM (as proposed above). Such a regression-based method would consist of five steps. The first step is the rule-based identification of ground vs. off-ground samples using the same methodology as defined in Section 2.1. Secondly, a nDSM can be approximated by calculating the difference between the input DSM and an initial DTM obtained by interpolating the pixels labelled as ground in the previous step. The third step then consists of training a regression FCN rather than the classification FCN proposed in Section 2.2. Changing the classification FCN to a regression FCN can be done by replacing the soft-max loss function with a ℓ_2 loss function to minimize the squared Euclidean distance between the height predicted by the network and the nDSM created from the rule-based labels in the previous step. The fourth step then consists of applying this trained (regression) FCN to the entire dataset to obtain a complete nDSM. Finally, the fifth step then consists of subtracting the FCN-nDSM from the input DSM to obtain the DTM of the entire area. This method was tested for the Kigali and Dar es Salaam datasets. The Mean Error (ME) and Root-Mean-Square Error (RMSE) for the entire scene as well as only the ground pixels are presented as quality metrics.

5.4 Results

5.4.1 Feature sets, reference labels and dilation

The results obtained by the proposed FCN according to various combinations of training labels and input channels is presented in Table 5.5 and Figures 5.5-5.7. The first observation is that networks which utilize both image-based and DSM-based input channels (i.e. RGBZ, RGBnZ, RGBDTM and RGBnDSM) outperform networks which utilize only DSM-based (Z, nZ) channels for the Kigali and Dar es Salaam datasets. Using only image-based (RGB) channels as input obtains good results for the Kigali dataset, though the inclusion of elevation information clearly improves the results in Dar es Salaam and Lombardia. When true reference labels are available, the RGBnDSM method has the highest performance. This is logical as the nDSM input channel constructed using the true reference labels essentially defines the height of objects above the ground. However, the nDSM feature constructed using the rule-based training labels is an imperfect representation as these rule-based training labels may be erroneous or incomplete thereby causing the nDSM feature to be inaccurate.

Table 5.5. *The accuracy of the proposed FCN strategies for classifying ground vs. off-ground pixels in the Kigali, Dar es Salaam, and Lombardia datasets. The labels of the training samples are either obtained from the reference data (ref) or the rule-based morphological method (mph) whereas the input feature channels are either derived from the image (RGB), DSM (Z, nZ) or both RGB and DSM (RGBZ, RGBnZ, RGBDTM, RGBnDSM). The average and standard deviation of the mPA and mUA for three folds of randomly selected training data is presented.*

Labels	Features	Mean Producer's Accuracy (%)						Mean User's Accuracy (%)					
		Kigali		Dar es Salaam		Lombardia		Kigali		Dar es Salaam		Lombardia	
ref	RGB	94.8	±0.2	94.7	±2.0	89.5	±0.5	95.2	±0.9	93.6	±2.5	90.0	±0.9
	Z	65.4	±3.0	77.5	±0.5	97.1	±0.3	87.6	±1.5	87.7	±0.9	97.0	±0.2
	nZ	62.2	±4.6	77.9	±3.8	95.3	±1.0	83.8	±3.6	84.9	±0.7	94.4	±2.6
	RGBZ	96.1	±0.7	95.8	±1.2	97.7	±0.0	**97.4**	±0.0	97.2	±0.3	97.9	±0.1
	RGBnZ	94.6	±0.7	96.1	±1.3	97.2	±0.2	96.3	±1.0	98.1	±0.8	97.4	±0.3
	RGBDTM	94.3	±0.8	93.6	±1.7	91.5	±0.6	96.0	±0.5	95.2	±0.7	91.9	±0.2
	RGBnDSM	**97.9**	±0.4	**99.0**	±0.3	**99.1**	±0.4	96.9	±0.6	**98.6**	±0.4	**99.3**	±0.2
ref	RGBnZ (no dilation)	91.9	±1.4	93.5	±1.0	95.5	±0.1	93.9	±0.5	95.6	±0.6	96.1	±0.3
mph	RGB	93.9	±0.3	92.6	±1.2	88.3	±0.2	88.0	±1.0	94.3	±0.2	88.7	±0.2
	Z	81.3	±1.2	75.8	±4.2	**95.2**	±0.2	77.0	±1.5	87.5	±1.5	94.1	±0.3
	nZ	74.3	±0.4	80.6	±3.5	94.1	±0.4	69.9	±1.9	85.0	±1.0	92.3	±0.9
	RGBZ	91.4	±1.3	**94.3**	±1.0	**94.7**	±0.0	88.5	±0.4	95.8	±0.4	93.5	±0.3
	RGBnZ	92.8	±0.3	**95.0**	±0.4	**94.7**	±0.2	83.9	±0.8	95.7	±0.5	**93.7**	±0.2
	RGBDTM	**94.0**	±0.2	91.6	±1.8	89.7	±0.5	**89.3**	±1.7	94.8	±0.2	88.9	±0.4
	RGBnDSM	92.7	±0.3	92.9	±1.5	93.9	±0.0	87.0	±0.2	**95.9**	±0.2	92.3	±0.2

Figure 5.5: *Classification maps of the Kigali dataset for the rule-based training labels (a), FCN-RGBnZ (b), gLidar (c) and Lastools (d).*

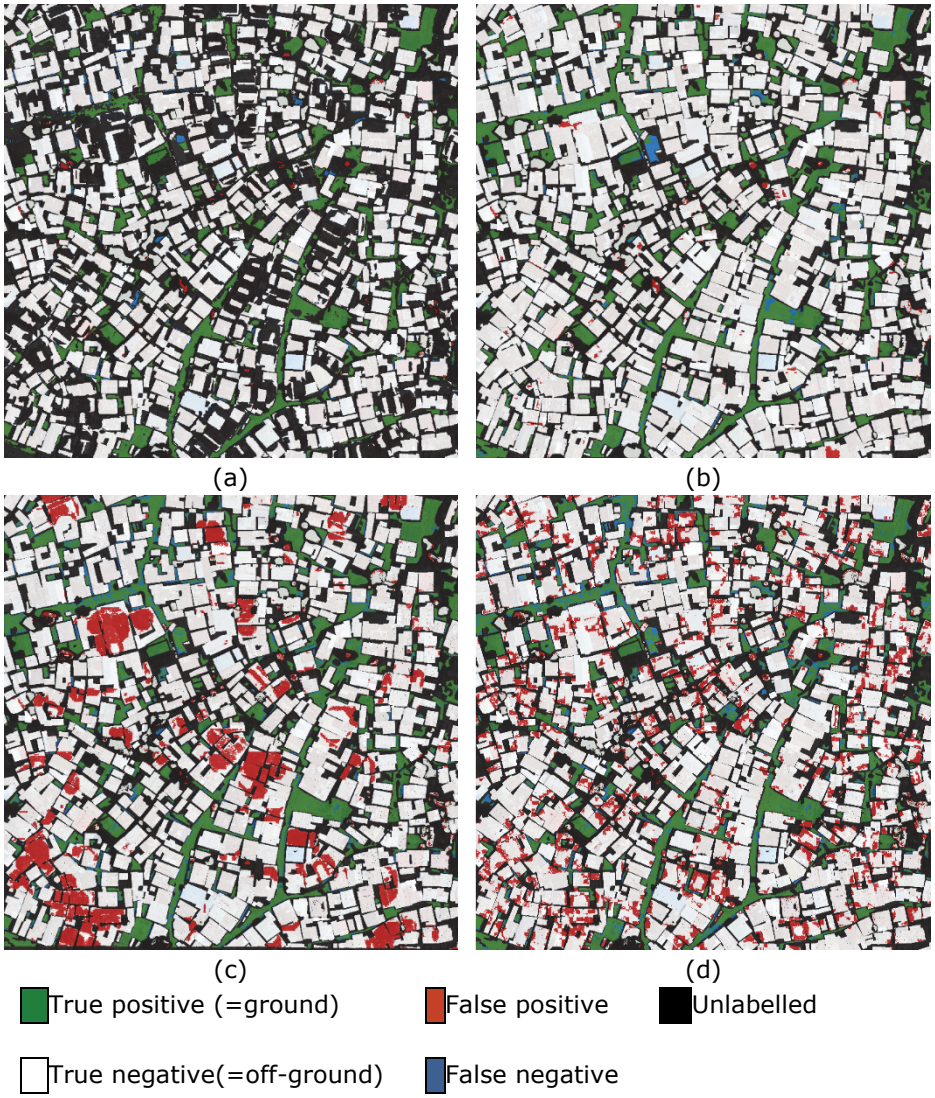


Figure 5.6: *Classification maps of the Dar es Salaam dataset for the rule-based training labels (a), FCN-RGBnZ (b), gLidar (c) and Lastools (d).*

Figure 5.7: *Classification maps of the Lombardia dataset for the rule-based training labels (a), FCN-RGBnZ (b), gLidar (c) and Lastools (d).*

Rather, the RGBDTM input channels achieve the highest mPA when using the rule-based training labels for the Kigali dataset. In this case, using only image-based features (RGB) works quite well for the Kigali dataset which may be due to the fact that the ground and elevated objects are more easily distinguished using spectral features in this dataset and that the topographic information is less informative due to the steep slopes in the area. Some of the errors in the top left corner (Figure 5.5b) are due to inconsistencies in the UAV flight operations, resulting in a blurring of the orthomosaic and a loss of texture. Previous research indicated that texture was an important cue for

distinguishing building roofs from ground (C. Gevaert *et al.*, 2016). One of the assumptions of our method is that the pixels along the edges of contiguous elevated objects will have a similar appearance as the central parts of those objects. This example in the top-left part of the Kigali dataset is a case where this assumption does not hold, as some pixels in the central parts of the contiguous buildings have a blurred texture (unlike the pixels along roof edges). This may cause errors in the classification results and interpolated DTM. The RGBnZ works best for the Dar es Salaam and Lombardia datasets. For the Lombardia dataset, using the Z channels as input for the FCN slightly outperforms the sets using both image-based and DSM-based combinations. Most of the errors in the Lombardia dataset are due to the assignation of incorrect labels to an elevated road during the rule-based label assignation which are used to train the FCNs (Figure 5.7a), causing systematic mislabeling of this road as off-ground (Figure 5.7b). Furthermore, there are some errors in the vegetation in the lower left corner, where errors in the rule-based labels caused by systematic tree height differences are propagated in the classification and interpolated DTM. The proposed FCN-RGBnZ method performs better than gLidar in these areas, although Lastools appears to perform best in this particular situation. The large extent of contiguous off-ground objects (forests) and relatively flat terrain in the Lombardia dataset suggests that increasing w_{big} could achieve better results.

These results indicate that although there are slight differences according to the scene characteristics of the various datasets, the RGBnZ input channels generally achieve a high and reliable classification accuracy when using the rule-based initialization of training labels. Indeed, some errors in the initial labelling of the Kigali dataset (Figure 5.5a) are corrected in the FCN-RGBnZ output (Figure 5.5b). This indicates that the proposed FCN does more than 'fill in the gaps' by relearning the top-hat heuristic used to generate the training labels. We furthermore see that for all three datasets, the network which does not include the dilation in the convolutional layers performs worse than the proposed network when using RGBnZ features. Using this proposed strategy with rule-based training labels RGBnZ features and dilated convolutional layers, we can accurately classify ground vs. off-ground objects with an mPA of 92.8% to 95.0% and a mUA of 83.9% to 93.7% for the three datasets. These results, which exploit simple rules to label the training samples, have an mPA of only 1.8% (Kigali), 1.1% (Dar es Salaam), and 2.5% (Lombardia) lower than FCNs trained using manually-labelled training samples.

5.4.2 Comparison with deeper network architectures

Results indicate that adding additional convolutional layers in this application does not lead to an increased accuracy. Table 5.6 displays the average accuracies obtained for each of the three folds of the Kigali, Dar es Salaam, and Lombardia datasets. The mean producer's accuracy remains around 93.4%

for Kigali, 95.8% for Dar es Salaam, and 94.8% for Lombardia. The differences in the accuracies reported in Table 5.6 and Table 5.5 are due to changes in the training rate and number of epochs. Table 5.7 presents the number of false positives and negatives. The networks generally show similar tendancies for the three datasets – Kigali has a larger number of false positives than false negatives, whereas Lombardia has relatively more false negatives. The additional depth of FCN-DK4, FCN-DK5 and FCN-DK6 comes with higher computing requirements, as illustrated in Table 5.8. The FCN-DK6 network has 120 000 parameters which require 462 KB of memory which takes around 5.42 hours to train. However, the FCN-RGBnZ network requires only 23 000 parameters which require 90 KB of memory and 1.92 hours of traning time. The smaller network can achieve a slightly higher accuracy than the deeper architectures in only 35% of the time.

Table 5.6: *The OA, mPA and mUA of FCN-RGBnZ (the proposed network), FCN-DK4, FCN-DK5, and FCN-DK6 for Kigali (K), Dar es Salaam (D), and Lombardia (L).*

FCN Network	OA (%)			mPA (%)			mUA (%)		
	K	D	L	K	D	L	K	D	L
FCN-RGBnZ (proposed)	93,5	97,6	95,1	93.5	95.9	94.9	83.7	95.3	94.2
FCN-DK4	93,4	97,8	94,7	93.4	96.0	94.7	83.5	95.9	93.7
FCN-DK5	93,6	97,8	94,8	93.6	95.6	94.7	83.9	96.3	93.9
FCN-DK6	93,2	97,9	94,7	93.2	95.6	94.7	83.0	96.4	93.8

Table 5.7: *The number of false negatives and false positives of FCN-RGBnZ (the proposed network), FCN-DK4, FCN-DK5, and FCN-DK6 for the three datasets.*

FCN Network	Kigali		Dar es Salaam		Lombardia	
	FN	FP	FN	FP	FN	FP
FCN-RGBnZ (proposed)	152887	979921	269509	345090	752471	456848
FCN-DK4	142741	1005187	278738	284432	845153	455564
FCN-DK5	152936	961631	312670	241046	800301	468322
FCN-DK6	155716	1040235	320196	226908	832636	456014

Table 5.8: *Characteristics of the four FCN network architectures.*

FCN Network	FCN-RGBnZ	FCN-DK4	FCN-DK5	FCN-DK6
Number of parameters	23 000	67 000	92 000	120 000
Memory requirement for parameters (KB)	90	260	361	462
Average training time (hours)	1.92	3.35	4.33	5.42
Final receptive field size (pixels)	57 x 57	81 x 81	121 x 121	169 x 169

5.4.3 Comparison with existing DTM extraction methods

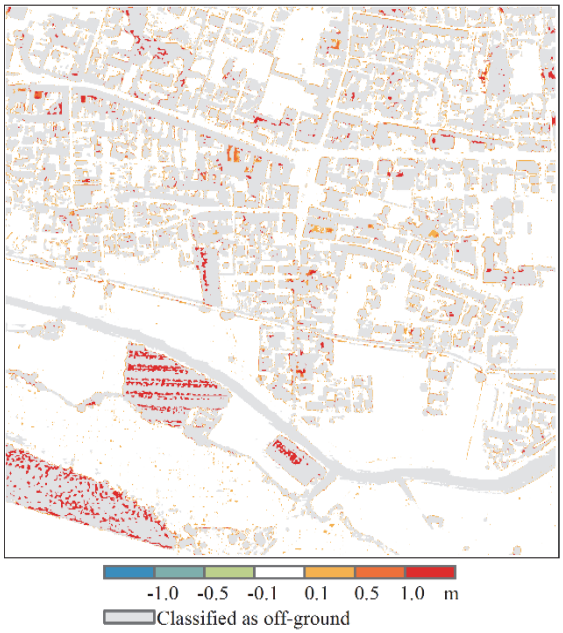
The proposed method also clearly outperforms the reference methods both visually (Figures 5.5-5.7) and quantitatively (Table 5.9). Note that two accuracy measures are provided for the rule-based labels in Table 5.9. As the morphological selection method does not label the entire image, we provide the accuracy of the labeled samples, and the accuracy where unlabeled samples are considered as errors in parentheses. The proposed method outperforms the reference methods for all three datasets with a single exception. LAStools slightly outperforms the automated method for the Lombardia dataset. However, it should be noted that the LAStools parameters were optimized separately for each dataset to maximize the accuracy on the testing data, whereas the proposed method utilized the same parameters for all datasets and is therefore more easily implemented in automatic workflows. The proposed method outperforms LAStools in the Kigali (increasing the mPA by 8.4% and mUA by 16.6%) and Dar es Salaam (increasing the mPA by 11.2% and mUA by 21.6%) datasets. In the Kigali dataset, both LAStools and gLidar clearly suffer from the steep slopes in the lower half of the image, where parts of the roofs are misclassified as terrain (Figure 5.5c,d) in a clear example of the problem illustrated in Figure5.2e. This effect is clearly lower using the proposed FCN-RGBnZ method, illustrating the importance of including RGB information in areas where the surface topography is complicated (Figure 5.5b). Indeed, when using only the height information (i.e. Z and nZ feature sets in Table 5.5), LAStools and gLidar outperform the FCN in Lombardia and have a higher mPA in Kigali and Dar es Salaam. In the Dar es Salaam dataset, contiguous roof-tops (Figure 5.2 g,h) appear to cause many problems errors for gLidar and LAStools (Figure 5.6c,d). In the Lombardia dataset, the proposed method outperforms the two reference methods in the correct classification of forested areas as off-ground. These areas in the bottom left and top right corners of the image are clearly visible as false positives in the gLidar results (Figure 5.7.c). However, the Lombardia dataset also clearly illustrates how samples on the elevated road crossing the center of the dataset were mislabeled in the first rule-based step (Figure 5.7a), causing systematic errors in the prediction map obtained by the proposed method (Figure 5.7b).

***Table 5.9**: The mPA and mUA of LAStools, gLidar, the rule-based labels (Step 1), and FCN-RGBnZ (Step 2) for Kigali (K), Dar es Salaam (D), and Lombardia (L). For the rule-based labels, we provide the mPA of the training samples which were labeled, and the mPA penalizing unlabeled pixels as classification errors in parentheses.*

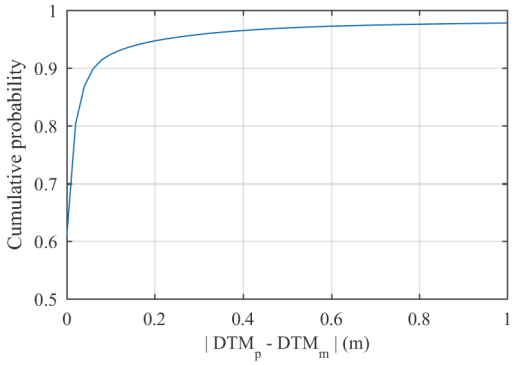
DTM extraction algorithm	mPA (%)			mUA (%)		
	K	D	L	K	D	L
LAStools	84.4[1]	83.8[2]	98.1[3]	67.3[1]	74.1[2]	97.7[3]
gLidar	85.3	85.7	93.1	66.2	75.5	94.4
Rule-based labels (Step 1)	95.1 (71.3)	96.2 (68.6)	97.7 (75.5)	90.8 (56.6)	97.4 (65.3)	97.4 (96.6)
FCN-RGBnZ (Step 2)	92.8	95.0	94.7	83.9	95.7	93.7

[1] The best LAStools results for the Kigali dataset were obtained using a step of 20 m, bulge of 0.3 m, and standard deviation of 30 cm.
[2] Using a step of 20 m, bulge of 0.3 m, and standard deviation of 40 cm.
[3] Using a step of 40 m, bulge of 1.8 m, and standard deviation of 0 cm.

A comparison between the DTM obtained through the proposed method and a manual editing is provided in Figure 5.8. Considering only areas labelled as ground by the proposed method, there was a mean error of 0.16 m and a mean absolute error of 0.18 m compared to the manually edited DTM. This indicates that there is a small bias of less than one GSD in results of the proposed method, which is slightly higher than the reference DTM provided. 93.1% of the pixels have an absolute difference of less than 10 cm in the two DTMs – which is less than half the GSD – and 96.9% have an absolute difference of less than 0.5 m (Figure 5.8b).



(a)



(b)

Figure 5.8: *A visualization of the predicted DTM (DTM_p) minus the manual DTM (DTM_m) for the Lombardia dataset (a), and the cumulative probability of this difference for pixels classified as ground by the proposed algorithm (b).*

5.4.4 Results on the ISPRS benchmark dataset

Table 5.10 displays the quantitative accuracies of FCN-RGBnZ applied to the ISPRS benchmark. Impervious surfaces are classified as ground with a User's Accuracy of 92.2% and a Producer's Accuracy of 74.5%. The three ISPRS classes of buildings, trees, and cars are classified as off-ground objects with a User's Accuracy of 87.4% and Producer's Accuracy of 96.9%. These results

indicate that there are more false negatives than false positives in the results, which can also be observed visually (see Figure 5.9). Some of these errors can be attributed to inconsistencies in the benchmark labels. For example, the central area of Figure 5.9c indicates false positives in the central area, where the ISPRS reference label is tree (Figure 5.9a). However, a visual analysis of the image (Figure 5.9b) suggests that these pixels could indeed be ground in between the trees. The results in Table 9 indicate a relatively large error due to pixels labelled as impervious surfaces to be classified as off-ground (i.e. false negatives). A visual analysis of the results indicates that such false negatives (Figure 5.9g) often occur in shadowed streets, where the reference label indicates impervious surface (Figure 5.9f), but the DSM actually shows relatively high elevation values (Figure 5.9e) and there are few visual cues in the image due to the shadows (Figure 5.9d). The different semantic labels and inconsistencies between the reference labels and input data make it difficult to compare the results of the FCN-RGBnZ method proposed for DTM extraction with the other contributions to the ISPRS benchmark.

***Table 5.10**: The User's Accuracy (=precision), Producer's Accuracy (=recall), and F1-scores for the FCN-RGBnZ algorithm applied to the ISPRS benchmark dataset. The top row presents the average percentage for all sixteen tiles, the rows below indicate the results of a tile with a high accuracy and lower accuracy.*

	Ground (impervious surfaces)			Off-ground (buildings, trees, and cars)		
	UA (%)	PA (%)	F1 score	UA (%)	PA (%)	F1 score
All tiles with reference labels	92.2	74.5	82.0	87.4	96.9	91.8
Tile with high accuracy (N° 34)	92.9	88.1	90.4	95.7	97.1	96.1
Tile with lower accuracy (N° 21)	91.0	71.5	80.1	87.1	96.5	91.5

DTM extraction algorithm	mPA (%)			mUA (%)		
	K	D	L	K	D	L
LAStools	84.4[1]	83.8[2]	98.1[3]	67.3[1]	74.1[2]	97.7[3]
gLidar	85.3	85.7	93.1	66.2	75.5	94.4
Rule-based labels (Step 1)	95.1 (71.3)	96.2 (68.6)	97.7 (75.5)	90.8 (56.6)	97.4 (65.3)	97.4 (96.6)
FCN-RGBnZ (Step 2)	92.8	95.0	94.7	83.9	95.7	93.7

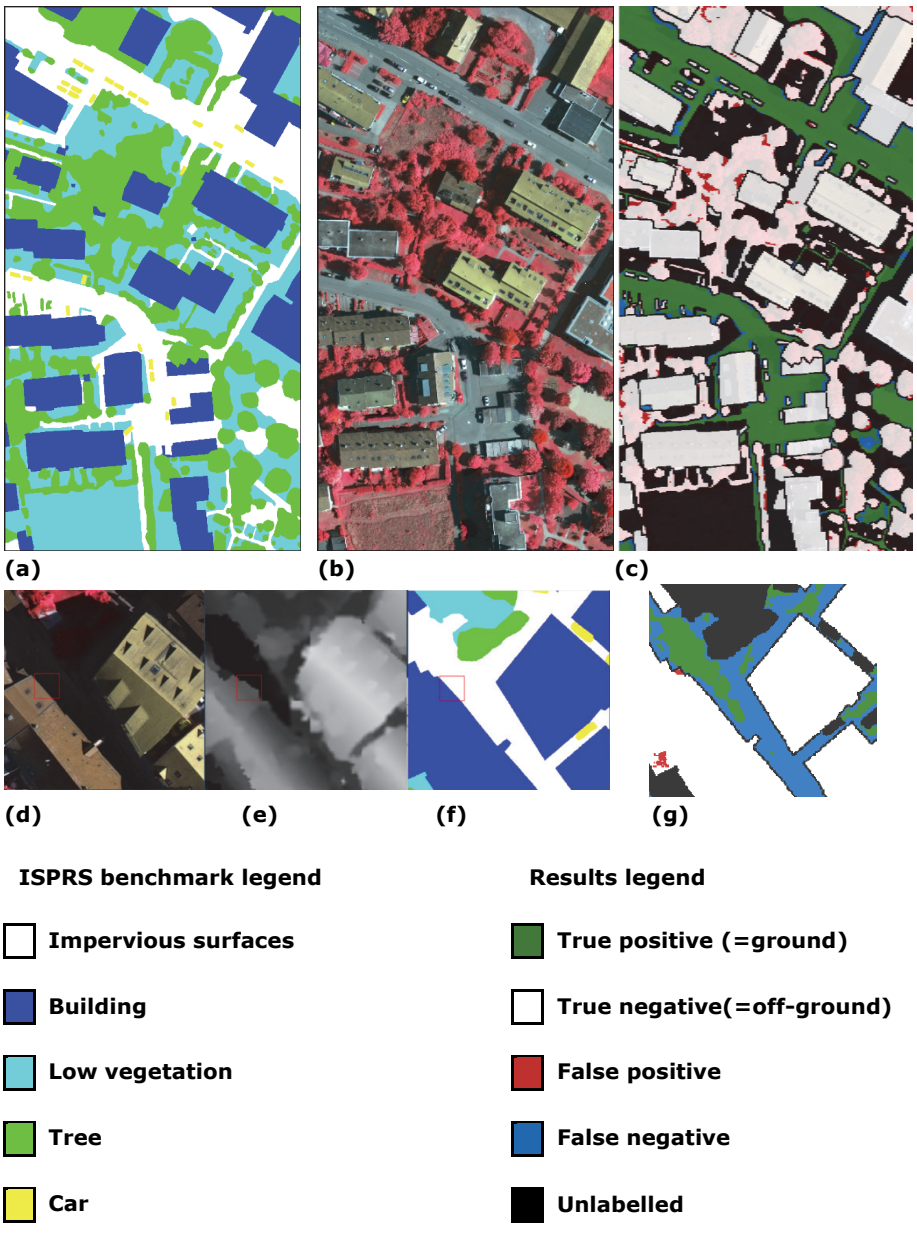


Figure 5.9: *Input ISPRS reference labels (a) and false-color images (b), and the FCN-RGBnZ results (c) of tile 34. The bottom row presents an example of causes of false negatives in tile 05. Note the narrow streets which are labelled as impervious surfaces in the reference data (f), but are classified as off-ground by our algorithm (g) due to the combination of shadows in the imagery (d) and elevated values in the DSM (e).*

5.4.5 Results of the regression-based DTM experiments

The error metrics in Table 5.10 indicate that the nDSM returned by the regression-based FCN are an average of 23 cm higher in the Kigali dataset than the reference nDSM values. This is 46 cm in the Dar es Salaam dataset. One difficulty in DTM prediction is that it isn't clear which 'terrain' height to assign to the terrain under building located on a slope. I.e. would it be correct to interpolate the height of the surrounding terrain, or should we assume the floor is flat and assign the elevation of the lowest floor to the entire building footprint? Due to such confusions, we also include error metrics of the nDSM predictions for pixels *labelled as ground* in the reference data. Table 5.10 indicates that the ME of the ground pixels is actually much higher than the global average, overestimating the reference elevation data by 1.74 m in Kigali and 2.18 m in Dar es Salaam. The RMSE of the ground pixels is also higher than that of the entire dataset in both cases. Further investigations indicated that although the average nDSM values of the predicted and reference datasets were similar, the variance of the predicted nDSM was much lower than that of the reference data. In essence, all height values are therefore closer to the mean nDSM value of the dataset. This in turn causes the overestimation of the height of ground pixels.

Dataset	ME[1] (m) Entire dataset	RMSE (m) Entire dataset	ME (m) Only ground	RMSE (m) Only ground
Kigali	0.23	1.62	1.74	1.85
Dar es Salaam	0.46	1.53	2.18	2.21

[1]The ME is calculated as the predicted nDSM minus the reference nDSM. Positive values therefore indicate that the predicted elevation overestimates the reference values.

Figure 5.10: *ME and RMSE of the nDSM predictions obtained with the regression-based FCN calculated over the entire dataset (i.e. both ground and off-ground objects) or only the pixels labelled as ground. All values are in meters.*

5.5 Discussion

In UAV applications, variations in flight heights and camera parameters are likely to cause a wider diversity in the spatial resolution of datasets. The representation of off-ground objects in datasets with a spatial resolution of 3 cm, 5 cm, or 20 cm for example, will be quite different. It is unclear how this wide variation of spatial resolution in UAV datasets will influence the parameters learned by a FCN. Hu and Yuan (2016) address this problem by summarizing the elevation information contained by point clouds in a grid of a fixed spatial resolution. Although this allows the utilization of a single FCN trained for various study areas and datasets, this strategy will not exploit the

full information contained in a dataset which has a higher point density (or spatial resolution) than the trained network. Therefore, an important characteristic of the method proposed here is that it demonstrates that it is feasible to train and apply a FCN on each dataset independently. The present manuscript demonstrates this using datasets from UAV or aerial imagery, but the method is not limited to these types of images. It would also be possible to apply the proposed method to satellite imagery with a lower spatial resolution and a larger extent. Applications which intend to cover a larger extent may benefit from larger sample sizes to train the network – this stresses the advantage of using the rule-based strategy to provide labels for training. The labelling and selection of training samples is completely automated in the proposed methodology. It is therefore in principle extendible to very large datasets. Furthermore, although training is time-consuming, FCNs are very fast in the testing phase and would therefore be a viable option in the classification of ground over large extents.

Although the point clouds obtained from dense matching tend to contain more random noise errors than LiDAR point clouds (Nex and Gerke, 2014), the simultaneous acquisition of both elevation and radiometric information can be seen as an advantage of UAV datasets and aerial photogrammetry in general. Making use of the complementary information in imagery may help distinguish ground from off-ground areas when the elevation information itself is not sufficient. However, it is important to note that the rule-based selection of training samples is only an estimation, and that mislabeled samples may cause systematic errors in the output of the FCN. For example, when considering scenes with steep slopes where ground and off-ground objects present a step-like pattern (i.e. Figure 5.2e), the rule-based selection of training samples based on morphological filters will not be able to distinguish between ground and off-ground objects. However, if such geometrically ambiguous areas form a minority in the dataset, then a sufficient number of correct ground vs. off-ground training samples can be collected. If a sufficient number of correct samples are captured and utilized to train the supervised classifier, and presuming the radiometric information from the imagery is capable of distinguishing between the ground and off-ground objects, then these initial errors may be corrected in the second step.

This second step, the exploitation of deep learning methods refers to a field of research which is currently developing rapidly. It is likely that emerging network architectures developed in the near future may further increase the accuracy of the proposed method. However, the observations of the present paper may serve to guide users towards the selection of a suitable network architecture. Firstly, one important issue is the redundancy of calculations when performing a pixel-based classification or semantic segmentation as it is known in the computer vision community. This motivated the selection of a

FCN architecture rather than a CNN architecture in the current paper. Other emerging options such as PixelNet (Bansal *et al.*, 2017) could be considered in the future. Similarly, due to the high spatial resolution compared to the size of off-ground objects, it is important to increase the receptive field of the network. In the present case, this was done through the use of dilated filters and adjusted DSM features. Alternative strategies could include multi-scale approaches (Farabet *et al.*, 2013) or skip-architectures (Song, Herranz and Jiang, 2017). Thirdly, the depth of the network architecture, or number of layers should also be considered. In general, the success of deep architectures may be attributed to their ability to learn complex patterns in very difficult classification tasks.

Network architecture may also be one of the underlying reasons behind the high errors obtained in the regression-based experiments. The main problem was the lower variance of the output nDSM predictions, causing the elevation of ground pixels to be overestimated. One hypothesis is that this has to do with the ℓ_2 loss function which penalizes outliers. In the case of DTM extraction, small errors overestimating the height of bare ground may be more concerning than larger errors underestimating the height of buildings. Further experiments could try using other loss functions such as the Huber loss (Huber, 1964) or Tukey's biweight function (Belagiannis et al., 2015) which others have found to be less sensitive to outliers when tuning deep regression networks. Another strategy could be the introduction of skip connections, which proved to be key to obtaining realistic height estimations from monocular imagery (Mou and Zhu, 2018). Further experimental analysis could focus on such direct height estimations to complement classification-based DTM extraction techniques such as the methodology proposed here.

In this application of DTM extraction, we are not interested in separating numerous abstract classes associated with complex appearance features like in other computer vision problems. In the considered application, the network should be able to capture features from both ground and non-ground, integrating radiometric and geometric variables. Results show that shallow networks with large receptive fields perform as good or better than deeper networks. On the other hand, shallower networks have less parameters, are easier and faster to train, less prone to overfitting, and generally more robust to different radiometric/geometric characteristics of the data set. The applicability of developments regarding the further reduction of parameters in deep learning networks could be analyzed in future works.

Furthermore, the DTM extraction algorithm proposed here has been designed to be trained and tested on a single dataset – focusing on UAV datasets which may have a limited extent and therefore limited number of training samples. If one were to instead combine UAV data from a large number of sources,

which may become feasible in the near future due to the wider availability of UAV imagery, e.g. OpenAerialMap, using a deeper network architecture trained on all of these images could be an alternative strategy. In this case it would be important that the selected datasets represent challenging situations (such as those depicted in Figure 5.2) in order to ensure that the network is able to handle them. Again, it depends on whether the user would like to have a quick DTM extraction tailored specifically to a single (UAV) dataset (i.e. the purpose of the current manuscript), or a general deep network trained applicable to a larger spatial extent.

Finally, another important consideration is how to assess the quality of a DTM extraction algorithm. In this case, we define the DTM as a classification problem, similar to Sithole and Vosselman (2004). Other studies use the vertical accuracy of a DTM compared to Ground Control Points (GCPs) collected in the field with GPS (Höhle and Höhle, 2009; Hugenholtz *et al.*, 2013). However, we should remember that the final product is an interpolated DTM surface. As such, false positive rates which introduce errors into the interpolation could be more malign than false negatives which lower the detail of the reconstructed surface. Further research could consider how to assess the quality of DTM extraction methods without the presence of alternative DTMs or the costly collection of GCPs in the field.

5.6 Conclusions

Existing algorithms for DTM extraction still face difficulties due to data outliers and geometric ambiguities of the scene due to contiguous off-ground areas or sloped environments. This work postulates that in such cases, the radiometric information contained in aerial imagery may be leveraged to distinguish between ground and off-ground objects. This is particularly relevant for, but not limited to, UAV datasets which simultaneously acquire both elevation and radiometric information.

The proposed method uses two simple rules based on morphological filters to select examples of ground and off-ground objects using the DSM. The underlying idea is not to use these rules to label the entire dataset, but rather to select reliable samples which together describe the variability in the geometric and radiometric attributes of both classes. These samples are then used to train a supervised classifier, which labels each pixel in the entire scene and may correct errors in the initial labelling. We propose using a FCN, as deep learning methods are currently state-of-the-art in supervised classification problems. Improvements to deep learning methods are rapidly evolving, therefore it is plausible that the network architecture presented here could be improved according to the continued developments in this field.

In this research we address a number of issues which are important when adapting deep learning methods to DTM extraction. Firstly, we bypass the costly requirement of large amounts of training data by employing simple rules to automatically select and label representative samples from the dataset itself. By training the FCN for each dataset, we can account for both differences in the spatial resolution of different datasets as well as the natural variability of objects in different parts of the world. Secondly, we illustrate how FCNs can be adapted to consider the topographical variations over a larger area without increasing the computational complexity of the algorithm. This is done both by considering dilated filters in the network architecture and through the inclusion of feature channels which summarize variations in the elevation over larger areas.

The proposed method is successfully tested using three photogrammetric datasets with different spatial resolutions and covering scenes containing areas which challenge DTM extraction methods, as well as the ISPRS benchmark dataset. The datasets used for testing are relatively small but the results can easily be applied to larger study areas, or imagery and DTMs with a lower spatial resolution. We demonstrate the improvements of the proposed method with respect to two reference DTM extraction algorithms.

Chapter 6 – Opportunities for UAV Mapping to Support Unplanned Settlement Upgrading[12]

[12] This chapter is based on:

Gevaert, C.M., Sliuzas, R., Persello, C., and Vosselman, G. (2016) 'Opportunities for UAV mapping to support unplanned settlement upgrading', *Rwanda Journal*, 1(1S), doi:10.4314/rj.v1i1S.4D

Abstract

The effort to improve sub-standard living conditions in unplanned settlements is often hindered due to a lack of adequate spatial information describing the baseline situation and changes occurring during and after the upgrading process. Low-cost Unmanned Aerial Vehicles (UAVs) could provide very detailed, up-to-date spatial information for small unplanned areas as and when required. To investigate the utility of such platforms in settlement upgrading, UAV flights were conducted over approximately 150 ha of unplanned settlements in the City of Kigali in May and June 2015. These activities were supplemented by an analysis of the spatial information needs of various stakeholders involved in the upgrading project. In the context of the upgrading project, the UAV imagery has four significant benefits: it could replace the 2008 25 cm ortho-imagery by up-to-date 3 cm imagery in current workflows for map updating. Moreover, it enables the extraction of additional information which was previously unavailable, such as detailed elevation data to support surface water runoff analysis and drainage capacity calculations. Additionally, it speeds up field work and provides a foundation for communication between different stakeholders.

When using UAVs it is also important to take many practical considerations, as well as the societal and ethical contexts, into account. The technological limitations, the requirement for specialized knowledge, and heavy computing requirements of data processing are factors to be addressed when using UAV technologies in this setting. First experiences in Kigali have indicated that while not generally perceived as a problem by the local population, fear of forced displacement and expropriation may raise concerns amongst the residents. Communication with the population before and during flights, and sharing the benefits of the acquired information are important to mitigate these fears. Moreover, the resolution and quality of the images is such that privacy concerns and issues such as their potential to be used to the detriment of residents of such areas should not be ignored.

6.1 Introduction

Spatial data is considered essential for unplanned settlement upgrading projects (Abbott, 2002; Kohli *et al.*, 2013; Taubenböck and Kraff, 2014). Obtaining an accurate base map of these areas provides a sound basis for designing technical interventions (Paar and Rekittke, 2011; UN-Habitat, 2012), as well as improving the communication between stakeholders (Barry and Rüther, 2005), and empowering local authorities and communities (Abbott, 2003).

Remotely sensed imagery is a key source of spatial information, as it can provide an objective, up-to-date overview of the physical situation in the settlement (Taubenböck and Kraff, 2014). Seven important roles of satellite imagery for unplanned settlement management can be identified: identification of unplanned settlements, identifying changes in the boundaries of these settlements over time, generation of surface data, land use classification, extraction of buildings and other objects for mapping purposes, and reconnaissance (Mason and Fraser, 1998). However, small buildings and narrow footpaths characteristic of unplanned settlements may hinder the interpretation of commercial satellite imagery with half-meter resolution (Kuffer, Barros and Sliuzas, 2014).

UAVs, also known as drones, Unmanned Aerial Systems (UAS) or Remotely Piloted Aircraft Systems (RPAS), are defined as small aircraft operated without an onboard pilot (Nex and Remondino, 2014). Similar to traditional aerial image acquisition, a UAV is mounted with a camera which takes images of the study area as it flies over. The individual images can be stitched together to create a 3D model in the form of a point cloud, as well as obtain a high-resolution Digital Surface Model (DSM) and orthomosaic. The orthomosaics obtained from the UAV imagery can reach a resolution of a few centimeters (Nebiker *et al.*, 2008). Although this is similar to the resolution which can be obtained by aerial photography, UAVs cost less and, at least for relatively small areas, are more flexible in acquiring data (Nex and Remondino, 2014). As such, this possibility of obtaining high-resolution spatial data in a relatively cheap and dynamic manner could be a practical approach to provide essential baseline information in unplanned settlement upgrading projects.

However, for a UAV to be useful for unplanned settlement mapping, the workflow must not only meet the technical requirements of the user, but the new technology must fit into the local context (Pannell *et al.*, 2011) and its use should be ethical and provide adequate protection of the privacy of those whose properties and even their bodies, are recorded on the images. It is therefore important to analyze the spatial information requirements of the potential end-users as well as the social context. The main users of UAVs for

unplanned settlement mapping are likely to be governmental institutions or organizations operating on their behalf as partner or consultant. Unfortunately, studies from cities in six different developing countries indicated that spatial data collection and products are often restrained to experts in the governing body and private sectors and that data sharing is limited (Baud *et al.*, 2014). Dependency on spatial technology may therefore increase social exclusion (Pfeffer *et al.*, 2013). It is important to analyze how the benefits of the UAV imagery may not only serve the governing bodies, but also how they may be distributed amongst stakeholders, in particular the local population as in the case of informal settlements they tend to be the most disadvantaged and vulnerable members of society

The objective of this paper is to analyze the potential of UAVs to support unplanned settlement upgrading projects, to describe the first perceptions of various stakeholders, to identify important factors which should be taken into account for the diffusion of this technology and its benefits amongst stakeholders, and to underline the ethical concerns of privacy and possible misuse of the obtained information. The paper is based on the results of UAV image acquisition in the context of upgrading projects in the City of Kigali, Rwanda in May/June, 2015. After a brief introduction to the UAV data acquisition workflow, this paper provides an overview of the context of the case study, and the UAV flights conducted over the area. Next, the spatial data requirements of upgrading projects are analyzed, and four key benefits of using UAV data are identified. The practical considerations and social and ethical context are then described. Finally, a discussion leads to the identification of key factors which should be taken into account to support the diffusion of UAVs and their related information extraction techniques, for unplanned settlement upgrading.

6.2 UAV data acquisition workflow

The process of using UAV for mapping purposes has been well documented (Colomina and Molina, 2014; Nex and Remondino, 2014). In general terms, we could summarize that the UAV information acquisition workflow consists of: (i) flight planning and execution, (ii) ground control point (GCP) acquisition, and (iii) data processing. The flight planning consists of, for example, selecting the UAV platform and defining the flight parameters. Depending on the available budget, project area and required image resolution, a suitable UAV platform and payload can be selected. For example, a fixed wing (airplane-like) UAV may cover a larger area with one flight, but requires an extended take-off and landing zone. On the other hand, rotary wing systems which are powered by vertically oriented propellers, or rotors, require more battery power which reduces their flight time. However, rotary-wing UAVs are capable of taking off and landing vertically, and are more agile in their image acquisition. Some

UAVs have a fixed payload (such as a built-in camera), whereas others allow the user to change the payload, thus giving the user more control of the spectral or spatial resolution of the acquired data products. After the platform is selected, the flight path must be defined. Many consumer-grade UAVs are capable of automatic flight planning, where a grid pattern is automatically generated from the flight height and overlap defined by the user. The flight height and payload will influence the spatial resolution of the output orthomosaic, whereas the overlap influences the quality of the data products. Finally, before executing flights one must first check the UAV flight regulations. In Rwanda, flight permission falls under the domain of the Rwanda Civil Aviation Authority (RCAA). At the time of publication, UAV regulations for the country of Rwanda are being drafted.

Secondly, most UAV systems require GCPs to improve and verify the geometric accuracy of the orthomosaic, DSM, and point cloud obtained from the UAV images. The precision of the GPS utilized must conform to the high spatial resolution of the UAV data products. The GCPs may be collected during the UAV flights by placing targets on the ground before the flights so they are visible in the UAV images, and measuring their coordinates. Alternatively, GCPs can be added afterwards by obtaining the coordinates of permanent structures which can also be identified in the UAV imagery.

Finally, the data processing stage converts the raw images into useable data products. There are a number of documents which provide theoretical (Hartley and Zisserman, 2003) or more practical (Nex and Remondino, 2014) explanations of this process. First, a unique descriptor is used to identify pixels from various images which represent the same object. After these so-called tie points are listed, the camera calibration and image orientation are performed through a bundle block adjustment. This identification of the interior and exterior image parameters is also known as "structure from motion". In this step, the GCPs can be added to improve the geometrical accuracy and verify the quality of the parameter estimations. The sparse 3D point cloud obtained from this step is then enriched through dense matching algorithms (Remondino *et al.*, 2014). Next, a DSM is interpolated from this point cloud. The original images are stitched together, using the elevation information from the DSM to form an orthomosaic. Various photogrammetric software are available which can execute the process from images to dense point cloud, DSM, and orthomosaic semi-automatically. Comparisons of such software have been researched (Remondino *et al.*, 2014; Sona *et al.*, 2014).

6.3 Study Area

In 2000, the Government of Rwanda established the Rwanda Vision 2020 (Rwanda, 2000). Kigali played a key role in this plan, which aimed to modernize the city and transform it into an important global city. A Conceptual Master Plan for the City of Kigali was developed in 2007 and updated by a detailed Kigali City Master Plan (KCMP) in 2013. In this plan, some unplanned settlements must be moved to make space for business districts, while others should be improved to meet the Kigali City Master Plan guidelines. A partnership was formed between the Rwanda Housing Authority (RHA), Affordable Housing Unit of the City of Kigali One Stop Center (CoK-OSC), and the Nyarugenge District OSC to conduct an upgrading project in Nyarugenge District. It is a pilot project which aims to not only identify infrastructure improvements in the area, but also to develop successful strategies for participatory engagement and designing slum upgrading projects in Kigali and the secondary cities of Rwanda. The study area covers parts of Agatare, Rwampara, Biryogo and Kiyovu cells of the Nyarugenge District in Kigali. The project targets issues including: storm water drainage, sewerage, drinking water supply, roads, electricity and public lighting, and housing improvement. The project area is roughly 86 ha in size, with an estimated 3,977 households and 18,914 individuals (GISTech Consultants LTD, 2015). The houses are generally made of mud and wood with corrugated iron roofs. A few localized improvements have previously been made regarding access and drainage, but the current project aims to provide a more comprehensive improvement. The project started on in December 2014 and the intervention design and cost plans were provided at the end of June, 2015. As of November 2015, activities in the project area include environmental and social safeguards reporting and a detailed design of the prioritized interventions.

In May 2015, a number of UAV flights were conducted in Kigali by the University of Twente – Faculty ITC with the support of CoK-OSC, the RCAA, and village representatives. The UAV was a DJI Phantom 2 Vision+ quadcopter (i.e. a rotary-wing UAV with four rotors) with a 14 Megapixel RGB camera with a fish-eye lens (FOV = 110°). Eighty-nine flights were made in total, taking more than 15,000 images to cover approximately 150 ha of unplanned settlements (Figure 6.1). Examples of some derived products, the dense point cloud and ortho-image, are given in Figures 6.2 and 6.3 respectively. The UAV orthomosaics were provided to the CoK-OSC and were used by the safeguard and detailed design consultants to carry out their activities. The orthomosaics were also made available at village level offices where they could be viewed and used by citizens engaged in public participation processes. This provides the opportunity to observe how the UAV images are actually used by various project stakeholders.

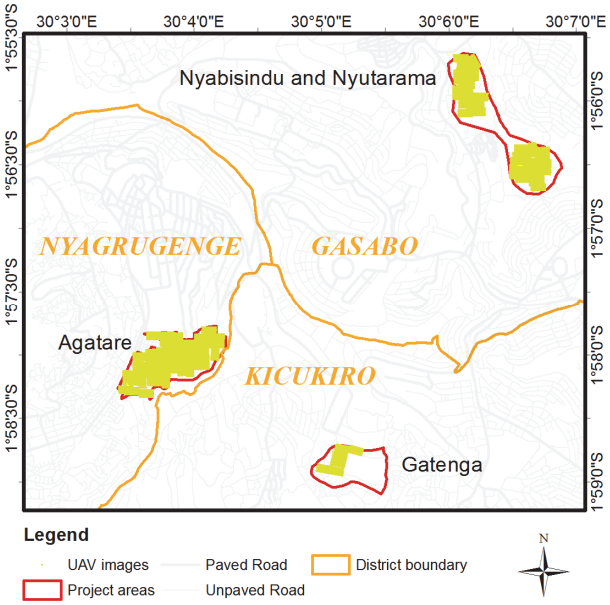


Figure 6.1: *The project areas covered by UAV flights over the three districts of Kigali in May 2015.*

Figure 6.2: *Sample of the ortho-image obtained from the UAV data over the Nyarugenge project area.*

Figure 6.3: *Sample of the 3D model (mesh) obtained from the UAV data over the Nyarugenge project area.*

Interviews were also conducted with both institutional and local stakeholders to identify important spatial information requirements for the upgrading projects, the perceived use of UAV for these activities, and to assess the attitudes of the local population towards the usage of UAV.

6.4 GIS requirements for upgrading projects

6.4.1 Information requirements for upgrading projects

An upgrading project consists of a number of phases: feasibility studies, detailed studies, developing project options, detailed design and project implementation (Davidson and Payne, 1983). The first two stages require information which describes the project area on four aspects: population and housing needs, the conditions of the project site, assessment of the current site development, and finally the institutional and financial framework. Although much of the socio-economic information required for upgrading projects may be stored in GIS databases, remotely sensed data such as airborne imagery or UAVs are limited to describing the physical characteristics of the settlement. We therefore restrain our analysis to the physical spatial information requirements which excludes the information regarding population and housing needs and the institutional and financial frameworks. However, remotely sensed data can provide enormous benefits in describing the physical characteristics of the project site environmental conditions (i.e. topography, vegetation, presence of hazards) and especially in describing the existing site

development. The latter includes characterizing the existing buildings, house layouts, plot coverage, land use, and accessibility. More extensive lists recommending which type of information should be collected for informal settlement upgrading projects have been published (Davidson and Payne, 1983; Goicoechea, 2008).

More specifically, we can focus on the information requirements defined by the case study area in Kigali. As the Nyarugenge upgrading project is intended to be a pilot project, there was not a well-established methodology regarding spatial data norms for upgrading projects in the City of Kigali. The spatial data requirements were therefore not clearly defined at the start of the feasibility and detailed design project which was completed in July 2015. Rather, they were based on the recommendations of a World Bank consultant (Banes, 2015), and were further specified in an interactive manner during the project execution by the contracting authorities (RHA, CoK-OSC, Nyarugenge District) and consultants. Table 6.1 lists the spatial data collected by the consultants, which could be equated to the current spatial data requirements. Initially the spatial information collected for the project was either: (i) provided by authorities, (ii) digitized from the 2008 ortho-imagery, or (iii) collected in field with a GPS. The collection of the spatial information was supported by Rapid Planning, a project supported by the German Federal Ministry for Education and Research and UN-Habitat to develop a trans-sectoral urban infrastructure planning methodology, for which Kigali is one of the three case cities (Consortium, 2015).

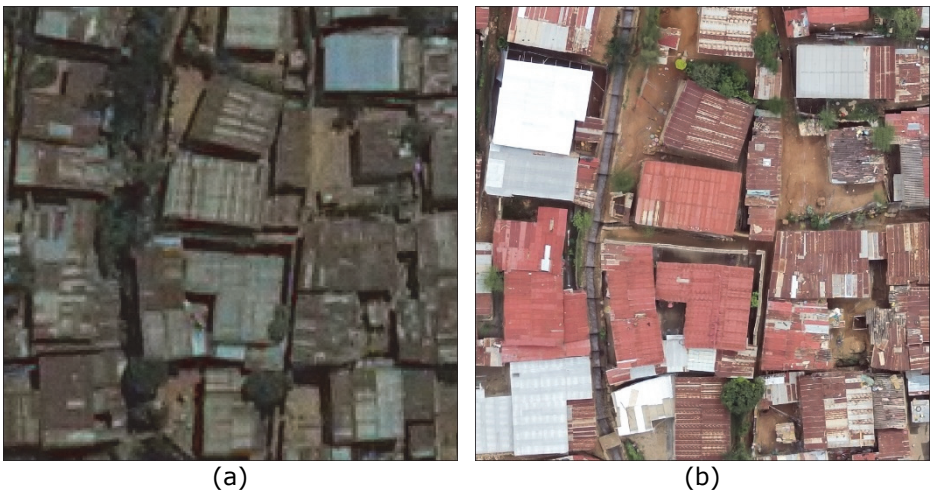
Table 6.1: *Spatial information collected by GIS Consultants for the Nyarugenge District Upgrading Project*

Spatial Layer	Source
Aerial imagery / Elevation	A 25 cm orthophoto from 2008 and DTM points at 10 m intervals provided by Rwanda Natural Resources Authority (RNRA)
Building Footprints	Footprints were digitized from satellite imagery and roof material was obtained by sampled questionnaires with the support of Rapid Planning,
Roads (type and material)	Digitized from satellite imagery, GPS in field
Drainage	Digitized from satellite imagery, GPS in field
Power lines	Provided by the power company, GPS in field
Water pipelines	Provided by the water utility company
Land use	Digitized from aerial photos, field survey
Services	GPS in field
Parcels	Provided by RNRA
Administrative boundaries	Provided by National Institute of Statistics Rwanda (NISR)

6.4.2 Opportunities of UAV to provide the required information

Previously, upgrading projects relied on the availability of satellite or airborne imagery. The benefits of UAV imagery as opposed to conventional imagery can be categorized into four main aspects: (i) by providing more accurate information through direct replacement of conventional imagery in existing workflows, (ii) by providing additional, previously inaccessible, information, (iii) by reducing field work, and (iv) by providing a basis for communication amongst stakeholders.

Regarding the first aspect, the imagery obtained from UAVs could be directly integrated by replacing satellite imagery in the existing spatial data workflows of the upgrading project. However, as opposed to airborne or satellite imagery, UAVs are a more flexible information acquisition platform. This potentially allows data to be collected more frequently, whether it's one large survey before a project starts or incremental surveys to observe and map changes during a project implementation phased over many years. Apart from the ability of providing up-to-date information, the spatial resolution of the UAV imagery is significantly better than that of conventional airborne imagery. Through direct replacement, rather than for example relying on an ortho-photo from 2008 with a 25 cm resolution, an up-to-date orthomosaic with a resolution of 3 cm could be used to digitize the objects of interest. This is apparent in Figure 6.4, which displays a part of the project area in the 2008 orthophoto (Figure 6.4a) versus the UAV orthophoto (Figure 6.4b). Notice how objects are more clearly visible in the UAV orthophoto, which facilitates digitization, and the appearance of new buildings, which increases the accuracy of the digitized information.



(a) (b)

Figure 6.4: *The added value of the UAV data is clearly visible when comparing the information provided by the 2008 orthomosaic (a) to the 2015 UAV orthomosaic (b). Note the enhanced visibility of objects in the scene as well as the appearance of new structures.*

The second main benefit is the provision of additional information. The higher spatial resolution of the orthomosaic as opposed to the aerial imagery allows for new objects of interest to be identified, such as lamp posts which indicate the presence of street lighting and the waste accumulation areas, which can be used to effectively plan waste collection points. Furthermore, more detailed attribute data of the existing objects can be obtained. For example, the roof material and condition of individual buildings can be observed and quantified. Such information may be used to monitor the implementation of the KCMP as well as provide baseline statistics describing the general conditions within the settlement. The 3D data obtained by the UAV provides information regarding building height and the local topography. Especially the latter may support the detailed design of interventions. For example, the drainage in the Agatare project area was mapped by GPS points in-field and measuring the width of the drain at regular intervals. Using the Digital Terrain Model (DTM) extracted from UAV imagery, the drainage capacity could be calculated more accurately (though still concealed in some areas by trees etc.), facilitating the design of more adequate interventions. In situations where the terrain isn't visible due to concealment by trees or other structures (i.e. *occlusions*), terrestrial imagery may be integrated with the top-view imagery to provide a more complete 3D input for drainage models (Meesuk *et al.*, 2015).

The ability to provide up-to-date imagery and to identify additional objects and attributes was perceived by the consultants to be a large benefit for the existing workflow by reducing the time and cost of subsequent field verification. This is the third main advantage of the utilization of UAVs, namely the reduction of field work. On the one hand, the ability to obtain more accurate and detailed information reduces the amount of data which must be collected or corrected in the field. On the other hand, as the project members have access to an up-to-date map, it is easier to plan their field work as the location of and accessibility to certain objects of interest are clearly visible. However, the reduction of the time required for the field work should be weighed against the time required to acquire and process the UAV imagery.

Finally, the high detail of the data products provides an intuitive environment which facilitates discussions between project planners and the local population. The UAV orthomosaics, overlaid with the designated project interventions, were presented by project members to the sector offices during the monthly community meeting in October 2015. Project members mentioned that, due to the detail, many local inhabitants could recognize their houses and other landmarks. This helped them locate the proposed interventions, thus providing a foundation for discussions between stakeholders. Although the upgrading project aims to limit expropriation, the introduction of utilities and improvement of infrastructure will require extra space and almost certainly results in some expropriation. As the spatial layout of buildings is more clearly

visible in the UAV orthomosaic, planners are better able to explain why interventions are designed in certain positions, and as a result why specific structures are selected for relocation. The printed maps were left at the sector offices, and hardcopies of the orthomosaics are therefore directly available to the local population. This is a simple way through which the residents could also benefit from the UAV imagery. Further analysis should identify additional uses of the imagery by the local citizens and as well identify concerns of an ethical nature related to mission planning, the mapping process, data ownership, data use, marginalization etc. (Rambaldi, Chambers, *et al.*, 2006) that may or have been recognized as problematic or potentially so.

6.5 Potential bottlenecks regarding the use of UAV

6.5.1 Practical considerations

Although from a scientific perspective, UAVs appear to be a promising method for spatial data collection, there are also practical considerations regarding the technological limitations of UAV platforms, data processing, and specialized knowledge. The technological characteristics of the UAV platform also affect the utility of UAVs as a data acquisition tool. Firstly, imagery cannot be acquired during rainy or windy weather conditions. Secondly, the range of the platform limits the extent of area covered per flight, and requires the take-off location to be close to the flight area. This may be difficult in unplanned settlements, where dense construction, narrow footpaths and overhanging power lines or vegetation make it difficult to find adequate take-off locations and which may slow down data acquisition.

Back in the office, there are also practical challenges to obtain the required information from the raw UAV images. Firstly, from a data quality perspective, a high image overlap, or redundancy, increases the quality of the 3D model. However, the large number of images may also incur data storage problems (Baiocchi *et al.*, 2014). Furthermore, processing the imagery currently requires specialized software and advanced hardware requirements. In recognition of the data processing bottleneck which may impede the utility of UAVs for informal settlement upgrading, the University of Twente – Faculty ITC, the French Institut National de L'Information Géographique et Forestière, and the Netherlands eScience Center have initiated a project aimed at speeding up the open source photogrammetric processing software MicMac (NLeSC, 2015). When completed this project could speed up data processing within a software environment which does not have licensing fees as well as reduce hardware requirements.

As with any new technology, there are costs associated with the initial adoption of UAVs for upgrading projects. Although the UAV itself may be relatively

inexpensive, there are additional expenses regarding obtaining the proper permissions and documentation, training the pilots, and training the GIS specialists and photogrammetrists who must then process the data. However, it should be noted that once the UAV imagery is collected for an upgrading project, there are many other sectors which may also benefit from the obtained data products. Examples include: updating existing (smaller scale or outdated) topographic map data, supporting cadaster, valuation for tax collection purposes, and inspection for the monitoring of illegal construction.

6.5.2 Social considerations

Such practical issues of technological limitations, data processing, and investment costs may be overcome. However, they overlook one of the most important stakeholders involved in UAV data acquisition – the inhabitants of the settlements being flown over and their privacy. Based on the experiences of flying the UAV in Kigali, the first observation was the interest of the inhabitants during the flights. Many people were curious, crowding around to watch the flights, taking pictures and asking questions. However, a number of people were also concerned. Most concerns were based on the fear that the UAV was being used to survey the area and plan for expropriation. In this particular project, there was a very limited time to execute the UAV flights after permission was obtained. There was therefore insufficient time to fully address such community concerns through proper planning and awareness raising (Rambaldi, Chambers, *et al.*, 2006). However, an effort was made to mitigate their fears by answering questions and explaining the purpose of the activities during the flight acquisition. Furthermore, the village leaders were notified about the flight activities beforehand so they could communicate this to the local population and mitigate concerns.

The utility of UAV data products, in the form of printed images, for the local residents was also analyzed through interviews. When shown some sample UAV images, a number of residents 'recognized' the images from previous land administration activities or the activities of the consultants involved in the Agatare project. Others indicated that it was very important to provide the users with some kind of training or explanation regarding how to interpret the images.

According to interviews with residents, the images were mainly considered to be useful for the village leaders. The general population mentioned uses such as giving a friend directions how to find their house, or as a memory to show their grandchildren how the neighborhood used to look like. Others mentioned that you could compare your house to your neighbor's house. This was also the most recurring theme in the utility of the images for village leaders. The UAV imagery was considered to be useful to identify which aspects (e.g. house typology) needed to be changed in order to comply with the KCMP. Village

leaders also mentioned that they could use the images to help explain the government plans to the population. These observations hint at the strength of the top down influence of the government in Rwanda, and to the extent to which plans such as the KCMP are communicated to the local leaders. As a hardcopy of the images is now available at the sector offices, future analysis may identify additional uses of the images by the local officials and citizens.

6.6 Discussion

Although there is a great enthusiasm regarding the detail of the imagery obtained from a UAV, practical aspects must be taken into account to determine which applications could maximally benefit from the UAV imagery. In the case of upgrading projects, the size of project areas and required level of detail appear to make it suitable for UAV image acquisition. However, the limited extent covered by the current UAV platform makes it less suitable for tasks which require covering large areas in a limited amount of time. Such applications could consider obtaining a UAV platform more suitable for the task in question, or smart sampling strategies.

The potential use of the DSM obtained from the imagery should also be further investigated. Comparing the quality of DSMs extracted from UAV imagery to traditional surveying methods and LiDAR is a being researched (Haarbrink and Eisenbeiss, 2008; Harwin and Lucieer, 2012). If the quality is sufficient, it could provide an enormous benefit to the upgrading project in the terms of surface water drainage analyses and the detailed design of infrastructure.

Currently the utility and adoption of UAVs for upgrading projects is limited by external factors. UAVs are an emerging technology, which are increasingly being used in developing countries for applications such as flood resilience in the Dar Ramani Huria project (http://ramanihuria.org/) and cadaster in the its4land project (http://www.its4land.com/) as well as unplanned settlement mapping. Unfortunately, legislation and protocols to obtain flight permission are often cumbersome or ambiguous. To resolve this issue, it is important to develop clear policies, guidelines and standards regarding flight regulations. In the case of Rwanda, legislation regarding UAV flight permissions is currently being drafted. One of the objectives of the its4land project is also to identify current UAV flight regulations in East Africa and provide guidelines for their future development. On the other hand, the UAV pilots should operate safely and in compliance with these standards to help establish trust and transparency between the various parties. This could ensure UAV users have the freedom to conduct flights and extract high-quality information to support e.g. urban planning activities, while ensuring responsible flights and respecting public safety.

6.7 Conclusions and recommendations

To conclude, a UAV has the potential to be a valuable tool in spatial information acquisition for urban upgrading projects. The provision of highly-accurate and up-to-date information allows for the mapping of the current situation in the area, including the identification of buildings, roads, land use, drainage, and other points of interest which are vital for upgrading projects. Using UAV imagery provides advantages on four fronts. Firstly, it allows more accurate information to be extracted by replacing conventional aerial or satellite imagery in existing project workflows. Secondly, UAV data products also have the potential to provide spatial information to the upgrading project which isn't available through conventional image sources, such as providing high-resolution elevation information for detailed drainage calculations and the design of implementation measures. Thirdly, the increased detail of the UAV imagery versus conventional satellite (or aerial) imagery saves time in field verification. Finally, the UAV images are also intuitively understandable by various project stakeholders, thus forming a foundation for effective communication of issues in the study area and planned interventions. Further analysis should investigate methods to obtain useful spatial information (i.e. semi-automatic classifications to identify different types of objects and their semantic attributes) from the data, which fit the needs of the stakeholders while taking local constraints into account.

To maximally benefit from the potential advantages of UAVs as a data acquisition platform for upgrading projects, the practical aspects of UAV data collection must be contemplated. The spatial data requirements should be analyzed to enable the selection of a UAV platform, photogrammetric software, and hardware which is suited to the task at hand. Furthermore, effort should be made to explicitly consider the ethical issues involved with obtaining such high-resolution imagery over these impoverished and marginalized areas, and to stress the importance of sharing the benefits of the information obtained through the flights with the local population.

Chapter 7 – Evaluating the Societal Impact of Using Drones to Support Urban Upgrading Projects[13]

[13] This chapter is based on:

Gevaert, C., Sliuzas, r., Persello, C., and Vosselman, G., (2018) 'Evaluating the societal impact of using drones to support urban upgrading projects', *ISPRS International Journal of Geo-Information,* 7(3), 91, doi: 10.3390/ijgi7030091

Abstract

Unmanned Aerial Vehicles (UAVs), or drones, have been gaining enormous popularity for many applications including informal settlement upgrading. Although UAVs can be used to efficiently collect highly detailed geospatial information, there are concerns regarding the ethical implications of its usage and the potential misuse of data. The aim of this study is therefore to evaluate the societal impacts of using UAVs for informal settlement mapping through two case studies in Eastern Africa. We discuss how the geospatial information they provide is beneficial from a technical perspective and analyze how the use of UAVs can be aligned with the values of: participation, empowerment, accountability, transparency, and equity. The local concept of privacy is investigated by asking citizens of the informal settlements to identify objects appearing in UAV images which they consider to be sensitive or private. As such, our research is an explicit example of how to increase citizen participation in the discussion of geospatial data security and privacy issues over urban areas and provides a framework of strategies illustrating how such issues can be addressed.

7.1 Introduction

Rapidly growing urban populations and an inability to meet affordable housing needs are some of the driving factors behind the emergence of informal settlements worldwide. It is estimated that 881 million people were living in slums in 2014, which corresponds to 29.7% of the urban population at that time (UN-Habitat, 2016). This proportion may be much higher nationally, as estimates of the urban population living in slum areas reaches 77% in Tanzania and 96% in Rwanda (UNHabitat, 2013). The need of improving these conditions is considered one of the main challenges in urban development (Barry and Rüther, 2005) and is a prominent issue on many urban agendas (AUC, 2015; United Nations, 2016; MININFRA, 2017; UN-Habitat III, 2017). The current paradigm regarding urban upgrading projects encourages in situ upgrading which aims to improve the living conditions within a neighborhood itself (Abbott, 2002; UN-Habitat, 2016; MININFRA, 2017) through the improvement of the physical infrastructure (Turley *et al.*, 2013), while advocating effective participation of the local community (UN-Habitat, 2013). The design of these infrastructural improvements, as well as urban governance in general (Baud *et al.*, 2014), requires geospatial information (Abbott, 2002; Sliuzas, 2003; MININFRA, 2017). This geospatial information generally consists of elements such as terrain elevation, building footprints, roads, drainage, power lines, water pipelines, land use, services, parcels, and administrative boundaries (Caroline Gevaert *et al.*, 2016). Information pertaining to natural hazards such as natural drainage, landslides, and inundated areas may also be relevant (Ramani Huria, 2016). However, consistent and up-to-date geospatial information is often lacking (Ordnance Survey, 2015), especially for informal settlements. These are sometimes excluded from official data collection (Carr-Hill, 2013) and often remain 'empty spots on the map' (Paar and Rekittke, 2011). Although Remote Sensing has emerged as a useful tool for the provision of spatial information for informal settlement management (Kuffer, Pfeffer and Sliuzas, 2016), the spatial resolution provided by satellite imagery is sometimes not sufficient for e.g., the detection of individual houses, infrastructure, and details of environmental conditions.

Unmanned Aerial Vehicles (UAVs), also known as drones or Remotely Piloted Aircraft Systems (RPAS), are a potential solution for this issue. A UAV equipped with a camera can take images of the area it flies over. These images can be used to obtain detailed elevation models and orthoimagery. UAVs have been gaining enormous popularity for many applications, from recreational uses by hobbyists to serving as a genuine data collection tool by businesses and local governments. Some projections estimate the global commercial UAV market to have a value of seven billion dollars in 2020 (Thibault and Aoude, 2016). Recently, there have been a number of projects using UAVs for mapping

informal settlements including in Rwanda (Caroline Gevaert *et al.*, 2016), Tanzania (Minja, Iliffe and Anderson, 2016), Uruguay (Birriel and González, 2015), and Albania (Kelm, Tonchovska and Volkmann, 2014).

The main motivation for the use of UAVs in urban upgrading projects is the perceived utility of the geospatial information that they obtain. It is important to understand the range of (technical) benefits of this geospatial information as it is a strong driver behind the deployment of UAVs and therefore provides context to the social impact analysis. We distinguish three categories which describe how UAVs can provide geospatial data to support urban upgrading projects (Table 1). The first category is data which can be directly derived from UAV imagery. Objects such as roads and building footprints may be first digitized over imagery in the office, and later updated and verified during field visits. Satellite imagery or outdated aerial imagery is often used as a basis for digitization, but having access to recent UAV imagery may greatly speed up both the digitization and the field verification (Caroline Gevaert *et al.*, 2016). The second category refers to spatial information for which the UAV imagery by itself cannot be used as a complete and accurate data source, but may be (partially) derived from the UAV data. This refers to the attributes of larger objects, such as identifying housing material which may be visible in some of the original UAV images. It also refers to objects which are relatively small (lamp posts) or sometimes lie below other objects (drains that run below covers or under roads). Other information can be derived from UAV data, but require advanced processing (e.g., the number of stories in each building can be approximated from the height difference between the ground and roofs) or local knowledge (e.g., water distribution points may be recognized by the piles of uniform water containers outside the building). Geospatial information in this second category can generally be used to support informed decision making when supplemented by additional data sources, advanced processing or local knowledge, but cannot by itself be used to provide the completeness and accuracy required for mapping. The local context greatly influences whether a certain type of information falls into the first (i.e., is clearly visible in the UAV data) or second category. Finally, the third category is geospatial data which cannot be identified from the UAV imagery. This can be 'invisible' data such as administrative boundaries which are social constructs and have no physical representation or information which may have a physical representation but are not visible in the UAV imagery, such as population counts.

Although UAVs have great potential for the provision of geospatial information in upgrading projects (Caroline Gevaert *et al.*, 2016), there are concerns regarding the ethical implications of their usage (Haarsma, 2017) and the potential misuse of data (Culver, 2014). Privacy is often stated as a concern as UAVs may infringe on: *privacy of location* when individuals can be identified

Table 7.1: *Examples of geospatial information derivable from UAV images*

Directly Derived from UAV Data	Partially or Indirectly Derived from UAV Data	Not Derived from UAV Data
Buildings	Attribute information (construction material, number of floors)	Administrative boundaries
		Population count
Roads	Utilities	Household income
Vegetation	Land use	
Elevation	Solid waste dump sites	

and located in the UAV images, *privacy of behavior* in a private space without being monitored by others, *privacy of space* as information is revealed regarding private areas such as back yards, *privacy of association* regarding group membership and affiliations, and *privacy of data and image* regarding the control of persons over images in which they are present (Finn, Wright and Friedewald, 2013). The perceptions of the local population (Sandbrook, 2015) are also a concern. Especially when flying over marginalized communities, it is important to notify citizens of the purpose of data collection, the rights to access, data processing and distribution (Pauner, Kamara and Viguri, 2015). The lack of a unified policy framework directing such practices leaves much of the responsibility to industry self-regulation, which may not sufficiently protect these marginalized communities (Clarke, 2014).

However, the existence of such a 'unified' framework is questionable due to two specific challenges. Firstly, the ethical use of UAVs is dependent on the application for which it is used (Finn and Wright, 2016). For example, one ethical concern is inadvertently capturing imagery of persons or privates spaces. This risk is higher when using UAVs for real estate applications than pipeline monitoring (Finn *et al.*, 2014). Especially in the case of the latter, ethical UAV operations can aim to avoid the inadvertent collection of persons in their data, but when using UAVs to videotape concerts for example this is unavoidable. On the other hand, as crowds at a concert are in a public space anyway, a "chilling effect" of being observed by the UAV will be limited (Finn *et al.*, 2014). For journalism, context (e.g., what is the reason why a protesting crowd captured by UAV imagery is protesting?) and conflict of interest (e.g., should footage of potentially conflicting activities be turned over to law enforcement?) are important ethical concerns (Culver, 2014). These examples clearly illustrate how ethics are dependent on the application.

Secondly, the concept of sensitive information or privacy may vary amongst people, groups, and cultures (Ordnance Survey, 2015). For example, a study of the privacy awareness behavior of almost 200,000 Facebook users from 30 countries showed a strong correlation with cultural dimensions, even when corrected for socio-economic indicators (Reed, Spiro and Butts, 2016). In another example, a European study of seven countries indicated that younger

age groups have a lower privacy concern but higher data protective behavior than older groups (Miltgen and Peyrat-Guillard, 2014). At a policy level, research has related Hofstede's concepts of "collectivist" and high "power distance" national cultures with a reluctance to implement open data initiatives (Saxena, no date).

These challenges suggest a need to use a case-by-case method to weigh the infringement of (mainly individual) moral rights and ethical values against the assumed common good achieved through the use of UAVs for data acquisition and provision (Culver, 2014). Such qualitative research is important for the ethical usage of UAVs, but also to promote the continuation of similar efforts in the future (Sandbrook, 2015). Previous studies have made efforts towards the development of conceptual framework, for example an analysis of six different emerging technologies identified seven different concepts of privacy (Finn, Wright and Friedewald, 2013). This enables other studies to focus on a single aspect of privacy, such as the effect of the surveillance capability of UAVs on the behavior privacy (Clarke, 2014). Still, there is a lack of empirical knowledge investigating ethical concerns (Sandbrook, 2015). One comprehensive study in Europe provides an overview of the perceptions and practices of industry, regulators and civil society (Finn and Wright, 2016). Another study in Tanzania focused on the perceptions and concerns of various private and public parties regarding the usage of UAVs for mapping purposes (Eichleay *et al.*, 2016), but does not analyze in detail the privacy concerns of the image products obtained from these flights and their potential distribution. The aim of this study is therefore to evaluate societal impacts of using UAVs for informal settlement mapping through two case studies in Eastern Africa. In analyzing their social impacts, we define 'ethical usage' of UAVs as the extent to which their use is aligned with the values of: participation, empowerment, accountability and transparency, and equity, as these values are characteristic of global policy frameworks regarding urban upgrading projects and urban governance in general. Specific emphasis is given to identifying if the local communities consider any of the objects captured by the UAV imagery to be sensitive.

The remainder of the manuscript is organized as follows. Section 2 provides background information of the two case study areas and UAV acquisitions and describes the questionnaires used to interview the residents. Section 3 summarizes and interprets the results of these questionnaires. In Section 4, these results are further interpreted by discussing the relations between UAV image acquisition to the values of privacy, empowerment, accountability, transparency, equity, and participation. Finally, Section 5 draws conclusions from these observations and analyses.

7.2 Materials and methods

We utilize a comparative case study approach to compare the use and impact of UAV imagery for mapping impoverished settlements in two projects: an urban upgrading project in Kigali, Rwanda and a participatory mapping program aimed at improving urban resilience in Dar es Salaam, Tanzania. The projects have several similarities. In both cases, the deployment of the UAVs was not initiated by the residents, but rather by external institutions (the University of Twente in Kigali, and The World Bank in Dar es Salaam) in conjunction with the local governments. As far as the authors are aware, this was the first utilization of UAVs for mapping purposes over both locations. The maps derived from the imagery could be sensitive as, in both locations, some residents could be subject to displacement. In Kigali, some residents may be displaced to make space for new roads or other infrastructure. In Dar es Salaam, many residents of houses that are located on river floodplains are threatened with displacement.

However, there are also some key differences between both case studies. The maps in Kigali were created by (foreign) engineering consultants and the local government, whereas a participatory mapping approach was adopted in Dar es Salaam. The distribution of the soft copies of the orthophoto mosaics and derived geospatial data was limited to the official project partners in Kigali, whereas in the case of Dar es Salaam they are freely accessible for the public through OpenStreetMap and a web portal (ramanihuria.org). The social context of both areas also differs, causing differences in the physical appearance of the informal settlements in the imagery as well as the local interpretation of objects and their significance by the residents. In the following section, both case studies are described in more detail.

7.2.1 Case study I – Kigali, Rwanda

After a city-wide inventory of the status of informal settlements in Kigali, the Agatare neighborhood was selected by national, municipal, and district authorities to serve as a pilot project for urban upgrading projects (GISTech Consultants LTD, 2015). The project employed a participatory approach to identify key issues in the neighborhood and propose infrastructural improvements. The methodology developed during this pilot project will serve to develop upgrading guidelines nation-wide (MININFRA, 2017). The project area is roughly 86 ha, with an estimated 3,977 households and 18,914 individuals at the time of the pilot project (GISTech Consultants LTD, 2015). Field investigations indicated that approximately 24% of respondents have a member in the household who works in the formal sector such as government of NGOs, whereas 27% of the household are employed informally (vendors, mechanics, construction, etc.). Roughly 40% of the households had an income of less than 100 euros per month at the time (GISTech Consultants LTD, 2015).

During the pilot study in May/June 2015, UAV imagery was acquired over the project area through a collaboration of the University of Twente—Faculty ITC and the City of Kigali One Stop Center (CoK-OSC). The UAV was a DJI Phantom 2 Vision+ quadcopter (i.e., a rotary-wing UAV with four rotors). Eighty-nine flights were made in total, taking more than 15,000 images. These images were then processed to obtain an orthomosaic with a spatial resolution of 3 cm. More information regarding the UAV flights and data processing can be found in (Gevaert *et al.*, 2017). Local village officials were notified when UAV flights would take place over their neighborhoods, with the idea that they would be able to further notify the residents of the area and answer any concerns they might have.

The raw UAV data and point clouds remain in the hands of the organization executing the UAV flights (University of Twente / Faculty ITC). The orthomosaic was provided to the CoK-OSC, and has been used by various consultants to support the design of interventions in subsequent stages of the upgrading project. Consultants of the upgrading project in Kigali described how the UAV imagery was beneficial for: reducing the time needed to collect data, providing previously unavailable information (such as solid waste accumulation sites), improving field work efficiency, designing more appropriate infrastructure interventions, communication between stakeholders, and mitigating expropriation (Caroline Gevaert *et al.*, 2016). Printed hardcopies of the UAV orthomosaic at scale of approximately 1:1000 overlaid with vector layers representing the planned project interventions and administrative boundaries, were also provided to the local sector offices by the CoK-OSC, where they are used by local officials and residents when discussing upgrading or general development issues.

7.2.2 Case study II – Dar es Salaam, Tanzania

Ramani Huria is a large-scale community mapping project in Dar es Salaam, funded by The World Bank and the United Kingdom's Department for International Development (DFID) under the Tanzanian Urban Resilience Program. Its objective is to improve urban resilience to flooding by providing accurate spatial information to support planning decisions. The project started with the mapping of the Tandale Ward in 2015. In 2017, the project was scaled up with the intent of mapping the entire city of Dar es Salaam. Local university students are mobilized to digitize buildings and urban infrastructure from imagery, which is then refined and populated with detailed attribute information in the field through the engagement of community members. One important aspect of the community outreach is their involvement in mapping areas prone to flooding.

To support the mapping process, UAV flights were conducted over a large part of the city in 2015 using a Sensefly eBee. The images were processed to obtain orthomosaics with a spatial resolution of 5 cm. The orthomosaics were published online by the Tanzania Open Data Initiative. Vector layers of buildings, roads, drainage, and inundated areas are published in OpenStreetMap. An atlas was made which translates the Ramani Huria data into three maps for each ward: general topographic, drainage, and potential inundation. Local ward offices were also provided with printed maps of their ward and copies of the atlas.

7.2.3 Methodology to analyze perceptions of the local community

Data on the potential social impact on the community was collected via resident questionnaires. The questionnaires consisted of two parts. The first consisted of open questions regarding residents' perceptions of the UAV flights and usefulness of the derived geospatial information. This part of the questionnaire aimed to answer the following questions:

- Citizen perceptions to the UAV flights: Did you see the UAV flights? What did you think?
- Citizen awareness of the implications of UAV flights: Have you seen other UAV flights and where? Have you seen aerial imagery before? What do you see in this example of aerial imagery?
- Citizen ability to control possibly sensitive data captured in the imagery: Were you aware the flights would take place? Are you aware of any issues being discussed at neighborhood level regarding privacy?
- The observed and perceived usage of UAV imagery and maps: Did you see the maps printed at the ward office? What were they used for and by whom? What do you think they could be used for?

The second part of the questionnaire aimed to identify which objects are considered as sensitive due to possible privacy or security concerns of the residents at various levels of abstraction (Figure 7.1). It showed examples of UAV orthomosaics at full resolution (i.e., 3 cm in Kigali and 5 cm in Dar es Salaam), the degraded orthomosaics downsampled to 50 cm (i.e., similar to high-resolution satellite imagery), and vector maps derived from the UAV imagery. In Kigali, the residents were also displayed a raw UAV image and a 3D mesh model. Such products were not available for Dar es Salaam. For both cities, an area of interest was selected which was representative of the settlement, and contained a clear view of resident's back yard, people, and cars. In the case of Dar es Salaam the areas also contained visible open drainage and roofless toilets. For each data format, the resident was asked to imagine that they lived in the depicted area. Then they were asked which objects or places visible in the image they would consider to be sensitive if seen by (a) their neighbors; (b) village/ward leaders; (c) other institutions; or

(d) the public (i.e., published online). The intention being to investigate how privacy relates to the identity of the viewer.

In Kigali, 54 interviews (57% female) were conducted in 2017 by selecting a section of the upgrading project area and interviewing residents who were home. Interviews were conducted in the local language (Kinyarwanda) by two local students and the support of one of the authors. The interviewers had experience in conducting questionnaires and were trained regarding the purpose of the study. The area was selected as: (i) it was subject to another set of UAV flights earlier the same year so perhaps more citizens have seen the flights; and (ii) it was transected by a new road which was implemented as part of the upgrading program. Interviews were conducted during working days and hours, the same as the UAV flights, in an effort to target parts of the population which were more likely to have seen the flights.

In Dar es Salaam, a digital format of the questionnaire was developed in ODK Collect. University students used the questionnaire to interview 26 community members (39% female) who were recruited as part of the Ramani Huria community outreach activities. The interviews were again conducted during working days and hours for the same reasons stated above. The questionnaire was translated to the local language (Swahili), and interviews were conducted in the same. The students conducting the interviews had previously received extensive training in surveying and interacting with community members.

7.3 Results

7.3.1 Perceived and actual usage of UAV data

Community members in both Kigali and Dar es Salaam could access the UAV products through hard copy maps distributed to local governmental offices. In Kigali, 41% of the respondents report seeing these maps at the sector offices. These maps were mainly being used by the sector officers for development, management, and explaining the Kigali City Master Plan to the local community. Only 18% of the reported cases mention to the maps being used by the general public. They mentioned using the maps for "locating" or "giving directions". However, when asked about what they thought the maps could be useful for (i.e., perceived usage of UAV data products), this proportion went up to 55%; an improvement but still quite low. However, in both cities, it is uncommon for low income residents to have access and use maps on a regular basis. Respondents in Kigali also reported that the maps could be useful for businessmen or tax and land tenure purposes.

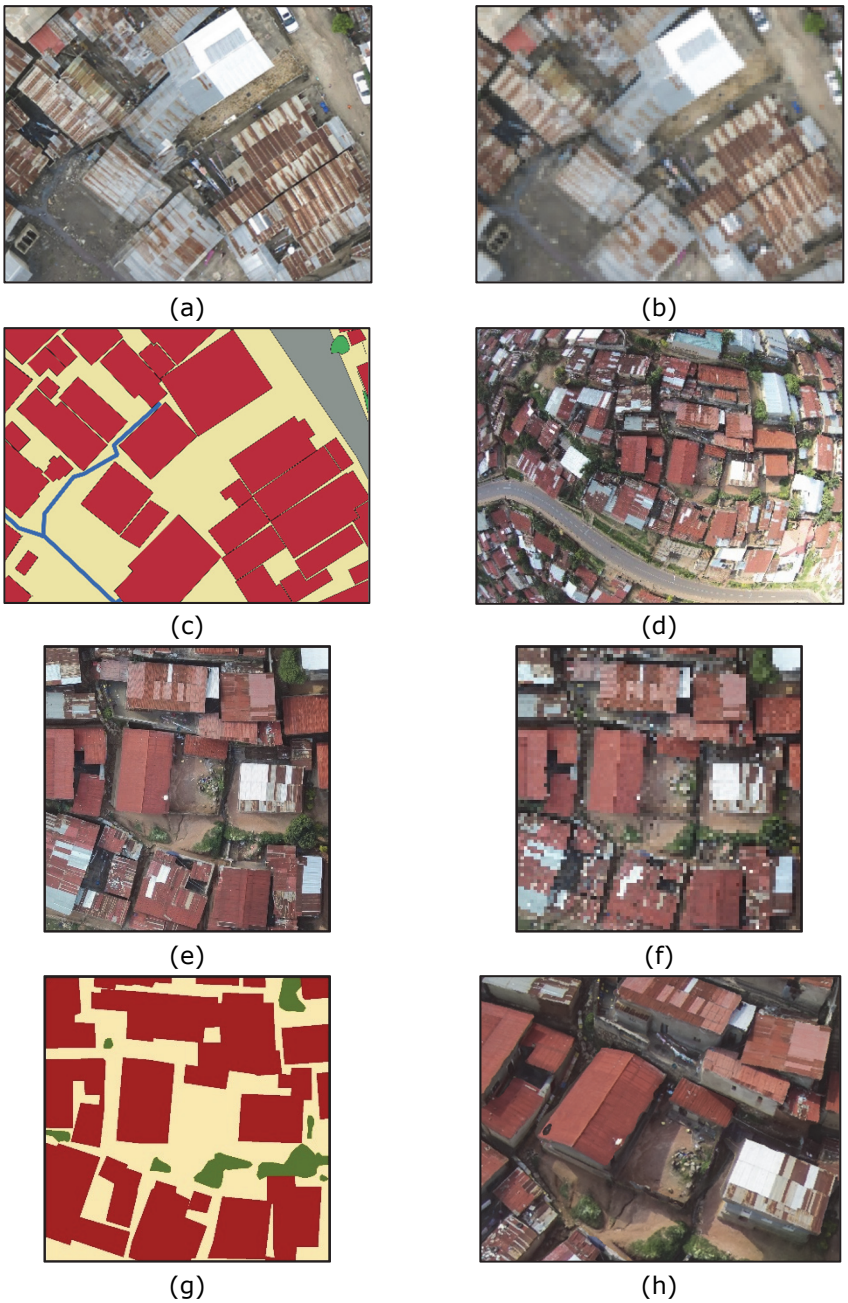


Figure 7.1: *The UAV orthomosaic (a), blurred orthomosaic (b); and vector map (c) images used for the Dar es Salaam questionnaire and the raw UAV image (d); UAV orthomosaic (e); blurred orthomosaic (f); vector map (g); and 3D mesh (h) images used for the Kigali questionnaire.*

In Dar es Salaam, only 35% of the respondents reported seeing the maps at the ward office being used, and only 11% of the reported use cases refer to usage by the general public for locating purposes. However, when asked for the perceived usage, 29% of the reported the leader using the imagery for planning purposes, 46% mentioned the public using the imagery for locating or giving directions, 33% reported using the images for map making, and 13% reported using the imagery for educational purposes.

7.3.2 Residents' perceptions regarding UAV flights

There are two main points of community members' perceptions to UAV flights which are of interest for the current investigation. Firstly, what are the initial reactions or emotions of the community members to the UAV flights? Secondly, are the community members aware of the purpose of the flights, and the implications thereof?

According to the questionnaires, 76% of the respondents saw UAV flights in Kigali, yet only one citizen reported knowing the flights would take place. In Dar es Salaam, only 31% of the respondents saw the flights, and 12% of the respondents were aware flights would take place. Few community members reported strong negative reactions when seeing the UAVs flying. Five respondents in Kigali reported negative reactions: believing that the UAV would destroy a building, spying, and two reported fear of expropriation. In Dar es Salaam, one respondent reported feeling afraid when seeing the flights, though it is unclear why. Fortunately, most respondents who saw the UAV flights had neutral or positive reactions. In Kigali, 59% thought the drone was taking pictures, 12% thought it was for the road construction, 17% mentioned other neutral reactions (e.g., "at first we thought they were toys"), and 7% mentioned other positive reactions (e.g., "we were happy to see a plane flying over our house" and "I was amazed"). In Dar es Salaam, 25% thought the UAV was taking pictures and 63% mentioned other neutral reactions. A study in Zanzibar similarly concluded that responses of the community to the UAV flights were generally positive, even if members were previously unaware of UAVs (Eichleay *et al.*, 2016).

The second question is whether community members perceived that the UAV was taking pictures or otherwise observing them. Seventy-six percent of the respondents in Kigali saw the UAV flying, and 59% of them realized that the UAV was capable of taking pictures or videos. In Dar es Salaam, 31% of the respondents saw the UAV, of which only 25% explicitly mentioned a camera. Thirty-eight percent recognized the UAV as a drone but did not specifically indicate that they were aware that it had a camera. It can be expected that as UAVs are becoming more prevalent in the general society, more community members will be aware that the UAVs flying overhead are likely to be observing them. For example, five respondents in Kigali reported seeing the use of a UAV

to capture videos of the Tour du Rwanda cycling event in 2016. In Dar es Salaam, half the respondents indicated seeing UAVs elsewhere such as weddings or concerts. The connection between UAVs and aerial photography will almost certainly become better known.

7.3.3 Privacy

Knowing that the UAV is observing the ground below is not the same as understanding the detail of the image products and the implications of its use for mapping. Regarding the five types of privacy discussed above, respondents did not list 'people' as being privacy sensitive even though the image samples used for the questionnaires in Kigali and Dar es Salaam were selected to include people. Upon further questioning, they reported that although the orthomosaics depicted persons, they could not be recognized and their visibility was not perceived as a cause of concern. This would indicate that *privacy of location* and *privacy of association* are not considered to be at risk in these specific locations and contexts. The *privacy of space* is relevant as the UAV images display private spaces such as backyards, which are not openly visible from the ground as the view is blocked by fences and walls. Similarly, the *privacy of behavior* is affected as objects in the backyard may shed light on the private activities performed by the household. The *privacy of data and image* is also relevant as the community members have little control over the usage of the data collected by the UAV.

Results of the questionnaires shed more light on the privacies of *space* and *behavior* in the two case studies by allowing community members to list objects which they considered to be private. The *privacy of data and image* is addressed by asking the respondents for the sensitivity of these objects for use cases by various end-users. Questionnaire responses indicate that there are substantial differences between both case studies regarding the type of objects and types of data products which were considered sensitive or private.

In Kigali, the main concern were old roofs, and 'rubbish' on roofs and in backyards (Figure 7.2a). Respondents did not wish these to be seen by the neighbors (13%), local leaders (11%), and other institutions (11%), whereas 24% of the respondents wouldn't like the orthomosaics showing these objects to be published online and available for the (international) public. In fact, although one respondent explicitly wished foreigners to see the old roofs, because "maybe they can provide help", in general more respondents wished to hide the dirty roofs from the foreigners more than the neighbors, local leaders, or other institutions. When further questioning queried the apparent lack of sensitivity of objects such as laundry hanging out to dry and cars in driveways, respondents answered that such behavior was 'normal' and so what would be their concern if others were to see it? With the exception of a single respondent declaring old roofs in the blurred image being considered as

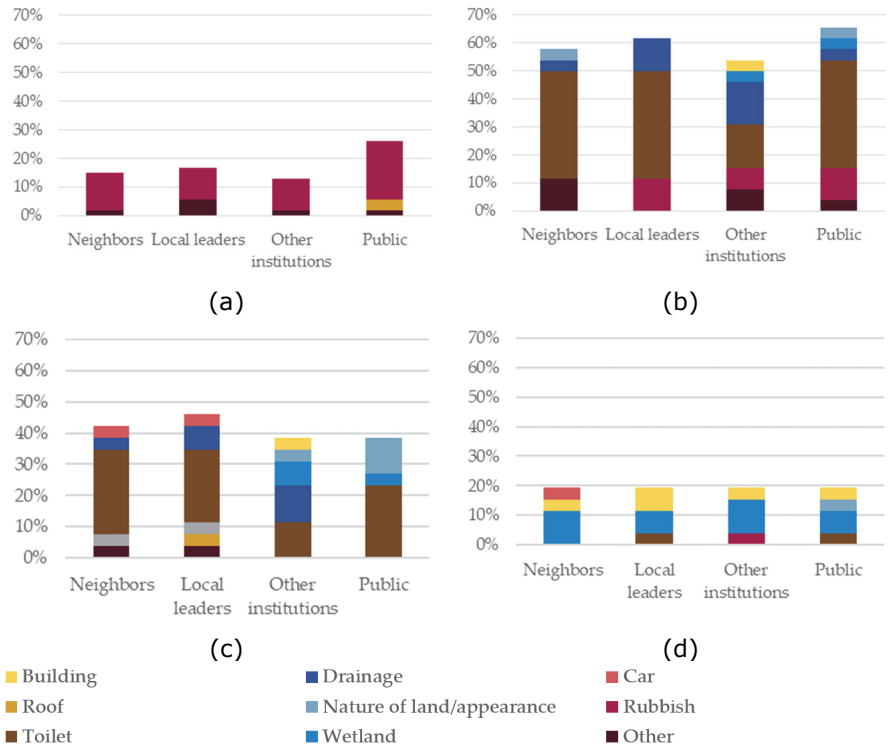


Figure 7.2: *Percentage of questionnaire respondents considering an object visibly sensitive in: the high-resolution UAV orthomosaic in Kigali (a) and Dar es Salaam (b). The privacy sensitive objects in the blurred orthomosaic (c) and vector map (d) also provided for the Dar es Salaam case study.*

private, the respondents in Kigali indicated that sensitive objects were not clearly visible in the blurred orthomosaics (i.e., simulated high-resolution satellite images) or vector maps.

Further investigation of the local context can help interpret the results of the questionnaire. First, the informal settlements in Kigali generally consist of multiple households on a single plot. Therefore, backyards are generally not considered as private spaces in these areas as many unrelated people are passing by on a daily basis. This explains why many objects on display in backyards are not considered to be sensitive if their visibility is shared with a wider public through the UAV data. On the other hand, the concern with the low-quality roofing or 'rubbish' indicates that the inhabitant of that house is deviating from the social norm of cleanliness. This causes a feeling of shame, as clearly indicated by one respondent: "It is embarrassing to show everyone that your house roof is old and dirty". Another: "It may cause problems if everyone sees that the roof of my house is old." The importance of 'cleanliness' is also emphasized by the government. For example, the Mayor of Kigali stated

"the development [in Rwanda] is very fast and the cleanliness is talked about the world over. Let everyone be responsible" (Kuteesa, 2016). This is reinforced by the legislation, as Article 107 of the Organic Law on Environment of 2005 states: "Any person who deposits, abandons or dumps waste [...] in a public or private place, is punished by a fine" (Rwanda, 2005). As such, the high-resolution orthomosaics may be evidence documenting which houses are not meeting the development or cleanliness aspirations of the local community and government and perhaps incur fines.

Responses in Dar es Salaam were quite different (Figure 7.2b). Only three respondents listed rubbish in the orthomosaic as being a sensitive object. Rather, toilets were listed as the main sensitive object in the orthomosaic. Ten of the respondents would not like these to be seen by neighbors, local leaders or the general public, and four would rather they not be seen by other institutions. It should be noted that in this case, the roofless structures housing the toilets were shadowed and so persons or objects inside these structure could not be distinguished. Yet the idea of a UAV peering into the toilets was enough to make it considered sensitive, even in the blurred orthomosaic images.

Other objects which were commonly named were wetland, drainage, and buildings (Figure 7.2d). Even in the vector map, 12–15% of the respondents did not want maps with these objects to be distributed. One interpretation of these results leads back to the flooding issues in the region. Flooding is a large problem in the city, and construction in the wetlands is restricted by the government. The delineation of buildings combined with the mapping of flooded areas is therefore an understandable cause of concern for the inhabitants of these buildings.

7.4 Discussion

7.4.1 Privacy, unintended usage, empowerment, and trust

The results of the questionnaires illustrate the importance of local context regarding privacy issues. Both case studies appear quite similar—both are related to flying UAVs over deprived areas for mapping purposes. Yet the objects considered as private by the local communities are quite different. Some concerns regarding sensitive objects and data distribution are linked to unintended usage of the UAV data. Residents in Kigali are concerned with being confronted with 'messy' roofs and backyards and the shame associated with having such transgressions being available to the wider public. In Dar es Salaam and Kigali there are concerns regarding expropriation. These uses are not the direct motivation of the UAV mapping activities, but may be perceived as such. We identify three strategies for addressing sensitive objects in UAV

data products (i) *avoidable*; (ii) *unavoidable but removable*; and (iii) *unavoidable and irremovable*, and describe how these strategies are related to the potential misuse of the data (Table 7.2).

Table 7.2: *Categories of sensitive objects and possible strategies to address residents' concept of sensitive objects.*

Sensitive Object Characteristics	**Examples**	**Privacy Protection Strategy**	**Geodata Containing Objects**
Avoidable	Rubbish Cars	Notify residents before flights Notify residents of image capture	None
Unavoidable but removable	Roof quality Toilets	Blur orthoimagery	Raw images
Unavoidable and irremovable	Houses in wetlands	Identify/mitigate potential misuse	All data products

Some objects which are considered sensitive are *avoidable*. For example, the sensitivity of rubbish on roofs and in back yards in Kigali. The presence of these objects in the imagery can be controlled by the local residents if they are informed the flights will take place and have an understanding of the implications of the UAV flights. From both case studies, it was clear that very few residents were aware the flights occur. Fifty-nine percent of the residents who saw the flights in Kigali realized the UAV was taking pictures or videos, though this was only 25% in Dar es Salaam. Ensuring that residents understand the observation capabilities of UAVs and when flights will take place also empower them to control which objects are visible in the UAV image. This is also the best way to protect sensitive objects against unintended uses, as they are not captured by the camera in the first place. On the other hand, it could be argued that in this case the privacy of space is improved at the cost of the privacy of behavior as the form of overt surveillance stifles illegal or discouraged behavior (Clarke, 2014).

In other cases, the sensitive objects are *unavoidable but removable*. Their presence in the images will be difficult to control by the local residents. Here, the privacy of the residents can be controlled through blurring certain parts of the image and / or restricting the data distribution. Indeed, there are reports of operators blurring people or cars in their UAV images before distribution, but coordinated efforts are difficult due to a lack of clear guidelines (Finn and Wright, 2016). The roofless toilets in Tanzania would fall into this category. This mode of respecting the privacy of the residents, however, depends on the ethical conduct of the UAV operators and requires trust between the institutions. It may also require a formal procedural check of image content to be made before public release and distribution. It has been noted that such self-regulation does not sufficiently protect the privacy of residents (Clarke,

2014). Limiting the distribution of data may serve to respect the privacy of the inhabitants and prevent misuse of the data for other purposes. However, the raw UAV data will likely be stored, and may therefore be (mis)used in the future for other purposes.
The third category of sensitive objects are *unavoidable and irremovable*. Their presence may be considered as private even when presented in vector format. The presence of buildings in flood-prone areas of Dar es Salaam is the most prominent example. This is the most difficult category to address, as the objects considered as private are directly related to the objects of interest for the mapping activities. Due to the scale of these objects, this is not a problem specific to the use of UAV imagery, as the data could be obtained through high-resolution satellite imagery or other participatory mapping applications. Indeed, the decision of which objects to put on a map may be a politicized decision. For example, informal settlements are sometimes purposefully excluded from official statistics (Carr-Hill, 2013).

7.4.2 Collaboration, transparency, and accountability

A map is politicized as there is always a person or body responsible for deciding which information is represented in the map. Once this information is categorized, it leaves little space for alternative interpretations. Imagery, however, can be interpreted in different ways. In the case of urban upgrading projects, the high-resolution imagery improved communication between stakeholders as it created a common visual format for communications. In Kigali, displaying the planned project interventions over the UAV imagery helped community members identify, locate and discuss issues in the area. Consultants could explain where planned project interventions are located and why, which improves the transparency behind the decisions of the upgrading interventions.

As the appearance of the settlement changes with time, the imagery provides concrete evidence of a previous state of the settlement. For example, when there are legal frameworks which guarantee citizens a right to compensation in case of expropriation, the image proves the existence of a dwelling, enables its size to be measured, and perhaps conveys some information on its materials and quality. Any building or plot characteristics present in the image could be used for valuation purposes. In Kenya, a community-based enumeration campaign was initiated to provide slum dwellers with documentation to give them legal protection in the case of expropriation (SDI, 2016). It is easy to imagine an application where UAVs are utilized to provide such information at a larger scale.

7.4.3 Equity and participation

An 'ethical usage' of UAV imagery depends not only on the mitigation of negative issues such as privacy, but also on a more equitable distribution of the benefits. Although geospatial technologies may support informed decision making regarding urban issues, it is important to ensure that those with no access to the knowledge nor the capacity to use it are not excluded (Pfeffer and Verrest, 2016).

Providing hard copies of the UAV imagery to local governmental offices is a successful way to return the information to community members. In Kigali, 70% of the interviewees could name a use case for the aerial imagery presented in the questionnaire. Out of these, 61% were examples of how the images could be used by businessmen or the general public. This means that 43% of the community members interviewed in Kigali could think of how the images could be useful for the general population. In Dar es Salaam, 46% of the respondents named examples of how the images could be useful for the general public. However, 13% of the respondents also indicated that the community members should be trained to use the images. It remains a question to which degree such training should accompany the distribution of the hardcopy maps to the community.

One key difference between the maps produced from the UAV imagery is *how* the map layers are generated. In Kigali, the GIS data was generated by institutions and consultants, whereas in Dar es Salaam the maps were created through Participatory GIS (PGIS) methods. The purpose of PGIS is to empower the community by providing them with control and access to spatial information (Rambaldi, Kyem, *et al.*, 2006). For example, in Dar es Salaam it was observed that involving community members in the collection of the spatial information also increased the awareness and responsibility amongst community members regarding possible flood mitigation levels at a local level (Minja, Iliffe and Anderson, 2016). One of the main differences between PGIS and Volunteered Geographical Information (VGI) is that PGIS involves the community in the process acquiring geospatial knowledge, whereas VGI combines the data obtained by the community (Verplanke *et al.*, 2016). As UAVs become more ubiquitous and cheaper, it is feasible to imagine community-based UAV mapping campaigns. This would involve community members in the process of UAV data capture, perhaps transferring some of the benefits of PGIS to UAV workflows.

7.4.4 Implications for policy development

The strategies identified in Table 2 can provide recommendations for policy development. *Avoidable* sensitive objects are closely related to residents being aware of the impending UAV activities and their implications. Relevant policy

therefore falls into the realm of legislation regarding UAV flight operations and the ethical responsibility of the pilot or organization responsible for the flights. Ethical policy frameworks and practices by the UAV operators themselves can (i) ensure that citizens are aware of the operations and consequences and (ii) limit the capture of unnecessary data—especially when this data is considered sensitive such as the inadvertent recording of persons. It is acknowledged that the former can be difficult in practice, in particular for large-scale UAV operations. Examples of these ethical practices are already included in UAV legislation in a number of countries (Stöcker *et al.*, 2017) as well as unofficial recommendations and guidelines for UAV operators (Finn *et al.*, 2014; UAViators, 2017).

Unavoidable objects (both *removable* and *irremovable*) fall into the domain of data protection and distribution policies. Good practices include ensuring securing of the data and avoiding storing unnecessary private information (Finn *et al.*, 2014). Cultural implications are likely to play a larger role in this group of policies compared to the operational policies discussed above. As illustrated by the two case studies presented here, the concept of which objects are considered as sensitive and which parties should have access to the data can differ greatly—even amongst seemingly similar situations. Sharing the data amongst multiple stakeholders can support the ethical values of transparency, accountability, and equity, though it is important to ensure that the voices of residents (the viewed) are incorporated in the decisions around exactly which stakeholders have access rights and what those rights entail. Due to the cultural nuances to the perceptions of data protection and sharing, it is important to include various stakeholders—especially the individuals whose persons or property is captured by the UAV data – to actively participate in the development of such policies.

7.5 Conclusions

The recognition of UAVs as a powerful geospatial information acquisition tool is ubiquitous. Concerns regarding their ethical usage are also on the rise. The social benefit of the high-resolution and comparatively low-cost geospatial information obtained through UAV platforms compared to more traditional methods must be weighed against social concerns such as privacy. This can be quite challenging depending on the degree to which the intentions of the mapping exercise are in conflict with objects considered as sensitive in the local context. Through an investigation of two case studies using UAVs to support urban upgrading projects, we contribute towards a better understanding of these issues and provide a framework with concrete ways of how to address them.

The use of two case studies illustrates the diversity of the concept of privacy in different contexts. You would expect similar questionnaire responses as both case studies use a UAV to map deprived areas for urban upgrading purposes in an East African context. However, the answers reveal that residents of both areas have strikingly different concepts of privacy—both regarding the objects which are considered sensitive as well as the level of abstraction in which these objects are still considered private. This underlines the importance of understanding the local context to respect the privacy.

Residents' privacy and the geospatial interests of the party collecting the UAV data do not necessarily need to cause problems. *Avoidable* objects which may be considered as private, such as rubbish on the roofs, may not necessarily be of interest for base-data collection in an informal settlement. If residents are notified of impending flights and aware that the flights imply the collection of images over their property, they have the power to move or cover these transient objects. Thus, they are empowered to control the collected data itself. Sensitive objects will not be captured in the raw data and the privacy is therefore ensured. This requires that residents understand that the UAV is collecting imagery. Although UAVs are a new technology, they are increasingly being used for various purposes; indeed the questionnaire results indicated that residents have already observed UAVs in multiple contexts. Therefore, it is expected that with time, citizens will become increasingly aware of the implications of a UAV flying over their property.

Unavoidable but removable objects which are not of specific interest to the mapping activities may be blurred in the orthoimagery before distribution. Assuming that the party conducting UAV flights bears good will and that there is trust between residents and this party, it is feasible to mitigate privacy concerns by defining guidelines for the level of abstraction and distribution. For example, toilets could be considered as sensitive in the imagery, but not in a vector map (see Figure 2). It is clear that these preferences are greatly dependent on the local context. So, how many people would need to be interviewed? How can one ensure that the respondents will feel free to express their opinions? How many individuals must present a certain view in order to sway the blurring and distribution guidelines? The practical execution of this ethical strategy is challenging, though PGIS solutions could assist.

The largest challenge is when *unavoidable and irremovable* objects of direct interest for the UAV mapping activities are those which are considered sensitive by the local population. For example, houses in wetlands may be considered sensitive even at a very high level of abstraction, such as a single point on a vector map. One could argue that, at least in these two study areas, such *unavoidable and irremovable* objects are so large that they are visible in satellite imagery anyway. As such, the sensitivity concerns are not limited to

the use of UAV imagery, but rather they apply to GIS data in general. Furthermore, in the present cases, the activities are considered sensitive by the locals because they oppose local (informal or formal) norms. Still, it is important to consider how to protect the residents' concerns in such cases. One strategy is to identify likely threats of data misuse and ensuring protective governance frameworks. The potential threat of being mapped may turn into evidence for ensuring citizens' rights in the presence of protective legislative guidelines. Furthermore, the increased availability of low-cost UAVs as well as further developments towards automated and open source photogrammetric software, such as OpenDroneMap, suggest that community-driven UAV mapping activities are feasible in the future. This would truly enable citizens to reap the benefits of the high resolution geospatial data for their own purposes—although legal flights will still be regulated by national authorities.

Chapter 8 - Synthesis

8.1 Conclusions per objective

The main objective of this manuscript was to analyze the suitability of UAVs to support informal settlement mapping projects by: analyzing synergies between the 2D and 3D data it provides, adapting machine learning methods, extracting DTMs, observing its use in actual case studies, and discussing the social impact. The following section describes the main conclusions per objective, and the next section reflects upon the implications of this work and discusses future investigations.

1) Identifying synergies between 2D and 3D information provided by UAVs
   Results indicated that the highest classification accuracies (above 90% for the selected datasets) were obtained when combining 2D features from the image with 3D features from the point cloud. Important image-based features included color and texture information summarized over image segments (super-pixels). The high spatial resolution of the UAV imagery was very beneficial here, as it captured the regular pattern of the roof materials (e.g., corrugated iron sheets). This allowed the classifier to correctly identify buildings, despite the differences in quality and color of roofs in the settlement. Important 3D features include: the ratio between the eigenvalues of neighboring pixels projected onto a 2D plane (low values indicate the points are more evenly distributed horizontally, whereas higher values suggest more linear structures such as walls); the variation of height of points in the same 2D neighborhood; and the maximum height of planar segments above neighboring points. The latter feature is important because of the heterogeneity of the roofs in informal settlements. Incremental roof upgrading may cause a single roof to have many different colored iron sheets. Image-based segmentation methods will not be able to capture the larger roof extent. However, if we look at the same building, a single roof will (usually) have a relatively planar structure. The 3D segmentation will therefore be more likely to include the entire roof as a single segment. Giving the maximum height above the surrounding neighbors is a good indicator of building roofs in sloped environments. Such planar segments can also be extracted from the DSM, but errors in the DSM interpolation are propagated to the classification results. The results of this chapter therefore show the usefulness of getting features directly from the point cloud.

2) Adapting supervised classification methods to deal with heterogeneous data
   Chapter 3 clearly illustrates the suitability of MKL for the classification of heterogeneous features from UAV datasets. Furthermore, we address an important gap in MKL research. Whereas previous studies analyze how to weigh and combine the different kernels, there was little understanding regarding how to group the features appropriately. Our contribution

developed a data-driven, automatic grouping strategy, which had a higher performance than previous grouping strategies for six different MKL methods.

3) Analyzing how reliable training labels can be obtained from existing geospatial data
The experiments indicate that both local and global cues are important for identifying mislabeled training samples. A methodology was developed to remove unreliable training samples obtained from existing geospatial data. A local criterion penalizes image segments which have a similar appearance to their neighbors, but a different class label. This is especially useful for label errors due to misalignments between the outlines in the outdated data providing the labels and the newly obtained (UAV) imagery. Secondly, a global contextual criterion checks the similarity between a sample's features and other such examples distributed over the entire dataset. This proves to be especially effective for errors caused by changes such as building construction or demolition. By combining both local and global cues, unreliable samples are flagged in classification results. These samples can be removed from the training set, leaving the more reliable samples to train another, improved, classification model. In this dataset, such a strategy showed that classification accuracies of above 90% could be obtained despite 30% of the initial training samples being mislabeled. These results implicate a considerable speed-up in the whole process of UAV image classification is possible by using outdated maps to provide training labels, and bypassing the need to manually label samples. Such strategies could also be used to control the quality of, e.g. large-scale digitization campaigns.

4) Analyzing how to extract Digital Terrain Models in challenging settings
A DTM extraction method specifically tailored to UAV datasets was developed. A rule-based labeling strategy was defined to automatically generate training labels from a dataset; this significantly speeds up the classification workflow while obtaining a similar final accuracy as when manually labeled samples are used. The deep learning architecture itself is fully convolutional and utilized dilated filters to consider a larger area while limiting the number of parameters to train. Even though the proposed network is shallower than those designed for other computer vision applications, it obtains similar performance while significantly lowering the training time. Furthermore, we again demonstrate how the integration of image-based information is a valuable addition to the geometric information. Indeed, the proposed method outperforms reference DTM extraction methods which only make use of geometric information and fail in the challenging informal settlement datasets.

5) Identifying opportunities of UAVs to support urban upgrading workflows
Consultation with various stakeholders in the urban upgrading project in Rwanda indicated that even without any advanced machine learning techniques, the UAV images were highly valued. The detail of the imagery facilitated the digitization and mapping. Additional information required for upgrading projects which are not typically identifiable in high-resolution imagery, such as solid waste accumulations, was visible. This greatly facilitated field work, by providing a more realistic and detailed view of the settlement and facilitating navigation of the complex network of footpaths. Finally, the images themselves were beneficial as a communication platform. More specifically, they facilitate engagement, interaction, and communication between the slum residents and other project stakeholders and the identification of vulnerable areas and prioritized interventions.

6) Analyzing the social impacts of using UAVs in the context of urban upgrading projects
Some concerns regarding the use of UAVs is the possible capture of objects considered as private in the imagery, the distribution of this sensitive information, and possible misuse of it. Indeed, investigations into two apparently similar projects indicate that perceptions are dependent on the local culture and context. In the ideal situation, the objects which are considered as sensitive can be removed or hidden by the residents themselves. This is the highest form of empowerment, allowing the citizens to control which data is captured. To support this, local leaders (e.g., Cell Executive Officers in Kigali and Ward Officers in Dar es Salaam) should be notified of upcoming flights and their purposes so they can communicate these to the residents. Bringing examples of imagery to the field when conducting the flights helps explain the purposes of the activities and alleviate some concerns. Other communication channels such as local radios can also be used to notify residents of upcoming flights so they can prepare accordingly. However, even if sensitive objects are captured in the imagery, the detail of the sensitive objects is not always needed for the objectives of the UAV flights. In this case, these objects could be blurred before distribution to other parties. Residents could then be involved in defining such data distribution rules. Adequate data distribution policies can support equity, as providing residents with access to the imagery is an opportunity for more equitable distribution of the benefits. For example, the field work in Kigali indicated that even if residents found it difficult to identify uses for the imagery when they were first collected in 2015, having access to the printed images at the cell offices enabled them to use the images for their own purposes. Two years later, residents were observed to be using these printed maps for some unanticipated purposes such as identifying

desirable areas to buy a house. Unfortunately, sometimes the objects considered as sensitive are exactly the information required by the upgrading project. In these cases, it is important to understand the reasons behind the residents' concerns and develop adequate protective policies. For example, some residents were concerned of expropriation if their houses were visible on a map. This underlines the need for suitable expropriation practices and not only data protection and distribution policies of the UAV imagery itself.

8.2 Reflections and outlook

The motivation behind the proposed research was, on the one hand, the extraction of geospatial information to support urban upgrading processes, and the second was on analyzing the suitability of UAVs as a data collection platform. The first aspect was addressed through detailed investigations using machine learning. A wide range of Machine Learning algorithms are available and have been addressed in this research. Understanding the strengths, weaknesses, and suitability of these algorithms for various applications is important. For example, SVM is especially suited for applications with many features, but relatively few samples are available or are costly to obtain. The use of MKL can make SVM especially suitable for heterogeneous datasets. This is why these were selected for the 2D vs. 3D feature analyses in Chapters 2 and 3. On the other hand, random forests are more suitable when very large and noisy training datasets are available, and fast model training is required. This was the case when iteratively using classification results to remove potentially noisy labels in Chapter 4. Deep learning methods are proving to be extremely powerful classifiers, but are costly to train. These are then more applicable for situations where many labeled training samples are available and training time is less important such as the DTM extraction in Chapter 5. Further research may identify how more complex 3D characteristics (which were more accurate than using DSM-based geometric information in Chapter 2) can be integrated into deep learning methods.

Taking a step back, how can we place this work on machine learning in society? The case studies in Chapters 6 and 7 indicated that the considered upgrading projects are still making use of manual digitization to map informal settlements. Why is this being done if the scientific research presented here and elsewhere claims that faster, (semi-)automatic methods are available which achieve accuracies of above 90%? In practice, machine learning requires more expertise and understanding than manual digitization. Skills which are not always available and even when they are, the produced maps are not always fit for purpose – even a 90% accuracy means that one in ten buildings will be incorrect. For applications such as cadaster or identifying which households must be expropriated, this is too high. However, manually

correcting 10% of the building outlines will still be faster than mapping everything from scratch and it is still quite useful for getting an idea of how many buildings there are in a neighborhood at the onset of the project (if the FP and FN are in balance). Better access to machine learning methods as they become increasingly available in specialized and non-specialized software and web applications and the concurrent increase in public awareness of such automatic methods suggest that workflows which combine machine learning with manual digitization will become more common in the near future.

Technological constraints such as limited flight times, legislative regulations such as within-line-of-sight constraints, and the big data challenges arising from the sheer size of the data captured at such a high resolution – all seem to indicate that UAVs are more suitable for limited study areas. Indeed, one of the main assumptions underlying this research design was that UAVs are particularly useful for upgrading projects due to the limited extent of a project area. This may seem at odds with the focus on machine learning techniques, whose (limited) inaccuracies may not exploit the full advantage of the higher resolution. However, a trained machine learning algorithm could provide a quick overview of the settlement at the initial reconnaissance stage and for monitoring purposes. The diversity of settlement characteristics implies that models will likely need to be retrained for different study areas, using either defined rules (as in Chapter 5) or existing vector data (as in Chapter 4), to provide initial training samples. Also, recent projects (mainly using fixed-wing UAVs) are collecting imagery of entire cities or islands[14] (Makoye, 2017). Research regarding how to efficiently process large UAV datasets is needed. Further investigations could also consider the interaction between manual methods, which enable a participatory approach and may provide a higher accuracy, and automatic methods which may be faster but currently require more expertise and are sensitive to hyperparameter tuning. Developments in large, holistic machine learning models using global training data coupled with widespread exposure to accessible GIS information (e.g., increased internet usage, mobile mapping, etc.) may bring such automated methods closer to the general community in the future.

The second underlying objective of this research was to consider the suitability of UAVs as a data collection platform – both regarding the geospatial information it can provide as well as from a societal context. The simultaneous acquisition of 2D and 3D data is useful for classification (Chapter 2 and 3) and DTM extraction (Chapter 5). Observations from the use of the images in the case study areas indicate that they are perceived as very useful by experts and to a limited degree by the residents of the settlements (discussed in Chapters 6 and 7).

[14] http://www.zanzibarmapping.com/

Regarding the societal context, the UAV flights conducted in 2015 was amongst one of the first UAV activities in Rwanda. Collaborating with the upgrading project and the enthusiasm of the local stakeholders indicated some of the practical challenges. One is the lack of or too strict regulations. Legislation and best practices are in various stages of development, and national aviation agencies are making clear efforts to make citizens aware of the rules connected to the use of cheap, off-the-shelf UAVs which are often considered as toys. Another challenge is posed by the high computing requirements of processing the images. Although both proprietary and open source software are developing relatively automated workflows, image processing, and computing remains a challenge and will likely remain a bottleneck as the number and size of UAV datasets continues increasing. Further developments which utilize cloud-based services or other smart strategies to reduce the computing requirements are likely to remain important as project sizes increase.

In sum, the contribution of this dissertation is two-fold. Firstly, it illustrates how machine learning methods can be manipulated to take advantage of UAV dataset characteristics and provide detailed, up-to-date geospatial information. Secondly, involvement with existing upgrading projects sheds light on the actual and perceived usage of the information provided by the UAV images as well as the societal context of using UAVs to map informal areas. As UAVs become cheaper, more automated and widespread; as the understanding of how to extract information from the data develops; and as the knowledge regarding UAV flight operations and relevant data processing workflows spreads – it is easy to imagine a future where UAVs play a more prominent role as a geospatial information acquisition tool to support urban upgrading projects.

A number of investigations can be done to bring this future closer. Firstly, smart ways of processing and storing large UAV datasets are needed. As UAV platforms are improving and more sensors are being added, the legislative framework is slowly taking shape, and photogrammetric software is becoming more automated – one important remaining bottleneck is how to efficiently process datasets of increasing sizes, especially considering inconsistent internet connectivity and limited computing requirements which are (at the moment) still commonality in many developing countries. Secondly, repeated UAV acquisitions will require further investigations into change detection methods. Synergies between periodically acquired satellite imagery and targeted UAV image acquisitions are especially interesting. Thirdly, investigations can target the extraction of additional information from the imagery. For example, identifying areas where solid waste accumulates, existing utilities, Floor Area Ratios, housing construction material, and a 3D understanding of building usage (as opposed to land use). A list of such indicators is given in Figure 1-1, but now that the first practical examples of

the use of UAV images by urban planners and other stakeholders exist it would be an opportune moment to consult these stakeholders on actual and perceived usage and update this list. Finally, the integration of this additional information into GIS analyses and 3D visualizations may improve decision-making workflows. For example, identifying cars in UAV imagery can be used to induce and plan transportation networks, the solid waste accumulation can be used to plan more effective waste collection services, an understanding of plot characteristics and locally preferred building characteristics can help propose building designs to individual land owners. The results of such analyses can be put into 3D visualizations of the neighborhood to further encourage the dialog between residents, engineers, and planners and design more suitable upgrading measures.

Bibliography

Abbott, J. (2002) 'A method-based planning framework for informal settlement upgrading', *Habitat International*, 26(3), pp. 317–333. doi: 10.1016/S0197-3975(01)00050-9.

Abbott, J. (2003) 'The use of GIS in informal settlement upgrading: its role and impact on the community and on local government', *Habitat International*, 27(4), pp. 575–593. doi: Doi 10.1016/S0197-3975(03)00006-7.

Achanta, R. *et al.* (2012) 'SLIC superpixels compared to state-of-the-art superpixel methods', *IEEE transactions on pattern analysis and machine intelligence*. IEEE, 34(11), pp. 2274–2282.

Arefi, H. and Hahn, M. (2005) 'A hierarchical procedure for segmentation and classification of airborne LIDAR images', in *International Geoscience and Remote Sensing Symposium*, p. 4950.

Arimah, B. C. (2010) *The Face of Urban Poverty: Explaining the Prevalence of Slums in Developing Countries*. Working paper//World Institute for Development Economics Research. doi: 978-92-9230-265-8.

AUC (2015) *Agenda 2063: The Africa we want*.

Audebert, N., Saux, B. Le and Lefèvre, S. (2016) 'Semantic Segmentation of Earth Observation Data Using Multimodal and Multi-scale Deep Networks', *arXiv preprint arXiv:1609.06846*.

Axelsson, P. (2000) 'DEM generation from laser scanner data using adaptive TIN models', *International Archives of Photogrammetry and Remote Sensing*, 33(B4/1; PART 4), pp. 111–118.

Bach, F. (2007) 'Consistency of the group Lasso and multiple kernel learning', *The Journal of Machine Learning Research*, 9, pp. 1179–1225.

Bachofer, F. (2016) 'Assessment of building heights from pléiades satellite imagery for the Nyarugenge sector, Kigali, Rwanda', *Rwanda Journal*. University of Rwanda, 1(1S).

Baiocchi, V. *et al.* (2014) 'Development of a software to optimize and plan the acquisitions from UAV and a first application in a post-seismic environment', *European Journal of Remote Sensing*, 47, pp. 477–496.

Banes, C. (2015) *Upgrading of Unplanned Settlements in Urban Areas of Rwandan Cities - Guidance Note for Identification, Survey, Planning, Design, Implementation, Monitoring, Maintenance and Project Management*. Kigali, Rwanda.

Bansal, A. *et al.* (2017) 'PixelNet: Representation of the pixels, by the pixels, and for the pixels'.

Bao, S. Y. *et al.* (2013) 'Dense object reconstruction with semantic priors', in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1264–1271. doi: 10.1109/CVPR.2013.167.

Barry, M. and Rüther, H. (2005) 'Data collection techniques for informal settlement upgrades in Cape Town, South Africa', *URISA Journal*, 17(1),

pp. 43–52.

Baud, I. *et al.* (2010) 'Understanding heterogeneity in metropolitan India: The added value of remote sensing data for analyzing sub-standard residential areas', *International Journal of Applied Earth Observation and Geoinformation*, 12(5), pp. 359–374. doi: 10.1016/j.jag.2010.04.008.

Baud, I. *et al.* (2014) 'Digital and spatial knowledge management in urban governance: Emerging issues in India, Brazil, South Africa, and Peru', *Habitat International*, 44, pp. 501–509. doi: 10.1016/j.habitatint.2014.09.009.

Beumier, C. and Idrissa, M. (2016) 'Digital terrain models derived from digital surface model uniform regions in urban areas', *International Journal of Remote Sensing*, pp. 1–17. doi: 10.1080/01431161.2016.1182666.

Birriel, P. and González, R. (2015) 'UAV, a Tool for Urbanism', *GIM International*, 29, pp. 15–17.

Blaha, M. *et al.* (2016) 'Large-Scale Semantic 3D Reconstruction: An Adaptive Multi-resolution Model for Multi-class Volumetric Labeling', in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3176–3184. doi: 10.1109/CVPR.2016.346.

Blaschke, T. (2010) 'Object based image analysis for remote sensing', *ISPRS journal of photogrammetry and remote sensing*. Elsevier, 65(1), pp. 2–16.

Boykov, Y. Y. and Jolly, M.-P. (2001) 'Interactive graph cuts for optimal boundary & region segmentation of objects in ND images', in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*. IEEE, pp. 105–112.

Breiman, L. (2001) 'Random Forests', *Machine Learning*, 45(1), pp. 5–32. doi: 10.1023/A:1010933404324.

Bruzzone, L. and Carlin, L. (2006) 'A multilevel context-based system for classification of very high spatial resolution images', *Geoscience and Remote Sensing, IEEE Transactions on*, 44(9), pp. 2587–2600.

Bruzzone, L. and Persello, C. (2009) 'A novel context-sensitive semisupervised SVM classifier robust to mislabeled training samples', *IEEE Transactions on Geoscience and Remote Sensing*. IEEE, 47(7), pp. 2142–2154.

Bruzzone, L. and Persello, C. (2010) 'Approaches based on Support Vector Machines to classification of remote sensing data', in *Handbook of Pattern Recognition and Computer Vision*. World Scientific, pp. 329–352.

Cabezas, R., Straub, J. and Fisher, J. W. (2015) 'Semantically-aware aerial reconstruction from multi-modal data', in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2156–2164. doi: 10.1109/ICCV.2015.249.

Carr-Hill, R. (2013) 'Missing Millions and Measuring Development Progress', *World Development*, 46, pp. 30–44. doi: 10.1016/j.worlddev.2012.12.017.

Castelluccio, M. *et al.* (2015) 'Land Use Classification in Remote Sensing

Images by Convolutional Neural Networks', *CoRR*, abs/1508.0.

Chang, C.-C. and Lin, C.-J. (2011) 'LIBSVM', *ACM Transactions on Intelligent Systems and Technology*, 2(3), pp. 1–27. doi: 10.1145/1961189.1961199.

Chaplot, V. *et al.* (2006) 'Accuracy of interpolation techniques for the derivation of digital elevation models in relation to landform types and data density', *Geomorphology*, 77(1–2), pp. 126–141. doi: 10.1016/j.geomorph.2005.12.010.

Chehata, N., Guo, L. and Mallet, C. (2009) 'Airborne Lidar feature Selection for urban classification using Random Forests', in *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 207–212.

Chen, B. *et al.* (2016) 'Building change detection with RGB-D map generated from UAV images', *Neurocomputing*. Elsevier, 208, pp. 350–364.

Chen, J. and Zipf, A. (2017) 'DeepVGI: Deep Learning with Volunteered Geographic Information', in *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, pp. 771–772.

Chen, L.-C. *et al.* (2015) 'Semantic image segmentation with deep convolutional nets and fully connected CRFs', in *International Conference on Learning Representations (ICLR 2015)*. San Diego.

Clarke, R. (2014) 'The Regulation of of the Impact of Civilian Drones on Behavioural Privacy', *Computer Law & Security Review*. Elsevier, 30(4), pp. 286–305. doi: https://doi.org/10.1016/j.clsr.2014.03.007.

Cohen, M. (2013) *The City is missing in the Millennium Development Goals*. Working Paper Series, Harvard Francois-Xavier Bagnoud Center for Health and Human Rights, Boston.

Colomina, I. and Molina, P. (2014) 'Unmanned aerial systems for photogrammetry and remote sensing: A review', *ISPRS Journal of Photogrammetry and Remote Sensing*. International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS), 92, pp. 79–97. doi: 10.1016/j.isprsjprs.2014.02.013.

Comaniciu, D. and Meer, P. (2002) 'Mean shift: a robust approach toward feature space analysis', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5), pp. 603–619. doi: 10.1109/34.1000236.

Consortium, R. P. (2015) *Rapid Planning - Sustainable Infrastructure, Environmental, and Resource Management for Highly Dynamic Metropolises*. Edited by D. Steinbach et al. Rapid Planning Consortium.

Cortes, C., Mohri, M. and Rostamizadeh, a (2009) 'Learning non-linear combinations of kernels', *Advances in Neural Information …*, pp. 1–9.

Cortes, C., Mohri, M. and Rostamizadeh, A. (2010) 'Two-stage learning kernel algorithms', in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 239–246.

Cristianini, N. *et al.* (2002) 'On kernel-target alignment', *Advances in Neural*

Information Processing Systems, 14.

Culver, K. B. (2014) 'From Battlefield to Newsroom: Ethical Implications of Drone Technology in Journalism', *Journal of Mass Media Ethics*, 29(1), pp. 52–64. doi: 10.1080/08900523.2013.829679.

Davidson, F. and Payne, G. (1983) *Urban Projects Manual*. Liverpool: Liverpool University Press.

Debella-Gilo, M. (2016) 'Bare-earth extraction and DTM generation from photogrammetric point clouds including the use of an existing lower-resolution DTM', *International Journal of Remote Sensing*, 37(13), pp. 3104–3124. doi: 10.1080/01431161.2016.1194543.

Demantké, J. *et al.* (2012) 'DIMENSIONALITY BASED SCALE SELECTION IN 3D LIDAR POINT CLOUDS', *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XXXVIII-5/, pp. 97–102. doi: 10.5194/isprsarchives-XXXVIII-5-W12-97-2011.

Demir, B., Persello, C. and Bruzzone, L. (2011) 'Batch-mode active-learning methods for the interactive classification of remote sensing images', *IEEE Transactions on Geoscience and Remote Sensing*. IEEE, 49(3), pp. 1014–1031.

Di, W. and Crawford, M. M. (2012) 'View generation for multiview maximum disagreement based active learning for hyperspectral image classification', *IEEE Transactions on Geoscience and Remote Sensing*, 50(5 PART 2), pp. 1942–1954.

Doshi, N. P. and Schaefer, G. (2012) 'A comprehensive benchmark of local binary pattern algorithms for texture retrieval', in *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, pp. 2760–2763.

Eichleay, M. *et al.* (2016) *Using Unmanned Aerial Vehicles for Development: Perspectives from Citizens and Government Officials in Tanzania*. Durham, USA.

Elmqvist, M. *et al.* (2001) 'Terrain modelling and analysis using laser scanner data', *International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences*, 34(3/W4), pp. 219–226.

Farabet, C. *et al.* (2013) 'Learning hierarchical features for scene labeling', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), pp. 1915–1929. doi: 10.1109/TPAMI.2012.231.

Feng, Q., Liu, J. and Gong, J. (2015) 'Urban flood mapping based on unmanned aerial vehicle remote sensing and random forest classifier-A case of yuyao, China', *Water (Switzerland)*. doi: 10.3390/w7041437.

Finn, R. *et al.* (2014) 'Study on privacy, data protection and ethical risks in civil Remotely Piloted Aircraft Systems operations', *Final Report, Luxembourg: Publications Office of the European Union*.

Finn, R. L. and Wright, D. (2016) 'Privacy, data protection and ethics for civil drone practice: A survey of industry, regulators and civil society organisations', *Computer Law & Security Review*, 32(4), pp. 577–586. doi:

10.1016/j.clsr.2016.05.010.

Finn, R. L., Wright, D. and Friedewald, M. (2013) 'Seven types of privacy', in *European data protection: coming of age*. Springer, pp. 3–32.

Flipse, S. M., van der Sanden, M. C. A. and Osseweijer, P. (2013) 'The Why and How of Enabling the Integration of Social and Ethical Aspects in Research and Development', *Science and engineering ethics*, 19(3), pp. 703–725.

Folleco, A. *et al.* (2008) 'Software quality modeling: The impact of class noise on the random forest classifier', in *Evolutionary Computation, 2008. CEC 2008.(IEEE World Congress on Computational Intelligence). IEEE Congress on*. IEEE, pp. 3853–3859.

Foody, G. M. (2004) 'Thematic Map Comparison: Evaluating the Statistical Significance of Differences in Classification Accuracy', *Photogrammetric Engineering & Remote Sensing*, 70(5), pp. 627–633. doi: 10.14358/PERS.70.5.627.

Frenay, B. and Verleysen, M. (2014) 'Classification in the Presence of Label Noise: A Survey', *IEEE Transactions on Neural Networks and Learning Systems*. IEEE, 25(5), pp. 845–869. doi: 10.1109/TNNLS.2013.2292894.

Furukawa, Y. and Ponce, J. (2010) 'Accurate, dense, and robust multiview stereopsis', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8), pp. 1362–1376.

Gallego, M. *et al.* (2013) 'Tabu search with strategic oscillation for the maximally diverse grouping problem', *The Journal of the Operational Research Society*. Palgrave Macmillan Journals on behalf of the Operational Research Society, 64(5), pp. 724–734.

Galleguillos, C. and Belongie, S. (2010) 'Context based object categorization: A critical survey', *Computer vision and image understanding*. Elsevier, 114(6), pp. 712–722.

Gehler, P. and Nowozin, S. (2009) 'On feature combination for multiclass object classification', in *Computer Vision, 2009 IEEE 12th International Conference on*, pp. 221–228. doi: 10.1109/ICCV.2009.5459169.

Gevaert, C. *et al.* (2016) 'Integration of 2D and 3D features from UAV imagery for informal settlement classification using Multiple Kernel Learning', in *International Geoscience and Remote Sensing Symposium (IGARSS)*. doi: 10.1109/IGARSS.2016.7729385.

Gevaert, C. *et al.* (2016) 'Opportunities for UAV mapping to support unplanned settlement upgrading', *Rwanda Journal*. University of Rwanda, 1(1S).

Gevaert, C. M. *et al.* (2015) 'Generation of Spectral-Temporal Response Surfaces by Combining Multispectral Satellite and Hyperspectral UAV Imagery for Precision Agriculture Applications', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(6). doi: 10.1109/JSTARS.2015.2406339.

Gevaert, C. M. *et al.* (2016) 'Classification of informal settlements through the integration of 2D and 3D features extracted from UAV data', *ISPRS Annals*

of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Copernicus GmbH, 3/WG3, pp. 317–324.

Gevaert, C. M. *et al.* (2017) 'Informal settlement classification using point-cloud and image-based features from UAV data', *ISPRS Journal of Photogrammetry and Remote Sensing*. Elsevier, 125, pp. 225–236. doi: 10.1016/j.isprsjprs.2017.01.017.

Gevaert, C. M., Persello, C. and Vosselman, G. (2016) 'Optimizing Multiple Kernel Learning for the Classification of UAV Data', *Remote Sensing*. Multidisciplinary Digital Publishing Institute, 8(12), p. 1025.

Gilbert, A. (2007) 'The Return of the Slum: Does Language Matter?', *International Journal of Urban and Regional Research*, 31(4), pp. 697–713. doi: 10.1111/j.1468-2427.2007.00754.x.

GISTech Consultants LTD (2015) *Project Brief for Upgrading of Informal Settlement in Agatare Cell / Nyarugenge Sector*. Kigali.

Goicoechea, A. (2008) *Peparing Surveys for Urban Upgrading Interventions - Prototype Survey Instrument and User Guide*, *Urban Paper Series - UP 6*. Washington D.C.: The World Bank.

Gönen, M. and Alpaydın, E. (2011) 'Multiple kernel learning algorithms', *Journal of Machine Learning Research*, 12(Jul), pp. 2211–2268.

Gould, S., Fulton, R. and Koller, D. (2009) 'Decomposing a scene into geometric and semantically consistent regions', in *Computer Vision, 2009 IEEE 12th International Conference on*, pp. 1–8. doi: 10.1109/ICCV.2009.5459211.

Gretton, A. *et al.* (2005) 'Algorithmic Learning Theory: 16th International Conference, ALT 2005, Singapore, October 8-11, 2005. Proceedings', in Jain, S., Simon, H. U., and Tomita, E. (eds). Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 63–77. doi: 10.1007/11564089_7.

Gretton, A. *et al.* (2006) 'A kernel method for the two-sample-problem', *Advances in Neural Information Processing Systems*, 19(157), pp. 0–43.

Gu, Y. *et al.* (2015) 'A novel MKL model of integrating LIDAR data and MSI for urban area classification', *IEEE Transactions on Geoscience and Remote Sensing*. IEEE, 53(10), pp. 5312–5326.

Guo, L. *et al.* (2011) 'Relevance of airborne lidar and multispectral image data for urban scene classification using Random Forests', *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(1), pp. 56–66. doi: 10.1016/j.isprsjprs.2010.08.007.

Guyon, I., Matic, N. and Vapnik, V. (1996) 'Discovering Informative Patterns and Data Cleaning.'

Haarbrink, R. B. and Eisenbeiss, H. (2008) 'Accurate DSM production from unmanned helicopter systems', *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 37, pp. 1259–1264.

Haarsma, D. (2017) 'Geo-ethics Requires Prudence with Private Data', *GIM International*, p. 1.

Hall, M. A. (1999) *Correlation-based feature selection for machine learning*. The University of Waikato.

Hanchuan Peng, Fuhui Long and Ding, C. (2005) 'Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8), pp. 1226–1238. doi: 10.1109/TPAMI.2005.159.

Hane, C. *et al.* (2013) 'Joint 3D scene reconstruction and class segmentation', in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 97–104. doi: 10.1109/CVPR.2013.20.

Hartfield, K. A., Landau, K. I. and Leeuwen, W. J. D. van (2011) 'Fusion of High Resolution Aerial Multispectral and LiDAR Data: Land Cover in the Context of Urban Mosquito Habitat', *Remote Sensing*, 3(12), pp. 2364–2383. doi: 10.3390/rs3112364.

Hartley, R. and Zisserman, A. (2003) *Multiple view geometry in computer vision*. Cambridge university press.

Harwin, S. and Lucieer, A. (2012) 'Assessing the Accuracy of Georeferenced Point Clouds Produced via Multi-View Stereopsis from Unmanned Aerial Vehicle (UAV) Imagery', *Remote Sensing*, 4(12), pp. 1573–1599. doi: 10.3390/rs4061573.

He, K. *et al.* (2015) 'Delving deep into rectifiers: Surpassing human-level performance on imagenet classification', in *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034.

He, K. *et al.* (2016) 'Deep residual learning for image recognition', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778.

Hingee, K. *et al.* (2016) 'Digital Terrain from a Two-Step Segmentation and Outlier-Based Algorithm', *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 233–239.

Hirschmüller, H. (2008) 'Stereo processing by semiglobal matching and mutual information', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2), pp. 328–341.

Hofmann, P. *et al.* (2008) 'Detecting informal settlements from QuickBird data in Rio de Janeiro using an object based approach', in Blaschke, T., Lang, S., and Hay, G. (eds) *Object-Based Image Analysis*. Springer Berlin Heidelberg, pp. 531–553. doi: 10.1007/978-3-540-77058-9_29.

Höhle, J. and Höhle, M. (2009) 'Accuracy assessment of digital elevation models by means of robust statistical methods', *ISPRS Journal of Photogrammetry and Remote Sensing*, 64(4), pp. 398–406. doi: 10.1016/j.isprsjprs.2009.02.003.

Hohn, M. E. (1991) *An Introduction to Applied Geostatistics*, *Computers & Geosciences*. New York: Oxford University Press. doi: 10.1016/0098-3004(91)90055-I.

Hsieh, C.-J., Si, S. and Dhillon, I. (2014) 'A divide-and-conquer solver for

kernel support vector machines', in *International Conference on Machine Learning*, pp. 566–574.

Hu, F. *et al.* (2015) 'Transferring Deep Convolutional Neural Networks for the Scene Classification of High-Resolution Remote Sensing Imagery', *Remote Sensing* . doi: 10.3390/rs71114680.

Hu, X. and Yuan, Y. (2016) 'Deep-Learning-Based Classification for DTM Extraction from ALS Point Cloud', *Remote Sensing*, 8(9), p. 730. doi: 10.3390/rs8090730.

Huang, M.-J. *et al.* (2008) 'A Knowledge-based Approach to Urban Feature Classification Using Aerial Imagery with Lidar Data', *Photogrammetric Engineering & Remote Sensing*, 74(12), pp. 1473–1485. doi: 10.14358/PERS.74.12.1473.

Hugenholtz, C. H. *et al.* (2013) 'Geomorphological mapping with a small unmanned aircraft system (sUAS): Feature detection and accuracy assessment of a photogrammetrically-derived digital terrain model', *Geomorphology*. Elsevier, 194, pp. 16–24.

Ioffe, S. and Szegedy, C. (2015) 'Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift', *ICML*.

Jankowska, M. M., Weeks, J. R. and Engstrom, R. (2012) 'Do the Most Vulnerable People Live in the Worst Slums? A Spatial Analysis of Accra, Ghana', *Annals of GIS*, 17(4), pp. 221–235.

Jeatrakul, P., Wong, K. W. and Fung, C. C. (2010) 'Data cleaning for classification using misclassification analysis', *Journal of Advanced Computational Intelligence and Intelligent Informatics*. Fuji Technology Press Co. Ltd., 14(3), pp. 297–302.

Kakembo, V. and van Niekerk, S. (2014) 'The integration of GIS into demographic surveying of informal settlements: The case of Nelson Mandela Bay Municipality, South Africa', *Habitat International*, 44, pp. 451–460. doi: 10.1016/j.habitatint.2014.09.004.

Kelm, K., Tonchovska, R. and Volkmann, W. (2014) 'Drones for Peace: Part II Fast and Inexpensive Spatial Data Capture for Multi-Purpose Use', in *2014 World Bank Conference on Land and Poverty*. Washington D.C., p. 26.

Kemker, R. and Kanan, C. (2017) 'Deep Neural Networks for Semantic Segmentation of Multispectral Remote Sensing Imagery', *arXiv preprint arXiv:1703.06452*.

Kilian, J., Haala, N. and Englich, M. (1996) 'Capture and evaluation of airborne laser scanner data', *International Archives of Photogrammetry and Remote Sensing*, 31(31), pp. 383–388.

Kit, O. and Lüdeke, M. (2013) 'Automated detection of slum area change in Hyderabad, India using multitemporal satellite imagery', *ISPRS Journal of Photogrammetry and Remote Sensing*. International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS), 83, pp. 130–137. doi: 10.1016/j.isprsjprs.2013.06.009.

Koeva, M. *et al.* (2016) 'Using UAVs for map creation and updating. A case

study in Rwanda', *Survey Review*. Taylor & Francis, pp. 1–14.

Kohli, D. *et al.* (2012) 'An ontology of slums for image-based classification', *Computers, Environment and Urban Systems*. Elsevier Ltd, 36(2), pp. 154–163. doi: 10.1016/j.compenvurbsys.2011.11.001.

Kohli, D. *et al.* (2013) 'Transferability of Object-Oriented Image Analysis Methods for Slum Identification', *Remote Sensing*, 5(9), pp. 4209–4228. doi: 10.3390/rs5094209.

Kraus, K. and Pfeifer, N. (1998) 'Determination of terrain models in wooded areas with airborne laser scanner data', *ISPRS Journal of Photogrammetry and Remote Sensing*, 53(4), pp. 193–203. doi: 10.1016/S0924-2716(98)00009-4.

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012) 'ImageNet Classification with Deep Convolutional Neural Networks', in Pereira, F. et al. (eds) *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., pp. 1097–1105.

Kuffer, M., Barros, J. and Sliuzas, R. V (2014) 'The development of a morphological unplanned settlement index using very-high-resolution (VHR) imagery', *Computers, Environment and Urban Systems*. Elsevier Ltd, 48, pp. 138–152. doi: 10.1016/j.compenvurbsys.2014.07.012.

Kuffer, M., Pfeffer, K. and Sliuzas, R. (2016) 'Slums from Space—15 Years of Slum Mapping Using Remote Sensing', *Remote Sensing*, 8(6), p. 455. doi: 10.3390/rs8060455.

Kuteesa, H. (2016) 'Maintain Kigali's cleanliness and security trademark, Mayor says', *The New Times*, p. 1.

Ladický, L. *et al.* (2012) 'Joint optimization for object class segmentation and dense stereo reconstruction', *International Journal of Computer Vision*, 100(2), pp. 122–133. doi: 10.1007/s11263-011-0489-0.

Lawson, C. L. (1972) 'Transforming triangulations', *Discrete Mathematics*. North-Holland, 3(4), pp. 365–372. doi: 10.1016/0012-365X(72)90093-3.

Li, Y. *et al.* (2014) 'Morphological operation based dense houses extraction from DSM', *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XL-3, pp. 183–189. doi: 10.5194/isprsarchives-XL-3-183-2014.

Liu, X. (2008) 'Airborne LiDAR for DEM generation: some critical issues', *Progress in Physical Geography*, 32(1), pp. 31–49. doi: 10.1177/0309133308089496.

Longbotham, N. *et al.* (2012) 'Very High Resolution Multiangle Urban Classification Analysis', *IEEE Transactions on Geoscience and Remote Sensing*, 50(4), pp. 1155–1170. doi: 10.1109/TGRS.2011.2165548.

Maas, A., Rottensteiner, F. and Heipke, C. (2016) 'Using label noise robust logistic regression for automated updating of topographic geospatial databases', in *23rd ISPRS Congress, July 12-19 2016, Prague, Czech Republic*. Göttingen: Copernicus GmbH.

Makoye, K. (2017) 'Drones help communities map flood risk in Dar es Salaam

slums', *Reuters World News*.

Mason, S. O. and Fraser, C. S. (1998) 'Image sources for informal settlement management', *The Photogrammetric Record*, 16(92), pp. 313–330.

Matic, N. *et al.* (1992) 'Computer aided cleaning of large databases for character recognition', in *Pattern Recognition, 1992. Vol. II. Conference B: Pattern Recognition Methodology and Systems, Proceedings., 11th IAPR International Conference on*. IEEE, pp. 330–333.

Meesuk, V. *et al.* (2015) 'Urban flood modelling combining top-view LiDAR data with ground-view SfM observations', *Advances in Water Resources*, 75, pp. 105–117. doi: http://dx.doi.org/10.1016/j.advwatres.2014.11.008.

Miltgen, C. L. and Peyrat-Guillard, D. (2014) 'Cultural and generational influences on privacy concerns: a qualitative study in seven European countries', *European Journal of Information Systems*, 23(2), pp. 103–125. doi: 10.1057/ejis.2013.17.

MININFRA (2017) *National Informal Urban Settlement Upgrading Strategy*.

Minja, D., Iliffe, M. and Anderson, E. (2016) 'Ramani Huria and Community Mapping - Towards Free and Open Map Data and Imagery for Dar es Salaam', in *2016 Tech4Dev*. Lausanne, p. 12.

Mnih, V. and Hinton, G. E. (2012) 'Learning to label aerial images from noisy data', in *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pp. 567–574.

Mongus, D., Lukač, N. and Žalik, B. (2014) 'Ground and building extraction from LiDAR data based on differential morphological profiles and locally fitted surfaces', *ISPRS Journal of Photogrammetry and Remote Sensing*, 93, pp. 145–156. doi: 10.1016/j.isprsjprs.2013.12.002.

Montgomery, M. R. (2008) 'The urban transformation of the developing world', *Science*, 319(5864), pp. 761–764. doi: 10.1126/science.1153012.

Moranduzzo, T. *et al.* (2015) 'Multiclass Coarse Analysis for UAV Imagery', *IEEE Transactions on Geoscience and Remote Sensing*. Institute of Electrical and Electronics Engineers Inc., 53(12), pp. 6394–6406.

Myint, S. W. *et al.* (2011) 'Per-pixel vs. object-based classification of urban land cover extraction using high spatial resolution imagery', *Remote Sensing of Environment*, 115(5), pp. 1145–1161. doi: 10.1016/j.rse.2010.12.017.

Nair, V. and Hinton, G. E. (2010) 'Rectified linear units improve restricted boltzmann machines', in *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814.

Nebiker, S. *et al.* (2008) 'A light-weight multispectral sensor for micro UAV—Opportunities for very high resolution airborne remote sensing', *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 37, pp. 1193–1200.

Nex, F. and Gerke, M. (2014) 'Photogrammetric DSM denoising', *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 40(3), p. 231.

Nex, F. and Remondino, F. (2014) 'UAV for 3D mapping applications: a review', *Applied Geomatics*, 6(1), pp. 1–15.

Niazmardi, S. *et al.* (2016) 'A Comparative Study on Multiple Kernel Learning for Remote Sensing Image Classification', in *IGARSS*. Beijing.

Nichol, J. (2009) 'Remote Sensing of Urban Areas', in *The SAGE Handbook of Remote Sensing*. 1 Oliver's Yard, 55 City Road London EC1Y 1SP: SAGE Publications, Inc., pp. 423–436. doi: 10.4135/9780857021052.n30.

NLeSC (2015) *Processing large datasets on consumer-grade computers*. Netherlands eScience Center. Available at: https://www.esciencecenter.nl/project/improving-open-source-photogrammetric-workflows-for-processing-big-datasets.

ten Oever, S. *et al.* (2016) 'The COGs (context, object, and goals) in multisensory processing', *Experimental brain research*. Springer, 234(5), pp. 1307–1323.

Ojala, T., Pietikainen, M. and Maenpaa, T. (2002) 'Multiresolution gray-scale and rotation invariant texture classification with local binary patterns', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7), pp. 971–987. doi: 10.1109/TPAMI.2002.1017623.

Ordnance Survey (2015) *Future trends in geospatial information management: the five to ten year vision*. 2nd edn.

Owen, K. K. and Wong, D. W. (2013) 'An approach to differentiate informal settlements using spectral, texture, geomorphology and road accessibility metrics', *Applied Geography*, 38, pp. 107–118.

Paar, P. and Rekittke, J. (2011) 'Low-Cost Mapping and Publishing Methods for Landscape Architectural Analysis and Design in Slum-Upgrading Projects', *Future Internet*, 3(4), pp. 228–247. doi: 10.3390/fi3040228.

Pannell, D. J. *et al.* (2011) 'Understanding and promoting adoption of conservation practices by rural landholders', *Changing Land Management: Adoption of New Practices by Rural Landholders*, p. 11.

Pauner, C., Kamara, I. and Viguri, J. (2015) 'Drones. Current challenges and standardisation solutions in the field of privacy and data protection', in *ITU Kaleidoscope: Trust in the Information Society (K-2015), 2015*. IEEE, pp. 1–7.

Pavlidis, P. *et al.* (2001) 'Gene functional classification from heterogeneous data', in *Proc. of the Fifth Annual International Conferences on Compututational Molecular Biology (RECOMB01)*.

Pérez-Garcia, J. L. *et al.* (2012) 'Progressive densification and region growing methods for LIDAR data classification', *International Archives of Photogrammetry, Remote Sensing, and Spatial Information Sciences*, 39(B3), pp. 155–160.

Persello, C. (2013) 'Interactive domain adaptation for the classification of remote sensing images using active learning', *IEEE Geoscience and Remote Sensing Letters*. IEEE, 10(4), pp. 736–740.

Persello, C. and Bruzzone, L. (2016) 'Kernel-Based Domain-Invariant Feature

Selection in Hyperspectral Images for Transfer Learning', *IEEE Transactions on Geoscience and Remote Sensing*, 54(5), pp. 2615–2626. doi: 10.1109/TGRS.2015.2503885.

Persello, C. and Stein, A. (2017) 'Deep Fully Convolutional Networks for the Detection of Informal Settlements in VHR Images', *IEEE Geoscience and Remote Sensing Letters*, 14(12), pp. 2325–2329. doi: 10.1109/LGRS.2017.2763738.

Pfeffer, K. *et al.* (2013) 'Participatory Spatial Knowledge Management Tools: Empowerment and upscaling or exclusion?', *Information, Communication & Society*, 16(2), pp. 258–285.

Pfeffer, K. and Verrest, H. (2016) 'Perspectives on the Role of Geo-Technologies for Addressing Contemporary Urban Issues: Implications for IDS', *European Journal of Development Research*. Springer, 28(2), pp. 154–166. doi: 10.1057/ejdr.2016.4.

Priestnall, G., Jaafar, J. and Duncan, A. (2000) 'Extracting urban features from LiDAR digital surface models', *Computers, Environment and Urban Systems*, 24(2), pp. 65–78. doi: 10.1016/S0198-9715(99)00047-2.

Pudil, P., Novovičová, J. and Kittler, J. (1994) 'Floating search methods in feature selection', *Pattern Recognition Letters*, 15(11), pp. 1119–1125. doi: 10.1016/0167-8655(94)90127-9.

Pugalis, L., Giddings, B. and Anyigor, K. (2014) 'Reappraising the World Bank responses to rapid urbanisation: Slum improvements in Nigeria', *Local Economy*, 29(4–5), pp. 519–540. doi: 10.1177/0269094214541377.

Puissant, A., Hirsch, J. and Weber, C. (2005) 'The utility of texture analysis to improve per-pixel classification for high to very high spatial resolution imagery', *International Journal of Remote Sensing*, 26(4), pp. 733–745. doi: 10.1080/01431160512331316838.

Qiu, S. and Lane, T. (2009) 'A framework for multiple kernel support vector regression and its applications to siRNA efficacy prediction', *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6(2), pp. 190–199.

Rakotomamonjy, A. *et al.* (2008) 'SimpleMKL', *Journal of Machine Learning Research*, 9, pp. 2491–2521.

Ramani Huria (2016) *The Atlas of Flood Resilience in Dar es Salaam*. Dar es Salaam.

Rambaldi, G., Kyem, P. A. K., *et al.* (2006) 'Participatory Spatial Information Management and Communication in Developing Countries', *The Electronic Journal of Information Systems in Developing Countries*, 25(1), pp. 1–9. doi: 10.1002/j.1681-4835.2006.tb00162.x.

Rambaldi, G., Chambers, R., *et al.* (2006) 'Practical ethics for PGIS practitioners, facilitators, technology intermediaries and researchers', *Participatory Learning and Action*, 54(1), pp. 106–113.

Ramona, M., Richard, G. and David, B. (2012) 'Multiclass feature selection with kernel gram-matrix-based criteria', *Neural Networks and Learning*

Systems, IEEE Transactions on. IEEE, 23(10), pp. 1611–1623.

Reed, P. J., Spiro, E. S. and Butts, C. T. (2016) 'Thumbs up for privacy?: Differences in online self-disclosure behavior across national cultures', *Social Science Research*. Academic Press, 59, pp. 155–170. doi: 10.1016/J.SSRESEARCH.2016.04.022.

Remondino, F. *et al.* (2014) 'State of the art in high density image matching', *The Photogrammetric Record*, 29(146), pp. 144–166. doi: 10.1111/phor.12063.

Remondino, F. and Campana, S. (2014) '3D recording and modelling in archaeology and cultural heritage : theory and best practices', *BAR international series 2598*.

Riegler, G. *et al.* (2017) 'OctNetFusion: Learning Depth Fusion from Data'.

Romero, A., Gatta, C. and Camps-Valls, G. (2016) 'Unsupervised Deep Feature Extraction for Remote Sensing Image Classification', *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1349–1362. doi: 10.1109/TGRS.2015.2478379.

Rottensteiner, F. *et al.* (2014) 'Results of the ISPRS benchmark on urban object detection and 3D building reconstruction', *ISPRS Journal of Photogrammetry and Remote Sensing*, 93, pp. 256–271. doi: 10.1016/j.isprsjprs.2013.10.004.

Rwanda, G. of (2000) 'Rwanda Vision 2020'. Kigali: Ministry of Finance and Economic Planning.

Rwanda, O. G. of the R. of (2005) *Organic Law Determining the Modalities of Protection, Conservation, and Promotion of the Environment in Rwanda*. Rwanda.

Sandbrook, C. (2015) 'The social implications of using drones for biodiversity conservation', *Ambio*, 44(S4), pp. 636–647. doi: 10.1007/s13280-015-0714-0.

Saxena, S. (no date) 'National Open Data frames across Japan, The Netherlands and Saudi Arabia: role of culture', *Foresight*, 0(ja), p. 0. doi: 10.1108/FS-07-2017-0038.

Schindler, K. (2012) 'An overview and comparison of smooth labeling methods for land-cover classification', *IEEE Transactions on Geoscience and Remote Sensing*. IEEE, 50(11), pp. 4534–4545.

Schölkopf, B., Smola, A. and Müller, K.-R. (1998) 'Nonlinear Component Analysis as a Kernel Eigenvalue Problem', *Neural Computation*, 10(5), pp. 1299–1319.

SDI (2016) *Nairobi Inventory*. Nairobi, Kenya.

Senthilnath, J. *et al.* (2017) 'Application of UAV imaging platform for vegetation analysis based on spectral-spatial methods', *Computers and Electronics in Agriculture*, 140, pp. 8–24. doi: http://dx.doi.org/10.1016/j.compag.2017.05.027.

Serna, A. and Marcotegui, B. (2014) 'Detection, segmentation and classification of 3D urban objects using mathematical morphology and

supervised learning', *ISPRS Journal of Photogrammetry and Remote Sensing*, 93, pp. 243–255. doi: 10.1016/j.isprsjprs.2014.03.015.

Shelhamer, E., Long, J. and Darrell, T. (2017) 'Fully convolutional networks for semantic segmentation', *IEEE transactions on pattern analysis and machine intelligence*. IEEE, 39(4), pp. 640–651.

Sherrah, J. (2016) 'Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery', *arXiv preprint arXiv:1606.02585*.

Shotton, J. *et al.* (2009) 'Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context', *International Journal of Computer Vision*. Springer, 81(1), pp. 2–23.

Simonyan, K. and Zisserman, A. (2014) 'Very deep convolutional networks for large-scale image recognition', *arXiv preprint arXiv:1409.1556*.

Sithole, G. and Vosselman, G. (2004) 'Experimental comparison of filter algorithms for bare-Earth extraction from airborne laser scanning point clouds', *ISPRS Journal of Photogrammetry and Remote Sensing*, 59(1–2), pp. 85–101. doi: 10.1016/j.isprsjprs.2004.05.004.

Sithole, G. and Vosselman, G. (2005) 'Filtering of airborne laser scanner data based on segmented point clouds', *International Archives of Photogrammetry, Remote Sensing, and Spatial Information Sciences*, 36(part 3), p. W19.

Sliuzas, R. (2003) 'Governance and the use of GIS in developing countries', *Habitat International*, 27(4), pp. 495–499. doi: 10.1016/S0197-3975(03)00002-X.

Sliuzas, R., Mboup, G. and de Sherbinin, A. (2008) *Report of the Expert Group Meeting on Slum Identification and Mapping*. CIESIN, UN-Habitat, ITC.

Sona, G. *et al.* (2014) 'Experimental analysis of different software packages for orientation and digital surface modelling from UAV images', *Earth Science Informatics*. Springer Berlin Heidelberg, 7(2), pp. 97–107. doi: 10.1007/s12145-013-0142-2.

Song, X., Herranz, L. and Jiang, S. (2017) 'Depth CNNs for RGB-D Scene Recognition: Learning From Scratch Better Than Transferring From RGB-CNNs', *Aaai*, pp. 4271–4277.

Srivastava, N. *et al.* (2014) 'Dropout: A simple way to prevent neural networks from overfitting', *The Journal of Machine Learning Research*. JMLR. org, 15(1), pp. 1929–1958.

Stöcker, C. *et al.* (2017) 'Review of the current state of UAV regulations', *Remote Sensing*, 9(5), pp. 1–26.

Strobl, E. V and Visweswaran, S. (2014) 'Markov Blanket Ranking using Kernel-based Conditional Dependence Measures', in *NIPS 2013 Workshop on Causality*.

Tarolli, P. (2014) 'High-resolution topography for understanding Earth surface processes: Opportunities and challenges', *Geomorphology*, 216, pp. 295–312. doi: 10.1016/j.geomorph.2014.03.008.

Taubenböck, H. and Kraff, N. J. (2014) 'The physical face of slums: a structural comparison of slums in Mumbai, India, based on remotely sensed data', *Journal of Housing and the Built Environment*, 29(1), pp. 15–38. doi: 10.1007/s10901-013-9333-x.

Thibault, G. and Aoude, G. (2016) 'Companies Are Turning Drones into a Competitive Advantage', *Harvard Buisness Review*, pp. 1–6.

Thongkam, J. *et al.* (2008) 'Support vector machine for outlier detection in breast cancer survivability prediction', in *Asia-Pacific Web Conference*. Springer, pp. 99–109.

Tokarczyk, P. *et al.* (2015) 'High-quality observation of surface imperviousness for urban runoff modelling using UAV imagery', *Hydrology and Earth System Sciences*. doi: 10.5194/hess-19-4215-2015.

Tomljenovic, I. *et al.* (2015) 'Building Extraction from Airborne Laser Scanning Data: An Analysis of the State of the Art', *Remote Sensing*, 7(4), pp. 3826–3862. doi: 10.3390/rs70403826.

Tomljenovic, I., Tiede, D. and Blaschke, T. (2016) 'A building extraction approach for Airborne Laser Scanner data utilizing the Object Based Image Analysis paradigm', *International Journal of Applied Earth Observation and Geoinformation2*, 52, pp. 137–148.

Tong, X., Xie, H. and Weng, Q. (2014) 'Urban Land Cover Classification With Airborne Hyperspectral Data: What Features to Use?', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(10), pp. 3998–4009. doi: 10.1109/JSTARS.2013.2272212.

Torres-Sánchez, J. *et al.* (2014) 'Multi-temporal mapping of the vegetation fraction in early-season wheat fields using images from UAV', *Computers and Electronics in Agriculture*, 103, pp. 104–113. doi: 10.1016/j.compag.2014.02.009.

Tóvári, D. and Pfeifer, N. (2005) 'Segmentation based robust interpolation-a new approach to laser data filtering', *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 36(3/W19), pp. 79–84.

Tsai, V. J. D. (1993) 'Delaunay triangulations in TIN creation: an overview and a linear-time algorithm', *International Journal of Geographical Information Systems*. Taylor & Francis, 7(6), pp. 501–524. doi: 10.1080/02693799308901979.

Tuia, D. *et al.* (2010) 'Learning relevant image features with multiple-kernel classification', *IEEE Transactions on Geoscience and Remote Sensing*. IEEE, 48(10), pp. 3780–3791.

Turley, R. *et al.* (2013) 'Slum upgrading strategies involving physical environment and infrastructure interventions and their effects on health and socioeconomic outcomes (Review)', *Cochrane Database of Systematic Reviews*. John Wiley & Sons, Ltd, (1). doi: 10.1002/14651858.CD010067.pub2.

UAViators (2017) *Humanitarian UAV Code of Conduct & Guidelines*.

Ulusoy, A. O., Black, M. J. and Geiger, A. (2017) 'Semantic Multi-view Stereo: Jointly Estimating Objects and Voxels', in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4531–4540. doi: 10.1109/CVPR.2017.482.

UN-Habitat (2012) *Streets as tools for urban transformation in slums: A Street-led Approach to Citywide Slum Upgrading*. Nairobi.

UN-Habitat (2013) 'Kigali Declaration', *2nd Tripartiate conference ACP / EC / UN-Habitat - Sustainable Urbanization for Poverty Eradication*. Kigali, Rwanda.

UN-Habitat (2015) *Habitat III Issue Paper 22 - Informal Settlements*. New York.

UN-Habitat (2016) *Slum Almanac 2015/2016*. Nairobi.

UN-Habitat and Earthscan (2003) *The challenge of slums : global report on human settlements*.

UN-Habitat III (2017) *New Urban Agenda*, *Conference on Housing and Sustainable Urban Development (Habitat III)*. doi: ISBN: 978-92-1-132757-1.

UNHabitat (2013) 'Urban Data', *Urban Data*, pp. 273–281.

United Nations (2015) *Transforming Our World: the 2030 Agenda for Sustainable Development*, *Resolution adopted by the General Assembly on 25 September 2015. A/RES/70/1*.

United Nations (2016) *The Sustainable Development Goals Report*, *United Nations*. doi: 10.18356/3405d09f-en.

Varma, M. and Babu, B. R. (2009) 'More generality in efficient multiple kernel learning', *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, 2009(June), pp. 1–8.

Verplanke, J. *et al.* (2016) 'A Shared Perspective for PGIS and VGI', *Cartographic Journal*. Taylor & Francis, 53(4), pp. 308–317. doi: 10.1080/00087041.2016.1227552.

Vetrivel, A. *et al.* (2015) 'Identification of damage in buildings based on gaps in 3D point clouds from very high resolution oblique airborne images', *ISPRS Journal of Photogrammetry and Remote Sensing*. Elsevier, 105, pp. 61–78.

Vetrivel, A. *et al.* (2015) 'Segmentation of UAV-based images incorporating 3D point cloud information', in *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*. International Society for Photogrammetry and Remote Sensing, pp. 261–268.

Volpi, M. and Tuia, D. (2017) 'Dense semantic labeling of subdecimeter resolution images with convolutional neural networks', *IEEE Transactions on Geoscience and Remote Sensing*. IEEE, 55(2), pp. 881–893.

Vosselman, G. (2012) 'Automated planimetric quality control in high accuracy airborne laser scanning surveys', *ISPRS Journal of Photogrammetry and Remote Sensing*, 74, pp. 90–100. doi: 10.1016/j.isprsjprs.2012.09.002.

Vosselman, G. (2013) 'Point cloud segmentation for urban scene classification', *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XL-7/W2(2), pp. 257–262. doi: 10.5194/isprsarchives-XL-7-W2-257-2013.

Wallace, L. *et al.* (2012) 'Development of a UAV-LiDAR system with application to forest inventory', *Remote Sensing*, 4(6), pp. 1519–1543.

Weidner, U. and Förstner, W. (1995) 'Towards automatic building extraction from high-resolution digital elevation models', *ISPRS Journal of Photogrammetry and Remote Sensing*, 50(4), pp. 38–49. doi: 10.1016/0924-2716(95)98236-S.

Weinmann, M. *et al.* (2015) 'Semantic point cloud interpretation based on optimal neighborhoods, relevant features and efficient classifiers', *ISPRS Journal of Photogrammetry and Remote Sensing*, 105, pp. 286–304. doi: 10.1016/j.isprsjprs.2015.01.016.

Wekesa, B. W., Steyn, G. S. and Otieno, F. A. O. (2011) 'A review of physical and socio-economic characteristics and intervention approaches of informal settlements', *Habitat International*, 35(2), pp. 238–245. doi: http://dx.doi.org/10.1016/j.habitatint.2010.09.006.

Weng, Q. (2012) 'Remote sensing of impervious surfaces in the urban areas: Requirements, methods, and trends', *Remote Sensing of Environment*, 117, pp. 34–49. doi: 10.1016/j.rse.2011.02.030.

Woebbecke, D. M. *et al.* (1995) 'Color Indices for Weed Identification Under Various Soil, Residue, and Lighting Conditions', *Transactions of the ASAE*, 38(1), pp. 259–269. doi: 10.13031/2013.27838.

Xiao, J. (2013) *Automatic building detection using oblique imagery*. University of Twente Faculty of Geo-Information and Earth Observation (ITC).

Xu, C., Tao, D. and Xu, C. (2013) 'A Survey on Multi-view Learning', *Cvpr*, 36(8), p. 300072.

Xu, S., Vosselman, G. and Oude Elberink, S. (2014) 'Multiple-entity based classification of airborne laser scanning data in urban areas', *ISPRS Journal of Photogrammetry and Remote Sensing*, 88, pp. 1–15. doi: 10.1016/j.isprsjprs.2013.11.008.

Xu, Z. *et al.* (2010) 'Simple and efficient multiple kernel learning by group lasso', *International Conference on Machine Learning (ICML)*, pp. 1191–1198.

Yan, M. *et al.* (2012) 'An object-based analysis filtering algorithm for airborne laser scanning', *International Journal of Remote Sensing*, 33(22), pp. 7099–7116. doi: 10.1080/01431161.2012.699694.

Yan, W. Y., Shaker, A. and El-Ashmawy, N. (2015) 'Urban land cover classification using airborne LiDAR data: A review', *Remote Sensing of Environment*, 158, pp. 295–310. doi: 10.1016/j.rse.2014.11.001.

Yeh, Y. *et al.* (2012) 'A Novel Multiple Kernel Learning Framework for Heterogeneous Feature Fusion and Variable Selection', *IEEE Transactions on Multimedia*, 14(3), pp. 563–574. doi: 10.1109/TMM.2012.2188783.

Yu, F. and Koltun, V. (2015) 'Multi-scale context aggregation by dilated convolutions', *arXiv preprint arXiv:1511.07122*.

Zhang, C. and Kovacs, J. M. (2012) 'The application of small unmanned aerial systems for precision agriculture: A review', *Precision Agriculture*, pp. 693–712.

Zhang, L., Zhang, L. and Du, B. (2016) 'Deep Learning for Remote Sensing Data: A Technical Tutorial on the State of the Art', *IEEE Geoscience and Remote Sensing Magazine*, pp. 22–40. doi: 10.1109/MGRS.2016.2540798.

Zhang, Q. *et al.* (2015) 'Classification of ultra-high resolution orthophotos combined with DSM using a dual morphological top hat profile', *Remote Sensing*. Multidisciplinary Digital Publishing Institute, 7(12), pp. 16422–16440.

Zhang, W. *et al.* (2016) 'An Easy-to-Use Airborne LiDAR Data Filtering Method Based on Cloth Simulation', *Remote Sensing*, 8(6), p. 501. doi: 10.3390/rs8060501.

Summary

Informal settlements, or slums, are considered to be one of the major development challenges of our time. One of the major obstacles for slum upgrading projects is the lack of data regarding current slum conditions such as the existing housing situation, accessibility through road networks, and hazardous environments. Informal settlements are often literally and symbolically "empty spots on the map". Unmanned Aerial Vehicles (UAVs) are capable of providing imagery at a higher resolution but lower cost than imagery from satellites or manned aircraft. The present research explores the use of this exciting new technology to help fill these gaps on the map. The work entails an exploration of how to tailor machine learning methods to the characteristics of UAV data and ensure their performance despite the challenging characteristics of informal settlements. By working in the field together with actual informal settlement upgrading projects, it is also possible to investigate how well UAVs match the practical needs of upgrading projects and understand its societal impact.

The first part focusses on the use of machine learning methods. For example, supervised classification methods can be used to recognize patterns in data from some labeled training samples, enabling a class label to be assigned to new data. The first step in supervised classification is usually to define relevant features to describe the samples. The first objective aimed to **identify synergies between 2D and 3D data provided by UAVs**. Experiments using UAV data of unplanned settlements in Kigali, Rwanda and Maldonado, Uruguay indicated that buildings, roads, vegetation, structures and clutter could be discriminated with accuracies over 90% when combining 2D features from the imagery with 3D features from the point cloud.

In recognition of the statistical differences between 2D and 3D features, the next step aimed to **adapt supervised classification methods to deal with heterogeneous data**. Support Vector Machines (SVMs) are a successful machine learning method but generally use a single kernel to describe the non-linear similarity between training samples. Multiple Kernel Learning, however, uses different kernels for different feature groups which allows it to identify more subtle similarities and differences between samples. This manuscript presents an algorithm which can automatically group the features and provide tailored kernel parameters. Experiments show that the proposed MKL method achieves higher accuracies than conventional single-kernel methods applied to the 2D and 3D features while requiring less user-interaction than previous MKL methods.

As supervised classification methods are improving, obtaining training samples is proving to be a bottleneck as it generally requires much manual work.

Therefore, research was done to **analyze how reliable training labels can be obtained from existing geospatial data**. Translating existing maps into training labels for newly-acquired UAV data will introduce errors due to (1) changes in the scene itself such as building constructions or demolitions, or (2) misalignments due to digitization at a lower spatial resolution or other geo-referencing issues. Experiments demonstrate the effectiveness of using these 'noisy' labels to train a classifier, remove samples with unreliable labels based on local and global contextual cues, train another classifier, etc. in an iterative process. An accuracy of above 90% could be obtained even if 30% of the initial training samples were mislabeled. This method can easily be applied to classify recurrent UAV imagery in projects which require frequent data coverage and check/improve the quality of manual digitization campaigns.

In addition to maps, detailed elevation models are important sources of information to support urban upgrading projects. Overlapping aerial images can provide a Digital Surface Model (DSM) which describes the elevation of the tops of objects, whereas many planning activities require the Digital Terrain Model (DTM) which provides the elevation of the underlying terrain. Unfortunately, unplanned settlements are often characterized by densely built-up areas and are often located in less desired areas such as steep slopes which cause difficulties. Therefore, it was also **analyzed how to extract Digital Terrain Models in challenging settings**. A method specifically tailored to aerial photogrammetric datasets was developed using cutting-edge deep learning techniques. Firstly, a simple rule uses the DSM to propose pixels which are likely to be ground or off-ground without any manual intervention. These samples are used to train a Fully Convolutional Network (FCN) specifically designed for this task, enabling it to differentiate between terrain and off-ground objects using imagery and DSM-based features. The proposed method was shown to significantly outperform two reference DTM extraction techniques, thus enabling DTM extraction to be performed in challenging settings while eliminating the requirement of collecting costly training samples.

Apart from generating maps automatically and extracting DTMs, the UAV data can be useful for upgrading projects in many ways. One task was to observe the use of the UAV data by stakeholders in Kigali, Rwanda to **identify opportunities of UAVs to support urban upgrading workflows**. Important observations included that even without advanced machine learning techniques, the images were considered to be highly valuable and were used by the upgrading projects in various ways. The higher resolution and recency of the imagery facilitated manual digitization exercises. Additional information required for upgrading projects which are not typically identifiable in satellite or aerial imagery, such as solid waste accumulations, was visible. The imagery enabled consultants to prepare better for the field and navigate the complex network of footpaths more effectively during operations. The data was also

valuable as a communications platform, enabling communication between stakeholders in understanding the existing situation and prioritizing interventions.

Some concerns regarding the use of UAVs is the possible capture of objects considered as private in the imagery, the distribution of this sensitive information, and possible misuse of it. A final sub-objective was therefore to **analyze the social impacts of using UAVs in the context of urban upgrading projects**. Residents of unplanned areas in Kigali, Rwanda and Dar es Salaam, Tanzania which was subject to UAV flights were asked what their perceptions of the flights were. They were also asked to point out objects in the imagery and maps which they considered to be sensitive. These consisted of *avoidable* objects which could be removed by residents if they are aware UAV flights will take place, *unavoidable but removable* objects which are captured in the UAV imagery but can be blurred before distribution to other stakeholders, and *unavoidable and irremovable* objects. The later causes the most concern as the objects considered as sensitive are exactly those who are targeted by the UAV operations. For example, houses located in hazardous areas may be subject to expropriation. The research further illustrates the importance of local context regarding these concerns and which actions can be taken to ensure more ethical UAV operations and equitable distribution of the benefits.

In sum, this manuscript illustrates how UAVs and machine learning methods can be manipulated to provide accurate and up-to-date geospatial information. The simultaneous provision of 2D imagery and 3D point clouds proves to be quite useful and stresses the importance of developing targeted geoinformatic workflows which make use of these synergies rather than applying standard algorithms developed for either imagery or point clouds. This enables automatic algorithms to return highly accurate maps, despite the challenging characteristics of unplanned neighborhoods. Secondly, involvement with existing urban upgrading projects throughout the research has enabled a unique view of the actual usage and effectiveness of the imagery for current urban upgrading projects by local governments, engineering consultants and residents. At the data collection phase, residents were intrigued but often unable to think of practical uses for the UAV imagery. Returning years later, it appeared that the images were being used for a wide range of unexpected applications. As UAVs becoming increasingly available and as data processing simplifies, it is feasible to imagine a future where UAVs become increasingly used to support urban upgrading projects.

Samenvatting

Sloppenwijken worden beschouwd als één van de grootste uitdagingen van onze tijdperk. Eén van de grootste obstakels voor projecten met als doel deze wijken proberen te verbeteren is het tekort aan data omtrent de huidige situatie van een sloppenwijk. Voorbeelden hiervan zijn: de huidige bebouwing, toegankelijkheid met betrekking tot het wegennetwerk en gevaarlijke omgevingen. Sloppenwijken zijn vaak letterlijk en symbolisch "lege plekken op de kaart". Onbemande luchtvaartuigen, of drones, kunnen beelden nemen met een hogere resolutie tegen lagere kosten dan beelden van satellieten of bemande luchtvaartuigen. Het huidige onderzoek gaat na in hoeverre deze nieuwe technologie deze gaten in de kaart weet op te vullen. Het werk omvat hoe kunstmatige intelligentie algoritmes aangepast kunnen worden aan de kenmerken van data afkomstig van drones en hun nauwkeurigheid behouden, ondanks de moeilijke kenmerken van sloppenwijken. Door samenwerking met reële projecten is het ook mogelijk na te gaan tot hoeverre drones de praktische behoeftes van deze projecten kunnen voorzien, zowel de maatschappelijk impact van hun gebruik.

Het eerste gedeelte onderzoekt het gebruik van automatische algoritmes. Classificatiemodellen kunnen patronen in data leren herkennen door het gebruik van monsters waarvan de klasse bekend is, die dan gebruikt kan worden om klassen toe te wijzen aan onbekende monsters. De eerste stap van dergelijke modellen is vaak het definiëren van de onderscheidende attributen. Het eerste doel is daarom het **identificeren van synergiën tussen 2D en 3D attributen afkomstig van drones**. Experimenten met drone data van sloppenwijken in Kigali, Rwanda en Maldonado, Uruguay tonen aan dat gebouwen, wegen, vegetatie, structuren en rommel met een nauwkeurigheid van meer dan 90% kunnen worden herkend als zowel 2D attributen van de beelden en 3D attributen afkomstig van de puntenwolk.

Omdat er statistische verschillen zijn tussen de 2D en 3D data, is de volgende stap om te onderzoeken hoe **klassificatie algoritmes aangepast kunnen worden om met heterogeen data te werken.** Support Vector Machine (SVM) is een succesvolle klassificatie methode. Echter maakt SVM gewoonlijk gebruik van een enkele kernelfunctie om de gelijkenis tussen monsters te bepalen. Multiple Kernel Learning (MKL), gebruikt, integendeel, verschillende kernelfuncties voor verschillende groepen attributen en kan zo meer subtiele gelijkenissen en verschillen vastleggen. Dit onderzoek presenteert een algoritme die automatisch attributen in toepasselijke groepen onderverdeeld en de gepaste kernelfunctie parameters bepaald. Experimenten tonen aan deze methode beter presteert dan de gebruikelijke enkel kernelfunctie SVM en tegelijkertijd minder invoer van de gebruiker nodig heeft als bestaande MKL werkwijzen.

Het verkrijgen van goede monsters om de modellen af te stemmen is moeilijk omdat ze vaak veel manuele interventie nodig hebben. Daarom werd er ook onderzoek gedaan naar **hoe bestaande ruimtelijke data gebruikt kan worden om betrouwbare etiketten toe te wijzen aan monsters**. De bestaande data kan fouten bevatten door (1) veranderingen in het landschap zelf zoals nieuwe of gesloopte gebouwen, en (2) verplaatsingen door het digitaliseren over beelden met een lagere resolutie of andere problemen met de georeferencing. Experimenten bewijzen dat de effectiviteit van deze werkwijze die de imperfecte etiketten te gebruiken, onbetrouwbare monsters wegnemen door gebaseerd op lokale en globale samenhorigheid, de betrouwbare monsters gebruiken om de classificatiemodel te verfijnen, enz. in een iteratief proces. Een nauwkeurigheid boven 90% kan worden verkregen, ook al hadden 30% van de initiële monsters een verkeerd etiket. Deze werkwijze kan gebruikt worden om herhaalde UAV beelden te classificeren in projecten die frequente beeldopnames eisen en om de kwaliteit van manuele digitaliserings-campagnes te verifiëren en verbeteren.

Niet alleen kaarten, maar ook ruimtelijke hoogtemodellen zijn een belangrijke bron van informatie ter ondersteuning van stedelijke verbeteringsprojecten. Overlappende beelden kunnen een digitale oppervlaktemodel (DSM) opleveren, terwijl veel ruimtelijke ontwikkeling projecten juist een digitale terreinmodel (DTM) nodig hebben. Dichte bebouwing en minder gewenste omgevingen zoals steile hellingen zijn vaak kenmerkend van sloppenwijken en bemoeilijken het proces om DTMs uit DSMs te verkrijgen. Daarom werd er ook **geanalyseerd hoe DTMs verkregen kunnen worden in uitdagende omstandigheden**. Een methode die gebruik maakt van de laatste kunstmatige intelligentie technieken werd specifiek ontwikkeld om dit te doen voor luchtbeelden. De eerste stap gebruikt een eenvoudige regel om onderscheid te maken tussen pixels die waarschijnlijk terrein en bovengrondse objecten representeren. Deze monsters worden dan gebruikt om een unieke *Fully Convolutional Network (FCN)* te verfijnen, die dan gebruikt kan worden om de hele DSM te classificeren. Deze voorgestelde methode werkte duidelijk beter dan twee referentie DTM extractie methodes ondanks de uitdagende omstandigheden en elimineert tegelijkertijd de eis om dure monsters te verkrijgen.

Naast het genereren van kaarten en digitale terreinmodellen, kan UAV data in verschillende aspecten nuttig zijn. Een taak was daarom ook om te observeren hoe de belanghebbende partijen in Kigali, Rwanda de data gebruiken om de **mogelijkheden van UAVs om stedelijke verbeteringsprojecten te ondersteunen**. Het werd bijvoorbeeld opgemerkt dat zelfs zonder geavanceerde kunstmatige intelligentie technieken, de beelden als zeer nuttig werden gewaardeerd en op verschillende manieren werden gebruikt. De hogere resolutie en actualiteit van de beelden vergemakkelijkten manuele

digitalisatie processen. Aanvullende informatie die voorgaand niet beschikbaar waren in satellietbeelden of luchtfoto's, zoals ophopingen van afval, kunnen worden geobserveerd. De beelden lieten consultants zich ook beter voorbereiden op veldwerk en in het veld zelf de complexe netwerken van voetpaden te navigeren. De UAV beelden vergemakkelijkte ook de communicatie tussen verschillende belanghebbende partijen betreft het begrijpen van het huidige situatie en de prioriteren van interventies.

Er zijn een aantal ethische zorgen betreft UAV beelden, zoals: het mogelijk in kaart brengen van objecten die als privaat beschouwd worden, de distributie van deze gevoelige informatie, en mogelijke misbruik ervan. Een laatste taak was dus om **de sociale impact van het gebruik van UAVs in de context van sloppenwijk verbeteringsprojecten te analyseren**. Inwoners van sloppenwijken in Kigali, Rwanda en Dar es Salaam, Tanzania waar UAV projecten plaatsvonden werden gevraagd betreft hun percepties van deze activiteiten. Ze werden ook gevraagd om objecten aan te wijzen in deze beelden die zij als gevoelig beschouwen. Deze objecten konden: *vermijdbaar* zijn die door de inwooners zelf weg genomen konden worden als ze voorafgaand zouden weten van de UAV vluchten en implicaties ervan, *onvermijdbaar maar verwijdbare* objecten die niet weggenomen kunnen worden maar wel in de beelden zelf vervaagd kunnen worden voor distributie, of *onvermijdbaar en onverwijderbaar*. Deze laatste groep behoort tot het meest verontrustende vanwege het feit dat deze objecten als privé beschouwd konden worden en precies degene zijn die door de vluchten als doel beschouwd worden. Bijvoorbeeld, huizen die in gevaarlijke gebieden liggen zouden onteigend kunnen worden. Het onderzoek toont ook aan hoe belangrijk de lokale context is zowel als welke acties ondernomen kunnen worden om het ethisch gebruik van UAVs en een gelijke distributie van de baten te verzorgen.

Tot slot toont dit onderzoek aan hoe drones en artificiële intelligentie gemanipuleerd kunnen worden om actuele en gedetailleerde ruimtelijke data te verkrijgen. De simultane provisie van 2D beelden en 3D puntenwolken blijkt ontzettend bruikbaar te zijn en benadrukt het belang om doelgerichte geoinformatische processen te ontwikkelen. Deze moeten gebruik maken van de mogelijke synergiën in plaats van standaard algoritmes voor ofwel beelden ofwel puntwolken toe te passen. Zo kan men hoogwaardige kaarten verkrijgen door (grotendeels) automatische processen, ondanks de moeilijke kenmerken van sloppenwijken. Ten tweede, door samen te werken met echte sloppenwijk verbeteringsprojecten heeft men tijdens het onderzoek een uniek beeld kunnen verkrijgen betreft het daadwerkelijke gebruik van de data door de lokale overheden, ingenieurs, en bewoners. Bewoners van de sloppenwijken waren tijdens de data collectie geïntrigeerd, echter konden zij in de eerste instantie weinig praktische gebruik van deze beelden bedenken. Twee jaar later bleek, integendeel, dat de beelden gebruikt weerden voor verschillende

onverwachte toepassingen. Aangezien drones steeds toegankelijker worden, en het verwerken van hun data steeds eenvoudiger, wordt het ook steeds eenvoudiger om een toekomst voor te stellen waarin drones steeds vaker gebruikt worden om stedelijke ontwikkelingsprojecten te ondersteunen.

Authors Biography

Caroline received a BSc. Degree in International Land and Water Management at the University of Wageningen (the Netherlands) in 2011. She then completed an MSc in Remote Sensing at the University of Valencia (Spain) in 2013 and an MSc in Geographical Information Science at Lund University (Sweden) in 2014. In September 2014, she started as a Ph.D. candidate at the Earth Observation Science department of the Faculty of Geo-Information Science and Earth Observation (Faculty ITC), University of Twente, the Netherlands – which has resulted in the present manuscript. She has published various papers in leading remote sensing journals and was nominated for the Student Paper Award at the Joint Urban Remote Sensing Event (JURSE) 2017 in Dubai. Next to her academic activities, she is currently working part-time as a Senior Remote Sensing and Machine Learning consultant at the World Bank.

This research has produced the following publications besides the present manuscript:

Gevaert, CM, Sliuzas, RV, Persello, C & Vosselman, G 2015, Opportunities for UAV mapping to support unplanned settlement upgrading. in *Proceedings of GeoTech Rwanda 2015, 18-20 November 2015, Kigali, Rwanda.* pp. 1-5, GeoTech Rwanda, Kigali, Rwanda, 18-20 November.

Gevaert, CM, Sliuzas, RV, Persello, C & Vosselman, G 2016, 'Opportunities for UAV mapping to support unplanned settlement upgrading' *Rwanda journal : Series D : Life and natural sciences*, vol 1, no. SE 2, 4, pp. 1-19. DOI: 10.4314/rj.v1i1S.4D

Gevaert, CM, Persello, C & Vosselman, G 2016, 'Optimizing Multiple Kernel Learning for the classification of UAV data' *Remote sensing*, vol 8, no. 12, 1025, pp. 1-22. DOI: 10.3390/rs8121025

Koeva, MN, Muneza, M, Gevaert, CM, Gerke, M & Nex, FC 2016, 'Using UAVs for map creation and updating : a case study in Rwanda' *Survey review*, pp. -. DOI: 10.1080/00396265.2016.1268756

Gevaert, CM, Persello, C, Sliuzas, RV & Vosselman, G 2016, Classification of informal settlements through the integration of 2D and 3D features extracted from UAV data. in L Halounova [et al] (eds), *Proceedings of the XXIII ISPRS Congress : From human history to the future with spatial information, 12-19 July 2016, Prague, Czech Republic. Peer reviewed Annals, Volume III-3, 2016.* vol. III-3, ISPRS Annals of Photogrammaetry and Remote Sensing, no. III-3, vol. XLI-B1, International Society for Photogrammetry and Remote Sensing (ISPRS), pp. 317-324, XXIII ISPRS Congress, Prague, Czech Republic, 12-19 July. DOI: 10.5194/isprs-annals-III-3-317-2016

Gevaert, CM, Persello, C, Sliuzas, RV & Vosselman, G 2016, 'Integrating UAV point clouds and imagery : an application for informal settlement mapping : abstract' NCG Symposium 2016, Enschede, Netherlands, 30/11/16, .

Gevaert, CM, Persello, C, Nex, FC & Vosselman, G 2017, 'A Deep Learning Approach to DTM Extraction from Imagery Using Rule-Based Training Labels : powerpoint.' NCG Symposium 2017, Delft, Netherlands, 2/11/17 - 2/11/17, pp. 1s-25s.

Gevaert, CM, Persello, C, Sliuzas, RV & Vosselman, G 2017, 'Informal settlement classification using point-cloud and image-based features from UAV data' *ISPRS journal of photogrammetry and remote sensing*, vol. 125, pp. 225-236. DOI: 10.1016/j.isprsjprs.2017.01.017

Gevaert, CM, Persello, C, Oude Elberink, SJ, Vosselman, G & Sliuzas, RV 2017, An automated technique for basemap updating using UAV data. in *Proceedings of Joint urban remote sensing event (JURSE) 2017, 6-8 March 2017, Dubai, United Arab Emirates.* IEEE, Piscataway, pp. 1-4, Joint Urban Remote Sensing Event, Dubai, United Arab Emirates, 6-8 March. (Nominated for best Student Paper Award)

Sliuzas, RV, Kuffer, M, Pfeffer, K, Gevaert, CM & Persello, C 2017, Slum mapping : from space to unmanned aerial vehicle based approaches. in *Proceedings of Joint urban remote sensing event (JURSE) 2017, 6-8 March 2017, Dubai, United Arab Emirates.* Dubai, pp. 1-4, Joint Urban Remote Sensing Event, Dubai, United Arab Emirates, 6-8 March. DOI: 10.1109/JURSE.2017.7924589

Gevaert, CM, Persello, C, Elberink, SO, Vosselman, G & Sliuzas, R 2017, 'Context-Based Filtering of Noisy Labels for Automatic Basemap Updating From UAV Data' *IEEE Journal of selected topics in applied earth observations and remote sensing*, no. 99, pp. 1-11. DOI: 10.1109/JSTARS.2017.2762905

Gevaert, C.M., Sliuzas, R., Persello, C. and Vosselman, G., 2018. 'Evaluating the Societal Impact of Using Drones to Support Urban Upgrading Projects' *ISPRS International Journal of Geo-Information*, *7*(3), p.91. DOI: 10.3390/ijgi7030091

Gevaert, C.M., Persello, C., Nex, F. and Vosselman, G., 2018. 'A deep learning approach to DTM extraction from imagery using rule-based training labels' *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 142, pp.106-123. DOI: 10.1016/j.isprsjprs.2018.06.001 (Featured article August 2018).