UNIVERSITY OF NOTTINGHAM

Automatic Image Annotation Applied to Habitat Classification

by

Mercedes Torres Torres, MSc.

Thesis submitted to the University of Nottingham
for the degree of Doctor of Philosophy

January 2015

Mi vida es un erial:
flor que toco se deshoja;
que en mi camino fatal
alguien va sembrando el mal
para que yo lo recoja.

Gustavo Adolfo Bécquer

Abstract

Habitat classification, the process of mapping a site with its habitats, is a crucial activity for monitoring environmental biodiversity. Phase 1 classification, a 10-class four-tier hierarchical scheme, is the most widely used scheme in the UK. Currently, no automatic approaches have been developed and its classification is carried out exclusively by ecologists. This manual approach using surveyors is laborious, expensive and subjective. To this date, no automatic approach has been developed.

This thesis presents the first automatic system for Phase 1 classification. Our main contribution is an Automatic Image Annotation (AIA) framework for the automatic classification of Phase 1 habitats. This framework combines five elements to annotate unseen photographs: ground-taken geo-referenced photography, low-level visual features, medium-level semantic information, random projections forests and location-based weighted predictions.

Our second contribution are two fully-annotated ground-taken photograph datasets, the first publicly available databases specifically designed for the development of multimedia analysis techniques for ecological applications. Habitat 1K has over 1,000 photographs and 4,000 annotated habitats and Habitat 3K has over 3,000 images and 11,000 annotated habitats. This is the first time ground-taken photographs have been used with such ecological purposes.

Our third contribution is a novel Random Forest-based classifier: Random Projection Forests (RPF). RPFs use Random Projections as a dimensionality reduction mechanism in their split nodes. This new design makes their training and testing phase more efficient than those of the traditional implementation of Random Forests.

Our fourth contribution arises from the limitations that low-level features have when classifying similarly visual classes. Low-level features have been proven to be inadequate for discriminating high-level semantic concepts, such as habitat classes. Currently, only humans posses such high-level knowledge. In order to obtain this knowledge, we create a new type of feature, called medium-level features, which use a Human-In-The-Loop approach to extract crucial semantic information.

Our final contribution is a location-based voting system for RPFs. We benefit from the geographical properties of habitats to weight the predictions from the RPFs according to the geographical distance between unseen test photographs and photographs in the training set.

Results will show that ground-taken photographs are a promising source of information that can be successfully applied to Phase 1 classification. Experiments will demonstrate

that our AIA approach outperforms traditional Random Forests in terms of recall and precision. Moreover, both our modifications, the inclusion of medium-level knowledge and a location-based voting system, greatly improve the recall and precision of even the most complex habitats. This makes our complete image-annotation system, to the best of our knowledge, the most accurate automatic alternative to manual habitat classification for the complete categorization of Phase 1 habitats.

List of Publications

1. Torres, M., Qiu, G. (2014), Habitat Image Annotation with Low-Level Features, Medium-Level Knowledge and Location Information, Multimedia Systems (accepted for publication).[Chapter 8 and Chapter 9]

2. Torres, M., Qiu, G. (2014), Crowd-sourcing Applied to Photograph-Based Automatic Habitat Classification, Proceedings of the 3rd ACM International Workshop on Multimedia Analysis for Ecological Data, November 07-07, 2014, Orlando, USA. [Chapter 7]

3. Torres, M., Qiu, G. (2013), Automatic Habitat Classification using Image Analysis and Random Forest, Ecological Informatics. Available online 2 September 2013. [Chapter 6]

4. Torres, M., Qiu, G. (2013), Habitat classification using random forest based image annotation, IEEE International Conference on Image Processing (ICIP2013), September 15-18, Melbourne, Australia. [Chapter 6]

5. Torres, M., Qiu, G. (2012), Grass, scrub, trees and random forest, Proceedings of the 1st ACM International Workshop on Multimedia Analysis for Ecological Data, November 02-02, 2012, Nara, Japan. [Chapter 5]

6. Torres, M. (2012), Automatic habitat classification using aerial imagery, GIS Research UK 20th Annual Conference (GISRUK 2012), 11 - 13 Apr 2012, Lancaster University, Lancaster, UK. [Chapter 5]

Acknowledgements

First of all, my sincere thanks go to my supervisor, Dr. Guoping Qiu. Thank you very much for your guidance, your patience and your wisdom during these last four years. There is no doubt in my mind that I could not have survived this without your help and encouragement. Moreover, thank you for giving me the amazing opportunity of visiting China, an experience that I will always remember fondly.

Many thanks to my second supervisor, Dr. Gary Priestnall. Our discussions during my writing-up process were invaluable and I am extremely thankful for all your help and input.

My deepest gratitude to The Ordnance Survey for all their support during these past four years. Specially, I would like to thank Glen Hart, to whom I owe more than I can express. Thank you very much for striking up a conversation with me about old photographs during that ice-cold Industry Day at the DTC. I would also like to thank Carolina, Andy, Kat, Charlotte and Pernille for their help and feedback during and after my internship. More importantly, I would like to thank them for agreeing to spend several days walking around in the wilderness that is rural England just so I could have photographs to work with.

None of this would have been possible without the support of the the University of Nottingham and the RCUK Digital Economy programme. In particular, I would like to thank the Horizon Digital Economy Research Lab, Professor Sarah Sharples and Professor Steve Benford, for giving me the opportunity of being part of the Doctoral Training Centre. I would also like to thank Dr. Michel Valstar for his input and feedback during my annual internal assessments.

Many thanks to Emma Juggins for all her help and for always answering all my questions promptly and patiently. I would like to thank all members in VIPLAB, IMA and the Horizon DTC. Special thanks go to: Min Zhang, Hao Fu, Orod Razeghi, Fang Zhou, Shi Cheng, Bozhi Liu, Daphne Lai, Wenwen Bao, Jianhua Shao and Qian Zhang.

Finally, I would like to thank my family and friends. Thanks to Carmen for being the optimist that I can never be, for cheering me up during the most difficult times of this four-year journey and for never failing to make me laugh. Your are the Leslie to my Ron and my respect for you is as big as your love for SaB. So, you know, it's pretty big. I would also like to thank Kathy, for her understanding and constant encouragement.

Most of all, I would like to thank my mother. Resumir en un párrafo lo que no cabría en una biblioteca entera es prácticamente imposible, así que tendrás que perdonarme si

cometo alguna torpeza. Gracias. Por tu ayuda, tu tiempo, tu humor y tu paciencia. Por cantar Yolanda conmigo mientras derrotábamos a marcianos invasores, por llamar a Madrid mientras yo estaba en el colegio para que pudiéramos deshacernos de aquel maldito cofre, por ayudarme con (hacerme) aquellos horribles proyectos de Plástica, por todas y cada una de las vacas. Por aprender qué es un Random Forest y cuántas clases de matojos existen sólo para que pudiera hablarlo contigo. Pero, sobre todo, gracias por animarme a continuar y expandir mi educación, aunque eso significara vivir a más de dos mil kilómetros de distancia y aprender que, después de todo, Delilah no era una balada.

Contents

List of Figures

List of Tables

Abbreviations

AIA	**A**utomatic **I**mage **A**nnotation
FGVC	**F**ine-**G**rained **V**isual **C**ategorization
GPS	**G**lobal **P**ositioning **S**ystem
HC	**H**abitat **C**lassification
HITL	**H**uman **I**n **T**he **L**oop
JNCC	**J**oint **N**ature **C**onservation **C**ommittee
LIDAR	**Li**dar **D**etection **A**nd **R**anging
RF	**R**andom **F**orests
RPF	**R**andom **P**rojection **F**orests

Para Georgina.

Te quiero, te adoro, te llevo a La Macarena y te preparo una buena cena.

Chapter 1

Introduction

Habitat classification is an essential ecological activity which helps humans structure environmental knowledge and develop their understanding of the natural world. There are many manual and automatic habitat classification schemes that have been developed to this day. Their methodologies vary greatly depending on the subject, i.e. animals, plants, insects, etc.; their geographical location, i.e. coastal, rural, urban, etc.; and the types of data used, i.e. satellite imagery, aerial photographs, maps, etc.

This thesis deals with the problem of automatic Phase 1 habitat classification using ground-taken geo-referenced photographs. Our research is focused on the classification of wildlife habitats, more specifically, vegetation habitats, within the United Kingdom. We will be following the Phase 1 classification scheme, standardised by the Joint Nature Conservation Committee (JNCC) [102] and widely used by ecologists in the United Kingdom.

From a Computer Vision point of view, and given the similarities between the classes that we aim to classify, such as different types of grasses, heathland, water or woodland, automatic habitat classification can be regarded as a Fine-Grained Visual Categorization (FGVC) problem [24]. With this in mind, we have approached Phase 1 habitat classification from an image annotation perspective. We have created the first automatic framework for Phase 1 classification, whose inputs are unseen ground-taken geo-referenced photographs and whose output is a list of all possible habitats from more probable to less probable. In summary, the main goal of this thesis is to study the performance of our image-annotation framework for the specific purpose of Phase 1 Habitat classification. Moreover, we aim to study the merits and limitations of ground-taken imagery as the main source of information for automatic habitat classification and the effect that pattern, colour and texture features have in this classification process.

This thesis is organised in ten chapters. In this chapter, we describe the motivations and the technical challenges behind automatic habitat classification using ground-taken imagery and we list the contributions made in this thesis. Chapter 2 presents an overview of the state-of-the-art methods related to our research in both Ecology and Computer Vision, with special emphasis in Habitat Classification and Image Annotation methods. Chapter 3 will describe what Phase 1 Habitat Classification is in more detail. In Chapter 4 we present a brief study on the limitations that remote sensed data, and aerial imagery in particular, present when automatically classifying Phase 1 habitats. Chapter 5 presents a brief overview of the automatic annotation framework proposed in this thesis. Chapter 6 describes in detail the type of ground-taken imagery we will be working with in our framework. Chapter 7 describes the main novel contribution of this framework, Random Projection Forests. Chapter 8 will introduce the concept of Medium-level Knowledge and how its inclusion in our framework can improve the classification process. Chapter 9 describes how we have used geographical information during testing to obtain more accurate results. Finally, Chapter 10 summarises our contributions in this thesis, discusses the merits and limitations of our approach and offers some recommendations for future work.

1.1 Motivation and Technical Challenges

The worldwide fragmentation and destruction of habitats and their economic, biological and ethical consequences are considered to be one of the biggest challenges currently affecting our society [41]. Habitats are defined in the European Union Habitats Directive as terrestrial or aquatic areas distinguished by geographic abiotic and biotic features, whether natural or semi-natural [43]. Their classification and characterization has been carried out for more than one hundred years [7] and environmental agencies of countries such as the United Kingdom, Spain, Germany, Switzerland, Denmark and The Netherlands [138] maintain projects related to habitat monitoring.

The purpose of classifying habitats is twofold: firstly, it helps to reduce the complexity present in the natural world. Secondly, by categorizing habitats, their characterization and comparison can be done much more efficiently and effectively. While there are multiple schemes that have been developed to date, one of the most widely used by ecologists is the Phase 1 Habitat Survey scheme [102]. This standardised hierarchical classification divides all habitats into ten broad categories and it was designed to provide a detailed record of the vegetation and wildlife present in a determined area.

In essence, Phase 1 habitat classification can be regarded as a preliminary ecological procedure which serves to monitor and describe the ecological properties of an area. It

must be carried out before any other ecological activities that might affect an area can be executed. Trained ecologists will extract as much information as possible about the area and their classification and assessment will directly influence any other ecological decisions that may affect the aforementioned area. Consequently, there are many applications to habitat classification, such as habitat monitoring and identification, landscape ecology, and monitoring and conservation of rare species [111, 128, 158].

However, one of the main drawbacks of Phase 1 Habitat Classification is that it relies very heavily on human surveyors [102]. This manual approach is laborious, expensive and time consuming, since ecologists have to be deployed to the areas that need mapping. Additionally, it can also be extremely subjective, since there are many similarities between some of the finer habitat classes. Having an accurate automatic Phase 1 classification would greatly facilitate this process. Approaches have been developed with the aim of automating the habitat classification process [54] but, to our knowledge, no automatic alternative uses ground-taken imagery and no automatic methods have been presented for Phase 1 classification to this date.

One of the main reasons why fully accurate results have not been obtained is because most of the methods developed use remotely sensed data. Aerial photography and satellite imagery, in particular, seem to be the most popular choices for input data [20, 23]. Given the level of detail that is necessary to distinguish between some of the habitats collected in the Phase 1 Habitat Survey scheme, both aerial and satellite imagery have been proven to be insufficient [180].

For this thesis we have chosen an alternative source of information: ground-taken imagery [182]. Geo-referenced ground-taken photographs present two main advantages over aerial and satellite imagery. Firstly, ground-taken photography has a greater degree of detail. For FGVC problems, such as habitat classification, this is a decisive trait, since details will be crucial to differentiate between similar habitat classes. Secondly, they can be obtained more easily than aerial and satellite imagery, since the only equipment necessary is a digital camera or a smartphone. Moreover, it is also possible to use the Internet to obtain this type of data, with crowd-sourcing websites such as Geograph [154] or Flickr [125].

However, the use of ground-taken photography also presents some challenges. One of the main challenges is the varied nature of the photographs. Remotely sensed data, such as aerial or satellite imagery, commonly follows the same pattern and layout. The imagery is taken under the same conditions every single time: the camera is at a constant distance from the subject of the images and the angles between the camera and the subject are always the same. On the other hand, the ground-taken photographs used in this thesis are extremely varied in terms of layout, orientation and perspective. This was done

purposely in order to create a robust database that recorded as many types of habitats under as many different circumstances as possible. Nevertheless, this lack of control on the conditions under which the photographs are taken results into two different issues. First, the habitat class of the subject of the photograph might not be clearly discernibly due to the perspective or the layout of the photograph. This can be problematic for our automatic framework. For example, in Figure 1.1, the perspective of the image makes distinguishing whether the scrub shown belongs to the Scrub class (in Class A) or to the Hedge class (in Class J) difficult. That is one of the reasons an extensive and varied database and very precise ground truth data is extremely important in our case. Second, the lack of consistency in the perspective makes the locations of the photograph different from the location of the subject of the photographs. As shown in Figure 1.2, the location of the photograph will be one set of coordinates, while the habitats that appear on it expand a greater territory. This means that if geographical location is introduced in the classification process, some considerations need to be taken into account when measuring the performance of the framework.

FIGURE 1.1: Limitations of perspective and layout in ground-taken photographs. Given the perspective and layout of the image, it is difficult to distinguish whether the scrub shown belongs to Class A (Woodland and Scrub) or Class J (Miscellaneous).

In this thesis we have developed an image-annotation framework for automatic Phase 1 habitat classification using ground-taken imagery. From an Image Processing perspective, approaching automatic habitat classification as an image annotation problem presents an interesting and compelling set of technical challenges. Image annotation is an increasingly popular topic in Computer Vision [76, 169] and image annotation frameworks have been applied to medical, ecological and biological research [151].

FIGURE 1.2: Perspective, layout and ground-taken photographs. Given the perspective and layout of the image, the subject of the photograph expands further than the geographical location of the photograph.

However, what makes the problem of habitat classification a more challenging task than common image annotation problems is the nature of the classes that need to be recognised. Instead of conventional and clearly separable classes, such as *building, flower, tree, dog, cow, road, body, boat, mountain, forest* [150, 167], Phase 1 combines two very interesting characteristics. Firstly, it is a hierarchical classification. Phase 1 has ten first-level classes and extends to four levels for a total of 150 different habitat classes. Additionally, some of these classes may have similar components or similar types of vegetation. For example, as mentioned previously, scrub can be present on its own, as class A.2, or as part of a boundary habitat (Hedges, J.2). It can also appear as part class D.1., Dry dwarf and shrub heath.

Secondly, its classes are difficult to identify even by human surveyors. When classifying

Phase 1 habitats, the aim is not to classify trees, grass or water, for example, but to classify *which* kind of trees (broad-leaved or coniferous), grasses (improved, semi-improved or unimproved) or water (standing or running) appear in the photographs. This task is difficult even for trained Phase 1 experts and it may require previous knowledge of the ecological properties of the area. In Computer Vision, this type of problem, in which the classes to classify are very similar visually, is commonly referred to as Fine-Grained Visual Categorization problems (FGVC) [25].

In summary, our goal is to test and study the advantages and disadvantages that our image-annotation framework and ground-taken imagery provide when automatically classifying Phase 1 habitats.

1.2 Contributions

In this thesis, we make the following contributions:

- Image-Annotation Framework: We approach automatic habitat classification as an image annotation problem. We have developed and tested an automatic image-annotation framework for Phase 1 habitat classification. Our framework combines five main elements: ground-taken imagery, low-level visual features, medium-level information, random projections forests and geographical location to annotate unseen photographs using the Phase 1 classification scheme. This is the first instance in which ground-taken photographs have been combined with an Automatic Annotation methodology for the ecological purpose of habitat classification. Moreover, our framework is, to our knowledge, the first automatic framework specially designed for the complete classification of Phase 1 habitats. Extensive experimentation shows that our framework can successfully classify Phase 1 habitats in terms of precision and recall. [Chapter 5].

- Habitat 1K and Habitat 3K: Two fully annotated databases specially created for ecological purposes. Habitat 1K is composed of 1,086 photographs and 4,223 annotations from five habitat classes: Woodland and Scrub (A), Grassland and Marsh (B), Tall Herb and Fern (C), Heathland (D) and Miscellaneous (J). Habitat 3K has 3,094 ground-taken geo-referenced photographs. This database has been ground-truthed by a Phase 1 expert and it includes 11,517 different instances of habitats from seven out of the ten possible habitat classes. These are: Woodland and Scrub (A), Grassland and Marsh (B), Tall Herb and Fern (C), Heathland (D), Open Water (G), Coastland(H), Rock Exposure (I) and Miscellaneous (J). The photographs of these databases do not follow any particular layout. Therefore,

different perspectives, such as landscape shots, detail shots or ground shots are all allowed. These databases have been made publicly available [1] and they are the first visual databases specifically designed for the development of multimedia analysis techniques for ecological applications. [Chapter 6]

- Low-level Visual Features Applied to Habitat Classification: We carry out a study on the use of some of the most popular low-level visual features. Particularly, we study the effect that texture (Tamura coefficients, Grey-Level Co-occurrence Matrix), pattern (Colour Pattern Appearance Model) and colour (Colour Histograms and Colour Moments) features have on Phase 1 habitat classification when using ground-taken imagery. This helps us better understand the benefits and limitations that ground-taken imagery present when classifying Phase 1 habitats. Results will show that pattern and colour features obtain the most stable precision and recall results in more than 80% of the testing scenarios. On the other hand, texture feature can obtain more accurate results than pattern and colour in particular cases, such as the classification of Heath mosaics with Random Projection Forests, but their general performance in all experiments is considerably less stable. [Chapter 7, Chapter 8 and Chapter 9]

- Random Projection Forests (RPF): Random Forests is an increasingly popular machine learning technique that have been successfully applied to a varied number of problems in the field of computer vision, such as image classification [132] and image segmentation [167]. In the field of Ecology, they have also been applied to habitat structure classification [11] and land cover [81]. We chose to use this ensemble classifier because they combine the benefits of two other popular Machine Learning techniques, NN-based methods and SVMs, without being as affected by their disadvantages. Random forests are simple to implement and easy to modify to be applied to multi-label problems, similarly to NN-based methods. On the other hand, similarly to SVMs, they are accurate and do not suffer from a less efficient testing phase. We propose an alternative to Random Forests that uses Random Projections, a popular dimensionality reduction technique. With RPF, we generate a random projection vector with values -1, 0, 1 in each of the nodes of our decision tree and we project each feature vector according to the corresponding random projection vector. The inclusion of projections makes the training and testing processes more efficient without sacrificing accuracy in the results. Results show that our initial design of Random Projection Forests, as shown in Chapter 7, is not only more efficient, but also outperforms Random Forests both in terms of recall and precision. This difference in performance is clearly noticeable when

[1] http://www.viplab.cs.nott.ac.uk/download/habitat_classification_database.html

classifying Woodland and Scrub (A), Grassland and Marsh (B) and Heathland (D) habitats. [Chapter 7]

- Medium-Level Features: Low-level features have been proven to be inadequate for discriminating high-level concepts, such as habitat classes belonging to Phase 1. These limitations caused significant lack of accuracy in second- and third-tier habitats, such as boundaries and heathland mosaics. On the other hand, humans are able to identify objects belonging to different classes quite effortlessly using semantic information. In an effort to incorporate this higher-level information to the classification process, we adopt a Human-In-The-Loop (HITL) approach [24] to extract semantic information for the annotation process. Human-In-the-loop is an interactive, hybrid human-computer interaction method for object classification which aims to benefit from the strengths of both humans (their ability to differentiate between classes rapidly) and computers (their ability to compute large amounts of data efficiently). We have developed an innovative way to implement this HITL approach and we have successfully incorporated it to our AIA framework: non-experts users are asked a series of 'yes'-or-'no' questions about the ground-taken photographs in our database and we transform their answers to these questions, along with the certainty level they have on these responses, into medium-level features. These features are then used as the input of our classifier. Additionally, we combine these medium-level features with low-level visual features to obtain more accurate results in the most challenging habitat classes: Tall Herb and Fern (C) and Heathland (D). Experiments show that the inclusion of medium-level features entails a considerable improvement over our initial design of Random Projection Forests, particularly in terms of precision, which improves up to 20%. This increase is particularly noticeable in Tall Herb and Fern habitats (C) and complex habitats such as Hedge and Trees (J.2.3) and Heathland mosaics. [Chapter 8]

- Location-Based Voting System: We include geographical information during the annotation process. We take advantage of the geographical properties of habitats to improve the accuracy of our framework. Geographically close areas have similar ecological characteristics, since habitat properties do not generally change abruptly. Therefore, near regions will have similar habitats. Since all the images in the database are geo-referenced, we use their GPS coordinates to calculate the distance between unseen photographs and the ground-taken photographs of the leaves they have reached in the RPF. Consequently, we weight the different decision trees in our RPF, with closer trees having more weight in the prediction than further trees. Experiments will show that this final modification of Random Projections Forests yields the most accurate recall and precision results from all

the scenarios tested in this thesis. In particular, complex mosaics and Coastland (H) habitats, which have proven specially difficult to classify, experience a considerable recall and precision improvement over past modifications. Consequently, this final contribution, to our knowledge, makes our Random Projection Forests with medium-level features and a location-based voting system the first and most accurate automatic framework specifically designed for the classification of the complete Phase 1 scheme. [Chapter 9]

1.3 Summary

In this chapter we have introduced the problem we aim to tackle in this thesis: automatic Phase 1 habitat classification. Moreover, we have described our contributions and we have introduced the methodology which we will be following: Automatic Image Annotation.

In the next chapter, we will present a comprehensive review of significant literature in the areas of Ecology and Computer Vision, with the aim of delimiting the clear research gap in current methodologies with regards to automatically classifying Phase 1 habitats.

Chapter 2

Literature Review

This thesis aims to incorporate work from two different disciplines: Ecology and Computer Vision. Accordingly, in this chapter we give an overview of the state-of-the-art methods related to our image annotation approach for the classification of habitats in both areas with the aim of presenting the clear research gap in literature.

This chapter is divided into two sections: Section 2.1 reviews current methods for habitat classification in Ecology. Section 2.1.1 reviews some of the most popular habitat classification schemes currently used and explains why we have chosen to work with the Phase 1 scheme in particular. We review merits and limitations of both manual and automatic approaches in Section 2.1.2 and Section 2.1.3 respectively. On the other hand, Section 2.2 examines related methods in the area of Computer Vision, focusing on current image annotation methods. In this section we review related state-of-the-art techniques for visual feature extraction, shown in Section 2.2.1, image annotation and fine-grained visual categorization problems, shown in Section 2.2.3, and machine learning, shown in Section 2.2.4, with special emphasis in the machine learning technique we have chosen, Random Forests, in Section 2.2.4.4. Finally, Section 2.3 briefly summarises the contents of this chapter.

2.1 Ecology

In Ecology, habitat classification is defined as the process of mapping all habitats present in an area according to a determined scheme [102]. The classification of habitats is a crucial activity for structuring knowledge and developing our understanding of the natural world. It has been carried out for more than two hundred years all over the

world [138] with the first recorded instance of habitat classification done by Linnaeus [51].

2.1.1 Habitat Classification Schemes

There are numerous terrestrial and freshwater habitat classification schemes that have been developed worldwide. While the overall aim of all these classifications is the same, to map the habitats present in a site, their characteristics vary depending on the nature of the vegetation that needs to be classified and on the geographical area of these sites. In this section we will introduce some of the most popular habitat schemes in Europe including Phase 1, the scheme we will be using in this thesis. Moreover, we will also compare these classifications to Phase 1 in order to better explain why we have chosen Phase 1.

These habitat classification schemes are:

- **European Nature Information System (EUNIS)**: This framework was first implemented in the late 1990s by the European Environment Agency and continues to be updated periodically [51]. EUNIS has a database that follows a very comprehensive classification scheme which records information about species, habitat types and sites. Their data was collected in the framework of NATURA2000 [39]. Moreover, it was also compiled from the literature [51].

  In this scheme, the concept of habitat is much broader than in Phase 1. In EUNIS, habitats are defined as: "Plant and animal communities as the characterising elements of the biotic environment, together with abiotic factors operating together at a particular scale". Table 2.1 shows the first-tier categories for only the Habitat classes. The EUNIS classification is a hierarchical scheme. It has 11 first-tier classes and four levels. After the fourth tier, the component units are drawn from other classification systems and these are combined in the common framework.

  Compared to Phase 1, which was designed specifically for habitats in the United Kingdom, the EUNIS classification scheme is a comprehensive pan-European system which aims to facilitate the collective description and collection of data across Europe through the use of standardised criteria for habitat identification. It covers all types of habitats from natural to artificial and from terrestrial to freshwater and marine. For this reason, this scheme is very useful when comparing species, habitats or sites of different European countries. However, since in our case we are only interested in habitats from the UK, this scheme is not a suitable candidate.

Table 2.1: EUNIS Habitat Classification Classes.

Code	Habitat Class
A	Marine habitats
B	Coastal habitats
C	Inland surface waters
D	Mires, bogs and fens
E	Grasslands and lands dominated by forbs, mosses or lichens
F	Heathland, scrub and tundra
G	Woodland, forest and other wooded land
H	Inland unvegetated or sparsely vegetated habitats
I	Cultivated agricultural, horticultural and domestic habitats
J	Constructed, industrial and other artificial habitats
X	Habitat complexes

- **The International Union for Conservation of Nature (IUCN) Habitats Classification Scheme**: First introduced in 1994 by the IUCN, this world-wide classification scheme is one of the most comprehensive approaches for the evaluation of the conservation level of habitats and wildlife. Its main goal is to collect information not only about the species present in an area, but also about their conservation status.

  This classification is a hierarchical scheme with eighteen broad classes and two levels. In comparison with Phase 1, this classification collects more information for some particular habitats, such as deserts and marine habitats. As shown in Table 2.2, IUCN's classification has six different classes devised to categorise marine or aquatic habitats (classes 9, 10, 11, 12, 13, 15), while Phase 1 only has three (classes F, G, H). However, IUCN's scheme fails to collect information about one of the most complex and useful habitats found in rural areas: boundaries. Phase 1 considers five different types of boundaries in its Miscellaneous category (Hedges, Fences, Walls, Dry ditches, Boundaries removed and Earth banks), while IUCN's classification does not distinguish between them and would consider all of them to be part of the Other category.

- **Fossit's Irish Habitat Classification**: Proposed in 2000 by Julie A. Fossit and The Heritage Council, this scheme presents a standard classification for identifying, describing and classifying wildlife habitats in Ireland [69]. It covers natural, semi-natural and artificial habitats. Moreover, it classifies terrestrial and freshwater environments, of inshore marine waters, and of urban and rural areas. As the previous schemes, this classification is hierarchical with eleven broad classes and has three tiers.

  Similarly to Phase 1, its various levels can be applied depending on the scale of the project, the details needed and the expertise of the surveyor. However, contrary

TABLE 2.2: IUCN's Habitat Classification Scheme

Code	Habitat Class
1	Forest
2	Savanna
3	Shrubland
4	Grassland
5	Wetlands
6	Rocky Areas
7	Caves and Subterranean
8	Desert
9	Marine Neritic
10	Marine Oceanic
11	Marine Deep Ocean Floor
12	Marine Intertidal
13	Marine Coastal/Supratidal
14	Artificial - Terrestrial
15	Artificial - Aquatic
16	Introduced Vegetation
17	Other
18	Unknown

TABLE 2.3: Fossil's Habitat Classification Scheme

Code	Habitat Class
F	Freshwater
G	Grassland and marsh
H	Heath and Dense Bracken
P	Peatlands
W	Woodland and Scrub
E	Exposed Rock
B	Cultivated and Built Land
C	Coastal
L	Litoral
S	Sublitoral
M	Marine Water Body

to Phase 1, Fossit created this classification as a first-step approach for general habitat recording rather than as a basis for detailed study and evaluation [69]. The main aim of this classification was to create a standard scheme, which Ireland lacked until Fossit's scheme.

As can be seen, this classification has many common classes with Phase 1 (i.e., Woodland and Scrub, Grassland and Marsh, Coastal). However, like previous classifications, it fails to take into account boundaries between habitats.

- **Phase 1 Habitat Classification**: A standardised classification scheme proposed by the Joint Nature Conservation Committee (JNCC) [102]. It was first introduced

TABLE 2.4: Phase 1 Habitat Classification Classes.

Code	Habitat Class
A	Woodland and scrub
B	Grassland and marsh
C	Tall herb and fern
D	Heathland
E	Mires
F	Swamp, marginal and inundation
G	Open water
H	Coastland
I	Rock exposure and waste
J	Miscellaneous

in the 1970s in the United Kingdom and it is specially designed for rapid wildlife mapping over large areas of the countryside. Similarly to all the previous schemes, this classification is hierarchical and it comprises ten categories, shown in Table 2.4. It has four tiers that enable ecologists to select the level of detail necessary on their survey depending on their expertise and the requirements of the project. A more in depth description on the characteristics and the challenges of Phase 1 classification can be found in Section 3.1. Additionally, the whole classification scheme can be found in [102]. In this thesis, we have chosen Phase 1 habitat classification because it is widely used by ecologists and because it was specifically designed to be applied in Great Britain and Ireland. This is very suitable for us because all the images in our ground-taken photograph database are from Great Britain.

2.1.2 Manual Habitat Classification

As mentioned in the previous section, Phase 1 classification relies heavily on human surveyors to manually classify and map areas. This requires training the surveyors and deploying them to a particular site that needs mapping. Then, using maps, the ecologists will survey the whole area and annotate the habitats found in their path. Figure 2.1 shows how a group of Phase 1 surveyors may classify habitats manually.

The process of manually classifying Phase 1 habitats is summarised in [102] as:

1. The surveyor visits every parcel of land within the survey area.
2. The vegetation that surveyors encounter in their path is mapped onto a habitat map (usually using 1:10,000 scale). Often this can be done from a road or footpath without the need to walk the ground but, depending on the area, surveyors might need to enter the sites.

FIGURE 2.1: Manual Habitat Classification. Trained human surveyors manually classify habitats in Titchfield Haven, United Kingdom, July 2011.

3. Phase 1 guidelines establish standard alphanumerical and colour codes for each habitat class. Surveyors use them to classify habitats into one of around 150 specified habitat types, allowing rapid visual assessment of the extent and distribution of habitat types.

4. Along with creating maps, surveyors are encouraged to take target notes. These notes record habitat descriptions, site-related information such as species, communities or presence of any species of conservation concern, and any other information of interest.

5. Once the area is mapped, statistics may be obtained regarding the extent and distribution of each habitat type.

6. The end products of a Phase 1 survey are: habitat maps, target notes and statistics, together with a descriptive and interpretive report.

As can be inferred, this manual approach has several drawbacks [102]. These include:

- Specific training: Phase 1 habitat classification requires additional training for ecologists. Consequently, time and resources must be allocated to train ecologists or to hire experts.

- Previous knowledge: Previous knowledge of the site is often required to accurately classify its habitats. Phase 1 habitat classification collects information about ecological characteristics that may not be completely visible to the surveyors. For example, geographical properties of the ground, such as whether it is calcareous, neutral or acid ground are required to classify grasslands. Trained ecologists will consult the geographical properties of the area they have to map in advance in order to obtain this information. Common sources of information include: Natural England [61], the Environment Agency [3], the MultiAgency Geographic Information for the Countryside [62] and the Joint Nature Conservation Committee [40]. Additionally, it is also common for ecologists to consult satellite or aerial imagery of the site before visiting the area to gather more information.

- Labour intensive: Ecologists need to cover the whole site that needs mapping on foot. If the area is large or difficult to access, this can be a very labour-intensive task.

- Time consuming: Depending on the size and the characteristics of the area that needs mapping, covering the whole site may be very time consuming. Moreover, the time that it takes to deploy the experts to and from the area of the survey also needs to be taken into account.

- Costly: Related to the two previous points, depending on the area that needs mapping, it could be necessary to either employ more than one ecologist or to allocate more time to map the areas that need classifying. Additional expenses to take into consideration also include the cost of transporting the surveyors to and from the site and other costs that may occur during the survey.

- Physical Output: In general, ecologists will use pen and paper when surveying a site and classifying its habitats. One of the outputs produced by ecologists are classification maps, an example of which is shown in Figure 2.2. In order to ensure that the information is not lost or misplaced, these maps are digitalised or scanned for safekeeping once the survey is finished. This process can be tedious and, if the weather conditions are bad, for example, if it is raining or snowing during the survey, this can negatively affect the state of the maps. Moreover, if maps are digitised instead of only scanned, the people in charge of digitizing these maps can introduce errors.

- Timing of the Survey: Phase 1 is recommended to be undertaken between the months of April and October, when deciduous and annual plant species are more easily identifiable, due to weather constrictions that may make habitat classification difficult [55]. This greatly restricts the time in which is possible to obtain new and updated data.

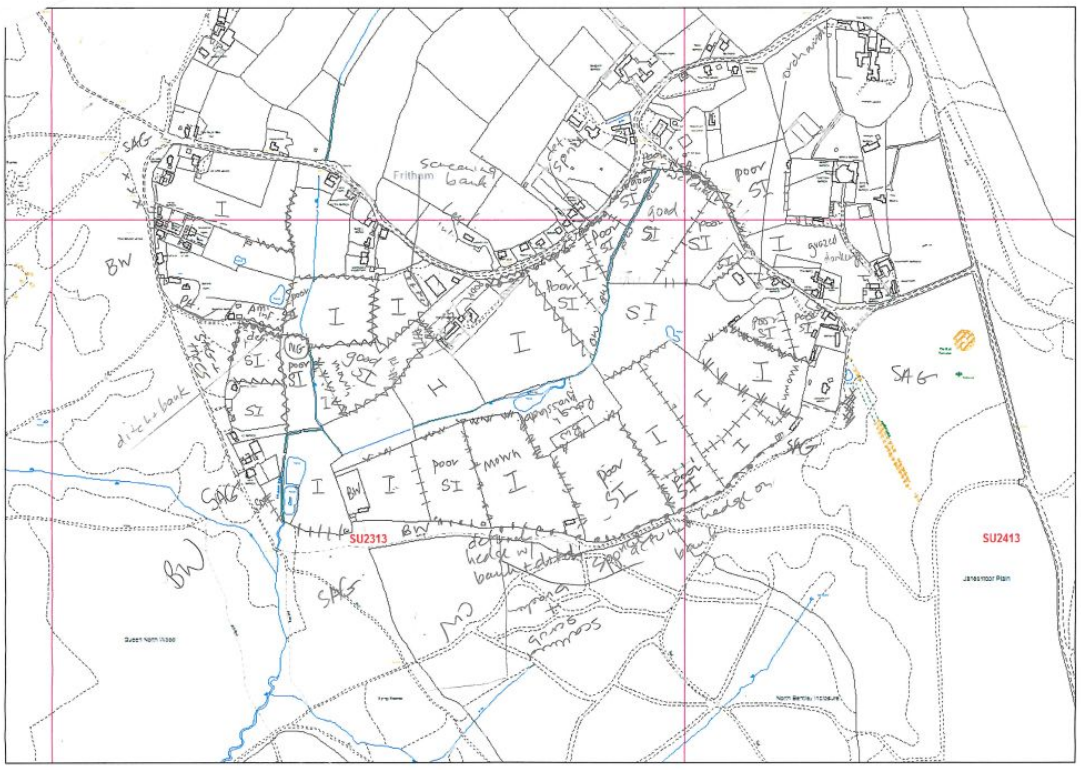


FIGURE 2.2: Habitat Map. Output from a trained Phase 1 ecologist for the area of New Forest, United Kingdom. BW stands for Broadleaved Woodland, I stands for Improved Grassland, SI stands for Semi-Improved Grassland and SAG stands for Acid-grassland Semi-Improved

- Subjective: Given the similarities between some of the habitats classified in Phase 1, such as Semi-Improved and Improved grasslands, their classification can be subjective or inconsistent.

2.1.3 Automatic Habitat Classification and Remote Sensing

In order to improve manual habitat classification, there are several automatic habitat classification methods that have been developed for different classification schemes. Examples of these include [35, 54]. It is interesting to notice that, to our knowledge, no previous work has been done for the automation of Phase 1 habitat classification in particular. Consequently, the approaches reviewed in this section use other classification schemes. Moreover, none of the automatic approaches developed to date use ground-taken photographs as the main source of data.

An automatic habitat classification approach, like the one proposed in this thesis and those reviewed here, could ideally eliminate the disadvantages presented in the previous section regarding manual habitat classification. Users would need no additional training to use the automatic system and they might not even need to be ecologists. For example,

the EUNIS framework enables non-expert users to search for habitat information by geographical site or by species [51].

Moreover, these users would not have to have previous knowledge about the site they want to classify. In fact, this external knowledge could be implemented into the automatic system in different forms. For example, [38] combined Light Detection And Ranging (LiDAR) height and intensity information with multi-spectral imagery to map coastal habitats in the Basque Country. Moreover, [163] combined Shuttle Radar Topography Mission (SRTM) data and Landsat TM imagery to classify habitats in neotropical environments. In our case, our framework could ultimately combine multiple sources of information such as aerial imagery and ground-taken photographs. As we will show in Chapter 8, contextual information could also be taken into consideration in the classification process and could even be used as part of the input. For example, if imagery was used as input, information such as the time of the year in which the image was taken, the geographical location of the site and even past results from other surveys from the same area could be added automatically. In particular, in our work, we have combined low-level visual features, medium-level contextual features and geographical location in the classification process.

Additionally, there would be no need to transport any humans to these sites which would save time, labour and money. The outputs would be already digitised and human errors would not be introduced during this process. Moreover, the system's decision making process would be uniform. Consequently, the classification would be equally uniform and there would not be any subjectivity involved in the process. Finally, the output statistics needed could be easily calculated using a computer and the already digitised information.

Most of the automatic approaches proposed in the literature use remote sensing imagery [54]. Consequently, remote sensing is defined as "the science and art of obtaining information about an object, area, or phenomenon through the analysis of data acquired by a device that is not in contact with the object, area, or phenomenon under investigation" [117]. Remote sensing data is data that has been obtained using remote sensing methods. Common types of remote sensing imagery that has been used for habitat and species classification include: aerial imagery [45], satellite photography [111], LiDAR [38] and hyperspectral imagery [212].

The use of remote sensing imagery has several advantages over manual approaches: they are more exhaustive, data can be periodically recorded and they can be used to record spectral information in non-visible regions of the electromagnetic spectrum [54]. Moreover, while the data collection phase might be time consuming and requires specialised equipment, the classification process is faster and more efficient.

2.1.3.1 Pixel-Oriented and Object-Oriented Classification Methods

Classification methods using remote sensing can be divided into two classes: pixel-oriented methods and object-oriented methods. In pixel-oriented methods, each pixel is classified individually and independently of the other pixels within the image. This type of classification is referred to as *spectral pattern recognition*. On the other hand, object-oriented classifiers use *spectral* and *spatial pattern recognition*. This means that when classifying a pixel, its surroundings and how the same pixel's value change over time are taken into consideration during the classification process. Object-oriented methods normally involve two steps: first, the image is segmented into discrete objects and then each object is classified. These methods approach classification in a way similar to how humans approach digital imagery interpretation, which uses different types of information, such as colour, shape, size, texture, pattern and context to group pixels into meaningful objects [117]. Pixel-oriented methods frequently obtain less accurate results and this can be attributed to the fact that, by taking only one pixel into consideration, there is a lot of spatial, temporal and contextual information that is being ignored in the classification process.

Object-oriented methods usually obtain more accurate results. Consequently, they are more popular when developing automatic habitat classification systems [117]. For example, in [54], Díaz Varela el al. used satellite imagery to classify habitats in the western end of the Cantabrian Coast, in Spain. They compared the performance of a Nearest-Neighbour and a maximum likelihood classifiers using an object-oriented and pixel-oriented approach and object-oriented methods outperformed pixel-oriented methods. [37] also followed an object-oriented methodology, using multi-temporal satellite imagery as part of their system for detecting shoreline changes for tideland areas and obtain an error rate of less that 15.5%. Moreover, [111] used a hierarchical inductive classification of satellite imagery to identify native grasslands in eastern Kansas. They used discriminant analysis of ground occurrence data that was extrapolated to distinguish high-quality from low-quality grasslands. [212] also followed an object-based methodology, by extracting texture measures from hyperspectral imagery and using a neural-network approach. They successfully applied it to map vegetation Everglades and obtained a 94% accuracy. In [5] coastal and marine ecological classification standard using satellite-derived and modeled data products for pelagic habitats in the Northern Gulf of Mexico. And [45] used aerial images to classify wetlands and deep-water habitats of the United States.

2.1.3.2 Limitations of Remote Sensing Data and Methods

Currently, most remote sensing methods focus on land cover or land use over habitat classification [7, 79, 81, 177]. While land cover and land use share some similarities with habitat classification, the classification schemes used and the applications are extremely different. For example land cover and land use classifications collect very little to none biodiversity or species information [179]. Consequently, they cannot be applied to habitat monitoring or rare species monitoring and conservation. Research has been done on how to translate between land cover methodologies and habitat classification [145, 179].

Those methods that focus on habitat classification tend to either focus on mapping particular habitat species instead of using complete schemes, such as [176, 212], or to create their own classifications depending on the site to map, such as [54, 111]. The former leads to relative or incomplete results [4] and the latter leads to results which are very dependant on the site and not easily comparable with other classifiers. For example, in [54], instead of using a standardised habitat scheme, the authors developed their own hierarchical classification with fifteen first-tier classes according to the geographical characteristics of the particular site they were classifying. While their results were very promising, their classification was tied to the particular site they were mapping. The same problem arises from [124], in which the authors combined aerial photography, CASI and HyMap data to classify forests in Australia. Once again, the authors created a new scheme instead of using an standardised classification.

Moreover, the use of remote sensing imagery to classify habitats is frequently hampered by the presence of complex habitats, such as mosaics with combinations of different habitats, complex canopy structures and transitions of vegetation types [67, 187]. These problems are very common in mountain areas, fragmented ecosystems, tropical environments or fine patterned landscapes [34, 99].

Additionally, in the specific case of Phase 1 classification, the use of remote-sensed imagery to categorise habitats presents additional disadvantages. Table 2.5 summarises the limitations of aerial and satellite images in comparison with manual Phase 1 habitat classification [102].

In conclusion, several automatic habitat classification methods have been developed to date. However, no automatic Phase 1 habitat classification system has been created. Additionally, most of the automatic systems use remote-sensed imagery, such as satellite or aerial imagery. Remote-sensed imagery on its own has several limitations, specially for the case of Phase 1, which requires a large level of detail. Moreover, it can be difficult to obtain. These are the reasons why we have created an automatic framework

TABLE 2.5: Comparison of manual Phase 1 habitat survey and automatic habitat classification using aerial photography and satellite imagery [102].

Classification Methods	Manual survey	Aerial Photography	Aerial Satellite Photography
Data Coverage	Complete ground cover possible	Incomplete for some dates. Variable quality	Complete cover but data can be obscured by clouds
Data Collection	Direct recording in the field by humans	Relies on tone and pattern of spectral reflectance	More limited range of tones but greater contrast than aerial photography
Equipment	No sophisticated or expensive equipment required	Needs complicated and expensive equipment	Needs complicated and expensive equipment
Accuracy	Accuracy depends on field surveyors	Images are accurate	Images are accurate
Interpretation	Interpretation problems	Interpretation can be difficult	Interpretation can be difficult
Habitat Coverage	Yields complete set of Phase 1 habitat categories	Yields limited set of habitat categories	Yields limited set of habitat categories
Canopy Information	Gives information on canopy and groundlayer	Information on canopy only (unless repeated at different seasons)	Information on canopy only (unless repeated at different seasons)
Species Information	Gives information on dominant and other plant species	Little species information	Very little species information
Conservation Evaluation	Can be used for conservation evaluation	Limited used for conservation evaluation	Limited use for conservation evaluation

for habitat classification and why we will be working with ground-taken images for the automatic classification of Phase 1 habitats in this thesis.

2.2 Computer Vision

From a Computer Vision point of view, automatic habitat classification using ground-taken imagery presents a collection of interesting challenges with a wide array of possibilities.

In this thesis, we have created a framework that follows an image-annotation approach and combines feature extraction, random projections and multi-label random forests. The literature for these areas of Computer Vision is vast, varied and ever-growing. In this section we review the most relevant and state-of-the-art methods with direct applications to our problem.

2.2.1 Feature Extraction

Features are one of the cornerstone concepts in modern Computer Vision. A feature is defined as "a piece of information which is relevant for solving the computational task related to a certain application" [88]. The selection of appropriate features is one of the most challenging tasks in Computer Vision problems, since it will directly influence the performance of the approaches chosen and it is data and problem dependent [114].

In image processing, the aim of extracting features is to collect the most compact but descriptive information about an image. By extracting meaningful features and using them as input in the classifiers, we do not have to work with all the pixels in an image, which can be time consuming and, depending on the task we wish to accomplish, unnecessary. Thus, extracting features is a dimensionality reduction mechanism whose goal is to improve efficiency and reduce storage space.

Defining feature vectors remains one of the most common and convenient means of data representation for classification and regression problems [88]. There is a large number of methods that have been developed with the goal of extracting meaningful and descriptive features [87] and features have been successfully applied to numerous, diverse and popular Computer Vision problems, such as object and scene recognition [169] and human action recognition [164]. Moreover, they have also been applied to multiple ecological problems, such as land use\land cover [74, 100], change detection [37] and habitat monitoring [31, 192].

Depending on their processing primitives, we can classify feature extraction methods into three categories: pixel-level, regional-level and image-level feature extraction. A pixel can be defined by two properties, its colour feature, normally represented by its RGB values, and its geometric position (x,y) within the image. Notable works on pixel-level feature extraction include [166], which proposes extracting features from a pixel while also taking into consideration its neighbours. The authors randomly crop rectangles from the neighborhood of the pixel and extract features from these. This feature together with the location of the rectangle are assembled as a two-tuple and considered to be the feature of the pixel.

Regional-level feature extraction is one of the most popular feature-extraction approaches. There are a multitude of regional-level features that have been proposed in the literature. Some of the most successful include: colour histograms, colour SIFT [191], texton histograms [195], Tamura features [175], Gray-Level Co-occurrence Matrices (GLCM) features [84], Histogram of Oriented Gradient (HOG) [48], geometry features [178] and Scale-Invariant Feature Transform (SIFT) [123] features. Additionally, [19] proposes a kernel descriptor that can turn pixel attributes into regional features.

Once regional information is extracted, they are combined to create a final regional feature. This combination may be as simple as concatenating the regional information. However, concatenation is the least recommendable method, as it would produce features of large dimensions and worsen "the curse of dimensionality" [14], which is one of the biggest problems in image processing.

Instead of concatenating the information, one of the most common approaches is to combine them using a "bag-of-words" methodology. A bag-of-words approach uses a pre-trained codebook and assigns each feature an index, referred to as a word, in this codebook. The final feature is then the histogram of all the words in the image [169]. Other methods include the covariance matrix representation [189, 215], graph representation [12] and fisher vector representation [142].

Moreover, recent works in feature extraction propose the creation of higher-level features, referred to as image-level features. These features are obtained by using the outputs of classifiers. [116] combines the results of several object-recognition classifiers as higher-level features and uses them for image classification. [156] develop this idea further, by using the output of many separate action detectors as higher-features, which are used for action recognition. In another example, [75] create high-level features with the output from a segmentation algorithm. These higher-level features are then combined with low-level features to generate object-consistent regions.

In our case, since we are working on automatically classifying different types of vegetation, we are specially interested in regional-level features. Particularly, we will be using colour, texture and pattern features, such as colour histograms, the Tamura coefficients and the GLCM matrix features. The reason for this is that we wish to mirror the surveyor's method on how to distinguish between habitats. However, we also study the performance of other well-known regional features, such as SIFT and HoG, in order to further study the performance of ground-taken imagery.

2.2.2 Random projections

As mentioned in Section 2.2.1, the "curse of dimensionality" is one of the main problems in Computer Vision. The concept of "Curse of dimensionality" was first introduced by Richard Bellman in 1961 [14]. It refers to the problems caused by increasing the number of dimensions of a mathematical space. It is a major obstacle in high dimensional data analysis because increasing the number of dimensions results in an exponential increase in sparsity between samples [201]. That is, as the number of dimensions increases, points that were close, spread further apart. This can result in inaccuracies during the classification process. Particularly in image processing, it often refers to the increase in the number of dimensions of the feature vectors the classifiers use.

There have been many approaches developed to remedy the issues brought by the "curse of dimensionality". [204] divided these methods into two categories: Function-Approximation approaches [91], popular in the past but not so widely used in current research, and Dimension-Reduction approaches, the most popular methodology used currently. A traditional example of dimension-reduction method is Principal Component Analysis (PCA) [49]. Arguably the most popular method currently used in image processing, PCA is a statistical procedure that reduces the dimensionality of the data by finding a low-dimensional subspace that maximises data variance.

Random Projections is another example of a dimension-reduction method. Dimension-reduction methods, particularly Random Projections, make use of the Johnson-Lindenstrauss lemma [30] in order to decrease the number of dimensions of the data. This lemma states that:

Given $0 < \varepsilon < l$, a set X of m points in $\mathbf{R}^N$, and a number $n > 8ln(m)/\varepsilon^2$, there is a linear map $f : \mathbf{R}^N \to \mathbf{R}^n$ such that

$$\big(1-\varepsilon\big)\|u-v\|^2 \leq \big\| f(u)-f(v)\|^2 \leq (1-\varepsilon)\|u+v\|^2 \tag{2.1}$$

for all $u, v \in X$

This entails that small set of points in a high-dimensional space can be embedded into a space of much lower dimension in such a way that distances between the points are nearly preserved. In the case of Random Projections, this dimension reduction is done by projecting the original data onto a subspace using a random matrix, whose columns have unit length [101]. This is commonly done by processing the scalar multiplication between the random projection matrix and the original data. The result data is labelled as the projection of the original data. These projections, whose dimensions will vary depending on the sparsity of the random projection matrix chosen, are then used as the new input for the classifiers. As can be seen, the required computation for applying Random Projections to data is quite small, since the only operation that needs to be executed is the scalar multiplication of the random projection matrix and the data itself.

Comparisons between both PCA and Random Projections are frequent, since, due to its popularity, PCA is commonly used as a benchmark for dimension-reduction methods' performance. In comparison with PCA, Random Projections are data independent and less computationally expensive [57]. Nevertheless, according to [71], Random Projections also can be outperformed by PCA, depending on the classifier used and, more importantly, depending on the number of dimensions of the reduced dataset. As shown in 2.1, the number of reduced dimensions needs to be at least $8ln(m)/\varepsilon^2$ for the random projections to be effective. If the number of reduced dimensions is too small, random projections can perform erratically.

However, given their relatively simple computation, Random Projections have steadily become more popular in the Computer Vision community and they have been used in a variety of problems in both image and signal processing [59]. They have been applied to hyperspectral imagery [57, 70], speech recognition [174] and face recognition [82, 203]. Moreover, [2], similarly to our work in [182] use random projections matrices with values -1,0,+1 and conclude that these values are specially suited for database dimension reduction.

2.2.3 Image Annotation

In this thesis, we approach habitat classification using ground-taken imagery as an automatic image annotation problem. In this scenario, the annotations are the different Phase 1 habitat classes. Consequently, our goal is to correctly identify which habitat classes are present in which photographs or, in other words, which annotations belong to which photographs.

Automatic image annotation (AIA) is an increasingly popular approach often used in the Computer Vision community. AIA was developed as mechanism to deal with the

exponential increase in visual data [185]. For example, Flickr surpassed 6 billion photographs in 2011, only six years after its foundation [125] and Geograph is hosting almost 4 million photographs from England, Ireland and the Isle of Man as of April 2014 [154]. Traditional image retrieval techniques proved to be lacking when dealing with such a large number of images, specially due to the gap between content-based image retrieval and classification and image semantics understandable by humans [213]. This gap is often referred to in literature as the semantic gap [185].

AIA methods can be regarded as particularly well-suited methods to bridge the semantic gap between low level features and high level semantics. In essence, AIA methods were developed to facilitate the search and navigation of large numbers of images. In [213], the authors propose AIA as an alternative to content-based and text-based annotation image retrieval.

The main aim of AIA methods is to automatically learn semantic concept models, in the form of annotations, from a large number of samples, images in our case. Then, new unseen images are labeled using these models. For this, semantically labelled images are collected and significant features, such as those discussed in Section 2.2.1, are extracted. These are used in conjunction with a machine learning algorithm that, once trained, will be used to annotate unseen samples.

AIA methods can be divided into three categories: single labelling annotations, multi-labelling annotations and annotations which use metadata to annotate images [213]. Our problem is inherently a multi-label problem, since the ground-taken photographs that we have collected contain a variable number of habitats. Moreover, we have used metadata in the decision-making process. Consequently, in this thesis, we have created a hybrid annotation framework which mixes approaches from the second and third categories.

There are many methods that have been developed for image annotation with general classes, also referred to as basic-level classes [209]. For example, [150] combined image annotation with semantic information and bag-of-features to classify photographs according to twenty-one classes such as *building, grass, tree, cow, water, chair, road* and *cat*. [167] used semantic texton forests to annotate and classify images with a similar classification scheme. [25] combined interactive and online learning to create a framework that was able to annotate bird images. [112] also developed a method for indoor and outdoor scene recognition based on partitioning an image into increasingly finer sub-regions and computing their histograms.

However, what makes the problem of habitat classification different from other image annotation problems is the nature of the classes that need to be recognised. Most of the existing AIA research focuses on object [22, 66, 150] or scene [112] recognition

and annotation. In those works, the classes are easily identifiable, they do not share semantic properties and their classification is regarded as basic-level categorization (i.e. distinguishing between a *boat* and a *cow*, a *chair* and a *building*).

However, instead of conventional and clearly separable classes, such as *building, flower, tree, dog, cow, road, body, boat, mountain, forest* [150, 167], Phase 1 is a hierarchical classification whose classes are difficult to identify and tell apart even for human surveyors [102]. As mentioned in Chapter 1, the aim in this case, instead of classifying trees, grass or water, for example, is to classify *which kind* of trees (broad-leaved or coniferous), grasses (improved, semi-improved or unimproved) or water (standing or running) appear in the photographs. In Computer Vision, this type of problems are referred to as fine-grained visual categorization problems (FGVC) [24]. FGVC, in contrast to the concept of basic-level categorization presented previously, is also known as subordinate-level categorization [209]. In FGVC problems, the aim is the accurate discrimination between classes that share similar semantics [205].

FGVC has gained much interest in the Computer Vision field in the last few years mainly due to its many applications and its technical challenges, since it tackles categorization problems that are difficult even for humans. Examples of FGVC applications include the categorization of leaves [108], flowers [136], dogs [120] and, more recently, birds [15]. As can be inferred, FGVC methods and approaches are extremely fitting for biological problems, specially those where taxonomy impose a set of mutually exclusive subcategories [15].

Additionally, FGVC and image annotation are deeply connected. This is due to the fact that most FGVC datasets and approaches work with different types of annotations and related metadata in order to extract as much information as possible from the images, which can help improve the performance of such difficult classification tasks. Some alternatives have been developed in order to eliminate the use of annotations or, alternatively, visual code-words, another popular approach applied to FGVC. An example of this is found in [210], in which the authors used a large number of random image templates instead in order to classify the unseen test samples. However, most of the state-of-the-art FGVC methods continue to use annotations as part of their framework due to their flexibility and the large amount of information they can provide [15, 58, 78]. For example CUB-200-2011, created by [199], is a dataset for birds with parts and attributes and Leeds Butterflies, created by [200], includes segmentations and text descriptions of butterflies.

Moreover, a methodology that has been successfully introduced in FGVC problems is the human-in-the-loop (HITL) approach [26]. Since FGVC problems are difficult for both human and computers, HITL methods aim to be an intermediate solution which

combines the strengths of both and to progressively minimise the amount of human labour [24]. HITL methodology can be easily applied to many different problems, such as criminology [140], port design [29] and aviation [171]. However, it is particularly suitable for FGVC problems. For example, [26] developed a HITL method for bird classification and [151] used HITL technology for skin-lesion image recognition.

In summary, automatic image annotation is a very broad topic whose research has been expanding and developing greatly during recent years. It has been regarded as a methodology whose purpose is to bridge the semantic gap often associated with content-based image classification. Moreover, it has obtained excellent results in many classifications problems [15, 112, 167]. In our case, and given the semantic similarities between the classes that we aim to categorise, a FGVC image-annotation approach seems the most appropriate option to apply.

2.2.4 Classification Methods

The term classifier belongs to Machine Learning, the discipline that studies the construction and behaviour of systems that can learn from data. In Computer Vision, classifiers are used to determine the most probable class of an unknown object. In our approach, we will use a classifier to annotate unseen ground-taken photographs. Given an unseen ground-taken photograph with a undetermined number of habitats present, our classifier's aim is to obtain a probability distribution or a histogram of all possible habitats in our unseen sample, sorted according to their probability of occurrence. Since we know all the possible habitats that are recorded in Phase 1, our problem is defined as a supervised classification problem [18]. Moreover, since our photographs can contain more than one habitat in them, we will be focusing on multi-label classifiers.

In the following sections, we will review some of the most popular classification methods used in the literature: Support Vector Machine (SVM), k-Nearest Neighbour (k-NN) and, finally, Random Forests. We aim to present some of the limitations that k-NN and SVMs have for the particular problem of automatic habitat classification and discuss how Random Forests can overcome these limitations.

2.2.4.1 Support Vector Machines

Support Vector Machine (SVM) is arguably one of the most used Machine Learning methods in Computer Vision. It is extremely popular due to several reasons: it has a straightforward geometric interpretation, a sound theoretical justification and, contrary to other methods, it is less likely to overfit. SVMs are parametric classifiers used

for supervised learning problems. Consequently, they can be applied to classification, regression and novelty detection problems [18].

The aim of the linear SVMs, the simplest form of SVMs, is to learn a hyperplane that can clearly separate the training data depending on the ground-truth or their labels. This separation is obtained by maximizing the margin between different classes. Additionally, by using a kernel approach, SVMs can work with non-linear data.

The traditional kernel-based SVMs use only one kernel matrix. Nevertheless, depending on the problem, this can be inadequate. While there are many types of kernel functions, these functions have many parameters which are difficult to tune. However, there are situations, for example if features need to be combined, in which more than one kernel is necessary.

From this necessity, Multiple Kernel Learning (MKL) methods were developed. MKL methods use a series of kernels and try to learn the optimal linear combination of them. Originally proposed in [110], it has spurred many modifications, such as [139, 206, 219]. Moreover, it has been used in many Computer Vision problems [194, 197]. A comprehensive comparison of several MKL methods can be found in [83], in which the authors found that there were not significant differences between these methods in terms of performance accuracy. MKLs are still being questioned in the literature. This is due to the implication that each kernel is fixed for all the samples in the training set [207], which can be considered a restrictive constraint. However, research on non-linear combinations of the different kernels also has been recently developed, particularly in [52, 207].

Despite their success, SVM methods, along with MKL, have several drawbacks. SVMs are inherently designed as two-class classifiers [18]. There have been several methods proposed to apply SVMs in multiple-label problems, such as our automatic habitat classification problem. One of the most popular approaches is referred to in the literature as the *one-versus-the-rest* approach and it was originally proposed in [193]. It consists in constructing as many SVMs as classes has the classification problem. The k^{th} model is trained with the data from class k as the positive examples and the data from the remaining classes as the negative example. However, this method can lead to inconsistent results if an input is assigned multiple classes at the same time [18]. Additionally, this division of the training set would be very imbalanced, since, given a class, the set with negative examples (that is, the set with all the positive examples for all the other classes) will generally be much larger than the set with positive examples. Moreover, it makes the assumption that the input only belongs to one class, which, for example, would not apply to our case, since the ground-taken photographs we will be working with can contain between 1 and 6 different habitats. Other modifications have been introduced, such as [113, 146, 202] but they involve a complicated training phase. This also results in

a significant increase in the training time and, also, in more computation requirements during testing [18].

2.2.4.2 Multi-label K-Nearest Neighbour

Another method that can be used in for classification is k-Nearest Neighbour (k-NN)[18]. Contrary to SVMs, k-NN is a non-parametric classifier. Used for classification and regression, k-NN is an instance-based learning method, where all computation is postponed until the classification, or the testing, phase. The k-NN algorithm is among the simplest of all machine learning algorithms but it has also obtained surprisingly accurate results [18]. The only requirement to use it is to store all the training samples and their labels at the same time in memory. When we want to classify a test sample, the only step that needs to be performed is to calculate the distances between the test sample and the k closest samples in the training set. The prediction, that is, the label to be assigned to the unseen testing sample, is the most popular label within the nearest k training samples, with k being a natural number.

As exemplified in [214], the k-NN classifier has been considered a baseline method, with a performance that cannot surpass that of discriminative classifiers, particularly SVMs. However, [21] challenged this notion by proving that NN-based classifiers could surpass SVM's performance for image classification tasks. [13, 129, 188] investigated and developed this idea further in their work.

In Computer Vision, k-NN has been successfully applied not only to image classification, but to image parsing [119, 178], scene completion [93] and even image annotation [86, 126]. Moreover, it has also been applied to Phase 1 habitat classification with aerial imagery in one of our works [180].

2.2.4.3 Limitations of Support Vector Machines and Multi-label K-Nearest Neighbour

K-NN presents some advantages over SVMs. First, its implementation is simpler and more straightforward. The training stage, which is complicated and time-consuming for SVMs, is practically non-existent for k-NN methods. This is extremely helpful when the size of the training set is large. Moreover, if the size of the training set were to change, which is very common in image annotation problems, when the databases used are constantly being updated, this would not affect the k-NN classifier. If the training set were to increase in size, the only step to carry out would be to include the new samples and if the training set were to decrease in size, we would only need to delete the

desired training samples. However, SVMs would require a new training phase with the updated training set. Additionally, k-NN classifiers can be used without any problems when the number of classes is very large. On the other hand, this presents a complicated challenge for SVMs classifiers. Moreover, k-NN methods can be easily modified to be used as a multi-label classifier, while, as mentioned previously, the same process for SVMs is much more complicated [18].

However, k-NN methods also present some drawbacks. First, all training samples must be allocated in memory at the same time. If the training set or its number of dimensions is large, this entails in large storage requirements. Nevertheless, the most obvious issue is the time complexity of the testing phase. If the size of the training sample is n in R^m, then the prediction time for one test sample will be $O(nm)$. Some efficient data structures have been created to accelerate the searching speed, for example kd-trees. However, these are only successful with low-dimensional data [60]. On the other hand, linear and kernel SVMs would require, respectively, a prediction time of order $O(m)$ or $O(cm)$, where c indicates the number of support vectors and it is commonly much smaller than n. In order to decrease the time complexity of k-NN methods, the Approximate Nearest Neighbour (ANN) approaches were developed. These include methods such as Locality-Sensitive Hashing (LSH) [141] and randomised kd-trees [133]. ANN is applied in problems where an approximate but faster guess is good enough than the actual correct, but also slower, prediction. The final drawback regarding k-NN methods, which also affects ANN methods, is related to the "semantic gap" problem. That is, just because the results retrieved are visually similar, this does not immediately guarantee the same semantic meaning. Or in other words, two samples that share similar visual or feature-related properties can belong to two completely different objects. This is consistent with the unsupervised nature of the k-NN algorithm and its ability to weight dimensions.

2.2.4.4 Random Forests

In the last few years, another machine learning method that has gained popularity is Random Forests. Random Forests, also known as Decision Forests, can be applied to both classification and regression problems [46]. They were first introduced in 1995 in [94], where the author applied it to handwritten digit recognition. However, it was in [28] where they were consolidated as powerful and accurate learning models.

Random Forests have been compared to other Machine Learning techniques and they have obtained successful results, as shown in [33]. Additionally, they have even been applied to a large number of computer vision problems, such as image classification [22, 127, 132], image labeling [106], action recognition [218], object detection [77] and

image annotation [75]. Moreover, they have also been used in Ecology problems, such as land-cover classification [81], urban trees mapping [147], habitat structure classification [11], groundwater-dependent vegetation pattern modeling [144], ecohydrological modeling [143] and land cover [81], genomic data analysis [37] and even age estimation [131]. More prominently, they also have been used in the Microsoft Kinect for XBox 360 [80, 168, 198]. Consequently, they have proven to be successful classifiers fit to be applied to a wide set of problems.

A Random Forest is composed of a set of independent decision trees. Each tree is trained separately on a random subset of the training data and the final prediction is obtained by combining the predictions of each of the independent decision trees. Therefore, Random Forests are, in essence, an ensemble method [157]. Another popular example of ensemble method is AdaBoost [72]. AdaBoost repeatedly calls a chosen weak learning algorithm, such as decision forests, a number of times [73]. Research has shown that using an ensemble of learners, also referred to as or weak classifiers, on unseen data can produce greater accuracy [157]. This is known as generalization [6].

A decision tree is a hierarchical structure composed of nodes and edges. Depending on their nature, nodes will have associated either a test function (internal nodes) or a prediction (leaf nodes). The most important aspect of the decision trees that compose the Random Forests is that each tree is randomly different from the other decision trees in the Random Forest. This leads to de-correlation between the predictions and improves generalization [46]. This randomness also helps with increasing the robustness of the model with regards to noisy data. Traditionally, randomness is introduced during training [28]. The two most widely-used methods are bagging and randomised node optimization [28, 94]. The former is popular because it yields greater training efficiency, while the latter is beneficial because it yields margin-maximization properties and because it uses all the training data to train each tree. However, these methods are not mutually exclusive and are often used together.

In a typical classification scenario, given a labelled training set, the aim is to learn a general mapping which associates previously unseen test data with their correct classes [46]. In this case, a decision tree in a decision forest will commonly be constructed following these steps: given $\{s_i\}_{i=1}^{N}$, a set of training samples, and $\{y_i\}_{i=1}^{N}$, its corresponding labels which belong to a classification C, the first step is to extract a set of features $\{\mathbf{F}_i\}_{i=1}^{N}$ from the training test. The samples that reach each internal node will go to the left child or the right child of the node depending on the results of the split function. The main aim is for the split function to be as discriminative and informative as possible. Traditionally, the Information Gain is used to divide the data. The Information Gain is

calculated as shown in 7.2. The split that maximises the information gain in the final distribution is the split chosen.

$$IG = \mathbf{H}(S) - \sum_{i \in \{1,2\}} \frac{|S^i|}{|S|} H(S^i) \tag{2.2}$$

With $\mathbf{H}(S)$ being the Shannon entropy, which is defined as:

$$\mathbf{H}(S) = -\sum_{c \in C} p(c) log(p(c)) \tag{2.3}$$

As one of the most significant parameters that define a Random Forest, research on possible split functions is vast: [22, 209] studied linear split functions, while [121] proposed semi-supervised splitting functions. However, no clear all-around solution for the problem of splitting data has been found to date. This is reasonable, since finding an accurate and informative splitting criterion is inherently data and problem dependent [28, 46].

As we mentioned previously, in a Random Forests, each decision tree will provide one prediction for the unseen test samples. That is, each tree in the Random Forest will cast a vote. The prediction given by the Random Forest as a whole is obtained by combining the independent predictions of the separate decision trees. The simplest combination, given N samples and a forest of size T, shown in 2.4, is the linear combination of all the predictions in the forest.

$$P(c) = \frac{1}{N} \sum_{t=1} N P^{T_t}(c) \tag{2.4}$$

However, this voting mechanism follows the assumption that all predictions are equally good. In other words, a linear combination reflects that all trees are equally accurate at labelling the unseen data. This is often not correct, as some trees might be better at classifying than others [152]. Research on voting mechanisms is not as developed as research for other modifications of Random Forests, such as optimal feature selection or split function generation. However, it is becoming increasingly popular. For example, [152] weights the predictions of each tree by using internal parameters to compare unseen samples with samples used during training. Those trees with more similar samples obtained a higher weight. [186] present a comprehensive classification of voting mechanisms and studies their genetical impact in the classification process.

In general, a Random Forest can be defined by a set of five different parameters [46]: the forest size, the maximum depth of the decision trees, the amount and type of randomness, the split function, the choice of features and the voting system.

A large portion of the research on Random Forests has focused on studying the effects that the modification of these parameters have in terms of accuracy and generalization. Notably, [47, 167] studied the effect of forest size in classification accuracy, while [168] studied the relationship between overfitting and the forests' depth. Additionally, [211] used stratified sampling to separate the features into strongly or weakly informative and combined them during the training phase. [63] used the proximity between leaf nodes, via a proximity matrix, to classify unseen examples. Furthermore, [89] created a framework for feature selection and [9] studied the characteristics of different importance measures used for feature selection in Random Forests with the goal of identifying the true predictor among a large number of candidate predictors.

However, these parameters are not the only elements that can be modified to improve Random Forests. It is possible to modify how the actual trees are constructed. Modifications to the construction method include the creation of Alternating Random Forests [160]. These forests are constructed by minimizing the losses by giving weight to the training samples. [153] created Rotation Forests which use Principal Component Analysis in random subsets of the training data before the training phase. [131] define Entangled Decision Forests, which are built breath-first according to a priority queue and [105] also use a breath-first approach to include contextual information during the training process. Finally, [16] created Dynamic Random Forests which, inspired by boosting algorithms, continuously resample the training data.

In this thesis, we will be working with Random Forests to automatically classify Phase 1 habitats. We believe that Random Forests are the best choice of classifiers given the supervised multi-label nature of our problem. Random Forests are a simple yet accurate and efficient alternative to other machine learning methods, such as k-NN and SVMs. They combine the best features of theses classifiers: accuracy and generalization (SVMs), and multi-class classification and simple implementation (k-NN). Moreover, the inherent hierarchical structure of the decision trees, similar to the hierarchical structure of the classification scheme we are using, and the discriminative power of the Random Forests can aid the decision-making process.

2.3 Summary

In this chapter we have reviewed significant literature in the two fields we are working in: Ecology and Computer Vision. From the Ecology perspective, we have reviewed some of the most popular habitat classifications schemes and we have introduced the classification scheme we will be using: Phase 1. Moreover, we have reviewed both manual and automatic habitat classification methods and we have described their limitations with regards to Phase 1 habitat classification. From the Computer Vision perspective, we have reviewed literature related to the main methods that are used in our framework: feature extraction, random projections, image annotation and supervised multi-label classifiers and we have discussed their drawbacks when applied to habitat classification.

In the next chapter we will describe in detail the characteristics of the Phase 1 habitat classification scheme in order to clarify some of the most important challenges of Phase 1 habitat classification.

Chapter 3

Phase 1 Habitat Classification

Habitat classification is the process of mapping an area following a determined habitat classification scheme. It is an essential ecological activity which provides crucial information about the wildlife of a site and its ecological properties. Moreover, it has many ecological applications such as landscape ecology, habitat monitoring and, more importantly, rare species conservation [102].

The aim of this chapter is to describe in detail why Phase 1 was the classification chosen, how it is organised and what are some of its main merits and limitations. As we have shown in the previous chapter, while many automatic approaches to habitat classification have been developed, no automatic approach has been proposed for the Phase 1 habitat classification scheme. This is mainly due to the level of detail necessary to distinguish between Phase 1 habitats. This level of detail cannot be found in remote-sensed imagery. In this thesis, we study the use of ground-taken photographs as the main source of information for Phase 1 classification.

This chapter is divided into four sections. Section 3.1 describes Phase 1 classification, previously introduced in Chapter 2, with more detail. Section 3.2 describes the merits and limitations of Phase 1 as a classification scheme. Section 3.3 describes how Phase 1 habitats can be divided from a Computer Vision approach. Finally, remarks and a brief summary are presented in Section 3.4.

3.1 Phase 1 Habitat Classification

The Phase 1 scheme was standardised by the Joint Nature Conservation Committee (JNCC) [102]. The first draft was produced in 1986 and the current version, which

was produced in 2010 and the version we will be using, is a revision of the 1986 draft [102]. It is specially designed for rapid wildlife mapping over rural and coastal areas in Great Britain. Therefore, it only collects information about the vegetation of a site. The information provided by the Phase 1 survey can be used to assist effective nature conservation, for example by highlighting areas in need of special protection, and by providing a clearly defined baseline for monitoring change. The information can also assist local authorities and planners in forming policy and strategy for the rural and coastal areas, and enable them to make well informed and speedy planning decisions. Importantly, it provides planning authorities with statistics that can be used to support the case for the conservation of threatened habitats, especially in work connected with planning appeals. All the guidelines, standards and definitions necessary to train ecologists are collected in [102], published by the JNCC.

In this thesis, we are using Phase 1 Habitat Classification for four main reasons.

- Phase 1 is one of the most widely-used schemes by ecologists all over the United Kingdom. Examples include [27, 32, 42, 161].
- Even though Phase 1 is extremely popular in the United Kingdom, there is no previous work on how to automate its classification process. This research gap presents an interesting opportunity to study how Computer Vision and Machine Learning methods can help to make the process easier and, ultimately, more accurate.
- It is the scheme used by The Ordnance Survey, who provided part of the ground-taken imagery that we are using and with whom we worked closely.
- It collects information about the types of habitats that can be found in particular in rural and coastal England, including boundaries. Contrary to other classification schemes, such as EUNIS, which included European habitats not present in the UK, all Phase 1 classes occur within Great Britain and Ireland.

As a classification, Phase 1 follows a strict hierarchical structure. Table 3.1 shows the first-level and second-level habitats of the classification.

Phase 1 is specially designed for rural and coastal areas of Great Britain, although it can also be applied to urban areas. The classification is composed of ten first-tier categories, shown in the previous chapter in Table 2.4. It has four levels and a total of 150 habitat types. Each habitat type is uniquely identified by an alphanumeric code and a colour or a pattern. The alphanumeric code also follows a hierarchical nomenclature: the first tier is identified by letters, from A to J, and the rest of the levels are identified

Table 3.1: Phase 1 Habitat Classification Classes. Two levels shown.

Habitat Class		
Code	First Tier	Second Tier
A	Woodland and Scrub	Woodland Scrub
B	Grassland and Marsh	Parkland Scattered Trees Neutral Grassland Calcareous Grassland Improved Grassland Poor Semi-Improved Grassland Recently-Felled Woodland Acid Grassland
C	Tall Herb and Fern	Bracken Ledges Other
D	Heathland	Dry Dwarf Shrub Heath Dry Dwarf Shrub Heath Lichen/bryophte heath Montane Heath/ Dwarf Herb Dry Heath/Acid Grassland Mosaic Wet Heath/Acid Grassland Mosaic
E	Mire	Bog Flush and Spring Fen Bare Peat
F	Swamp and Marginal Inundation	Swamp Marginal and Inundation
G	Open Water	Standing Water Runing Water
H	Coastal	Intertidal Saltmarsh
I	Rock Exposure and Waste	Natural Artifical
J	Miscellaneous	Cultivated/Disturbed Land Boundaries

by numbers, which are appended to their corresponding letter. For example, Neutral Grassland Unimproved is also identified by the colour orange and by the code B.2.1. In this case, the B indicates that the habitat belongs to the Grassland and Marsh category, the 2 indicates that it belongs to the Neutral Grasslands and the 1 indicates that it is Unimproved. [102] contains the full classification, with the alphanumeric codes.

3.2 Merits and Limitations of Phase 1

It is important to notice that election of Phase 1 as the classification scheme to be used in our framework entails several benefits and challenges. Phase 1 is, by nature, a very detailed classification scheme. Surveyors not only record the Phase 1 habitat classes that are present in a site, but are also encouraged to make target notes. These target notes specify interesting or out-of-ordinary information, such as particular vegetation species, percentages of appearance of different plants in complex habitats or relevant comments about the distribution and relationships between different habitats, etc. As can be inferred, this type of information provides a great deal of relevant information which can help end users gain a much deeper understanding of the ecological properties of a site. However, target notes are difficult to incorporate into a Machine Learning framework, such as ours, since these notes generally do not follow an specific layout nor are they present in all surveys. Therefore, there needs to be a trade off between a fast and efficient automatic classification and the amount of information that this classification will provide.

Furthermore, falling in line with the FGVC nature of the problem, some of the habitats recorded might be difficult to classify even for trained surveyors. Distinguishing between grasses (class B), particularly, can be extremely challenging, since data regarding the geological properties of the ground are needed, i.e. acid, neutral or calcareous ground results in acid (B.1), neutral (B.2) or calcareous (B.3) grass. These different grasses, with extremely different ecological properties, may look exactly the same to the untrained or unexperienced eye. To avoid this situation, surveyors might research information about the site they have to classify in advance. They will incorporate that knowledge, and their previous experience, to the classification process. Incorporating that previous experience and external information presents a challenge. In this thesis, we propose two modifications specifically aimed to include this type of contextual information albeit in another manner. These modifications are medium-level features and a location-based voting system. The medium-level features, described in Chapter 8 will incorporate semantic information about what humans perceive in the photos while the assignation

of weight to the predictions of the classifier according to their GPS location, shown in Chapter 9, will prioritise previously-collected information from the same area.

Additionally, Phase 1 is also very versatile and can be adjusted to be used depending on the conditions of the survey. Given the requirements or the aims of the survey, the needs of the end user and the surveyor's experience, it may not be necessary to use all four levels in the classifications process, and using two or three levels might be enough.

3.3 Phase 1 and Computer Vision

For the purposes of our research, we have divided the habitats recorded with Phase 1 in two ways depending on different characteristics. This is very useful to better understand the results obtained by our framework in Chapter 7, Chapter 8 and Chapter 9.

First, we can divide habitats into natural or artificial habitats. Natural habitats are habitats which are not composed of artificial or treated materials. They might be maintained by humans, but they were created naturally. Artificial habitats have been created by humans. Following this division, habitats such as Amenities (J.1.2 or canary yellow) or fences are artificial habitats, while habitats such as Sand Dune (H.1.6), Grassland and Marsh (B) and Standing Water (G.1) would be considered natural.

Another more helpful division separates habitats depending on their complexity. In this case, we regard habitats as either simple or complex. We define simple habitats as those habitats composed by vegetation belonging to one and only one of the ten first-tier classes. There may be more than one type of vegetation in these habitats, but all of them must belong to the same first-tier category. Moreover, there are no requirements about the layout that these habitats must have to follow or impositions about how they have to be used. On the other hand, complex habitats are composed by habitats from different first-tier classes or habitats which have to follow a particular layout. For example, Mixed Woodland (A.1.3) is a habitat composed by Coniferous (A.1.2) and Broad-leaved (A.1.1) woodland. Since both types of vegetation belong to the same class, A.1.3 would be defined as a simple habitat. However, a Dry Heath/Acid Grassland Mosaic (D.5), composed of Heath (class C) and Grassland (class B), would be considered a complex habitat. Hedges (J.2.3), in particular, present an interesting and challenging case. While they are composed by Woodland (A.1) and Scrub (A.2) habitats, both habitats from class A, Hedges are required to follow a very specific layout: they have to be arranged in a single row and they have to be used to separate two sites or two other habitats. Consequently, Hedges would be considered complex habitats. This last division will be extremely useful when studying the performance of our framework, since automatically

classifying complex habitats, instead of classifying each simple habitat separately, is easy for human surveyors but can be quite challenging for a computer. It is important to notice that these surveyors may incorrectly classify the habitats, but they can generally rapidly classify which habitats are composed of several types of vegetation, which is a arduous task for computers. How our automatic system classifies simple habitats versus complex habitats will offer insight into its accuracy and the usefulness of ground-taken photographs.

3.4 Concluding Remarks

In this chapter we have presented a detailed description of Phase 1 Habitat Classification, the classification scheme we will be working with in this thesis. Moreover, we have given four main reasons why the Phase 1 scheme was chosen to be used in our framework and we have discussed its merits and limitations as a classification. As a final note, from now on, in this thesis we will use the term habitat classification to refer to Phase 1 habitat classification specifically.

In the next chapter we will present a brief study on the limitations that remote-sensed imagery, particularly aerial photographs, when applied to the automatic classification Phase 1 habitats in more detail.

Chapter 4

Automatic Habitat Classification Using Aerial Imagery

In Chapter 2 we discussed the limitations of both remote sensed data and content-based image retrieval and classification approaches when automatically classifying habitats. The aim of this chapter is to study and discuss these limitations in the particular case of content-based automatic habitat classification using aerial photography. Moreover, we present specific results that help clarify the reason behind these limitations. For this, we will study the performance of aerial imagery and local invariant features when classifying and retrieving four of the habitats that appear more frequently in rural England: Woodland, Scrub, Grassland and Arable land.

Moreover, in order to obtain more information about the use of aerial imagery for Phase 1 classification, we have approached automatic habitat classification under two different scenarios: a classification scenario and a retrieval scenario. In the classification scenario, the objective is to correctly classify the query image using photos from a database. In the retrieval scenario, the objective is to retrieve the photographs from the same habitat as the query image. We evaluate the performance of aerial imagery in these two scenarios by calculating the recall of the system.

This chapter expands the work published by the author of this thesis in [180] and it is divided in five sections. Section 4.1 describes the types of data we have worked with and shows several visual examples. Section 4.2 describes the methodology followed for the automatic classification of Phase 1 habitats. As mentioned previously, we have approached this as both a retrieval and as a classification problem. Moreover, we have used a k-NN methodology in both scenarios. Section 4.3 shows the testing scenarios for

both the classification and the retrieval approaches. Finally, Section 4.4 shows the recall results for both scenarios and Section 4.5 presents a discussion on these results.

4.1 Data

We have used five different types of data. Examples of each are shown in Figure 4.1. These data include:

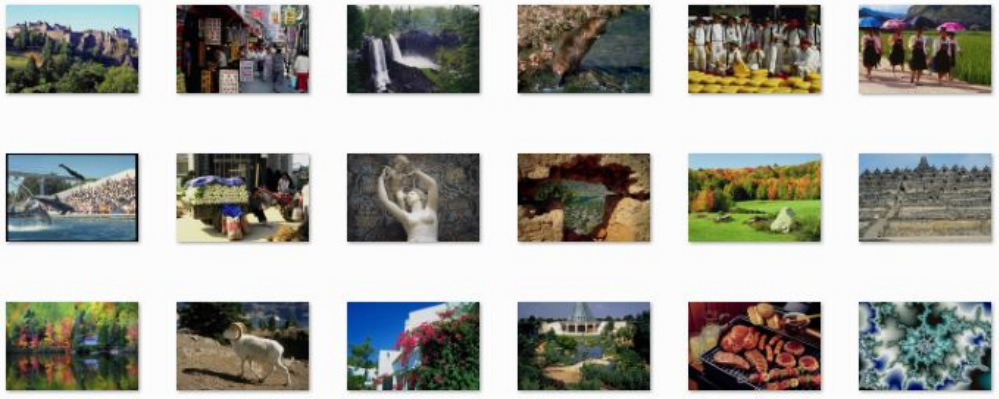
- Corel Dataset: We used 20,546 images from the Corel dataset to generate the codebook during training [50]. These photographs are extremely diverse and contain objects, scenery, patterns, people, animals, paintings and food. This variety guaranteed that robust codebook would be built. All of them are 256x384 pixels. Moreover, there are black and white and colour images.

- OS Master Map - Topography Layer: Master Map is a database that collects information about every fixed feature in Great Britain larger than a few metres [173]. It is one continuous digital map. The topography layer, in particular, represents topography at a scale of 1:1250. Moreover, it is subdivided into a number of themes: land area classifications' buildings, roads, tracks and paths, rail, water, terrain and height, heritage and antiquities, structures, and administrative boundaries. It is organised by polygons which represent the area on the ground that the feature covers, in National Grid coordinates

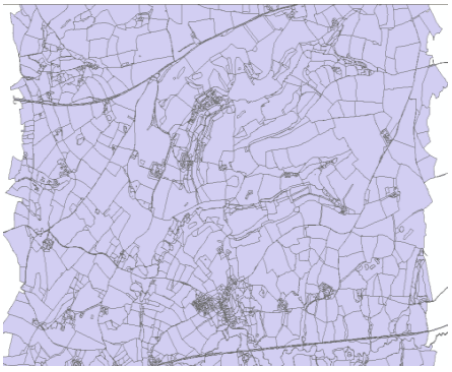
- OS Master Map - Imagery Layer: An aerial photograph composed by a variable number of plots with different lighting conditions [173]. These raster images are usually large in size and, consequently, difficult to manipulate. Therefore, using the whole image during testing would be time consuming and would not yield accurate results, since all the plots with the different habitats would be combined.

- Query set: Instead of using the whole raster image in the testing phase and then using a spatial extension in the retrieval process [208], we divided the raster image using the topography layer from OS MasterMap [173]. This process is referred to as "clipping" the raster images. The query set is composed by all the images obtained from clipping the imagery layer with the topography layer. The Phase 1 ground-truth associated with this data was classified by an expert.

- Test set: This set is a ground-truth catalogue of the Phase 1 habitats we are aiming to retrieve and classify. It was classified by an expert in Phase 1 Habitat Survey [102] and collected and organised by the author of this thesis. It is composed of 1072 images and it includes the following habitats: arable without crops (231

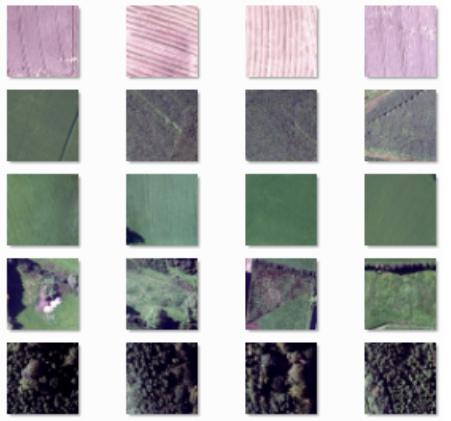


(a) Corel Database

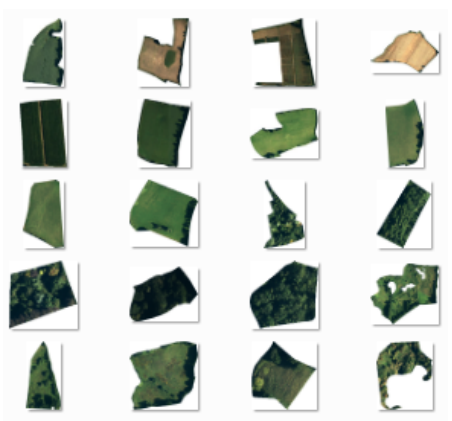


(b) Master Map

(c) Raster Image



(d) Query Set



(e) Test Set

FIGURE 4.1: Data Used In Our Content-Retrieval Approach. We use these four types of data in our content-retrieval system.

images), arable with crops (115), grassland (285), scrub (80) and woodland (361). All the images have the same dimensions, 456x456 pixels and, as shown in Figure 4.1 the lighting conditions were purposely chosen to be very diverse.

4.2 Methodology

The methodology that we have used to study the performance of aerial imagery is based on the famous work of [169], which used a bag-of-visual-words approach. In [169], visual words were extracted to describe video frames and to detect and retrieve objects under varying conditions. Moreover, it was also an extension on our past work, presented in [181], in which we followed a similar approach to classify and retrieve historical photographs from the same landmarks taken in different times.

As discussed in Chapter 2, the use of visual words is an extremely useful methodology for image classification which helped alleviate the problem of "the curse of the dimensionality". They are frequently used because they enable users to describe images using only a compact numerical vector. It does not matter how large or small these images are, they all are described by these numerical vectors, whose size is determined by the number of visual words users choose. Consequently, the complicated task of comparing two or more images is reduced to calculating the distances between their respective frequency vectors. To obtain these frequency vectors, a codebook, along with the visual words of each image are needed.

In relation to [169], we introduced three improvements aimed to increase efficiency and decrease the effect of high-dimensionality feature vectors. These improvements are:

1. Not extracting Maximally Stable (MS) regions: In [169], Sivic and Zisserman extract SIFT descriptors only from the MS regions within the images. In our case, we do not use MS regions. As a result, more SIFT descriptors are extracted from the images. This larger number of features is then used when creating the codebook, which makes it more detailed and robust.

2. Reduced number of visual words: Sivic and Zisserman used a 16,000-visual-word codebook [169]. In our case, we were able to obtain accurate results only using 100 visual words in our codebook. This makes the training process much faster and more efficient.

3. Frequency Sensitive Competitive Learning (FSCL): Instead of using traditional k-means, we use FSCL [148] when creating the codebook. Consequently, we avoid choosing local minima as the centroids.

Figure 4.2 shows the complete overview of the system. As can be seen, it can be divided into three phases: (a) preprocessing, (b) training and (c) testing. The training and preprocessing phase can be carried out off-line. Moreover, the training only needs to

be done once. Preprocessing, however, may have to be done several times if new raster data, or new locations, are introduced. Even though we are approaching automatic habitat classification from two different perspectives, retrieval and classification, the methodology followed is the same in both cases. The main difference appears during the testing phase, in which the goals are different, as shown in Section 4.2.1 and Section 4.2.2. These phases can be described as:

- Preprocessing: This first phase can be carried out concurrently with the training phase. The main aim of this phase is to create the query images. Therefore, the goal is to prepare the testing samples for the retrieval and classification process. During this phase, the Imagery Layer from the testing site is clipped with the Topography Layer, which contains a polygon for each different feature, or habitat, present in the raster image. By the end of this phase, the query set is completed.

- Training: in this case, since we are using a k-NN approach, the training phase is quite simple and straightforward. The main aim of the training phase is to extract relevant features and create a codebook. Given the varied nature of the aerial photography, we chose to extract Scale-Invariant-Feature-Transform (SIFT) descriptors [196]. These descriptors are suitable candidates to describe images because they detect lighting-, perspective-, orientation - and scale-invariant regions. After the features are extracted, a codebook is produced.

  A codebook is a glossary of the most descriptive visual words, called in this case code words. While [169] used a 16,000-code-word codebook, we chose a much smaller number in order to obtain a balance between resources needed and performance accuracy. Therefore, in our case, a 100-code-word codebook was calculated. For this, k-means clustering was applied to the Corel Database [50]. We chose the Corel Database for two main reasons:

  1. It is extremely varied. These photographs include scenery, patterns, people and objects. Moreover, there are black and white and colour photographs. Figure 4.1 shows a sample of the types of images used to generate the codebook. This diversity implies that the resulting code words will be very robust, descriptive and significant.
  2. It is completely independent of the testing images. Consequently, the same codebook could be used with different testing sets. It could even be used with other types of imagery, such as satellite and ground-taken photographs.

  Finally, instead of using the histogram of the images as the feature vectors, we use its inverse frequency vector. The inverse frequency vector describes each aerial

image and it is generated by measuring the frequency of appearance of the code words in relation to its own visual words [169]. By using the inverse frequency, visual words that appear less, and therefore are more descriptive of the different habitats, will have more weight when describing the images.

- Testing: The final phase is testing of the images. This phase is different depending on whether we are aiming to retrieve or classify the habitat images. For retrieving, the query image is known during testing and the aim is to retrieve as many instances of the same habitat from the test set as possible. For classifying, the aim is to make a prediction about the unknown test image, using a k-NN methodology.

4.2.1 Retrieval

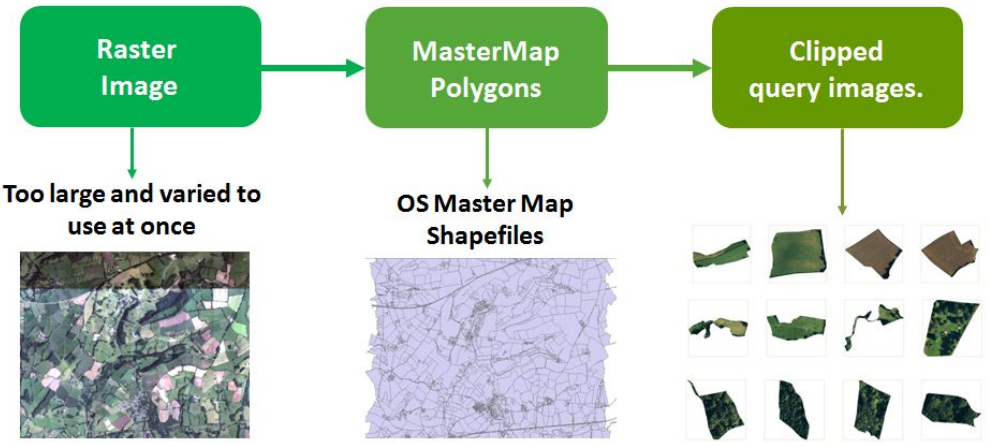
In this case, as shown in Figure 4.3, the habitat class of the query image is known during testing. The objective is to retrieve all the photos from the database (query set) that belong to the same category as the query image. This is done by calculating the Euclidean distance between the frequency vectors that describe the query image and the images in the test set. Once the distances are calculated, these are indexed from closest to further away.

4.2.2 Classification

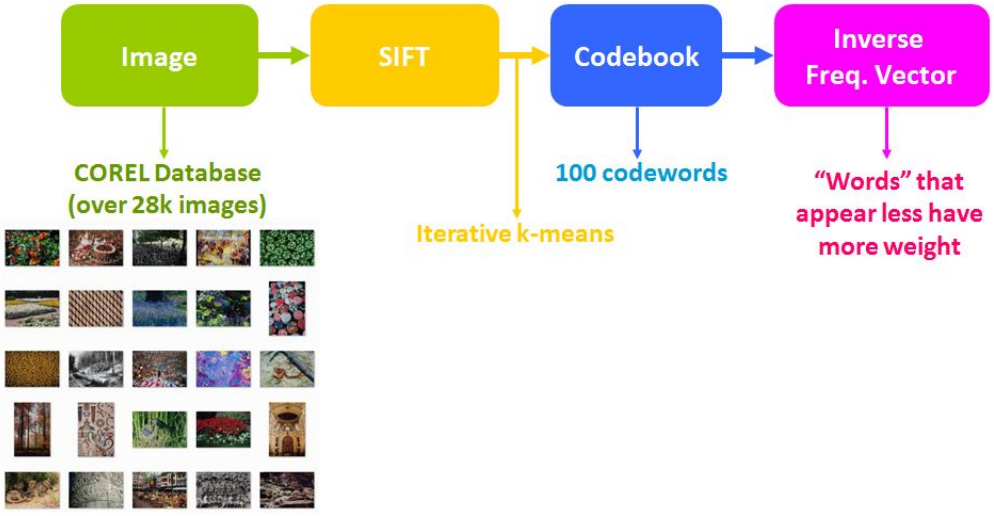
Following a classification approach, as shown in Figure 4.4, the habitat class of the testing (query) image is unknown. Consequently, the objective is to make a prediction about its class by using its closest images in the test set. k-NN is used to decide the class of the query image by averaging the k first results [44].

4.3 Experiments

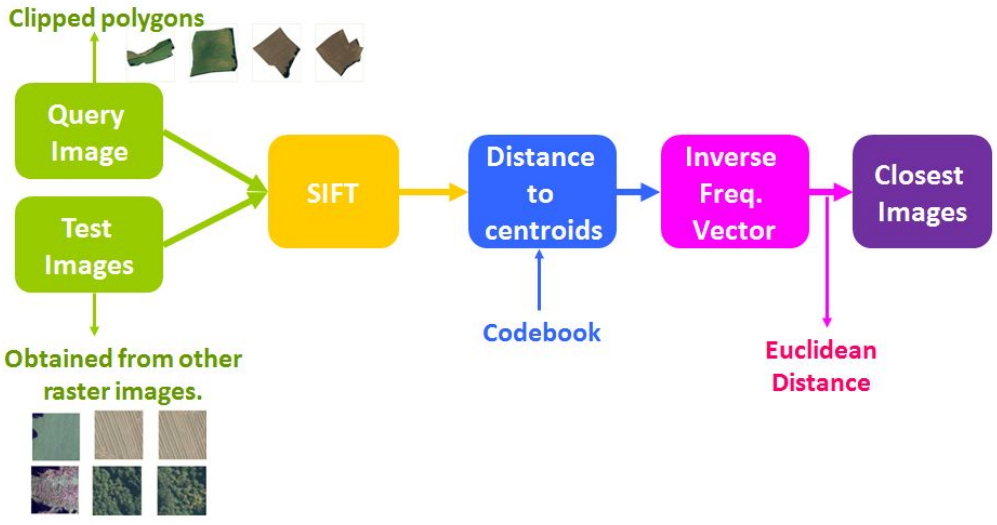
To test the two scenarios, imagery from two different locations was classified by an expert and used in our system. These locations are referred to as the query area and the test area. As their name indicates, the query area is used to generate the query set and the test area is used to generate the test set. Both areas belong to the Hampshire region in the United Kingdom. Figure 4.5 shows the two different areas on a map. Additionally, Table 4.1 shows the number of images corresponding to the four habitats retrieved and classified in both sites.



(a) Preprocessing



(b) Training



(c) Testing

FIGURE 4.2: Automatic Habitat Classification and Retrieval Using Aerial Imagery. Overview of the whole system.

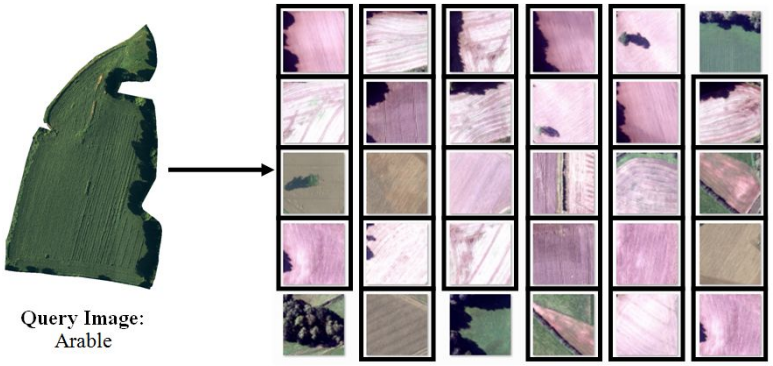


FIGURE 4.3: Retrieval Using Aerial Imagery. As shown, the class of the query image is known. The objective is to retrieve all instances of the same habitat in the test set.

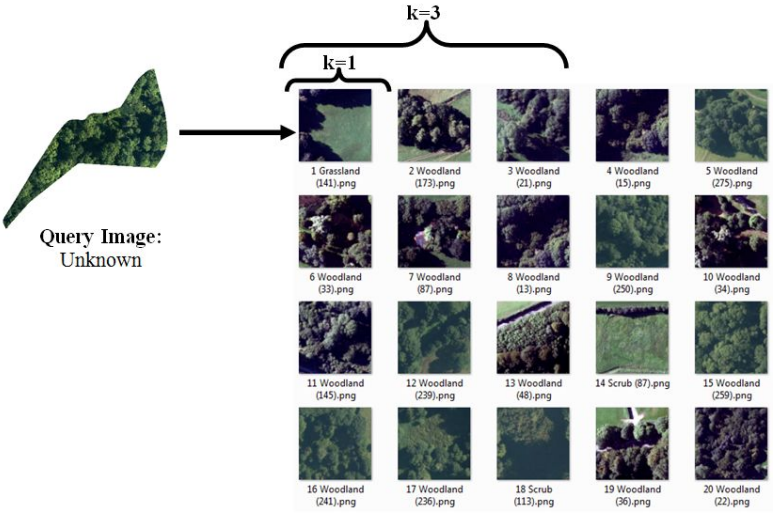


FIGURE 4.4: Classification Using Aerial Imagery. As shown, the class of the query image is unknown. The objective is to predict which habitat is present in the image.

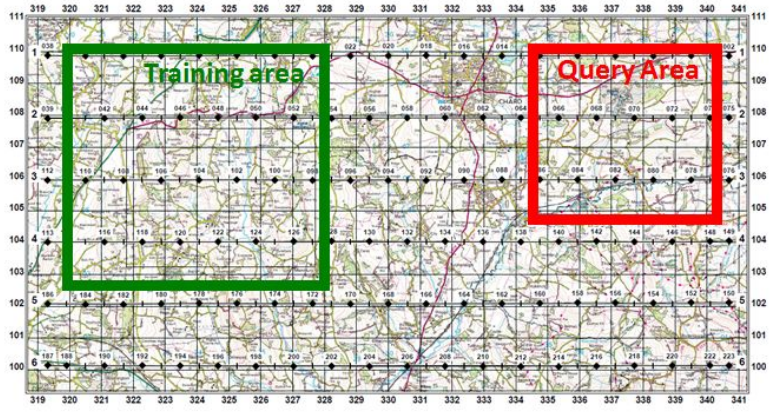


FIGURE 4.5: Training and Query Areas. Both areas are in the Hampshire county, in the UK.

TABLE 4.1: Training and Testing Set. Number of images of each habitat extracted from the query and the test area.

Habitat	Query Area	Test Area
Arable	68	346
Grassland	411	285
Scrub	12	80
Woodland	259	361

4.4 Results

There are diverse metrics that have been used to measure the performance of Computer Vision problems. In our case, we have used the recall metrics to assess the accuracy of the system [216].

Recall, also called sensitivity in the literature, is defined as the fraction of relevant instances that are retrieved. Following [216], the are calculated as follows: let N_h be the number of the images in the test set whose habitats are labeled by an expert, and N_c the number of images whose habitats our system correctly suggests. Recall is defined as shown in Equation 4.1:

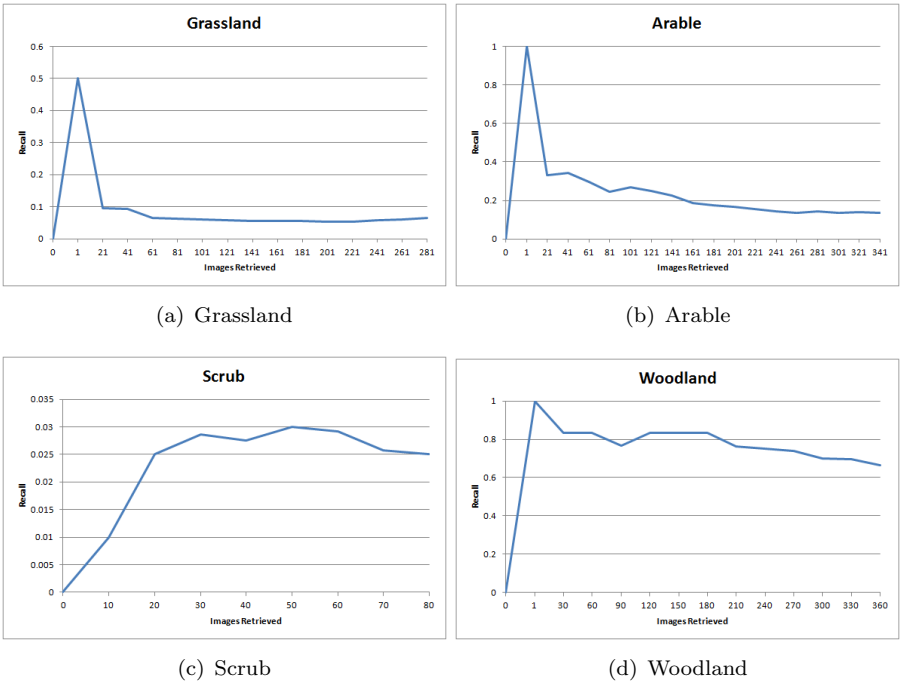
$$recall(w) = N_c/N_h \tag{4.1}$$

It is important to notice that this measure is often paired with the more strict metric precision. However, in our case, since we only aimed to get an understanding of the behaviour of aerial imagery and the effects of k-NN and SIFT features, we decided to use recall to evaluate the system.

4.4.1 Retrieval

The retrieval accuracy of the approach, shown in Figure 4.6, was measured by calculating its recall. An average of the number of correct answer retrieved was calculated by varying the number of retrieved images from one to the number of images of that habitat class in the test set. Figure 4.6 shows the results obtained.

Results show that as the number of results retrieved increases, the proportion of correctly retrieved photos decreases, which is consistent with the approach followed. Moreover, recall results concerning grassland and scrub are significantly low. This is mainly due to the fact that scrub and grassland habitats can have similar intensity properties and, consequently, the visual words extracted from the images can be similar. Therefore, using aerial imagery to distinguish between them can be a difficult task. This is not the only identification problem that aerial imagery entails. Figure 4.7 shows four different



(a) Grassland (b) Arable

(c) Scrub (d) Woodland

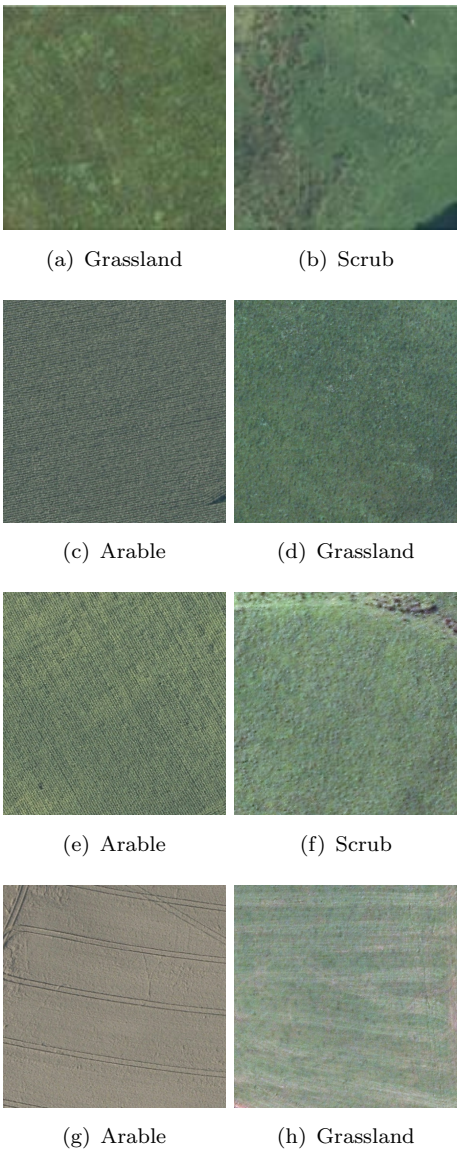
FIGURE 4.6: Recall for the retrieval of habitats Grassland, Arable, Scrub and Woodland.

cases in which distinguishing between habitats, even manually, is difficult. In each row, two images from the test set are shown. Even though the images belong to different habitat classes, their similar intensity and visual properties make their identification problematic, even for humans.

On the other hand, woodland intensity characteristics are very distinguishable from the other habitat classes. Consequently, its recall ability is high, over 65%, in all cases.

Additionally, we present five different sample results, shown in Figure 4.8, Figure 4.9, Figure 4.10, Figure 4.11 and Figure 4.12. In all cases we present the query image, which belongs to the query set, and the first five results obtained, which in turn belong to the test set.

As can be seen in Figure 4.8, the fact that our retrieval system only takes into consideration intensity level and not colour features to create the codebook makes possible the retrieval of arable land without crops when the query image is arable land with crops. Figure 4.9 is a serves to illustrate one of the limitations of aerial imagery discussed previously and shown in Figure 4.7. Similarities in intensity levels and visual properties make the distinction between grassland and scrub habitats a difficult task. Some of those limitations also affect the results shown in Figure 4.11 and in Figure 4.10, in which four out of the first five results are accurate. Finally, Figure 4.12 shows accurate results for Woodland retrieval, a direct consequence of the noticeably different visual



(a) Grassland (b) Scrub

(c) Arable (d) Grassland

(e) Arable (f) Scrub

(g) Arable (h) Grassland

FIGURE 4.7: Aerial Imagery Limitations. Habitats of each row have similar properties, which makes their classification difficult even for humans.

characteristics of woodland habitat samples in relation with the samples from the other habitats.

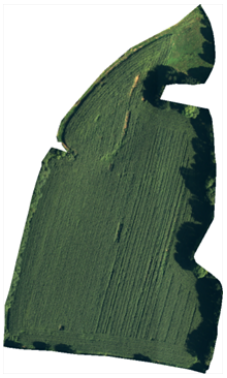
4.4.2 Classification

The classification accuracy of the method, shown in Table 4.2, was measured by applying k-NN and varying k, the number of neighbours taken into account when classifying the query image.

As can be seen, as k increases, the number of correctly classified images decreases. This is particularly noticeable in grassland habitats where the classification accuracy drops

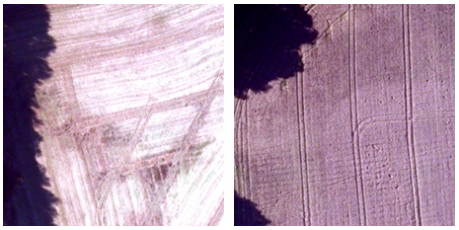
TABLE 4.2: Habitat classification using k-NN. Percentage of correctly classified images as k increases

Habitats	Values of k												
	1	3	5	7	9	11	13	15	17	19	21	23	25
Arable	10.98%	12.72%	11.56%	11.56%	10.12%	10.12%	10.40%	8.96%	8.67%	8.09%	8.67%	8.38%	8.09%
Grassland	57.19%	42.81%	8.07%	5.61%	5.61%	5.26%	5.96%	5.96%	6.32%	5.26%	5.26%	4.91%	5.26%
Scrub	5.00%	3.75%	6.25%	3.75%	3.75%	5.00%	5.00%	2.50%	2.50%	2.50%	3.75%	2.50%	3.75%
Woodland	18.84%	34.07%	38.78%	43.49%	45.43%	46.81%	46.26%	47.37%	47.65%	49.03%	50.42%	50.42%	50.69%



(a) **Query:** Arable

(b) Arable (c) Arable (d) Arable



(e) Arable (f) Arable

FIGURE 4.8: Retrieval Visual Example. The first five results retrieved by our framework are correct.

from 42.81% (122 correctly classified images) with $k = 3$ to 8.07% (23 correctly classified images) with $k = 5$. This is a consequence of intensity similarities between different habitats, particularly scrub and grassland, as previously discussed in Section 4.4.1. On the other hand, and in conjunction with the results obtained in the retrieval scenario, results related to woodland habitats, whose characteristics are more distinguishable, increase as k increases, achieving a 50.42% of correctly classified photos when looking at the first 25 results.

4.5 Discussion

From the results shown in Section 4.4, it can be appreciated that aerial imagery and content-based image retrieval approaches based on low-level visual features, such as visual words, can be applied to habitat classification. However, they have limitations

(a) **Query:** Grassland

(b) Grassland

(c) Scrub

(d) Scrub

(e) Woodland

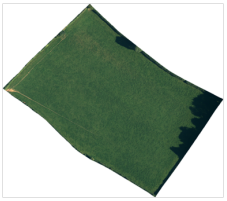
(f) Woodland

FIGURE 4.9: Retrieval Visual Example. Our system is unable to retrieve more than one correct result.

for both the retrieval and classification of Phase 1 habitats. The visual similarities between aerial images that represent different habitats, particularly grassland and scrub, as shown in Figure 4.7, present a problem when using remote-sensed data. Moreover, the limitations can be caused by similar visual properties are exacerbated by the fact that this content-retrieval framework only extracts low-level visual features. Therefore there is a large amount of information that is not used in the system, particularly semantic information, which can be crucial to distinguish between habitats.

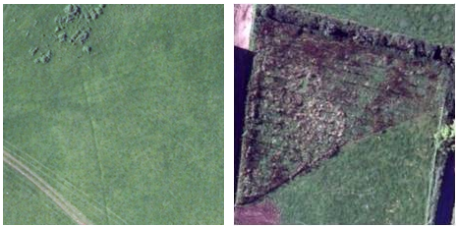
In essence, the system presented in this chapter offers a brief study on aerial imagery, local low-level features when applied to habitat classification and it can be seen as a starting point. Moreover, it can also be used to study traditional classification and retrieval methods, such as k-NN based approaches, and its limitations when automatically classifying habitats. As discussed in Chapter 2, NN-based methods, while useful for some classification tasks, have multiple limitations when applied to Fine-Grained Visual Categorization problems.

Particularly, if the aim is to extend k-NN to manage and work with large databases, as is our case, k-NN methods present three main technical challenges. NN-based methods



(a) **Query:** Grassland

(b) Grassland (c) Grassland (d) Grassland



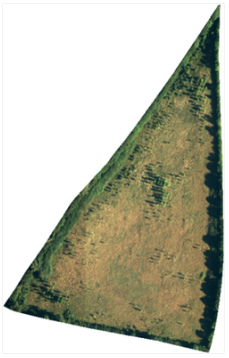
(e) Grassland (f) Scrub

FIGURE 4.10: Retrieval Visual Example. In this case, the system correctly retrieves four of the five first results.

require all training samples to be stored and available at the same during testing. Consequently, the first challenge is the design of efficient data structures that enable the storage of thousands, or even millions, of training samples. The second challenge comes from the necessity of retrieving the closest k neighbours during testing. As the value of k increases, this retrieval process will take more time. Finally, k-NN methods, specially when used only with low-level visual features, can aggravate the "semantic gap" problem [90]. This challenge comes from the fact that two objects from completely different classes can have similar visual properties and, therefore, be considered neighbours by NN-based methods' standards. On the other hand, random-forest based methods like the one we have developed in this thesis, do not present any of these issues.

4.6 Concluding Remarks

In this chapter, we have studied the use of remote sensing data, in particular aerial imagery, and content-based image retrieval and classification for the automatic classification of four Phase 1 habitats: Woodland, Grassland, Arable land and Scrub. Recall



(a) **Query:** Scrub

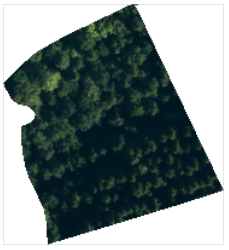
(b) Scrub (c) Woodland (d) Scrub

(e) Scrub (f) Scrub

FIGURE 4.11: Retrieval Visual Example. As shown, our system mistakes Woodland for Scrub in the second result.

results show that aerial imagery is insufficient to classify Phase 1 habitats, particularly in the case of distinguishing between Grassland and Scrub habitats.

In the next chapter, we will present a novel alternative framework to classify habitats based on automatic image annotation, feature extraction and ground-taken imagery. This will be the first main contribution of the thesis.



(a) **Query:** Woodland

(b) Woodland

(c) Woodland

(d) Woodland

(e) Woodland

(f) Woodland

FIGURE 4.12: Retrieval Visual Example. All results retrieved by our framework are correct.

Chapter 5

Automatic Image-Annotation Framework

5.1 Introduction

As shown in Chapter 4, remote-sensed data and content-based retrieval methodologies present some limitations when applied to automatic habitat classification. In this thesis, we present an alternative to this methodology by approaching automatic habitat classification as an image annotation problem. Moreover, instead of using remote-sensed imagery, which lacks the level of detail necessary to distinguish between some Phase 1 wildlife species, we use geo-referenced ground-taken imagery as the main source of data.

The aim of this chapter is to introduce the first contribution of this thesis: our automatic image-annotation framework for the classification of habitats using ground-taken photographs. This chapter is structured as follows: Section 5.3 describes in more detail how automatic image annotation works and presents an overview of the whole automatic image-annotation framework. Section 5.4 describes briefly the components of our framework. These components will be discussed and described in more detail in the following chapters. Finally, Section 5.5 presents a summary of the chapter and some concluding remarks.

5.2 Image Annotation: Methodology and Challenges

As previously discussed in Chapter 2, Automatic Image Annotation (AIA) is the process of automatically assigning metadata, such as keywords or labels, to a digital image.

Also referred to as Automatic Image Tagging [10] or Linguistic Indexing [115], AIA approaches have been gaining popularity in recent years due to the exponential increase in size that visual databases have experienced. A clear example of this can be found in Flickr, the image hosting website, which currently has over 6 billion photographs and over 1.3 million daily uploads of annotated public photos [125]. As can be inferred, searching databases of this size in an efficient and accurate manner is an extremely difficult task. AIA methods have been used traditionally as image-retrieval tools to organise and search images from a visual database using either visual features or, more efficiently, keywords [217]. An example of annotation-aided image retrieval using keywords is shown in Figure 5.1. Figure 5.1 shows an screen capture of how the database Geograph [154], which has almost 4 million photographs, can be efficiently browsed or searched using the annotations that users have created along with the photographs they have uploaded.

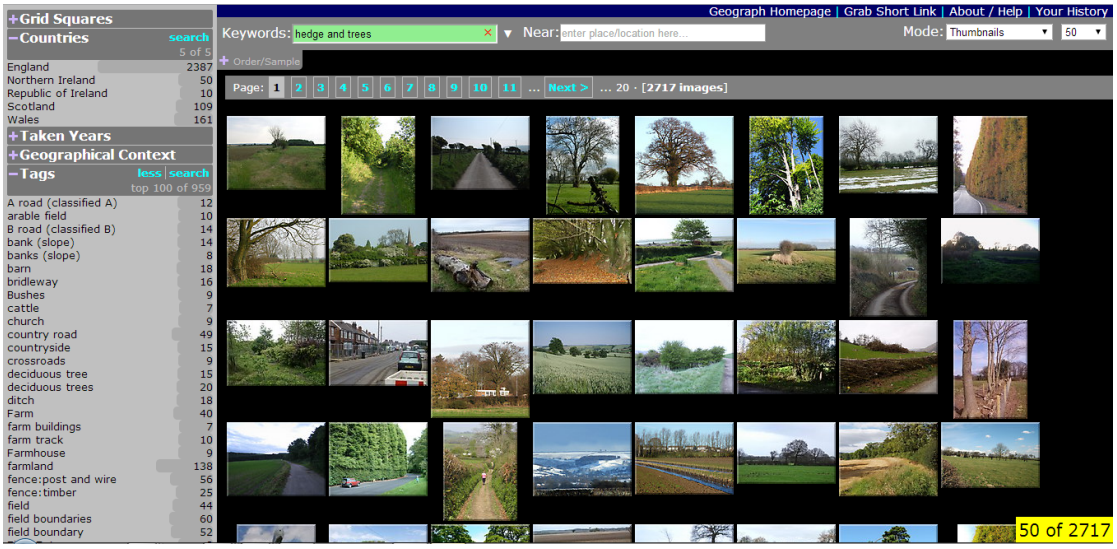


FIGURE 5.1: Geograph Search-By-Keyword Functionality. Photographs in the Geograph database can be searched using a combination of keywords.

AIA methods have obtained very successful results in retrieval tasks. However, AIA can also be applied to other Computer Vision problems. Particularly, it has been applied with much success to image classification [15, 108, 120]. In this case, AIA is regarded as a multi-class image classification problem in which the number of classes and the number of samples are relatively large [136]. In this thesis we follow this idea and we consider AIA a multi-class image classification problem in which the classes, or annotations, correspond to the habitat classes in the Phase 1 classification scheme.

AIA methodology applied to image classification follows a similar structure as other Computer Vision approaches, such as face or object recognition [203] or natural scene recognition [112]. An overview of this process is shown in Figure 5.2. First, an image

database is chosen or, as is our case, created. This database may be fully [183] or partially [216] annotated. These annotations serve as the ground-truth in the classification process. Once the database is chosen, image analysis is carried out on the images and feature vectors are extracted. Then, a machine learning technique is used to train a classifier. In our case, since we know all the categories present or possible annotations, our classifier is a supervised classifier. The chosen machine learning classifier has two inputs: the features extracted in the previous step and the annotations in the database that serve as the ground-truth. In essence, the main goal is to train a classifier that, using the annotations and visual information in our database, will automatically and correctly annotate new unseen images. Consequently, these approaches can be regarded as methods that learn the correlations between certain image features and certain words or annotations [104].

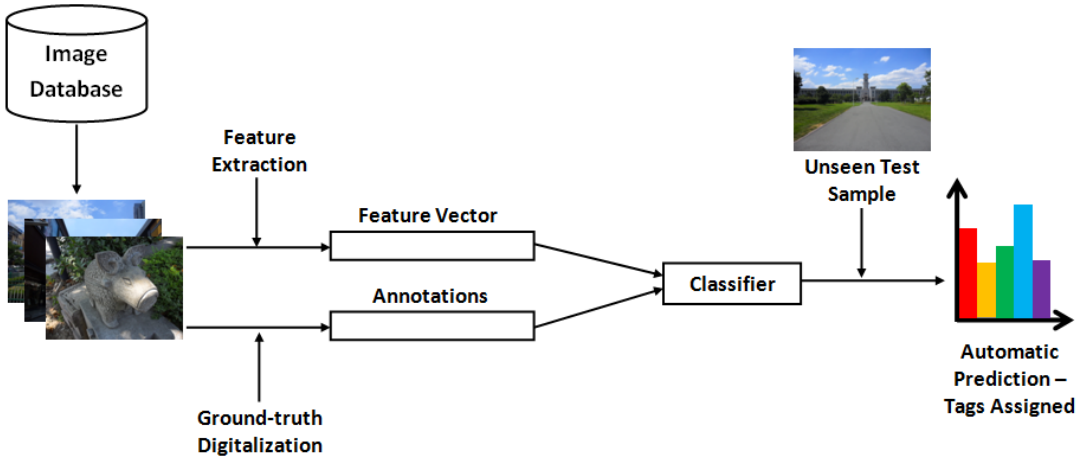


FIGURE 5.2: Overview of AIA as Image Classification. The common steps followed to be able to automatically annotate and classify images are shown.

From a Computer Vision perspective, AIA presents a series of interesting challenges. Notably, acquiring the appropriate ground-truth can be difficult and, most of all, time consuming. To be used in conjunction with AIA methods, visual databases are required to store not only pertinent images but also their corresponding annotation information. Therefore, space needs to be allocated to store the additional metadata contained in the annotations. Moreover, depending on the size of the database, the number of classes, the number of annotators and the annotation process of the ground-truth, the collection and organization of the ground-truth can be time consuming. However, in contrast to manually classifying images, or habitats in our case, it will only be needed to be done once before training the classifier, not every time a survey of a site is needed.

Nevertheless, the most interesting challenges involved in AIA are related to the variable nature of annotations and the dataset that is used. The characteristics of these

annotations will determine not only the type of problem to solve, but also its nature, i.e. supervised or unsupervised, and even the type of classifier that can be used, i.e. single-label or multi-label. It is important to notice that the annotation process can vary extensively depending on the task, the type of classification chosen and the manner in which the database was collected. For example, as commented previously, the annotations might have been added by one person or multiple people. If only one person is responsible for annotating all the images in the dataset, completing the task can take a significant amount of time. Nevertheless, the classification will be more consistent, since only one point of view will be reflected. On the other hand, employing several people, or even using crowd-sourcing methods, like websites such as Geograph [154] do, dramatically decreases the time needed for the manual annotation process. However, the larger the group of people responsible for annotating the images and the larger the dataset, the more difficult it will be to assess the quality of the annotations present in the database and consequently, the accuracy of the ground-truth. In our case, in order to have a more consistent classification, we have employed only one person, the author of this thesis, in the ground-truth annotating process for one of the databases, Habitat 1K. For the other database, Habitat 3K, we have used a crowd-sourcing mechanism, which was then refined by the author of this thesis.

Additionally, the degree of completeness of the annotations will determine whether or not the classification of images from the database is a supervised, in which case all data will be completely annotated, or a semi-supervised problem, in which case some of the data might be unlabelled [18]. An example of the first case is our dataset Habitat 3K [182], which is completely annotated with the pre-determined vocabulary given by the 150 Phase 1 habitat classes. [136] also follows this type of supervised approach, with either 17 or 102 flower classes taken into consideration in the classification process. An example of the second case is presented when trying to classify images from the dataset collected with the popular tool LabelMe[155]. LabelMe allows free annotations and, as a consequence, the database can never be considered completely annotated. Free annotations make the classification process a particularly challenging task, since objects of the same category can be annotated with different labels, such as synonyms or plural and singular labels. An example of this is shown in Figure 5.1, in which it can be seen that the tags “deciduous tree” (singular) and “deciduous trees” (plural) are both present when annotating photographs in Geograph [154]. While this will not prove a problem for humans, who are capable of recognizing that, for example, “automobile” and “car” represent the same concept, as do “cat” and “cats”, training a machine to learn this can be difficult.

Furthermore, the images in the dataset may have a fixed number of annotations or a variable number of annotations. Following the examples presented above, the dataset

collected by [136] belongs to the first case, since all the photographs contain one and only one annotation regarding the type of flower present in the photograph. On the other hand, our dataset Habitat 3K [183], as well as the LabelMe dataset [155], the Geograph dataset [154] and the dataset used for the popular Pascal Challenge [65], belong to the second category. The latter case is a much more challenging classification task, since it is impossible to know during testing how many results or predictions need to be taken into account before presenting the results. Solutions for this include choosing a fixed number, for example the average of the annotations [182], or even establishing a threshold on the probabilities of the predictions, so that the only predictions that are returned are the predictions whose probabilities of occurrence is larger than the threshold.

Moreover, the number of annotations per image will also determine if the classifier needed will be a single-label classifier, such as traditional Support Vector Machines (SVMs), or a multi-label classifier, such as the Random Projection Forest classifier presented in this thesis. Whether the problem is a single-label or a multi-label task will directly inform the classifier choice in the AIA approach, since, as it was discussed in Chapter 2, there are classifiers which are difficult to expand to include multiple labels, for example SVMs, and classifiers that are easily transformed into multi-label classifiers, such as NN-based methods and Random Forests [18].

Finally, annotations can be localised within the images or they can be global, as shown in Figure 5.3. For example the LableMe dataset belongs to the first category, while the dataset presented in [136] and the Geograph database [154] belong to the second. Moreover, the location of the annotations can be recorded in different ways. For example, the dataset created for the Pascal Challenge [65], located objects using the smallest bounding box around the object while in Habitat 1K and Habitat 3K, the datasets created in this thesis, the annotations are localised using polygons. The use of polygons gives more flexibility and accuracy when extracting local features. However, the annotating process is more time consuming. In our case, we have studied the effect of both scenarios, as shown in Chapter 7, by extracting global features from the whole photographs and from each annotation polygon separately and evaluating them with our framework in both cases. However, results, as it will be presented in Chapter 7 showed that, for the case of pattern features, the use of polygons as input did not dramatically impact the performance of the classifier, but it did hinder its efficiency.

5.3 Image Annotation Framework

In this thesis, we have followed an AIA methodology and we have created an image-annotation framework that can be applied to the automatic classification of habitats

(a) Sky (blue), Hedge (yellow), SI Grassland (red)

(b) Woodland, Hedges, SI Grassland

FIGURE 5.3: Localised and Global Annotations. (a) Shows an image with localised annotations and (b) shows a photograph with global annotations. Both images belong to our Habitat 3K database. SI stands for Semi-Improved.

using ground-taken imagery. The image-annotation framework and its main components are shown in Figure 5.4. As can be seen, our framework is composed of five main elements: the source data, feature extraction of low-level and medium-level features, the classifier and a weighed voting system.

In essence, our approach can be regarded as a method that takes into consideration "closeness" between photographs during the classification process. That is, during training, we take into consideration *visual closeness* by extracting significant low-level and medium-level features. Then, during testing, we take into consideration *geographical closeness* to assign weight to the predictions offered by each decision tree in the Random Projection Forest.

As with the vast majority of Machine Learning classifiers, the classification process is divided into two phases: training and testing. In our framework, these can be described as:

- **Training:** First, significant features are extracted from the ground-taken photographs in the training set. They can be low-level visual features or a combination of low-level visual features and medium-level knowledge. These features, in combination with the annotations (the ground-truth data), are used as the training input of our classifier, Random Projection Forests. At the end of this phase, a Random Projection Forest has been trained and it is prepared to annotate unseen ground-taken photographs.

- **Testing:** Similarly to the training phase, significant features are extracted from the testing subset in our ground-taken photograph database. These are injected in the root node of our classifier and propagated through the internal nodes of all the decision trees in our Random Projection Forest. Each tree in the forest will provide

a prediction about the classes present in the testing photographs. A prediction takes the form of a list of all the possible Phase 1 habitats sorted according to their corresponding probability of appearance in the photo. If the geographical location of the test photograph is not used, all predictions will have the same weight in the final prediction and be linearly combined. That is, each tree will cast a unit vote for the final classification. However, if the geographical location of the test photograph is used, each prediction will be weighted depending on the distance between the samples in the leaf nodes and unseen the test photograph. The final classification will be obtained by linearly combining these weighted predictions. At the end of this phase, a prediction in the form of a unique list of all the habitats, from most probable to least probable, is produced.

5.4 Components

As shown in Section 5.3, our image-annotation framework for automatic habitat classification is composed of: source data (ground-taken photographs), the features extracted from this data (low-level and medium-level features), a classifier which uses these features (Random Projection Forests) and a location-based voting system for predictions (calculated according to the GPS location of the images). In this section, we will briefly introduce all of them. Each element will be further described in the following chapters.

5.4.1 Source data: Ground-taken Imagery Annotated Database

The first element of our framework are annotated ground-taken photographs. They constitute the second contribution of this thesis. Ground-taken photographs offer two main advantages over remote-sensed imagery. These are:

- **Easier Collection:** Ground-taken photographs are easier and cheaper to obtain. There is no special equipment required, such as special cameras or access to satellites or planes. Ground-taken photographs can be obtained by using a digital camera and visiting a site of interest or by using crowd-sourcing mechanisms, such as websites like Flickr [125] or, as in our case, Geograph [154]. The first option offers more control over the characteristics of the photographs. However, habitats will be limited to the sites that can be visited by the collectors or the users. For example, if users were located in Nottingham, obtaining photographs from coastland habitats might be a challenge. On the other hand, the second option offers a wider array of possible habitats to take into consideration. Users only need to

search for coastland tag in the Geograph database. Nevertheless, as it will be discussed in later chapters, using third-party photographs implies a lack of control over the conditions under which the photographs were taken and the quality of the ground-data.

- **Finer Level of Detail:** Ground-taken photographs can provide more detail. This is extremely useful when classifying second or even third-tier habitats. In Chapter 4 it was shown how aerial imagery is insufficient to distinguish between Grasses or Scrub. Moreover, satellite imagery can include clouds or their shadow, which would make the classification process even more challenging. Additionally, in both cases, the layout of the images is always the same, the camera always being orthogonal to the ground. This lack of variation, while useful to make the source data uniform, can negatively affect the classification of finer habitats. However, ground-taken imagery can include photographs from different types of habitats under many different conditions.

In this thesis, we work with two different datasets: Habitat 1K and Habitat 3K. Figure 5.5 shows four examples of the photographs we are using as our main source data.

Habitat 1K contains 1,086 images and over 4,000 habitat annotations. The database was ground-truthed by a Phase 1 expert and annotated by the author of this thesis. The photographs have a resolution of 3648x2736 pixels. They were taken during the months of February, June and July in the Hampshire county, in England by research staff from The Ordnance Survey and by the author of this thesis. All photographs are geo-referenced. In this dataset, the lighting and perspective conditions, while diverse, are more controlled. Additionally, given their geographical location, habitats from classes A (Woodland and Scrub) and B (Grassland and Marsh) appear more frequently.

Habitat 3K contains 3,094 ground-taken photographs and over 11,000 habitat annotations. Habitat 3K contains all the photographs from Habitat 1K and an additional 2,000 photographs obtained from Geograph [154]. It was annotated by the author of this thesis using ground-truth data from the Geograph tag system and the ground-truth obtained previously. Similarly to Habitat 1K, all the photographs are geo-referenced. The aim of creating Habitat 3K was twofold. First, we wanted to include new habitat types, specially those which were difficult to reach given our location. Consequently, Habitat 3K has many more habitats present, such as Rock Exposure and Waste habitats or Coastland habitats. Secondly, we wanted to increase the size of our database and to introduce more variation in the habitats already present in our original database. In comparison with Habitat 1K, the lighting and perspective conditions of this dataset are much more varied, a clear consequence of using third-party photographs. Moreover, the photographs

are much more scattered through England. Instead of having many photographs of few different sites, as it was the case of Habitat 1K, we have a few photographs from many different sites.

The characteristics of Habitat 1K and Habitat 3K are described in more detail in Chapter 6.

5.4.2 Feature Extraction: Low-Level and Medium-Level Features

Feature extraction is an extremely popular Computer Vision approach which is specially used in Image Processing problems to work with large amounts of images in an efficient manner. Since all the images in the database, no matter how similar or different their characteristics, are described using the same parameters, feature extraction also serves as a homogenization process. Moreover, it can be seen as a method of dimensionality reduction, which helps the "Curse of dimensionality" [109].

The main aim of extracting features is to collect the most descriptive but compact information from an image. It is important to notice that the selection of features is an extremely decisive task. However, it is also highly problem-dependent [56]. Different types of problems will call for different types of features and extracting and combining a vast number of diverse features will not necessarily yield better accuracy than extracting a small but representative number of features, as will be demonstrated in Chapter 7 and in Chapter 8. For example, low-level shape features will be specially suited for tasks such as face recognition [203], while colour features might be more suited for problems such as bird classification [25]. Therefore, the aim in extracting features is to find a balance between the dimensionality of the features extracted and the quality of the information collected.

In order to work more efficiently with the ground-taken photographs, we extract low-level features from them. Moreover, we have created a new type of feature, referred to as Medium-Level Features, with the aim to extract more relevant information from the images. Consequently, the second element of our framework are the features we extract from our annotated ground-taken database.

5.4.2.1 Low Level Feature Extraction

Low-level features collect local or global statistics about different aspects of an image. Low-level visual features are one of the most popular types of features commonly extracted in Image Processing problems. Extracting low-level visual features enables us

to work with a large number of high-definition photographs in an efficient and accurate manner. Moreover, they also allow for an easier comparison between images with different characteristics.

Commonly, low-level visual features can be divided into, at least, three groups [109]: colour features, such as colour histograms, texture features, such as the Tamura coefficients [175], and shape features, such as the Hough transform [98]. However, there is a large number of other features which extract other types of relevant information, such as pattern features [148].

As mentioned in the previous section, feature selection is dependent of the problem to solve. In our case, since we are aiming to classify different types of natural habitats, we will focus on extracting colour, texture and pattern features. This is due to the fact that examining colour, texture and pattern similarities between habitats is similar to the process followed by trained ecologists when surveying a site. In particular, we have used pattern features [148] as a guideline for the behaviour of our classifier under different testing scenarios. We chose to do this because the pattern features we extract, called Colour Pattern Appearance Model (CPAM) features, have two main advantages over colour and texture features: they are more compact, with only a 128-dimension feature vector, and, at the same time, they collect a large amount of information on both the colour and pattern texture of the images. Moreover, they have obtained successful results in image classification tasks [148, 151].

Low-level visual features are one of the components of the ground-taken photograph databases we have created as part of this thesis. Consequently, low-level feature extraction will be described in more detail in Chapter 6.

5.4.2.2 Medium Level Feature Extraction

While low-level features have been proven to be effective for image classification and image annotation tasks [169], they have some limitations with regards to the type of information they can effectively extract. In particular, low-level features are not suitable for the extraction of higher level or semantic information which can be crucial when classifying FGVC problems. This entails that objects that are easily identifiable to humans, might be complicated for computers to differentiate due to their similar visual properties. This is normally referred to as the "semantic gap" problem [18]. For example, a human can easily differentiate between a water habitat (class G) and the sky. However, given their similar colour, texture and pattern properties, it might be more difficult for a computer to classify both correctly, as will be shown in Chapter 7. Semantic features

were developed as a medium to bridge the semantic gap and to include higher level information in the decision-making process.

In our case, semantic features can be very useful when automatically classifying habitats. In order to include higher-level information, we create and extract a second type of feature: medium-level features, which are the third contribution of this thesis. We also refer to them as medium-level knowledge. We follow the method described in [151] to incorporate medium-level information in the classification process using a Human-in-the-Loop approach.

To collect this medium-level knowledge, users were shown photographs from the Habitat 1K or Habitat 3K dataset and they were asked twenty three yes-or-no questions about the different types of natural objects that they can identify within the images. These natural objects included: trees with leaves, trees without leaves, trees with and without leaves, bushes, grass with flowers or non-uniform grass, uniform grass, reed, fern, herbs, heath, water, crops, boundaries, walls, fences, the sky, other (i.e. cars, people, buildings, animals). Along with the answer to each question, users are asked to measure the degree of confidence they have on their own assessment, which ranged from 0(not sure at all) to 5 (completely sure).

Medium-level knowledge and medium-level features will be described in full detail in Chapter 8.

5.4.3 Machine Learning Classifier: Random Projection Forests

As discussed in Chapter 2, Random Forests (RFs) are ensemble classifiers. RFs are increasingly popular in Computer Vision due to their simple implementation and accurate results. In our case, we have decided to work with Random Forests because their characteristics fit perfectly with our problem, automatic habitat classification.

Random Forests combine all the benefits that NN-based methods and SVMs entail without being critically affected by their more significant limitations. Similarly to NN-based methods, Random Forests' parameters are easy to tune and simple to implement. As with SVMs, they are efficient. However, contrary to these methods, Random Forests do not require complicated computation producers, like SVMs, or large storage of space in memory, like NN-based methods, to be applied. Moreover, they can be easily modified to be used on multi-label image annotation problems and to include semantic data.

Moreover, in order to improve some of the efficiency issues of RFs, we have created a new type of RF: Random Projection Forests (RPFs). These RPF constitute the third contribution of this thesis. They combine traditional RF and Random Projections,

introduced in Chapter 2 as a widely successful dimensionality-reduction mechanism. In RPF, at each split node, we will project the feature vectors that have reached that node, reducing them to only a scalar value. Results, shown in Chapter 7 will show that our novel approach is not only more efficient that traditional RFs, but also more accurate, particularly in the case of second- and third-tier habitats.

Random Projection forests will be described in more detail in Chapter 7.

5.4.4 Location- Based Voting System

Traditionally in Random Forest, each tree in the ensemble casts a unit vote, also referred to as a prediction, on the classes present in the unseen test image. These unit votes are commonly linearly combined to create the final prediction. This voting system method assumes that all trees in the ensemble are equally accurate classifiers. However, literature has shown that not all trees in a random forest are equally good at classifying unseen samples images [152].

In our case, we take advantage of the geographical properties of habitats to determine which trees might be more accurate in the classification process. Geographically close areas have similar ecological characteristics, since habitat properties do not generally change abruptly. For example, ground properties will not change from calcareous (class B.2) to neutral (class B.3) suddenly. Therefore, near regions will have similar habitats. Moreover, even if abrupt changes in habitat types were to occur, for example the sudden change between an inland cliff (class I.1.1) and acid grassland (class B.1) at the bottom of the cliff, a robust annotated database, such as the Habitat 1K database we have created, would have enough geo-referenced photographs of the site to accurately reflect that particular combination of habitats.

Since all the images in the database are geo-referenced, we benefit from this premise and we use their GPS coordinates to assign weight to the predictions. Weights are assigned according to the distance between the test sample and the images that are in the leaf node the sample has reached. By minimizing the distance and assigning weight, the predictions of trees with closer leaves influence the final classification more. However, it is important to notice that while some trees' predictions will weight more than others, all predictions are taken into consideration in our framework. Therefore, the final contribution of this thesis is a novel voting system which takes into consideration the geographical location of the photographs during the testing phase.

The inclusion of geographical location in the testing phase will be described in more detail in Chapter 9.

5.5 Concluding Remarks

In this chapter we have described what Automatic Image Annotation is and how it can be successfully applied to automatic habitat classification. Additionally, we have discussed some of the main challenges that AIA methods present. Moreover, we have presented our first contribution: an image-annotation framework for the automatic classification of habitats. We have given an overview of the whole framework and we have briefly introduced its components: its source data, feature extraction, the Machine Learning classifier used to annotate the photographs and the weighted voting system.

In the following chapters, we will describe in detail how each element of the framework works and how it relates to the other components of our system. The next chapter will describe the first element of our framework: the ground-taken photograph database, which is the first fully annotated database created for the classification of habitats.

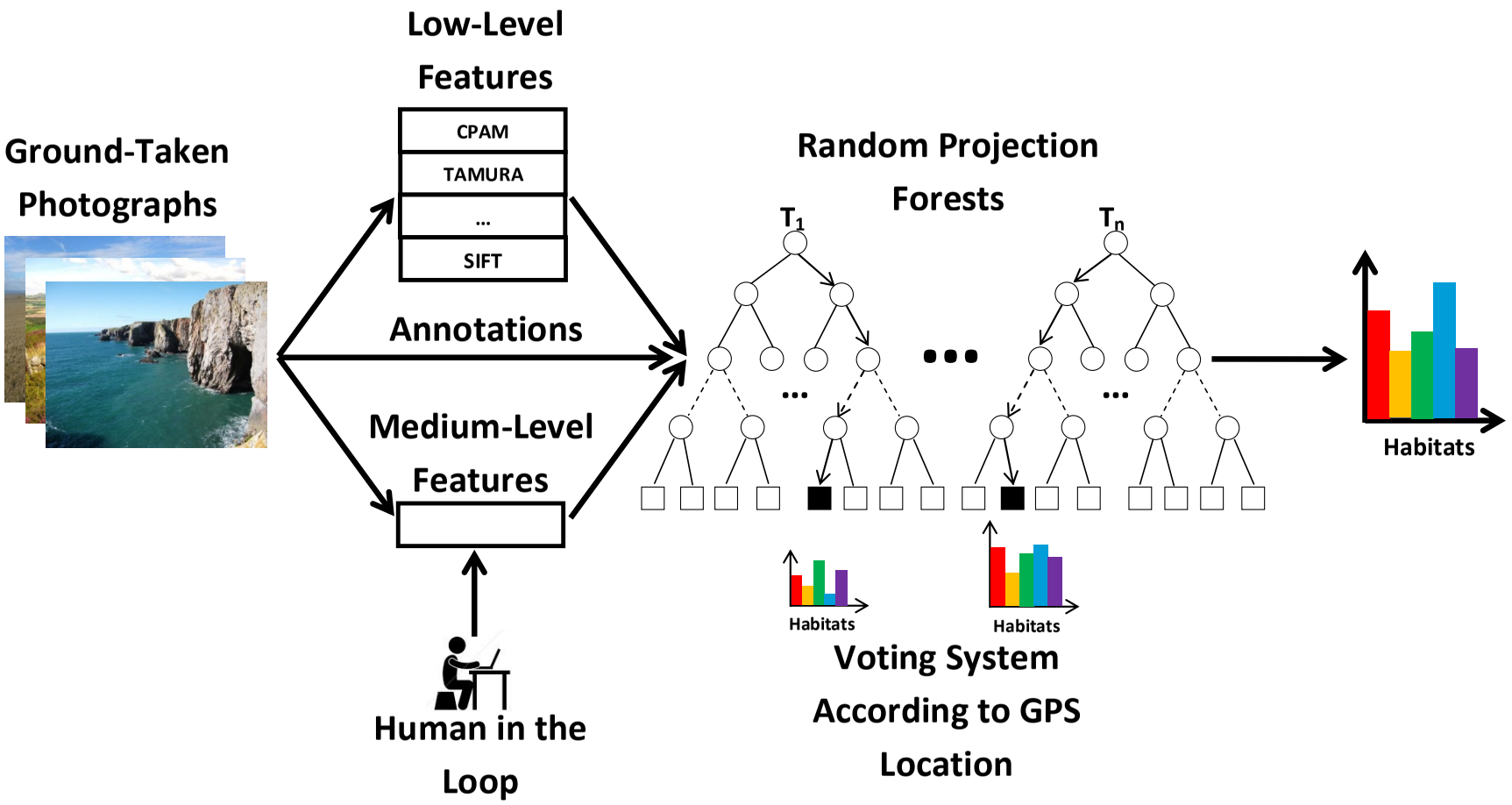


FIGURE 5.4: Image Annotation-Based Habitat Classification. Our framework consists of four elements: the photographs, the features extracted, the classifier and the location-based voting system.

FIGURE 5.5: Ground-Taken Photographs Used In Our Framework. Photographs (a) and (b) belong to Habitat 1K and (c) and (d) belong to Habitat 3K.

Chapter 6

Ground-Taken Photograph Database

As shown and discussed in Chapter 3, remote-sensed data has proven to be insufficient to accurately classify Phase 1 habitats. In this thesis, we study the use of an alternative source of data: ground-taken photographs. Our image annotation framework uses these types photographs to automatically classify habitats. For this purpose, we have created two different annotated datasets: Habitat 1K and Habitat 3K. These are, to our knowledge, the first ground-taken photograph datasets specially created and used for the purpose of automatic habitat classification. Moreover, our framework is also, to our knowledge, the first type of system which uses these types of photographs for the problem of habitat classification.

This chapter is divided into five sections. Section 6.1 describes the overall characteristics of the photographs we will be using and Section 6.2 gives a brief description of the three components of the databases we have created and annotated: the ground-taken photographs, the annotations and the low-level visual features we have extracted from them. Section 6.3 gives a detailed description of the first component: the ground-taken photographs from our datasets. Section 6.4 describes how the annotations were collected, created and stored and how they can be used in conjunction with the visual datasets. Additionally, Section 6.5 describes the third element of the databases, the low-level visual features extracted from the ground-taken photographs, which are also publicly available. We finish this chapter with concluding remarks and a brief summary in Section 3.4.

6.1 Ground-Taken Imagery: Definition

In this thesis, we use ground-taken photographs to automatically classify Phase 1 habitats. We define the term "ground-taken photograph" formally as:

> *A ground-taken photograph is a digital photograph taken by a human on the ground. There are no limitations to the subject of the photograph. Additionally, there are no limitations to its layout in terms of orientation or perspective. These photographs may be taken with any type of digital or mobile camera. These photographs may or may not be geo-referenced.*

This definition of ground-taken photograph is very broad, as it includes both indoors and outdoors photographs. Moreover, it does not restrict its subject. Multiple examples of ground-taken photographs are shown in Figure 6.1.

However, given our goal, the ground-taken photographs that we will be working with need to have some additional restrictions. We are interested in outdoor ground-taken photographs, specifically those taken in rural and coastal areas in the United Kingdom, Europe, for which Phase 1 was specifically designed by the JNCC [102]. There must be at least one discernable habitat instance, either natural (i.e. grasslands, dunes, etc.) or artificial (i.e. walls, fences, parks). Moreover, these instances must be the focus of the photograph.

Thus, for example, in Figure 6.1, we are not interested in working with any of the ground-taken photographs from the first row: (a) would not be used in our database because it contains an indoor scene, (b) is an outdoor scene but it is not a photograph taken in the United Kingdom and contains no habitats and (c), even though it was taken in the United Kingdom, does not have the habitats as the main focus of the photograph. It is important to notice that there are no restrictions as to the layout of the photographs. Consequently, all three photographs from the second row in Figure 6.1 can be used in our framework: (d) shows a ground shot of New Forest, (e) shows a landscape shot of Titchfield Haven which includes three habitats and (f) shows an artificial or man-made boundary habitat, a wall, taken in rural England. These photographs are, in fact, part of our Habitat 3K database.

As can be seen, there are no limitations to the number of habitat classes within a photograph, nor to how they might appear on said photographs. This has been done purposely with the aim of including as much variety and as much information as possible in our database. Additionally, by including the same habitats under many different circumstances (i.e. different times of the year, different perspectives, different orientations, etc.) our database will become more representative and robust.

(a) (b) (c)

(d) (e) (f)

FIGURE 6.1: Ground-taken Photographs.

We have chosen to work with ground-taken photographs in our framework for two main reasons. First, remote-sensed imagery has been proved to be insufficient for the automatic classification of Phase 1 habitats. Second, satellite and aerial photographs are more difficult to obtain. On the other hand, ground-taken photographs can be obtained more easily. Web sites such as Geograph [154] or Flickr [125], can be used to obtain ground-taken imagery with habitats on them. By using these crowd-sourcing sites, we benefit from their large collection of photographs to create a vast and robust database in a relatively effortless manner.

6.2 Annotated Ground-Taken Databases For Automatic Habitat Classification

Following the definition given in Section 6.1, we have compiled two different datasets to study the use of ground-taken imagery for the automatic classification of habitats. These datasets are called Habitat 1K and Habitat 3K. Moreover, we have an intermediate ground-taken image database, referred to as Geograph 2K. Each database has a particular purpose: Habitat 1K was collected with the aim of studying the characteristics of ground-taken photographs and its applicability to habitat classification when the conditions of the images were controlled. The main four first-tier habitats are represented in Habitat 1K: Woodland and Scrub (A), Grassland and Marsh (B), Tall Herb and Fern (C), Heathland (D) and Miscellaneous (J). Geograph 2K was collected through crowd-sourcing methods in order to add more variation and more habitat instances to

our database. From the union of Habitat 1K and Geograph 2K, we created Habitat 3K. Habitat 3K stores information about seven of the ten first-tier habitats: Woodland and Scrub (A), Grassland and Marsh (B), Tall Herb and Fern (C), Heathland (D), Open Water (G), Coastland (H), Rock Exposure and Waste (I) and Miscellaneous (J)

All databases used are composed of the same three elements:

- Ground-taken photographs: These photographs are digital photographs of outdoor scenes taken in rural or coastal England and whose focus are the habitats present, as established in Section 6.1
- Annotations: Each photograph will be annotated with the habitats present in it. The annotations are stored in an XML file, which is easy to create, work with and manipulate.
- Low-level visual features: These low-level visual features include colour, pattern and texture information.

6.3 Ground-Taken Photographs

6.3.1 Habitat 1K

Habitat 1K is the first version of the annotated database created as a contribution for this thesis. It was created with the aim of studying the usefulness of ground-taken photographs for automatic habitat classification under somewhat controlled conditions. Therefore, Habitat 1K can be seen as a starting point to the use of ground-taken imagery for automatic Phase 1 classification.

It contains 1086 ground-taken photographs of rural England. The photographs were taken in the Hampshire County during the summer of 2011 and the winter and summer of 2012. Consequently, most of the habitats present in the database belong to classes A (Woodland and Scrub) and B (Grassland and marsh).

6.3.1.1 Specifications

The specifications of the Habitat 1K are summarised in Table 6.1.

Additionally, Table 6.2 summarises the number of instances of each Phase 1 habitat present in the database and Figure 6.2 shows the same information as an histogram. Moreover, all the photographs are geo-referenced. The geographical location of all images

TABLE 6.1: Specifications of database Habitat 1K

Characteristics	Description
Number of Images	1086
Resolution	3648x2736pixels
Number of Habitats	4223
Average annotations per image	3.88
Maximum annotations per image	6
Minimum annotations per image	1
Camera Model	Sony Cybershot DSCHXvb

taken in 2011 and 2012 are shown projected in a map in Figure 6.3 and Figure 6.4, respectively.

6.3.1.2 Collection and Ground-truth

This database was collected during two different seasons in four different sites. Photographs from New Forest and Titchfield Haven were taken by the author of this thesis during the months of June and July of 2011. On the other hand, photographs from Christmas Commons were taken by researchers in The Ordnance Survey [173] in February 2012. Researchers in The Ordnance Survey also took the photographs from Wildgrounds Nature Reserve in July 2012.

All four sites were surveyed by the same Phase 1 expert, which guaranteed an agreement in the classification. Figure 6.5 shows two of the Phase 1 maps produced by this expert during the surveys: (a) shows the classification map from Titchfield Haven and (b) and shows the maps from New Forest. The information obtained was digitised by the author of this thesis using The OS MasterMap [173] and ArcGIS [64].

6.3.1.3 Visual Examples

Figure 6.6 shows a collection of sixteen photographs taken from our Habitat 1K database. Photographs (a) to (d) were taken in New Forest, while photographs (e) to (h) were taken in Titchfield Haven. Moreover, Figure 6.7 shows photographs from Christmas Commons, from (a) to (d), and photographs from Wildgrounds Nature Reserve, from (e) to (h). These photographs are a prime example of the level of variability of the conditions are with regards to perspective, layout, lighting, number of habitats present. As can be seen, even though these are variable, there is still some control over the conditions of the photographs.

TABLE 6.2: Habitat 1K. Habitats Instances in our database Habitat 1K

Habitat	Number of instances
Woodland broad leaved	406
Woodland coniferous	48
Woodland mixed	206
Scrub dense	298
Scrub scattered	22
Acid grassland unimproved	2
Acid grassland semi improved	150
Neutral grassland unimproved	127
Neutral grassland semi improved	391
Improved grassland	299
Marshy grassland	36
Poor semi improved grassland	3
Bracken continous	55
Bracken scattered	10
Tall ruderal	30
Dry dwarf shrub heath acid	40
Dry dwarf shrub heath basic	7
Dry heath acid grassland mosaic	88
Fen	1
Standing water	1
Running water	17
Cultivated arable	66
Hedge and trees species rich	111
Hedge and trees species poor	232
Fence	235
Wall	14
Dry ditch	9
Bare ground	12
Sky	1048
Others	259

6.3.1.4 Merits and Limitations of Habitat 1K

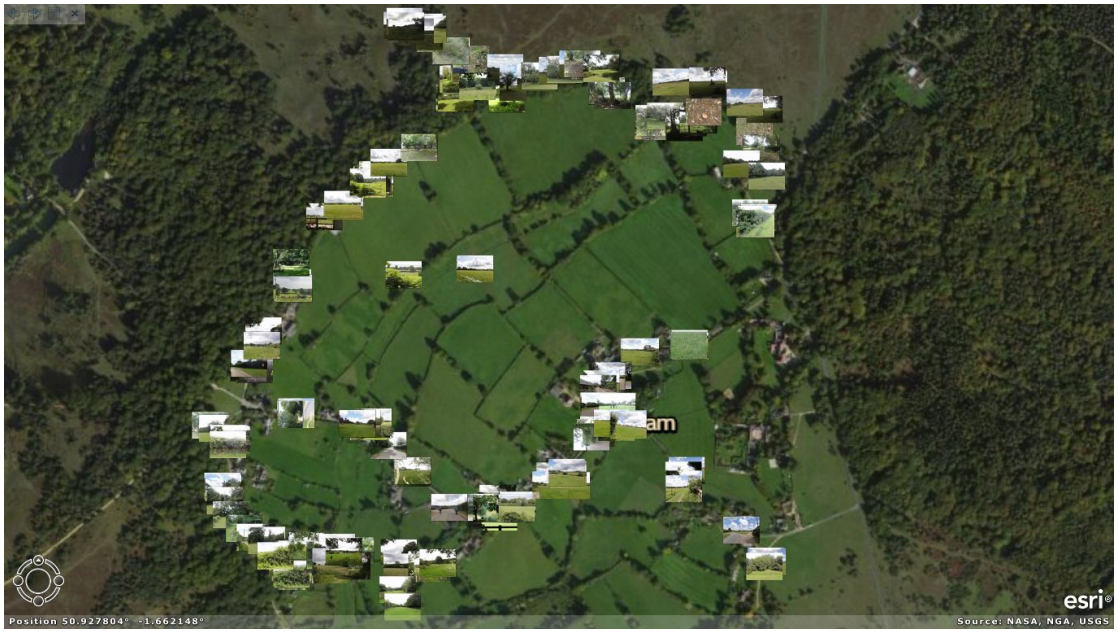
Habitat 1K was created with the specific goal of assessing how useful ground-taken photographs could be for automatic Phase 1 classification. Moreover, it was also created to study the performance of our image-annotation framework. For this, the database had to comply with two main requirements related to balance. First, it had to have a manageable size. Large enough to obtain reliable results but also small enough to be managed by a single person. Second, it had to be robust and variable enough in terms of perspective, layout, lighting and types habitats present. In summary, the aim was to create manageable database which presented a balance between different conditions, in order to make the database robust, but the goal was also to create a database in which there was control over these aforementioned conditions, to fully study if important and

relevant information could be extracted from this type of photographs in contrast to remote sensing imagery. Consequently, Habitat 1K should be used when there is an interest in testing a smaller number of habitats under controlled conditions.

Creating this type of database also helped to identify some of the advantages and limitations of ground-taken photography in comparison with remote sensing imagery. For example, similarly to aerial imagery, ground-taken photograph is subject to lighting conditions. However, the level of detail in terms of pattern, texture and colour is much more visible in ground-taken photographs.

Nevertheless, the main challenge of ground-taken photographs, which is not present in remote sensed data, is the discordance between the location of the photographer and the location of the habitats being photographed. That is, the location of a photograph does not necessarily have to reflect the location of the items that appear in the photographs. This is due to the angle and orientation properties of the photograph. While the angle in remote sensed data is constant, perpendicular to the ground [117], in the case of ground-taken photographs, as shown if Figure 6.8, it can vary a great deal. This will be one of the main challenges of using geographic location in the classification process and it will further discussed in Chapter 7, Chapter 8 and 9.

FIGURE 6.2: Habitat 1K. Instances of each habitat in our database.



(a) New Forest - July 2011

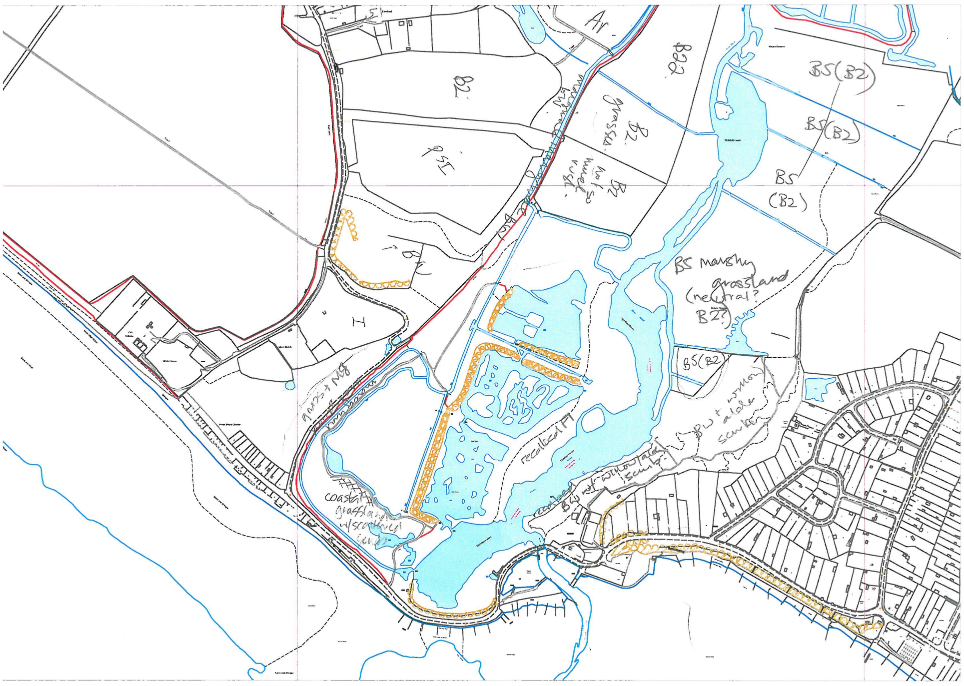
(b) Titchfield Haven - July 2011

FIGURE 6.3: Habitat 1K. Ground-taken images taken in 2011 projected on a map.

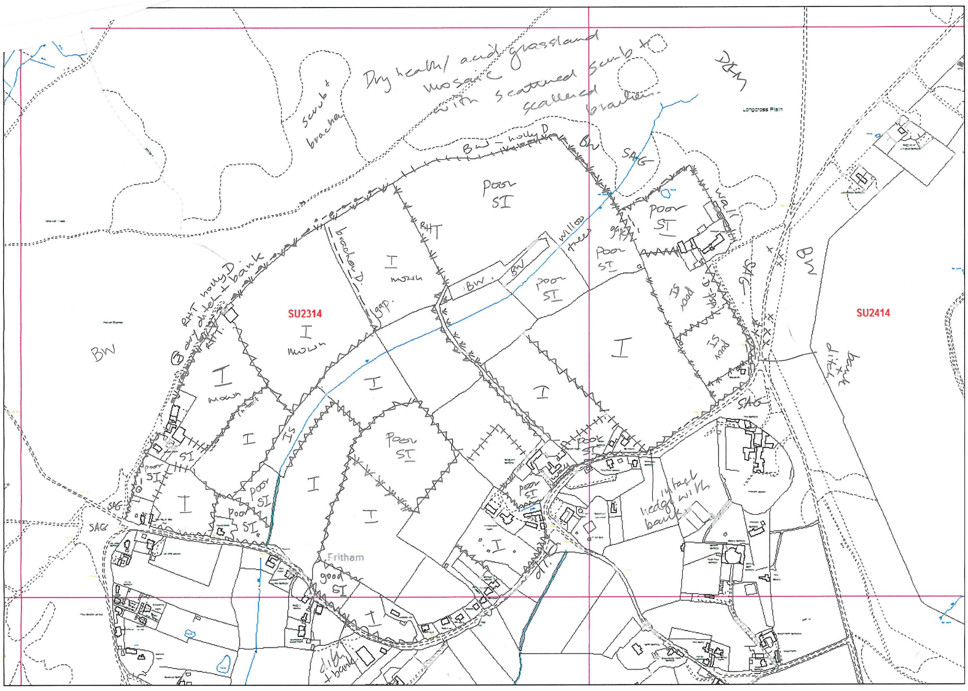
(a) Christmas Common - February 2012

(b) Wildgrounds Nature Reserve - July 2012

FIGURE 6.4: Habitat 1K. Ground-taken images taken in 2012 projected on a map.



(a)



(b)

FIGURE 6.5: Phase 1 Habitat maps filled by an expert.

(a) (b)

(c) (d)

(e) (f)

(g) (h)

FIGURE 6.6: Habitat 1K. (a) to (d) show photographs from New Forest, taken in July 2011. (e) to (h) show photographs from the Titchfield Haven in July 2011.

(a) (b)

(c) (d)

(e) (f)

(g) (h)

FIGURE 6.7: Habitat 1K. (a) to (d) show photographs from Christmas Commons, taken in February 2012. (e) to (h) show photographs from the Wildgrounds Nature Reserve in July 2012.

(a)

(b)

FIGURE 6.8: Habitat 1K. Differences in perspective. (a) shows a ground-shot while (b) shows a landscape shot. Both types of perspectives are present in our Habitat 1K database.

6.3.2 Geograph 2K

Results from studying the use of Habitat 1K with our approach, as shown in Chapter 7 and onwards, proved that ground-taken photographs were indeed a promising source of information for automatic classification. Consequently, the next logical step was to include more variability in our database in order to further test our image-annotation approach. As mentioned in the previous section, one of the main characteristics of Habitat 1K was its manageable size. This size was convenient for a preliminary study on the merits of ground-taken photographs and our annotation framework for automatic habitat classification. However, it also left out many other habitat types and different types of sites.

In an effort to include more variability on the photographs conditions and the habitats present in them and to increase the number of images in the database, Geograph 2K was created. Geograph 2K has 2094 photographs and it contains photographs from all over Great Britain. Moreover, it not only includes rural areas but also coastal environments.

6.3.2.1 Specifications

The specifications of the Geograph 2K are summarised in Table 6.3.

TABLE 6.3: Specifications of database Geograph 2K

Characteristics	Description
Number of Images	2008
Resolution	640x480pixels
Camera Model	Different Models
Number of Habitats	7121
Average annotations per image	3.55
Maximum annotations per image	5
Minimum annotations per image	1

As shown in Table 6.3, the resolution of the images is much lower. However, the number of habitats in significantly larger. Table 6.4 summarises the number of instances of each Phase 1 habitat present in the database and Figure 6.9 shows the same information as an histogram. It can be seen that Geograph 2K contains eight out of the ten Phase 1 first-tier classes: Woodland and Scrub (A), Grassland and Marsh (B), Tall herb and fern (C), Heathland (D), Open Water (G), Coastland (H), Rock Exposure and Waste (I) and Miscellaneous (J). These are the first collected instances of classes G, H and I. Moreover, contrary to the Habitat 1K database, we had no control over the location of the photographs. Consequently, these are much more sparsely distributed.

Table 6.4: Geograph 2K. Habitats instances in the database Geograph 2K

Habitat	Number of instances
Woodland broad leaved	617
Woodland coniferous	179
Woodland mixed	248
Scrub dense	216
Recently felled woodland broad leaved	3
Acid grassland unimproved	28
Acid grassland semi improved	226
Neutral grassland unimproved	13
Neutral grassland semi improved	417
Improved grassland	98
Marshy grassland	163
Bracken continuous	132
Tall ruderal	76
Dry dwarf shrub heath acid	331
Dry dwarf shrub heath basic	9
Dry heath acid grassland mosaic	348
Wet heath acid grassland mosaic	1
Fen	1
Marginal vegetation	1
Standing water	0
Running water	885
Intertidal mud sand	157
Intertidal shingles cobbles	118
Intertidal boulders rocks	104
Boulders above high tide	1
Sand dune dune grassland	2
Sand dune dune heath	4
Sand dune open dune	2
Maritime cliff slope hard cliff	175
Maritime cliff slope soft cliff	47
Maritime cliff slope coastal grassland	1
Maritime cliff slope coastal heathland	3
Inland cliff acid neutral	132
Scree acid neutral	13
Cultivated arable	77
Cultivated introduced shrub	1
Intact hedge species rich	1
Hedge and trees species rich	176
Hedge and trees species poor	7
Fence	90
Wall	91
Dry ditch	1
Buildings	139
Sky	1654
Others	133

6.3.2.2 Collection and Ground-truth

Geograph 2K was collected by the author of this thesis using the Geograph crowd-sourcing website. Geograph maintains a database of over four million ground-taken photographs. Moreover, it also stores associated metadata, such as geographical location and the time of the photo, for all of these photographs. Geograph's aim is to collect, publish, organise and preserve representative images and associated information for every square kilometre of Great Britain, Ireland, and the Isle of Man [154]. Its photographs, as well as its associated information, are freely available to the public.

Geograph photographs can be tagged and annotated. Consequently, we collected 2094 additional photographs using this search-by-tag feature and by searching for the ground-taken photographs with any of these tags: Arable, Boundary, Coastal, Flat landscapes, Grassland, Heath, Scrub, Hedge, Lakes, Park and Public Gardens, Rivers, Streams, Drainage, Rocks, Scree, Cliffs, Wall, Woodland, Forest, while excluding theses tags: Housing, Dwellings, Suburb, Urban fringe, Business, Retail, Services, Docks, Harbours, Roads, Road transport.

The photographs were taken year-round in England, Ireland and the Isle of Man by different people and using different types of cameras. Moreover, they were classified and digitised using their tags by the author of this thesis.

6.3.2.3 Visual Examples

Figure 6.10 shows six different examples of the types of photographs present in Geograph 2K. As can be seen, the habitats present in this database are much more varied than those in Habitat 1K. It is also important to notice that the lighting conditions, perspective and layouts present are also much more varied.

6.3.2.4 Merits and Limitations of Geograph 2K

Geograph 2K was created with the aim of increasing the number of photographs in our database and the types of habitats present in it. In comparison with Habitat 1K, Geograph 2K has double the number of images and it includes habitat from three new classes: Open Water, Coastland and Rock Exposures and Waste. Moreover, it includes photographs from all over Great Britain. The photographs were also taken during different years and under different weather and seasonal conditions.

This makes Geograph 2K much more varied than Habitat 1K and, consequently, more robust. However, there is a trade off between this increase in variety and the control we

have over the photographs in the database. First, all the images have a drastically lower resolution. Habitat 1K photographs have a resolution of 3648x2730 pixels, while Geograph 2K photographs have a resolution of 640x480 pixels. This was a "necessary evil", as the 2000 photographs were downloaded from the Internet. Moreover, the ground-truth of this set of photographs was also obtained through Geograph and then modified and refined by the author of this thesis. This means that the ground-truth information was obtained through crowd-sourcing methods, with the users who uploaded the images being the ones introducing the tags and classifying the habitats. This makes this classification process less consistent than the process followed with Habitat 1K.

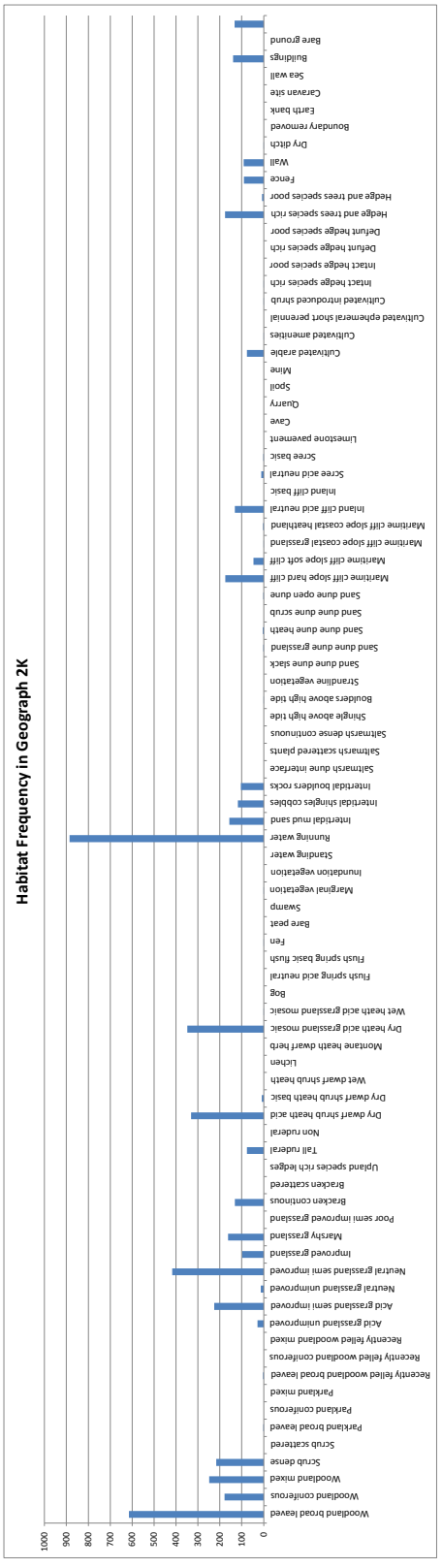


FIGURE 6.9: Geograph 2K. Instances of each habitat in the Geograph 2K database.

(a) (b)

(c) (d)

(e) (f)

FIGURE 6.10: Geograph 2K. (a) to (g) show photographs from the database Geograph 2K. Differences in perspectives, layout and lighting are clearly identifiable. This is mainly due to the crowd-sourcing nature of the photographs, which were taken at different times by different people.

6.3.3 Habitat 3K

The second database contribution of this thesis is Habitat 3K. Habitat 3K is the combination of the two databases previously introduced, Habitat 1K and Geograph 2K. It combines the characteristics of both databases and it contains over 11,000 habitat instances from eight out of the ten first-level classes. Habitat 3K was created with the goal of further testing our approach under more variable conditions and taking into consideration more habitats.

6.3.3.1 Specifications

The specifications of the Habitat 3K are summarised in Table 6.5.

TABLE 6.5: Specifications of database Habitat 3K

Characteristics	Description
Number of Images	3094
Resolution	640x480p and 3648x2736p
Camera Model	Different Models
Number of Habitats	11344
Average annotations per image	3.66
Maximum annotations per image	6
Minimum annotations per image	1

Additionally, table 6.6 summarises the number of instances of each Phase 1 habitat present in the database and Figure 6.11 shows the same information as an histogram. Moreover, Figure 6.12 shows how the three databases compare to each other in terms of first-tier habitat classes.

Table 6.6: Habitat 3K. Habitats instances in the database Habitat 3K

Habitat	Number of instances
Woodland broad leaved	1023
Woodland coniferous	227
Woodland mixed	454
Scrub dense	514
Scrub scattered	22
Recently felled woodland broad leaved	3
Acid grassland unimproved	30
Acid grassland semi improved	376
Neutral grassland unimproved	140
Neutral grassland semi improved	808
Improved grassland	397
Marshy grassland	199
Poor semi improved grassland	3
Bracken continuous	187
Bracken scattered	10
Tall ruderal	106
Dry dwarf shrub heath acid	371
Dry dwarf shrub heath basic	16
Dry heath acid grassland mosaic	436
Wet heath acid grassland mosaic	1
Fen	2
Marginal vegetation	1
Standing water	1
Running water	902
Intertidal mud sand	157
Intertidal shingles cobbles	118
Intertidal boulders rocks	104
Boulders above high tide	1
Sand dune dune grassland	2
Sand dune dune heath	4
Sand dune open dune	2
Maritime cliff slope hard cliff	175
Maritime cliff slope soft cliff	47
Maritime cliff slope coastal grassland	1
Maritime cliff slope coastal heathland	3
Inland cliff acid neutral	132
Scree acid neutral	13
Cultivated arable	143
Cultivated introduced shrub	1
Intact hedge species rich	1
Hedge and trees species rich	287
Hedge and trees species poor	239
Fence	325
Wall	105
Dry ditch	10
Buildings	139
Bare ground	12
Sky	2702
Others	392

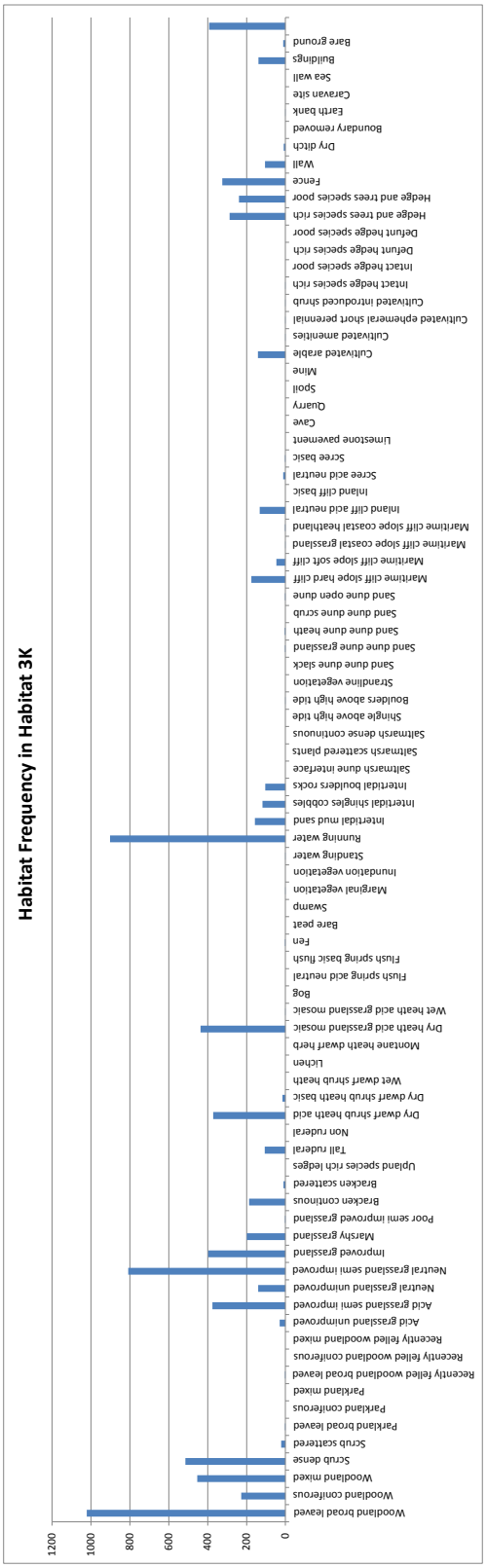


FIGURE 6.11: Habitat 3K. Instances of each habitat in the Habitat 3K database.

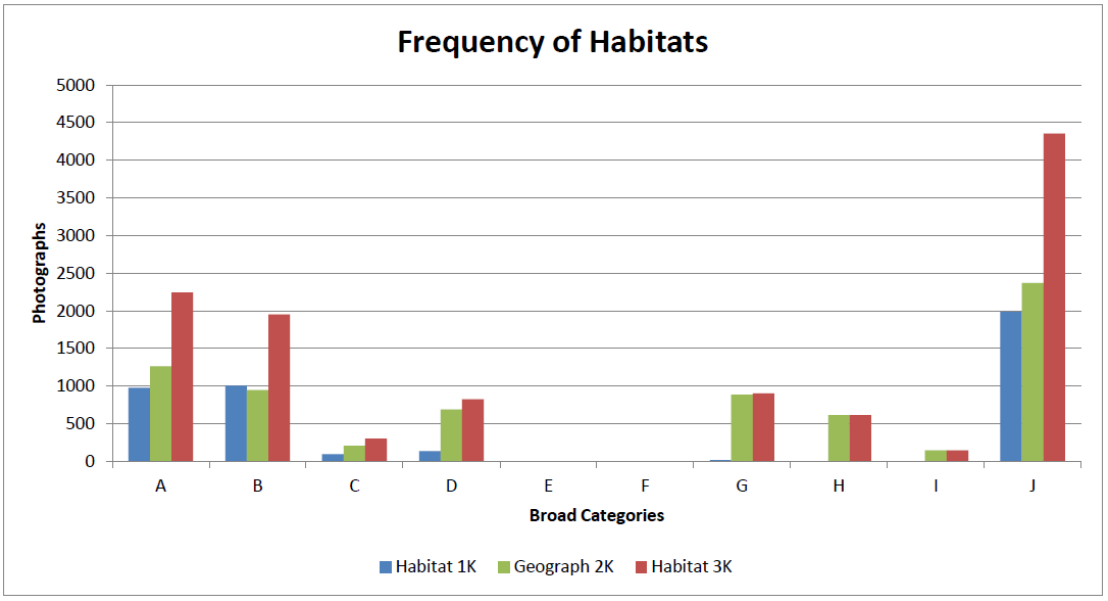


FIGURE 6.12: Datasets Comparison.Instances of each first-tier habitat in Habitat 1K, Geograph 2K and Habitat 3K databases.

6.3.3.2 Visual Examples

Figure 6.13 shows a collection of six images taken from the Habitat 3K database. As can be seen, there are clear differences is resolution of the images, lighting conditions and the layout of the photographs.

6.3.3.3 Merits and Limitations of Habitat 3K

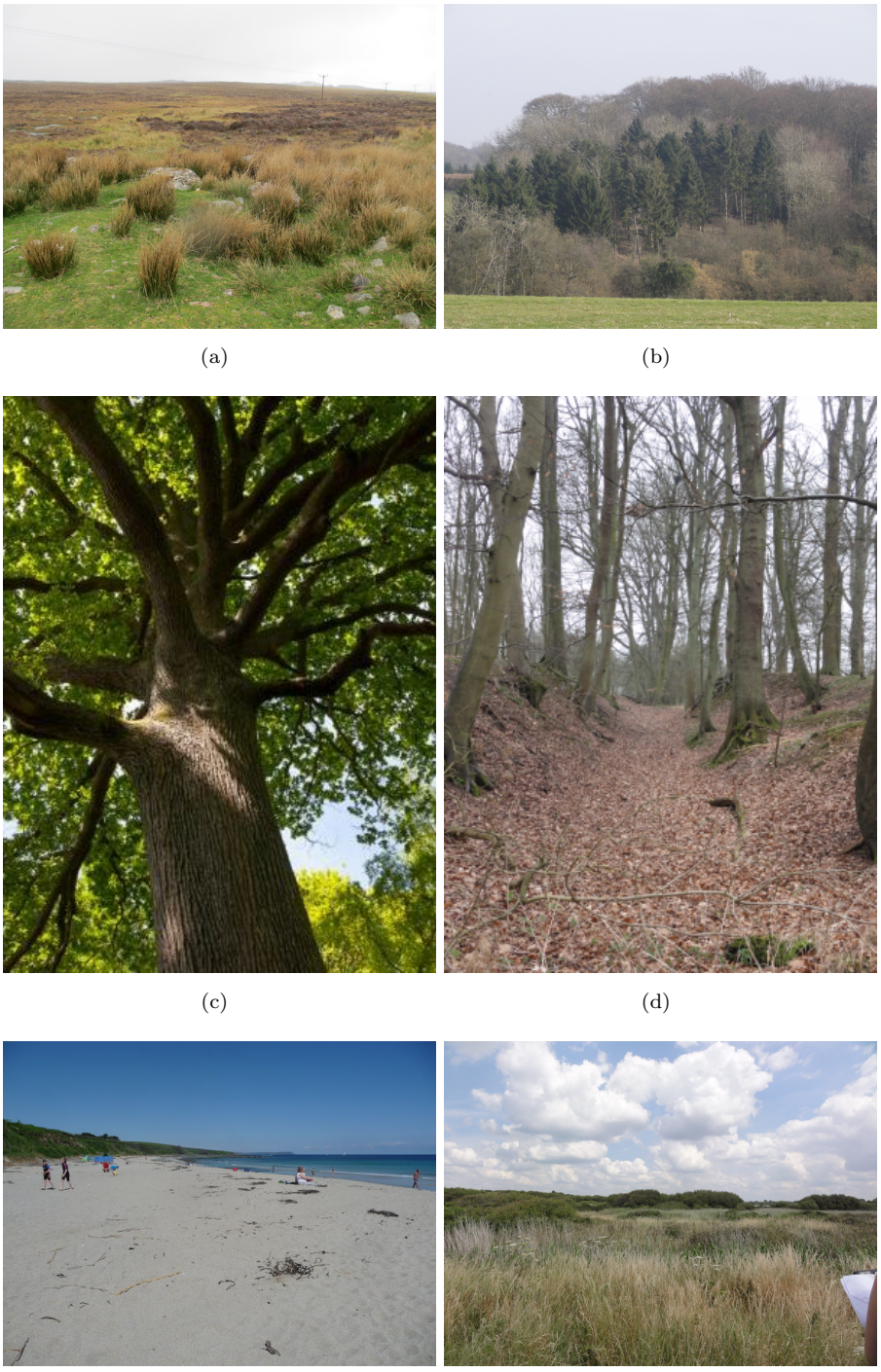
Habitat 3K is the result of combining the ground-taken photographs from Habitat 1K and Geograph 2K together. Consequently, it combines all the merits and limitations from both datasets, as discussed in Section 6.3.1.4 and Section 6.3.2.2. Its size is three times the size of Habitat 1K and it contains more than twice the number of habitats. Moreover, much like Geograph 2K, it contains habitats from eight out of the ten possible Phase 1 first-tier habitat classes, including Coastland habitats. Consequently, Habitat 3K should be used when interested in testing under a mixture of conditions: diverse and varied conditions over two thirds and some controlled condition over a third of the database.

Additionally, it contains a mixture of high and low resolution photographs, taken during all twelve months of the year in Great Britain. Finally, its classification is a mixture of the classification done by an expert in Phase 1 and the classification obtained from Geograph's tagging system.

6.4 Annotations

All images in both Habitat 1K and Habitat 3K were annotated by the author of this thesis following the same procedure and using the same tool. This image annotation tool was developed by the University of Bonn and it was specially designed for MATLAB [107]. Its interface is shown in Figure 6.14.

In essence, each annotation has of two main components: a polygon, which delimits where in the image a habitat appears, and its corresponding label, which follows the Phase 1 classification scheme. Examples of five annotated images are shown in Figure 6.15, Figure 6.16, Figure 6.17, Figure 6.18 and Figure 6.19. However, as can be seen in Figure 6.14, there is more information that can be included in the annotation, such as the degree of confidence in the annotation, the occlusion index of the object and the degree of representativeness of the object with regards to its class.



(a) (b)

(c) (d)

(e) (f)

FIGURE 6.13: Habitat 3K. Photographs from the 1st column belong to Geograph 2k. Photographs from the 2nd column belong to Habitat 1K. The differences in lighting and perspective are clearly identifiable.

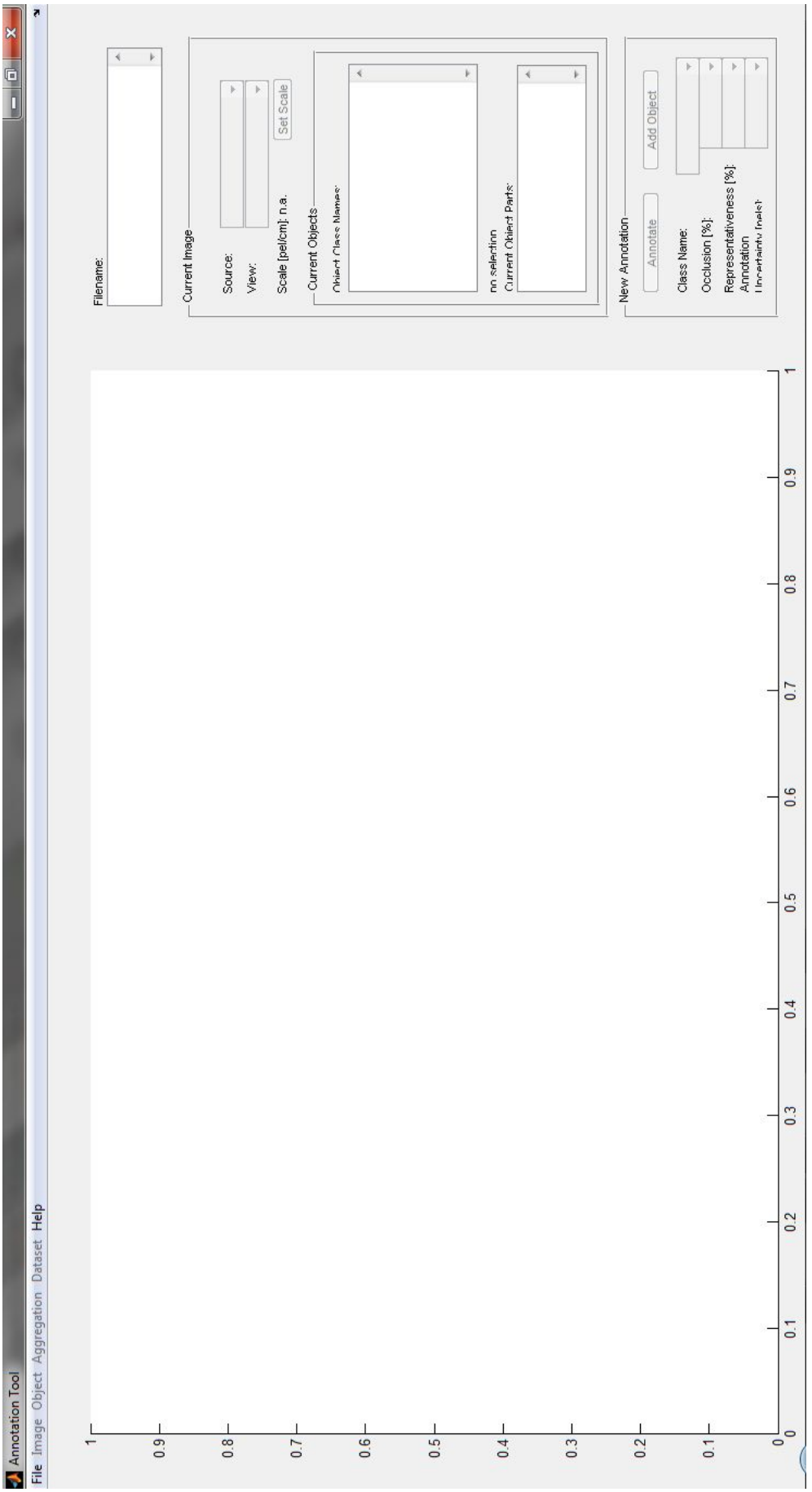


FIGURE 6.14: Image annotation tool. The tool also collects information about the source and the type of view of the whole image and the representatives, occlusion and the uncertainty of each annotation.

FIGURE 6.15: Annotated Image From Habitat 1K.

FIGURE 6.16: Annotated Image From Habitat 1K.

FIGURE 6.17: Annotated Image From Habitat 1K.

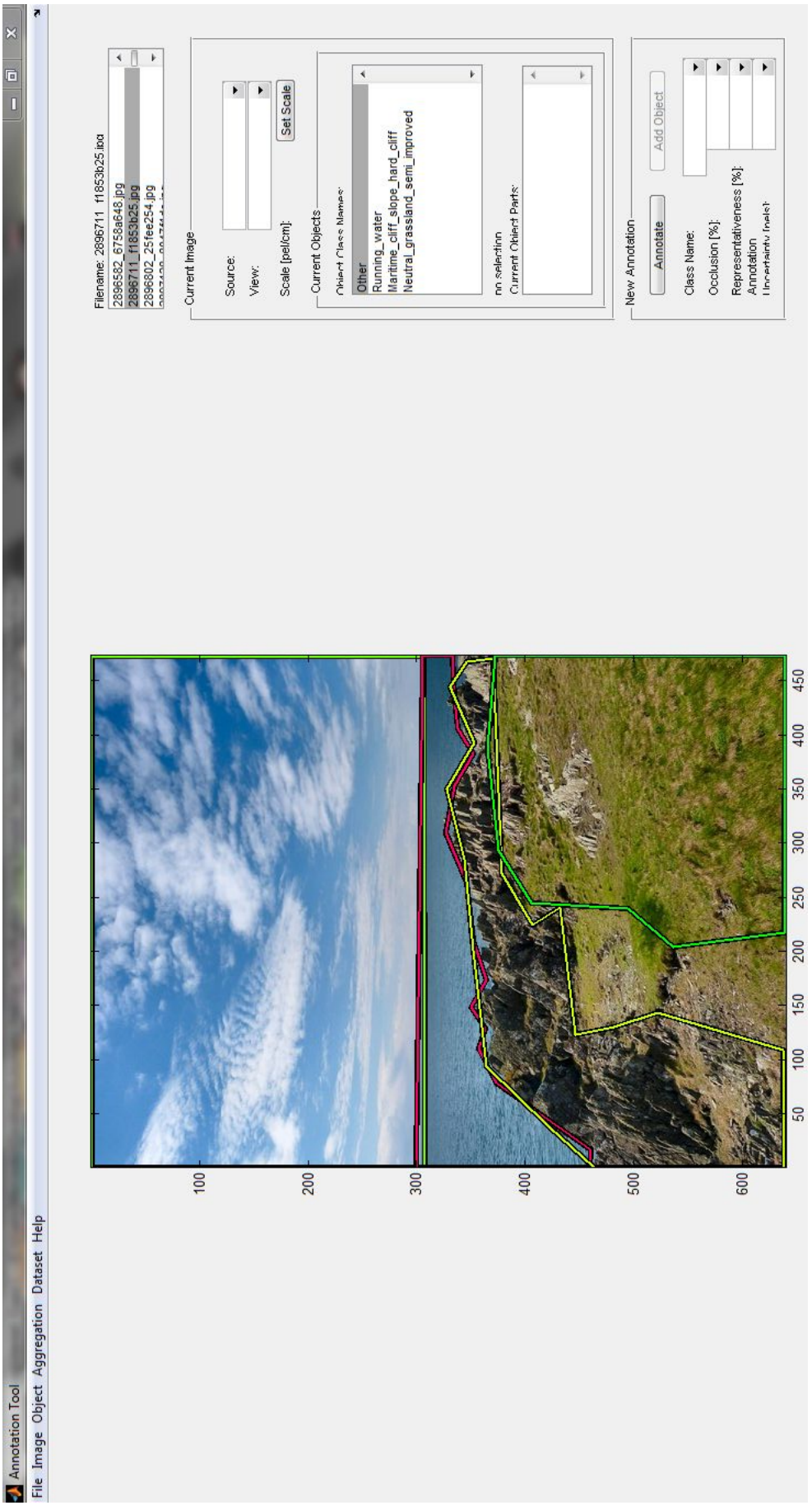


FIGURE 6.18: Annotated Image From Habitat 3K.

FIGURE 6.19: Annotated Image From Habitat 3K.

The information about the annotations is conveniently stored in an XML file per photograph. An example of this file is shown in Code 6.1. This XML file corresponds to the ground-taken photograph shown in Figure 6.6(c).

```
<annotation>
    <filename>DSC03245.png</filename>
    <folder>All Images</folder>
    <sourceImage>Habitat 1K</sourceImage>
    <sourceAnnotationXML>Version 2.40</sourceAnnotationXML>
    <rectified>0</rectified>
    <viewType>ground_taken</viewType>
    <scale>n/a</scale>
    <imageWidth>3648</imageWidth>
    <imageHeight>2736</imageHeight>
    <transformationMatrix>n/a</transformationMatrix>
    <object>
        <name>Improved_grassland</name>
        <objectID>73494601513464</objectID>
        <occlusion>0</occlusion>
        <representativeness>80</representativeness>
        <uncertainty>n/a</uncertainty>
        <deleted>0</deleted>
        <verified>0</verified>
        <date>18-Mar-2012</date>
        <sourceAnnotation>Mercedes</sourceAnnotation>
            <polygon>
                <pt>
                    <x>1</x>
                    <y>1</y>
                </pt>
                <pt>
                    <x>3648</x>
                    <y>1</y>
                </pt>
                <pt>
                    <x>3648</x>
                    <y>2736</y>
                </pt>
                <pt>
                    <x>533.2525</x>
                    <y>2736</y>
```

```
                </pt>
                <pt>
                    <x>1</x>
                    <y>2736</y>
                </pt>
            </polygon>
            <objectParts>n/a</objectParts>
            <comment> </comment>
    </object>
</annotation>
```

LISTING 6.1: XML annotation file from a ground-taken photograph

As can be seen, this information in the XML file includes: who made the annotations, when it was made, which classification scheme it follows, the location of the image, the location of annotation file, the locations of the different polygons within the image and its corresponding classes. Having this information stored in an XML file makes its use and manipulation easier when working with MATLAB, the environment we have used to develop our framework.

However, while easy to work with and to manipulate, this approach presents a clear limitation. It assumes that the limits of all habitats are clearly distinguishable and separable in our photographs This is not always the case, as the frontiers between habitats might be fuzzy. An example of this is shown in Figure 6.10(e), in which the limits between the sand and the water are not clear cut.

6.5 Low-Level Features

Visual-database retrieval and search are becoming increasingly popular activities. However, image databases are increasing their size exponentially [125, 154]. As a consequence, indexing and retrieving thousands or even millions of images is a difficult task that needs to combine both high accuracy and low execution time. This has inspired a wide variety of research approaches, such as content-based image retrieval [118, 169], image classification [148] and image annotation [76]. Not incidentally, most of these approaches have the same preliminary step: dimensionality reduction by feature extraction [68].

In Pattern Recognition, local features are defined as points or regions of interest in the images. The use of features involves two main tasks which are connected: feature selection and feature extraction. As discussed in [56], feature selection and extraction

can be regarded as the most important step in the pattern recognition framework, since the selected features will directly influence the design of the classifier and, consequently, the results of the system.

The main aim of feature selection and extraction is to find the most compact and descriptive and relevant combination of features to use during the classification process. Selecting the correct features is not only a crucial step but also quite a challenging task, since the right grouping of features is problem-dependent [56]. Consequently, a set of features that might be successfully applied to face recognition, might yield less than accurate results for a different task, such as object recognition.

Once relevant features have been selected, feature extraction is carried out to efficiently reduce the dimensionality of the data into a compact and descriptive feature vector. In our AIA framework, we have chosen the extraction of low-level visual features as our first step. Low-level visual features collect statistics about different aspects of an image, such as color [162, 165], texture [85, 175], pattern [148] or shape [158, 196] information. Extracting low-level visual features enables us to work with a large number of high-definition photographs in an efficient and accurate manner. Moreover, as discussed in Chapter 2, feature extraction helps combat "the curse of dimensionality" [18] in the classification process. Moreover, features also allow for an easier comparison between images with different characteristics.

Using mathematical notation and applying it to our case, in which we work with colour ground-taken photograph and global features, the aim of feature extraction is to transform a N-by-3-dimensional matrix, the colour ground-taken photograph, $X = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3]^T$, with $\mathbf{x}_i = [x_{i1}, x_{i2}, ..., x_{in}]$ by finding f, such that X is mapped into a M-dimensional vector $Y = [y_1, y_2, ..., y_m]^T$, with $m < n$. Y can therefore be expressed as $Y = f(X)$.

In this thesis, we extract a total of eleven different low-level visual features. We have divided them into four main categories according to the nature of the information that is extracted: pattern features, color features, texture features and other features. Other features include a set of six features commonly used in Pattern Recognition problems. In particular, we are interested in studying how pattern, color and texture features perform in our framework. Moreover, we are also interested in how their performance compares to the performance of other popular features used in Pattern Recognition.

6.5.1 Pattern Features

A pattern is defined by the Oxford dictionary as "an arrangement or sequence regularly found in comparable objects or events" [8]. Pattern information combines both colour

and spatial information. Consequently, pattern features are extremely useful when distinguishing between similar habitat classes. An example of this is shown in Figure 6.20, which shows different types of heath mosaics, easily distinguishable to the human eye because of their differences in pattern.

(a) (b)

FIGURE 6.20: Pattern Information. Pattern is crucial when distinguishing between habitats. These two Heath habitats are easily identifiable to humans due, in part, to their pattern information.

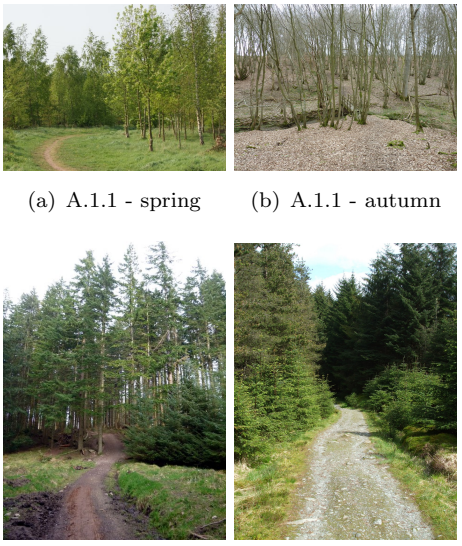
We have chosen to extract Color Pattern Appearance Model (CPAM) [148] features for this purpose. CPAM features were one of the earliest bag-of-visual-words style image content representation features. Moreover, they have been successfully applied to image retrieval [148] and image annotation [216]. CPAM features are extracted using two codebooks, referred to as achromatic and chromatic codebook. Together, they are used to capture both color and texture patterns of tiles within the photographs. For this reason, because it collects colour and texture pattern information in an extremely compact manner, we chose to use CPAM feature as the main guideline to assess the performance of our framework. As mentioned in Section 5.4.2, CPAM features were the first features that we extracted of all testing scenarios to assess the validity of each experimentation approach. Depending on the results, we either decide to continue further testing or not. This is clearly exemplified in Chapter 7, when testing the use of blocks within the images as the input of our classifier yielded surprisingly inaccurate results and served to identify why the use of tiles was not appropriate for the task of habitat classification. Consequently, we discarded the idea of using of blocks as input in further experiments with our system.

In essence, CPAM features are global histograms capturing the frequencies of the codewords that have been used to encode patches of the image for both codebooks. In our experiments, we used a 64 codewords achromatic codebook and a 64 codewords chromatic codebook. Consequently, using CPAM features enable us to encode each of the ground-taken photographs in our database as a 128-dimension vector which collects pattern information.

6.5.2 Colour Features

Colour is defined as by the Oxford dictionary as "the property possessed by an object of producing different sensations on the eye as a result of the way it reflects or emits light" [8]. Colour features are one of the most popular features extracted in Pattern Recognition problems [165] because color properties in general provide extensive information about the nature and characteristics of the objects that need to be classified.

In this thesis, we extract colour features because, as mentioned, colour information is very powerful descriptive tool to distinguish objects in general and habitats in particular. For example, it can be used to distinguish between broad-leaved(A.1.1) and coniferous woodland (A.1.2). As shown in Figure 6.21, broad-leaved woodland is commonly bright green during spring and summer or completely brown during autumn and winter while coniferous woodland is dark green during all four seasons.



(a) A.1.1 - spring (b) A.1.1 - autumn

(c) A.1.1 (l) and A.1.2. (r) - spring (d) A.1.2. - winter

FIGURE 6.21: Colour Information. These two Woodland habitats can easily be differentiated due to their different colour properties. (l) stands for left and (r) for right.

In our case, we extract two simple but powerful global colour features:

- Colour Histograms: A histogram is defined as the statistical representation of the frequency of appearance of a pixel value [165]. The use of histograms as colour features has been researched at length in works such as [162]. Histograms are extremely useful because they collect global information about the colour distribution within an image. In our case, we will be extracting 256-bin colour histograms from each of the channels of three different colour spaces. These colour spaces are: RGB, HSV and CIEL*a*b*. Consequently, each photograph will generate nine colour

feature vectors, which will be arranged in a 256-by-9 matrix. Moreover, since not all the images in our Habitat 3K database have the same size, we normalise their histograms by we dividing each bin by the number of pixels of the photograph.

It is important to notice that, if we were to have applied NN-based methods to annotate the images in our database, the use of such large 256-bin histograms would have proven problematic and would have made the testing phase quite inefficient [97]. A common solution would have been to create smaller colour histograms, for example 10-bin colour histograms, and group pixel values together. This solution would not be useful in our case since most of the colours that appear in the photographs are different shades of basic nature colours, particularly green and brown. Consequently, in our case, we are specially interested in collecting slight changes or differences in colour, since they can mean, as shown previously in Figure 6.21, that broad-leaved woodland is present in the photograph, instead of coniferous woodland. However, since we are using Random Forests, we can use 256-bin histograms without sacrificing efficiency. Random Forests take a random number of features in each node instead of all of them at once. Moreover, as we will propose in Chapter 7, we can even improve efficiency by taking all features into consideration during training but projecting them at each node [183]. Consequently, Random Forests are proven once again to be a much more suitable choice for automatic habitat classification.

- Colour Moments: The second type of colour feature we extract are colour moments [190]. They have been successfully applied in popular Computer Vision problems, such as object category retrieval [118]. Similarly to colour histograms, colour moments assume that the colour within an image can be represented as a probability distribution. All probability distributions are characterised by a number of unique moments. Therefore, the colour characteristics of an image, which follows a probability distribution, can be used to calculate its unique moments. We calculate six possible moments. As with colour histograms, we extract these measures from the photographs three different colour spaces: RGB, HSV and CIEL*a*b.

6.5.3 Texture Features

Defined in [1] as "an ensemble of repetitive subpatterns, which follow a set of well defined placement rules", the concept of texture is difficult to define formally. However, it is an easy concept for humans to identify [175]. As studied in [53], texture features are related to higher frequencies in the image spectrum.

Along with pattern and colour features, texture features also offer important and discriminative information for the classification of habitats. For example, as shown in Figure 6.22, the texture of the photographs alone, without taking into consideration the colour, is enough for humans to clearly identify that the habitats shown in each figure, scrub and bracken respectively, are different. To exemplify this, we have converted both images to grayscale.

(a) (b)

FIGURE 6.22: Texture Information. Although difficult to formally define, differences in texture are easily identifiable to humans. These two habitats are clearly from separate classes, due to their different texture properties.

In our case, we will extract two of the most popular texture features developed to date:

- Grey Level Co-occurrence Matrices (GLCM): One of the most popular texture features, GLCM measure the frequency with which two pixels appear next to each other within a pre-determined distance [85]. We will use a distance of 1 in each direction, obtaining 8 different directions: north, south, east, west, northeast, northwest, southeast and southwest. Consequently, each image in our database will generate 8 GLCMs matrices.

- Tamura Coefficients: Introduced in [175] by Tamura et al, These coefficients relate to the human visual perception process. [175] developed six possible coefficients that range from most relevant to least relevant. These coefficients are: coarseness, contrast, directionality, line-likeness, regularity and roughness. In our case, we will use the first three, which have been proven to perform accurately when used together [96]. Coarseness, selected in [175] as the most important of the coefficients, aims to identify the largest texture in a image. Contrast determines the variations in the grey levels of the images and how polarised are black and white distributions. Finally, the directionality coefficient aims to identify global properties within the images, such as pronounced curves or long lines.

6.5.4 Other Features

Along with the three types of features previously described, we chose to use six of the most popular low-level features developed to date with the aim of comparing their performance against pattern, colour and texture features. These features are have been used in a multitude of works [137, 158, 169, 196] and have been applied to problems such as image classification [137] and object recognition and image retrieval [169]. The main aim of extracting and testing these features is to further study the effect that feature selection has on habitat classification and to get a better understanding of what colour, texture and pattern features can do for a more accurate classification process.

The features we have extracted are: Geometric Blur (GB) [158], Global Image Descriptor (GIST) [137], Pyramid Histogram of Oriented Gradients (PHOG) [158], Scale-invariant Feature Transform (SIFT) [196], Pyramid Histogram of Visual Words (PHOW) [167], Self-similarity Feature (SSIM) [158].

6.6 Concluding Remarks

In this chapter we have introduced the notion of ground-taken photographs. Moreover, we have presented the second contribution of this thesis: the public and fully annotated datasets Habitat 1K and Habitat 3K. We have described their characteristics and limitations for the specific problem of Phase 1 classification and we have shown numerous visual examples. We have described how the annotation process works and how annotations are stored and manipulated. Finally, we have described the type of low-level visual features that will be used in our framework and the motivation behind their selection.

In the next chapter we will present the second element of our framework and our next contribution: Random Projection Forests. This Machine Learning classifier combines Random Projections and Random Forests and it is used to predict the habitats present in unseen photographs.

Chapter 7

Random Projection Forests

In Chapter 6, we introduced the type of source data we will work with and the first element of our framework: ground-taken photographs. Moreover, we described the type of low-level information that we extract from them in order to work with the photographs in an efficient and homogenous manner. In this chapter, we describe in detail how these data and these features are used in the context of automatic habitat classification. Consequently, we introduce the second element in our image-annotation framework and our third overall contribution: our Random-Projection-based classifier. This Machine Learning classifier, used to automatically annotate unseen ground-taken photographs, is referred to as Random Projections Forests (RPFs). RPFs are a modification of the traditional Random Forests as defined in [28]. They combine Random Forests and Random Projections, previously discussed in Chapter 2 as a dimension-reduction method, to automatically classify and annotate images more efficiently. We have carried out extensive experiments to assess the performance of Random Projection Forests in comparison to Random Forests for the task of automatic habitat classification. Moreover, we have studied the effects of pattern, colour and texture features on the classification process with both classifiers and both of our databases. Recall and precision results showed that Random Projection Forests are suitable candidates for our image-annotation framework and that they are more efficient and more accurate than RFs when automatically classifying Phase 1 habitats.

This chapter is divided into eight sections. Section 7.1 describes the motivation behind using Random Forests. Section 7.2 shows how traditional random forests are constructed and discusses it most relevant limitations. Section 7.3 presents our third contribution: Random Projection Forests. It describes in detail how random projections forests are constructed and discusses its advantages in comparison to traditional random forests. Section 7.4 describes how Random Projection Forests can be applied to automatic image

annotation problems in general and to the problem of habitat classification using ground-taken photographs in particular. Furthermore, Section 7.5 describes the first series of experiments that we carried out. The aim of these experiments was to test the effects of combining ground-taken images, low-level visual features, particularly pattern, colour and texture features, and random projection forests. Section 7.6 shows the results obtained from these experiments and compares them with traditional random forests. Moreover, it presents a discussion on the results obtained, with particular focus on the effects of colour, texture and pattern features when automatically classifying habitats. To conclude, Section 7.7 summarises the contents of the chapter.

7.1 Motivation: Limitations of NN-based Methods and SVMs

As mentioned in Chapter 2, there are multiple Machine Learning approaches that can be used for the task of image annotation and classification. Two of the most widely used currently are Nearest Neighbour methods and Support Vector Machines. Particularly, Nearest Neighbour (NN) methods have proven to be a popular choice in the Computer Vision community given its simplicity and its relatively non-existent training phase [18]. However, as shown in Chapter 4, NN-methods cannot be easily extended to use large amounts of data. Moreover, using NN-based methods to classify photographs presents a series of limitations in terms of efficiency. First, since NN methods require all training samples to be available during testing, the use of a large dataset would entail large storage requirements. Moreover, as the number of retrieved neighbours, represented by the parameter k, increases, the retrieval process will take more time. Finally, the combination of NN methods and feature extraction can negatively affect the "semantic gap" problem [90], since two objects might have similar visual properties, which might make them neighbours in the K-NN space, but they might belong to two completely different classes.

Other type of Machine Learning approaches that have been used are Support Vector Machines (SVMs). However, as discussed in Chapter 2, SVMs also present a set of limitations that make them unsuitable for the task of automatic image annotation. Firstly, SVMs are notoriously complicated to train, since they require fine tuning of a wide set of parameters. However, their main drawback is that they are single-label classifiers by nature. That is, SVMs are designed to return only one result. This, combined to their complicated nature, makes modifying them to be used in multi-label problems, such as habitat classification, a complicated and challenging task.

In this thesis, we use Random Forests to counter these limitations. Firstly, the hierarchical tree structure of the random forests allows for efficient classification of visually similar samples. Moreover, the use of the annotations stored in our ground-taken databases, Habitat 1K and Habitat 3K, can transform our classifier from an unsupervised to a supervised classifier. Consequently, the use of annotation information and their relationship with the photographs will guide the generation of the decision trees, which will make the decision process take into consideration not only visual features but semantic concepts as well. Additionally, Random Forests combine the simplicity of NN-based methods in terms of implementation and, more importantly, they can be easily modified to be applied in multi-label problems, such as Phase 1 classification.

Furthermore, previous work has shown that ensemble classifiers tend to obtain higher accuracy on previously unseen data [76]. Moreover, random forests have been successfully applied to a varied number of problems in the field of computer vision, such as image labeling [76], image classification [132] and even image segmentation [167]. They have also been applied to the field of Ecology, in tasks such as habitat structure classification [11], groundwater-dependent vegetation pattern modeling [144], ecohydrological modeling [143] and land cover [81].

In summary, in this thesis, we have chosen to research the use Random Forests because they present a promising alternative to the two most popular classifiers nowadays, NN-based methods and SVMs. Random Forests are able to combine their merits and lessen their limitations.

7.2 Random Forests

Random forests are composed of an ensemble of randomly trained decision trees. Decision trees have been used for quite a long time [149] with successful results in image classification tasks [22, 28, 81, 92, 147]. As shown in [28, 46], binary decision trees are composed by a collection of nodes and edges. These components follow a hierarchical structure in which there are no loops. Figure 7.1 shows an example of a binary decision tree with three levels.

As can be seen in Figure 7.1, nodes are usually numbered breadth-first, starting with the root node at 1. Moreover, trees have two different types of nodes: internal nodes, represented by circles, and terminal nodes, represented by squares. Internal nodes are also referred to as split nodes, because their function is to divide or split the received data into its children nodes. The root node is a special case of an internal node because it is were the data is injected into the classifier. On the other hand, terminal nodes are

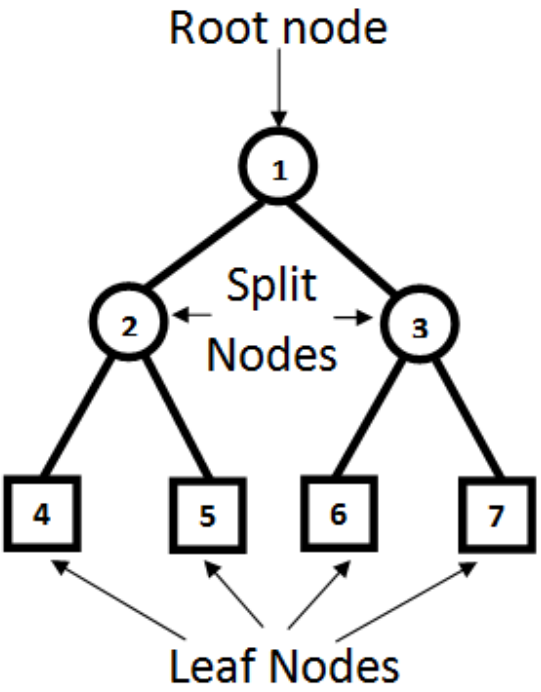


FIGURE 7.1: Decision Trees. Decision trees are composed of nodes and edges. Split nodes will separate the input data and leaf nodes will offer a prediction on the classes present within the data.

often referred to as leaf nodes. Split nodes, except the root have one incoming edge and, given that we are working with binary trees, two outgoing edges. Moreover, leaf nodes receive one incoming edge but do not produce any outgoing edges.

Decision trees are a Machine Learning technique used to make predictions on unseen data. These predictions are stored in the leaf nodes. As discussed in [46] decision trees can be regarded as a mechanism to iteratively split complex problems into a hierarchy of simpler ones. In turn, a Random Forest is a classifier which is composed of an ensemble of randomly trained decision trees. First introduced in [94] and further consolidated in [28], decision forests were shown to obtain better generalization than boosting and C4.5-trained trees on several tasks [95].

A random forest is defined by a series of parameters: its size, the maximum allowed tree depth, its type of randomness, the choice of weak learner model, the training objective function and the features selected. A variation in those parameters will affect the performance of the RF as a whole. However, this variation should not dramatically affect the performance of the RFs. That is, the aim is to generate stable RFs in which small variations of the input parameters should yield small variations in the results obtained. In order to do this, it is important to extract a significant group of features and to select an appropriate split function for the internal nodes.

Like many other Machine Learning techniques, constructing and using RF consists of two phases: the training phase and the testing phase.

- Training: This first phase is commonly carried out off-line and aims to generate stable RFs by optimizing the parameters of the split functions [46]. Traditionally, randomness is introduced in this phase. The two most popular methods to introduce randomness in the training phase are training data sampling [28], such as bagging, and randomised node optimization [94]. This guarantees that each decision tree will be randomly different from the other decision trees in the random forest. Additionally, each tree will stop being constructed when one of these two stopping criteria is met: the trees have reached their maximum allotted depth or the number of samples in the nodes is less than a threshold, commonly 1. As mentioned previously, each decision tree in the forest will be composed of two types of nodes: internal and leaf nodes.

  - Internal Node: The split nodes are in charge of dividing samples by optimizing the split function. This process has the following steps:

    1. For each Random Forest, a random number M between 1 and the maximum number of input features is selected. M indicates the number of splits that will be considered in each internal node. A large M will result in more accurate decision trees, since more splits will be tested. However, it will also require more computation resources.
    2. For each split node and until M random features have been selected, a random feature is chosen.
    3. For each selected random feature, the values of the samples related to that feature are extracted.
    4. A variable number of thresholds, L, is selected. Typically, threshold values will range between the minimum and the maximum feature value from the samples.
    5. For each possible threshold value T, samples are split into left or right child. This split is done following 7.1.

$$\begin{cases} p_i \geq L_j & \text{go to left child} \\ otherwise & \text{go to right child} \end{cases} \tag{7.1}$$

    Where p_i is the value of the selected feature in the i^{th} sample of the split node and L_j is the j^{th} threshold taken into consideration.

    6. Once all samples have been divided into right or left child node, the Information Gain of that split is calculated. The IG assesses which split produces the highest confidence in the final distributions [46]. It is

calculated as:

$$I = H(S) - \sum_{i \in \{1,2\}} \frac{|S^i|}{|S|} H(S^i) \tag{7.2}$$

with S being the set of randomly selected features and $H(S)$ being the Shannon entropy [28].

7. Steps 2 to 6 are repeated M times.
8. The feature-and-threshold combination that produced the split with the highest IG is then selected. Samples are split into left and right child nodes according to this combination.

The calculation of the Information Gain is not trivial and, depending on the parameters of the Random Forest, it will have to be repeated a large number of times. The larger the dimension of the feature vector and the larger the number of features activated in the calculation of the IG, the less efficient RFs become and the longer the training phase will take. As a result, the process of training a Random Forest can be computationally expensive. In a Random Forest in which M features will be selected in each split node and in which L thresholds will be tested, for each decision tree T, with N split nodes, the calculation of the IG will have to be repeated $MxLxTxN$ times. For example, in a Random Forest with 150 trees of depth 9 (512 nodes, 264 of those split nodes) in which 10 thresholds and 50 random splits are considered, a small number considering that feature vectors can have thousands of values, the IG will have to be calculated 19,800,000 times.

– Leaf Node: The leaf nodes will learn a prediction during training. In classification tasks, each leaf will store the normalised probability distribution of each class, or habitat in our case, according to the samples that have reached that leaf. Consequently, if we apply it to our case the probability in each leaf l is calculated as

$$P^{T_k}(h) = \frac{|h_i|}{|h_k|} \tag{7.3}$$

with T_k being the kth decision tree in the random forest T, $|h_i|$ being the frequency of the hth habitat in the ith leaf node and $|h_k|$ being the frequency of appearance of the habitat h in the leaves of the kth decision tree.

- Testing: The aim of this phase is to give a prediction about previously unseen images. Contrary to the training phase, the testing phase does not include randomness of any kind, which makes it completely deterministic. In this phase, the features extracted from the unseen data are injected into the root node of each of the trees in the forest. These features are then propagated through the internal nodes in each tree. At each split node, the split function is applied to the incoming

set of features and, depending on the result, they are directed towards the left or right child node. This process is repeated until a leaf node is reached. Since we are working with an ensemble of decision trees, each tree will offer a prediction. Finally, the predictions will be combined into one single prediction using a the voting mechanism. As discussed in Chapter 2, there are many voting mechanisms that have been developed to date. The most common method is to linearly combine and normalise the predictions of each tree. Therefore, each tree in the random forest will cast a unit vote. However, as we will explore in Chapter 9, research has proven that not all the decision trees in a random forest are equally good at classifying unseen data. Consequently, it is possible to implement a voting mechanism that will assign weight to the different predictions before linearly combining them in order to improve accuracy.

7.3 Random Projection Forests

The traditional implementation of Random Forests presents some limitations when the dimensions of its basic parameters, i.e. size, depth and number of randomly selected features in each node, increase. Particularly, increasing the random number of features taken into consideration in each node can be quite time-consuming when the feature vector dimensions' increase. In order to fix these limitations, we have created Random Projection Forests (RPFs).

RPFs are the third contribution of this thesis. They were designed to be more efficient and accurate than traditional RF. RPFs are more efficient than RFs in terms of execution time during training and testing, as will be shown in Section 7.6, particularly when increasing two of its parameters: the size of the forest and the number of random features to be taken into consideration in the split nodes.

In Random Projection Forests, randomness is introduced in two ways. First, we use different random subsets of the training data to train different decision trees, referred to as bootstrapping [18]. Then, we use Random Projections [101] to reduce the dimensionality of the feature vectors. Random Projections have been used in conjunction with Random Forests in [103]. However, [103] follow a simple approach by projecting the input feature vectors before training traditional Random Forests. This choice is not ideal, since it limits the effect of the randomness that Random Projections could infuse Random Forests with and, consequently, does not benefit from Random Projections as much as they could. In our case, we generate a random projection in each internal node and we use it to project the samples that reach said node. Similarly to traditional

random forests, RPFs are composed of two different types of nodes: split nodes and leaf nodes.

As with Random Forests, the input of RPFs are the annotations of our database and the feature vectors extracted from the photographs themselves. Each forest is generated by training each binary decision tree breadth-wise until one of the stopping criteria is met. Similarly to the stopping criteria introduced in Section 7.2, RPFs will stop being constructed when the number of samples that reach a node is 1 or when the tree has reached its maximum allowed depth.

- Split nodes : These nodes store a test function that splits the data. As mentioned in the previous section, during training, the aim is to optimise the threshold of the split functions in each node so the trees can be as accurate as possible [46]. Our approach is based on random projections [17], previously discussed in Chapter 2. Random Projections are a dimensionality reduction mechanism that enables us to project large feature vectors into scalar values using orthogonal vectors.

  In our case, we use random projections to split incoming samples of an internal node to its two child nodes. Let $\mathbf{F} = (f_1, f_2, \ldots, f_n)$ be the n-dimensional input feature vector of a node, $\mathbf{R} = (r_1, r_2, \ldots, r_n)$ be an n-dimensional random vector, generated as follows

$$r_i = \begin{cases} -1 & \text{with probability } \frac{1}{3} \\ 0 & \text{with probability } \frac{1}{3} \\ +1 & \text{with probability } \frac{1}{3} \end{cases} \tag{7.4}$$

  with $i = 1, 2, ..., n$.

  We then project the input onto the random vector. This is done by calculating the inner product between the feature vector $\mathbf{F}$ and the random projection vector $\mathbf{R}$ as $p = \mathbf{F}\mathbf{R}^T$. Once the feature vector has been projected, each feature vector is reduced to a single scalar value, and samples are distributed to the left or the right child node according to a threshold as:

$$\begin{cases} p \geq T & \text{go to left child} \\ otherwise & \text{go to right child} \end{cases} \tag{7.5}$$

  where T is a threshold value.

  As can be seen, each feature vector, once projected, will be reduced to only one scalar value, the projection itself. This makes our RPFs much more efficient than traditional RFs. Since the projected feature vectors are simple scalar values, the

calculation of the threshold value is quite simple. After projecting all the samples that have reached an internal node, we generate a user-input number of equidistant thresholds, 10 by default, that range from the minimum projection to the maximum. Then, we select the threshold that maximises the Information Gain (IG). The IG, which can be calculated as shown in Equation 7.2, is then used to select the split function which produces the highest information gain in the final distributions [28].

The computational requirements needed to train a Random Projection Forest are much smaller than those required to train a Random Forest. Instead of considering M splits in each internal node, all samples are projected into one scalar value, an operation that only requires a multiplication. Moreover, in a Random Projection Forest in which L thresholds will be tested, for each random-projection decision tree T with N split nodes, the IG will be calculated $LxTxN$ times. Following the example introduced in Section 7.2, in a Random Projection Forest with 150 trees of depth 9 (512 nodes, 264 of those split nodes) in which 10 thresholds are considered, the IG will have to be calculated 396,000 times. That is 19,404,000 less IG calculations than in the corresponding scenario with Random Forest.

- Leaf nodes: At this stage, the leaf nodes are the same as those of traditional RFs. In our case, they store a normalised probability distribution of the occurrence of all possible habitats. This probability is calculated as shown in Equation 7.3.

The whole procedure of building a random projection decision tree is summarised in Algorithm 1. The pseudocode describing how to build a Random Projection Forest is shown in 2.

7.4 Random Projection Forests For Image Annotation

We have designed Random Projection Forests with the aim of applying them to automatically annotate unseen ground-taken photographs with the habitats present in them. In Section 7.2 we described how RPF are constructed, or, in other words, their training phase. In this section we describe how they can be applied to Image Annotation or, alternatively, their testing phase.

The testing procedure for RPFs is similar to the that of the traditional RFs. Once features are extracted from the unseen test image, these are injected in each of the root nodes of the projection trees that form the RPF. At each split node, the feature vector will be projected by calculating the inner product between the feature vector and that

Algorithm 1 Random Projection Decision Tree: Training. *Thresholds* is an integer that indicates the number of thresholds that will be tested with the Information Gain

```
procedure TRAIN_PROJECTION_DECISION_TREE(depth, features, thresholds)
    no_nodes ← 2^depth − 1
    rpdt = initialise_forest(no_nodes, features)        ▷ Initialises tree and root
    for n = 1 to no_nodes do
        samples = rpdt(n).features                      ▷ Incoming samples
        if n < no_nodesandsize(samples) > 1 then        ▷ n is a split node
            rpdt(n).rp = calculate_random_projection()  ▷ RP assigned to node n
            rpdt(n).p = rpdt(n).rp ∗ rpdt(n).features'  ▷ Features are projected
            rpdt(n).max_threshold = calculate_IG_maximum(rpdt(n), thresholds)
            divide_samples(rpdt(n))   ▷ Divides samples according to max_threshold
        else                                            ▷ n is a leaf node
            calculate_tree_probabilities(rpdt)
        end if
    end for
     return rpdt
end procedure
```

Algorithm 2 Random Projection Forests: Training.

```
Input: size,depth, samples, thresholds
Output: forest
  for i = 1 to size do
      features = calculate_bootstrap_sample(samples)
      forest(i) = train_projection_decision_tree(depth, features, thresholds)
  end for
  calculate_forest_probabilities(forest)
       return forest
```

particular node's random projection vector. Then, the feature vector will be propagated to either the left or the right child node according to the result of the comparison between the projected vector and the threshold, as shown in Equation 7.5. This process will be repeated until the feature vector reaches a leaf node in each of the trees in the forest. In this implementation of RPFs, each tree will cast a unit vote. Consequently, the predictions of each tree in a RPF of size N will be linearly combined and then normalised, as shown in 7.6.

$$P(h) = \frac{1}{N} \sum_{t=1}^{N} P^{T_t}(h) \tag{7.6}$$

where $P(h)$ is the probability of occurrence of the habitat h in the unseen photograph and $P^{T_t}(h)$ is the probability of occurrence of the habitat h according to the decision tree t.

7.5 Experiments

A series of experiments was carried out to evaluate the use of ground-taken photographs, random projection forests and low-level visual features when applied to automatic habitat classification.

We set up these experiments with the goal of studying the effects on the performance of an specific set of parameters. Moreover, we decided to compare Random Projection Forests against traditional Random Forests to obtain a more in-depth study of its effects. These parameters are:

- Depth variations: As mentioned previously in this chapter, stability is a crucial trait in Random Forests. In order to measure how stable our RPFs are, we carried out a small experiment, in which we compared results obtained using RPFs and RF with depths varying from 5 to 10.
- Input: Once the depth is set, we study the results obtained by varying the input of our framework. To do this, we use three different types of input. Figure 7.2 shows the differences in input information in each case.

  These three categories are:

  - Whole images: Features are extracted from the photographs as a whole. Consequently, each photograph in our database produces one feature vector.
  - Annotation Segments: Features are extracted from each different annotated polygon within a photograph. Consequently, each photograph in our database will produce a variable number of feature vectors, depending on the number of habitats present in it.
  - Blocks: The ground-taken photographs are divided in square blocks of varying sizes and features are extracted per tile. The size of these tiles are 64 and 1024 pixels. Consequently, we will obtain 1974 and 24 feature vectors per image, respectively.

- Colour, pattern and texture features: Human surveyors will normally take into account colour, pattern and texture information in their classification. Consequently, we are extremely interested in studying if these features in particular can also be applied in our automatic system. In order to do this, we will compare performances of these features versus the performance of the "Other Features" presented previously Chapter 6. The features extracted were previously described in Chapter 6. These are: colour features (Color Histogram and Color Moments),

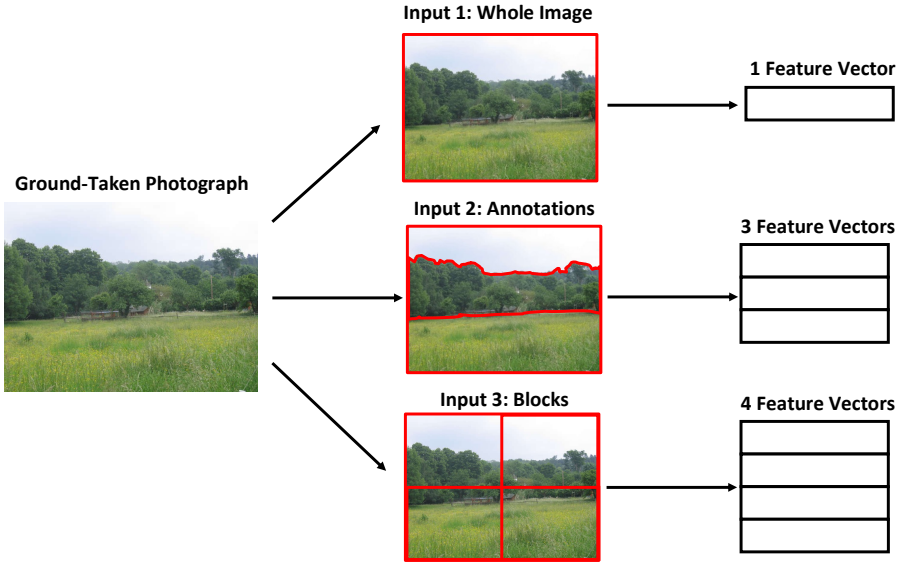


FIGURE 7.2: Input Feature Vectors For The Classifiers. Each input type generates a different number of feature vectors per photograph.

texture features (Tamura Coefficients and GLMC), pattern features (CPAM) and a combination of six of the most common visual features currently used in Computer Vision problems (GB, GIST, SIFT, SSI, PHOW, PHOG).

- Database: Given the different nature of the databases created in this thesis, Habitat 1K being collected under controlled circumstances and Habitat 3K being collected using crow-sourcing methods, we also aim to study their performance when the input data and the features extracted are modified. Moreover we aim to study the effect that increasing the number of habitats presents and the number of instances of each habitat also results in improved results.

In essence, the experiments helped us determine the best configuration of these parameters to obtain an equilibrium between accuracy and efficiency when automatically classifying habitats.

7.5.1 Performance Metrics

In order to assess the performance of RPFs and low-level visual features when automatically classifying habitats, two separate metrics were calculated: recall and precision.

Firstly, the recall and precision have been measured using the implementation proposed in [216]. Let N_h be the number of photographs in the test set whose habitats are correctly labeled by an expert and are part of our ground-truth. Let N_{sys} be the number of photographs that are suggested for each habitat in our system, and N_c the number of images whose habitats our system correctly suggests. The precision and recall are defined as shown in Equation 7.7 and Equation 7.8.

$$recall(w) = N_c/N_h \tag{7.7}$$

and

$$precision(w) = N_c/N_{sys} \tag{7.8}$$

Moreover, in order to measure the robustness and the performance of our approach, in all experiment scenarios in this chapter and following chapters, the database was randomly divided ten different times. Each time the training set contains $\frac{2}{3}$ of the photographs and a test set contains the rest of the images. Therefore, the recall and precision results shown in the next sections are an average of the results obtained with the ten randomly-generated training and testing sets.

7.6 Results

Before starting our series of experiments to assess the effects of the low-level features and our databases as previously described, we had to test first two crucial aspects of our RPFs: their efficiency and their stability. RPFs would be considered efficient if their training and testing execution times were better than RFs executions times. Moreover, they would be considered stable if small changes in some parameters, particularly the sizes and the depth of the trees, produced only small changes in the performance of the forest.

First, in order to asses the efficiency of RPFs, we calculated the execution times of training forests of sizes from 1 to 150. We compared these results with those obtained from using RF. In this case, colour, texture and pattern features from the images as a whole were used as the input of both sets of forests. Moreover, we trained ten sets of forests and calculated the average execution times. Additionally, in the case of RFs, we took into consideration $\frac{2}{3}$ of the features extracted. The choice to select $\frac{2}{3}$ of the features was not random: we chose this particular number because it is the same amount

of features that are projected with our approach. When we project the feature vector, only $\frac{1}{3}$ of the features will be ignored, since they are projected with the value 0. The other $\frac{2}{3}$ of the features will be projected with the values 1 and -1. Consequently, we are comparing efficiency when the same number of features are activated in the split nodes.

Experiments showed that our RPFs performed equally or more efficiently in all cases. When the sizes of the forests was relatively small, less than 20 in general, both approaches would take similar times. However, once the number of trees in the forest increased, RPFs would take less time to generate. This is consistent with the operations taken into account in each split node. RPFs only require an arithmetic operation, a multiplication, while RFs will test several sets of random features to find the configuration with a higher Information Gain. Table 7.1 shows a particular example of this. To make the visualization easier, Table 7.1 show the average execution times of RFs and RPFs with trees of depth 9. As can be seen, execution times are similar, with a difference of less than 0.1 seconds in favour of RFs, when the size of the forest is small but, as it increases, RPFs take less time to train its forest, even reaching a difference of over 4.5 seconds.

TABLE 7.1: Average Execution Times. These results were obtained training Random Forests and Random Projection Forests of deph 9 and with a varying size between 1 and 150.

	Execution Time (s)	
Size	RF	RPF
1	0.5460	0.5772
10	4.7892	4.8360
20	9.6721	9.4069
30	14.6329	12.4785
40	19.2193	18.0181
50	24.5078	21.6513
60	29.2502	27.3158
70	33.7586	31.6682
80	38.7974	36.8942
90	43.4463	40.2327
100	48.7659	45.2403
110	54.4599	50.5599
120	59.0308	55.0528
130	63.2740	58.5316
140	68.6872	63.1180
150	72.6965	68.1724

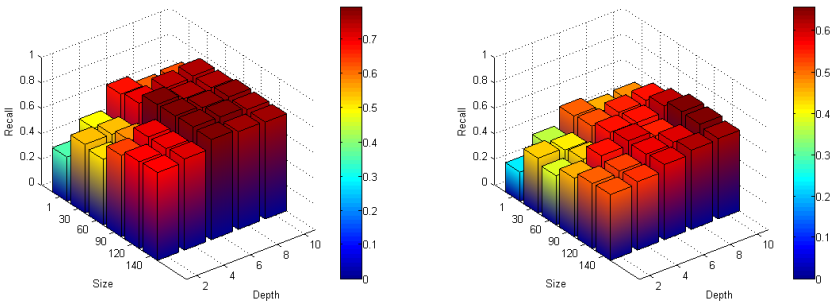
In order to obtain more information about the stability and performance of our framework, we also compared our Random Projection Forests with Random Output Space Projections (ROP) [103], which used random projections of the output to train random forests. We calculated execution times, recall and precision using pattern, colour, texture and all features together with trees with depth from 2 to 10 and with forests with sizes

from 1 to 150. Table 7.2 shows the average execution time, recall and precision of all scenarios. As can be seen, Random Output Space Projections had a much more efficient execution time than both Random Forests and Random Projection Forests. However, their performance when classifying habitats, which are extremely visually similar, was considerably less accurate, particularly in terms of precision. For this reason, we decided to keep using Random Projections in each of our nodes instead that in the input only and we also decided to compare our framework with traditional Random Forests in terms of recall and precision.

TABLE 7.2: Execution time in seconds, recall and precision averages of Random Forests (RF), Random Output Space Projections and Random Forests (ROP) [103] and Random Projections Forests (RPF).

	RF	ROP	RPF
Time	36.595	25.196	33.984
Recall	0.313	0.21	0.408
Precision	0.26	0.12	0.265

Second, in order to assess the stability of our RPFs we decided to test the approach and study its average recall results for first-tier habitats when the depth of the trees ranged from 2 to 10 and the size of the trees varied from 1 to 140. Following the same configuration as in the previous experiment, texture, pattern and colour features from the images as a whole were used to obtain the recall of our approach when the size and depth of our RPF were varied. Moreover, to get a better understanding of their stability, we compared our results with those results obtained from using RFs under the same circumstances. Figure 7.3 shows the results in the particular case of habitats of class A (Woodland and Scrub). As can be seen, RPFs are considerably stable, as RFs, since small changes in size and depth result in small changes in the results.



(a) RPF (b) RF

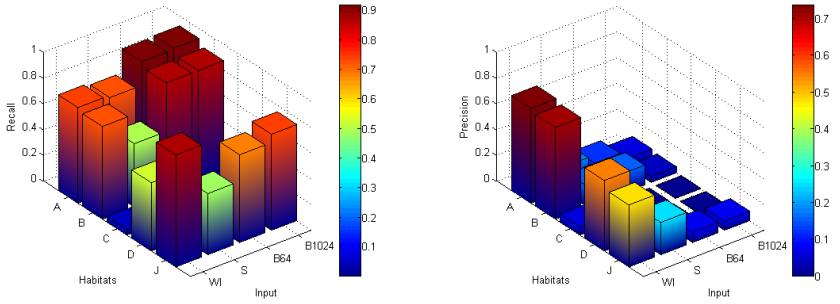
FIGURE 7.3: Stability of RPFs and RFs. We show the recall when classifying Woodland and Scrub (A) habitats with Habitat 1K.

It is important to notice that, given the results from the previous experiments and, in order to present the remaining results in a more compact and comprehensive manner,

we decided to choose a fixed depth of 9 for the trees. That is, each tree in the RPFs or in the RFs will be composed of 512 nodes, unless the opposite is stated.

The two previous experiments led to us conclude that RPFs were a viable alternative to RFs. They are stable and have better execution times than traditional Random Forests. The next logical step was to begin testing the performance of the classifier itself when annotating habitat classes with more depth.

However, before starting to annotate unseen test samples, we decided to complete one more experiment in which we studied the effect of different types of input photographs. We decided to test this in order to find the best configuration for the rest of the experiments in terms of source data. As mentioned previously, we contemplated three cases: the whole photograph as an input, using the polygon annotations and using blocks within the images. Moreover, we tested both first and second-tier habitats using our Habitat 1K database. Figure 7.4 shows the results obtained for first-tier habitats. We show the precision and recall results we obtained for all three cases when the depth of the trees was set at 9 and the size of the forests was 150.



(a) Recall (b) Precision

FIGURE 7.4: Effect of Input in RPFs. Results show that using the Whole Image (WI) obtains better results than using Segmented Annotations (S) and square Blocks of 64 (B64) or 1024 (B1024) and pixels.

Results were consistent thorough all scenarios and, contrary to our preliminary thoughts, using the whole image as the input yielded the best results both in terms of efficiency and accuracy. This is particularly noticeable in the precision results shown in Figure 7.4. As can be seen, using the whole image as the input yielded the best precision results in all testing cases.

In particular, tiles proved to be very inaccurate when classifying habitats in terms of precision. While their recall was more accurate than the use of whole images for Woodland and Scrub (A) and Grassland and Marsh (B) habitats, their precision results in general were quite low, reaching less that 1.5% in the cases of Grassland and Marsh

(B), Tall Herb and Fern (C) and Heathland (D). Moreover, 64x64 tiles were specially inaccurate when classifying second-tier complex habitats, such as Hedges and Trees (J.2.3). This is not surprising, since Hedges and Trees (J.2.3) and Heathland mosaics (D) are composed of several types of vegetation that belong to other habitats as well. Correspondingly, if a block only represented a small portion of one type of vegetation that could be part of several habitats, such as acid grass being part of Acid Grassland habitats, Dry Heath/Acid Grassland mosaics and Dry Heath/Scrub mosaics, classifying it correctly becomes virtually impossible when only taking into consideration low-level visual features. This was mainly because the small size of the tiles was insufficient for the information of such habitats to be collected. Moreover, using 64x64 tiles proved to be a challenge, since the training phase became less efficient. The use of tiles entailed that 1,974 tiles were generated by each photograph, with a total of 2,143,764 tiles in Habitat 1K, 714,588 of those used for testing and 1,429,176 used for training.

However, as can be seen, when the size of the blocks increased from 64x64pixels to 1024x1024 pixels, recall results improved considerably. This is consistent with the results obtained from the whole images: the larger the area we extract features from, the more accurate the results. 1024x1024 pixel tiles divide the photographs from our Habitat 1K database in twenty-four tiles, which were large enough to contain more information, such as the combination of several simple habitats to create a complex habitat.

On the other hand, the use of the annotated polygons yielded better results than the use of blocks and, as shown in Figure 7.4 for the case of Tall Herb and Fern (C), even better results than using the whole images. However, their precision results were lower than using the whole image. Moreover, the trade-off between accuracy and efficiency was not good enough to choose annotated polygons in further experiments. That is, the recall improvement over the results of using whole images was small even though the computation resources required were larger.

Using the whole photographs as the input entailed a much faster training phase and resulted in the majority of the most accurate results. Moreover, complex habitats obtained better recall and precision results, since all the information within the images was taken into consideration. Consequently, due to their balance between efficiency and accuracy, we decided to continue our experiments using the photographs as a whole.

Once this experiment was finished and the whole images were chosen as the most suitable candidates for the input of our system, we decided to test our framework's performance when classifying habitats present in unseen ground-taken photographs. As mentioned previously, we were particularly interested in studying the effects that low-level features had in relation to our two datasets, Habitat 1K and Habitat 3K, and in further comparing the performance of RFs with RPFs.

In order to assess the effect of low-level visual features, Random Forests and Random Projections Forests, we have tested ten scenarios with each or our databases. These scenarios are:

1. RPF with colour features. This scenario is referred to as RPF - Color in the figures.
2. RPF with pattern features. We refer to this as RPF - Pattern in the figures.
3. RPF with texture. This is called RPF - Texture in the figures.
4. RPF with all three features linearly combined. This scenario is referred to as RPF - All in the figures.
5. RPF with other features. In order to make visualization easier, we have not included these results in the graphs. However, the findings from this set of experiments will be commented and compared with the results obtained in the other experiments.
6. RF with colour features. This scenario is referred to as RF - Color in the figures.
7. RF with pattern features. We refer to this as RF - Pattern in the figures.
8. RF with texture features. This is referred to as RF - Texture in the figures.
9. RF with all three features linearly combined. In order to make visualization easier, these results are not included in the graphs. However, the findings from this set of experiments and how they compare with the other feature combinations will be discussed.

Moreover, we divided the results obtained according to the hierarchical structure of Phase 1. Consequently, first we calculated recall and precision results for first-tier habitats and then for second- and, in some cases third-, tiers. We have divided these results into two additional sections: Section 7.6.1 presents results obtained when only classifying first-tier habitats and Section 7.6.2 presents results obtained when looking into second- and third-tier classes. Finally, we present some visual examples obtained during our testing in Section 7.6.3.

7.6.1 First-Tier Classes

Figure 7.5 shows the recall and precision results obtained in the testing scenarios introduced previously when using features extracted from whole images from Habitat 1K, referred to H1K from now on, as the input. Additionally, Figure 7.6 shows the same metrics when testing our framework with features extracted from whole photographs

from Habitat 3K, referred to as H3K, as the input. We tested forests with sizes ranging from 1 to 150 and with depths ranging from 2 to 10. However, in order to present the results in a clear and concise manner, we set their depth to 9 in the previous figures. Nevertheless, the performance of both systems was similar and stable in all cases.

Looking at the graphs as a whole, it can be noticed that recall results are higher than precision, regardless of the approach followed and the features extracted, with Miscellaneous (J) being the only exception. This is consistent with the more relaxed nature of the recall measure in comparison with precision, as mentioned in Section 7.5.1. Moreover, at a broad glance, it is clear that Woodland and Scrub (A), Grassland and Marsh (B) and Miscellaenous (J) are the most successfully classified habitats with H1K, while H3K obtains higher accuracy for Woodland and Scrub (A) and Open Water (G) habitats. This is due to two main reasons: in both cases, both successfully classified habitats are either the habitats with most instances (Woodland and Scrub, Grassland and Marsh) or they are very visually different from the rest of the habitats present in the database (Open Water). The first reason entails that the habitats are presented under many different circumstances and conditions and the second reason makes those habitats stand out from the other habitats present in the database, therefore their classification is more straightforward.

On the other hand, Tall Herb and Fern (C) and Heathland (D) are the most challenging habitats to classify when using H1K and Tall Herb and Fern (C) and Rock Exposures and Waste (I) obtain the least accurate results when using H3K. Following the ideas discussed previously, this is to be expected. In both cases, both habitats are the classes with the least instances in the database. For example, in H1K, Grassland and Marsh have 1008 instances versus a mere 95 instances collected from Tall Herb and Fern habitats. Likewise, Woodland and Scrub have 2243 instances in H3K, with Rock Exposure and Waste having only 145. The effect that the number of instances has on the performance is also exemplified by the behaviour of Heathland habitats (D) in H1K and H3K. Their classification in general improves greatly in H3K given their much larger number of instances in the database, 824 in H3K against 135 in H1K.

Moreover, similar visual properties between habitats also entail a lower performance. All inaccurately classified habitats can easily be confused with other habitats. Tall herb and Fern (C) can be easily mistaken for Scrub (A.4), and Rock Exposure and Waste (I) share many similarities with Coastland (H) habitats, particularly Maritime Cliffs (H.2). This last case is what produces such unstable precision results when classifying Coastland (H) habitats, as shown in Figure 7.6.

The impact of visual similarity in the classification process is further shown in Table 7.3 and Table 7.4, which present the confusion matrices for all first-tier habitats for both

H1K and H3K when classified using only colour features. In order to obtain a more accurate resuls, we used the annotations as the input of the classifier in each case and we took into consideration only the most probable result obtained with our framework. In Table 7.3 and Table 7.4, each row represents the habitat of the annotation and each column represents the most probable annotation that our system predicted for that case. Ideally, the matrix would be a diagonal matrix, in which each habitat is correctly classified. However, as can be seen, similarities in visual properties result in common misclassifications. This is clear large number of cases in which Tall Herb and Fern (C) and Heathland (D) are misclassified as Woodland and Scrub (A) and Grassland and Marsh (B) due to their similar visual characteristics of both habitats with Scrub (A.4) habitats.

If we look into the experiments more in depth, we can also find quite interesting results. First of all, in general and regardless of the low-level features taken into consideration, Random Projections Forests is able to outperform Random Forests when classifying first-tier habitats in the majority of the cases. Moreover, this improvement is specially noticeable and particularly important when looking at the systems' precision. An example of this is the case of Woodland and Scrub (A) habitat, in which the precision of using RPFs clearly surpasses RFs in both H1K and H3K. These results, combined with the results obtained previously regarding the efficiency and stability of both approaches, serve to illustrate the validity of RPFs and their applicability not only to habitat classification, but also, potentially, to other classification tasks.

Another interesting result can be seen when comparing the different types of features extracted. As we previously discussed, the pattern features we have chosen [148] combine both color and pattern texture information. Consequently, they were the best candidates to extract information from the images in a compact and descriptive manner. As a result, the fact that they obtain most of higher recall and precision measures is not a surprise and supports our decision of having chosen them as guidelines to study the initial performance of our system. Moreover, it is interesting to notice that these pattern features perform equally well with both classifiers, RPFs and RPs, and they generally generate the best set of results obtained with each classifier.

However, what is more intriguing is the performance that texture features have obtained in our whole system. Initially, we regarded texture features as as as useful and informative as colour features. This is in part supported by its recall results when used with H3K, where they perform slightly less accurately than our other features. In the case of Open Water (G), they even obtain one of the most accurate recall measures, close to 98%. Nevertheless, it is their performance in terms of precision in all cases what clearly

Table 7.3: Confusion Matrix of H1K and RPFs trained with colour features. Correct classification percentages are shown in bold, while common misclassification scenarios are shown in italics.

	A	B	C	D	E	F	G	H	I	J
A	**62.65%**	4.08%	16.33%	4.69%	0%	0%	0%	0%	0%	12.24%
B	8.33%	**66.77%**	12.50%	11.41%	0%	0%	0%	0%	0%	0.99%
C	*44.21%*	13.68%	**1.05%**	*35.79%*	0%	0%	0%	0%	0%	5.26%
D	*34.8%*	22.22%	24.441%	**2.22%**	0%	0%	0%	0%	0%	16.30%
E	0%	0%	0%	0%	**0%**	0%	0%	0%	0%	100.00%
F	0%	0%	0%	0%	0%	**0%**	0%	0%	0%	0%
G	0%	0%	0%	0%	0%	0%	**33.33%**	0%	0%	*66.67%*
H	0%	0%	0%	0%	0%	0%	0%	**0%**	0%	0%
I	0%	0%	0%	0%	0%	0%	0%	0%	**0%**	0%
J	4.08%	4.98%	0.60%	2.01%	0%	0%	0%	0.00%	0.00%	**88.32%**

TABLE 7.4: Confusion Matrix of H3K and RPFs trained with colour features. Correct classification percentages are shown in bold, while common misclassification scenarios are shown in italics.

	A	B	C	D	E	F	G	H	I	J
A	**52.92%**	9.99%	17.83%	13.73%	0%	0%	0%	0%	0%	5.53%
B	5.10%	**52.14%**	5.35%	18.37%	0%	0%	0%	0%	0%	19.04%
C	*42.57%*	6.60%	**3.63%**	*34.32%*	0%	0%	0%	0%	0%	12.87%
D	29.98%	*43.81%*	0.97%	**5.83%**	0%	0%	0%	0%	4%	15.41%
E	0%	0%	0%	0%	**0%**	0%	0%	0%	0%	100.00%
F	0%	0%	0%	0%	0%	**0%**	0%	0%	0%	100%
G	0%	3%	0%	4%	0%	0%	**35.55%**	0%	22%	35.66%
H	6%	11%	0%	0%	0%	0%	0%	**3%**	37%	43%
I	8%	0%	0%	3%	0%	0%	0%	67%	**0%**	22%
J	4.11%	2.66%	0.53%	1.01%	0%	0%	0%	0.00%	0.16%	**91.53%**

implies that texture alone is not descriptive enough to accurately classify first-tier habitats. This is exemplified in the classification of Woodland and Scrub (A), Grassland and Marsh (B) and Miscellaneous (F) habitats for both H1K and H3K, in which they obtain up to 20% less accuracy than any other method. Moreover, this inaccurate performance independent from the classifier used, as texture features obtain the less accurate results with both RFs and RPFs.

The performance of texture features is particularly striking when it is compared with the results obtained from using colour features. Even though the colour features we have extracted, colour histograms and colour moments, are quite simple, their performance is undoubtedly better than the performance of the texture features. Moreover, colour features are more stable, with the differences between its recall and precision results not being as abysmal. In some cases, such as the classification of Tall Herb and Fern (C) in H1K or Woodland and Scrub (A) both in H1K and H3K, they can even outperform pattern features. In summary, this dissonance in performances between texture features and colour and pattern features serves to emphasise the importance of colour information in the classification process. Moreover, it has helped determine that, contrary to pattern and colour features, texture alone is not a suitable candidate for habitat classification.

Another important point is given by the performance obtained by uniting the colour, texture and pattern features. Instead of increasing dramatically the results with our system, its results are generally worse than those obtained using pattern, and sometimes colour, features alone. This supports the idea the combination of multiple features might not be the best solution to a classification problem such as ours, not only because the training phase will be more computationally expensive, but also because the results obtained might not be the most accurate.

This notion is also supported by the results obtained from uniting the "Other" visual features. These results are not in the graphs to make the visualization of the most relevant features easier. However, their performance was comparable to the lowest performance of the texture features, with recall not surpassing 30% and with a precision of less than 20% accuracy with their best configuration, which was found when classifying Woodland and Scrub (A) with RPFs. Consequently, these features are less accurate than colour, pattern in all cases. These results also help stress the crucial significance of feature selection and its problem-dependent nature. As can be seen, regardless of the classifier used or the database chosen, SIFT, GIST, GB, PHOG, PHOW and SSI features are unsuited for the task of habitat classification.

Finally, we can also compare results obtained when using Habitat 1K and Habitat 3K. As can be seen, as the number of instances in our databases grow, so do the general recall ability and precision performance of the system. This is consistent with the image

annotation framework we have created, in which the more robust the input database and the more significant the features extracted, the more accurate our classification should be. As mentioned previously, this is clearly visible in the case of Heathland (D) habitats. Its precision in particular improves dramatically, from around 10% precision at most in H1K to almost 30% in the case of RPF and pattern features.

However, it is also important to comment that the results obtained in this section, while a starting point, are far from perfect. Even though recall is high, with more than 50% recall in four out of the five habitats collected in H1K and four out of the seven habitats stored in H3K, it is also very low, less than 10%, for the rest of the habitats. Moreover, precision results are similar or, in the case of H1K, even lower. This is not surprising, since as we mentioned previously, the use of low-level visual features only in a FGVC problem such as habitat classification, in which classes are extremely visually similar, will entail a loss of information, particularly semantic information, that could be crucial in the classification process. As a direct consequence of this, we can only expect these results to be less accurate when classifying second- and third-tier habitats, since the similarities between classes on these levels are even more pronounced.

7.6.2 Second-Tier and Third-Tier Classes

Figure 7.7 shows the recall and precision results obtained in the same testing scenarios as the previous section when using features extracted from whole photographs from H1K as the input. Additionally, Figure 7.8 shows the same metrics when testing our framework with features extracted from whole photographs from H3K. Testing was done varying the size of the forests between 1 and 150 and the depth was varied between 2 and 10. However, as can be seen in the previous figures, we have set the size of the forests to 120 and the depth of the forests to 9. Since the performance of both systems was similar and stable in all cases, this was done in order to make the visualization of the results easier.

Looking at all the graphs as a whole, we can see that, similarly to the classification of first-tier habitats, the recall metric is more accurate than the precision measure in all cases. Moreover, as projected in the previous section, both metrics have experimented a significant decrease in accuracy. Precision metrics, in particular, are the most affected. This is to be expected, since we are only extracting visual information while, at the same time, trying to classify classes which are extremely visually similar.

Notably, precision for habitats within the classes Tall Herb and Fern (C) and Heathland (D) in H1K are particularly inaccurate. Moreover, second-tier habitats from Tall Herb and Fern (C), Heathland (D), Coastland (H), Rock Exposure and Waste (I) obtain the

lowest precision when using H3K as the input. This is consistent with the comments from the previous section in which we discussed the limitations that visual features have when classifying FGVC classes.

Additionally, it can also be appreciated that complex and artificial habitats obtain particularly low precision and a average recall results. Heathland mosaics (D.1. and D.2.) and Mixed Woodland (A.2) obtain some of the lowest precision results, even reaching 0% in some cases. Fence (J.2.4) habitats experiment similar results. Nevertheless, Hedges and Trees (J.2.1), another complex habitat, obtains quite good recall results with RPFs and texture or pattern features but generate a precision close to 1%.

On the other hand, the recall and precision of the two classes with more instances in our databases, Woodland and Scrub (A) and Grassland and Marsh (B) do not experiment such a dramatic decrease between recall and precision. Broad-leaved Woodland (A.1) and Acid (B.1) and Neutral (B.2) Grassland obtain the highest recall results in H1K and in H3K. As mentioned in the previous section, this is mainly due to the fact that both databases have a larger number of these habitat classes in them.

Moreover, following the previous trend, it can be seen that RPFs keep outperforming RFs in all cases, particularly when measuring the recall of the different approaches. This helped further establish RPFs as a more adequate candidate for the classification of habitats, albeit both systems proved to be generally inaccurate for the task of second- and third-level habitat classification.

In terms of the effectiveness of the features extracted, the experiments revealed a similar situation to the previous scenarios in terms of the performance of pattern features. CPAM features obtained the best precision and recall results in general in almost all scenarios, with Neutral Grassland (B.2) in H3K and both Heathland mosaics (D.1 and D.2) in H3K and H1K being the clearest exceptions. However, contrary to the previous set of experiments, colour and texture features experimented a shift in performance. Texture features obtained much more accurate results, oftentimes even outperforming colour features, such as in the classification of Intertidal Mud/Sand (I.1) mosaics Neutral Grassland (B.2) and, in one particular occasion, in the case of the recall for Dry Heath and Acid Grassland Mosaics (D.2), even outperforming the use of pattern features. It is because of this that we decided to keep using texture in our future experiments. However, it should be noted that the use of texture features, both with RPFs and RFs, also produced some of the most variable and unstable results. For example, their recall accuracy for Broad-leaved Woodland (B.1) and Dry Heath/Grassland mosaic clashes with their inability to classify Fences (J.1.3), Marshy Grassland (B.3) and even Mixed Woodland (A.2). On the other hand, the use of pattern features produce more stable, albeit less accurate on occasion, results.

On the other hand, colour features experienced a decline in accuracy in both recall and precision metrics. These two situations are linked, since the colour characteristics between members of the same classes can be too subtle while texture characteristics might be more pronounced. For example, the texture of a Dry Heath/Acid Grassland mosaic is quite different from a Dry Heath/Scrub mosaic, while the colour characteristics might be similar, since Scrub and Acid Grassland can share the same shades of green.

Regarding the effects of each of the databases, it can be seen that the situation was reversed in comparison to the classification of first-tier habitats. In this case, results for H1K habitats were more accurate than results from H3K. This is consistent with the nature and purpose of both datasets. H1K photographs were taken under more controlled situations. Moreover, all four sites visited were from the same geographical area, Hampshire. Consequently, the variation of second- and third- tier habitats was not as large. For example, most of the woodland photographed was Broad-leaved (B.1), and most of the grassland was Neutral Grassland (B.2). These are, not incidentally, the two most accurate classified habitats. On the other hand, H3K photographs were taken under an extremely varied number of conditions. They were taken by a number of different people, located all across Great Britain, using different equipment and during different times and years. Consequently the variation present, which greatly helped first-tier classification, harmed second- and third-tier classification because the instances for each different combination of conditions were not enough.

Moreover, another set of interesting results can also be found when looking at the new categories introduced with H3K. For example, Open Water (G) obtains some of the highest recall results of the whole framework. However, its precision results are lower. This is mainly due to the reflection effects of the water. Consequently, the colour, pattern and even texture between some of the instances in the Open Water category were similar to those of the habitats which were reflected in the water.

Another important result comes from the classification of Inland Cliffs (I.1.1) and Maritime Cliffs (H.3). This is a great example on the limitations of visual features, since both habitats are composed of essentially the same type of geographical object, a cliff. It is only their location with respect to water what makes them different habitats. As a result, it is clear that visual features alone cannot help their correct classification.

In summary, these sets of experiments helped determine that, while RPFs are viable alternative to RFs, the current design had clear limitations for the accurate classification of second- and third-tier Phase 1 habitats. These limitations were mainly due to the type of features we were extracting. That is, the information that was being extracted was not enough to clearly differentiate between extremely similar classes, such as a Maritime Cliff (H.3)and an Inland Cliff (I.1.1) or between Tall Herb (C.1) and Scrub (A.4). The

same way the "Other" set of features had proven to be insufficient to classify first-tier habitats, the same low-level features that had obtained reasonably good results for first-tier habitats now had clear limitations when applied to a finer level of classification.

This motivated us to contemplate the integration of another type of features in our framework. In particular, we chose to include semantic information, which is crucial when distinguishing between practically identical classes. This information would extract relevant information that was already in the photographs but that low-level visual features could not collect. A way to introduce semantic information in the classification, such as the approach we present in Chapter 7, would entail a higher performance in terms of recall and, more importantly, precision in both classification scenarios.

7.6.3 Visual Results

Figure 7.10 and Figure 7.5 present two particular visual examples of our H1K and H3K databases, respectively. Moreover, Table 7.5 and Table 7.6 results obtained from our experiments. Both of them show the unseen test photographs and gives the first five results obtained with RPF and RF when extracting pattern features, colour features, texture features and all the features together. Additionally, correct results are shown in bold and italics.

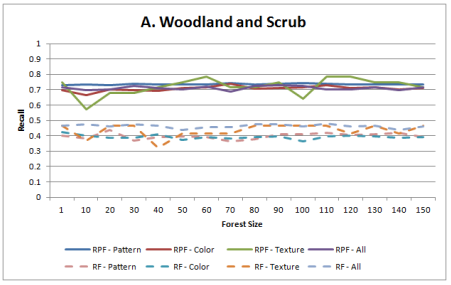
These examples serve to further illustrate the finding from these experiments. As can be seen in both examples, RPFs are more accurate than RFs in all cases. Moreover, the best classification results are obtained using pattern features.

7.7 Concluding Remarks

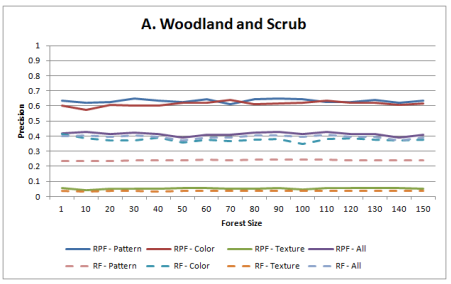
In this chapter, we have described in detail the Machine Learning approach we have developed and used to automatically classify habitats: Random Projection Forests. This is the third contribution of this thesis. In particular, we have described how they are built, how they can be applied to image annotation and how they compare to Random Forests. Moreover, we have carried out extensive testing to demonstrate their stability and to study their performance when combined with low-level visual features.

Finally, we also present the first part of our fourth contribution: a study on the effects that colour, texture and pattern features have on automatic habitat classification. Results have shown that, while low-level visual features can be used as the first step in the classification, they present some limitations when classifying second- and third- tier habitats, which have extremely similar visual properties.

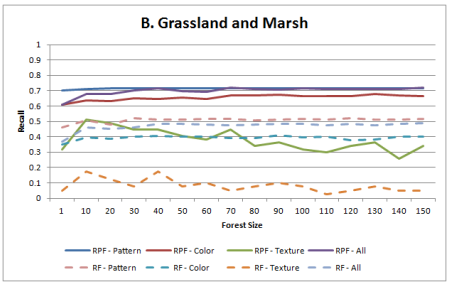
In the next chapter, we develop and present a new type of feature, specially created to help with this problem by including semantic information in the classification process as part of the input. We refer to these features as medium-level features.



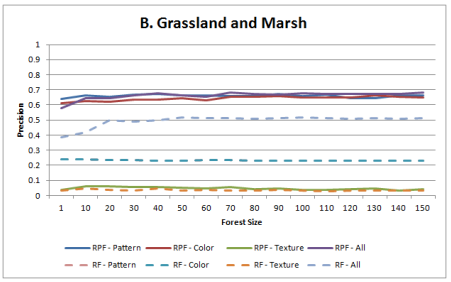
(a) A. Recall



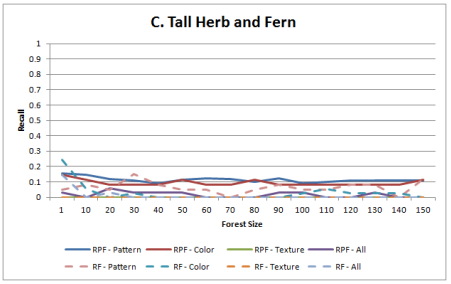
(b) A. Precision



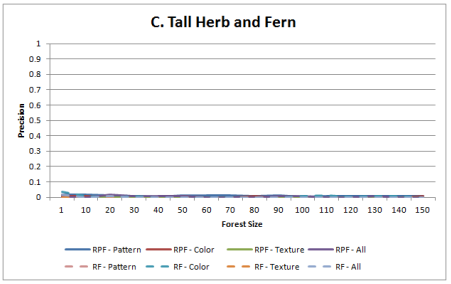
(c) B. Recall



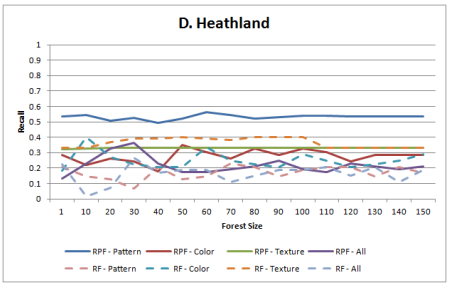
(d) B. Precision



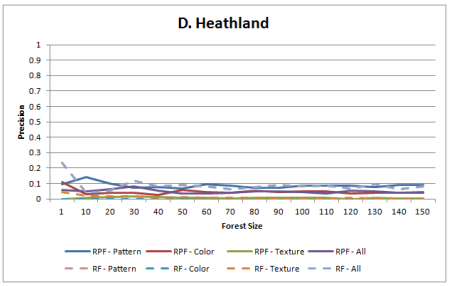
(e) C. Recall



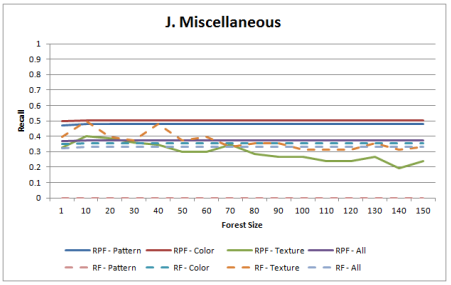
(f) C. Precision



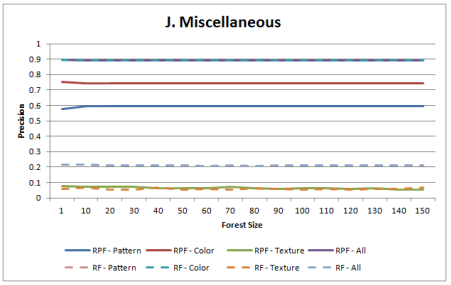
(g) D. Recall



(h) D. Precision

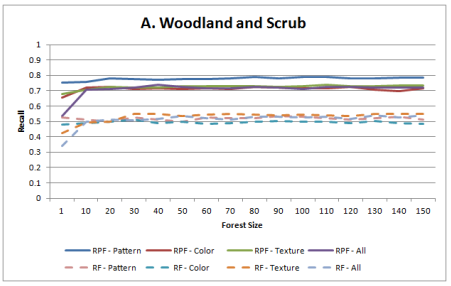


(i) J. Recall

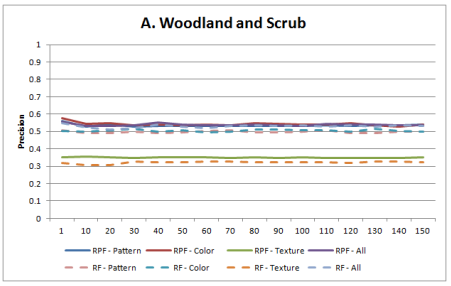


(j) J. Precision

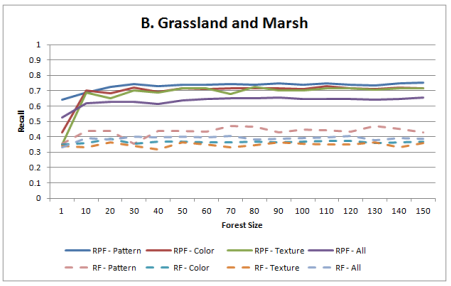
FIGURE 7.5: Random Projection Forests. Recall and precision results for first-tier habitats from Habitat 1K



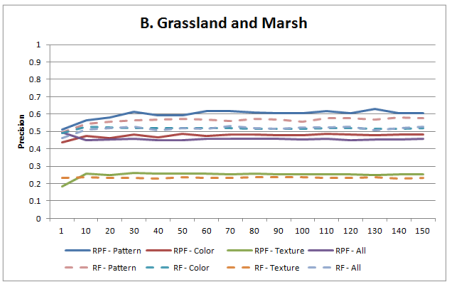
(a) A. Recall



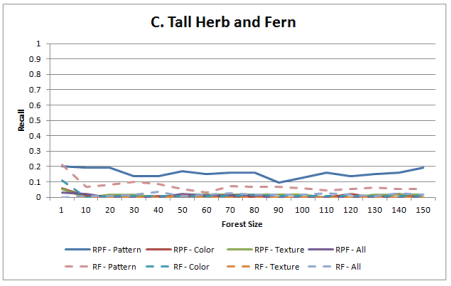
(b) A. Precision



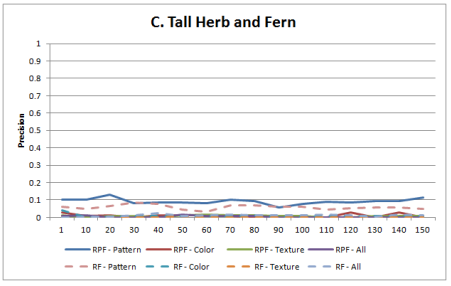
(c) B. Recall



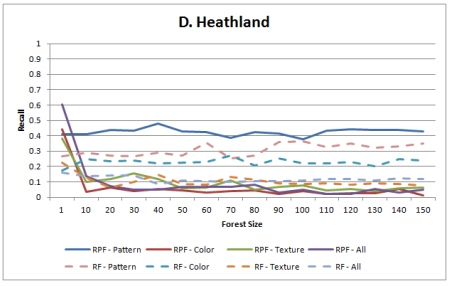
(d) B. Precision



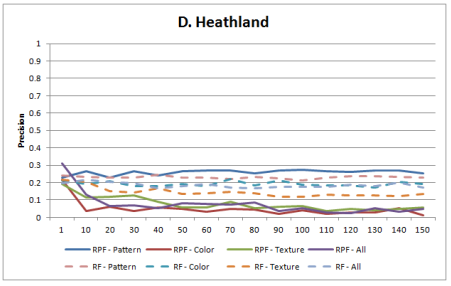
(e) C. Recall



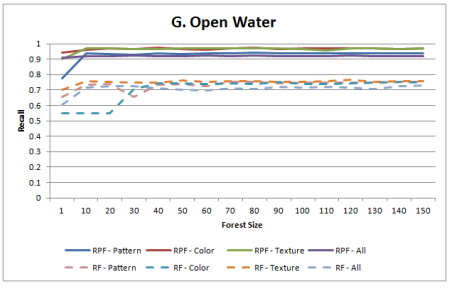
(f) C. Precision



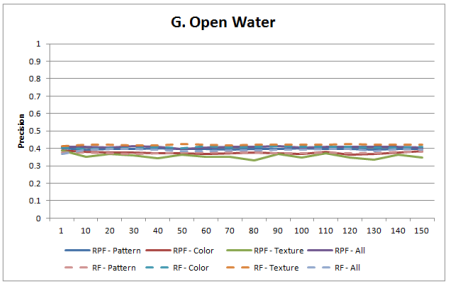
(g) D. Recall



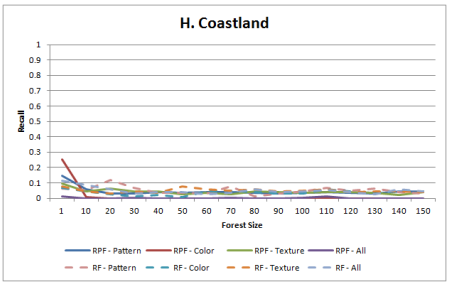
(h) D. Precision



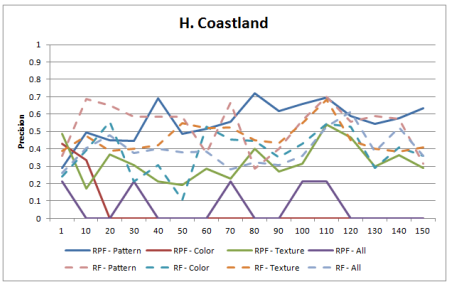
(i) G. Recall



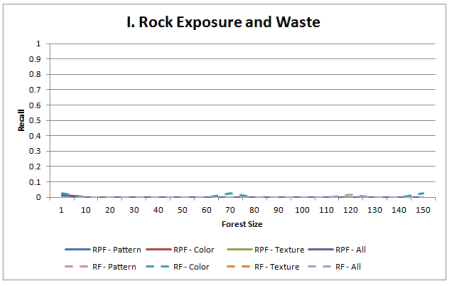
(j) G. Precision



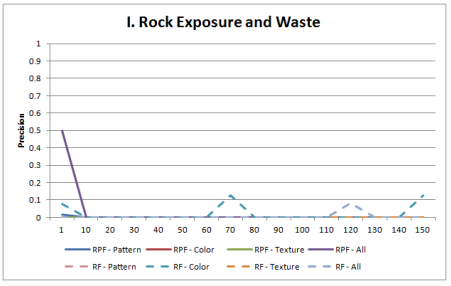
(k) H. Recall



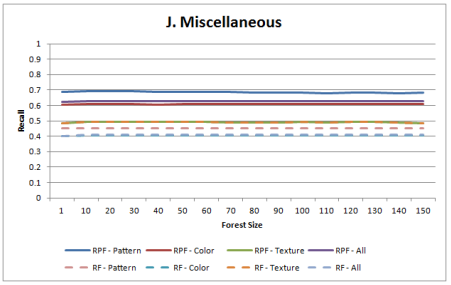
(l) H. Precision



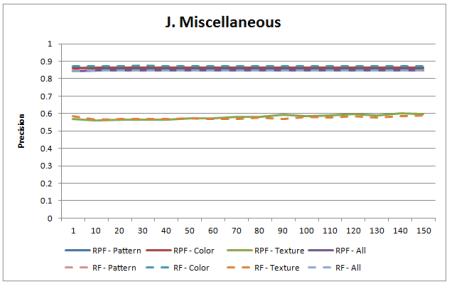
(m) I. Recall



(n) I. Precision

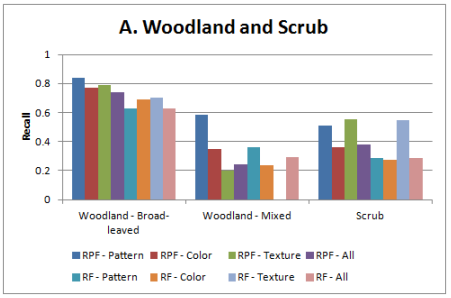


(o) J. Recall

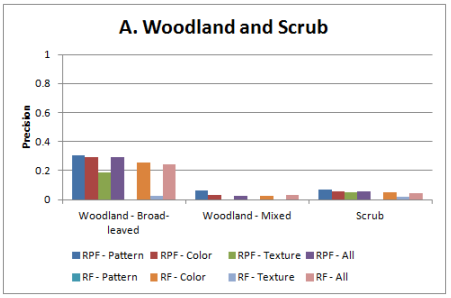


(p) J. Precision

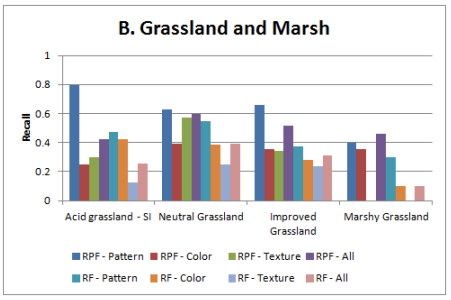
FIGURE 7.6: (Cont.) Random Projection Forests. Recall and precision results for first-tier habitats from Habitat 3K



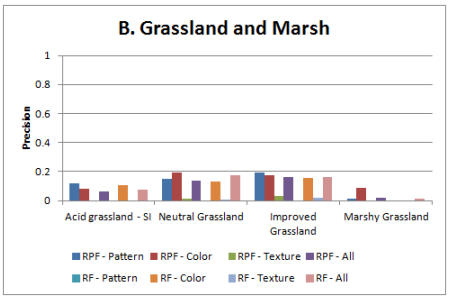
(a) A. Recall



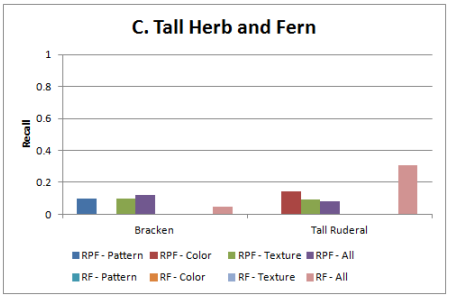
(b) A. Precision



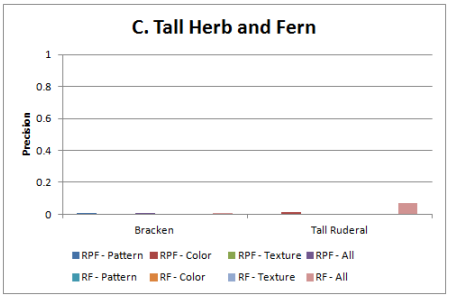
(c) B. Recall



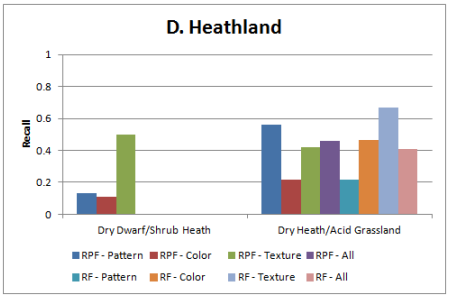
(d) B. Precision



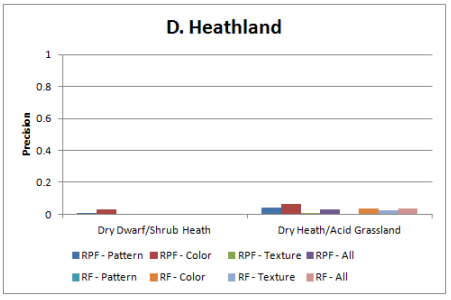
(e) C. Recall



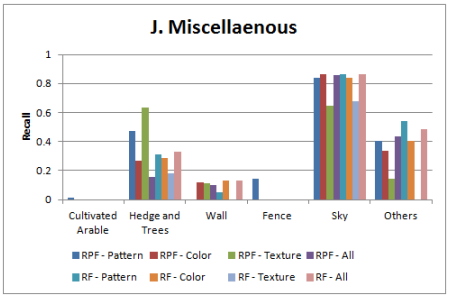
(f) C. Precision



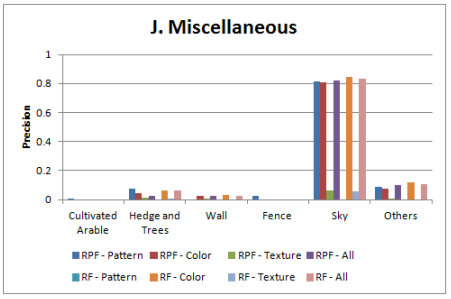
(g) D. Recall



(h) D. Precision

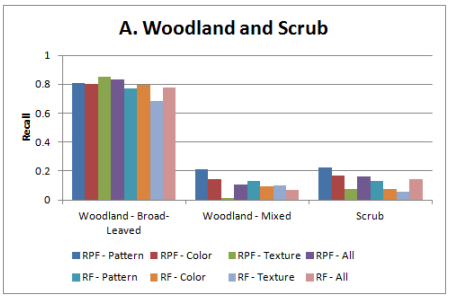


(i) J. Recall

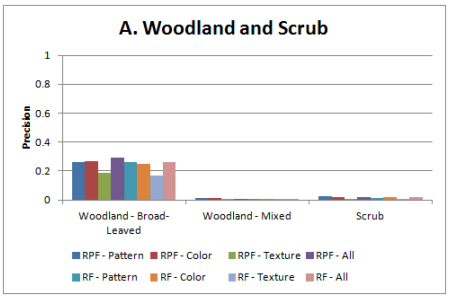


(j) J. Precision

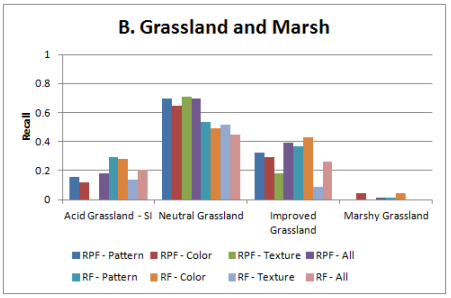
FIGURE 7.7: Random Projection Forests. Recall and precision results for second- and third-tier habitats from Habitat 1K



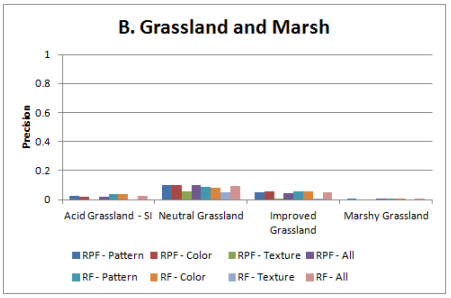
(a) A. Recall



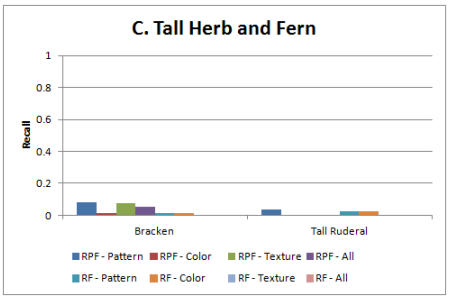
(b) A. Precision



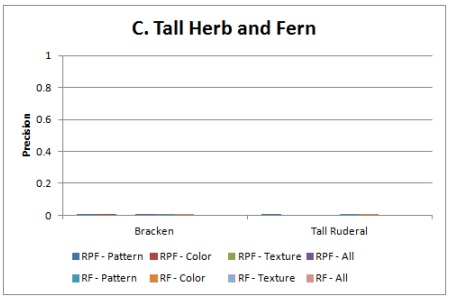
(c) B. Recall



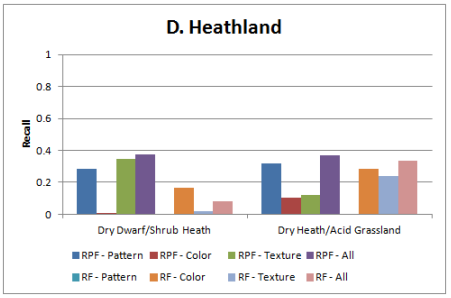
(d) B. Precision



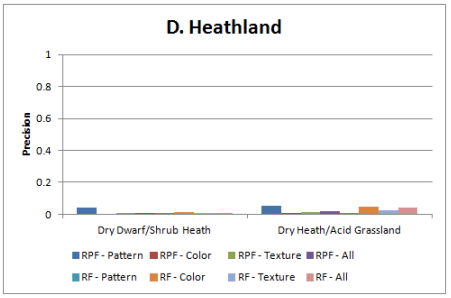
(e) C. Recall



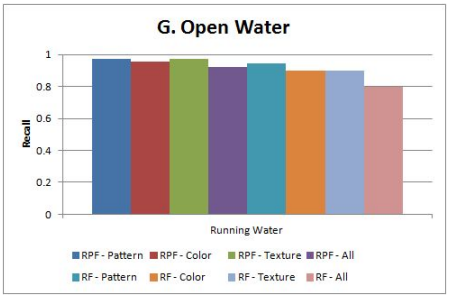
(f) C. Precision



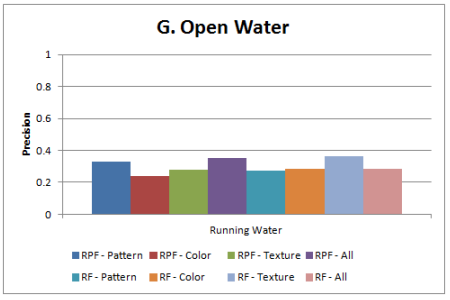
(g) D. Recall



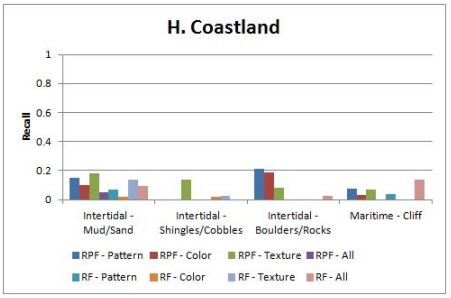
(h) D. Precision



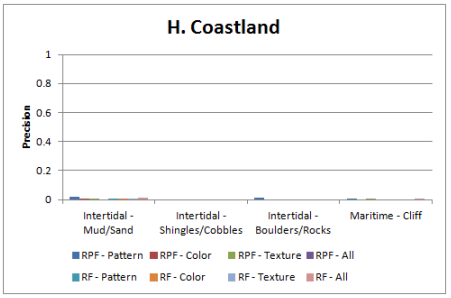
(i) G. Recall



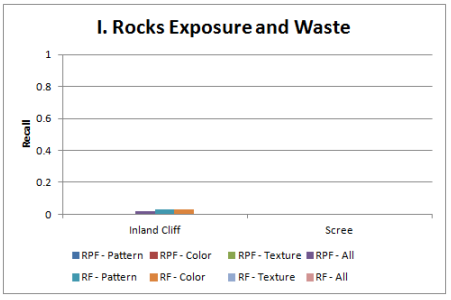
(j) G. Precision



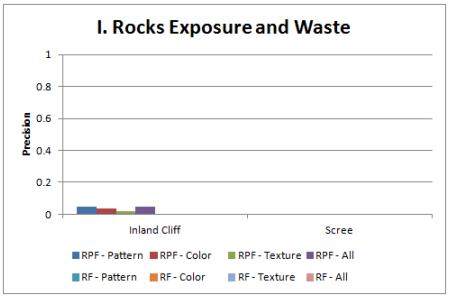
(k) H. Recall



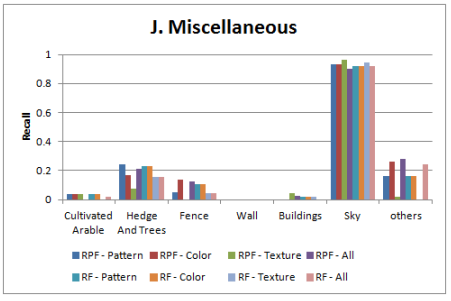
(l) H. Precision



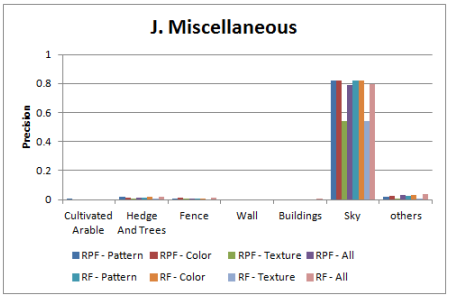
(m) I. Recall



(n) I. Precision



(o) J. Recall



(p) J. Precision

FIGURE 7.8: (Cont.) Random Projection Forests. Recall and precision results for second-tier habitats from Habitat 3K

FIGURE 7.9: Visual Example From H1K. Habitats present are: Acid Grassland - Semi-Improved, Scrub and Bracken.

TABLE 7.5: Results. We show the five most probable results obtained with our experiments.

		RF	RPF
Features Extracted	**Pattern**	Neutral Grassland	***Scrub***
		Dry Dwarf/Shrub Heath	***Bracken***
		Woodland - Broad-leaved	***Acid Grassland - SI***
		Scrub	***Sky***
		Sky	Neutral Grassland
	Color	Improved Grassland	***Acid Grassland - SI***
		Neutral Grassland	Neutral Grassland
		Sky	***Scrub***
		Scrub	Woodland - Broad-leaved
		Woodland - Broad-leaved	***Sky***
	Texture	Woodland - Broad-leaved	***Acid Grassland - SI***
		Woodland - Mixed	Woodland - Broad-leaved
		Scrub	***Sky***
		Neutral Grassland	***Scrub***
		Sky	Improved Grassland
	All	Neutral Grassland	***Acid Grassland - SI***
		Scrub	***Sky***
		Acid Grassland - SI	***Scrub***
		Dry Dwarf/Shrub Heath	Neutral Grassland
		Sky	Dry Dwarf/Shrub Heath

FIGURE 7.10: Visual Example From H3K. Habitats present are: Woodland - Broad-leaved, Running Water, Scrub, Acid Grassland - Semi-Improved

TABLE 7.6: Results. We show the five most probable results obtained with our experiments.

		RF	RPF
Features Extracted	Pattern	***Woodland - Broad-leaved***	***Scrub***
		Scrub	***Woodland - Broad-leaved***
		Bracken	Neutral Grassland
		Sky	***Sky***
		Neutral Grassland	***Running Water***
	Color	***Sky***	***Sky***
		Bracken	***Scrub***
		Marshy_grassland	***Woodland - Broad-leaved***
		Woodland - Broad-leaved	***Running Water***
		Scrub	***Acid Grassland - SI***
	Texture	Neutral Grassland	***Sky***
		Sky	***Acid Grassland - SI***
		Improved Grassland	***Scrub***
		Scrub	***Woodland - Broad-leaved***
		Woodland - Broad-leaved	Improved Grassland
	All	***Woodland - Broad-leaved***	***Woodland - Broad-leaved***
		Woodland - Mixed	***Sky***
		Neutral Grassland	Improved Grassland
		Sky	***Scrub***
		Improved Grassland	Dry Dwarf/Shrub Heath

Chapter 8

Medium-Level Features

As discussed in the Chapter 7, the use of low-level visual features has some limitations. In particular, low-level visual features cannot, by nature, collect semantic information, which can be crucial to distinguish between habitats that belong to completely different classes but that share similar visual characteristics. In this chapter, we propose the use of semantic features, referred to as medium-level features, in combination with low-level visual features to improve the performance of our Random Projection Forests. The generation, selection and extraction of medium-level features constitute the fifth contribution of this thesis. Medium-level features are extracted from ground-taken photographs using a Human-in-the-Loop approach. We have created a set of thirty-six questions regarding the objects present in the photographs and we use the answers to these questions and the certainty users have on their answers to create medium-level features. Experiments were carried out to test the addition of semantic features to our framework and their effect when combined with low-level features. As will be shown in the results, the inclusion of medium-level knowledge in our framework improves the accuracy of the classification, with recall and precision improving significantly in the case of complex habitats.

This chapter is structured as follows. Section 8.1 explains the motivation behind adding Medium-Level Knowledge in our framework and how they can be used to help with the problems brought by the "Semantic Gap". Section 8.2 describes how this Medium-Level Knowledge is extracted and how it can be transformed into features, referred to as Medium-Level Features. Moreover, it also describes how it can be incorporated in our image annotation framework. Finally, it also describes in detail the set of medium-level annotations that we have created. Moreover, Section 8.3 gives a brief description of the medium-level annotations and features that were extracted for our two habitat classification databases, Habitat 1K and Habitat 3K, along with some statistics and

some visual examples. Section 8.4 describes the type of experiments we have carried out to assess their performance when combined to RPFs. Moreover, Section 8.5 presents the results obtained from the experiments and discusses them in depth. Finally, Section 8.6 presents a brief summary of the contents of the chapter and some brief final remarks.

8.1 Motivation

Low-level feature selection and extraction methods have been successfully applied to many popular Computer Vision problems, such as face recognition [203], image retrieval [169] and even image annotation [76]. However, as shown in the results obtained in Chapter 4 and Chapter 7, relying solely on low-level visual features entails some limitations.

As mentioned in Chapter 6, low-level features commonly extract only visual information in the form of global or local statistics. However, there are objects that, while belonging to completely different classes, might have similar visual properties. This makes their automatic classification process extremely complicated if only visual features are taken into consideration. For example, based on colour, texture or pattern features alone, it is impossible to distinguish a tree that belongs to a Woodland (A.1) habitat or a tree that belongs to a Hedge and Trees (J.1.2.) formation. In these cases, there is a clear gap between the visual characteristics of the objets within a photograph and their semantic meaning.

This phenomenon is known in the Computer Vision field as the "Semantic Gap" [18]. The semantic gap is defined by [170] as "the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation". This concept clearly identifies the limitations that visual information has when classifying objects. Moreover, it also points that there is a lack of "interpretation" information taken into consideration during the classification. In other words, the Semantic Gap can be caused or aggravated by traditional feature extraction methods, which focus only on visual information extraction, while there is a lot of semantic or interpretation information that could aid the classification process that it is not extracted following these traditional feature extraction approaches.

In an effort to bridge this gap, the introduction of semantic information in the classification process has been proposed. However, low-level feature extraction methods are not suited for the collection of such semantic information. As a result, a new type of feature, often referred to higher-level features, has been proposed [36]. Higher-level features are designed to incorporate semantic information about the objects within an image. They

(a) Semi-Improved Grassland (b) Improved Grassland

FIGURE 8.1: Visual Similarity of FGVC Problems. The two images belong to different Grass categories. However, they are extremely visually similar.

can be used on their own or they can be combined with other types of features [134, 135]. Moreover, they can take many forms, as shown in [135, 151, 184]. Furthermore, can be extracted automatically [36] or, as is our case, they can be extracted using humans [184]. Additionally, they can be applied to a wide range of problems, not only in the field of Computer Vision [122], but also in other fields, such as Signal Processing [130].

It is particular crucial to notice that the semantic gap problem is even more pronounced and has more effect in Fine-Grained Visual Categorization problems, such automatic habitat classification. As described in Chapter 2, FGVC problems aim to accurately classify between classes that are visually similar and have similar semantics [205]. For example, current research on FGVC includes the automatic classification of different types of leaves [108], flowers [136], dogs [120] and birds [15, 25]. As can be seen, the classes to identify in FGVC problems share very similar visual properties and it is often that they can be indistinguishable to the untrained eye. Figure 8.1 shows an example of this based on our problem, automatic habitat classification. It can be seen how similar Semi-Improved Grassland and Improved Grassland can be both visually, both of them are mainly green objects with similar texture, and semantically, they are both types of grasses.

In our case, we employ humans to extract semantic information in an effort to improve the classification. We refer to this semantic information as medium-level knowledge, and, from them, we create medium-level features. We introduce medium-level features in our framework to incorporate crucial semantic information that low-level features are unable to extract in the classification process. Additionally, the aim of using humans to collect this semantic information is to create a system that can benefit from both humans' strengths, such as being able to differentiate between different classes just by looking at a photograph, and computers' strengths, such as being able to carry out complicated calculations at a fast speed. Consequently, in order to take into consideration visual

and semantic information during the classification of habitats, we combine low-level and medium-level features.

In summary, we are adding semantic information, in the form of medium-level features, to our image annotation framework in order to bridge the limitations of introduced by low-level features and the semantic gap. The combination of low-level and medium-level features is designed to help classify habitats which share very similar visual properties and improve accuracy of our framework as a whole.

8.2 Medium-Level Annotations and Features

As mentioned in the previous section, higher-level knowledge can take many forms and can be applied in different ways through the classification process. In this chapter, we propose the inclusion of semantic information in the classification process as an extension of our framework in order to improve accuracy. We refer to this semantic information as medium-level knowledge or medium-level information.

In particular, we expand the Random Projection Forest design presented in Chapter 7 to include higher-level semantic information as part of the input. To do this, and following the automatic image annotation approach we have created, we collect medium-level knowledge as annotations. Figure 8.2 shows an overview of how the process of creating medium-level information is carried out and how medium-level annotations are transformed into medium-level features. As can be seen, the process can be divided into two phases: the generation of the knowledge as annotations and the generation of the corresponding features.

8.2.1 Knowledge and Annotation Generation

In this first phase, human users are needed to generate medium-level knowledge. These users are not required to have previous knowledge of habitat classification. They do not need to be Phase 1 experts, or even ecologists. The inclusion of users in the classification process is an approach that has been used in the Computer Vision community for several years [25]. This methodology is commonly referred to as a "Human-in-the-loop" (HITL) approach.

First proposed in [24], HITL approaches have been successfully applied in FGVC problems, as shown in [25, 26]. Since FGVC classification is challenging for both humans and computers, HITL methods were proposed to be an intermediate solution which would progressively minimise the amount of human labour necessary to classify FGVC classes

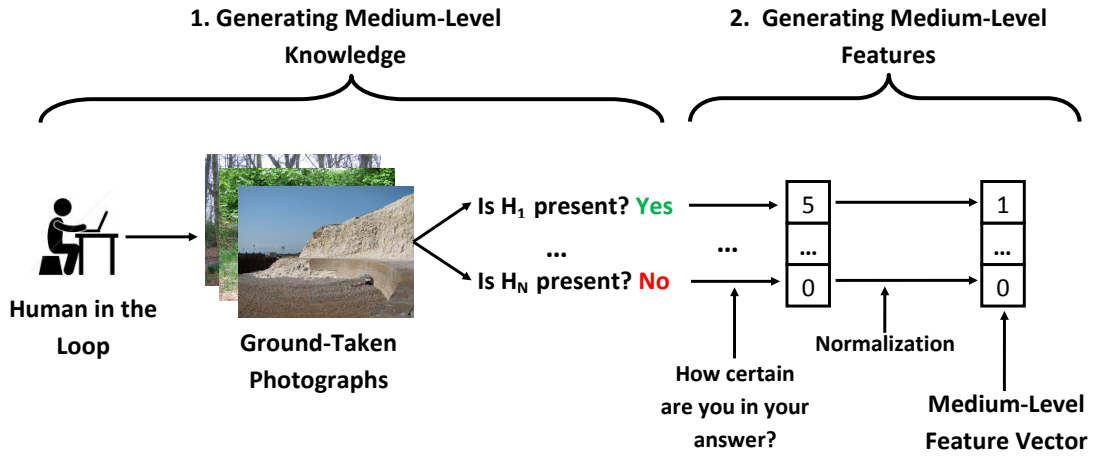


FIGURE 8.2: Medium-Level Information and Features. In our case, N is equal to 36 and certainty in measured between 0 (not sure at all) and 5 (completely sure).

[24]. HITL methodology can be easily applied many different problems, such as criminology [140], port design [29] and even aviation [171]. However, it is particularly suitable for FGVC because it brilliantly utilises the ability that humans have to differentiate between objects. For example, [24] developed a HITL method for bird classification and [151] used HITL technology for skin-lesion image recognition.

In HITL methods, a user is shown a photograph and, then, asked a series of questions regarding the contents of said photograph. As can be inferred, the selection of the questions is crucial. A set of sufficiently descriptive and discriminative questions must be prepared, since they will determine the information that will be collected from the photographs. These questions do not have to follow any particular pattern. They can be completely open, multiple-choice or they can be simple "yes"-or"no" questions [24, 151].

In our case, we have developed a set of twenty-three "yes"-or"no" questions. This object-based set of questions that aims to collect information about which habitats are present within an image. Consequently, all the questions follow the same pattern: "Is/Are there any X object/s in the photograph?", with X being each of the thirty-six objects. The list of questions is presented in Table 8.1.

TABLE 8.1: Questions Asked To Users. With this questions, we extract Medium-Level Knowledge which will be then transform into Medium-Level Features

Is/Are there any X object/s in the photograph?	
Trees - leaves	Grass - not green
Trees - no leaves	Sand - mud
Trees - mixed leaves	Small rocks
Scrub	Big rocks
Grass - with flowers	Water - standing
Grass - uniform	Cliff - water
Grass - reed	Cliff - no water
Bracken or fern	Spring
Tall herb	Summer
Heath	Atumn
Water - runing	Winter
Arable land	Brown
Boundary - scrub/trees	Yellow
Wall	Red
Fence	White
Sky	Blue
Grass - bright green	Green

Moreover, in order to make the extraction of information more efficient and less tiresome for the users, the questions are all asked at the same time with the help of a drop-down menu. This information is then converted into annotations. This whole process is iterative and follows these steps:

1. The users are presented with a ground-taken photograph.
2. For each distinguishable object that they are able to identify in the photographs:
   (a) Users create a polygon that contains said object If they are unable to create the polygons, due to habitat regions not being clear enough to delimit where they start or finish, their annotations will refer to the whole photograph.
   (b) Users answer the twenty-three questions by choosing which objects are present in the photograph if they want.
   (c) For each answer, users also score their level of confidence in their response. The level of confidence follows a scale between 0 and 5, with 0 being "not sure at all" and 5 being "completely sure". If their confidence is not filled, we assume a confidence of 5. This answers will be used in the next step to create the medium-level features.
3. Once the users have finished with all the objects in the image that they can distinguish, the information they have provided is converted into an annotation and

stored. The coordinates of the polygon are stored in an XML file, in a similar fashion as the ground-truth annotations were stored.

8.2.2 Feature Generation

To create the medium-level features, we use the confidence measures collected in the previous step. For each image x in the database, all the users responses stored in a 23-dimension feature vector $H(x) = (h_1, h_2, , h_{23})$ that is generated as follows:

$$h_i = \begin{cases} c_i & \text{if the answer to } q_i \text{ is “yes”} \\ 0 & \text{if the answer to } q_i \text{ is “no”} \end{cases} \tag{8.1}$$

Where c_i is the degree of confidence that the user has in that the object of question i is present in the photograph x. Consequently, the vector H is what we will refer to as medium-level features.

It is important to notice that our framework presents two modifications over traditional HITL approaches, such as the methods presented in [24–26]. First, in the HITL methodology described in [24], the answer to one question directly influences the selection of the following questions. This process is repeated iteratively until a prediction can be made. This type of approach is consequent for the classification tasks chosen in [24]. That is, bird classification. In [24], only one object within the photographs is being classified and the questions asked about the birds in the photographs revolve around their characteristics, such as the colour of their feathers, the shape of their beak, etc. Questions need to be prioritised and changed because not all possible combinations of characteristics are possible and because an species might be determined by a variable number of answers. For example, birds with an orange beak might always have black feathers on their wings but the shape of their heads might be a defining quality. Consequently, asking about the shape of their head is crucial and might give an accurate prediction only with those two answers, while inquiring about the colour of the wings might collect unnecessary information for the classification process. In a way, we can regard the questions and the objects of these questions as dependent of each other.

In our case, we have chosen to simplify this process. Users are shown all the questions at the same time and they only have to choose which objects they see in the images, where they are localised and their level of confidence in their answer. Consequently, one answer does not affect other questions. The motivation behind this decision is rooted in the notion that we are classifying several objects, or habitats, in each photograph and the presence of one type of habitat in the image does not necessarily determine the

presence of another habitat. For example, the appearance of bright-green grass does not interfere with the appearance of sand. That it, we assume that all the objects present in an image are independent from each other. Consequently, all questions must be asked every time.

Our second modification is with regards to the extent to which humans are used in the framework. [24–26] include human input iteratively. In our case, we asked the questions once to the users. Then, we transform their answers into features which are used as the input of the classifier. The main reason behind this decision is efficiency, since engaging users in multiple cycles of image annotation was time-consuming and labour intensive.

8.3 Medium-Level Features in Habitat 1K and Habitat 3K

The extraction of medium-level feature for Habitat 1K and Habitat 3K was done following the steps described in the previous section. The annotation process was done using the same annotation tool used to ground-truth our ground-taken databases [107]. We modified the tool to include the twenty-three questions instead of the Phase 1 classification scheme. To collect the information, we recruited three people who annotated the photographs with medium-level information in four different sessions. Each image was annotated once by one of the participants. Consequently, each photograph in our ground-taken database generated one medium-level feature vector. An alternative method would have been to collect multiple feature vectors from each photograph. This would have given us different points of view and additional information about the ground-taken photographs. However, time constraints prevented this. Moreover, it is important to point out that, following traditional HITL methodologies [151], none of the users were trained ecologists.

Table 8.1 shows the frequency of appearance of the answers in both databases. Additionally, Figure 8.3 shows four examples of annotated photographs in which the annotations were global, as in the first column, and localised, shown in the second column.

8.4 Experiments

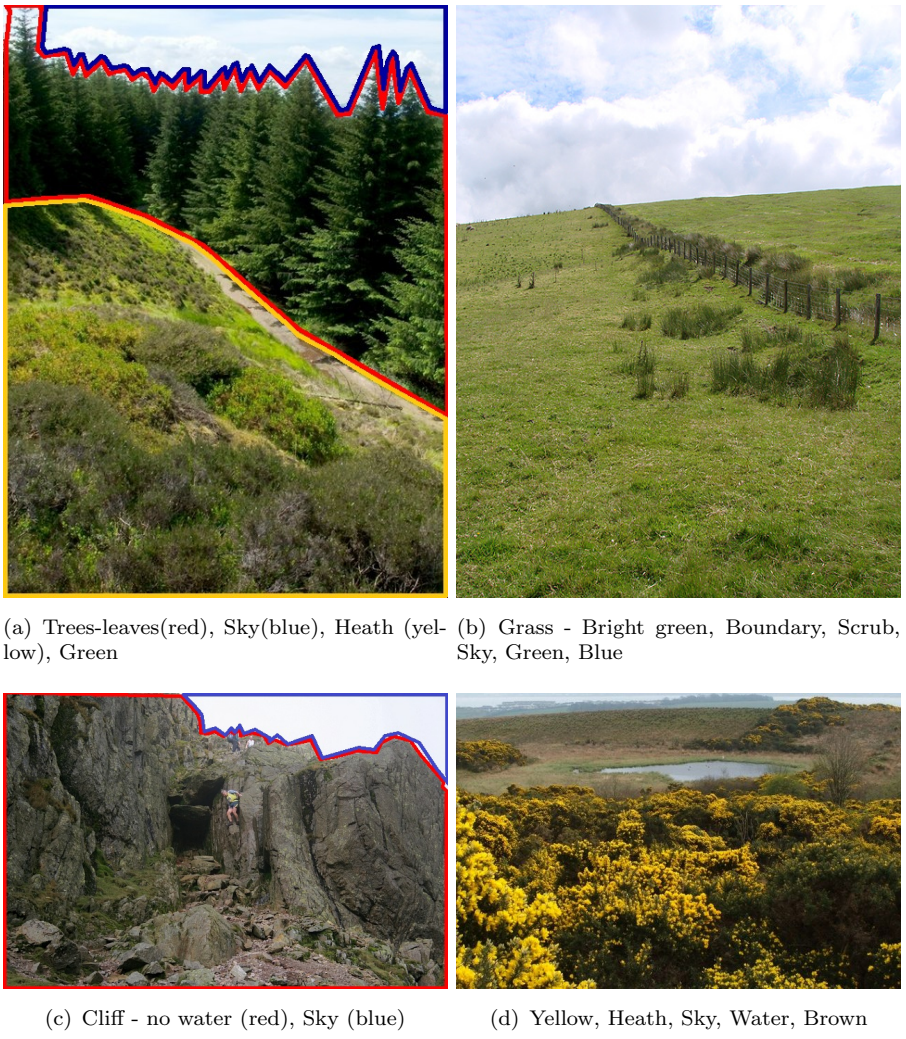
A series of experiments were carried out to test the inclusion of medium-level features to our framework. Following the findings from Chapter 7, we decided to focus our experiments on extracting features from the images as a whole and comparing the performance of the modified RPF framework with the original RPF system.

TABLE 8.2: Frequency of Appearance of Each Annotation in H1K and H3K.

Objects	Habitat 1K	Habitat 3K
Trees - leaves	43	193
Trees - no leaves	364	913
Trees - mixed leaves	238	375
Scrub	380	958
Grass - with flowers	541	1252
Grass - uniform	137	184
Grass - reed	69	197
Bracken or fern	130	250
Tall herb	23	119
Heath	91	745
Water	19	118
Arable land	67	119
Boundary - scrub/trees	217	436
Wall	12	95
Fence	153	241
Sky	916	2557
Grass - bright green	134	134
Grass - not green	0	39
Sand - mud	0	167
Small rocks	0	152
Big rocks	0	28
Cliff - water	0	84
Cliff - no water	0	183
Spring	352	75
Summer	431	122
Autumn	17	0
Winter	169	0
Brown	984	1007
Yellow	0	52
Red	0	28
White	0	4
Blue	1043	30
Green	967	47

Correspondingly, we set up these experiments with the specific goal of studying the effect of medium-level features, RPFs and global feature vectors. Similarly to Chapter 7, we studied this by generating results on the performance of RPFs when varying an specific set of parameters. These parameters are:

- Medium-level features: Results from Chapter 7 demonstrated the clear limitations of low-level visual features, particularly when classifying second- and third- tier habitats. In this chapter, we have introduced the concept of medium-level features, which were extracted using and HITL approach and store semantic information.



(a) Trees-leaves(red), Sky(blue), Heath (yellow), Green

(b) Grass - Bright green, Boundary, Scrub, Sky, Green, Blue

(c) Cliff - no water (red), Sky (blue)

(d) Yellow, Heath, Sky, Water, Brown

FIGURE 8.3: Photographs Annotated With Medium-Level Tags. Users decided to use global tags for photographs (a) and (c) and a mixture of global and localised tags for photographs (b) and (d).

In these experiments, we aim to test their efficacy when compared to RPFs which do not use them.

- Colour, pattern and texture features: Following our findings from the previous chapter, we extract and compare the performance of our classifier when colour features (Colour Histogram, Colour Moments), texture features (Tamura, GLCM), pattern features (CPAM) and all of them combined are extracted and combined with medium-level features. We project that visual features will continue producing high recall results for first-tier habitat classification while medium-level features will increase precision accuracy when classifying second- and third- level habitats. We also compare performances of these features against the performance of the "Other Features", a combination of six of the most common visual features currently used in Computer Vision problems (GB, GIST, SIFT, SSI, PHOW, PHOG). For

clarity purposes, these results will not be shown in the graphs, but we will describe their performance in their respective sections.

- Database: Given the different nature of the databases created in this thesis, Habitat 1K being collected under controlled circumstances and Habitat 3K being collected using crowd-sourcing methods, we also aim to study the effect of semantic information on their performance.

Moreover, we decided to compare the original design of Random Projection Forests against Random Projection Forests with medium-level features to obtain a more in-depth study of their effect. Additionally, to ensure consistency between the results, we follow the same methodology as in Chapter 7 and we calculate the recall, precision and the confusion matrix of results obtained.

8.5 Results

In order to assess the effect of medium-level visual features and Random Projections Forests, we have tested ten scenarios with each or our databases. These scenarios are:

1. RPF with colour features and medium-level features. This scenario is referred to as MLF - Color in the following figures.
2. RPF with pattern features and medium-level features. We refer to this as MLF - Pattern in the following figures.
3. RPF with texture and medium-level features. This is called MLF - Texture in the figures.
4. RPF with all three features linearly combined and medium-level features. This scenario is referred to as MLF - All in the following figures.
5. RPF with other features and medium-level features. In order to make visualization easier, we have not included these results in the graphs. However, the findings from this set of experiments will be commented and compared with the results obtained in the other experiments.
6. RPF with colour features. This scenario is referred to as RPF - Color in the following figures.
7. RPF with pattern features. We refer to this as RPF - Pattern in the following figures.

8. RPF with texture. This is called RPF - Texture in the following figures.

9. RPF with all three features linearly combined. This scenario is referred to as RPF - All in the following figures.

10. RPF with other features. In order to make visualization easier, these results are not included in the graphs. However, the findings from this set of experiments and how they compare with the other feature combinations will be discussed.

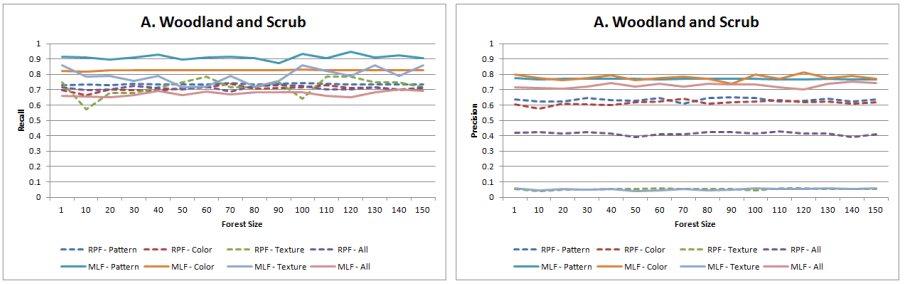
Similarly to Chapter 7, we divided the results obtained according to the level of detail of the habitats classified. We have calculated the recall and precision for first tier-habitats in Section 8.5.1, while Section 8.5.2 presents results for second- and third- tier habitats. We compare each set of results with the Random Projections Forests results obtained in the previous chapter. Finally, we present some visual examples obtained during our testing in Section 8.5.3.

8.5.1 First-Tier Classes

Figure 8.4 shows the recall and precision results obtained in the testing scenarios introduced previously when using features extracted from whole images from H1K as the input. On the other hand, Figure 8.5 shows the same metrics when testing our framework with features extracted from whole photographs from H3K as the input. We tested forests with sizes ranging from 1 to 150 and with depths ranging from 2 to 10. However, in order to present the results in a clear and concise manner, we set their depth to 9 in the mentioned figures. Nevertheless, the performance of both systems was similar and stable in all cases.

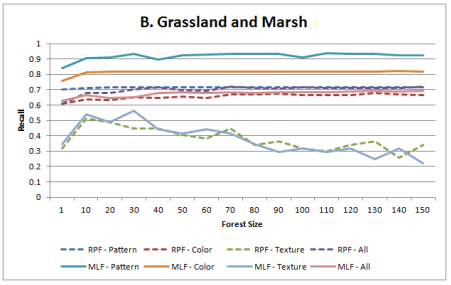
Looking at the results as a whole, we can see that, similarly to the results obtained in Chapter 7, the recall results tend to be higher than the precision results in most cases. The biggest difference in results is found in the case of H3K and Open Water (G) habitats, which experience a recall close to 100% in all experiments but, in terms of precision, these results drop to 40%. This situation also occurred in Chapter 7. However, it is interesting to notice that for the rest of the experiments, the differences between recall and precision result are not as pronounced.

Moreover, Tall Herb and Fern (C) and Heathland (D) continue being the most difficult classes to classify for H1K, while Rock Exposure and Waste (I) obtains the most inaccurate results when using H3K. This follows the trend discussed in Chapter 7 and should not be surprising, since the number of instances of these habitats in their respective databases are much lower. On the other hand, Woodland and Scrub (A), Grassland and

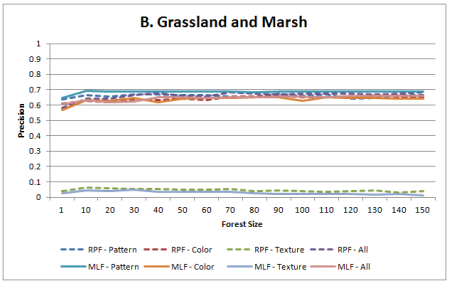


(a) A. Recall

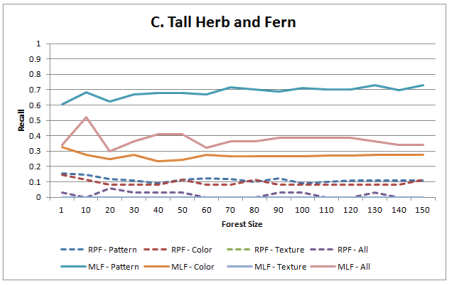
(b) A. Precision



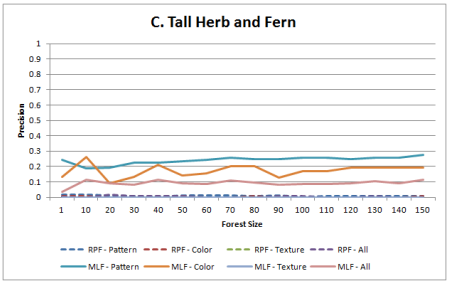
(c) B. Recall



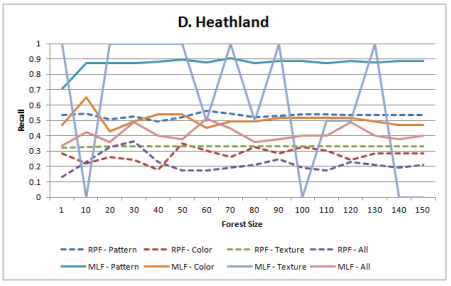
(d) B. Precision



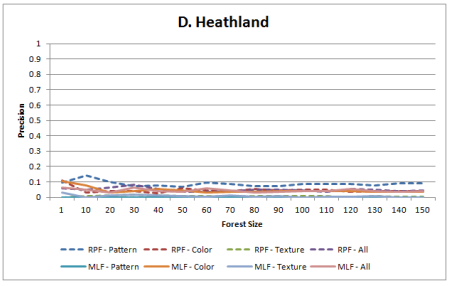
(e) C. Recall



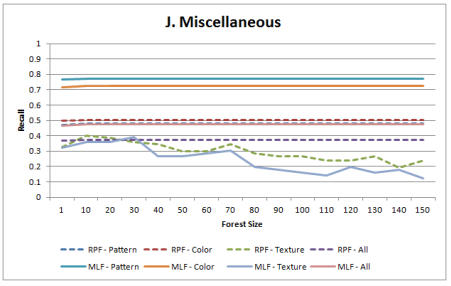
(f) C. Precision



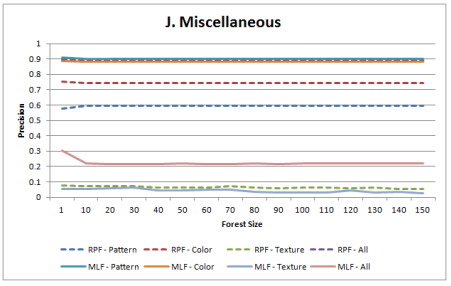
(g) D. Recall



(h) D. Precision



(i) J. Recall



(j) J. Precision

FIGURE 8.4: Medium-Level Features. Recall and precision results for first-tier habitats from Habitat 1K

Marsh (B) and Miscellaneous (J) continue being the most accurately classified classes in all cases. Additionally, it can be clearly observed that the addition of medium-level knowledge has aided Heathland (D) classification, with Heathland results, as a whole, experimenting an increase in accuracy in all new testing scenarios.

Looking at the experiments more closely, it can be seen that the inclusion of medium-level features matches or improves the performance of their equivalent experiment with only low-level features in all scenarios tested. This supports our belief that their selection and extraction is clearly useful for the FGVC problem that is automatic habitat classification. Medium-level features help with the visual similarities between some of the most problematic classes, such as Tall Herb and Fern (C) and Heathland (D). In fact, looking at the confusion matrices for the experiments with medium-level features, shown in Table 8.3 and Table 8.4, we can see that the misclassification of some habitats has been reduced by the introduction of semantic information. An example of this is shown when classifying Inland Cliff (I.1.1) habitats versus Maritime Cliff (H.3) (included in Coastland) habitats. The usefulness of medium-level features is clearly visible in the case of Tall Herb and Fern (C) in H1K, in which the combination of pattern and medium-level features present a great improvement over the results obtained with only pattern features. As mentioned previously, Tall Herb and Fern (C), while a simple habitat in nature, it is one of the most difficult first-tier habitats to classify due to its visual similarities with other habitats, such as Scrub (A.4), as shown in Table 8.3 and Table 8.4, and also due to their lack of frequency of appearance in H1K. However, the inclusion of semantic information affects its classification positively.

Another set of interesting results comes from comparing the different types of features extracted and how they interact with our medium-level features. As was the case of the results presented in Chapter 7, pattern features continue being the most accurate ones in most of the cases. This is not surprising, since the feature vectors obtained from extracting pattern and medium-level features contain color, texture pattern and semantic information in the most compact way. Their combination with medium-level features produces the majority of the most accurate results, both in terms of recall and precision. Additionally, colour features continue performing adequately well, obtaining similar results as the use of all the features put together.

Finally, texture features keep obtaining the least accurate classification results in all testing experiments except one. This clear exception is found in the case of Heathland (D) in H1K, in which texture and medium-level features, which generally perform quite inaccurately, actually outperform pattern and medium-level features. This is due to the clear influence of medium-level annotations. The HITL approach we have followed

TABLE 8.3: Confusion Matrix of H1K once medium-level features have been added to Random Projection Forests

	A	B	C	D	E	F	G	H	I	J
A	**72.86%**	4.08%	6.12%	4.69%	0%	0%	0%	0%	0%	12.24%
B	7.84%	**68.75%**	12.50%	9.92%	0%	0%	0%	0%	0%	0.99%
C	*44.21%*	13.68%	**20.00%**	16.84%	0%	0%	0%	0%	0%	5.26%
D	*34.81%*	22.22%	7.41%	**18.52%**	0%	0%	0%	0%	0%	17.04%
E	0%	0%	0%	0%	**0%**	0%	0%	0%	0%	100.00%
F	0%	0%	0%	0%	0%	**0%**	0%	0%	0%	0%
G	0%	0%	0%	0%	0%	0%	**33.33%**	0%	0%	*66.67%*
H	0%	0%	0%	0%	0%	0%	0%	**0%**	0%	0%
I	0%	0%	0%	0%	0%	0%	0%	0%	**0%**	0%
J	4.08%	4.98%	0.60%	2.01%	0%	0%	0%	0.00%	0.00%	**88.32%**

TABLE 8.4: Confusion Matrix of H3K once medium-level features have been added to Random Projection Forests

	A	B	C	D	E	F	G	H	I	J
A	**57.82%**	9.99%	13.51%	13.73%	0%	0%	0%	0%	0%	4.95%
B	4.89%	**58.11%**	4.63%	15.80%	0%	0%	0%	0%	0%	16.57%
C	*33.99%*	5.61%	**13.86%**	*34.32%*	0%	0%	0%	0%	0%	12.21%
D	27.55%	*37.74%*	0.24%	**20.63%**	0%	0%	0%	0%	2%	11.41%
E	0%	0%	0%	0%	**0%**	0%	0%	0%	0%	*100.00%*
F	0%	0%	0%	0%	0%	**0%**	0%	0%	0%	100%
G	0%	2%	0%	3%	0%	0%	**43.08%**	0%	18%	33.55%
H	4%	10%	0%	0%	0%	0%	0%	**20%**	28%	*38%*
I	7%	0%	0%	3%	0%	0%	0%	*59%*	**9%**	22%
J	3.40%	2.25%	0.32%	0.85%	0%	0%	0%	0.00%	0.05%	**93.14%**

enabled us to use the classifying strengths of humans, who might not be able to differentiate between different classes of heath mosaics, but who are very good at differentiating between what constitutes heath and what makes it different from all other habitats. On the other hand, the "other features" remain one of the most inaccurate features, with recall remaining at 39% at most and precision remaining at10% the best case, which was Woodland and Scrub (A) classification with RPFs and medium-level features, once again proving the problem-dependent nature of feature extraction and the importance of feature selection.

Looking at the results from an input point of view, we can see that the performances of H1K and H3K, particularly in the case of recall, are quite similar. The figures show a more balanced set of results between the databases than in the previous chapter. This is again a main consequence of introducing medium-level features. However, it is important to notice that, while medium-level features improve accuracy in all accounts in our framework, there are some cases in which the results obtained are still too inaccurate. Examples of this include the classification of Tall Herb and Fern (C), Coastland (H) and Rock Exposure and Waste (I) habitats with H3K, in which precision results average only a 10% accuracy. For these cases, the inclusion of semantic information has proven to be an development in the right direction, but it is still lacking. As a direct consequence of this, we can expect the results for second- and third-tier habitats to improve, albeit slightly, when combining medium-level features and low-level features.

8.5.2 Second-Tier and Third-Tier Classes

Figures 8.6 show the recall and precision results obtained in the same testing scenarios as in Section 8.5.1. Additionally, Figures 8.7 show the same metrics when testing our framework with H3K. Similarly to the other testing scenarios, we are using the whole photographs when extracting the features. We tested the forests varying their size between 1 and 150 and their depth between 2 and 10. However, in order to make the results easier to visualise, we have set the size of the forests to 120 and the depth of the forests to 9 in the graphs, since the performance of both systems was similar and stable in all cases.

Looking at the results as a whole, it is clear that the relationship between the recall and precision metrics, seen in the previous sets of experiments, is maintained. In all cases, recall measures are higher than precision metrics. Moreover, habitats from Tall Herb and Fern (C) and Heathland (D) continue being the most difficult to classify with our framework in H1K. In the case of H3K, Coastland (H) habitats, particularly Intertidal mosaics (H.2) and Rock Exposure and Waste (I) habitats obtain the less accurate results.

On the other hand, habitats from Woodland and Scrub (A) and Grassland and Marsh (B) remain the most accurately classified. It is important to notice that there is a particular increase in the classification results for Mixed Woodland (B.2). This increase in maintained in all scenarios where medium-level features are included except in recall results for texture and medium-level features. Consequently, we can conclude that this improvement is mainly due to the inclusion of semantic information.

Looking at the experiments more closely, it is clear that the inclusion of medium-level features has aided the classification of second- and third- level habitats a great deal. Those experiments in which medium-level features were used obtain more accurate metrics, particularly in terms of precision, with a raise close to 20%, as exemplified in Tall Herb and Fern (C) results. Medium-level features have also obtained higher recall results, albeit these improved results are not as consistently drastic as those seen in precision measures. These improvements are more noticeable in the case of complex habitats, such as Mixed Woodland (B.1.2), Heathland mosaics (D.1 and D.2) and, particularly, Hedge and Trees (J.2.3) both in H1K and H3K. All of these complex habitats experiment a significant increase in their recall and, to some degree, in their precision as well. This is consistent with the type of information that we have extracted. Complex habitats are in essence the types of habitat that most benefit from semantic information. This is mainly due to their shares visual similarities with other multiple habitats, such as the habitats of the vegetation that composes them.

Artificial habitats have clearly benefited from the introduction of semantic information in the classification process. The main artificial habitats in both datasets are boundary habitats, Wall (J.2.5) and Fence (J.2.4.) habitats. We can see in the results that the have experimented an improvement in recall and precision close to 30% and 25%, respectively. This is a reasonable consequence of the inclusion of medium-level features. When considering only visual features, these habitats are generally difficult to classify accurately because, in contrast to other types of habitats, such as Grasslands (B) and Woodlands (A) which appear very prominently in all the photographs in which they are depicted, they occupy a smaller fraction of the images. Consequently, most of the visual information extracted from our global features will revolve around those larger habitats. However, when we introduced semantic information and asked for the appearance of "boundaries", we were extracting information specially centered around these particular habitats. Moreover, humans are exceptionally good at distinguishing artificial habitats, such as fences and walls, from natural habitats, such as grass and cliffs. Therefore, the certainty levels of the answers for these questions were always the highest possible and, consequently, they had more weight during training.

Moreover, looking at the performance of the medium-level features alone, it can be seen that there is a clear variation between their effect on the different types of habitat classes, regardless of their combination with other extracted features and the database used. For example, while Coastland (H) habitats obtain slighter better recall and precision results, their increase in accuracy is not as significant as the case of Heathland (D) habitats. In a way, it seemed like the medium-level features had different levels of impact depending on the habitat types. In an effort to understand the variable effect of medium-level features, we revised the feature vectors generated by our medium-level knowledge and found that, in most cases, lower increases in accuracy were caused by uncertain answers to our set of questions. That is, the users who had annotated the photographs had chosen lower certainty levels, generally between 0 and 2, when classifying these habitats. Moreover, some of the users, in an effort to collect as much information as possible, had created and labelled the same polygons containing habitats with two or more annotations, all of them with low levels of certainty. This was particular prominent in the case of Cliffs, both Maritime and Inland, and Intertidal habitats. Close to two thirds photographs containing cliffs did not contain visual clues about whether or not the cliff was situated near water. Consequently, some of the users decided to annotate the images with two annotations, "Cliff - water" and "Cliff - no water" at the same time, assigning both answer low certainty levels. This was also particularly prominent in Intertidal mosaics, in which users' were not sure about the distinctions between "Shingles" and "Sand". This practice, a direct consequence of involving humans in the classification process, led to some lower quality features being extracted. However, an easy method to solve this problem, which we could not carry out due to time constraints, is to have more than one user classify each photograph in the database. That way, each photograph would generate several medium-level feature vectors, ideally between four and seven, which could then be combined using weights so more common answers would receive higher weight than less frequent or more uncertain answers. By weighting medium-level features a single user's uncertainties would not affect the classification process as directly. We consider this improvement as part of the future work that will be discussed in Chapter 10.

Considering the other different features we have selected and extracted, pattern features continue being the best option of a more accurate classification. Moreover, texture features remain the most inaccurate and unstable features both when used on their own and when combined with medium-level features. On the other hand, colour and all the features combined together obtain reasonably good results which, while not as accurate as pattern features, outperform texture features in all cases in both datasets.

Finally, comparing the performance of both datasets, it can be observed that, as studied in Chapter 7, the habitats with more instances in each database are the ones which are

more accurately classified. Moreover, while the differences in results were more striking when we compared RF and RPFs, in this case, the results obtained from both datasets are quite similar in overall precision and recall terms.

In summary, these series of experiments served to corroborate that the inclusion of medium-level features helps our framework obtain higher recall and precision results. The inclusion of semantic information has been of particular help with complex habitats, such as Trees and Hedges (J.2.3) habitats. In these cases, the recall, and the precision to a lesser extent, experiment a noticeable increase. However, there are still improvements that could be done, specially in terms of improving precision results of habitats such as Coastland (H), Rock Exposure and Waste (I) and Heathland (D). Even though these habitats experiment a tangible increase, they still obtain low precision results. Since the inclusion of external semantic information yielded such promising results, we decided to study which other types of information could be used to aid the classification process. With this in mind, we started to consider the inclusion of metadata from the photographs, as the next type of information to include in our classifier.

8.5.3 Visual Results

Figure 8.8 and Figure 8.9 present two particular examples from H1K and H3K, respectively. Moreover, Table 8.5 and Table 8.6 show the five most probable results obtained from with experiments. Additionally, correct results are shown in bold and italics.

Both of these examples serve to further illustrate the effects of medium-level features in the classification process.

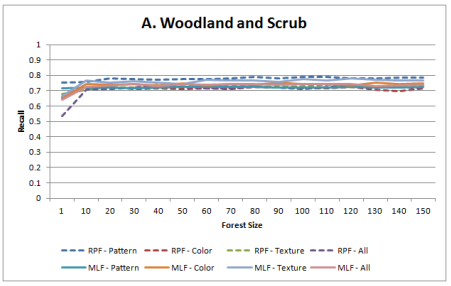
Table 8.8 shows how only the inclusion of semantic information is able to correctly classify the artificial habitat of Fence (J.2.4) in all cases. A similar situation is shown in Table 8.6, in which medium-level features ensure the classification of the unseen sample as Maritime Cliff (H.3) in three of the four testing scenarios.

8.6 Concluding Remarks

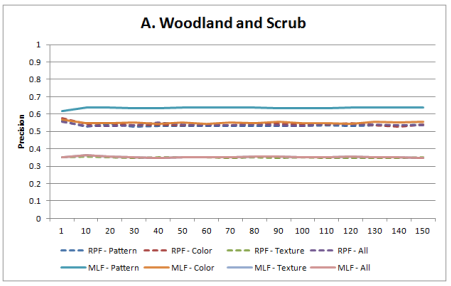
In this chapter, we have presented the second type of features that are extracted from the ground-taken photographs: medium-level features. These features are the fifth contribution of this thesis. We propose the inclusion of semantic information as a method to overcome the limitations that visual features present when distinguishing between visually similar classes, such as the case of habitat classification. We have used a Human-In-The-Loop approach to extract semantic information and we have transformed this

information into medium-level features that are used as the input of our Random Projection Forest classifier.

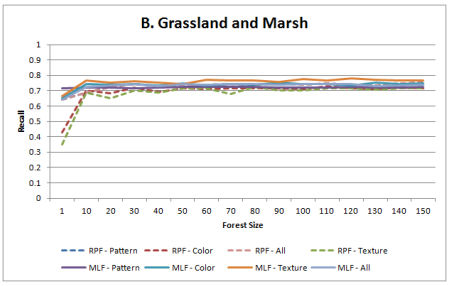
Experiments have shown that the inclusion of medium-level features improved the performance of our Random Projection Forest classifier, with their combination with pattern features yielding the most stable results. Complex and artificial habitats, in particular, have benefited considerably with their addition in our framework. In the next chapter we will present our final contribution: a location-based voting system for our classifier designed to use the geo-references from the ground-taken photographs to improve the performance of our classifier.



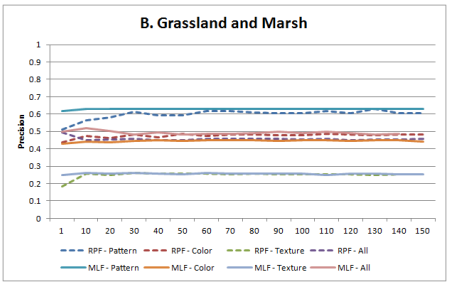
(a) A. Recall



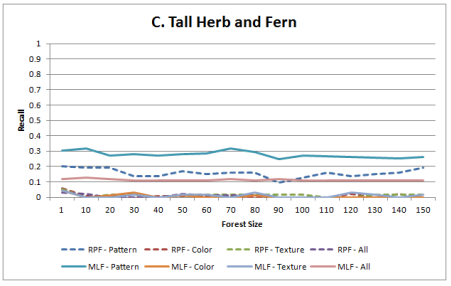
(b) A. Precision



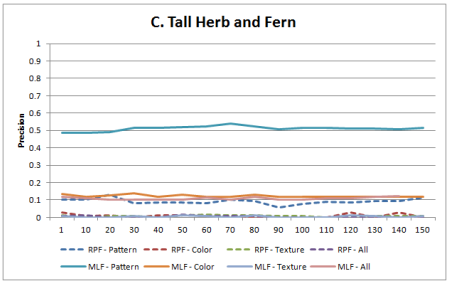
(c) B. Recall



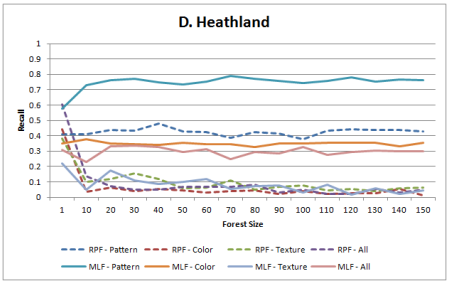
(d) B. Precision



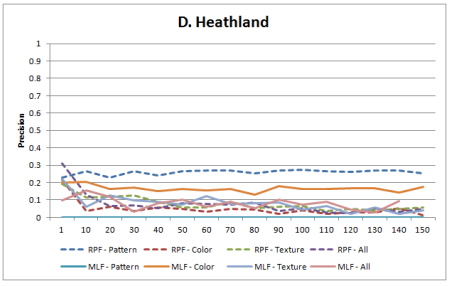
(e) C. Recall



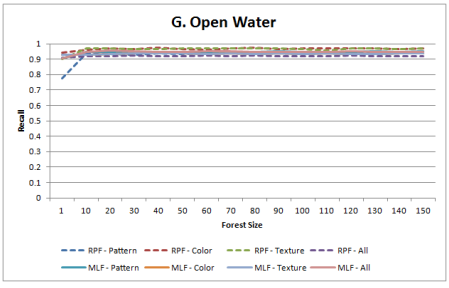
(f) C. Precision



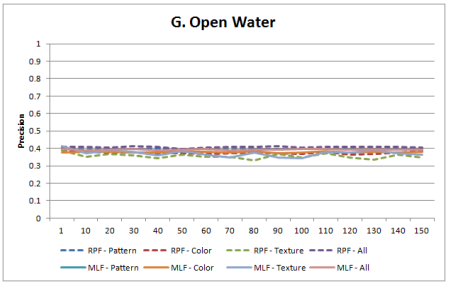
(g) D. Recall



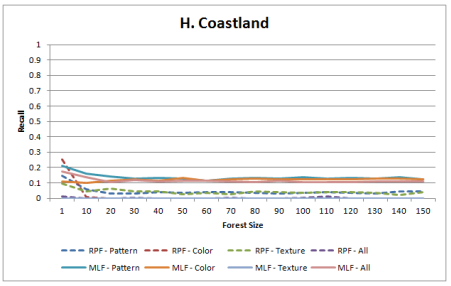
(h) D. Precision



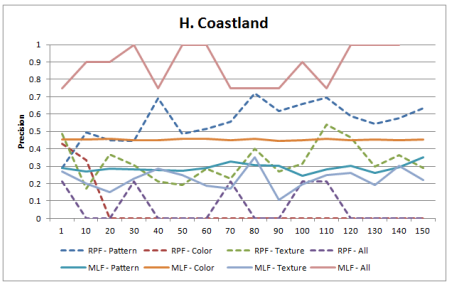
(i) G. Recall



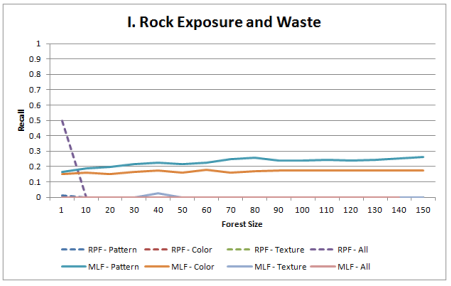
(j) G. Precision



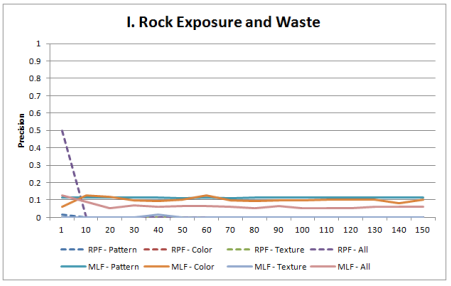
(k) H. Recall



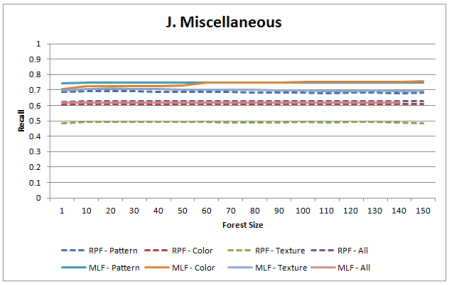
(l) H. Precision



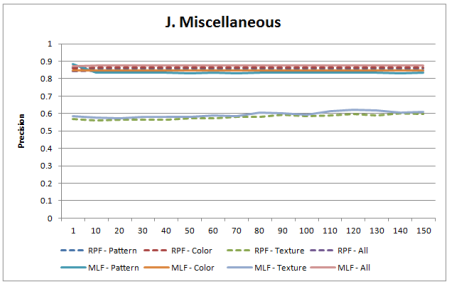
(m) I. Recall



(n) I. Precision

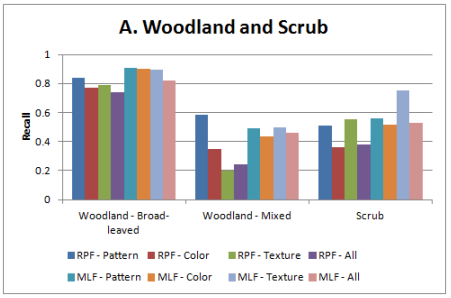


(o) J. Recall

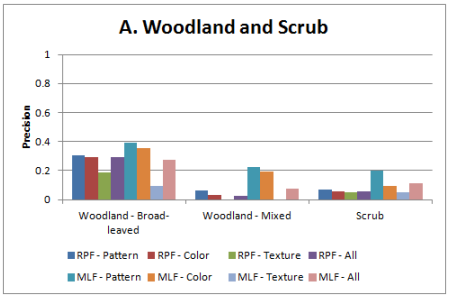


(p) J. Precision

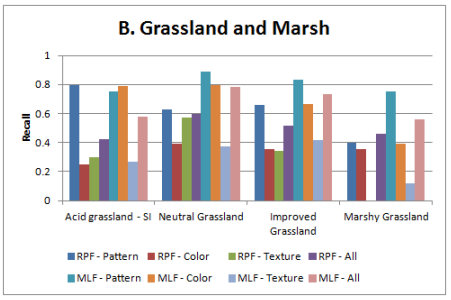
FIGURE 8.5: (Cont.) Medium-Level Features. Recall and precision results for first-tier habitats from Habitat 3K



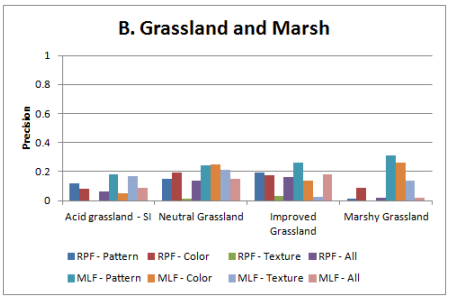
(a) A. Recall



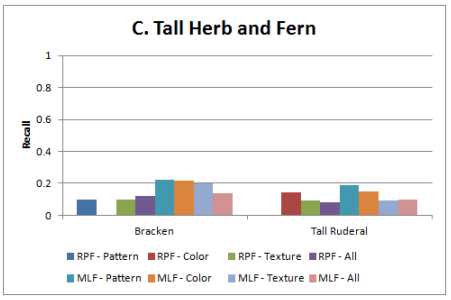
(b) A. Precision



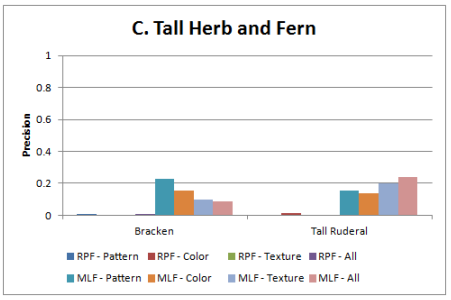
(c) B. Recall



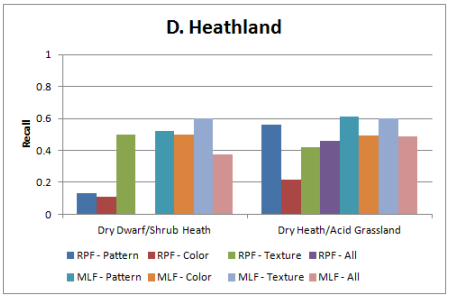
(d) B. Precision



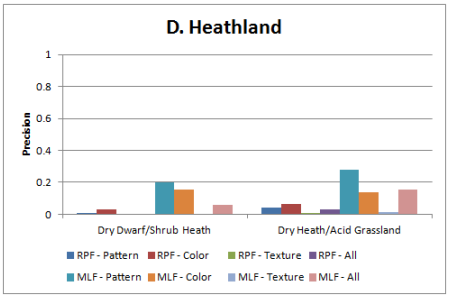
(e) C. Recall



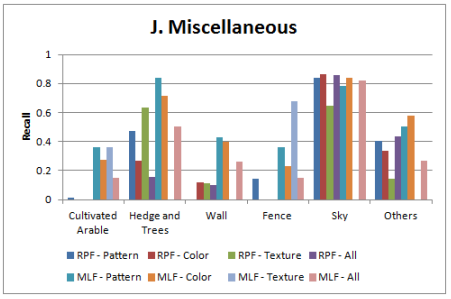
(f) C. Precision



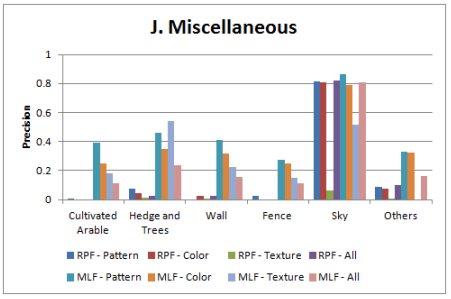
(g) D. Recall



(h) D. Precision

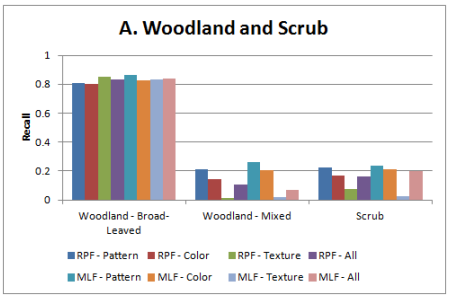


(i) J. Recall

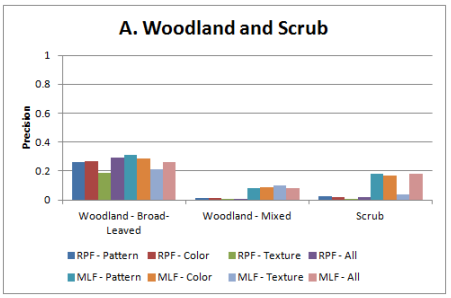


(j) J. Precision

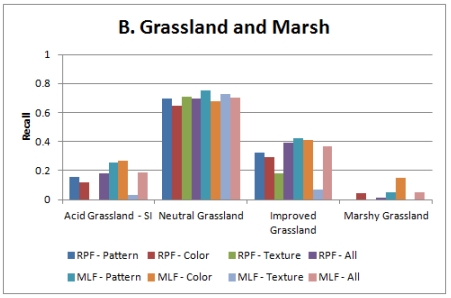
FIGURE 8.6: Medium-Level Features. Recall and precision results for second- and third-tier habitats from Habitat 1K



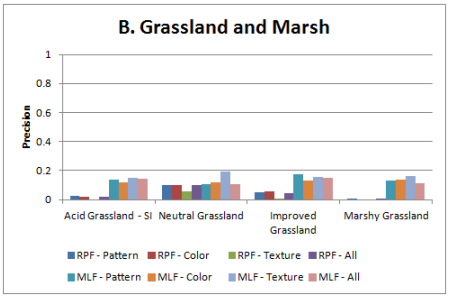
(a) A. Recall



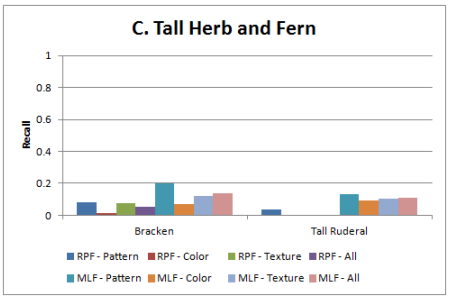
(b) A. Precision



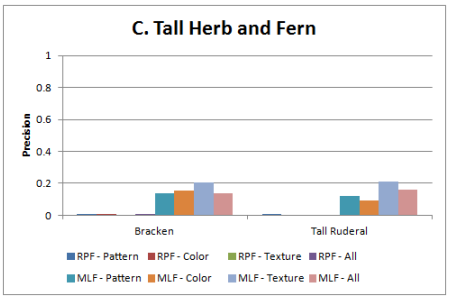
(c) B. Recall



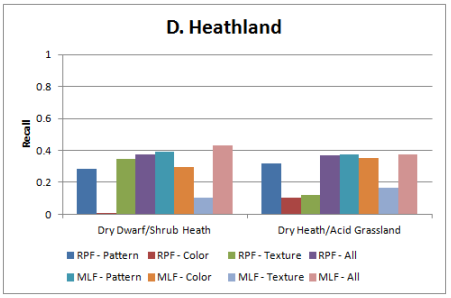
(d) B. Precision



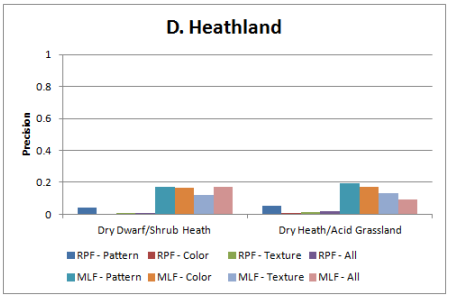
(e) C. Recall



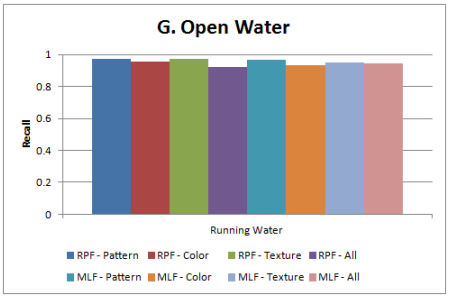
(f) C. Precision



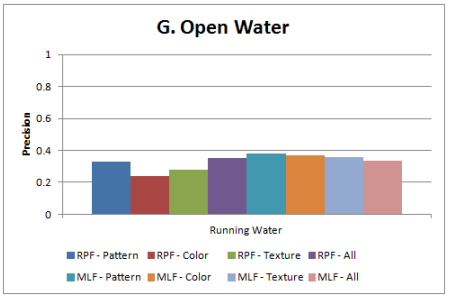
(g) D. Recall



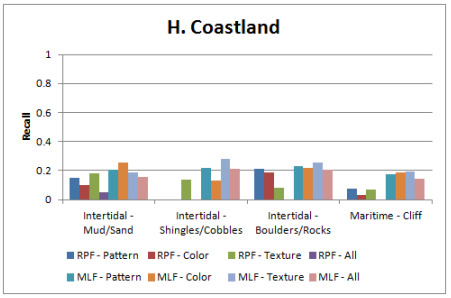
(h) D. Precision



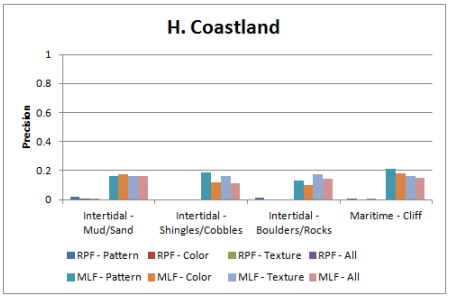
(i) G. Recall



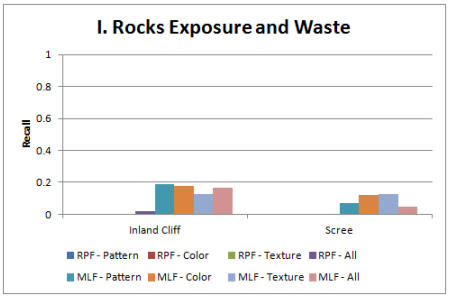
(j) G. Precision



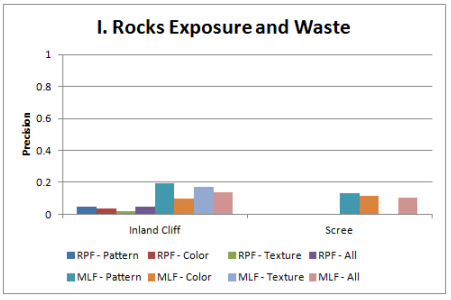
(k) H. Recall



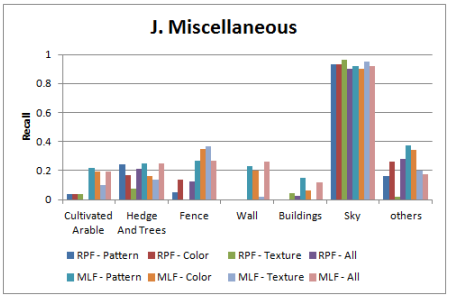
(l) H. Precision



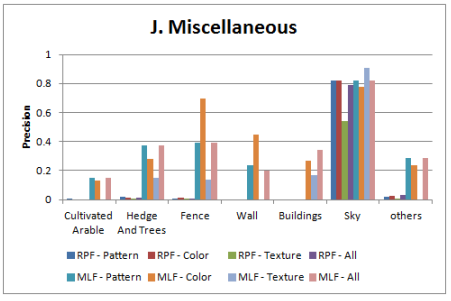
(m) I. Recall



(n) I. Precision



(o) J. Recall



(p) J. Precision

FIGURE 8.7: (Cont.) Medium-Level Features. Recall and precision results for second-tier habitats from Habitat 3K

FIGURE 8.8: Visual Example From H1K. Habitats present are: Improved Grassland, Woodland - Broad-leaved and Fence.

TABLE 8.5: Results. We show the five most probable results obtained with our experiments. Note how the use of medium-level features is the only approach which can successfully classify the Fence habitat.

Features Extracted		RPF	MLF and RPF
	Pattern	***Woodland - Broad-leaved***	***Woodland - Broad-leaved***
		Neutral Grassland	***Fence***
		Acid Grassland - SI	***Improved Grassland***
		Scrub	Scrub
		Improved Grassland	Tall Ruderal
	Color	Scrub	***Fence***
		Woodland - Broad-leaved	***Woodland - Broad-leaved***
		Improved Grassland	Scrub
		Neutral Grassland	***Improved Grassland***
		Acid Grassland - SI	Neutral Grassland
	Texture	***Woodland - Broad-leaved***	***Woodland - Broad-leaved***
		Neutral Grassland	***Fence***
		Improved Grassland	Scrub
		Dry Heath/Acid Grassland	***Improved Grassland***
		Scrub	Neutral Grassland
	All	***Improved Grassland***	***Woodland - Broad-leaved***
		Woodland - Broad-leaved	***Improved Grassland***
		Dry Heath/Acid Grassland	Neutral Grassland
		Scrub	***Fence***
		Bracken	Scrub

FIGURE 8.9: Visual Example From H3K. Habitats present are: Running Water and Maritime Cliff.

TABLE 8.6: Results. We show the five most probable results obtained with our experiments. Note how the use of medium-level features is the only approach which can successfully classify the Maritime Cliff habitat in three of the four scenarios tested.

		RPF	**MLF and RPF**
Features Extracted	**Pattern**	***Sky***	***Running Water***
		Others	Scree
		Neutral Grassland	***Sky***
		Dry Dwarf/Acid Grassland	***Maritime Cliff***
		Running Water	Inland Cliff
	Color	***Sky***	***Maritime Cliff***
		Running Water	***Sky***
		Wall	Scree
		Dry Heath/Acid Grassland	***Running Water***
		Building	Wall
	Texture	***Sky***	***Sky***
		Running Water	Inland Cliff
		Improved Grassland	***Maritime Cliff***
		Intertidal Boulders/Rocks	Wall
		Wall	***Running Water***
	All	Dry Dwarf/Acid Grassland	***Sky***
		Sky	Inland Cliff
		Running Water	Wall
		Intertidal Boulders/Rocks	***Running Water***
		Others	Others

Chapter 9

Location-Based Voting System

As mentioned in Chapter 6, Random Forests are characterised by a set of parameters, such as the size of the forest, the nature of the features used as the input, the split function and the voting system used to combine the predictions obtained in each tree of the forest. In the two previous chapters, we introduced modifications to two of these parameters, the input features and the split function of each internal node, and we studied their effects on automatic habitat classification.

In this chapter, we propose a modification on the last of the parameters mentioned above: the voting system. We present a novel voting system based on the use of the geographical information stored in photographs of our database. We benefit from the natural properties of habitats, which entail that neighboring areas have similar geological and ecological properties. As a result, their habitats can be extremely similar. Therefore, the predictions generated by leaves with photographs which are close to the unseen test photograph should have more weight in the decision making process. Consequently, this chapter presents the last contribution of this thesis and, incidentally, the last element of our image-annotation framework: a voting system based on the inclusion of geographical location during testing.

Experiments were carried out to evaluate the effect of location-based weighted voting during testing. We have calculated the recall and precision of the complete framework and results show that the whole system outperforms all other methods tested in this thesis, including traditional Random Forests. This makes our complete image-annotation system, which combines Random Projection Forests with low- and medium-level features and location-based testing, to our knowledge, the most accurate automatic alternative to manual habitat classification for the complete categorization of Phase 1 habitats.

This chapter is organised as follows: Section 9.1 describes the motivation behind the idea of using geographical information during testing. Section 9.2 explains how GPS coordinates can be used to weight the different predictions offered by the decision trees in our Random Projections Forests. Moreover, Section 9.3 describes the experiments carried out, Section 9.4 shows the results obtained from these experiments and discusses their significance in comparison with the results obtained in previous chapters. Finally, Section 9.5 offers concluding remarks.

9.1 Motivation

As discussed in Chapter 2, in traditional Random Forests, each tree in the ensemble casts a unit vote on the classes present in the unseen test photographs. This implies that all the decision trees in the forest are equally good at classifying an unseen test photograph. However, this is often not the case, as some trees have been proven to be better at classifying than others [152]. In this situation, it would be ideal to be able to somehow identify and select the most accurate trees and to prioritise their predictions over the predictions from less accurate trees. That is the goal of a weighted voting system. In essence, the aim of modifying the traditional voting system used in RFs is to find a mechanism in which more accurate trees are given more importance in the decision-making process, while, at the same time, not ignoring the other decision trees in the forest completely.

In our case, we decided to focus on modifying the voting system as our final contribution for two main reasons. First, to use the data that was already stored our databases to the fullest. In other words, we wanted to extract and use as much of the information already stored in our database as possible. The same way that the use of low-level features only, as shown in Chapter 6, entailed that important semantic information, already present in the photographs, was not taken into account when annotating images, we felt that the current implementation of Random Projection Forests did not take into consideration other extremely crucial information already stored in our database, the geographical information of the images, which could improve our results.

As can be seen, in comparison to other FGVC-oriented databases, such as the CUB-200-2011 Database [199] and the Leeds Butterflies Database [200], Habitat 1K and Habitat 3K present an interesting difference which we have exploited in our location-based voting system. Photographs in most of the FGVC-oriented databases are not related to one another. Taking an example from [199], a photograph of a bird is in no way related to other photograph of a bird, whether it is the same kind of bird or not. In other words, there is no apparent way of linking the two photographs. In our dataset, however, this is

not the case. Photographs are extremely related to one another. This relationship can be measured by their geographical location, which is stored as their GPS positioning. Therefore, the information that we have extracted from a photograph can, in fact, affect the classification and annotation of the photographs which are linked or related to it.

Second, to benefit from the particular properties of automatically classifying habitats. The reason we choose to work with geographical location instead of, for example, the time of the day a photograph was taken, is related to the intrinsic characteristics and the nature of the problem. It is not usual for habitat types to change quickly within an area. It can happen in some rare cases, for example the abrupt change between a Maritime Cliff (H.8) and the Ocean (G.2). However, in most cases, the geological properties of an area will result in similar habitat classes. For example, as exemplified by the photographs taken in New Forest as part of Habitat 1K, all the woodland present in the area was Broad-leaved Woodland (A.1.1). Similarly, most of the grassland captured around the lake in Titchfield Haven was Marshy Grassland (B.5). Correspondingly, since most habitat properties do not generally change abruptly, geographically close areas will have similar ecological characteristics. Therefore, in our case, we decided to take advantage of this geographical property of habitats during the testing phase.

Moreover, the benefits of this location-based voting system could be applied to the cases in which abrupt changes were to happen, such as the Maritime Cliff and the Ocean example mentioned above. In this case, the only requirement to successfully apply this location-based voting system would be to have a sufficiently robust database that contemplated this type of abrupt change. This would not be difficult, since abrupt changes often happen between the same types of habitats. Therefore, in a way by storing multiple photographs with these "abrupt" changes occurring, they would be stop being considered "abrupt" and the modified voting system could be applied.

Research has been developed on voting systems, with some alternatives suggesting the inclusion of weights for the predictions as a particular convenient methodology [152]. Following this approach, the random forest will be constructed in a traditional fashion and, during testing, the most accurate trees' predictions will have more weight in the final prediction. That is, their vote will be more important.

There are two main points that are important to notice when modifying the voting system to include weighted predictions.

First of all, as can be inferred, the notion of a "more accurate" decision tree needs to be clarified. That is, what constitutes an accurate tree, or an inaccurate tree, needs to be clearly specified. This is extremely problem-dependent, since the type of source data to be used can vary tremendously. Moreover, not only the nature of the problem will

affect this choice, but also, the type of measures that need to be extracted to evaluate the RF's performance. For example, if only recall is to be calculated, like we did in Chapter 4, the definition of an "accurate" tree would not need to be as strict, since recall is usually a more relaxed measure. However, if precision is to be included in the performance metrics, "tree accuracy" will have to be determined very carefully, since precision tends to be a very difficult measure and giving priority to an incorrect set of trees could result in performance metrics being disastrously low. Consequently, one of the main challenges of weighting predictions is to actually decide how to structure the assignation of weights.

Second of all, it is crucial that the less accurate trees should not be ignored. As introduced in Chapter 2, one of the strengths of ensemble classifiers, and of RFs in particular, is the fact the ensemble benefits from having many weak learners generating and offering predictions. Moreover, RFs benefit not only from being an ensemble classifier, but for also introducing randomness in the classification process. The combination of weak learners and randomness is one characteristics that makes RF such a robust classifier. Consequently, discarding trees' predictions only because they are non-compliant with the "accuracy" measures that have been established would, eventually, hurt the overall performance of the whole forest.

To solve the first, as mentioned above we use the GPS coordinates of the photographs to establish how accurate the trees within the forest are. This is not a problem since all the images in the database are geo-referenced. This is done by calculating the distance between the test sample and the images that are in the leaf node the sample has reached. Then, we weight the prediction that each tree casts according to their distance. By minimizing the distance and assigning weight, the predictions of trees with closer leaves influence the final classification more.

Moreover, to solve the second issue, our implementation of the weights makes sure that all the predictions are taken into account. This is done by varying the weights between 1 and 2, instead of the usual [0,1] interval. Consequently, even the least accurate tree in the forest, that is, the tree whose leaf images are the furthers away from the test sample, will be taken into consideration in the decision-making process.

In summary, we will use the geographical information already stored in our database to create a new voting system that will prioritise the predictions of the trees which are closer to the unseen test photograph. In essence, our system can be seen as taking into consideration two types of closeness: we take into consideration visual closeness during training and, then, geographical closeness during testing.

It is important to notice that there are some limitations, discussed in much more depth in Section 9.4, that come from using this location-based voting mechanism. However, they are mainly related to the type of data that we have used. Remote sensed data, such as aerial or satellite imagery, lack the level of detail that we needed to classify within Phase 1 species. However, they present a clear advantage over ground-taken photographs and that is the structured layout of the images. While aerial and satellite images orientation and perspective is always the same, orthogonal to the ground, layouts and perspectives can vary a great deal in ground-taken imagery. This results is a dichotomy between the geographical location of the place where the picture was taken and the actual geographical location of the habitats present within the photograph. This limitation is particularly manifested in the Geograph 2K database, since the collection of those photographs was done using crow-sourcing methods and there was less control over the characteristics of the photographs in terms of layout and perspective.

9.2 RPFs and Location-Based Voting

As mentioned in the previous section, the voting-mechanism modifications introduced in this chapter affect only the testing phase of the Random Projection Forests. Correspondingly, the training phase will be the exact same as previously described. Consequently, in order to include this modification, we first need to construct the RPFs as shown in Chapters 7 and Chapter 8. As in the other testing scenarios, we are using RPFs and previously annotated ground-taken photography, commonly referred to as the training set, with the aim of annotating unseen photographs with the habitats present in them.

Once the training phase has finished and the RPFs have been constructed, we start the testing phase. The testing methodology followed is similar to the original RPF design. During testing, features are extracted from the previously unseen images and the resulting vector representing the test photograph is injected at the root node of all the trees of a forest. These features can be the low-level features or the combination of low-level features and medium-level features. The only requirement is complete agreement between the features extracted to create the RPFs and the features extracted from the unseen test photograph.

At each split node, the inner product between the test image feature vector and the nodes random projection vector is calculated and it is distributed to either the left or the right child node based on whether the inner product is greater or smaller than the nodes optimal threshold value. This process is repeated until the data reaches one leaf node in each of the decision trees in the forest.

It is at this step that the location information stored in our database becomes relevant. This location information is stored in the Exif tags associated to each of the photographs in our database. The Exchangeable Image File Format (Exif) tags store a variable number of metadata regrading the characteristics of the photographs. It covers a broad spectrum of information, such as date and time of the photograph, the make and model of the camera used, the shutter speed and the geographical location. In particular, the geographical location is stored as the latitude and longitude coordinates of the photograph.

We use that latitude and longitude coordinates to calculate distances between the test photograph and each of the photographs in each of the leaves that the test sample reached when it was injected in the roots. In particular, we have chosen the to use the Haversine distance to calculate the distance between photographs since it is more accurate than the Euclidean distance [172]. We use MATLAB and the distance function implementation developed in [172] to extract this information and to calculate the distances between photographs.

Once all the distances are calculated, we attribute weight depending on the mean distance between the test sample and the samples of each leaf. Our weights are in the [1,2] range, with 1 being the furthest and 2 being the closest. As mentioned previously, the reason the weights vary from 1 to 2 instead of varying from 0 to 1 is because we want to take all trees into consideration, even those which might be geographically further away.

Finally, the final probability distribution for all the habitats in a forest with N trees is calculated as shown in Equation 9.1.

$$P(h) = \sum_{t=1}^{N} w(t) P^{T_t}(h) \tag{9.1}$$

Where $P(h)$ is the final probability of occurrence of the habitat h in the unseen test photograph, $P^{T_t}(h)$ is the probability of the habitat h in each of the leaf nodes that the test vector reaches and $w(t)$ is the weight of each prediction. This weight is calculated as shown in Equation 9.2.

$$w(t) = 1 + \frac{1}{N-1} O(t) \tag{9.2}$$

where

$$O(t) = \frac{No(t)}{max_t o(t)} \tag{9.3}$$

and

$$o(t) = \frac{1}{N_t} \sum_{i=1} N_t Distance[GPS(TI), GPS(I_i^T)] \tag{9.4}$$

where N_T is the number of images and I_i^T is the ith image in the leaf node of tree T that the testing image reached, TI is the testing image and $GPS(x)$ is the GPS location of image x.

By following the previous equations, the predictions from leaves whose samples belong to the closest photographs to the unseen test image will have more weight in the final prediction.

It is important to notice that this approach will work more accurately when the photographs in the database are close to each other, as is the case of Habitat 1K. In Habitat 1K, four areas were thoroughly mapped. Consequently, each photograph will have at least one more photograph in the same area. However, if the photographs were to be very scattered, as is the case of Geograph 2K and, consequently, some of the photographs from Habitat 3K, using the distance alone could prove to be counterproductive, since the closest samples could still be considered to be very far away in reality. Results obtained using our Habitat 3K database, shown in Section 9.4 confirmed this. In this case, instead of using a distance measure alone, it would be more appropriate to determine a radio or a threshold before assigning weights. This way, only predictions from trees whose leaf samples are within the radio would be weighted and the rest of the predictions would carry a weight of 1.

9.3 Experiments

A series of experiments were carried out to test the addition of the location-based voting system. Following the structure of both Chapter 7 and Chapter 8, we decided to focus our experiments on extracting features from the images as a whole. Moreover, we compare performances between RPFs with medium-level features and RPFs with medium-level features and the location-based voting mechanism. We set up these experiments with the specific goal of studying the effect of our novel voting mechanism. Correspondingly, we studied this by generating results on the performance of RPFs when varying an specific set of parameters. These parameters are:

- Location-based voting system: We study the effect of our location-based voting system by comparing its performance in terms of recall and precision with the results obtained in the previous chapter.

- Colour, pattern, texture and medium-level features: Following our findings from the previous two chapters, we extract and compare the performance of our classifier when colour features (Colour Histogram, Colour Moments), texture features (Tamura, GLCM), pattern features (CPAM) and all of them combined are extracted and combined with medium-level features. We hope that the combination of visual semantic and geographical information will increase the accuracy of second- and third- tier habitats. We also compare performances of these features against the performance of the "Other Features", a combination of six of the most common visual features currently used in Computer Vision problems (GB, GIST, SIFT, SSI, PHOW, PHOG). As in previous chapter, the results regarding the "Other features" have not been included in the graphs to help with the visualization of the most relevant features, but they will described and discussed.

- Database: Given the different nature of the databases created in this thesis, Habitat 1K being collected under controlled circumstances and Habitat 3K being collected using crow-sourcing methods, we also aim to study the effect of semantic information on their performance. We are particularly interested in the performance of the new voting system when applied to H1K, since all the photographs are very close to each other. We project less accurate results for H3K given the geographical sparsity of the photographs.

Moreover, we decided to compare Random Projection Forests with medium-level features against Random Forests with a location-based voting system to obtain a more in-depth study of the effect of the weighted predictions. Furthermore, to ensure consistency between the results, we follow the same methodology as in Chapter 7 and Chapter 8 and we calculate the recall, precision and confusion matrix of results obtained.

9.4 Results

In order to assess the impact of our location-based voting mechanism and Random Projections Forests, we have tested ten scenarios with each or our databases. These scenarios are:

1. RPF with colour features and medium-level features. This scenario is referred to as MLF - Color in the result figures.

2. RPF with pattern features and medium-level features. We refer to this as MLF - Pattern in the result figures.

3. RPF with texture and medium-level features. This is called MLF - Texture in the result figures.

4. RPF with all three features linearly combined and medium-level features. This scenario is referred to as MLF - All in the result figures.

5. RPF with other features and medium-level features. In order to make visualization easier, we have not included these results in the graphs. However, the findings from this set of experiments will be commented and compared with the results obtained in the other experiments.

6. RPF with colour features, medium-level features and the location-based voting system. This scenario is referred to as GPS - Color in the result figures.

7. RPF with pattern features, medium-level features and the location-based voting system. We refer to this as GPS - Pattern in the result figures.

8. RPF with texture, medium-level features and the location-based voting system. This is called GPS - Texture in the result figures.

9. RPF with all three features linearly combined, medium-level features and the location-based voting system. This scenario is referred to as GPS - All in the result figures.

10. RPF with other features, medium-level features and the location-based voting system. In order to make visualization easier, we have not included these results in the graphs. However, the findings from this set of experiments will be commented and compared with the results obtained in the other experiments.

Similarly to Chapter 7 and Chapter 8, we divided the results obtained according to the level of detail of the habitats classified. We have calculated the recall and precision for first tier-habitats in Section 9.4.1, while Section 9.4.2 presents results for second- and third- tier habitats. We compare each set of results with the Random Projections Forests results obtained in the previous chapter. Finally, we present some visual examples obtained during our testing in Section 9.4.3.

9.4.1 First-Tier Classes

Figure 9.1 shows the recall and precision results obtained in the testing scenarios introduced previously when using features extracted from whole images from H1K as the input. On the other hand, Figure 9.2 shows the same metrics when testing our framework with features extracted from whole photographs from H3K as the input. We tested

forests with sizes ranging from 1 to 150 and with depths ranging from 2 to 10. However, in order to present the results in a clear and concise manner, we set their depth to 9 in the previous figures. Nevertheless, the performance of both systems was similar and stable in all cases.

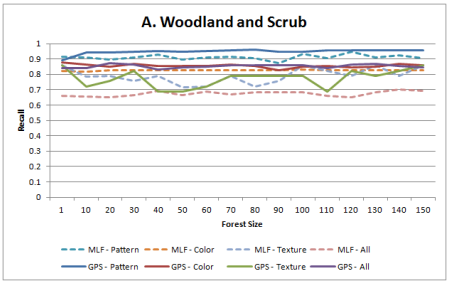
Looking at the results as a whole, it can be appreciated that the recall measures remain more accurate that the precision measures, similarly to the results presented in Chapter 7 and Chapter 8. There is only one exception to this case, present as well in Chapter 7, and that is the classification of Miscellaneous habitats in both H1K and H3K. In this case, the voting system is able to return a much higher precision than recall, reaching even 90% accuracy.

Moreover, it can be seen that this difference in results is not as significant as in the previous chapters. In the majority of cases, this is due to precision results experimenting a noticeably increase in accuracy. This is clearly exemplified in the case of Open Water (G). In Chapter 8 we discussed the dip between its recall and precision results, which was close to 50%. However, with the voting system, this difference has decreased to close to 40%, with the recall remaining at around 100% and the precision increasing from 40% to 60%. This is a direct consequence of our location-based voting system and the fact that most of the coastland photographs in H3K being from the same area, the south England.

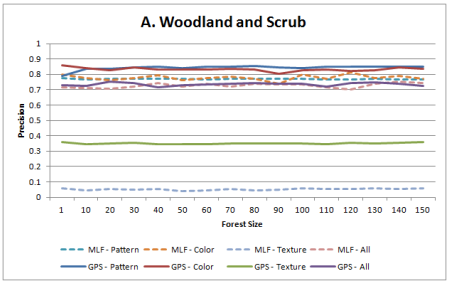
Moreover, Woodland and Scrub (A) and Grassland and Marsh (B) continue being the most accurately classified habitats in H1K, and, along with Open Water (G), the most successfully classified habitats in H3K. On the other hand, Rock Exposure and Waste (I) and Tall Herb and Fern (C), even though they experiment a slight improvement in their results, remain the most difficult to classify.

If we look into the experiments more in depth, we find that, regardless of the combination of features extracted or the databases used, our location-based voting system is able to outperform the RPFs with medium-level features in most cases. In those rare occasions in which the use location during testing does not outperform RPFs, such as is the case of Woodland and Scrub (A) classification using texture features, it is shown that the inclusion of the location during testing matches the performance of the RPFs without the weighted voting system. This accuracy, in turn, makes the complete system tested in this chapter, composed of low-level visual features, medium-level features, RPFs and location-based voting, the most accurate alternative for automatic habitat classification presented in this thesis and, to our knowledge, developed to date.

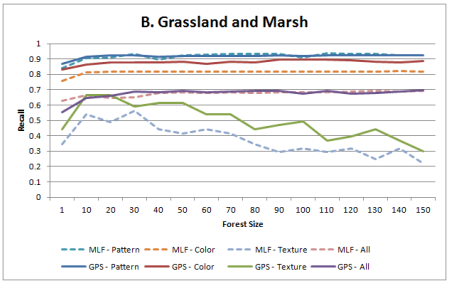
Another set of interesting results can be extracted when looking into the combinations of features and location-based voting. Similarly to the trend presented in Chapter 7 and



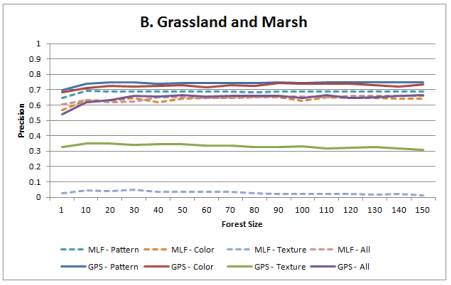
(a) A. Recall



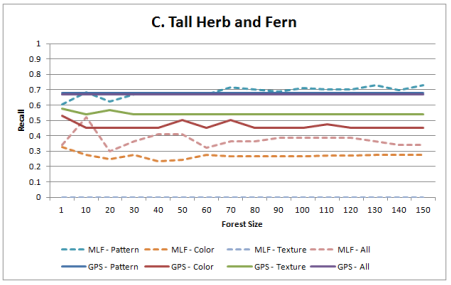
(b) A. Precision



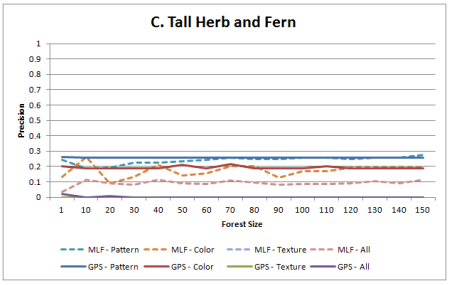
(c) B. Recall



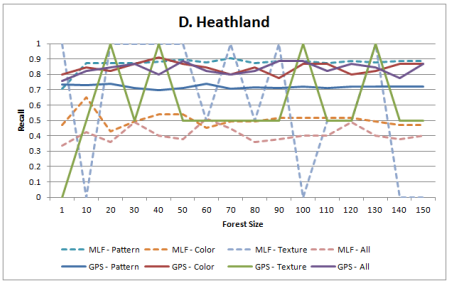
(d) B. Precision



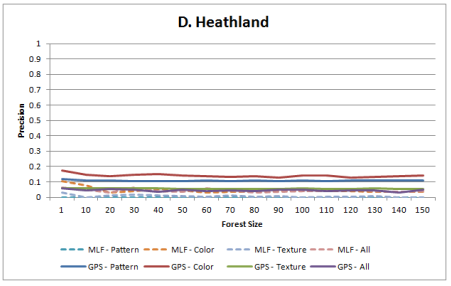
(e) C. Recall



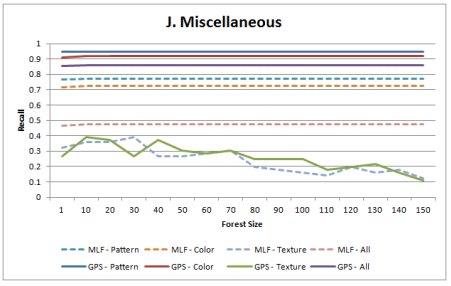
(f) C. Precision



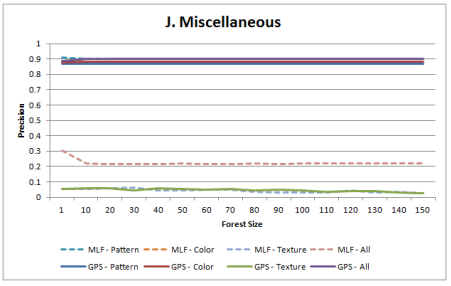
(g) D. Recall



(h) D. Precision



(i) J. Recall



(j) J. Precision

FIGURE 9.1: Location-based Voting System. Recall and precision results for first-tier habitats from Habitat 1K

Chapter 8, pattern features continue to provide the most accurate results in terms of precision and recall. Moreover, colour and all the features together continue having similar performances in terms of recall. However, colour features stand out when combined with our weighed voting system. Their combination actually obtains much higher precision than the combination of all features together and location-based testing. Additionally, in occasion, colour features are able to outperform the precision obtained from pattern features, if only slightly, as shown when classifying Heathland (D) habitats with H3K. This serves to further exemplify the importance of colour information in the classification process. However, pattern features continue generating the best overall balance between recall and precision results together. Finally, texture features, regardless of the inclusion of geographical location during testing, continue obtaining the less accurate results.

The combination of pattern features and weighted testing is particularly successful at classifying Tall Herb and Fern (C) habitats, one of the most difficult habitats to classify. This is not surprising in the case of H1K, where most instances of Tall Herb and Fern (C) were localised in New Forest, one of the surveyed sites. However, this improvement is also present in H3K. This seemed to indicate that our assumptions about the geographical location of the photographs present in H3K and their impact in the classification process needed revising.

We looked more closely to this phenomenon and observed an interesting situation when comparing the performances of both datasets more in depth. As mentioned previously, when setting up the experiments, we were expecting a definite increase in H1K precision and recall results, as a direct consequence of the database containing numerous photographs from only four particular sites. H1K photographs could be clearly separated geographically and, moreover, presented a comprehensive description of the habitats present in those for sites. Accordingly, results confirmed our expectations, particularly in two of the most difficult habitats to classify, Tall Herb and Fern (C) and Heathland (D). As can be seen in the results, both habitat classes experimented a raise over 10% in precision and recall accuracy.

On the other hand, we expected the improvements on H3K to be less significant since the pictures were more sparsely located. Instead of a small number of areas which were thoroughly mapped, H3K contained photographs distributed for the whole of Great Britain. However, we found that the inclusion of geographical location impacted quite positively the results of H3K, particularly in terms of precision. An example of this is shown when classifying Rock Exposure and Waste (I), Heathland (D) and Tall Herb and Fern (C). In order to study this phenomenon we looked more closely at the geographical distribution of our dataset and found that, while they photographs were indeed more

sparsely distributed, in the cases where the recall and precision metrics experimented a significant increase in accuracy, we had inadvertently chosen photographs who where located within the radio we had established during testing. In retrospect, this was not surprising, since the habitats included in this phenomenon were less frequent than those which were not, for example, Inland Cliff occur less frequently thorough Great Britain than Woodland habitats. Therefore, it is more likely that the larger the number of instances that these less-frequent habitats are, the more likely their photographs were taken around the same area. Consequently, small clusters of ten to fifteen photographs were formed and our framework had been successful enough that, when testing unseen samples, these had reached leaves in which photographs from these clusters were present. In essence, by classifying only the visual and semantic information from the images, we had still managed to include geographical information in an indirect way.

In summary, from all the modifications presented in this thesis during the last three chapters, the combination of RPFs with medium-level features and the inclusion of geographical location in the testing phase has generated the most accurate performance when classifying first-tier habitats. This whole framework outperformed traditional Random Forests in all cases and obtained recall and precision results over 50% in most of the habitats present in both datasets. Consequently and taking into consideration these results, we projected that the inclusion of geographical location during testing would improve the recall, and particularly the precision, of second- and third- tier habitats.

9.4.2 Second-Tier and Third-Tier Classes

Figures 9.3 show the recall and precision results obtained in the same testing scenarios as in Section 8.5.1. Additionally, Figures 9.4show the same metrics when testing our framework with H3K. Similarly to the other testing scenarios in Chapter 7 and Chapter 8, we are using the whole photographs when extracting the features. We tested the forests varying their size between 1 and 150 and their depth between 2 and 10. However, in order to make the results easier to visualise, we have set the size of the forests to 120 and the depth of the forests to 9 in the graphs, since the performance of both systems was similar and stable in all cases.

Looking at the results as a whole, it can be appreciated that recall measures remain being the most accurate in all scenarios tested. However, following the trend discussed in the previous section, the differences in the precision and recall results is smaller than in previous second- and third-tier testing. Moreover, habitats from the classes Woodland and Scrub (A) and Grassland and Marsh (B) continue being the most easily classified, due to the large amounts of photographs from them in both databases.

On the other hand, Tall Herb and Fern (C) and Heathland (D) habitats remain the most difficult to classify in H1K, their recall and precision, particularly in the case of Heathland mosaics (D.1 and D.2) has improved considerably. Similarly, Coastland (H) and Rock Exposure (I) habitats remain the less accurate in the case of H3K, but they have also experimented a noticeable improvement in terms of recall and precision. In particular, it can be seen that the problem we had before between classifying between the Inland Cliff habitat and the Maritime Habitat is not as pronounced, with their recall being close to 25% in most cases and their precision being close to 20% in all cases. While these results are still low, they serve to demonstrate the impact that taking into consideration the geographical location of the photographs have.

Moreover, if we look at the results more in depth, we can see that the inclusion of location-based voting during testing as affected very positively the classification of second- and third- tier habitats. This modification obtains more accurate results in all cases, regardless of the database used as input. Consequently, it not only outperforms RPFs with medium-level tags, but also the original design of RPFs and the traditional RF implementation. As discussed in the previous section, this makes this whole framework the most accurate system of all presented in this thesis.

In terms of feature combination, we can see that pattern features remain the most accurate features in the majority of cases. This is clearly noticeable in the case of Heathland (D) in H1K and Coastland (H) habitats in H3K. Similarly to the results obtained when classifying habitat from the first-tier classes, colour features and all features put together obtain similar results, with all features together obtaining a slight better recall but a considerable less accurate precision in most cases. Furthermore, texture features continue being the least accurate and most unstable features when used with or without location-based voting. This further proves that pattern features are the best option for our framework because the collect the most relevant information in the most compact manner.

Regarding the different types of habitats, we can see that complex habitats in particular have benefited from the inclusion of geographical location in our framework. This is noticeable in the Hedge and Trees (J.2.3) results, which experiment an increase of recall and precision of over 10% in H1K and, perhaps even more strikingly clear, in the Heathland mosaics results, which increase their accuracy close to 15% in H1K. On the other hand, artificial habitats, particularly the habitats Wall (J.2.5) and Fence (J.2.4) obtain only slightly better results, not remotely close to the significant increase in accuracy that medium-level features entailed, as seen in Chapter 8. This is understandable, since the photographs from these habitats are fewer and were taken in many different locations,

while photographs from complex mosaic habitats, such as the Heathland mosaics, were more abundant and less geographically distributed.

Moreover, if we compare the performances of H1K against H3K, we can see that there are clear differences in their precision results. Although, in general, it can be appreciated that precision results remain low in all cases, it is in experiments with H3K where precision results obtain their lowest results, with some cases not even reaching 15% accuracy. In these cases, it can be appreciated that the inclusion of the geographical location during testing has not aided the classification process and, in some particular instances, such as Scree and Inland Cliffs, it has even damaged their classification. This is clearly due to one of the main limitations of ground-taken photography, previously discussed in Chapter 6. As we mentioned in Chapter 6, the ground-taken photographs that we are working with have a variety of layouts and they were taken from multiple perspectives. As a direct result of this, the location of where the photograph was taken, which is part of the metadata information stored in our database, might not accurately reflect the location of the objects present in that photograph. For example, a photographs taken with a wide perspective, cannot accurately store the geographic location of the habitats present in that photograph. An example of this in our database concerns photographs which show Inlnd Cliffs. Since Cliffs are generally large habitats, they can appear in photographs that were taken kilometers away from them. Therefore, in those cases, the geographical location of the photographs can hinder the classification process.

The reason this phenomenon affects H3K more prominently is because, as we have discussed at length, we had no control over the conditions under which the photographs from H3K were taken. This resulted in photographs from H3K representing much more variable conditions than photographs from H1K. In essence, this problem of the geographical location of the photographs versus the geographical location of the habitats within the photographs can be regarded as a direct consequence of using crowd-sourcing methods to collect photographs. On the one hand, using Geograph enabled us to collect a larger number of photographs in a much shorter period of time and, more importantly, it enabled us to collect instances from habitats that, given our geographical location, were impossible to access for us, i.e. cliffs and coastland habitats. On the other hand, we were required to to relinquish control over their characteristics and we had to accept a broader variability on their layout, lighting and perspective conditions.

In summary, this final modification to RPFs has shown a definite impact in the performance of our AIA framework, particularly in the case of complex habitats and second- and third-tier classification. Moreover, while some habitats still obtain low precision and recall, they have experimented clear improvements in recall and precision as a consequence of each modification we have introduced. This can be seen more clearly in Table

9.1, which presents the average recall and precision for all of the approaches presented in the previous chapters. We have averaged all other parameters (databases used, features extracted, forest size and tree depth) in order to show more clearly the effect of each of the added contributions.

TABLE 9.1: Average precision and recall results for all modifications of our framework. Each modification has entailed an improvement over the results obtained in the previous version of the framework.

	RF	RPF	MLF	GPS
Recall	0.313	0.408	0.7125	0.7315
Precision	0.26	0.265	0.38	0.43

In general, experiments have shown that ground-taken photographs are a promising source of information that can be successfully applied to Phase 1 habitat classification. Moreover, the FGVC nature of the problem makes an AIA framework specially fitting and Random Forest-based methodologies, such as the Random Projection Forests we have created, are specially suitable to be used in this framework, since they combine efficiency and accuracy. We have also seen how low-level visual features, specially pattern features, can be used to certain extent as the first step of the classification. The limitations these features present have been lessened with the inclusion of semantic information in the form of medium-level features. These features have helped establish that the inclusion of human input in the classification process, while requiring additional precautions, can be extremely beneficial for complex or similarly visual habitats. Finally, we have shown that, given the nature of the problem and the classes we aim to identify, we can benefit from the geographical properties of habitats. We have done so by introducing a location-based voting system that prioritises predictions of leaves whose samples are closer to the testing samples. This final improvement has provided an increase in recall and precision results in most cases and has made our Random Projection Forests with ground-taken photographs, medium-level features and location-based voting the most accurate automatic alternative to manual Phase 1 habitat classification. As a final note, it can also be seen that the larger the number of instances of each habitat, the more accurate the results both in terms of recall and accuracy, as exemplified by results obtained in all testing scenarios by Woodland and Scrub (A) and Grassland and Marsh (B). Consequently, we can only foresee that larger datasets and the more geographically close the photographs, the more accurate that the results generated in all three levels of habitat classes will be.

9.4.3 Visual Results

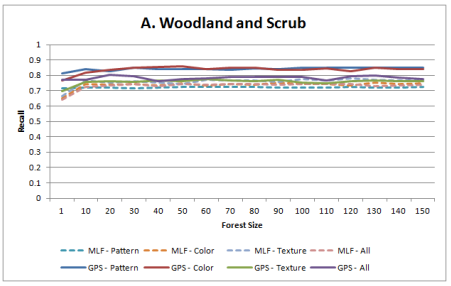
Figure 9.5 and Figure 9.6 present two particular examples from H1K and H3K, respectively. Moreover, Table 9.2 and Table 9.3 show the five most probable results obtained from the experiments.

Results from Figure 9.5 in particular serve to illustrate the positive effect that location-based voting has had in H1K. Without geographical information, RPFs are only able to classify Marshy Grassland (B.5) in one set of experiments, when using texture features. However, considering that there are a large number of photographs from the same area, Titchfield Haven, depicting Marshy Grassland in our database, using geographical location during testing makes possible the correct classification of Marshy Grassland in all cases.

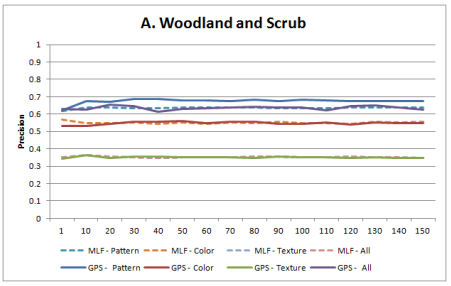
9.5 Concluding Remarks

In this chapter we have presented the last element of our framework and our last contribution: a location-based voting system for Random Projections Forests. We have explained the motivation behind our decision to include geographical information in the classification process and we have described how it can be implemented to be used during the testing phase. Moreover, we have carried out a series of experiments designed to measure the impact of this last modification in our system in comparison to the use of medium-level features and RPFs. Results show from all possible scenarios testing in the previous chapters, the inclusion of location-based voting mechanism to our RPFs with medium-level features has produced the most efficient and accurate results in this thesis and, to our knowledge, ever developed.

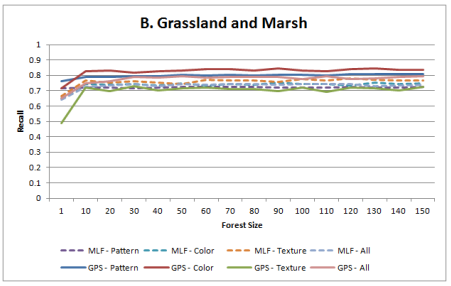
In the next chapter, we will summarise the contents of this thesis, we will reiterate our contributions and, more importantly, we will discuss some of the limitations from our current approach and offer some suggestions for further development.



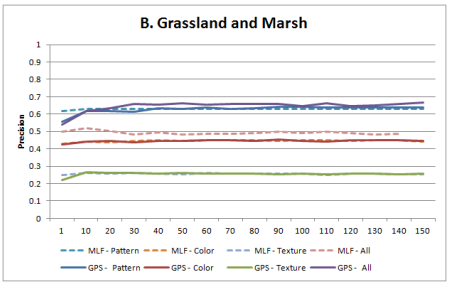
(a) A. Recall



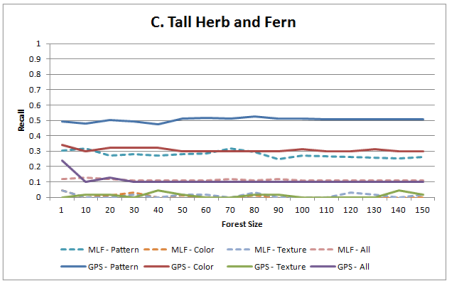
(b) A. Precision



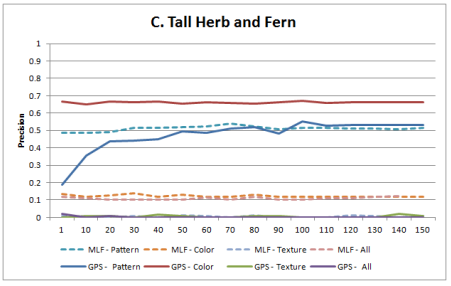
(c) B. Recall



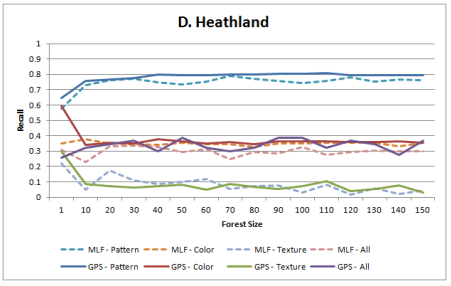
(d) B. Precision



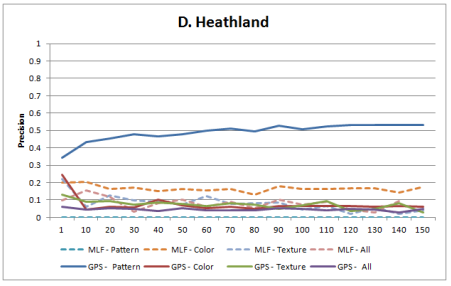
(e) C. Recall



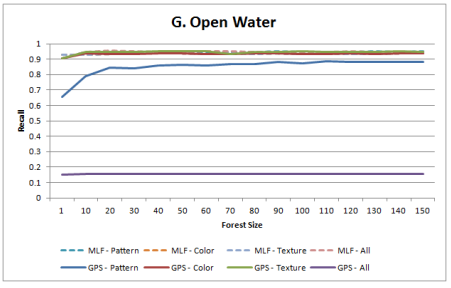
(f) C. Precision



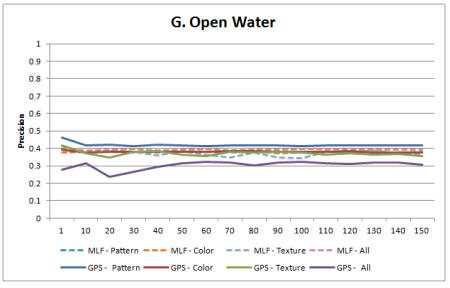
(g) D. Recall



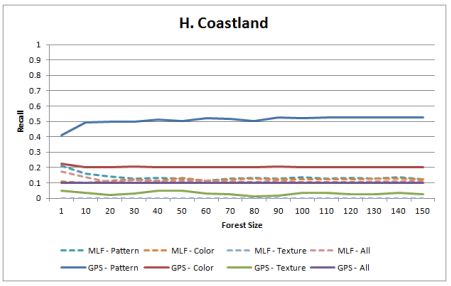
(h) D. Precision



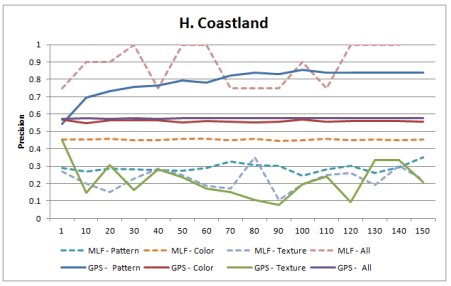
(i) G. Recall



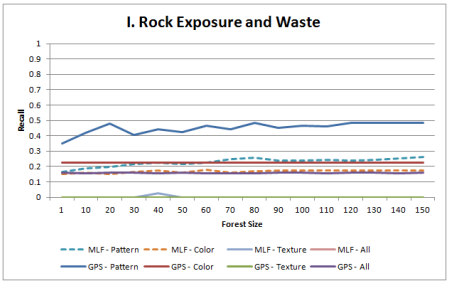
(j) G. Precision



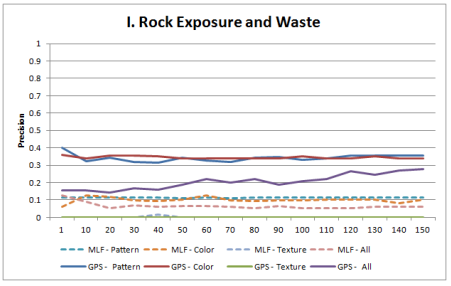
(k) H. Recall



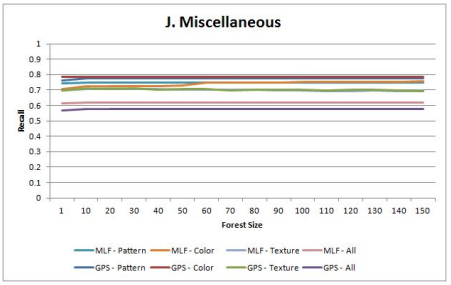
(l) H. Precision



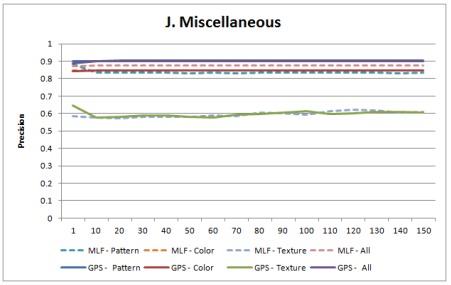
(m) I. Recall



(n) I. Precision

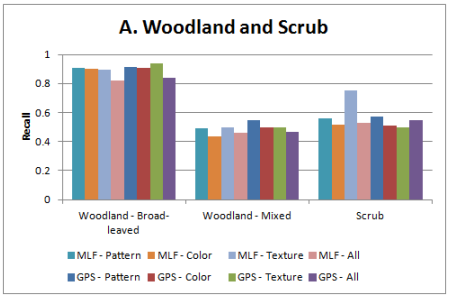


(o) J. Recall

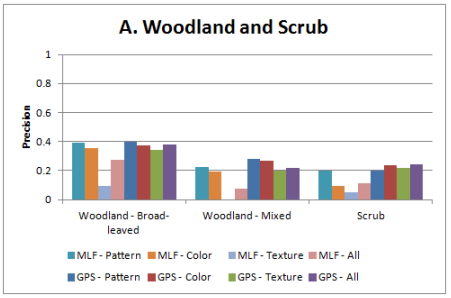


(p) J. Precision

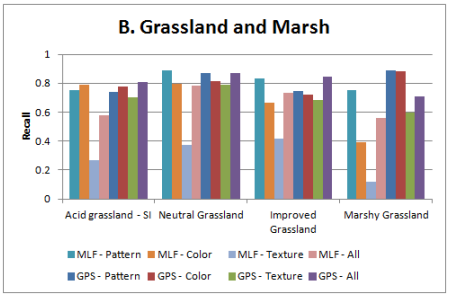
FIGURE 9.2: (Cont.) Location-based Voting System. Recall and precision results for first-tier habitats from Habitat 3K



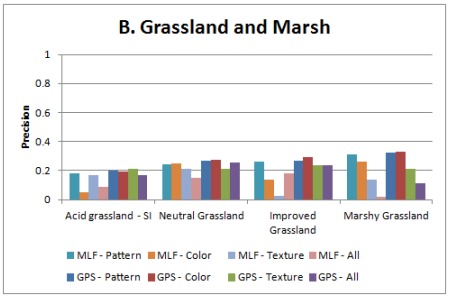
(a) A. Recall



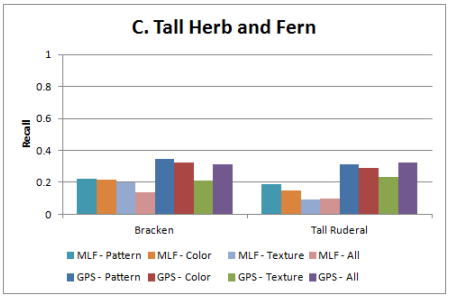
(b) A. Precision



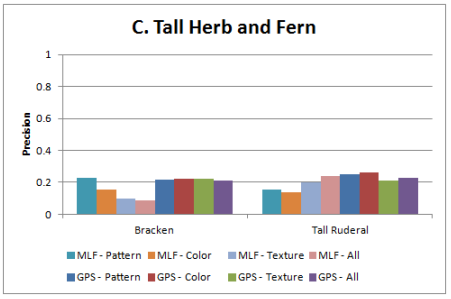
(c) B. Recall



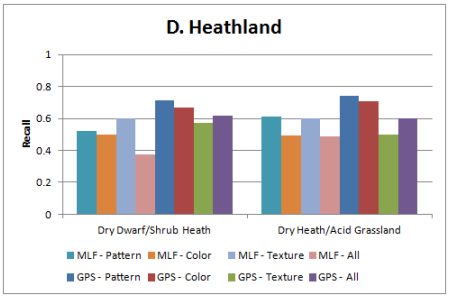
(d) B. Precision



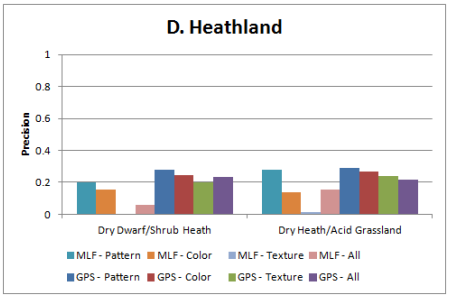
(e) C. Recall



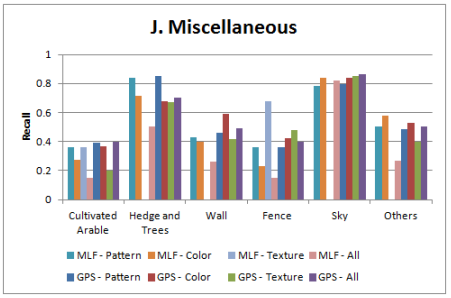
(f) C. Precision



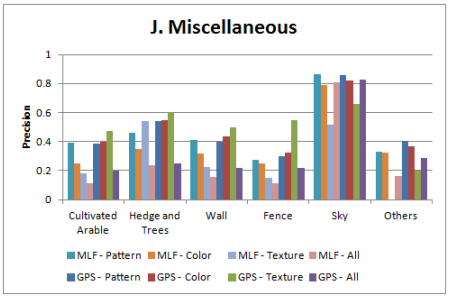
(g) D. Recall



(h) D. Precision

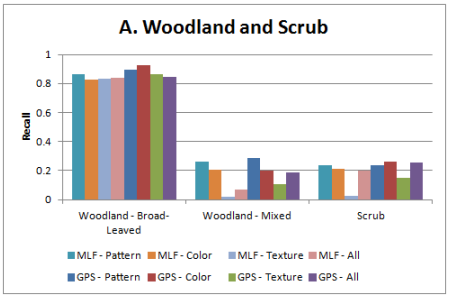


(i) J. Recall

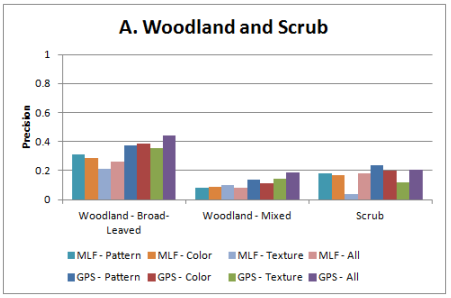


(j) J. Precision

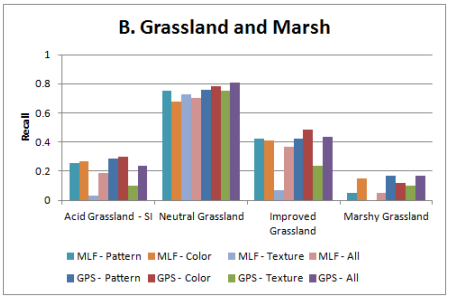
FIGURE 9.3: Location-based Voting System. Recall and precision results for second- and third-tier habitats from Habitat 1K



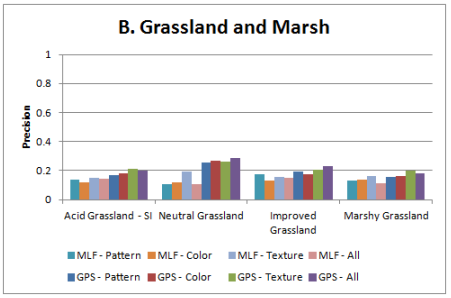
(a) A. Recall



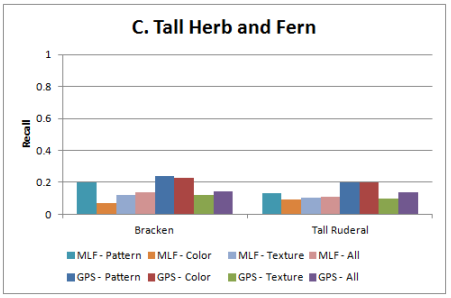
(b) A. Precision



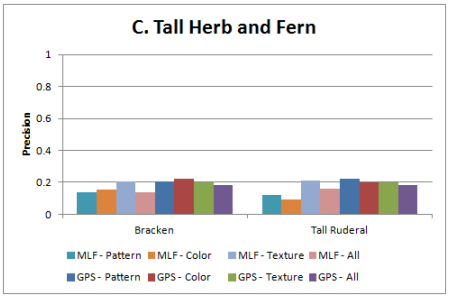
(c) B. Recall



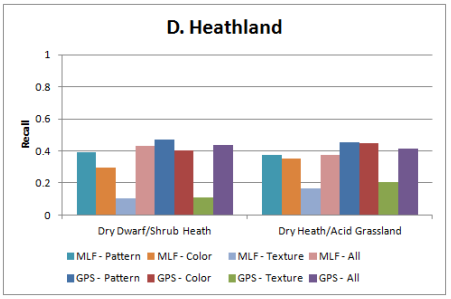
(d) B. Precision



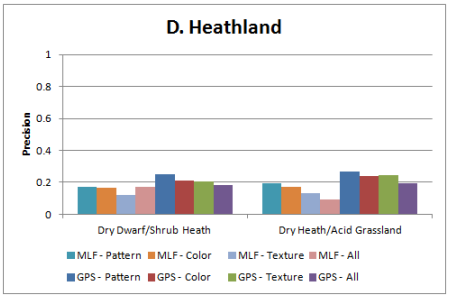
(e) C. Recall



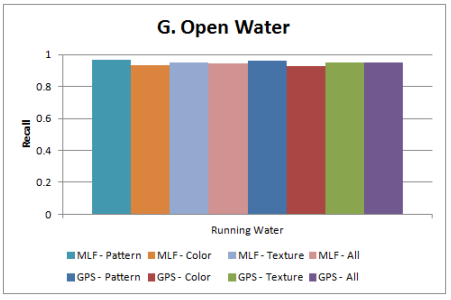
(f) C. Precision



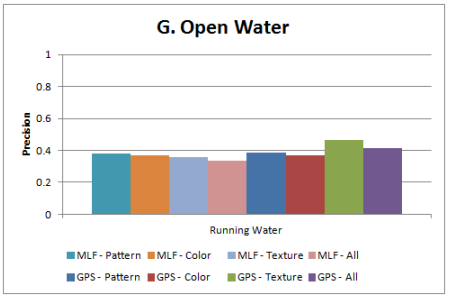
(g) D. Recall



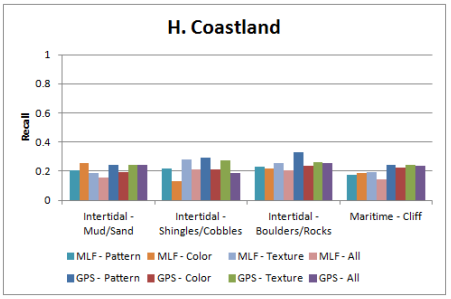
(h) D. Precision



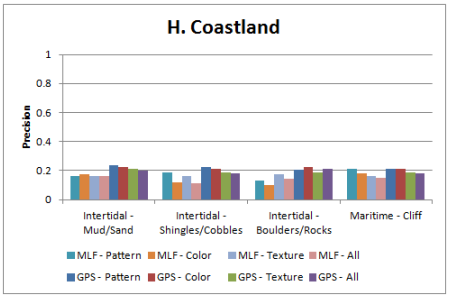
(i) G. Recall



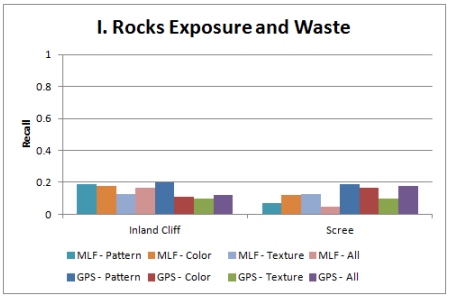
(j) G. Precision



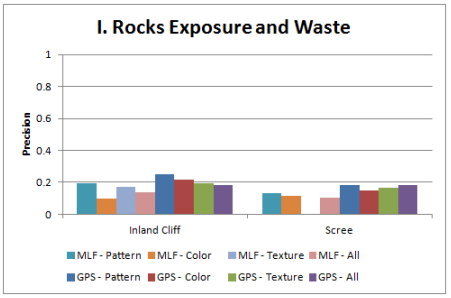
(k) H. Recall



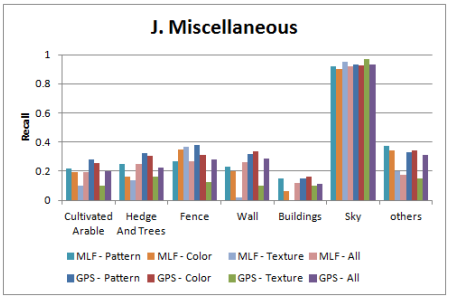
(l) H. Precision



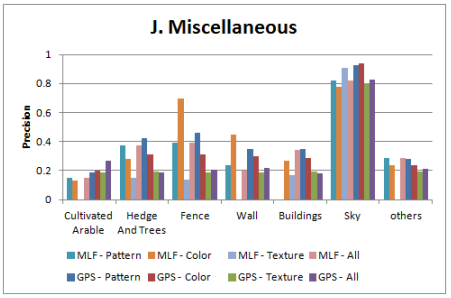
(m) I. Recall



(n) I. Precision



(o) J. Recall



(p) J. Precision

FIGURE 9.4: (Cont.) Location-based Voting System. Recall and precision results for second-tier habitats from Habitat 3K

FIGURE 9.5: Visual Example From H1K. Habitats present are: Improved Grassland, Woodland - Broad-leaved and Fence.

TABLE 9.2: Results. We show the five most probable results obtained with our experiments.

Features Extracted		MLF and RPF	MLF and RPF and GPS
	Pattern	***Scrub***	***Marshy Grassland***
		Dry Heath/Acid Grassland	***Sky***
		Dry Dwarf/Shrub Heath	Dry Heath/Acid Grassland
		Woodland - Broad-leaved	***Scrub***
		Sky	***Dry Dwarf/Shrub Heath***
	Color	Woodland - Broad-leaved	***Sky***
		Scrub	***Marshy Grassland***
		Improved Grassland	***Dry Dwarf/Shrub Heath***
		Sky	Dry Heath/Acid Grassland
		Neutral Grassland	***Scrub***
	Texture	***Sky***	***Marshy Grassland***
		Dry Dwarf/Shrub Heath	***Sky***
		Dry Heath/Acid Grassland	Woodland - Broad-leaved
		Tall Ruderal	***Scrub***
		Marshy Grassland	Neutral Grassland
	All	***Sky***	***Marshy Grassland***
		Woodland - Broad-leaved	***Dry Dwarf/Shrub Heath***
		Scrub	***Sky***
		Tall Ruderal	***Scrub***
		Dry Dwarf/Shrub Heath	Bracken

FIGURE 9.6: Visual Example From H3K. Habitats present are: Running Water, Marshy Grassland, Scrub, Dry Dwarf/Shrub Heath.

TABLE 9.3: Results. We show the five most probable results obtained with our experiments.

Features Extracted		MLF and RPF	MLF and RPF and GPS
	Pattern	***Sky***	Woodland - Broad-leaved
		Woodland - Broad-leaved	***Woodland - Mixed***
		Neutral Grassland	***Scrub***
		Improved Grassland	***Acid Grassland - SI***
		Scrub	***Sky***
	Color	***Sky***	***Sky***
		Woodland - Broad-leaved	***Scrub***
		Acid Grassland - SI	Tall Ruderal
		Neutral Grassland	Woodland - Broad-leaved
		Scrub	***Woodland - Mixed***
	Texture	Woodland - Broad-leaved	***Scrub***
		Scrub	***Acid Grassland - SI***
		Sky	Woodland - Broad-leaved
		Woodland - Mixed	***Sky***
		Acid Grassland - SI	***Woodland - Mixed***
	All	Woodland - Broad-leaved	***Scrub***
		Tall Ruderal	***Woodland - Mixed***
		Woodland - Mixed	Neutral Grassland
		Sky	***Sky***
		Scrub	Woodland - Broad-leaved

Chapter 10

Concluding Remarks

In this thesis, we have studied the problem of automatic Phase 1 Habitat classification using ground-taken photographs. For this purpose, we have developed an automatic image annotation framework. This framework combines ground-taken photographs, low and medium-level feature extraction and Random Projection Forests with a location-based voting system to enable us to annotate unseen photographs with the habitats present in them.

This final chapter is organised as follows: Section 10.1 summarises the contributions of this thesis, while Section 10.2 explores some of the limitations of our framework and suggests future work that can be carried out to improve its performance. Finally, Section 10.3 presents a complete summary of the work presented in this thesis.

10.1 Contributions

In this thesis, we have proposed an automatic image annotation framework for the classification of Phase 1 habitats. We make the following contributions:

- Image-Annotation Framework [Chapter 5]: We have approached automatic habitat classification as an automatic image annotation (AIA) problem. We have developed an automatic image-annotation framework for Phase 1 habitat classification. Our framework, shown in Figure 10.1, combines five main elements to annotate unseen photographs using the Phase 1 classification scheme. These elements are: ground-taken photography, low-level visual features, medium-level semantic information, random projections forests and location-based weighted predictions. Extensive experimentation shows that our framework can successfully classify Phase

1 habitats in terms of precision and recall, making it the first and most accurate automatic system specifically designed for the classification of the complete Phase 1 scheme.

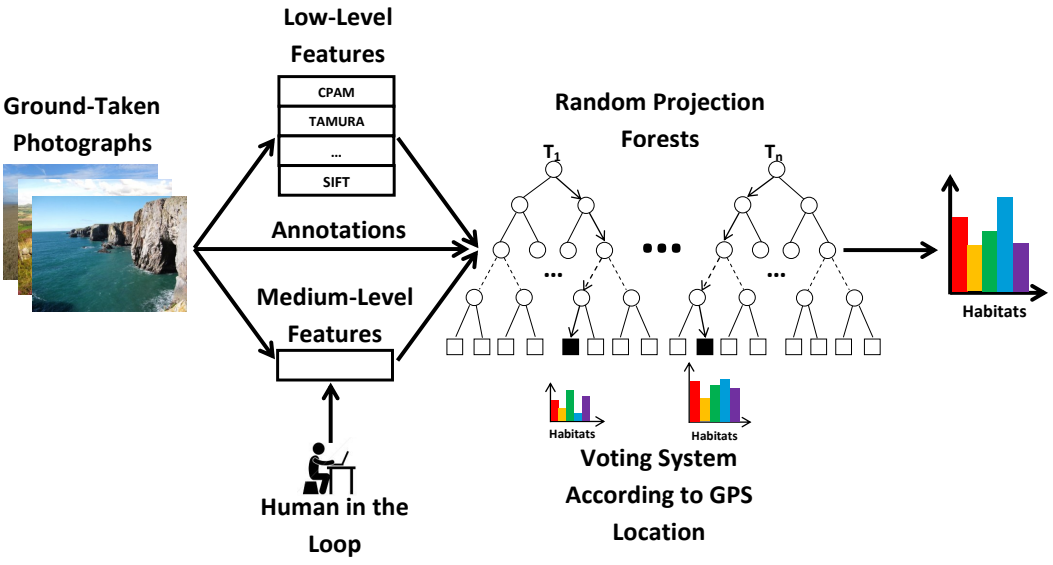


FIGURE 10.1: Image Annotation-Based Habitat Classification. Our framework consists of: ground-taken photographs, low- and medium feature extraction, random projection forests and a location-based voting system.

- Habitat 1K and Habitat 3K [Chapter 6]: We have compiled, organised and annotated two databases specially created for ecological purposes. Habitat 1K is composed of 1,086 photographs and 4,223 annotations from five habitat classes: Woodland and Scrub (A), Grassland and Marsh (B), Tall Herb and Fern (C), Heathland (D) and Miscellaneous (J). Photographs were taken under controlled environmental conditions by the author of this thesis. Habitat 3K has 3,094 ground-taken geo-referenced photographs. This database was collected using a crowd-sourcing mechanism and it has been ground-truthed by a Phase 1 expert and the author of this thesis. As a direct consequence of this, the environmental conditions of Habitat 3K are widely variable. It includes 11,517 different instances of habitats from seven out of the ten possible habitat classes. These are: Woodland and Scrub (A), Grassland and Marsh (B), Tall Herb and Fern (C), Heathland (D), Open Water (G), Coastland (H), Rock Exposure and Waste (I) and Miscellaneous (J). The photos of both these databases do not follow any particular layout, with

all types of shots, i.e. ground shots, detail shots or landscape shots, being allowed. Moreover, they have been made publicly available and they are the first image databases specifically designed for the development of multimedia analysis techniques for habitat classification.

- Low-level Visual Features Applied to Habitat Classification [Chapter 7, Chapter 8, Chapter 9]: We carry out an study on the effects of a number of the most popular low-level visual features. Particularly, we study the effect that texture (Tamura coefficients and Gray-Level Co-occurrence Matrices), pattern (Colour Pattern Appearance Model) and colour (Colour Histograms and Color Moments) features have on Phase 1 habitat classification when using ground-taken imagery. This helps us better understand the benefits and limitations that ground-taken imagery present when classifying Phase 1 habitats. Results show that pattern and colour features obtain the most stable precision and recall results in more than 80% of the testing scenarios. On the other hand, texture features can obtain more accurate results than pattern and colour in particular cases, such as the classification of heath mosaics with Random Projection Forests, but their general performance in all experiments is considerably less stable.

- Random Projection Forests (RPF)[Chapter 7]: Random Forests is an increasingly popular machine learning technique. We chose to use this ensemble classifier because they combine the benefits of two other popular Machine Learning techniques, NN-based methods and SVMs, without being affected by their disadvantages. Like NN-based methods and contrary to SVMs, Random forests are simple to implement and easy to modify to be applied to multi-label problems. On the other hand, similarly to SVMs and contrary to NN-based methods, they are accurate and do not suffer from a less efficient testing phase. Additionally, random forests have been successfully applied to a varied number of problems in the field of computer vision, such image classification [132] and image segmentation [167]. In the field of ecology, they have also been applied to habitat structure classification [11] and land cover [81]. We propose a novel design of Random Forests that uses Random Projections. With RPF, we generate a random projection vector with values {-1, 0, 1} in each of the nodes of our decision tree and we project each feature vector according to the corresponding random projection vector. The inclusion of projections makes the training and testing process more efficient without sacrificing accuracy in the results. Results show that our initial design of Random Projection Forests is not only more efficient, but also outperforms Random Forests both in terms of recall and precision. This difference in performance is clearly noticeable when classifying Woodland and Scrub (A), Grassland and Marsh (B) and Heathland (D) habitats.

- Medium-Level Features [Chapter 8]: Habitat classification is a Fine-Grained Visual Categorization (FGVC) problem in which classes, particularly second- and third-tier classes, share many visual similarities. Consequently, the use low-level visual features entails a series of limitations in the classification process. In order to combat these limitations, we propose the inclusion of semantic information, which can be crucial to distinguish between habitats, during the training phase. We adopt a Human-In-The-Loop (HITL) approach, shown in Figure 10.2, to obtain medium-level semantic information [24] and we include that information in the classification process in the form of features. HITL is an interactive, hybrid human-computer method for object classification which aims to benefit from the strengths of both humans (their ability to distinguish between objects by incorporating semantic and contextual information) and computers (their ability of computing large amounts of data efficiently). In our approach, non-experts users are asked a series of 'yes'-or-'no' questions about the ground-taken photographs in our database and they are also required to grade the degree of certainty they have in their answer. Additionally, we combine these medium-level features with low-level visual features to obtain more accurate results in the most challenging habitat classes: Tall Herb and Fern (C) and Heathland (D). Experiments show that the inclusion of medium-level features entails a considerable improvement over our initial design of Random Projection Forests, particularly in terms of precision, which improves up to 20%. This increase is particularly noticeable in Tall Herb and Fern habitats (C) and complex habitats such as Hedge and Trees (J.2.3) and Heathland mosaics.

- Location-Based Voting [Chapter 9]: In order to exploit the geographical properties of the habitats we are classifying, we include geographical information during the annotation process. We take advantage of the geographical properties of habitats considering the following: geographically close areas have similar ecological characteristics, since habitat properties do not generally change abruptly. Therefore, near regions will have similar habitats. Since all the images in the database are geo-referenced, we use their GPS coordinates to calculate the distance between unseen photographs and the ground-taken photographs of the leaves they have reached in the RPF. Consequently, we weight the different decision trees in our RPF, with closer trees having more weight in the prediction than further trees. Experiments show that this final modification of Random Projections Forests yields the most accurate recall and precision results from all the scenarios tested in this thesis. In particular, complex mosaics and Coastland (H) habitats, which have proven specially difficult to classify, experience a considerable recall and precision improvement over past modifications. Consequently, this final contribution, to our

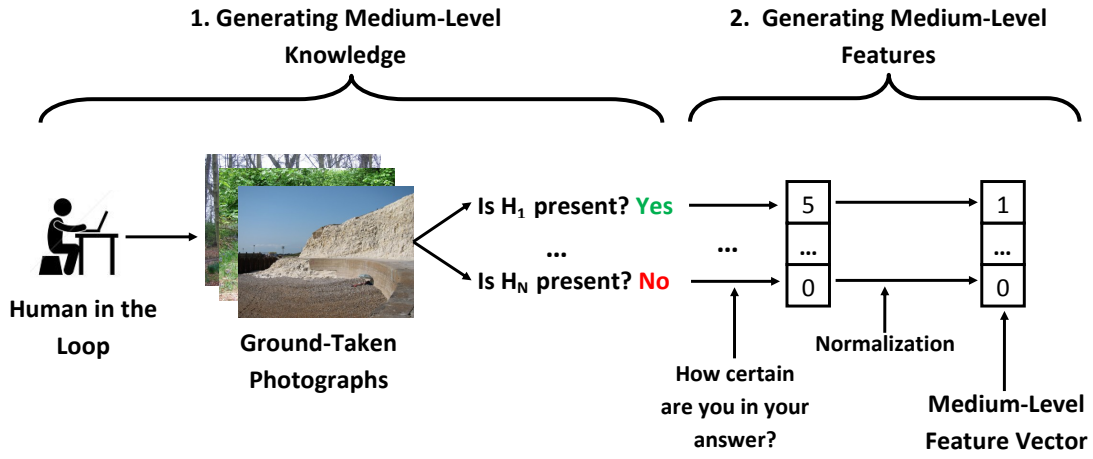


FIGURE 10.2: Medium-Level Information and Features. In our case, N is equal to 36 and certainty in measured between 0 (not sure at all) and 5 (completely sure).

knowledge, makes our Random Projection Forests with medium-level features and a location-based voting system the first and most accurate automatic framework specifically designed for the classification of the complete Phase 1 scheme.

10.2 Limitations and Suggestions For Improvement

As we mentioned in the previous section, the image-annotation framework presented in this thesis was designed as an alternative to current Phase 1 classification, which is carried out manually. Nevertheless, the current automatic design has some limitations with regards to its performance, particularly in the case of second- and third-tier precision results.

In this section, we discuss these limitations and offer possible improvements that could be developed as further work. These limitations can be linked to four main aspects of our framework: the input data, the features extracted, the classifier and the use of location information. These are:

- Ground-taken Photographs: We have proven that ground-taken photographs are a valid source of information for the automatic classification of Phase 1 habitats.

However, the two current databases created and collected as part of this thesis could be further improved to obtain more accurate results. Both Habitat 1K and Habitat 3K contain an imbalance between two of their classes, Woodland and Scrub (A) and Grassland and Marsh (B) and the rest of the classes present. Moreover, within both categories, Broad-leaved Woodland (A.1) and Neutral Grassland (B.2) amount the largest number of instances, with over 500 instances of difference with habitats such as Bracken (C.1) or Fence (J.2.4). Considering that both A.1 and B.2 habitats are amongst the most accurately classified in our framework and that those habitats with the lowest number of instances, such as Tall Herb and Fern (C), obtain the least accurate results, we project that increasing the number of instances of the other habitats, particularly those which have been proven to be more difficult to classify, such as Tall Herb and Fern (C), Coastland (H) and Rock Exposures and Waste (I) would only benefit current performance results.

Moreover, the inclusion of more ground-taken photographs is not the only aspect regarding our source data that could be improved. As discussed in Chapter 6 and demonstrated in Chapter 9, ground-taken photographs, while easier to obtain and more detailed than remote-sensed data, present a clear limitation in terms of geographical information. That is, the position of a photograph might not accurately reflect the position of the habitats present within the photograph. This makes the use of geographical information a complicated endeavour which can result in inaccurate classification results.

We propose the inclusion of remote-sensed data in the classification process, not as a source of information per se, but as a tool to correctly obtain the location of the habitats present within the photographs. Research has been developed on how to accurately project different elements within geo-referenced photographs onto maps, as shown in [159], and we consider that the further development and application of these methods could greatly benefit the performance of our current system.

- Semantic Information: As mentioned in Chapter 9, semantic information is crucial when trying to classify FGVC problems such as habitat classification. In this type of classification problems, in which the classes are extremely visually similar, there is a significant need for extracting other kinds of descriptive and discriminative information to aid the classification process. In this thesis, we presented a new type of semantic features, medium-level features, which were extracted using a HITL approach. However, as we discussed in Chapter 9, uncertain answers from the users employed to obtain this information could affect negatively the performance of our system. This problem was exacerbated in our current system because we only used one user to obtain one feature vector. Nevertheless, as further work, we

propose employing more users, at least five, to extract different opinions on the answers to the questions that help us create our medium-level features. This way, the answers would be combined and the uncertainty of one user would not affect as directly the classification process.

- Random Projection Forests: In Chapter 7 we explained that, in the current framework design, we consider all habitats independent from each other. This assumption was done consciously, since we had not carried out experiments to investigate the relationships between habitats. However, as experiments helped to identify, this is not always the case. In general, there are several types of habitat configurations that are more likely to appear together depending on the geographical-location of the photographs. For example, Neutral Grassland (B.2) habitats in New Forest are more likely to appear with Hedges and Trees (J.2.3) than with Running Water (G.2) or Scree (I.1.2). This information could greatly aid the second- and third-tier classification of Phase 1 habitats. Moreover, it is information that is already present our datasets, in the form of the frequency of appearance of particular annotations with other specific annotations. We would only have to include this information during training, using the geographical location of the photographs, to benefit from knowledge that is already in our database. Therefore, we propose the exploitation of habitat relationships as further work for our classifier.

- Location-based information: As explained in Chapter 9, our current system only takes into consideration geographical location during testing. However, as we introduced in the previous point, there are other aspects to geographical location, and the consequent information that they could provide, that could aid the classification of visually similar Phase 1 classes. For example, Woodland in the area of Titchfield Haven is more likely to be Broad-Leaved than Coniferous. Consequently, another promising improvement to the framework would be to further exploit the geographical location of the photographs and their relationship during training to accurately classify second and third- tier habitats.

10.3 Summary

In summary, we have have created an automatic image-annotation framework for the classification of Phase 1 habitats. Contrary to the habitat classification schemes reviewed in Chapter 2, our framework is, to our knowledge the first system created to date which classifies all possible Phase 1 habitats. In Chapter 2 and Chapter 3 we explained our motivation for having chosen Phase 1 as our classification scheme and we discussed its main merits and limitations.

Our complete framework was presented in Chapter 5. Following chapters expand on each element of our framework with Chapter 6 focusing on our source data, ground-taken photographs, and the two databases we have collected and annotated, called Habitat 1K and Habitat 3K; Chapter 7 introducing our novel classifier, Random Projection Forests; Chapter 8 detailing a new type of semantic features, medium-level features, and lastly, Chapter 8 introducing our location-based voting system. Each chapter gives a detailed description of each component of the framework and expands on the motivations behind their design, creation and their inclusion to our system.

Furthermore, we carried out extensive experiments with the aim of studying the performance of ground-taken photographs, low- and medium-level features, Random Projection Forests and a location-based voting system. Results to these experiments, shown in Chapter 7, Chapter 8 and Chapter 9, served to demonstrate the validity of RPFs as classifiers, particularly for the case of Phase 1 classification. We compared the performance of traditional Random Forest and each of the modifications introduced in our design of RPFs and found that RPFs with pattern features, semantic information and a location-based voting system produced the most stable and accurate results.

However, our current design has some limitations with regards to its performance, particularly in terms second- and third-tier habitat classification. With the aim of improving this part of the classification process, we propose as further work the expansion of our current databases to include a more balanced number of habitats present and the inclusion of more robust semantic features. These features would use several the answers from users to estimate the presence of the semantic tags within the photographs. Additionally, we propose the use of relationships between habitats during training and the inclusion of geo-referenced multi-source data, such as satellite photographs, to help with the perspective limitations of ground-taken photographs.

In essence, we regard our current image-annotation framework as a first step towards a completely automatic Phase 1 habitat classification process. We consider that there is still a lot of research that could be done and we envision that the inclusion of the suggested further work will only help to improve the performance of the presented system.

Bibliography

[1] T. Acharya and A.K. Ray. *Image processing: principles and applications*. John Wiley & Sons, 2005.

[2] D. Achlioptas. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *Journal of computer and System Sciences*, 66(4):671–687, 2003.

[3] Environment Agency. Environment agency, 2014. URL https://www.gov.uk/government/organisations/environment-agency.

[4] R. Alexander, A.C. Millington, et al. *Vegetation mapping: From patch to planet.* John Wiley & Sons, 2000.

[5] R.J. Allee, J. Kurtz, R.W. Gould Jr., D.S. Ko, M. Finkbeiner, and K. Goodin. Application of the coastal and marine ecological classification standard using satellite-derived and modeled data products forpelagic habitats in the northern gulf of mexico. *Ocean & Coastal Management*, 88(0):13–20, 2014.

[6] Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural computation*, 9(7):1545–1588, 1997.

[7] J.R. Anderson. *A land use and land cover classification system for use with remote sensor data*, volume 964. US Government Printing Office, 1976.

[8] S. Angus. Oxford dictionary of english. URL http://www.oxfordreference.com.

[9] K.J. Archer and R.V. Kimes. Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, 52(4):2249–2260, 2008.

[10] M. Bambha, T. Anderson, S. Pearman, and V. Roussilhon. Automatic image tagging, March 31 2010. US Patent App. 12/752,099.

[11] D. Bargiel. Capabilities of high resolution satellite radar for the detection of semi-natural habitat structures and grasslands in agricultural landscapes. *Ecological Informatics*, 13:9–16, 2013.

[12] R. Behmo, N. Paragios, and V. Prinet. Graph commute times for image representation. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[13] R. Behmo, P. Marcombes, A. Dalalyan, and V. Prinet. Towards optimal naive bayes nearest neighbor. In *Computer Vision–ECCV 2010*, pages 171–184. Springer, 2010.

[14] R. E. Bellman. *Adaptive control processes - A guided tour*. Princeton University Press, Princeton, New Jersey, U.S.A., 1961.

[15] T. Berg, J. Liu, S.W. Lee, M.L. Alexander, D.W. Jacobs, and P.N. Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Proc. Conf. Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[16] S. Bernard, S. Adam, and L. Heutte. Dynamic random forests. *Pattern Recognition Letters*, 33(12):1580–1586, 2012.

[17] E. Bingham and H. Mannila. Random projection in dimensionality reduction: Applications to image and text data. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 245–250, 2001.

[18] C.M. Bishop et al. *Pattern recognition and machine learning*, volume 1. Springer New York, 2006.

[19] X. Bo, L.and Ren and D. Fox. Kernel descriptors for visual recognition. *NIPS*, 1 (2):3, 2010.

[20] M. Bock. Remote sensing and gis-based techniques for the classification and monitoring of biotopes: Case examples for a wet grass-and moor land area in northern germany. *Journal for Nature Conservation*, 11(3):145–155, 2003.

[21] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[22] A. Bosch, A. Zisserman, and X. Munoz. Image classification using random forests and ferns. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, Oct 2007.

[23] D. Boyd, C. Sanchez-Hernandez, and G. Foody. Mapping a specific class for priority habitats monitoring from satellite sensor data. *International Journal of Remote Sensing*, 27(13):2631–2644, 2006.

[24] S. Branson, C. Wah, F. Schroff, B. Babenko, P. Welinder, P. Perona, and S. Belongie. *Visual recognition with humans in the loop*, volume 6314 LNCS (PART 4) of *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2010.

[25] S. Branson, P. Perona, and S. Belongie. Strong supervision from weak annotation: Interactive training of deformable part models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1832–1839, 2011.

[26] S. Branson, G. Van Horn, C. Wah, P. Perona, and S. Belongie. The ignorant led by the blind: A hybrid human-machine vision system for fine-grained categorization. *International Journal of Computer Vision*, pages 1–27, 2014.

[27] A. Bratt. Phase 1 Habitat Survey Report - Honiton Community Centre. Technical report, Acorn Ecology Limited, 2011.

[28] L. Breiman. Random forests. In *Machine Learning*, pages 5–32, 2001.

[29] W.F.J. Bronaugh. human-in-the-loopsimulation: the right tool for port design. *Port Technology International*, 32:1–2, 2007.

[30] R.A. Brualdi. *Contemporary mathematics - American Mathematical Society*. Conference in Modern Analysis and Probability. American Mathematical Society, 1984.

[31] O. Buck, B. Peter, A. Völker, and A. Donning. Object based image analysis to support environmental monitoring under the european habitat directive: a case study from decover. In *ISPRS Hannover Workshop*, 2011.

[32] C. Bundy. Swarkestone Quarry, Barrow Upon Trent, Derbyshire - Extended Phase 1 Habitat Survey. Technical Report RT-MME-105181, Middlemarch Environmental Ltd, 2009.

[33] R. Caruana, N. Karampatziakis, and A. Yessenalina. An empirical evaluation of supervised learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pages 96–103. ACM, 2008.

[34] L. Cayuela, J.M.R. Benayas, and C. Echeverría. Clearance and fragmentation of tropical montane forests in the highlands of chiapas, mexico (1975–2000). *Forest Ecology and Management*, 226(1):208–218, 2006.

[35] C. Chailloux, A-G Allais, P. Simeoni, and K. Olu. Automatic classification of deep benthic habitats: Detection of microbial mats and siboglinid polychaete fields from optical images on the mosby mud volcano. In *OCEANS 2008*, pages 1–7, Sept 2008.

[36] S.F. Chang, W. Hsu, L. Kennedy, L. Xie, A. Yanagawa, E. Zavesky, and D.Q. Zhang. Columbia university trecvid-2005 video search and high-level feature extraction. In *NIST TRECVID workshop, Gaithersburg, MD*, 2005.

[37] L. C. Chen and J. Y. Rau. Detection of shoreline changes for tideland areas using multi-temporal satellite images. *International Journal of Remote Sensing*, 19(17): 3383–3397, 1998.

[38] G. Chust, I. Galparsoro, A. Borja, J. Franco, and A. Uriarte. Coastal and estuarine habitat mapping, using {LIDAR} height and intensity and multi-spectral imagery. *Estuarine, Coastal and Shelf Science*, 78(4):633 – 643, 2008. ISSN 0272-7714.

[39] European Commission. Natura 2000 network, 2000. URL http://ec.europa.eu/environment/nature/natura2000/index_en.htm.

[40] The Joint Committee. Joint nature conservation committee, 1994. URL https://www.gov.uk/government/organisations/environment-agency.

[41] G. Cornelis Van Kooten, B. Stennes, E. Krcmar-Nozic, and R. Van Gorkom. Economics of afforestation for carbon sequestration in western canada. *Forestry Chronicle*, 76(1):165–172, 2000.

[42] Charnwood Borough Council. Extended Phase 1 Vegetation and Habitat Survey for Lane East and West of Snells Nook Lane, Loughborough. Technical report, Charnwood Borough Council, 2011.

[43] Council of the European Communities. Council Directive 1992/43/EEC, 1992.

[44] T. Cover and P. Hart. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27, 1967.

[45] L.M. Cowardin, Biological Services Program (U.S.), U.S. Fish, and Wildlife Service. *Classification of wetlands and deepwater habitats of the United States*. Washington, D.C. :Fish and Wildlife Service, U.S. Dept. of the Interior,, 1979. URL http://www.biodiversitylibrary.org/item/22732. http://www.biodiversitylibrary.org/bibliography/4108.

[46] A. Criminisi, J. Shotton, and E. Konukoglu. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends in Computer Graphics and Vision*, 7(2-3):81–227, 2011.

[47] A. Criminisi, J. Shotton, D. Robertson, and E. Konukoglu. Regression forests for efficient anatomy detection and localization in ct studies. In *Medical Computer*

Vision. Recognition Techniques and Applications in Medical Imaging, pages 106–117. Springer, 2011.

[48] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[49] S. Dasgupta. Experiments with random projection. In *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*, pages 143–151. Morgan Kaufmann Publishers Inc., 2000.

[50] Corel Database. Uci, 1791. URL https://archive.ics.uci.edu/ml/datasets/Corel+Image+Features.

[51] C.E. Davies, D. Moss, and M.O Hill. EUNIS Habitat Classification Revised 2004. Report to the European Topic Centre on Nature Protection and Biodiversity, 2004.

[52] I.M. de Diego, A. Muñoz, and J.M. Moguerza. Methods for the combination of kernel matrices within a support vector framework. *Machine Learning*, 78(1-2): 137–174, 2010.

[53] A. Del Bimbo. Visual information retrieval. 1999.

[54] R. A. Díaz Varela, P. Ramil Rego, S. Calvo Iglesias, and C. Muñoz Sobrino. Automatic habitat classification methods based on satellite images: A practical assessment in the nw iberia coastal mountains. *Environmental monitoring and assessment*, 144(1-3):229–250, 2008.

[55] S. Dickinson. Wild frontier ecology - what is phase one?, 1998. URL http://www.wildfrontier-ecology.co.uk/.

[56] S. Ding, H. Zhu, W. Jia, and C. Su. A survey on feature extraction for pattern recognition. *Artificial Intelligence Review*, 37(3):169–180, 2012. ISSN 0269-2821.

[57] Q. Du, J.E. Fowler, and B. Ma. Random-projection-based dimensionality reduction and decision fusion for hyperspectral target detection. In *Geoscience and Remote Sensing Symposium (IGARSS), 2011 IEEE International*, pages 1790–1793. IEEE, 2011.

[58] K. Duan, D. Parikh, D. Crandall, and K. Grauman. Discovering localized attributes for fine-grained recognition. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3474–3481. IEEE, 2012.

[59] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification (2Nd Edition)*. Wiley-Interscience, 2000.

[60] Charles Elkan. Results of the kdd'99 classifier learning. *ACM SIGKDD Explorations Newsletter*, 1(2):63–64, 2000.

[61] Natural England. Natural England, 2006. URL http://www.naturalengland.org.uk/.

[62] Natural England. Magic: Multi-agency geographic information for the countryside, 2013. URL http://www.magic.gov.uk/NaturalEngland.

[63] C. Englund and A. Verikas. A novel approach to estimate proximity in a random forest: An exploratory study. *Expert Systems with Applications*, 39(17):13046–13050, 2012.

[64] ESRI. Arcgis, 2012. URL https://www.arcgis.com.

[65] M. Everingham, S.M.A. Eslami, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge–a retrospective.

[66] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 524–531. IEEE, 2005.

[67] G.M Foody. Approaches for the production and evaluation of fuzzy land cover classifications from remotely-sensed data. *International Journal of Remote Sensing*, 17(7):1317–1340, 1996.

[68] W. Förstner. A framework for low level feature extraction. In *Computer Vision, ECCV'94*, pages 383–394. Springer, 1994.

[69] J.A. Fossit. A guide to habitats in ireland. *Kilkenny: Heritage Council*, 2000.

[70] J.E. Fowler, Q. Du, W. Zhu, and N.H. Younan. Classification performance of random-projection-based dimensionality reduction of hyperspectral imagery. In *Geoscience and Remote Sensing Symposium, 2009 IEEE International, IGARSS 2009*, volume 5, pages V–76. IEEE, 2009.

[71] D. Fradkin and D. Madigan. Experiments with random projections for machine learning. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 517–522. ACM, 2003.

[72] Y. Freund and R.E. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory*, pages 23–37. Springer, 1995.

[73] Y. Freund, R. Schapire, and N. Abe. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999.

[74] M.A. Friedl and C.E. Brodley. Decision tree classification of land cover from remotely sensed data. *Remote sensing of environment*, 61(3):399–409, 1997.

[75] H. Fu and G. Qiu. Integrating low-level and semantic features for object consistent segmentation. *Neurocomputing*, 119:74–81, 2013.

[76] H. Fu, Q. Zhang, and G. Qiu. Random forest for image annotation. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part VI*, ECCV'12, pages 86–99, Berlin, Heidelberg, 2012. Springer-Verlag. ISBN 978-3-642-33782-6.

[77] J. Gall and V. Lempitsky. Class-specific hough forests for object detection. In *Decision Forests for Computer Vision and Medical Image Analysis*, pages 143–157. Springer, 2013.

[78] S. Gao, I.W. Tsang, and Y. Ma. Learning category-specific dictionary and shared dictionary for fine-grained image categorization. 2014.

[79] D. Geneletti and B.G.H. Gorte. A method for object-oriented land cover classification combining landsat tm data and aerial photographs. *International Journal of Remote Sensing*, 24(6):1273–1286, 2003.

[80] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon. Efficient regression of general-activity human poses from depth images. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 415–422. IEEE, 2011.

[81] P.O. Gislason, J.A. Benediktsson, and Johannes R. Sveinsson. Random forests for land cover classification. *Pattern Recogn. Lett.*, 27(4):294–300, 2006. ISSN 0167-8655.

[82] N. Goel, G. Bebis, and A. Nefian. Face recognition experiments with random projection. In *Defense and Security*, pages 426–437. International Society for Optics and Photonics, 2005.

[83] M. Gönen and E. Alpaydın. Multiple kernel learning algorithms. *The Journal of Machine Learning Research*, 12:2211–2268, 2011.

[84] C.C. Gotlieb and H.E. Kreyszig. Texture descriptors based on co-occurrence matrices. *Computer Vision, Graphics, and Image Processing*, 51(1):70–86, 1990.

[85] C.C. Gotlieb and H.E. Kreyszig. Texture descriptors based on co-occurrence matrices. *Computer Vision, Graphics, and Image Processing*, 51(1):70–86, 1990.

[86] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 309–316. IEEE, 2009.

[87] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, March 2003. ISSN 1532-4435.

[88] I. Guyon, S. Gunn, M. Nikravesh, and L.A. Zadeh. *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[89] A. Hapfelmeier and K. Ulm. A new variable selection approach using random forests. *Computational Statistics & Data Analysis*, 60:50–69, 2013.

[90] J.S. Hare, P. H. Lewis, P.G.B. Enser, and C.J. Sandom. Mind the gap: Another look at the problem of the semantic gap in image retrieval. In *Electronic Imaging 2006*, pages 607309–607309. International Society for Optics and Photonics, 2006.

[91] T.J. Hastie and R.J. Tibshirani. *Generalized additive models*, volume 43. CRC Press, 1990.

[92] M.M. Hayes, S.N. Miller, and M.A. Murphy. High-resolution landcover classification using random forest. *Remote Sensing Letters*, 5(2):112–121, 2014.

[93] J. Hays and A.A. Efros. Scene completion using millions of photographs. In *ACM Transactions on Graphics (TOG)*, volume 26, page 4. ACM, 2007.

[94] T.K. Ho. Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, volume 1, pages 278–282. IEEE, 1995.

[95] T.K. Ho. The random subspace method for constructing decision forests. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(8):832–844, 1998.

[96] P. Howarth and S. Rüger. Robust texture features for still-image retrieval. *IEE Proceedings-Vision, Image and Signal Processing*, 152(6):868–874, 2005.

[97] P. Howarth and S. Rüger. Fractional distance measures for content-based image retrieval. In *Advances in Information Retrieval*, pages 447–456. Springer, 2005.

[98] J. Illingworth and J. Kittler. A survey of the hough transform. *Computer vision, graphics, and image processing*, 44(1):87–116, 1988.

[99] J. Imbernon and A. Branthomme. Characterization of landscape patterns of deforestation in tropical rain forests. *International Journal of Remote Sensing*, 22 (9):1753–1765, 2001.

[100] D. Jiang, Y. Huang, D. Zhuang, Y. Zhu, X. Xu, and H. Ren. A simple semi-automatic approach for land cover classification from multispectral remote sensing imagery. *PloS one*, 7(9), 2012.

[101] X. Jin and R. Bie. Improving software quality classification with random projection. In *Cognitive Informatics, 2006. ICCI 2006. 5th IEEE International Conference on*, volume 1, pages 149–154. IEEE, 2006.

[102] Joint Nature Conservation Committee. Handbook for Phase 1 habitat survey - a technique for environmental audit, 2010.

[103] A. Joly, P. Geurts, and L. Wehenkel. Random forests with random projections of the output space for high dimensional multi-label classification. *arXiv preprint arXiv:1404.3581*, 2014.

[104] F. Kang. *Automatic Image Annotation*. PhD thesis, East Lansing, MI, USA, 2007.

[105] P. Kohli, M. Pelillo, and H. Bischof. Context-sensitive decision forests for object detection. 2012.

[106] P. Kontschieder, S. R. Bulo, H. Bischof, and M. Pelillo. Structured class-labels in random forests for semantic image labelling. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2190–2197, 2011.

[107] F. Korč and D. Schneider. Annotation tool. Technical Report TR-IGG-P-2007-01, June 2007. URL http://www.ipb.uni-bonn.de/html_pages_software/annotation-tool/publ/Korc-TR-IGG-P-2007-01.pdf.

[108] N. Kumar, P. N. Belhumeur, A. Biswas, D. W. Jacobs, W. J. Kress, I. C. Lopez, and J. V. B. Soares. Leafsnap: A computer vision system for automatic plant species identification. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part VI*, ECCV'12, pages 502–516, 2012. ISBN 978-3-642-33782-6.

[109] J. Laaksonen, E. Oja, M. Koskela, and S. Brandt. Analyzing low-level visual features using content-based image retrieval. In *Proceedings of International Conference in Neural Information Processing*, pages 14–18, 2000.

[110] G.R.G. Lanckriet, N. Cristianini, P. Bartlett, L.E. Ghaoui, and M.I. Jordan. Learning the kernel matrix with semidefinite programming. *The Journal of Machine Learning Research*, 5:27–72, 2004.

[111] C. L. Lauver and J. L. Whistler. A hierarchical classification of landsat tm imagery to identify natural grassland areas and rare species habitat. *Photogrammetric Engineering & Remote Sensing*, 59(5):627–634, 1993.

[112] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the IEEE*

Computer Society Conference on Computer Vision and Pattern Recognition, volume 2, pages 2169–2178, 2006.

[113] Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99(465):67–81, 2004.

[114] C. Li, L. Shao, C. Xu, and H. Lu. Feature selection under learning to rank model for multimedia retrieve. In *Proceedings of the Second International Conference on Internet Multimedia Computing and Service*, ICIMCS '10, pages 69–72, New York, NY, USA, 2010. ACM. doi: 10.1145/1937728.1937745.

[115] J. Li and J.Z. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(9):1075–1088, 2003.

[116] L.J. Li, H. Su, E.P. Xing, and F.F Li. Object bank: A high-level image representation for scene classification & semantic feature sparsification. *NIPS*, 2(3):5, 2010.

[117] T.M. Lillesand, R.W. Kiefer, and J.W. Chipman. *Remote sensing and image interpretation*. John Wiley & Sons, 2008.

[118] S. Ling, L. Ping, and I. Kirenko. Object category retrieval for multimedia databases. In *Consumer Electronics, 2006. ISCE '06. 2006 IEEE Tenth International Symposium on*, pages 1–3, 2006. doi: 10.1109/ISCE.2006.1689406.

[119] C. Liu, J. Yuen, A. Torralba, J. Sivic, and W.T. Freeman. Sift flow: Dense correspondence across different scenes. In *Computer Vision–ECCV 2008*, pages 28–42. Springer, 2008.

[120] J. Liu, A. Kanazawa, D. Jacobs, and P. Belhumeur. Dog breed classification using part localization. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part VI*, volume 7572 LNCS (PART 1) of *ECCV'12*, pages 172–185, 2012.

[121] X. Liu, M. Song, D. Tao, Z. Liu, L. Zhang, C. Chen, and J. Bu. Semi-supervised node splitting for random forest construction. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 492–499. IEEE, 2013.

[122] Y. Liu, D. Zhang, G. Lu, and W.Y. Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262–282, 2007.

[123] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[124] R. Lucas, K. Medcalf, A. Brown, P. Bunting, J. Breyer, D. Clewley, S. Keyworth, and P. Blackmore. Updating the phase 1 habitat map of wales, uk, using satellite sensor data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(1):81–102, 2011.

[125] Ludicorp. Flickr, 2004. URL https://www.flickr.com.

[126] A. Makadia, V. Pavlovic, and S. Kumar. Baselines for image annotation. *International Journal of Computer Vision*, 90(1):88–105, 2010.

[127] R. Maree, P. Geurts, J. Piater, and L. Wehenkel. Random subwindows for robust image classification. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 34–40. IEEE, 2005.

[128] S. Martínez, P. Ramil, and E. Chuvieco. Monitoring loss of biodiversity in cultural landscapes. new methodology based on satellite data. *Landscape and Urban Planning*, 94(2):127–140, 2010.

[129] S. McCann and D.G. Lowe. Local naive bayes nearest neighbor for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3650–3656. IEEE, 2012.

[130] C. McKay and I. Fujinaga. Automatic genre classification using large high-level musical feature sets. In *ISMIR*, volume 2004, pages 525–530, 2004.

[131] A. Montillo and H. Ling. Age regression from faces using random forests. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 2465–2468. IEEE, 2009.

[132] F. Moosmann, E. Nowak, and F. Jurie. Randomized clustering forests for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(9):1632–1646, 2008.

[133] M. Muja and D.G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *VISAPP (1)*, pages 331–340, 2009.

[134] S.Y. Neo, J. Zhao, M.Y. Kan, and T.S. Chua. Video retrieval using high level features: Exploiting query matching and confidence-based weighting. In *Image and Video Retrieval*, pages 143–152. Springer, 2006.

[135] C.W. Ngo, Y.G. Jiang, X.Y. Wei, W. Zhao, Y. Liu, J. Wang, S. Zhu, and S.F. Chang. Vireo/dvmm at trecvid 2009: High-level feature extraction, automatic video search, and content-based copy detection. *Proc. of TRECVID2009*, pages 415–432, 2009.

[136] M. E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Proceedings - 6th Indian Conference on Computer Vision, Graphics and Image Processing, ICVGIP 2008*, pages 722–729, 2008.

[137] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3): 145–175, 2001.

[138] B.P. Olsen. Automatic change detection for validation of digital map databases. In *In: International Archives of Photogrammetry and Remote Sensing, Vol. XXX IV, Part B2*, 2004.

[139] F. Orabona, L. Jie, and B. Caputo. Online-batch strongly convex multi kernel learning. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 787–794. IEEE, 2010.

[140] A. Park, J. Clare, V. Spicer, P. L. Brantingham, T. Calvert, and G. Jenion. Examining context-specific perceptions of risk: Exploring the utility of "human-in-the-loop" simulation models for criminology. *Journal of Experimental Criminology*, 8(1):29–47, 2012.

[141] L. Paulevé, H. Jégou, and L. Amsaleg. Locality sensitive hashing: A comparison of hash function types and querying mechanisms. *Pattern Recognition Letters*, 31 (11):1348–1358, 2010.

[142] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier. Large-scale image retrieval with compressed fisher vectors. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3384–3391. IEEE, 2010.

[143] J. Peters, B. De Baets, N.E.C. Verhoest, R. Samson, S. Degroeve, P. De Becker, and W. Huybrechts. Random forests as a tool for ecohydrological distribution modelling. *Ecological Modelling*, 207:304–318, 2007.

[144] J. Peters, B. De Baets, R. Samson, and N. E. C. Verhoest. Modelling groundwater-dependent vegetation patterns using ensemble learning. *Hydrology and Earth System Sciences*, 12(2):603–613, 2008.

[145] Z. Petrou, T. Stathaki, I. Manakos, M. Adamo, C. Tarantino, and P. Blonda. Land cover to habitat map conversion using remote sensing data: a supervised learning approach. In *International Geoscience and Remote Sensing Symposium, IGARSS 2014*, April 2014.

[146] J.C. Platt. *Probabilities for SV Machines*, pages 61–74. MIT Press, 2000.

[147] A. Puissant, S. Rougier, and A. Stumpf. Object-oriented mapping of urban trees using random forest classifiers. *International Journal of Applied Earth Observation and Geoinformation*, 26:235–245, 2014.

[148] G. Qiu. Indexing chromatic and achromatic patterns for content-based colour image retrieval. *Pattern Recognition*, 35(8):1675–1686, 2002.

[149] J.R. Quinlan. *C4. 5: programs for machine learning*, volume 1. Morgan kaufmann, 1993.

[150] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 14–21, 2007.

[151] O. Razeghi, G. Qiu, H. Williams, and K. Thomas. *Computer aided skin lesion diagnosis with humans in the loop*, volume 7588 LNCS of *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2012.

[152] M. Robnik-Šikonja. Improving random forests. In *Machine Learning: ECML 2004*, pages 359–370. Springer, 2004.

[153] J.J. Rodrguez, L.I Kuncheva, and C.J. Alonso. Rotation forest: A new classifier ensemble method. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 28 (10):1619–1630, 2006.

[154] G. Rogers. Geograph, 2005. URL http://www.geograph.org.uk/.

[155] B.C. Russell, A. Torralba, K.P. Murphy, and W.T. Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1-3):157–173, 2008.

[156] S. Sadanand and J.J. Corso. Action bank: A high-level representation of activity in video. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1234–1241. IEEE, 2012.

[157] R.E. Schapire. The strength of weak learnability. *Machine learning*, 5(2):197–227, 1990.

[158] D. W. Schemske, B. C. Husband, M. H. Ruckelshaus, C. Goodwillie, I. M. Parker, and J. G. Bishop. Evaluating approaches to the conservation of rare and endangered plants. *Ecology*, 75(3):584–606, 1994.

[159] F. Schmid, L. Frommberger, C. Cai, and C. Freksa. What you see is what you map: Geometry-preserving micro-mapping for smaller geographic objects with

mapit. In *Geographic Information Science at the Heart of Europe*, pages 3–19. Springer, 2013.

[160] S. Schulter, P. Wohlhart, C. Leistner, A. Saffari, P.M. Roth, and H. Bischof. Alternating decision forests. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 508–515. IEEE, 2013.

[161] S. Searle. Phase 1 Habitat Survey Report - Millwater School at Bicton College. Technical report, Acorn Ecology Limited, 2011.

[162] S. Sergyán. Color histogram features based image classification in content-based image retrieval systems. In *Applied Machine Intelligence and Informatics, 2008. SAMI 2008. 6th International Symposium on*, pages 221–224. IEEE, 2008.

[163] S. E. Sesnie, P. E. Gessler, B. Finegan, and S. Thessler. Integrating Landsat TM and SRTM-DEM derived variables with decision trees for habitat classification and change detection in complex neotropical environments. *Remote Sensing of Environment*, 112(5):2145–2159, 2008.

[164] L. Shao and R. Mattivi. Feature detector and descriptor evaluation in human action recognition. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, CIVR '10, pages 477–484, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0117-6. doi: 10.1145/1816041.1816111.

[165] L.G. Shapiro and G.C. Stockman. Computer vision prentice hall. *Englewood Cliffs, NJ*, 2001.

[166] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *Computer Vision-ECCV 2006*, pages 1–15. Springer, 2006.

[167] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2008.

[168] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116–124, 2013.

[169] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the International Conference on Computer Vision*, volume 2, pages 1470–1477, October 2003.

[170] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(12):1349–1380, 2000.

[171] N. M. Smith, P. U. Lee, T. Prevt, J. Mercer, E. A. Palmer III, V. Battiste, and W. Johnson. A human-in-the-loop evaluation of air-ground trajectory negotiation. In *Collection of Technical Papers - AIAA 4th Aviation Technology, Integration, and Operations Forum, ATIO*, volume 1, pages 206–218, 2004.

[172] M. Sohrabinia. Latlon distance calculator, 2012. URL http://www.mathworks.com/matlabcentral/fileexchange/38812-latlon-distance.

[173] The Ordnance Survey. The ordnance survey, 1791. URL http://www.ordnancesurvey.co.uk/.

[174] T. Takiguchi, J. Bilmes, M. Yoshii, and Y. Ariki. Evaluation of random-projection-based feature combination on speech recognition. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 2150–2153. IEEE, 2010.

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

[176] D.C. Thompson and G.H. Klassen. Caribou habitat mapping in the southern district of keewatin, nwt: an application of digital landsat data. *Journal of Applied Ecology*, pages 125–138, 1980.

[177] M. Thompson. A standard land-cover classification scheme for remote-sensing applications in south africa. *South African Journal of Science*, 92(1):34–42, 1996.

[178] J. Tighe and S. Lazebnik. Superparsing: scalable nonparametric image parsing with superpixels. In *Computer Vision-ECCV 2010*, pages 352–365. Springer, 2010.

[179] V. Tomaselli, P. Dimopoulos, C. Marangi, A. Kallimanis, M. Adamo, C. Tarantino, M. Panitsa, M. Terzi, G. Veronico, F. Lovergine, H. Nagendra, R. Lucas, P. Mairota, C. Mcher, and P. Blonda. Translating land cover/land use classifications to habitat taxonomies for landscape monitoring: a mediterranean assessment. *Landscape Ecology*, 28(5):905–930, 2013. ISSN 0921-2973.

[180] M. Torres. Automatic habitat classification using aerial imagery. In *GIS Research UK 20th Annual Conference (GISRUK)*, volume 1, pages 11–13, 2012.

[181] M. Torres and G. Qiu. Picture the past from the present. In *Proceedings of the Third International Conference on Internet Multimedia Computing and Service*, pages 51–54. ACM, 2011.

[182] M. Torres and G. Qiu. Grass, scrub, trees and random forest. In *MAED 2012 - Proceedings of the 2012 ACM Workshop on Multimedia Analysis for Ecological Data, Co-located with ACM Multimedia 2012*, pages 1–6, 2012.

[183] M. Torres and G. Qiu. Automatic habitat classification using image analysis and random forest. *Ecological Informatics*, 2013.

[184] M. Torres and G. Qiu. Habitat image annotation with low-level features, medium-level knowledge and location information, multimedia systems. *Multimedia Systems*, 2014.

[185] A.M. Tousch, S. Herbin, and J.Y. Audibert. Semantic hierarchies for image annotation: A survey. *Pattern Recognition*, 45(1):333–345, 2012.

[186] E.E. Tripoliti, D.I. Fotiadis, and G. Manis. Modifications of the construction and voting mechanisms of the random forests algorithm. *Data & Knowledge Engineering*, 87:41–65, 2013.

[187] N.M. Trodd. Analysis and representation of heathland vegetation from near-ground level remotely-sensed data. *Global Ecology and Biogeography Letters*, pages 206–216, 1996.

[188] T. Tuytelaars, M. Fritz, K. Saenko, and T. Darrell. The nbnn kernel. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1824–1831. IEEE, 2011.

[189] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In *Computer Vision-ECCV 2006*, pages 589–600. Springer, 2006.

[190] K.E.A. van de Sande, T. Gevers, and C.G.M. Snoek. A comparison of color features for visual concept classification. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 141–150. ACM, 2008.

[191] K.E.A Van De Sande, T. Gevers, and C.G.M. Snoek. Evaluating color descriptors for object and scene recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1582–1596, 2010.

[192] J. Vanden Borre, D. Paelinckx, C.A. Mücher, L. Kooistra, B. Haest, G. De Blust, and A.M. Schmidt. Integrating remote sensing in natura 2000 habitat monitoring:

Prospects on the way forward. *Journal for Nature Conservation*, 19(2):116–125, 2011.

[193] V. Vapnik. Statistical learning theory. 1998, 1998.

[194] M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[195] M. Varma and A. Zisserman. A statistical approach to texture classification from single images. *International Journal of Computer Vision*, 62(1-2):61–81, 2005.

[196] A. Vedaldi and B. Fulkerson. Vlfeat: An open and portable library of computer vision algorithms. In *Proceedings of the International Conference on Multimedia*, MM '10, pages 1469–1472, New York, NY, USA, 2010. ACM.

[197] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 606–613, 2009.

[198] Microsoft Corp. Redmond WA. Kinect for xbox 360.

[199] C. Wah, S. Branson, P. Perona, and S. Belongie. Multiclass recognition and part localization with humans in the loop. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2524–2531, 2011.

[200] J. Wang, K. Markert, and M. Everingham. Learning models for object recognition from natural language descriptions. 2009.

[201] L. Wang. High dimensional data analysis, 2010. URL http://lilywang.myweb.uga.edu/Research/highdimension.pdf.

[202] J. Weston and C. Watkins. Multi-class support vector machines. Technical report, Citeseer, 1998.

[203] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2):210–227, 2009.

[204] Y. Xia, H. Tong, W.K. Li, and L.X. Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):363–410, 2002.

[205] L. Xie, Q. Tian, S. Yan, and B. Zhang. Hierarchical part matching for fine-grained visual categorization. Technical report, Technical Report, Department of Computer Science and Technology, Tsinghua Univerity, 2013.

[206] F. Yan, K. Mikolajczyk, M. Barnard, H. Cai, and J. Kittler. lp norm multiple kernel fisher discriminant analysis for object and image categorisation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3626–3632. IEEE, 2010.

[207] F. Yan, K. Mikolajczyk, J. Kittler, and M.A. Tahir. Combining multiple kernels by augmenting the kernel matrix. In *Multiple Classifier Systems*, pages 175–184. Springer, 2010.

[208] Y. Yang and S. Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 270–279. ACM, 2010.

[209] B. Yao, A. Khosla, and L. Fei-Fei. Combining randomization and discrimination for fine-grained image categorization. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1577–1584. IEEE, 2011.

[210] B. Yao, G. Bradski, and L. Fei-Fei. A codebook-free and annotation-free approach for fine-grained image categorization. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3466–3473. IEEE, 2012.

[211] Y. Ye, Q. Wu, J. Zhexue Huang, M.K. Ng, and X. Li. Stratified sampling for feature subspace selection in random forests for high dimensional data. *Pattern Recognition*, 46(3):769–787, 2013.

[212] C. Zhang and Z. Xie. Combining object-based texture measures with a neural network for vegetation mapping in the everglades from hyperspectral imagery. *Remote Sensing of Environment*, 124:310–320, 2012.

[213] Dengsheng Zhang, Md Monirul Islam, and Guojun Lu. A review on automatic image annotation techniques. *Pattern Recognition*, 45(1):346–362, 2012.

[214] H. Zhang, A.C. Berg, M. Maire, and J. Malik. Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2126–2136. IEEE, 2006.

[215] L. Zheng, G. Qiu, J. Huang, and H. Fu. Salient covariance for near-duplicate image and video detection. In *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pages 2537–2540. IEEE, 2011.

[216] N. Zhou, W. K. Cheung, G. Qiu, and X. Xue. A hybrid probabilistic model for unified collaborative and content-based image tagging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011.

[217] X. Zhou, M. Wang, Q. Zhang, J. Zhang, and B. Shi. Automatic image annotation by an iterative approach: incorporating keyword correlations and region matching. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 25–32. ACM, 2007.

[218] F. Zhu, L. Shao, and M. Lin. Multi-view action recognition using local similarity random forests and sensor fusion. *Pattern Recogn. Lett.*, 34(1):20–24, 2013. ISSN 0167-8655. doi: 10.1016/j.patrec.2012.04.016.

[219] A. Zien and C.S. Ong. Multiclass multiple kernel learning. In *Proceedings of the 24th international conference on Machine learning*, pages 1191–1198. ACM, 2007.