UNIVERSITY OF CALIFORNIA, MERCED

Proximate Sensing: Geographic Knowledge Discovery in On-line Photo Collections

A dissertation submitted in partial satisfaction of the

requirements for the degree

Doctor of Philosophy

in

Electrical Engineering & Computer Science

by

Chi Yan Daniel Leung

Committee in charge:

Professor Shawn Newsam, Chair
Professor Qinghua Guo
Professor Ming-Hsuan Yang

2013

The Dissertation of Chi Yan Daniel Leung is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Professor Qinghua Guo

Professor Ming-Hsuan Yang

Professor Shawn Newsam, Chair

University of California, Merced

2013

DEDICATION

To My Family...

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge the following funding sources that have financially supported my research for this dissertation: NSF CAREER grant (IIS-1150115) and the US Department of Energy Early Career Scientist and Engineer/PECASE award.

There are many individuals who have helped me throughout this journey. Without their help, my achievement would not be possible. First and foremost, I would like to thank my advisor Professor Shawn Newsam for inviting me to be a part of the pioneering graduate student group at UC Merced. He introduced me to a new horizon where I can apply as well as further advance my engineering skills. As an advisor, not only did Professor Newsam care about my academic performance, but he also cared about my future career, personal well-being, and family. I will always remember how he helped me set up meetings with faculty members from other colleges so I could gain a better understanding of what I want to be in the future. I will also remember the lovely little outfits he passed on from his children to my children. I am glad to have had such a wonderful mentor during the past six years, and I hope his example will be my role model when I have my own students in the future.

I would also like to extend my appreciation to the members of my graduate committee for their dedication and advice. Professor Qinghua Guo has helped me, a student with limited background in geographic information systems, discover the potential of my research on proximate sensing with his insightful knowledge on remote sensing. Professor Ming-hsuan Yang has introduced me to a broad range of research in the computer vision community. Professor Yang always challenges his students to elevate their research performance through hard work and creative thinking. I will remember the question he always asks whenever I start a research project: ”What’s new in your work?”

Finally, I would like to thank all of my family members for their support. Emily, you are the most loving and supportive wife I can ever pray for. Caleb, you were born when I started my graduate program. Although sometimes you and your sister can make my poor student life even harder, your faces and laughter have relieved the stress from my work. Elizabeth, you were born a few days after I took my qualifying exam. It was

an unforgettable moment during the exam when I was nervous about the exam while excited about your birth. To my parents, thank you for your love and discipline that made my success today. To my parents-in-law, thank you for your steadfast love and support.

VITA

1999	B. S. in Electrical Engineering, University of Wisconsin-Madison
2006	M. S. in Electrical Engineering, California State University-Fresno
2007-2009	Graduate Teaching Assistant, University of California, Merced
2013	Ph. D. in Electrical Engineering and Computer Science, University of California, Merced

PUBLICATIONS

S. Newsam, D. Leung, "Georeferenced social multimedia as volunteered geographic information", *S. Wang and M. F. Goodchild eds. CyberGIS: Fostering a New Wave of Geospatial Discovery and Innovation, Springer, Dordrecht, Netherlands*, 2013.

D. Leung, S. Newsam, "Can off-the-shelf object detectors be used to extract geographic information from geo-referenced social multimedia?", *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL GIS): Workshop on Location Based Social Networks (LBSN)*, 12, 2012.

D. Leung, S. Newsam, "Exploring geotagged images for land-use classification", *ACM International Conference on Multimedia(ACM MM): Workshop on Geotagging and Its Applications in Multimedia (GTAM)*, 3, 2012.

D. Leung, S. Newsam, "Proximate sensing: inferring what-is-where from georeferenced photo collections", *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2955, 2010.

S. Newsam, D. Leung, J. Floreza, O. Caballero, J. Pulido, "CBGIR: content-based geographic image retrieval", *ACM International Conference on Advances in Geographic Information Systems (ACM GIS)*, 526, 2010.

D. Leung, S. Newsam, "Proximate sensing using georeferenced community contributed photo collections", *ACM International Conference on Advances in Geographic Information Systems (ACM GIS): Workshop on Location Based Social Networks (LBSN)*, 57, 2009.

ABSTRACT OF THE DISSERTATION

Proximate Sensing: Geographic Knowledge Discovery in On-line Photo Collections

by

Chi Yan Daniel Leung

Doctor of Philosophy in Electrical Engineering & Computer Science

University of California, Merced, 2013

Professor Shawn Newsam, Chair

On-line photo sharing websites such as Flickr[1] not only allow users to share their precious memories with others, they also act as a repository of all kinds of information carried by their photos and tags. As geo-tagged photos can be easily created with the help of global position systems (GPS), we contend that the hundreds of millions of these geo-referenced images being acquired by millions of citizen sensors are a valuable

[1] http://flickr.com

source of geographic information. The objective of this dissertation is to perform geographic knowledge discovery using community-contributed geo-referenced photo collections such as available at Flickr. We present a novel knowledge discovery paradigm termed proximate sensing and demonstrate how it can be used to perform land cover and land use classification using widely applied image features in computer vision as well as text associated with the photos.

For land cover classification, a case study is performed using a supervised classification framework. More than one million photos are collected from two on-line photo sharing websites over a 100x100 km study region in the United Kingdom. The study region is further divided into 10,000 sub-regions, where each sub-region is classified into developed or undeveloped regions by analyzing the ground-level photos. The classification results are then compared based on the image features used as well as the source of the photos.

A case study of land use classification is conducted to further validate the concept of proximate sensing. More than 16,000 images are collected from Flickr over two university campuses. The images are classified into academic, sports, and residential facilities; a land use map of each campus is generated according to the classification results. Furthermore, we explore the idea of extracting geographic information semantically by applying state-of-the art object and concept detectors directly to the photo collections. Maps of object distributions are generated according to the detection results of different object detectors. The spatial analysis performed on these object maps suggests that it is possible to extract useful geographic information using these object detectors, and an experiment of land use classification is conducted to validate this finding.

Chapter 1

Introduction

1.1 Volunteered Geographic Information

Volunteered geographic information (VGI) has become a fast growing phenomenon in the Internet, where individuals can provide geographically relevant information freely. Termed by geographer Goodchild [Goo07], VGI describes how individuals act like a network of sensors to record geographic information. A successful example of a VGI oriented project is OpenStreetMap[1], which allows any registered user to contribute geographic data within a map interface. Other projects such as the Audubon Society's Christmas Bird Count[2] and Pop-versus-Soda[3] illustrate the observation of nature and human preferences across difference geographic areas.

In the VGI examples mentioned above, the objective of using the geographic information in each project is specific and unique. Users do not contribute non-map information such as events happening at a certain location to OpenStreetMap, or report non-bird wildlife observations to the bird count project. Social media websites such as Flickr, Facebook[4], and Twitter[5] on the other hand serve users with a wide variety of interests.

[1] http://www.openstreetmap.org
[2] http://birds.audubon.org/christmas-bird-count
[3] http://popvssoda.com
[4] http://www.facebook.com
[5] http://twitter.com

Although contributing geographic information is not the main goal of the users of these social media websites, the content they post whether as an image, blog, or tweet may be related to any place on the Earth. As a result, we argue that geo-referenced[6] social media can also be considered as a form of VGI.

1.2 Proximate Sensing

In remote sensing, geographic information is extracted from overhead, distant images such as taken from airplanes or satellites. These overhead images often provide good visual cues to classify different types of land cover; however, they are not conveniently accessible to the public due to copyright and other commercial issues. Moreover, geographic information conveyed by these images may not be reliable over time because most of them are not updated frequently. In contrast, ground-level images provide a different visual perspective of what-is-where. They also represent the "people's perspective" on interpreting the significance of a geographic location since different photographers can capture different scenes at the same location. Moreover, these ground-level images often contain indoor scenes that record how a building structure is used. Since many cameras, especially smartphone cameras, are equipped with global positioning system (GPS) nowadays, photos taken by these devices are stamped with time and location information. Internet users frequently contribute these geo-referenced photos, along with textual description or tags, to social media networks. Flickr alone has a collection of hundreds of millions of geo-referenced photos and millions of photos are added each month. With such a large amount of frequently updated geo-referenced photo collections freely available in the public domain, we ask the question whether they can be used to perform geographic analysis instead of using overhead images. Take, for example, the four photos referenced on the map in figure 1.1. The content of these photos provides rich geographic information about the locations at which they were taken. In this respect, we proposed the term *proximate sensing* [LN09, LN10, NL13] to refer to geographic discovery using ground-level images of nearby objects and scenes.

[6] We use the term geo-referenced to indicate that a multimedia object has least approximate location metadata associated with it.

Figure 1.1: We conjecture that the visual content of georeferenced images can be used to derive maps of what-is-where on the surface of the earth.

Proximate sensing has a great potential to facilitate a broad variety of applications. One of the popular applications is Street View from Google Maps[7], which allows users to view the surrounding scene of a location in a street level. Besides displaying what-is-where on the surface of the earth, proximate sensing can also be applied in scientific analysis such as monitoring the effects of climate or geological change where ground-level images taken over a period of time can be compared. Furthermore, proximate sensing can assist existing remote sensing techniques to enhance the performance of many geographic analyses such as land cover and land use classification.

1.3 Related Work

There is a growing body of research on analyzing geo-referenced community contributed photo collections. Methods have been developed which leverage the collections to 1) annotate novel images; 2) annotate geographic locations; and 3) perform geographic knowledge discovery. Proximate sensing is an example of this last class.

[7]http://www.maps.google.com/streetview

Leveraging Collections to Annotate Novel Images

Automated annotation is essential for managing large image collections. Methods have been developed that leverage large sets of geo-referenced images to semantically annotate novel images whose location is known. This is particularly useful for images captured using GPS enabled cameras as the system generated annotation allows the images to be organized and searched at a more meaningful way than with low-level image descriptors such as color or texture. Methods have been developed for suggesting tags such as "surfer", "wave", and "Santa Barbara" for a photograph of someone surfing in Santa Barbara, California [MKM08]; for assigning a constrained set of event/activity labels such as "a visit to the beach" or "wedding" [JL08]; for annotating groups of images at the event ("skiing") or scene ("coast") level [CLK08]; for annotating the identities of people appearing in an image [NYG05]; and for linking images, such as a photograph of the Arc de Triomphe, to relevant Wikipedia articles [QLV08].

Collections of geo-referenced images have also been used to annotate the locations of novel images–that is, to estimate where in the world the photo was taken. Methods have been developed to geo-locate Web cameras distributed around the United States based on image variations relating to the diurnal cycle and weather [JSR07]; to geo-locate a single image using only its visual content [HE08] as well as textual tags [GJY09] by performing similarity search against a reference collection; and to estimate coarse image location by first clustering a reference collection and then indexing the novel image based on its visual content and textual tags [CYL09, CBH09, CPC08].

Leveraging Collections to Annotate Geographic Locations

Collections of geo-referenced images have also been used to annotate geographic locations, a task in which on-line photo collections are considered more explicitly as VGI as the objective is more in line with the problem of determining what-is-where on the surface of the Earth. Methods have been developed for visually annotating prominent landmarks with representative images at the city [CBH09] and world-wide [ZZS09] scales; to suggest representative tags as well as images for geographic loca-

tions [KN08, KNA07, NYG05]; and to automatically generate tourist maps showing popular landmarks as vectorized icons [CBG09].

Leveraging Collections for Geographic Knowledge Discovery

The Mirriam-Webster dictionary describes geography as "a science that deals with the description, distribution, and interaction of the diverse physical, biological, and cultural features of the Earth's surface". Accordingly, we consider geographic knowledge discovery to be a process that derives knowledge about what-is-where on the surface of the Earth in the broad sense of the term "what". Simply put, it can be used to generate maps not only of the physical aspects of our world, such as the terrain, but also of the abstract aspects, such as culture and natural or man-made behavior, of the world. While there has been relatively little work on using geo-referenced on-line photos for geographic knowledge discovery, we feel it has significant potential for realizing the full worth of geo-referenced on-line photos as VGI, particularly as an alternate to traditional means of geographic knowledge inquiry.

Examples of work in this area include using large collections of geo-referenced images to discover spatially varying (visual) cultural differences among concepts such as "wedding cake" [YYQ09]; to discover interesting properties about popular cities and landmarks such as the most photographed locations [CBH09]; to estimate weather satellite images using widely distributed Web cameras [JSR07]; and to create a map-like partitioning of a country-sized region into geographically coherent subregions [CPC08].

Our work on proximate sensing as applied to geo-referenced on-line photo collections, however, represents a more comprehensive framework for geographic knowledge discovery particularly of phenomena often not observable through other means.

1.4 Land Cover and Land Use Classification

Land cover and land use classification, and their changes, are two fundamental geographic tasks. While land cover and land use are related and often overlap, their distinctions are important. Land cover "is the physical material at the surface of the earth. It is the material that we see and which directly interacts with electromagnetic radiation and causes the level of reflected energy that we observe as the tone or the digital number at a location in an aerial photograph or satellite image. Land covers include grass, asphalt, trees, bare ground, water, etc. ... Land use, by contrast, is a description of how people use the land. Urban and agricultural land uses are two of the most commonly recognised high-level classes of use. Institutional land, sports grounds, residential land, etc. are also all land uses" [FCW05]. The scope of this dissertation will focus on these two geographic tasks.

1.5 Image Content Analysis

Effective image content analysis is key to the goal of using on-line photo collections for geographic discovery. This section describes the range of analysis methods that can be brought to bear on this problem.

1.5.1 Low-level Analysis

Much progress has been made over the past several decades on extracting so-called low-level features from images and videos. Standard low-level features include color histograms which summarize the distribution of pixels in an image in a (typically) three-dimensional color space, and texture features which characterize the spatial distribution of pixel intensities, typically by applying spatial filters tuned to different scales and orientations. These features are usually extracted globally from an image and thus do not contain information about the spatial layout of an image. In Chapter 2, we will see

how low-level features such as edge and color histograms will assist us in performing land cover classification.

Local analysis based on low-level features extracted from perceptually salient regions has advanced a number of image analysis tasks over the last decade. Local invariant features avoid the challenging problem of segmentation and instead focus on image patches which can be reliably detected and characterized independent of a range of image transformations, including geometric transformations such as rotation and scaling, as well as photometric transformations that result from changes in illumination, etc. The most popular of these features is David Lowe's Scale Invariant Feature Transform (SIFT) [Low99, Low04]. The invariance provided by these features stands to be critical for analyzing on-line photos since the images in these collections exhibit great diversity not only in content but also viewpoint and environment. We will SIFT features to perform land use classification in Chapter 3.

1.5.2 Mid-level Analysis

It is well known that low-level features do not characterize an image at a semantic level. They will therefore be limited in their capacity to extract geographically relevant information from on-line photos. Mid-level analysis potentially offers richer representations which, while not at the level of objects, concepts, events, and activities, still helps to narrow the semantic gap.

Of particular interest is the work by Oliva and Torralba [OT01] on modeling the shape of the scene in an image using so-called gist features. This method bypasses the segmentation and processing of individual objects or regions and instead uses the spatial envelope of an image to assign a set of perceptual dimensions such as naturalness, openness, roughness, expansion, and ruggedness that represent the dominant spatial structure of a scene. Such dimensions could clearly be informative for geographic discovery. The authors indeed show their approach generates a multidimensional space in which scenes sharing membership in semantic categories (e.g., streets, highways, coasts) are projected close together. We will discuss the effectiveness of this feature on land cover classifica-

tion in Chapter 2.

1.5.3 High-level Analysis: Object and Concept Detection in Images and Videos

Automated object and concept detection can clearly facilitate geographic discovery in social multimedia. While such semantic-level understanding remains a challenging problem, significant progress has been made in computer vision research over the past decade on generic object and concept detection. This progress is in large part a result of image analysis based on local invariant features which, besides the invariance properties mentioned above, are robust to occlusion, a major challenge in object detection. A good survey on state-of-the-art techniques in object and concept detection can be found in [PHS06]. The other development that has advanced the field is the availability of standard training and evaluation datasets such as Caltech-256 [GHP07] which contains over 30,000 images of 256 object classes and the MIT_CSAIL Database of Objects and Scenes [TMF04] which contains over 72,000 images of 107 object classes.

Fortunately, a wide range of pre-trained object and concept detectors for images have recently been made available. Ready-to-be-applied detectors include:

- MediaMill 101 - Born out of the TRECVID video retrieval competition, MediaMill [SWG06] provides trained classifier models for 101 concepts such as animal, dog, basketball, sports, food, and many others which are likely to be relevant to performing geographic discovery. Local color-texture features are used.

- Columbia-374, VIREO-374, CU-VIREO374 - Columbia-374 [YCK07] also emerged out of TRECVID. It provides pre-trained detectors for 374 concepts. The Columbia-374 detectors utilize three visual features: edge direction histograms, Gabor texture features, and grid color moments. VIREO-374 [JYN10] provides pre-trained detectors for the same concepts as Columbia-374. However, it utilizes local invariant features. CU-VIREO374 [JYC08] fuses the global features of Columbia-374 with the local features of VIREO-374.

- VIREO-WEB81 - VIREO-WEB81 [ZWN10] provides detectors for 81 concepts. It differs from the TRECVID-based detectors above in that it is trained using approximately 260K Flickr images manually annotated with 81 concepts. It includes concepts such as book, cat, computer, dancing, food, person, running, and sports. The detectors utilize quantized local features and grid-based color moment and wavelet texture features.

In Chapter 4, we will explore the possibility of using object detectors to solve land use classification problems.

1.6 Overview of Dissertation

This dissertation focuses on two examples of using proximate sensing for geographic knowledge discovery: it demonstrates how proximate sensing can be use to perform land cover and land use classification. In the following chapter, a case study of using proximate sensing to perform land cover classification is presented. Although land cover classification is often possible using remote sensing techniques, we choose this application as a proof of concept experiment due to the availability of ground truth data. We also investigate the effects of different image features as well as the source of the images have on the classification performance. A case study of using proximate sensing to perform land use classification is presented in Chapter 3, where we evaluate how ground-level images can help distinguish different types of land use. We further investigate the application of land use classification by exploring the idea of using different object detectors in Chapter 4. We demonstrate how this top-down image understanding approach can benefit land use classification. The last chapter concludes this dissertation with a summary as well as a discussion on the future direction of this work.

1.7 Summary of Contributions

The contributions of this dissertation are summarized as follows:

- The conjecture that large collections of geo-referenced photo collections can be used to derive maps of what-is-where on the surface of the Earth.
- The first work to use the geo-referenced on-line photo collections to infer what-is-where on the surface of the Earth on a large scale.
- A novel framework for using state-of-the-art techniques in multimedia content analysis, in particular automated image understanding and statistical text analysis, to perform geographic knowledge discovery in large collections of on-line photos.
- The use of proximate sensing to complement the shortcoming of remote sensing for land use classification.
- The first investigation into using object detectors for geographic knowledge discovery.

Chapter 2

Land Cover Classification

This chapter describes a case study in which geo-referenced community contributed images are used to perform land cover classification into developed and undeveloped regions. Although land cover classification can often be obtained through analysis of satellite images, we focus on this problem as a case study to establish the validity of proximate sensing for geographic knowledge discovery. This is also a problem for which there is ground truth data to facilitate the evaluation. As we will discuss in the next chapter, we do not propose proximate sensing as a replacement for traditional remote sensing but as a complementary technique especially for phenomena not easily observable from above.

In this chapter, we introduce a series of experiments to validate the concept of using proximate sensing to perform land cover classification. We first create a small dataset of 5000 geo-referenced photos and train a binary classifier using edge histograms as image features. With the successful validation of our concept, we then extend our experiment to a larger pair of datasets consisting of almost 1 million photos and with a larger geographical coverage. In this experiment, we focus on how the quality of different data sources as well as different image features affect the performance of classification. We finally present an experiment on removing images that are not geographically informative in an attempt to improve quality the of the dataset for classification.

The work presented in this chapter was published as a peer-reviewed workshop paper at the International Workshop on Location Based Social Networks in 2009 [LN09] and as an oral paper in the IEEE Conference on Computer Vision and Pattern Recognition in 2010 [LN10]. An extended journal version of the work is current being prepared.

2.1 The Initial Experiment

To validate our hypothesis that land cover classification is possible using ground-level images, we perform land cover classification into developed and undeveloped regions as an initial test bed. We choose our study area as an approximately 33x42km section of California between and partly encompassing the cities of San Jose and Santa Cruz. We chose this region because 1) it encompasses a wide diversity of land cover types and 2) it contains a large number of geo-referenced Flickr images due to its proximity to the technology-savvy populace of Silicon Valley. Figure 2.1 contains a map downloaded from The National Map of the United States Geological Survey of this region indicating the distribution of the National Land Cover Database (NLCD) classes. The southern part of San Jose is visible in the top section of this image and Santa Cruz is visible in the middle lower section.

The 15 NLCD classes in the study area can naturally be divided into developed–21, 22, 23, and 24–and undeveloped superclasses–11, 31, 41, 42, 43, 52, 71, 81, 82, 90, and 95. This binary partitioning results in the developed/undeveloped map shown in Figure 2.3(a). The goal of this work is to estimate these superclasses using ground-level images.

We used the Flickr application programming interface (API) to download geo-referenced images for the study region. We downloaded the medium sized versions of the images which have a maximum dimension of 500 pixels. In order to localize the analysis, we partitioned the region into 3x3km tiles and downloaded a maximum of 50 Flickr images per tile. This tile size was chosen as a tradeoff between localizing the analysis and ensuring that most of the tiles contained enough images for the analysis to be meaningful.

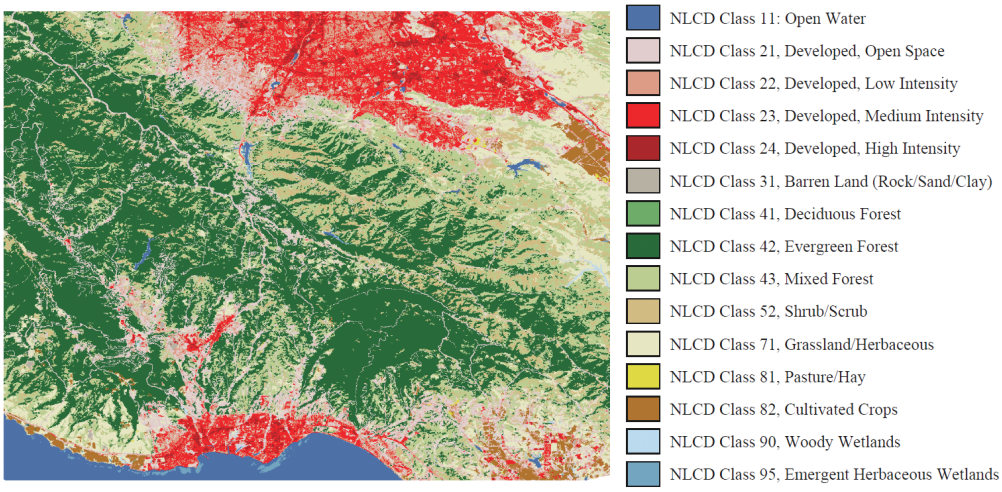


Figure 2.1: The study region consists of an approximately 33x42km section of California between and partly encompassing the cities of San Jose and Santa Cruz. The southern part of San Jose is visible in the top section of this NLCD map and Santa Cruz is visible in the middle lower section.

A total of 5509 images were downloaded for the 14x11=154 tiles for an average of 35.8 images/tile. There were 10 tiles which contained no images. Figure 2.2 illustrates a mosaic of sample images used in this experiment across the study region. We can see the diversity of images collected from Flickr. Figure 2.3(b) shows the distribution of Flickr images for the study region. Overlaid on this figure is a smoothed outline of the larger developed regions from the NLCD map. This outline serves as a landmark for comparing results and was machine generated by applying a Gaussian smoothing filter to the binary map in Figure 2.3(a), thresholding the result and using an edge detector to determine the resulting boundary.

Since the analysis using the Flickr images is performed at the 3x3km tile scale, we derived two ground truth items from the NLCD map at the same scale and grid. The first indicates the ratio of developed to total land cover in each of the tiles and is computed in a straightforward manner from the NLCD binary map. It is termed the *NLCD ratio map* and is shown using a heatmap in Figure 2.4(a). The second ground truth item is a binary classification of each tile as developed or undeveloped and is computed by applying a threshold of 0.5 to the NLCD ratio map. It is termed the *NLCD binary classification map* and is shown in Figure 2.4(b). The following sections explore how well these ground

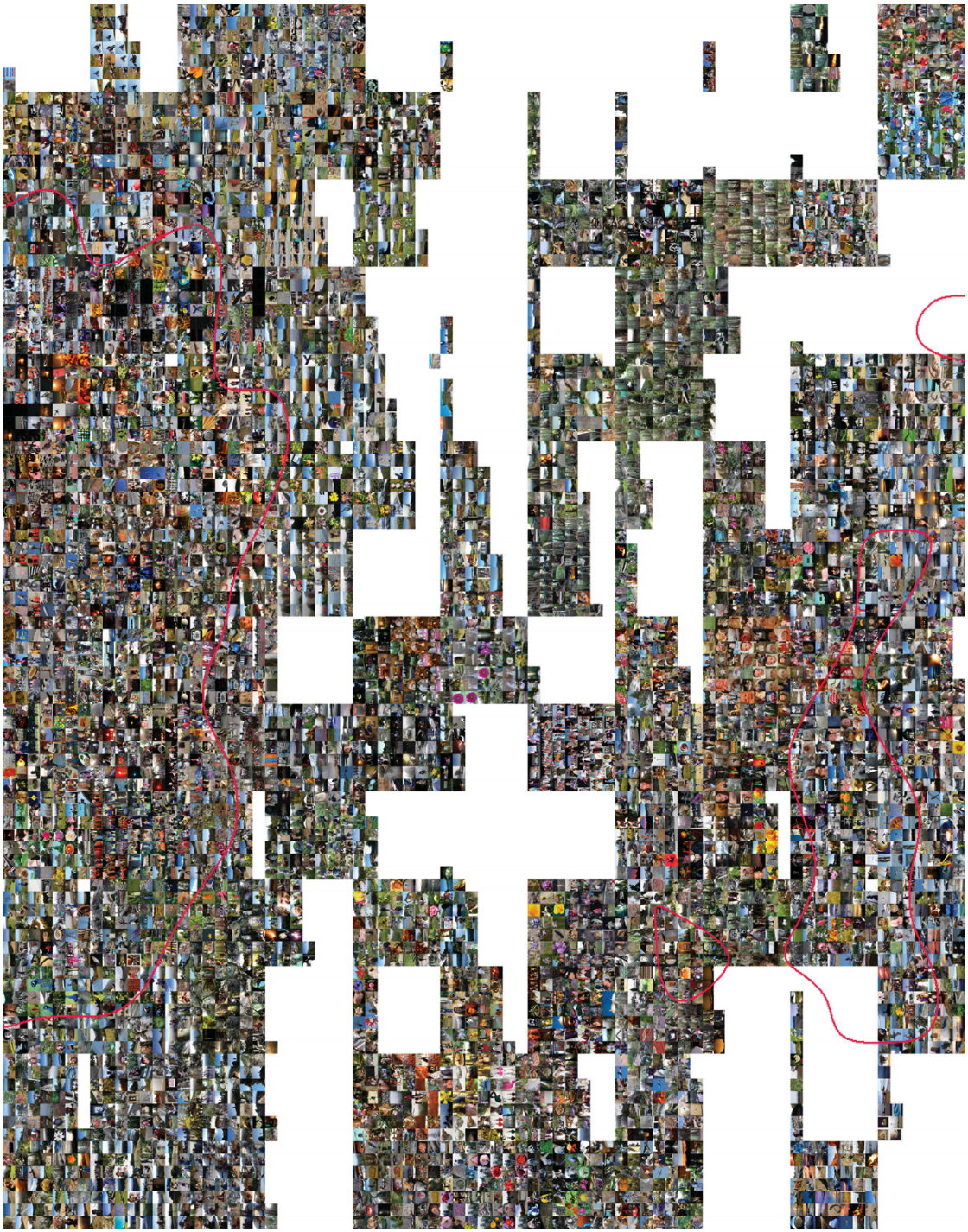
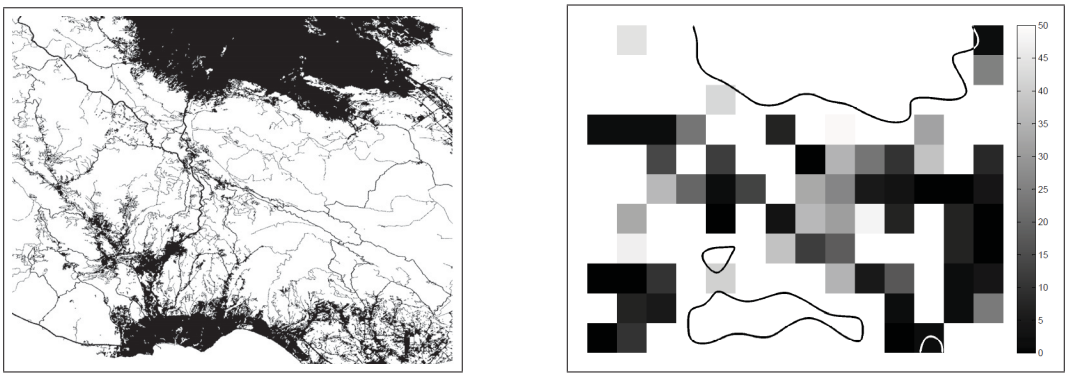


Figure 2.2: A mosaic of sample images used in this experiment.



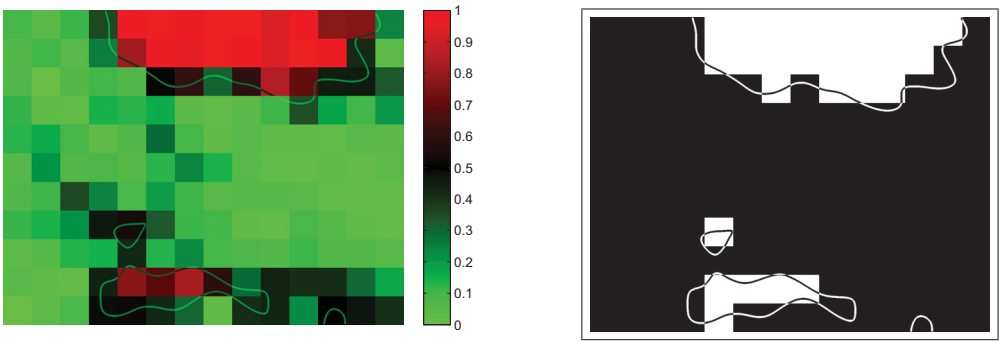
(a) (b)

Figure 2.3: (a) The 15 NLCD classes in the study region naturally divide into developed and undeveloped regions shown in black and white respectively. (b) The distribution of Flickr images for the study region. No images were available for 10 of the tiles. Overlaid on this figure is a smoothed outline of the larger developed regions from the NLCD map to serve as a landmark for comparing results.

truth maps can be estimated using the geo-referenced Flickr images.

2.1.1 Results: Manually Labeled Images

This experiment explores how well land cover classes can be predicted using geo-referenced ground level images *which have been manually labeled.* Specifically, we evaluate how accurately the ground truth NLCD ratio and binary classification maps can be estimated from a manually labeled set of Flickr images. Since the goal is to determine whether a geographic location is developed or undeveloped, we use the same binary labels for the images. Two "users" viewed the 5509 images and independently labeled them as developed if they depicted scenes containing constructed materials such as used in houses, buildings, etc., and labeled them as undeveloped if they were of open areas and/or contained mostly trees and vegetation. These criteria will of course result in "incorrectly" labeled images which will limit the approach. For example, indoor scenes will always be classified as being "developed" even though they might have been taken inside isolated homes in rural regions. Nonetheless, the results below show even this simple labelling is effective for estimating the ground truth NLCD ratio and binary classification maps. Figure 2.5 shows examples of labeled Flickr images.

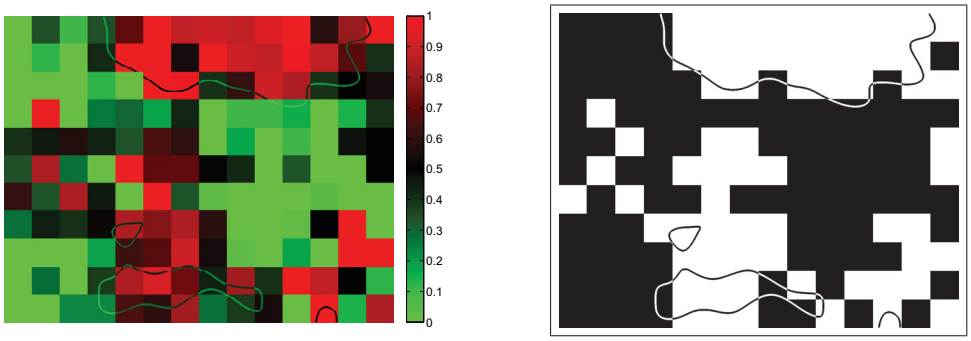


(a) (b)

Figure 2.4: Ground truth data derived from the NLCD binary map. (a) NLCD ratio map indicating the ratio of developed to total land cover for each of the tiles. (b) NLCD binary classification map with tiles labeled as developed (white) or undeveloped (black).

(a) Developed. (b) Developed. (c) Undeveloped. (d) Undeveloped.

Figure 2.5: Examples of Flickr images manually labeled as developed or undeveloped.



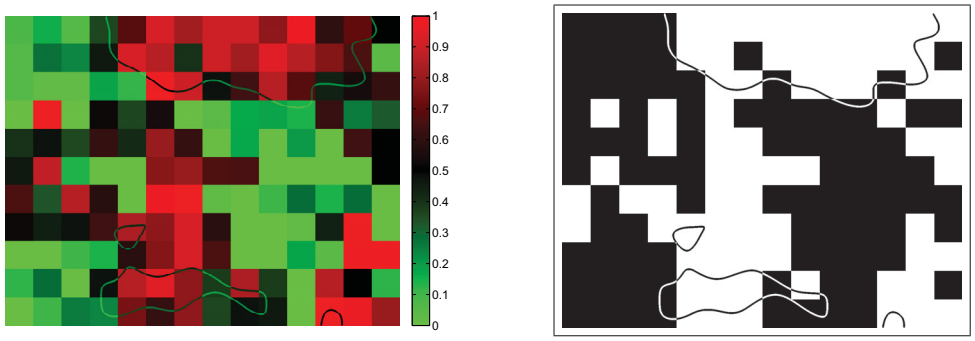
(a) (b)

Figure 2.6: Results for manual labelling of Flickr images for user 1. (a) Ratio map indicating the ratio of Flickr images labeled as developed to the total number of images per tile. (b) Binary classification map indicating tiles labeled as developed (white) or undeveloped (black).

We derived ratio maps from the labeled Flickr images by computing the ratio of images labeled as developed to the total number of images in each tile. Figures 2.6(a) and 2.7(a) show the ratio maps corresponding to manual labelling performed by users 1 and 2. These maps are quite similar to the ground truth NLCD map in Figure 2.4(a) indicating that labeled, geo-referenced ground level images can be used to predict land cover information.

We derived binary classification maps from the user ratio maps the same way the NLCD binary classification map was produced: tiles whose ratio was greater than 0.5 were classified as developed and the remainder were classified as undeveloped. Figures 2.6(b) and 2.7(b) show the binary classification maps corresponding to manual labelling performed by users 1 and 2. Again note the similarity between these maps and the ground truth NLCD binary classification map in Figure 2.4(b).

We quantitatively evaluated the similarity between the NLCD and user ratio maps by calculating the correlation coefficient between the tile ratio values taken as observations of random variables. Specifically, if random variables X and Y represent the ground truth NLCD and user tile ratio values respectively then the correlation coefficient for the user ratio map is computed as $\rho_{XY} = cov(X,Y)/\sigma_X\sigma_Y$ where $cov(X,Y)$ is the covariance of X and Y and σ_X (σ_Y) is the standard deviation of X (Y). ρ_{XY} ranges from -1 to 1 with a value of 0 indicating no correlation, and values of -1 and 1 indicating



(a) (b)

Figure 2.7: Results for manual labelling of Flickr images for user 2. (a) Ratio map indicating the ratio of Flickr images labeled as developed to the total number of images per tile. (b) Binary classification map indicating tiles labeled as developed (white) or undeveloped (black).

Table 2.1: Quantitative evaluation of how well land cover can be estimated using *manually* labeled, geo-referenced ground level images. The first data row gives the correlation coefficient between the ground truth NLCD ratio map and the ratio map derived from the labeled images. The second data row gives the percent of tiles that have the same label in the NLCD binary classification map and the classification map derived from the labeled images. Columns user 1 and user 2 correspond to the manually labeled Flickr images. Random corresponds to a random labelling of the images. The other two columns correspond to labellings in which all the images are labeled as developed or undeveloped.

	User 1	**User 2**	**Random**	**All labeled dev.**	**All labeled undev.**
Correlation coefficient (ρ)	0.651	0.604	0.186	N/A	N/A
% with same label	73.4%	69.5%	55.6%	37.0%	63.0%

strong negative and positive correlation respectively.

We quantitatively evaluated the similarity between the NLCD and user binary classification maps by calculating the percentage of tiles with the same label.

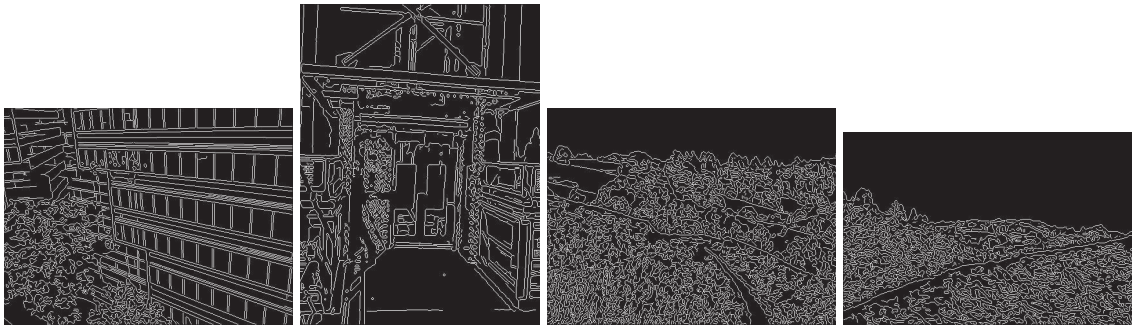
Table 2.1 shows the correlation coefficient between the NLCD and user ratio maps, and the percent agreement between the NLCD and user binary classification maps. Values are also given for control cases including a random labelling of the Flickr images as well as the cases where all the Flickr images are labeled as developed or undeveloped.

2.1.2 Results: Automatically Labeled Images

This experiment explores how well land cover classes can be predicted using geo-referenced ground level images *which have been automatically labeled.* The ratio and binary classification maps are now derived using Flickr images which have been labeled as developed or undeveloped using a trained classifier.

The dimensionality of the raw Flickr image space is too high for performing effective classification so instead we represent the images using compact descriptors. We choose edge histogram descriptors which quantify the distribution of edges at different orientations. This is motivated by the observation that images of developed scenes typically have a higher proportion of horizontal and vertical edges than images of undeveloped scenes. This is evident in the edge images in Figure 2.8 corresponding to the example Flickr images. Following the method outlined in [MOV98], we apply a set of five 2x2 linear filters to detect edges at roughly horizontal, vertical, 45° diagonal, 135° diagonal, and isotropic (non-orientation specific) directions. A threshold is applied to the outputs of these filters and the ratios of edges in various directions are summarized in a five bin L1 normalized histogram. In summary, each Flickr image is represented by a five dimensional edge histogram feature vector. The caption for each subfigure in 2.8 provides these vectors for the sample Flickr images.

We chose a support vector machine (SVM) as the classifier as it has proven effective as a general purpose classifier. We divide the 5509 Flickr image dataset into two similarly sized sets: a training set which is used to train the classifier and a test set which is used to produce the ratio and binary classification maps. The training/test partitioning is done at the tile level–half of the images for each tile are grouped into a single training set and the other half are kept with the tile for the evaluation. The training set contains 2740 images, 833 labeled as developed and 1907 labeled as undeveloped. An SVM classifier with a radial basis function kernel is trained on this set using five fold cross validation and a grid search for parameter selection. A separate classifier is trained for the user 1 and user 2 labeled datasets. The user 1 SVM achieves a validation classification rate of 72.2% (on the training data) and a classification rate of 70.2% on the held-out test images. The

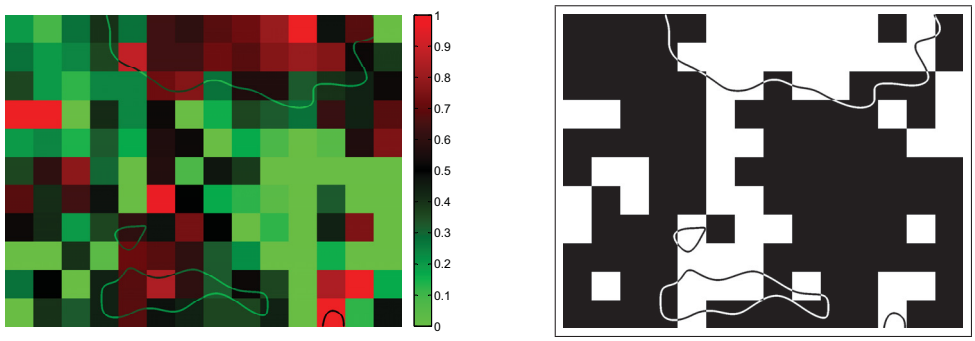


(a) (0.20 0.08 0.18 0.41 0.13) (b) (0.12 0.22 0.15 0.36 0.15) (c) (0.09 0.14 0.12 0.17 0.47) (d) (0.17 0.09 0.13 0.29 0.33)

Figure 2.8: Edge images corresponding to example Flickr images in Figure 2.5. The captions under each subfigure contain the five dimensional edge histogram feature vectors. The components of these vectors indicate the relative strength of edges in the horizontal, vertical, 45° diagonal, 135° diagonal, and isotropic (non-orientation specific) directions.

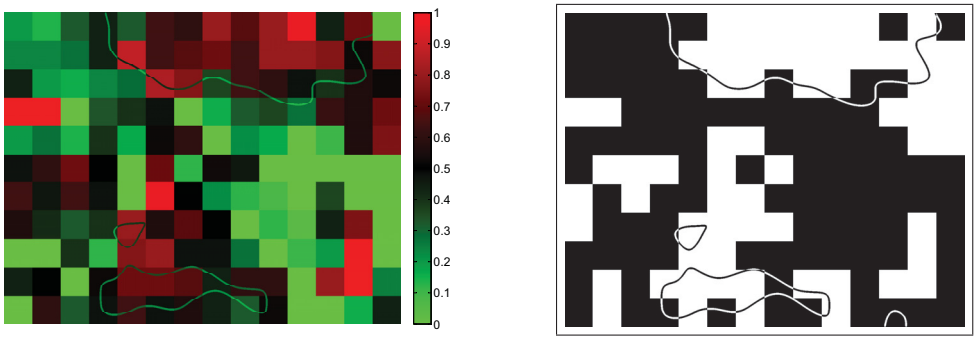
user 2 SVM achieves a validation classification rate of 68.9% and a classification rate of 69.0% on the held-out test images. An SVM trained using randomly labeled images achieves a validation classification rate of 50.7% and a classification rate of 49.2% on the held-out test images as expected.

After the classifiers have been applied to the test images for the tiles, we use the same approach as in Section 2.1.1 to generate the ratio and binary classification maps. The tiles in the SVM ratio maps indicate the ratio of images classified as developed by the SVMs to the total number of images in the tile. The tiles in the SVM binary classification maps indicate whether a majority of the tiles are classified as developed. These maps are shown in figures 2.9 and 2.10. We also quantitatively evaluate the results by computing the correlation coefficient and percent agreement with the ground truth NLCD maps. These results are listed in table 2.2. We also show results for an SVM trained using a randomly labeled image set.



(a) (b)

Figure 2.9: Results for automatic labelling of Flickr images using classifier trained using user 1 labeled images. (a) Ratio map indicating the ratio of Flickr images labeled as developed to the total number of images per tile. (b) Binary classification map indicating tiles labeled as developed (white) or undeveloped (black).



(a) (b)

Figure 2.10: Results for automatic labelling of Flickr images using classifier trained using user 2 labeled images. (a) Ratio map indicating the ratio of Flickr images labeled as developed to the total number of images per tile. (b) Binary classification map indicating tiles labeled as developed (white) or undeveloped (black).

Table 2.2: Quantitative evaluation of how well land cover can be estimated using *automatically* labeled, geo-referenced ground level images. The first data row gives the correlation coefficient between the ground truth NLCD ratio map and the ratio map derived from images labeled using an SVM classifier. The second data row gives the percent of tiles that have the same label in the NLCD binary classification map and the classification map derived from the classified images. Columns SVM 1 and SVM 2 correspond to SVMs trained using the Flickr images labeled by user 1 and user 2. Random corresponds to an SVM trained using a randomly labeled set of images.

	SVM 1	**SVM 2**	**Random**
Correlation coefficient (ρ)	0.559	0.509	0.022
% with same label	77.3%	71.4%	35.7%

2.1.3 Discussion

Even though the proposed method represents an initial, straightforward approach to using geo-referenced community contributed photos for land cover estimation, the results are significant. The user and SVM generated ratio and binary classification maps are similar both in terms of levels and spatial distribution to the ground truth NLCD data. The maps derived from the Flickr images tend to overestimate how developed a region is. This makes sense since photos will frequently depict "developed" scenes even when taken in relatively remote locations–images taken indoors are a prime example. The correlation coefficients between the user and SVM generated ratio maps and the ground truth NLCD data average 0.628 and 0.534 respectively. This indicates a high amount of correlation. And, the percentage of similarly classified tiles between the user and SVM generated binary classification maps and the ground truth NLCD data average 71.45% and 74.35%. This latter value is especially significant since the maps are derived in a completely automated fashion.

It is interesting that the user generated ratio maps are more similar to the ground truth than the SVM ratio maps while the opposite is true for the binary classification maps. A possible explanation for this is that while humans can manually classify individual Flickr images as developed or undeveloped more accurately than the SVM classifiers since they incorporate a higher-level understanding of the images, such as their context, into their decision, the SVM classifiers produce better results when the individually labeled images are aggregated to derive a binary classification since they are better at learning the overall distributions of the two classes.

2.2 Larger Scale Experiments

We extend our initial experiment to a larger pair of datasets consisting of almost 1 million photos. We also expand the geographical coverage of our study region. The objective of this experiment is to answer the following questions:

- Can existing geographic knowledge be used to provide labeled training data in a weakly-supervised manner? Since we have the ground truth data, we want to investigate the possibility of labeling the training data using this ground truth data.
- What is the effect of the photographer's intent when he or she captures the photograph? As the photo collections come from two social media that serve users with different purposes, we would like to find out how a photographer's intent affects the integrity of the geographic information in both datasets.
- Does it help to filter out uninformative images? Since not all on-line photos are geographically informative, we will investigate if removing these uninformative images will improve the classification performance.

2.2.1 Datasets

The study area is the 100x100 km of Great Britain corresponding to the TQ square in the British national grid system. This region encompasses the London metropolitan area and thus includes a range of developed and undeveloped land cover classes. Figure 2.11 shows the location of the study region in correspondence to the map of Great Britain.

We use the publicly accessible Countryside Information System (CIS) to download the Land Cover Map 2000 (LCM2000) of the United Kingdom's Centre for Ecology & Hydrology for the TQ study region. We focus on the LCM2000 Aggregate Class (AC) data which provides the percentages of ten land cover classes at the 1x1 km scale. Figure 2.12 shows the dominant classes for the TQ region.

We focus on binary classification into developed and undeveloped regions so the ten land cover classes are further aggregated into a developed superclass consisting of LCM AC:7 Built up areas and gardens, and an undeveloped superclass consisting of the remaining nine classes. We derive two ground truth datasets, one which indicates the fraction developed for each of the 10K 1x1 km tiles in the TQ region and another which simply indicates a binary label for each tile by applying a threshold of 0.5 to the frac-

Figure 2.11: Location of TQ square in correspondence to the map of Great Britain.

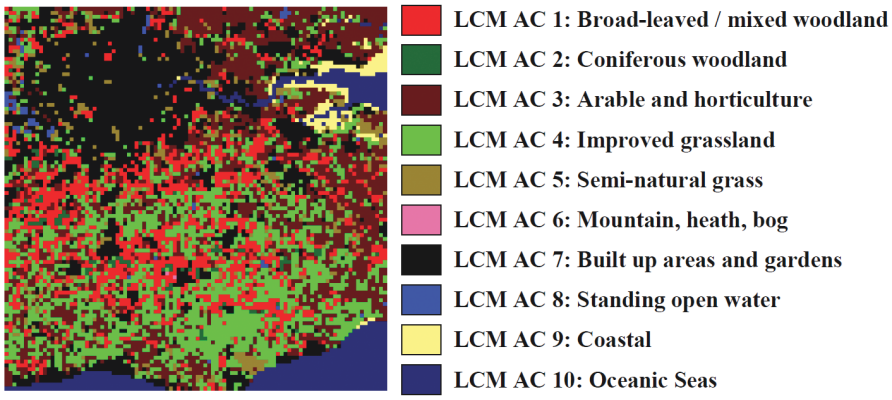
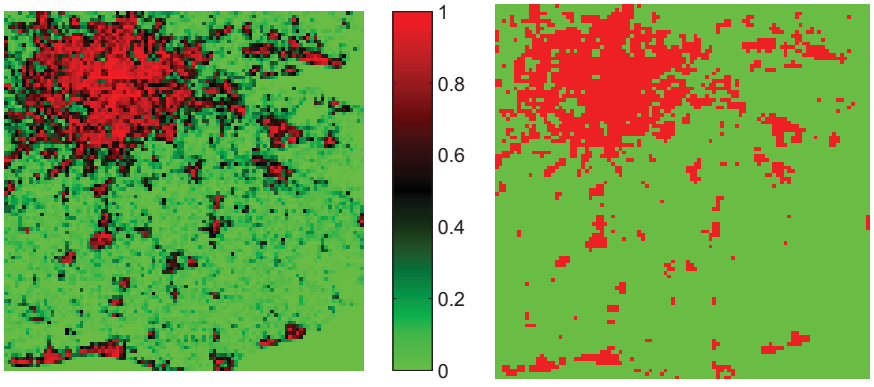


Figure 2.12: The dominant Land Cover Map 2000 Aggregate Classes (AC) for the TQ study area. This area measures 100x100 km and encompasses the London metropolitan area which appears towards the north-west.



(a) (b)

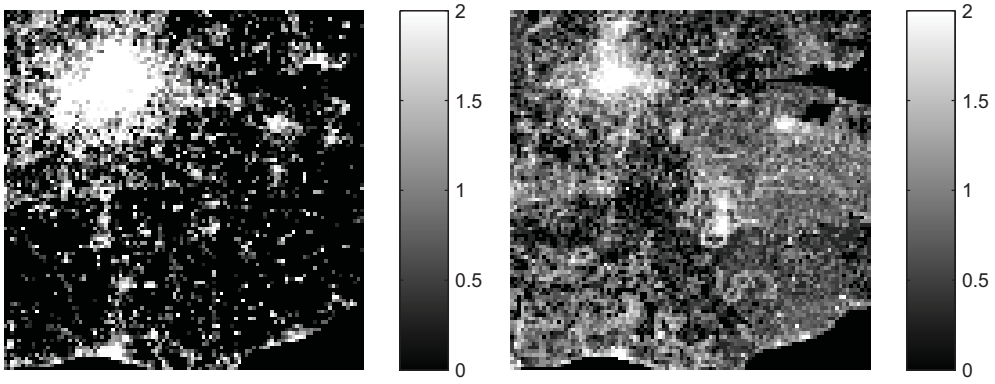
Figure 2.13: Ground truth derived from the LCM 2000 AC data. (a) Map of fraction developed values for each 1x1 km tile. (b) Map of binary labels in which red and green are used to indicate developed and undeveloped tiles respectively. The binary labels are derived from the fraction values by applying a threshold of 0.5.

tion developed. We refer to the first dataset as the ground truth *fraction values*[1] and the second as the ground truth *binary labels*. Figure 2.13 shows the ground truth datasets as maps.

We compile two geo-referenced image collections for the study area. First, we use the Flickr application programming interface (API) to download approximately 920K Flickr images located within the TQ region. The longitude and latitude information provided by the Flickr API is then used to assign each image to a 1x1 km tile. Figure 2.14(a) shows the distribution of the Flickr images. While Flickr contains a large collection of geo-referenced images, its spatial coverage is not uniform. For our study area, 5,420 of the 10K 1x1 km tiles do not contain any Flickr images. The 4,580 tiles with images contain an average of 200.7, a median of 10, and a maximum of 53,840 images.

We download a second set of potentially more geographically informative images from the Geograph Britain and Ireland (GBI) project whose aim is to "collect geographically representative photographs and information for every square kilometre of Great Britain and Ireland". This project contains over two million photos contributed by over 10K users and allows us to investigate the effect of photographer intent. We use the GBI API to download approximately 120K Geograph images for the study area. While there are

[1]These fraction values are the same as the ratio values in the previous experiment.



(a) (b)

Figure 2.14: The distribution of images for the TQ study region in the (a) Flickr and (b) Geograph datasets. On a base-10 logarithmic scale.

fewer Geograph images, they are more uniformly distributed than the Flickr images as shown in Figure 2.14(b). Now, only 614 of the 10K tiles do not contain any Geograph images and all but a few of these correspond to ocean. The remaining 9,386 tiles contain an average of 12.9, a median of 5, and a maximum of 1,458 images.

In order to investigate whether the Geograph or Flickr images are better for binary land cover classification, we use a common evaluation dataset consisting of the 4,441 tiles which contain images from both datasets. These tiles contain over 90K Geograph and over 900K Flickr images. This evaluation dataset is split into disjoint training and test sets with 400 and 4,041 tiles respectively.

Figure 2.15 shows sample images from the Flickr and Geograph datasets. Pairs of even/odd rows show Flickr/Geograph images for the same 1x1 km tiles. The top two pairs of rows are for tiles with a developed fraction of 1.0 while the bottom two pairs of rows are for tiles with a developed fraction of 0. The Geograph images tend to be more geographically informative although both support land cover classification as demonstrated in the experiments below.

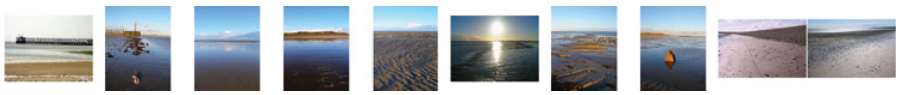
(a) Sample Flickr images from a 1x1 km tile with a developed fraction of 1.0.

(b) Sample Geograph images from the same tile as above.

(c) Sample Flickr images from a 1x1 km tile with a developed fraction of 1.0.

(d) Sample Geograph images from the same tile as above.

(e) Sample Flickr images from a 1x1 km tile with a developed fraction of 0.



(f) Sample Geograph images from the same tile as above.

(g) Sample Flickr images from a 1x1 km tile with a developed fraction of 0.

(h) Sample Geograph images from the same tile as above.

Figure 2.15: Sample images from the Flickr and Geograph datasets.

2.2.2 Experiments

The goal in all experiments is to investigate how well the visual feature based classification of individual geo-referenced images can be used to create developed/undeveloped land cover maps similar to the ground truth maps. We constrain the labels of the images to the same developed and undeveloped superclasses–that is each image is labeled as depicting a developed or undeveloped scene. The (developed) fraction assigned to a 1x1 km tile is then simply the ratio of the images with the label developed to the total number of images in the tile. Different approaches for labelling the images are compared based on how well the image generated fraction maps match the ground truth fraction map. Different quantitative measures of similarity are considered. We compute the correlation coefficient between the tile fraction values taken as observations of random variables as as described in Section 2.1.1. We also compute the mean absolute difference (MAD) and the root mean squared difference (RMSD) between the ground truth and image generated tile fraction values.

The binary label (developed or undeveloped) assigned to a 1x1 km tile is determined by applying a threshold to the image generated fraction for that tile. The similarity between the ground truth and an image generated binary classification map is measured in two ways. First, the overall classification rate is computed as the percentage of tiles with the same label. We also compute the average classification rate of the two classes (developed and undeveloped).

We deliberately choose simple features to characterize the visual content of the images. We annotate the geo-referenced images using edge histogram descriptors which quantify the distribution of edges at different orientations. These are the same features used in the previous experiment as described in Section 2.1.2.

We use a support vector machine (SVM) classifier to label individual images based on their edge histogram descriptors. Given a labeled training set, an SVM classifier with a Gaussian radial basis function kernel is trained using five fold cross validation and grid search for optimal parameter selection. Once trained, the classifier is used to label a set of target images which in all cases is disjoint from the training set. These labels are

then used to generate fraction and binary classification maps which are compared with the ground truth maps. The framework of the this experiment is illustrated in Figure 2.16. Even though the experiments below consider different training and target sets, the ground truth comparison is always based on the 4,553 tiles for which there are both Flickr and Geograph images. 38.9% of these tiles are developed in the ground truth so that the chance overall binary classification rate is 61.1% achievable by labelling all images and therefore all tiles an undeveloped.

Manually Labeled Training Set

Here, the training set contains 2,740 Flickr images which have been manually labeled. A non-expert labeled an image as developed if it depicts a scene containing constructed materials such as used in houses, buildings, etc., and labeled it as undeveloped if it is of open areas and/or contains mostly trees and vegetation. These criteria will of course result in "incorrectly" labeled images such as indoor scenes always being labeled as "developed" even though they might have been taken inside isolated homes in rural regions. The SVM trained with the manually labeled training set is then used to classify a target set consisting of the remaining images from the 920K Flickr image set. The individual image labels are used to generate the fraction and binary classification maps shown in Figure 2.17. Notice the similarity between these maps and the ground truth maps in Figure 2.13.

Line 1 of table 2.3 shows the quantitative similarity between the ground truth and image generated maps. The columns titled Fixed Threshold indicate the agreement between the binary classification maps when a fixed threshold of 0.5 is applied to the fraction values of the image generated fraction map to generate the binary classification map.

Prior Information

The threshold used to derive the binary classification map can be adjusted so that the fraction of developed tiles matches that of the ground truth if such prior information is

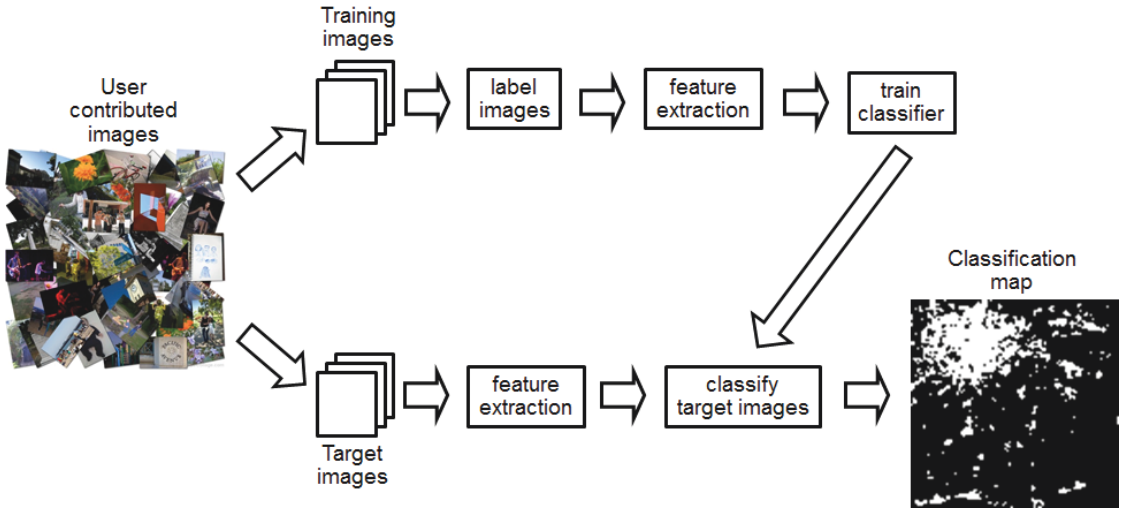
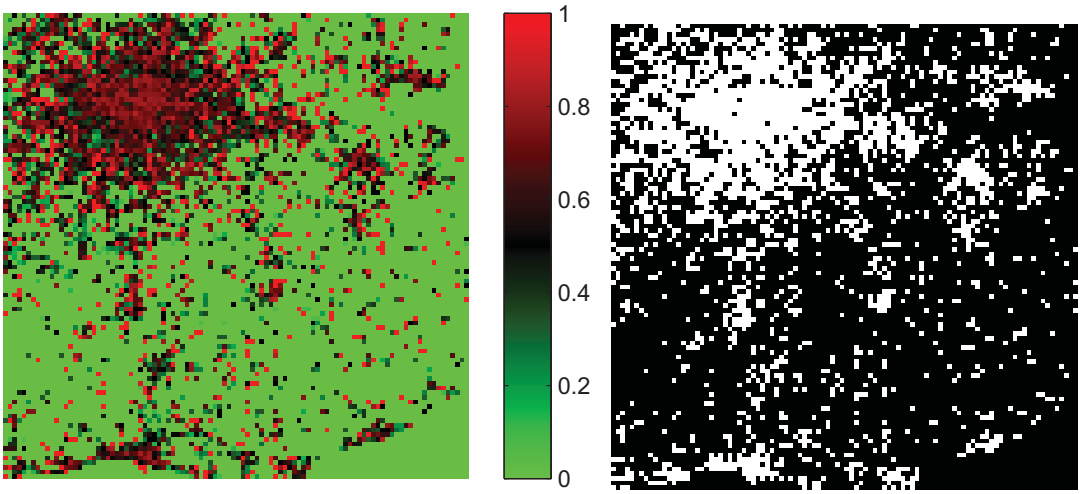


Figure 2.16: An overview of using the visual content of ground-level images to map developed and undeveloped regions.



(a) (b)

Figure 2.17: Land cover maps automatically generated using an SVM classifier trained with manually labeled Flickr images. The target set is also Flickr images. (a) Fraction map indicating the percent developed for each 1x1 km tile. (b) Binary classification map indicating the tiles labeled as developed (white) or undeveloped (black). Compare with the ground truth maps in Figure 2.13

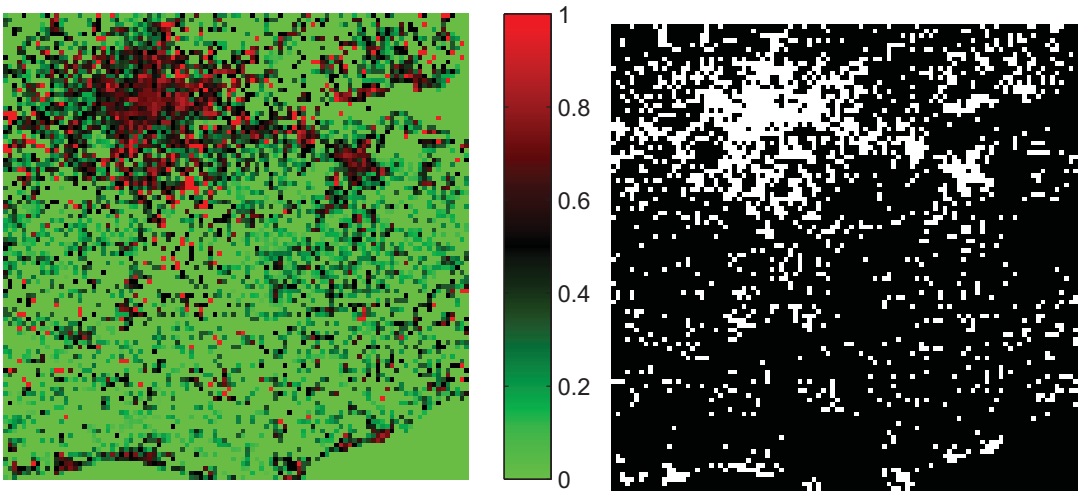
known. The columns titled Adaptive Threshold give the performance when the fraction of developed tiles in the image generated binary classification matches the 38.9% of the ground truth. For the manually labeled Flickr training set it results in decreased performance; subsequent experiments show that it can result in significant improvement.

Weakly-supervised Training

This experiment investigates the performance of a classifier trained in a weakly-supervised manner. The training set is constructed without any manual labelling by selecting two images at random from each 1x1 km tile and labelling them with the majority class of the tile. Selection is limited to tiles with four or more images so that at least two images remain in the disjoint target set. For the Flickr dataset, this results in a training set termed "Flickr-small" containing 5,872 images. Line 2 of table 2.3 shows the results for the Flickr-small classifier when applied to the Flickr dataset. The performance is shown to be better than that of the classifier trained with the manually labeled dataset, an interesting and significant result indicating that training sets can be generated from regions for which maps exists and then used to train classifiers for mapping unmapped regions. That the results are better than the manual case suggests that the automatically generated training set more accurately characterizes the differences between images from developed and undeveloped regions than the intuition humans use when labelling images.

Photographer Intent

This experiment investigates the effect photographer intent has on the image generated maps. The training and target sets are now selected from the Geograph dataset which contains images captured by photographers who intend their photographs to be geographically representative. We again train the classifier with a weakly-supervised dataset which now has 10,576 images as there are more tiles with four or more Geograph images than tiles with four or more Flickr images. Line 3 of table 2.3 shows the result of applying this Geograph-small classifier to a held out Geograph target set. The maps generated from the Geograph dataset are significantly better than those generated



(a) (b)

Figure 2.18: Land cover maps automatically generated using an SVM classifier trained with a large set of Geograph images labeled in a weakly-supervised manner. The target set is also Geograph images. (a) Fraction map indicating the percent developed for each 1x1 km tile. (b) Binary classification map indicating the tiles labeled as developed (white) or undeveloped (black). Compare with the ground truth maps in Figure 2.13

from the Flickr dataset indicating that photographer intent is a significant factor. We comment more on the implication of this result this in the Discussion section below.

Training Set Size

This experiment investigates the effect of the size of the weakly-supervised training set. We construct a 11,465 Flickr-large training set by selecting five images from tiles that contain more than ten Flickr images. Line 4 of table 2.3 shows the results of using this training set. The performance is worse than that of the Flickr-small training set likely because of a bias present in the smaller number of tiles with more than ten images. These tiles are more likely to be of developed regions as confirmed by the higher ratio of images labeled developed in the training set (indicated in parenthesis in the column titled Training Set Size in table 2.3). A Geograph-large training set is also constructed and shown to perform better than the Geograph-small set (see line 5 of the table). Figure 2.18 shows the maps generated using a classifier trained with the Geograph-large training set.

Training Set Quality

This experiment investigates the effect of the quality of the weakly-supervised training set. We now select training images from tiles that have very high or very low developed fractions according to the ground truth map. The intuition here is that such tiles should result in more accurate training sets. Lines 6 and 7 of table 2.3 show that these Flickr-good and Geograph-good training sets do not result in improved performance. This finding indicates that it is not necessary to constrain the weakly-supervised training sets in this way.

Relative Importance of Training and Target Sets

The results above clearly indicate that the Geograph dataset is more effective than the Flickr dataset. This experiment investigates whether this improvement is due to the training or target set. Lines 8 through 14 in table 2.3 list the results when the training and target sets are from different image collections. These results make it clear that photographer intent is more important for the target set than the training set. While this finding is somewhat unfortunate since the overall (worldwide) coverage of the Flickr dataset is broader than that of the Geograph dataset, it does identify some interesting research challenges which will be discussed in the Discussion section below.

Filtering Images With Faces

This experiment investigates whether removing images with faces improves the results. The motivation here is that photographs of people are less likely to be geographically informative, especially close-in portraits. The fact that few of the Geograph images contain people empirically suggests this is true. We used a standard face detection algorithm [VJ01, LM02] to filter Flickr images containing one or more faces. We then repeated a set of experiments using this face-free target set. Unfortunately, as lines 15 through 17 in table 2.3 show, this did not provide any improvement over the target set with faces.

Table 2.3: The experimental results. The number in parenthesis in the Training Set Size column indicates the fraction of images labeled as developed in the training set. Please see the text for other details.

				Binary Maps						
				Overall Class. Rate		Avg. Class. Rate				
	Training Set	Target Set	Training Set Size	Fixed Threshold %	Adaptive Threshold %	Fixed Threshold %	Adaptive Threshold %	Fraction Maps ρ	MAD	RMSD
1	Manual (Flickr)	Flickr	2740 (0.51)	66.4	64.9	68.8	63.0	0.374	0.287	0.383
2	Flickr small	Flickr	5872 (0.52)	67.2	66.9	68.7	65.2	0.380	0.279	0.373
3	Geograph small	Geograph	10576 (0.26)	68.2	74.0	60.8	72.6	0.520	0.271	0.358
4	Flickr large	Flickr	11465 (0.56)	57.7	61.5	64.0	59.6	0.372	0.336	0.441
5	Geograph large	Geograph	13374 (0.36)	73.8	**74.7**	70.2	**73.2**	**0.552**	0.235	0.313
6	Flickr good	Flickr	5070 (0.49)	67.0	68.1	67.4	66.6	0.329	0.285	0.374
7	Geograph good	Geograph	5603 (0.47)	**74.2**	74.6	**71.5**	73.1	0.551	**0.231**	**0.308**
8	Geograph small	Flickr	10576 (0.26)	60.0	72.3	49.8	70.9	0.230	0.354	0.457
9	Geograph large	Flickr	13374 (0.36)	60.0	68.7	51.4	66.9	0.301	0.312	0.404
10	Geograph good	Flickr	5603 (0.47)	60.7	68.3	53.8	66.6	0.330	0.294	0.381
11	Manual (Flickr)	Geograph	2740 (0.51)	66.1	73.5	70.1	72.0	0.531	0.273	0.356
12	Flickr small	Geograph	5872 (0.52)	67.8	74.1	70.8	72.3	0.526	0.264	0.345
13	Flickr large	Geograph	11465 (0.56)	56.3	72.6	63.3	71.2	0.486	0.340	0.428
14	Flickr good	Geograph	5070 (0.49)	69.9	73.1	**71.5**	71.7	0.496	0.2545	0.3310
15	Flickr small	Flickr (no faces)	5872 (0.52)	66.8	66.7	66.8	64.2	0.367	0.301	0.414
16	Geograph small	Flickr (no faces)	10576 (0.26)	59.8	72.2	49.0	69.7	0.225	0.377	0.493
17	Geograph good	Flickr (no faces)	5603 (0.47)	59.9	68.0	52.0	65.2	0.312	0.321	0.428

2.3 Aditional Features

Thus far, we have only considered simple edge histograms as the image feature to perform land cover classification. In this section, we study the effect of other features on classification performance. The features considered include color histogram, gist, and textual features derived from the annotations associates with the images. We follow an experimental setting similar to those mentioned above along with the same datasets and use edge histogram features as a baseline measurement.

Color histogram

In order to investigate whether color is a discriminating feature for our two-class problem, we extract color histogram descriptors from each image. We transform the images to the hue-lightness-saturation (HLS) colorspace and quantize each dimension into 4 bins for a total feature vector length of 64. The histograms are normalized to have an L1 norm of one to account for different image sizes.

Gist

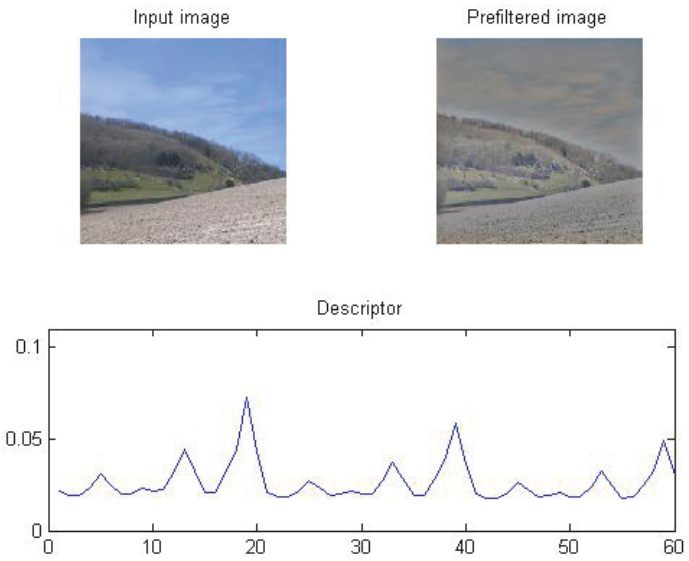
The final visual features we consider are gist descriptors described in Section 1.5.2. Gist features are similar to texture features extracted using Gabor filters [WMN00] in that they characterize the spectral energy of an image using Gaussian shaped filters tuned to different scales and orientations. A prefilter for normalizing the local contrast with respect to luminance variance is applied before gist features are extracted. A visualization of gist features extracted from two sample images obtained from the two land cover classes is shown in Figure 2.19. It is clear to see that the gist features are distinctive between the different land cover classes. We extract 60 dimensional gist feature vectors from each image.

The experiments below compare the performance of the three visual features which, to summarize, include: a 64 dimensional color histogram feature, a five dimensional edge histogram feature, and a 60 dimensional gist feature for each image.

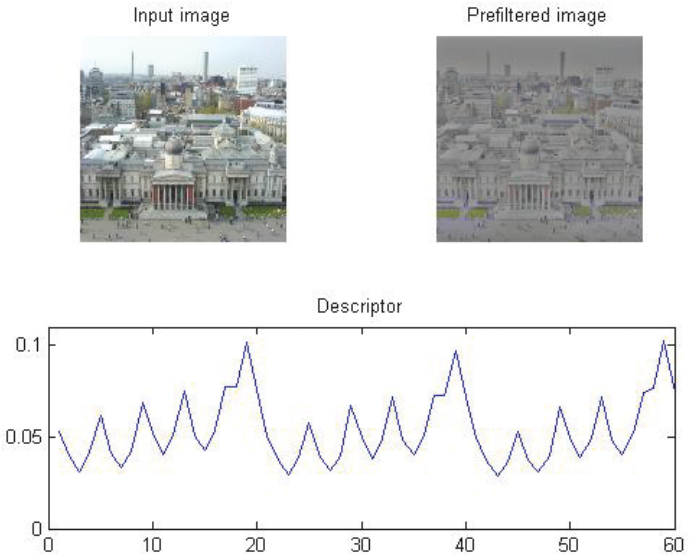
Text

Flickr and Geograph images commonly have user-supplied text associated with them. In the case of the Flickr images, this includes the image titles, descriptions, and tags. For example, the left-most image in Figure 2.15(a) is titled "Roosting dragon"; has the description "Or it might be a vampire bat? In Chancery Lane. Originally uploaded for Guess Where London."; and, is tagged with: "gwl, Guess Where London, stucco, dragon, Guessed by Citymuso, 115A, Chancery Lane, WC2, Holborn, Camden, London, England". The Geograph images have titles, descriptions, and categories. The left-most image in Figure 2.15(b) is titled "The Old Bailey, London"; has the comment "The Central Criminal Court, home of justice in England and Wales."; and is categorized as "Building of civic importance". We therefore investigate whether this user-supplied text is effective for land cover classification.

The text analysis is performed at the tile level since there is typically not enough text associated with the individual images for effective classification. Each of the text com-



(a)



(b)

Figure 2.19: Visualization of gist features of two sample images obtained from (a) an undeveloped region (b) a developed region. The plots indicate the responses of Gabor filters in 60 directions and scales.

ponents associated with an image obtained within each 1x1 km tile region is parsed into a set of terms (words) which are then pooled among terms from other images within the same tile. At the moment, all terms are given equal weight although different weightings based on the relative importance of the components would be an interesting extension.

It is unlikely that classification at the term level would be effective due to the sparse appearance of terms among the dictionary, so we apply a latent semantic approach from text document analysis in which a hidden topic $z \in Z = \{z_1, ..., z_K\}$ is associated with the observed occurrence of a word $w \in W = \{w_1, ..., w_M\}$ in a document (tile) $d \in D = \{d_1, ..., d_N\}$. This latent layer also helps overcome the problems of synonymy and polysemy.

We use a generative probabilistic technique termed probabilistic latent semantic analysis (pLSA) [Hof99, Hof01] to learn the hidden topics. A pLSA model can be expressed as

$$P(w|d) = \sum_{z \in Z} P(w|z)P(z|d),$$

where $P(w|d)$ is the observed word distributions over documents.

To learn the distribution of words over hidden topics, we use Expectation Maximization (EM) algorithm. In E-step, the posterior probabilities for the hidden topics are evaluated:

$$P(z|d, w) = \frac{P(z)P(d|z)P(w|z)}{\sum_{z'} P(z')P(d|z')P(w|z')},$$

while in M-step the parameters of E-step are estimated based on the result of E-steps:

$$P(w|z) = \frac{\sum_d n(d, w)P(z|d, w)}{\sum_{d,w'} n(d, w')P(z|d, w')},$$

$$P(d|z) = \frac{\sum_w n(d, w)P(z|d, w)}{\sum_{d',w} n(d', w)P(z|d', w)},$$

$$P(z) = \frac{1}{R} \sum_{d,w} n(d, w)P(z|d, w), R \equiv \sum_{d,w} n(d, w).$$

Instead of defining the number of hidden topics as the number of ground truth classes as

most classifications using pLSA approach usually do, we use pLSA as a tool to reduce the dimensionality of the term histogram of each tile by computing the distributions of hidden topics over the tiles, $P(z|d)$. In our experiment, we show that distributions of hidden topics provide an explicit representation of the tiles that is more robust than distributions over terms. To evaluate the hidden topic distribution of a novel tile, the EM algorithm is applied with fixed $P(w|z)$ learned from the training step.

We first determine reasonably sized term-dictionaries for each of the datasets. After applying stopping and stemming, a total of 106,213 unique terms result from the over 900K Flickr images in the 4,441 tiles with Flickr and Geograph images that have text, and 31,056 unique terms result from the over 90K Geograph images from the same tiles. The dictionary for the Flickr dataset is selected as the 2,708 most frequent Flickr terms, and the dictionary for the Geograph dataset is selected as the 2,702 most frequent Geograph terms.

A term histogram is computed for each tile based on the terms from all the images in the tile. The histograms for 200 training tiles are combined into a term-document matrix and pLSA is used to learn the term-topic distributions for a 60 topic model (this number was chosen empirically based on performance). Finally, a topic distribution is computed for each of the 4,041 tiles in the test set using the pLSA machinery.

To summarize, each tile is represented with a 60 dimensional text feature vector that consists of the distribution over the latent topics.

2.3.1 Experiments

The goal of the experiments is to use the geo-referenced images as represented by their visual or text features to produce developed/undeveloped land cover maps. We formulate this as a supervised classification problem in which support vector machines (SVMs) are trained on a labeled subset of the data and then used to assign labels to a disjoint held-out set. We compare applying the SVMs 1) at the image level, in which case the image labels (developed and undeveloped) are aggregated to produce the final

tile level fraction values and binary labels, and 2) applying them directly at the tile level. These two modes are described in sections 2.3.1 and 2.3.1 below. Performance is evaluated by comparing the predicted maps to the ground truth maps derived from the Land Cover Map 2000.

The SVMs are implemented using the LIBSVM package [CL01]. We use radial basis function (RBF) kernels and determine optimal values for the two parameters, the penalty term C and the kernel width γ, through grid-search on a random partitioning of the training set.

Image Level Classification

In this set of experiments, the SVMs are used to classify individual images as being developed or undeveloped. These labels are then aggregated to determine the tile fraction values and binary labels for comparison with the ground truth.

Training the SVMs requires a set of images labeled as developed or undeveloped. We construct a weakly labeled training set by propagating the tile labels to the images as follows. First, we identify the 100 most developed and the 300 least developed tiles according to the ground truth fraction values. These are training tiles. We use 300 least developed tiles because the least developed tiles generally contain fewer images than the most developed tiles. We then randomly sample approximately 2,500 images from the 100 most developed tiles and label them as developed. We similarly sample and label as undeveloped approximately 2,500 images from the 300 least developed tiles. This results in labeled training sets containing 5,031 and 5,026 images for the Flickr and Geograph datasets respectively.

Such a weakly labeled training set has two important advantages over a manually labeled one. First, it requires very little human effort. Second, it avoids the subjective interpretation of what is meant by developed at the image level. Indeed, we showed in the previous experiments above that a weakly labeled training set outperforms one in which the labels are assigned manually in a binary land cover classification problem. We now select the labeled images from the 100 most and least developed tiles.

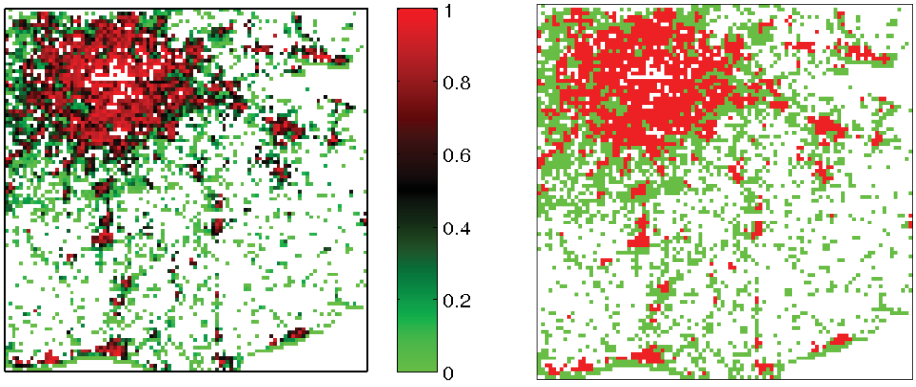
The SVMs are trained using the visual features and labels of the 5K+ training images. They are then used to label each of the images in the 4,041 test tiles that remain after the 400 training tiles have been removed (thus the training and test sets are distinct). These labels are aggregated to compute two tile level values: a fraction developed which is simply the fraction of images in the tile labeled as developed by the SVM; and a binary label which is determined by applying a threshold to the fraction developed. We explore using a threshold fixed at 0.5 as well as an adaptive threshold that is chosen so that the ratio of developed to undeveloped tiles in the predicted set matches that of the ground truth (this ratio thus represents prior knowledge of the problem).

Tile Level Classification

In this set of experiments, the SVMs are used to label the tiles directly. A single visual feature is computed for each tile by averaging the features from all the images located in that tile. This has the simple interpretation of a tile level histogram for the edge and color features. For the gist features, it is the average over all the images of the spectral energy in each of the frequency channels corresponding to the Gabor filters. The text features are already computed at the tile level so no aggregation is needed.

The training set is the 100 features (visual or text) corresponding to the 100 most developed tiles and the 100 features corresponding to the 100 least developed tiles. These 200 training tiles are a subset of the 400 training tiles used in the image level classification above.

Once trained, the SVMs are used to label each of the 4,041 test tiles as developed or undeveloped again using a single feature. Note that this results in only a binary label for each tile; the fraction developed value is not estimated when the SVM labelling is done at the tile level (we have not yet considered using the classifier margin for this).



(a) (b)

Figure 2.20: Ground truth data for the 4,041 tiles in the test set. (a) Fraction map indicating the percent developed for each 1x1 km tile. (b) Binary map indicating the tiles labeled as developed (red) or undeveloped (green).

2.3.2 Results

The results are evaluated by comparing the predicted fraction values and binary labels to that of the ground truth for the 4,041 tiles in the test set. The ground truth fraction values and binary labels for the test set are shown in Figure 2.20.

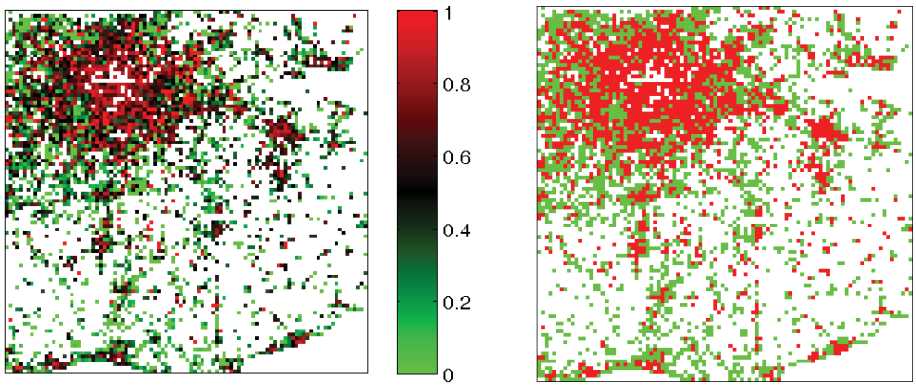
Similar to the previous experiments above, the predicted fraction values are evaluated based on their correlation with the ground truth values. The binary labels are evaluated using classification rates. The overall classification rate is the percentage of tiles assigned the same label–developed or undeveloped–as the ground truth.

As mentioned earlier, when the SVM labelling is performed at the image level, the tile labels are determined by applying either a fixed or an adaptive threshold to the aggregated image labels. In the fixed case, a tile is labeled as developed if the fraction value is greater than 0.5 (i.e., more than 50% of its images are labeled as developed). The adaptive threshold is chosen (through brute force) so that the resulting ratio of developed to undeveloped tiles matches that of the ground truth. This prior knowledge helps compensate for systematic biases in the predicted image labels.

Since 2,386 of the 4,041 test tiles are undeveloped in the ground truth, labelling all images or all tiles as undeveloped results in a “chance” classification rate of 59.0%.

Table 2.4: The results of the image level classification.

Dataset	Visual Feature	Binary Prediction Overall Class. Rate Fixed Threshold %	Binary Prediction Overall Class. Rate Adaptive Threshold %	Fraction Prediction ρ
Geograph	Color	70.9	73.0	0.519
Geograph	Edge	70.7	73.1	0.528
Geograph	Gist	**75.0**	**75.0**	**0.614**
Flickr	Color	63.5	64.3	0.317
Flickr	Edge	65.6	65.1	0.376
Flickr	Gist	68.8	69.2	0.425



(a) (b)

Figure 2.21: The predicted fraction values (a) and binary labels (b) that result from using gist features to classify Geograph images as developed or undeveloped. Compare with the ground truth in Figure 2.20.

Results of the Image Level Classification

Table 2.4 summarizes the results when the SVM classification is performed at the image level. The predicted fraction values and binary labels for the best case corresponding to Geograph images classified using gist features are shown visually in Figure 2.21. Compare this with the ground truth values and labels in Figure 2.20.

Based on these results, we conclude the following about the case when the classifiers are applied at the image level:

- The Geograph dataset outperforms the Flickr dataset.

Table 2.5: The results of the tile level classification.

Dataset	Feature	Binary Prediction
		Overall Class. Rate (%)
Geograph	Color	68.8
Geograph	Edge	72.2
Geograph	Gist	74.0
Geograph	Text	74.2
Flickr	Color	70.5
Flickr	Edge	69.7
Flickr	Gist	68.0
Flickr	Text	49.4

- The gist features perform best overall. The edge histogram features perform better than the color histogram features. This ordering is true for both datasets.
- The adaptive threshold improves the overall classification rate in most cases.

Results of the Tile Level Classification

Table 2.5 summarizes the results when the SVM classification is performed at the tile level. Based on these results, we conclude the following about this case:

- The Geograph dataset outperforms the Flickr dataset except when using the color histogram features.
- The relative performance of the visual features depends on the dataset. For the Geograph dataset, the ordering is the same as for the image level classification with the gist features performing best overall followed by the edge histogram features. This ordering is reversed for the Flickr dataset.
- The text features perform better than the visual features for the Geograph dataset but much worse for the Flickr dataset.

When comparing the tile level to the image level classification, we conclude:

- Aggregating the color and edge histogram features at the tile level results in worse performance for the Geograph dataset but significantly better performance for the

Table 2.6: Detailed classification results. The first column indicates the dataset. The second column indicates whether the classification is performed at the image or tile level. The third column indicates the feature. The fourth column indicates whether a fixed or adaptive threshold is used to derive the binary label from the fraction value. Columns five through eight indicate the true-positive, true-negative, false-positive, and false-negative rates in terms of the number number of tiles and the percentage. The test dataset contains 4,041 tiles of which 1,655 have positive labels (labeled as developed in the ground truth).

Data.	Lev.	Fea.	Thresh.	TP	TN	FP	FN
Geo.	Im.	Col.	Fix.	1235 (74.6%)	1632 (68.4%)	754 (31.6%)	420 (25.4%)
Geo.	Im.	Col.	Ad.	1089 (65.8%)	1860 (78.0%)	526 (22.0%)	566 (34.2%)
Geo.	Im.	Edge	Fix.	1310 (79.2%)	1549 (64.9%)	837 (35.1%)	345 (20.8%)
Geo.	Im.	Edge	Ad.	1094 (66.1%)	1859 (77.9%)	527 (22.1%)	561 (33.9%)
Geo.	Im.	Gist	Fix.	1228 (74.2%)	1804 (75.6%)	582 (24.4%)	427 (25.8%)
Geo.	Im.	Gist	Ad.	1228 (74.2%)	1804 (75.6%)	582 (24.4%)	427 (25.8%)
Fli.	Im.	Col.	Fix.	1277 (77.2%)	1291 (54.1%)	1095 (45.9%)	378 (22.8%)
Fli.	Im.	Col.	Ad.	933 (56.4%)	1664 (69.7%)	722 (30.3%)	722 (43.6%)
Fli.	Im.	Edge	Fix.	1358 (82.1%)	1292 (54.1%)	1094 (45.9%)	297 (17.9%)
Fli.	Im.	Edge	Ad.	976 (59.0%)	1656 (69.4%)	730 (30.6%)	679 (41.0%)
Fli.	Im.	Gist	Fix.	1258 (76.0%)	1524 (63.9%)	862 (36.1%)	397 (24.0%)
Fli.	Im.	Gist	Ad.	1031 (62.3%)	1764 (73.9%)	622 (26.1%)	624 (37.7%)
Geo.	Tile	Col.	NA	1305 (78.9%)	1475 (61.8%)	911 (38.2%)	350 (21.1%)
Geo.	Tile	Edge	NA	1302 (78.7%)	1614 (67.6%)	772 (32.4%)	353 (21.3%)
Geo.	Tile	Gist	NA	1208 (73.0%)	1784 (74.7%)	602 (25.2%)	447 (27.0%)
Geo.	Tile	Text	NA	1061 (64.1%)	1936 (81.1%)	450 (18.9%)	594 (35.9%)
Fli.	Tile	Col.	NA	1055 (63.7%)	1794 (75.2%)	592 (24.5%)	600 (36.3%)
Fli.	Tile	Edge	NA	1026 (62.0%)	1792 (75.1%)	594 (24.9%)	629 (28.0%)
Fli.	Tile	Gist	NA	1228 (74.2%)	1518 (63.6%)	868 (36.4%)	427 (25.8%)
Fli.	Tile	Text	NA	1012 (61.1%)	986 (41.3%)	1400 (58.7%)	643 (38.9%)

Flickr dataset.

- Aggregating the gist features at the tile level results in reduced performance for both datasets.

Detailed Results

Finally, table 2.6 presents detailed results in the form of the true-positive, true-negative, false-positive, and false-negative counts and rates for each of the experimental configurations. We make the following conclusions based on these results:

- The image level SVM classifiers appear to overclassify the images as being developed. This bias is evident in the fact that the threshold applied to the fraction values in order for the the ratio of the developed to undeveloped tiles in the predicted set to match that of the ground truth is always greater than 0.5. The optimal threshold (not shown) ranges from 0.51 for the Geograph-color case to 0.66 for the Flickr-edge case. The one exception is the Geograph-gist case for which a threshold of 0.5 is optimal.

- The edge histogram features result in the highest image level overclassification, followed by the color histogram features.

- The tile level SVM classifiers have mixed biases. With regard to the visual features, the Geograph-color, Geograph-edge, and Flickr-gist appear to be biased towards the developed class in that they have higher false-positive than false-negative rates. The Geograph-gist, Flickr-color, and Flickr-edge are biased towards the undeveloped class. The text features are heavily biased towards the undeveloped class for the Geograph dataset but heavily biased towards the developed class for the Flickr dataset.

2.3.3 Discussion

The experiments above demonstrate that large collections of geo-referenced community contributed photos can be used to derive maps of what-is-where on the surface of the Earth. Binary land cover maps produced in an automated fashion using the visual and text features of the images were shown to be qualitatively and quantitatively similar to a ground truth dataset. We now discuss further insights into the proposed framework.

We showed that image level classifiers could be learned in a weakly supervised manner by propagating the tile level labels to individual images located in those tiles. This has clear benefits in terms of the manual effort required to train the classifiers. While our training and test datasets are disjoint, they come from the same 100×100 km region. We plan to explore how well classifiers trained on one region generalize to other regions

especially for which land cover maps are not available or are out-of-date.

The Geograph dataset was shown to outperform the Flickr dataset demonstrating that photographer intent is important for treating community contributed photos as VGI. This is not an unexpected result since the Geograph images are contributed by users whose goal is to collect geographically representative photographs. Since non-specialized collections such as Flickr have better coverage, this finding raises the interesting research question of how to use one dataset to improve another. Specifically, can the Geograph dataset be used to derive filters or other mechanisms for improving the Flickr dataset. We envision both top-down approaches such as removing images with faces which are generally less geographically informative, as well as bottom-up approaches based on the statistics of low-level image features.

The gist features were shown to outperform the color and edge histogram features for classifying individual images as developed or undeveloped. This is in agreement with other studies on using gist features for scene classification. The visual features are complementary so combining them should result in improved performance.

While aggregating gist features at the tile level proved to be ineffective, using tile level color and edge histograms was surprisingly effective especially for the Flickr dataset in which the tile level labeling significantly outperforms the image level labeling. This is interesting because it indicates that the aggregation helps remove the noise in non-specialized datasets such as Flickr. We plan to explore richer representations of the aggregate features such as Gaussian mixture or non-parametric kernel density models.

The text based analysis was shown to be effective for the Geograph but not the Flickr dataset. This shows that photographer intent, here in terms of how individual images are described and tagged, seems to have even more of an effect on the text associated with the images than their visual content. Put differently, the fact that visual aspect of the photo collections appears to be less affected by contributor intent suggests that it is preferable over textual forms as a source of VGI.

2.4 Removal of Geographically Uninformative Images

From the previous experiments, we show that learning models created using the Flickr collection perform worse than the models created using the Geograph collection in land cover classification. It is clear that not all images in Flickr are geographically informative because users of Flickr do not necessary share photos with the same intention as Geograph's users do. Our limited experimental results discussed in Section 2.2.2 suggest that removing photos containing human faces does not improve the overall classification results.

In this section, we focus on improving the fidelity of geographic information extracted from a collection of Flickr photos, which will be used to train a land cover classifier. When constructing the training samples for the classification, photos that are not geographically informative are automatically filtered out based on the physical properties of the cameras recorded in their metadata. These properties include the model of the cameras and the use of camera flash.

To be able to analyze a photo with visual features, the image quality of the photo must be good. We assume that most photos taken by stand-alone cameras will produce better quality than those taken by cameras in most mobile phones due to the less advanced camera components used in mobile phones. As a result, we remove any photos taken by cameras in mobile phones from the training data.

We also try to avoid photos taken indoors because they contain less information in regard to whether the locations of where they are taken belong to developed or undeveloped regions. Since most cameras fire their flash when taking indoor scenes, we remove any photos taken when the camera flash is on from the training data. Even though cameras may fire their flash when taking outdoor scenes, these photos usually capture close-by objects under low-light conditions. Therefore, this type of outdoor photo is not as informative and thus can be discarded as well.

Table 2.7: Image level classification results on Flickr dataset using edge histogram features.

Training Set	Overall Class. Rate (%)
Flickr	65.6
Flickr Camera	65.7
Flickr Flash off	66.1
Flickr Flash on	64.8

2.4.1 Emperiments

To extract the camera properties, we use Flickr API to obtain the EXIF metadata of each photo in the training set. We construct the *Camera* training set where photos taken by cameras in mobile phones are removed based on the camera models listed in the metadata. We also construct the *Flash off* training set in a similar fashion.

We follow the same experimental setup described in Section 2.3.1 to evaluate the quality of the training data. Edge histograms are used as the image features. We expect the results will stay true to other image and textual features discussed in this chapter. Classifiers are trained according to the refined sets mentioned above. They are then applied to the original whole test set.

From table 2.7, we see that removing photos taken by cameras in mobile phones or taken when camera flash is off results in better classification performances. To validate our hypothesis on camera flash usage, we also construct a training set (*Flash on*) containing only photos taken when camera flash is fired. We can see that this training data performs worse than the original training data.

Although the performance gain from removing geographic uninformative information from the training data is subtle, our experiments illustrate that, with a more detailed analysis, metadata from the photo collections can provide information that is useful to the applications of proximate sensing.

2.5 Summary

In this chapter, we propose a framework of using geo-referenced on-line photo collections to perform a binary land cover classification into developed and undeveloped regions as an example of proximate sensing. By comparing different sources of photo collections as well as different features extracted from these photos and their metadata, we have demonstrated that it is possible although challenging to extract geographic information from these on-line photo collections. This sets the course of our next investigation into using proximate sensing in land use classification, which will be discussed in the following chapter.

Chapter 3

Land Use Classification

In traditional remote sensing, overhead images are used to distinguish between different types of land cover such as vegetation, structures, or other features that cover the land; however, it much more difficult to determine the type of land use a certain land cover class belongs to. For example it is easy to locate a region with large buildings and parking lots in the satellite view mode in Google Maps, but it is much more challenging to use the satellite view to determine whether the region belongs to a shopping center or a warehouse. To find out the answer, one can switch to the Street View mode and see the images of nearby objects and scenes taken from the ground level. Although we do not use images from Google's StreetView in our experiments due to the limited visual perspectives of the scenes the images capture, they serve as a good example of how ground-level images can assist us in determining the land use of a location.

The Standard Land Use Coding Manual [Sta65] of the Urban Renewal Administration in the US Department of Commerce defines the following eight top-level land use classes: residential; manufacturing; transportation, communications, and utilities; trade; services; cultural, entertainment, and recreational; resource production and extraction; and undeveloped land and water areas. While some of these classes might be recognizable using remote sensed imagery, their subclasses are much more difficult. Trade is partitioned into several subclasses including building materials, hardware, and farm equipment; food; automotive; apparel and accessories; furniture; and eating and drink-

ing. Services is partitioned into finance, insurance, and real estate; personal; repair; professional (which is further partitioned into medical, dental, etc.); governmental; and educational.

In this chapter, we use proximate sensing to establish a framework that uses ground-level images of nearby objects and scenes to perform land use classification. We begin with an experiment of classifying land use on two university campuses. SIFT features are extracted from the photos collected according to the locations on the campuses, and they are used to train classifiers to identify three types of land use: academic buildings, residential buildings, and sports facilities. A land use map is created according to the classification results from each campus.

We then extend our experiments to a larger dataset with eight urban land use classes. The scope of this experiment is to highlight the benefit of using proximate sensing in land use classification. Eight classifiers are trained using gist features extracted from training images collected from the whole Great Britain area. These classifiers are then applied to photos collected from the TQ square region, and we evaluate their classification performance as well as create land use maps according to the results.

A portion of the work presented in this chapter was published as a peer-reviewed workshop paper at the Workshop on Geotagging and Its Applications in Multimedia in 2012 [LN12b]. The more recent work on eight urban classes is being prepared for conference submission.

3.1 Land use Classification in University Campuses

In this section, a case study is conducted to validate the concept of applying proximate sensing in land use classification on university campuses. We focus on land use classification on university campuses for three reasons: 1) these regions exhibit a variety of land use classes in a compact geographic region; 2) ground-truth maps can be derived from campus maps; and 3) there tends to be excellent social multimedia coverage due to the technical savvy of students, educators, and researchers.

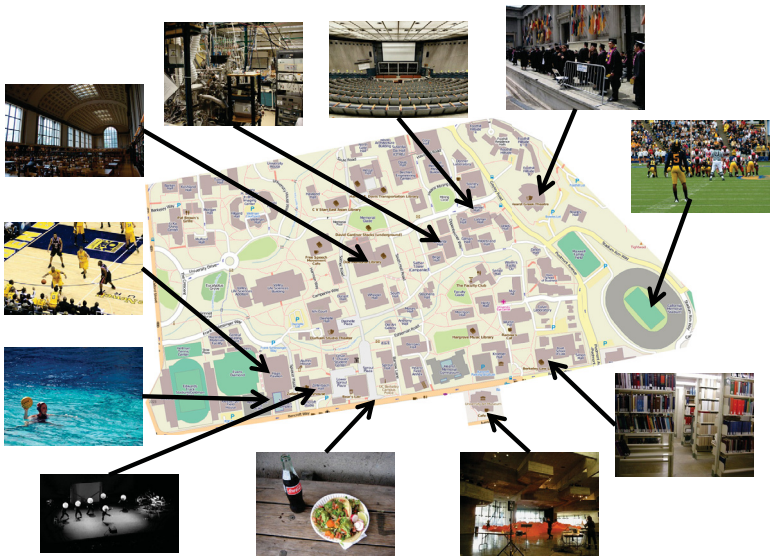


Figure 3.1: Sample Flickr images for the University of California, Berkeley campus. These are the actual locations of the images. These images clearly provide evidence on how different parts of the campus are used.

Consider the set of Flickr images geolocated on the University of California, Berkeley campus in Figure 3.1. These images clearly provide evidence on how different parts of the campus are used. While the content of these images could be used to identify a wide range of land uses such as libraries, classrooms, different kinds of sports facilities, laboratories, office space, entertainment venues, etc., we first focus on labeling regions as belonging to one of the three coarse classes: academic, residential, and sports.

The novel contribution of this work is to use proximate sensing to complement the shortcoming of remote sensing for land use classification.

3.1.1 Dataset

Two university campuses (University of California, Berkeley and Stanford University) are selected as our study areas to learn the land use classification models. We use the Flickr application programming interface (API) to download Flickr images located within the campus regions. For each campus, a land use map is derived manually according to its campus map. Three land use classes are considered: academic buildings,

residential buildings, and sports facilities. Each downloaded image is then assigned a land use class label according to its geographic location on the map. Figure 3.2 shows sample images from each class.

Table 3.1: Datasets Used in Visual Image Level Classification

	Academic	**Sports**	**Residence**	**Total**
Berkeley Training	5029	2153	463	7645
Berkeley Test	2000	1500	50	3550
Stanford Training	1524	2772	747	5043
Stanford Test	200	200	100	500

Table 3.2: Datasets Used in Visual Group Level Classification

	Academic	**Sports**	**Residence**	**Total**
Berkeley Training	1517	365	122	2004
Berkeley Test	200	50	30	280
Stanford Training	504	204	186	894
Stanford Test	50	30	30	110

Table 3.3: Datasets Used in Textual Group Level Classification

	Academic	**Sports**	**Residence**	**Total**
Berkeley Training	1425	348	123	1896
Berkeley Test	150	30	20	200
Stanford Training	421	193	141	755
Stanford Test	50	20	20	90

Besides training the classifiers at the image level, we also train the classifiers at the group of images level. Since the content of user contributed photos as well as the distribution of user contributions are very diverse, many photos contributed by the same user may bias the training data. As a result, we split the dataset into groups based on users (owners of photos), geographic locations, and the time when the photos are taken. For each campus, we first split all images into 20 sub-regions based on their geographic locations using k-means clustering. These sub-regions are independent from the land use classes. Finally, within each sub-region we group all the images taken by the same user on the same day. Our grouping methodology is based on the assumption that same user takes photos of similar scenes in a nearby location within a short period of time. Tables 3.1-3.3 provides the details of the two campus datasets.

(a) Sample Academic Images

(b) Sample Sports Images

(c) Sample Residence Images

Figure 3.2: Sample images from the university land use datasets.

3.1.2 Features

Bag of Visual Words

We extract a bag of visual words (BOVW) feature with soft-weighing scheme [JNY07] from each image. BOVW builds upon the local invariant features described in Section 1.5.1 by quantizing the features into visual words and then summarizing their distribution in an image using a histogram. Instead of assigning a visual word nearest to a keypoint detected, the soft-weighing scheme assigns 4 nearest visual words to a detected keypoint. A dictionary of 500 visual words is used in our implementation.

Text

Since Flickr images commonly have user-supplied text associated with them, we also study the effectiveness of this text for land use classification. To obtain the text features, we create a dictionary of terms based on the words extracted from the image title, descriptions, and tags associated with each image. After applying stopping and stemming, a total of 2457 unique terms are recorded, and out of these, the 1949 most frequent terms

are selected as the dictionary.

The text analysis is performed at the group level since there is typically not enough text associated with the individual images for effective classification. We follow the same pLSA scheme described in Chapter 2 to extract the text features.

3.1.3 Experiments

The goal of the experiments is to use the geo-referenced on-line photos as represented by their visual or text features to perform land use classification. We formulate this as a supervised classification problem in which support vector machines (SVMs) are trained on a labeled subset of the data and then used to assign labels to a disjoint held-out set. We compare applying the SVMs 1) at the image level and 2) applying them at the group level. These two modes are described in the following subsections. Performance is evaluated at two levels. First at the image or group level, and the second by comparing the predicted land use maps to the ground truth maps derived manually from the campus maps. The workflow of the experiments is illustrated in Figure 3.3.

The SVMs are implemented using radial basis function (RBF) kernels. We determine optimal values for the two parameters, the penalty term and the kernel width, through grid-search on a random partitioning of the training set.

Image Level Classification

In this set of experiments, the SVMs are used to classify individual images as being academic, sports, or residential. We use a one-versus-all approach to train the SVMs for each campus and class. To generate a predicted land use map, we first divide each campus into a map of 50x50 regions (tiles). The trained SVMs are then used to label each of the test images within each tile. As a result, each tile is represented by three ratios of images being classified into the three respective classes–e.g., academic versus other. We use the label of the highest ratio to assign a land use label to each tile for comparison with the ground truth maps.

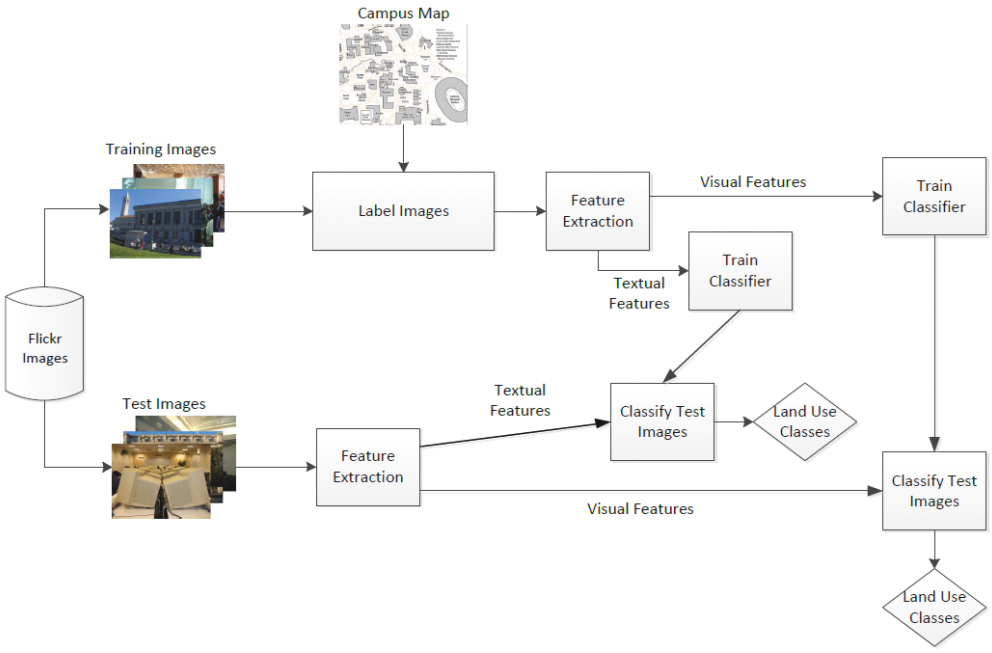


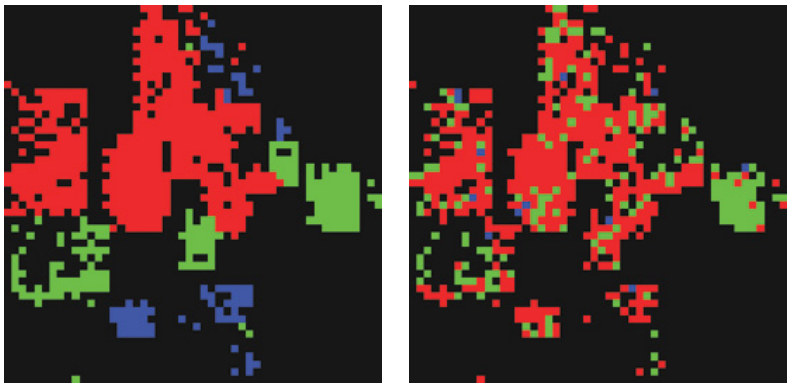
Figure 3.3: Framework of the proposed approach.

Table 3.4: Visual image level classification accuracy

	Berkeley Test Sets			Stanford Test Sets		
Training Sets	**Academic**	**Sports**	**Residential**	**Academic**	**Sports**	**Residential**
Berkeley Academic	**82**	17	36	**62**	27	39
Berkeley Sports	18	**84**	68	42	**72**	65
Berkeley Residential	44	57	**97**	59	59	**80**
Stanford Academic	**64**	36	69	**75**	31	58
Stanford Sports	28	**73**	54	28	**85**	44
Stanford Residential	44	57	**96**	55	54	**84**

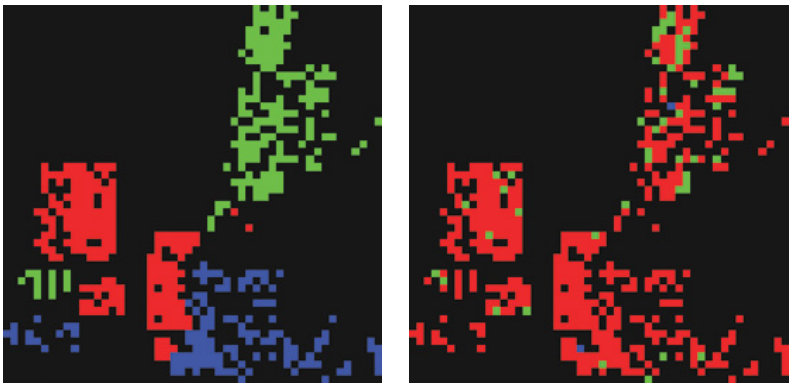
Group Level Classification

In this set of experiments, the SVMs are used to label the groups directly. A single visual feature is computed for each group by averaging the features from all the images within the group. The text features are already computed at the group level so no aggregation is needed. Due to the fact that not all images have accompanying text, the size of the training sets for the text features is slightly reduced from that of the image features.



(a) (b)

Figure 3.4: Land use classification of the Berkeley campus. (a) Ground truth map. (b) Predicted map using classifiers trained on the Stanford dataset. Academic, sports, and residential are denoted by red, green, and blue.



(a) (b)

Figure 3.5: Land use classification of the Stanford campus. (a) Ground truth map. (b) Predicted map using classifiers trained on the Berkeley dataset. Academic, Sports, and Residential are denoted by red, green, and blue.

Table 3.5: Visual group level classification accuracy

	Berkeley Test Sets			**Stanford Test Sets**		
Training Sets	**Academic**	**Sports**	**Residential**	**Academic**	**Sports**	**Residential**
Berkeley Academic	**74**	19	18	**58**	24	33
Berkeley Sports	25	**86**	84	46	**78**	67
Berkeley Residential	28	82	**90**	55	73	**73**
Stanford Academic	**60**	36	39	**68**	30	45
Stanford Sports	29	**81**	76	40	**84**	60
Stanford Residential	29	82	**89**	55	73	**73**

Table 3.6: Textual group level classification accuracy

	Berkeley Test Sets			Stanford Test Sets		
Training Sets	**Academic**	**Sports**	**Residential**	**Academic**	**Sports**	**Residential**
Berkeley Academic	**80**	12	16	**66**	21	28
Berkeley Sports	21	**88**	85	39	**81**	70
Berkeley Residential	25	85	**90**	44	78	**78**
Stanford Academic	**66**	33	37	**72**	23	46
Stanford Sports	22	**87**	82	37	**83**	70
Stanford Residential	25	85	**90**	44	78	**78**

Table 3.7: Precision and recall rates

		Visual Image		Visual Group		Textual Group	
Training Sets	**Test Sets**	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**
Berkeley Academic	Berkeley Academic	0.80	0.92	0.76	0.94	0.80	0.98
Berkeley Sports	Berkeley Sports	0.92	0.67	0.81	0.26	0.75	0.30
Berkeley Residential	Berkeley Residential	0.13	0.14	1.0	0.03	0	0
Berkeley Academic	Stanford Academic	0.52	0.93	0.52	0.98	0.62	0.96
Berkeley Sports	Stanford Sports	0.79	0.41	0.80	0.27	0.67	0.30
Berkeley Residential	Stanford Residential	0.50	0.05	0	0	0	0
Stanford Academic	Berkeley Academic	0.83	0.46	0.74	0.68	0.82	0.70
Stanford Sports	Berkeley Sports	0.67	0.71	0.48	0.38	0.61	0.37
Stanford Residential	Berkeley Residential	0.03	0.06	0	0	0	0
Stanford Academic	Stanford Academic	0.72	0.63	0.62	0.78	0.74	0.78
Stanford Sports	Stanford Sports	0.79	0.85	0.83	0.50	0.78	0.38
Stanford Residential	Stanford Residential	0.84	0.25	0	0	0	0

3.1.4 Results

The different approaches are evaluated based on their accuracy, precision, and recall. Accuracy is the number of correctly predicted labels (both positive and negative) for a particular classifier normalized by the size of the particular test set. It is reported as a percentage. For example, if the binary classifier trained on the Berkeley Academic dataset classifies 320 of the 500 images in the Stanford test set correctly then the accuracy is 60%. An "accuracy" value can also be computed when a classifier is used to detect a class other than the one it is trained for. For example, the binary classifier trained on the Berkeley Academic dataset can be applied to the Stanford test with the sports images as the positive labels. In this case, a low accuracy value would be a good result. Precision is the fraction of images that are assigned a particular class that actually have that class and recall is the fraction of the images with a particular class that are assigned that class.

Table 3.4 summarizes the visual image level classification accuracy of each classifier trained using one class and applied to detect another class for both the intra- and inter-campus cases. These results clearly demonstrate that the classifiers are learning discriminating visual features for the three different land use classes. Classifiers trained on a particular class are always more likely to detect that class than another. As might be expected, the intra-campus results are better than the inter-campus ones. However, the approach is seen to generalize quite well from one campus to another. The high values for the residential class can be explained by the relatively few images in this test set especially for the Berkeley campus.

Table 3.5 summarizes the visual group level accuracy. The classification performance at the visual group level is comparable to that at the visual image level. This is significant since grouping the images results in smaller number of training samples, greatly reducing the computational cost of the SVM learning.

Table 3.6 summarizes the textual group level classification accuracy. We can see that despite the diversity of text accompanying the images, pLSA is able to extract sufficient discriminating semantic information to distinguish the classes.

Table 3.7 summarizes the precision and recall rates for each approach. These results corroborate those of the accuracy results: the classifiers are able to distinguish between different classes; the intra-campus results are better than the inter-campus but the generalization is still good; and that the visual group and textual group results are comparable to that of the visual image. The poor precision and recall values for the Berkeley residential class are again a result of there being too few images in this dataset.

Finally, we produce land use maps using the visual image classifications. We first divide the test images into a map of 50x50 regions (tiles) according to their geographic locations. The trained SVMs are then used to label each of the test images. As a result, each tile is represented by three ratios of images being classified as the three respective classes. We use the label of the highest ratio to assign a land use label to each tile label. Figures 3.4 and 3.5 show the classification maps compared to the ground truth maps of each campus. Since there are not enough test images to generate a map for each campus,

we use the cross-campus classifiers to classify the entire image set of each campus. We note that most of the academic class regions are correctly identified due to the strong performance of this classifier. The Stanford sports classifier is also able to locate a significant amount of the sports class regions correctly on the Berkeley campus. On the other hand, we can see that many labels of the residential class are missing due to the failure of this classifier.

3.1.5 Discussion

The work in this section represents a proof-of-concept of land use classification using geo-referenced on-line photos. Of course, land use maps at least in the form of campus maps already exist for university campuses. We expect our approach to generalize to other areas for which land use maps are not available. However, the ground truth for these regions will be more difficult to derive which is part of the reason we focused on university campuses here. In the following section, we will extend our work to creating land use maps of different urban land use classes in a larger dataset.

3.2 Mapping Urban Land Use in Great Britain

In the previous section, we have presented an example of how to use geo-referenced on-line photos to perform land use classification. We assume all the photos taken within a region belong to a same land use class. Due to the small study area and errors introduced during manual labeling using existing maps however, the accuracy of the location where each photo is taken cannot be verified. As a result, photos belonging to a specific class may be misplaced in a wrong region and thus weakening the classification performance.

In this section, we will overcome this problem by hand-picking the training sets and enlarging our study area. The objective of our experiment presented in this section is to locate eight types of land use classes in a large study region in Great Britain.

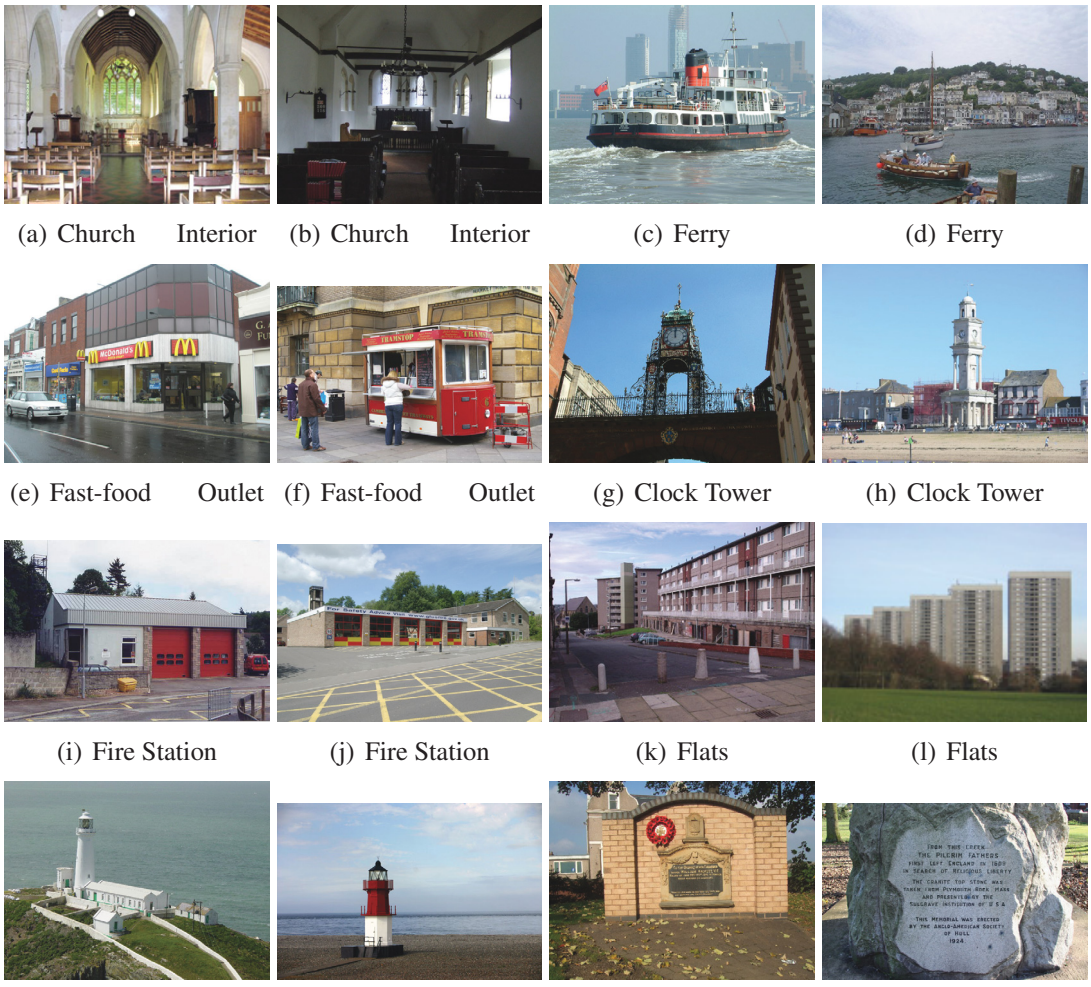
3.2.1 Dataset

The study area is the same TQ square of Great Britain we have used in Chapter 2, and we use photos from the Geograph dataset in our experiment. Since land use information is not available for our study area, we construct our training sets manually. Instead of assigning a class label to all the photos within a specific region, we use the Geograph API to perform keyword search to collect relevant photos for each training set. As a result, the photos collected can better represent their corresponding training classes.

Since our goal is to map land use classes in the TQ region, we use photos collected from outside the TQ region as the training sets. Eight land use classes are considered: church (specified by church interior), clock tower, fast-food outlet, ferry, fire station, flat, lighthouse, and memorial. The reason of choosing these eight types of land use classes is to illustrate the strength of proximate sensing. Most of these classes can be found in urban areas where they are difficult to be distinguished from overhead images. By using ground-level images, we expect that these classes can be identified. Figure 3.6 shows the sample photos from each of the eight classes. To evaluate the classification performance, a test set for each class is constructed by using keyword search to collect photos within the TQ region. Therefore, the training and test sets are mutually exclusive. Table 3.8 lists the number of photos used in this experiment.

Table 3.8: Data used in urban land use classification

Class	Training	Test
Church interior	2000	1373
Clock tower	894	232
Fast-food outlet	461	76
Ferry	1913	51
Fire station	2000	294
Flats	1969	1186
Lighthouse	2000	36
Memorial	1263	86
Background	17600	NA



(a) Church Interior (b) Church Interior (c) Ferry (d) Ferry

(e) Fast-food Outlet (f) Fast-food Outlet (g) Clock Tower (h) Clock Tower

(i) Fire Station (j) Fire Station (k) Flats (l) Flats

(m) Lighthouse (n) Lighthouse (o) Memorial (p) Memorial

Figure 3.6: Sample photos of the eight urban land use classes.

3.2.2 Experiments

We set up this experiment as a supervised multi-class classification problem with one-versus-all approach. A binary support vector machine (SVM) is trained for each class. The training set of each class is constructed such that the ratio between positive and negative samples is 1 to 2, where the negative samples are evenly selected from the rest of the classes as well as from the background photos. After the training stage, the resulting class models are then used to predict labels of the photos from each test set. The SVMs are implemented using radial basis function (RBF) kernels. We determine the optimal values for the two parameters, the penalty and the kernel width, through grid-search on a random partitioning of the training set.

We use gist as the input features of the SVMs due to their superior performance over other image features as observed in Chapter 2.

To create land use maps, we apply classifiers trained for the eight classes to all the photos collected within the TQ region. The TQ region is subdivided into 10,000 tiles and the label of each tile is the number of photos that are classified as belonging to a certain land use class. In an other word, each map of a certain land use class displays the spatial distribution of positive classifications across the TQ region. There are about 120,000 test photos in the TQ region.

Soft classification is used to produce soft-margin results instead of binary results in creating the land use maps. These soft-margin results represent how confident a classifier is when it assigns a label to a test photo. A higher positive value means a positive label is predicted with a higher confidence. To reduce the amount of false positive classification, we only consider a result with a margin higher than 1.0 to be a positive prediction. Each test photo is tested against all eight classifiers, and the class of that photo is determined by the class of the classifier producing the highest soft-margin result. As a result, each test photo is labeled as one of the eight land use classes.

3.2.3 Results

Figure 3.7 shows examples of test photos that are correctly classified into the eight land use classes. The classification results are listed in Table 3.9. We can see that the classifier of each class performs best when tested against its own class. In an other word, these classifiers are able to distinguish land use classes from each other. From the results, we also see that some classifiers have a higher false positive rate when tested against photos of certain classes. For example the model of fast-food outlet performs worst in fire station's test data, and the model of clock tower performs worst in lighthouse's test data. This may be due to the lack of positive training samples as well as the similarities between the classes. However, with sufficient amount of training samples, models of fire station and lighthouse perform much better when tested against photos from the other classes.

We also notice that the model of church interior performs very well even when it is tested against the photos from other classes. Although churches may have similar structural parts as clock towers and memorials, photos of church interior distinguish churches from the other two structures. This illustrates that using photos taken from inside a building can provide additional information and assist proximate sensing in land use classification.

Furthermore, we observe that some classifiers are able to provide information about the spatial relationship between certain land use classes. The model of lighthouse has a higher false positive rate when tested against photos from the ferry class (similar result is observed vice versa). This suggests that the two classes are spatially correlated. In fact, photos from the two classes usually contain similar scenes: water, land, and sky, and both classes sometimes appear in the same photo. Even though the classification performance of those two class models is not high, the false positive rates among those two classes can be further exploited to provide extra spatial information that will be useful to enhance the performance of land use classification.

The land use maps of the eight land use classes are displayed in Figure 3.8. The yellow lines indicate the shoreline of the study region. Since the land use data of the region

is not available, we cannot perform any quantitative analysis to verify the accuracy of the maps. We can only assume the accuracy of each map according to the accuracy of each classifier. For the classifiers with high accuracy such as church interior, we see that the spatial distribution of the class is more centralized toward certain areas. For the classifiers with low accuracy on the other hand, the spatial distribution becomes more scattered due to the high false positive rate.

From the land use maps, we can see that most of the positively identified photos appear to be located in the urban areas of the study region (the bright area at the top left-hand corner is the greater London metropolitan area). Furthermore, the maps of ferry and lighthouse classes are able to locate the shoreline of the study region as well as the river front of the River Thames, which is located at the top of the map. Once again these two maps show the spatial correlation between the two classes.

Table 3.9: Results of urban land use classification. Each row represents the results of one class model tested against test set of different classes.

	Church interior	**Clock tower**	**Fast-food outlet**	**Ferry**	**Fire station**	**Flats**	**Lighthouse**	**Memorial**
Church interior	**87.18**	14.66	11.84	11.76	9.52	10.54	0	15.12
Clock tower	19.67	**45.69**	11.84	25.49	19.39	14.84	22.22	15.12
Fast-food outlet	22.36	21.12	**75**	15.69	41.16	31.7	16.67	9.3
Ferry	2.84	10.78	11.84	**45.1**	5.1	9.78	27.78	9.3
Fire station	5.54	12.93	14.47	9.8	**54.76**	15.68	25	9.3
Flats	8.59	27.16	36.84	16.69	37.42	**58.01**	0	11.63
Lighthouse	2.55	16.81	2.63	23.53	3.06	3.37	**52.78**	13.95
Memorial	6.05	14.22	1.31	13.73	1.7	6.66	5.56	**20.93**

3.3 Summary

In this chapter, we have illustrated that land use classification can be performed using proximate sensing. Although large scale public land use data is not available to use as ground truth, we are able to validate our hypothesis by using the two datasets we have created. We hope that further studies on this topic will eventually assist in creating more diverse and accurate land use data.

Thus far we have used low- and mid-level image features to help us extract geographic information for land use classification. We will explore the idea of using object detectors as high-level features to tackle this problem in the next chapter.

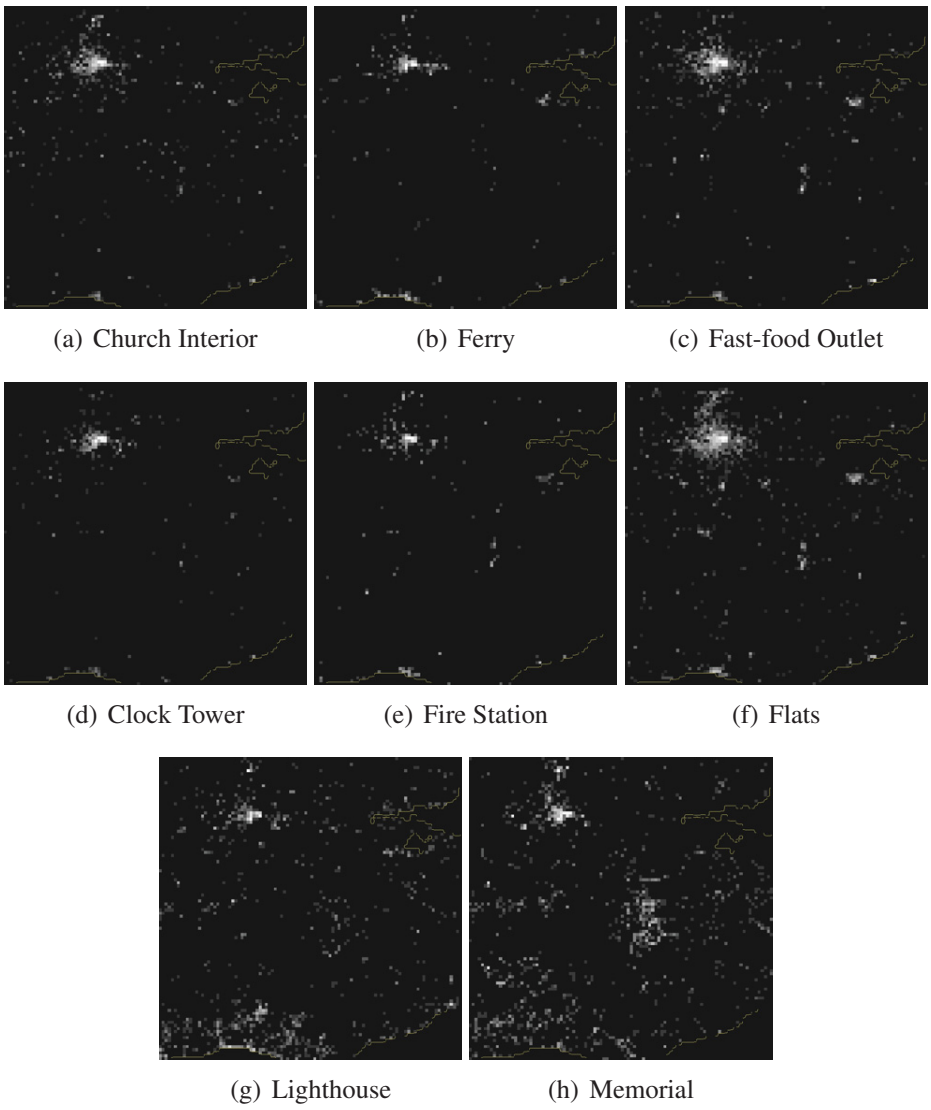
(a) Church Interior (b) Church Interior (c) Ferry (d) Ferry

(e) Fast-food Outlet (f) Fast-food Outlet (g) Clock Tower (h) Clock Tower

(i) Fire Station (j) Fire Station (k) Flats (l) Flats

(m) Lighthouse (n) Lighthouse (o) Memorial (p) Memorial

Figure 3.7: Correctly classified test photos of the eight urban land use classes.



(a) Church Interior (b) Ferry (c) Fast-food Outlet

(d) Clock Tower (e) Fire Station (f) Flats

(g) Lighthouse (h) Memorial

Figure 3.8: Land use maps of the eight urban land use classes generated using gist features. Each subregion is represented by the number of positively labeled photos in log scale.

Chapter 4

Object Detection for Land Use Classification

On-line photo sharing websites such as Flickr and Picasa have become popular channels for people to share their precious memories with one another. Although these photo collections capture many memories, they also contain other information that may be interesting particularly in different contexts. We usually think of the 5 W's and 1 H (Who, What, Where, When, Why, and How) when we read literature, but each of the photos in the collections can also provide us with some of these six types of information. Therefore, we can say that these online photo sharing websites act as a repository of all kinds of information. This allows individuals to perform knowledge discovery by crowdsourcing of information through these photo collections. With more than 200 million geo-referenced photos available from Flickr, our goal is to map what-is-where on the surface of the Earth using the "What" and "Where" aspects of the information. In particular, we explore the idea of extracting geographic information semantically for land use classification by applying object detectors directly to the photo collections.

In the previous chapters, our experiments show that low- and mid-level image visual features extracted from the geo-referenced on-line photos enable us to perform geographic knowledge discovery. However, these features do not characterize the image at a semantic level. As we have mentioned the 5 W's and 1 H at the beginning, it is very difficult

to extract these types of semantic information by using the pixel values of the photos. As a result, a high-level or top-down approach to solve the problem is considered.

In this chapter, we will study if off-the-shelf (pre-trained) object detectors can be used to extract useful geographic information to perform land use classification. We propose a novel framework of using state-of-the-art object detectors to perform geographic knowledge discovery in large collections of geo-referenced on-line photos. This framework can be applied to any land use classes, especially classes that cannot be discerned by using overhead images.

A portion of the work presented in this chapter was published as a peer-reviewed workshop paper at the International Workshop on Location Based Social Networks in 2012 [LN12a].

4.1 Experiment

Our focus in this work is to investigate whether object detectors can extract geographic information that is useful for land use classification from the geo-referenced photo collections. As a first step, we explore whether the object detectors can produce maps of objects with distinctive spatial distributions within a study region.

Our study region is the 10x11km center of metropolitan London, Great Britain. This region includes commercial, residential, as well as recreational areas. We divide the study region into 110 1x1km sub-regions (tiles) and collect photos according to the coordinates of each tile using the Flickr API. We then apply detectors of 177 objects to these photos. Table 4.1 shows a list of objects detectors used in this experiment.

The object detectors we apply in this work are the Object Bank representation developed by Li et al. [LSX10]. It is an implementation of the latent SVM detectors [FMR08] and texture classifiers [HEH05] for 177 objects in different scales and spatial pyramid levels. To detect an object at different sizes, we set the scale level to the maximum of 12 and select the highest detection rate value among the 12 levels as the detection rate for each

Table 4.1: List of objects detectors used in this experiment.

shield	fruit	people	fork	elephant	dog
plate	Ferris wheel	shoe	candle	pen	room light
television A	bottle	pup tent	baggage	mountain	boot
keyboard	coral	hat	aircraft	roller coaster	cat
sky	aquarium	floor	streetlight	lion	bird
kitchen	balloon	elevator car	computer mouse	blind	railing
rug	soccer ball	window	propeller	table	flipper
television B	baseball	shower curtain	dishwasher	cupboard	mouse
clock	bus	gravel	cow	turtle	wall
toilet seat	motorcycle	truck	cabinet	bathtub	saddle
basketnall hoop	cesspool	tower	gallery	desk	wing
beach	jersey	pool table	human	stick	drawer
door	fence	writing desk	newspaper	horse	cloud
vase	tree	camera	blanket	bench	snake
button	clam	cross	suit	duck	light
computer screen	printer	sailboat	key	backboard	ocean
spectacles	garage	snail	computing system	bride	boat
basketball	radio	goggles	hook	aqualung	basketball court
helmet	bridal gown	bed	towel	bear	animal
pot	flower	bag	sail	public toilet	swing
ball	car	cell	face veil	monkey	microphone
pool ball	filter	loudspeaker	umbrella	rabbit	squash racket
buckle	curtain	drum	ship	bus stop	knife
wheel	microwave	laptop	train	telephone	grass
seashore	building	sofa	lamp	groom	rock
desktop computer	switch	airplane	skyscraper	bridge	glove
mirror	French horn	rack	box	oxygen mask	faucet
computer monitor	mug	table-tennis table	baseball glove	bouquet	
stove	soil	dressing table	male horse	attire	
electric refrigerator	bookshelf	guitar	chair	shelf	

object. Since our focus is to detect whether an object appears in a photo or not, the spatial location of that object is not as relevant and therefore we only consider the first level of the spatial pyramid. As a result, each photo will be represented by a distribution of detection rates of the 177 objects. A threshold value is selected for each of the objects so that a particular object is considered as present in a photo if the detection rate of this object is higher than the corresponding threshold value. To generate a map of an object, we count the number of photos labeled as containing the object within each geographic tile and normalize the counts by the total number of photos within the tile. This forms a distribution of that object across the tiles, hence the object map. Figure 4.1 shows the framework of producing object maps.

In order for the results from the object detectors to be geographically informative, maps of the detected objects should display distinctive spatial distributions. To study this behavior, we perform spatial co-occurrence analysis on each object map. We treat each object map as a grayscale image and evaluate its co-occurrence matrix by measuring the distribution of spatially co-occurring object counts across the study region. We then calculate the homogeneity of the co-occurrence matrix of each object. Homogeneity is a measurement of closeness of distribution of the object counts in an object map. It ranges from 0 to 1, where a 1 indicates that locations with similar number of objects detected are clustered together. Objects with less homogeneity (or more heterogeneity) suggest that these objects are not present evenly across the study region.

Besides the distinctiveness of the object distributions, it is interesting to investigate the spatial correlations between objects since related objects should appear in the same land use region. To measure the correlation between objects we compute the correlation coefficients between the 10 objects that are the most heterogeneously distributed in the study region. The correlation coefficient ranges from -1 to 1, where a 1 (or -1) suggests that there is positive (or negative) linear relationship between the objects.

4.2 Results

The 10 most heterogeneously distributed objects are listed in Table 4.2, and their corresponding object maps are shown in Figure 4.3. From Figure 4.3, we can see that these 10 objects have different spatial distributions across the study region and we believe that these spatially distinctive distributions might provide meaningful geographic information that could be useful for land use classification.

Table 4.2: The 10 most heterogeneously distributed objects.

Objects	Homogeneity
Light	0.78
Sky	0.78
Fence	0.785
Desk	0.79
Gallery	0.79
Soil	0.795
Basketball hoop	0.8
Clock	0.8
Desktop computer	0.8
Boot	0.805

Table 4.3 shows the correlation coefficients for pairs of the 10 most heterogeneously distributed objects. While we find pairs of objects such as desks and desktop computers, plates and fruits, that are related logically, we also find some illogical pairs such as clams and gallery, and plates and basketball hoops. As we further investigate this problem, we discover that the detectors are often not detecting what they are designed to detect. In other words, the false positive rate of the detectors is high. Figure 4.2 illustrates some of the false positives from the detections.

4.3 Discussion

Our experimental results show promising opportunities of performing land use classification by detecting objects and concepts from user contributed geo-referenced photos; challenges clearly remain however.

Table 4.3: Correlation coefficients for pairs of the 10 most heterogeneously distributed objects.

	Light	Sky	Fence	Desk	Gallery	Soil	Basketball hoop	Clock	Desktop computer	Boot
Light	1.0000	0.2285	0.4196	0.1192	0.2191	0.5729	0.2406	0.2546	0.2438	0.5046
Sky	0.2285	1.0000	0.3861	0.2836	0.0650	0.3893	0.1466	0.1627	0.1838	0.3662
Fence	0.4196	0.3861	1.0000	0.4229	0.5803	0.5456	0.4455	0.4083	0.4369	0.5152
Desk	0.1192	0.2836	0.4229	1.0000	0.4428	0.1274	0.4329	0.3712	0.4853	0.5455
Gallery	0.2191	0.0650	0.5803	0.4428	1.0000	0.1691	0.4537	0.4232	0.6080	0.2406
Soil	0.5729	0.3893	0.5456	0.1274	0.1691	1.0000	0.2323	0.2108	0.1468	0.3750
Basketball hoop	0.2406	0.1466	0.4455	0.4329	0.4537	0.2323	1.0000	0.8988	0.6825	0.4016
Clock	0.2546	0.1627	0.4083	0.3712	0.4232	0.2108	0.8988	1.0000	0.5812	0.3565
Desktop computer	0.2438	0.1838	0.4369	0.4853	0.6080	0.1468	0.6825	0.5812	1.0000	0.5347
Boot	0.5046	0.3662	0.5152	0.5455	0.2406	0.3750	0.4016	0.3565	0.5347	1.0000

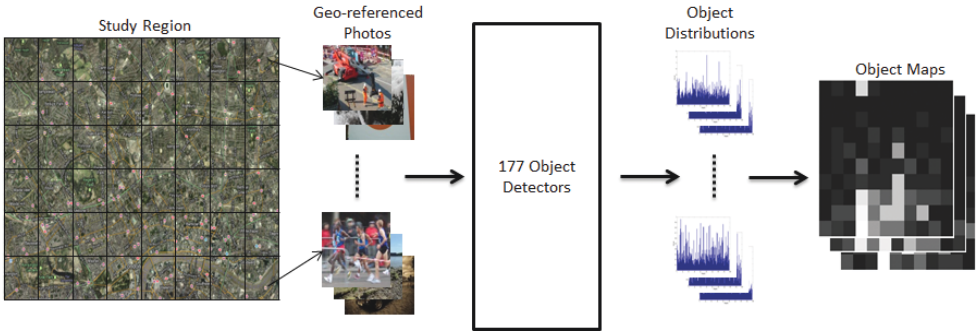
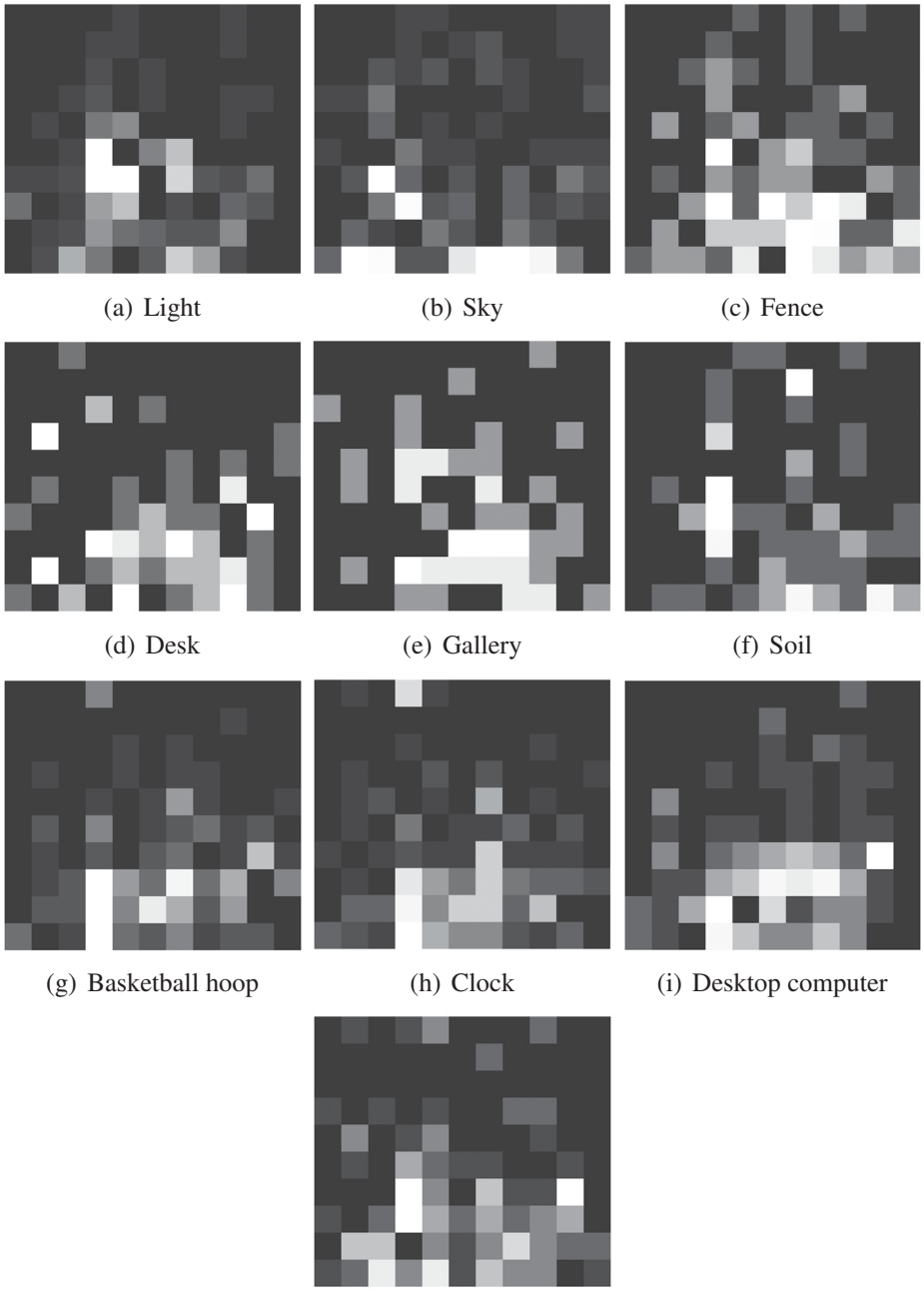


Figure 4.1: Framework for producing object maps.

(a) (b)

Figure 4.2: Examples of false detections. (a) A basketball hoop is detected. (b) A boot is detected.



(a) Light (b) Sky (c) Fence

(d) Desk (e) Gallery (f) Soil

(g) Basketball hoop (h) Clock (i) Desktop computer

(j) Boot

Figure 4.3: Spatial distributions of the 10 most heterogeneously distributed objects. Each block corresponds to a 1x1km region in the study area. The intensities of the blocks indicate the distribution of the detected objects.

Noise in datasets

Although the object detectors we applied are considered to be state-of-the-art based on evaluation using standardized datasets in the computer vision community, they fail to perform as well in the real-life photo collections that contain many different types of photos and different styles of photography. This poses a challenge to using user-contributed photo collections for geographic knowledge discovery because many of these photos are not geographically informative. One aspect of our future work will focus on how to pre-process the photo collections so that non-useful photos will be removed from the collections before any image analysis takes place. One way of achieving this might be to employ image processing techniques to remove photos with poor image quality such as blurred and low-contrast photos. Furthermore, we can analyze the textual information accompanying the photos and discard photos without any geographically informative text.

Latent information

Because the semantic information from the photo collections may not be extracted correctly due to the inaccuracy of the object detectors, we cannot determine the land use class of any region directly based on the detected object appearances. However, the distinctiveness of the spatial distributions among objects suggests that the detectors are able to observe differences across the study region. Although the detected "objects" may not have any semantic meaning, they can serve as a mid-level, or latent, information that sits between low-level and high-level image analysis. In our future work, we will investigate the use of the resulting object distributions within each geographic tile as input features to perform land use classification in a machine learning framework.

4.4 Experiments on Land Use Classification Using Object Detectors

Since the Object Bank detectors are able to capture the spatial information of different objects, we use them here to perform the same urban land use classification described in Chapter 3.2. For each of the photos in both training and test sets, we extract the Object Bank feature in 12 scales and 3 spatial pyramid levels resulting in a feature of 44604 detection rates of 177 objects. To reduced the size of the feature, we pick the highest detection rate from the results of each object detector. As a result, a 177 dimensional feature is constructed as the input of the SVM classifiers. We follow the same framework as described in Chapter 3.2.2 to train the classifiers and we compare the results to the classifiers trained using gist features.

Results

Figure 4.4 shows the classification results of the classifiers using Object Bank features. As expected, classifiers trained using Object Bank features are able to identify photos of their own classes. In fact, we see that besides fast-food outlet, the classifiers of the rest of the classes outperform the classifiers trained using gist features as compared to Table 3.9 in Chapter 3.2.

After training the classifiers, we also apply them to the photos taken from the TQ region to generate land use maps. The land use maps of the eight classes are shown in Figure 4.4. We observe that the spatial distributions of each class is more concise due to the low false positive rate from the classifiers. From the map of ferry, we can even see the resemblance of River Thames stretched between the coast and London.

Table 4.4: Results of urban land use classification using Object Bank features. Each row represents the results of one class model tested against test set of different classes.

	Church interior	Clock tower	Fast-food outlet	Ferry	Fire station	Flats	Lighthouse	Memorial
Church interior	**89.17**	12.07	5.56	2	0.68	2.45	0	16.47
Clock tower	13.65	**75.43**	9.72	0	4.79	9.63	32.35	22.35
Fast-food outlet	11.7	4.74	**56.94**	16	49.32	19.17	11.76	4.71
Ferry	1.01	3.02	5.56	**56**	1.03	5.66	29.41	2.35
Fire station	0.58	0	8.33	2	**66.44**	10.64	0	3.53
Flats	2.96	8.62	26.39	8	29.8	**64.78**	0	4.71
Lighthouse	0.79	12.93	1.39	14	1.03	2.62	**70.59**	10.59
Memorial	11.05	14.22	2.78	4	1.37	2.45	20.59	**42.35**

(a) Church Interior (b) Ferry (c) Fast-food Outlet

(d) Clock Tower (e) Fire Station (f) Flats

(g) Lighthouse (h) Memorial

Figure 4.4: Land use maps of the eight urban land use classes generated using Object Bank features. Each subregion is represented by the number of positively labeled photos in log scale.

4.5 Summary

In this chapter, we applied off-the-shelf object detectors to a collection of geo-referenced photos for the purpose of extracting semantic information from the collection. Although the detectors themselves have high detection errors, the maps they produce indicate a large range of spatial variation among objects and therefore may be used as a discriminative tool for land use classification. This assumption is further verified by the experimental results of applying Object Bank detectors to urban land use classification in Great Britain.

Chapter 5

Conclusion

In the 45 minutes or so it has taken you to read this dissertation, over four thousand geo-referenced photos have been uploaded to Flickr. These photos along with the other forms of geo-referenced on-line media contributed during this brief period are a valuable source of VGI that contain timely geographic information. The challenge to the multimedia content analysis and related computer science research communities is how to make sense of this data.

In this dissertation, we have presented one possible framework based on proximate sensing and showed how it could be used to perform binary land cover classification into developed and undeveloped regions based on the visual and textual aspects of geo-referenced on-line photos. Our extensive studies on land cover classification using more than a million images show that photographer intent of a photo collection plays a big role in the capability of extracting geographic information from the collection. As a result, removing geographic uninformative images while crowdsourcing information from the on-line photo collections is a major challenge of proximate sensing.

To demonstrate the benefits of proximate sensing, we have presented a framework of applying it on land use classification, a more challenging problem in remote sensing due to the limitations of overhead images. Our evaluations on two independent datasets show that ground-level on-line photos can supplement the overhead images to perform

land use classification.

The last problem this dissertation addresses is to perform land use classification using a top-down approach. We apply off-the-shelf object detectors to map different objects in a study area and perform spatial analysis on these object maps. Our study finds that although some of these object detectors produce high detection errors individually, the object maps they have created are actually spatially distinctive. To follow-up with this finding, we apply these detection results as intermediate features to train the land use classifiers. Our experiment shows that not only can the object detectors be used in land use classification, but they also outperform the performance of using low- and mid-level visual features.

The next step of our future work will be improving the classification performance. Thus far, we only use generic features and classification algorithms to validate our concept of proximate sensing in geographic discovery. Although our experiments produce encouraging results, we need to fine-tune each step of our framework so that the results can be applied into real life situations. In terms of image features, both low-level and high-level have proven to be useful in extracting geographic information from our image datasets. It will be interesting to see if combinations of these as well as other features will produce a better result. In terms of classifiers, other algorithms shall be researched especially when training images are scarce to provide enough data for the SVMs we use. Another interesting direction is to explore the potential use of the metadata from the images. Even though our initial experiments on using EXIF data do not indicate the benefit of using the metadata, we believe further mining of this data with proper tools should help us extract geographic information more effectively. Finally, we have seen spatial relationships between land use classes or objects in our experiments. Tobler's first law of geography states that all things are related, but nearby things are more related than distant things [Tob70]. Prior knowledge of the spatial distribution of any land cover or land use regions could improve the classification performance. There is a wealth of spatial models which could be incorporated into the proximate sensing framework ranging from linear estimation like kriging to generative probabilistic models based on Markov random fields. However, it is not clear whether these models will spatially scale-down

to the granularity of the analysis that is made possible geo-referenced on-line media, or whether new models are required. This could be an interesting research topic for the spatial analysis community.

With the experiments presented in this dissertation, we have demonstrated that it is possible to perform geographic knowledge discovery using geo-referenced ground-level on-line photo collections. We feel this is a good start to the problem but that many interesting challenges and opportunities remain ahead.

Bibliography

[CBG09] Wei-Chao Chen, Agathe Battestini, Natasha Gelfand, and Vidya Setlur. "Visual summaries of popular landmarks from community photo collections." In *Proceedings of the ACM International Conference on Multimedia*, pp. 789–792, 2009.

[CBH09] David Crandall, Lars Backstrom, Dan Huttenlocher, and Jon Kleinberg. "Mapping the World's Photos." In *Proceedings of the International World Wide Web Conference*, pp. 761–770, 2009.

[CL01] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[CLK08] Liangliang Cao, Jiebo Luo, H. Kautz, and T.S. Huang. "Annotating collections of photos using hierarchical event and scene models." In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[CPC08] M. Cristani, A. Perina, U. Castellani, and V. Murino. "Geo-located image analysis using latent representations." In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[CYL09] Liangliang Cao, Jie Yu, Jiebo Luo, and Thomas S. Huang. "Enhancing semantic and geographic annotation of Web images via logistic canonical correlation regression." In *Proceedings of the ACM International Conference on Multimedia*, pp. 125–134, 2009.

[FCW05] Peter Fisher, Alexis J. Comber, and Richard Wadsworth. "Land Use and Land Cover: Contradiction or Complement." In Peter Fisher and David J. Unwin, editors, *Re-presenting GIS*, pp. 85–98. Wiley, 2005.

[FMR08] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. "A Discriminatively Trained, Multiscale, Deformable Part Model." In *Proceedings of*

the IEEE International Conference on Computer Vision and Pattern Recognition, pp. 1–8, 2008.

[GHP07] G. Griffin, A. Holub, and P. Perona. "Caltech-256 Object Category Dataset." Technical Report 7694, California Institute of Technology, 2007.

[GJY09] A. Gallagher, D. Joshi, Jie Yu, and Jiebo Luo. "Geo-location inference from image content and user tags." In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, Workshop on Internet Vision*, pp. 55–62, 2009.

[Goo07] Michael F. Goodchild. "Citizens as sensors: The world of volunteered geography." *GeoJournal*, **69**(4):211–221, 2007.

[HE08] J. Hays and A.A. Efros. "IM2GPS: Estimating geographic information from a single image." In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[HEH05] Derek Hoiem, Alexei A. Efros, and Martial Hebert. "Automatic photo pop-up." In *ACM SIGGRAPH 2005 Papers*, SIGGRAPH '05, pp. 577–584, New York, NY, USA, 2005. ACM.

[Hof99] Thomas Hofmann. "Probabilistic latent semantic indexing." In *SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 50–57, 1999.

[Hof01] Thomas Hofmann. "Unsupervised Learning by Probabilistic Latent Semantic Analysis." *Machine Learning*, **42**(1-2):177–196, 2001.

[JL08] Dhiraj Joshi and Jiebo Luo. "Inferring generic activities and events from image content and bags of geo-tags." In *Proceedings of the International Conference on Content-based Image and Video Retrieval*, pp. 37–46, 2008.

[JNY07] Yu-Gang Jiang, Chong-Wah Ngo, and Jun Yang. "Towards Optimal Bag-of-Features for Object Categorization and Semantic Video Retrieval." In *Proceedings of ACM International Conference on Image and Video Retrieval*, 2007.

[JSR07] N. Jacobs, S. Satkin, N. Roman, R. Speyer, and R. Pless. "Geolocating Static Cameras." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1–6, 2007.

[JYC08] Yu-Gang Jiang, Akira Yanagawa, Shih-Fu Chang, and Chong-Wah Ngo. "CU-VIREO374: Fusing Columbia374 and VIREO374 for Large Scale Semantic Concept Detection." Technical report, Columbia University ADVENT #223-2008-1, 2008.

[JYN10] Yu-Gang Jiang, Jun Yang, Chong-Wah Ngo, and A.G. Hauptmann. "Representations of Keypoint-Based Semantic Concept Detection: A Comprehensive Study." *IEEE Transactions on Multimedia*, **12**(1):42–53, 2010.

[KN08] Lyndon Kennedy and Mor Naaman. "Generating diverse and representative image search results for landmarks." In *Proceedings of the International World Wide Web Conference*, pp. 297–306, 2008.

[KNA07] Lyndon Kennedy, Mor Naaman, Shane Ahern, Rahul Nair, and Tye Rattenbury. "How Flickr helps us make sense of the world: Context and content in community-contributed media collections." In *Proceedings of the ACM International Conference on Multimedia*, pp. 631–640, 2007.

[LM02] Rainer Lienhart and Jochen Maydt. "An Extended Set of Haar-Like Features for Rapid Object Detection." In *Proceedings of the IEEE International Conference on Image Processing*, pp. 900–903, 2002.

[LN09] Daniel Leung and Shawn Newsam. "Proximate sensing using georeferenced community contributed photo collections." In *ACM International Conference on Advances in Geographic Information Systems: Workshop on Location Based Social Networks*, pp. 57–64, 2009.

[LN10] Daniel Leung and Shawn Newsam. "Proximate Sensing: Inferring What-Is-Where From Georeferenced Photo Collections." In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 2955–2962, 2010.

[LN12a] Daniel Leung and Shawn Newsam. "Can Off-The-Shelf Object Detectors Be Used to Extract Geographic Information From Geo-referenced Social Multimedia?" In *ACM International Conference on Advances in Geographic Information Systems: Workshop on Location Based Social Networks*, pp. 12–15, 2012.

[LN12b] Daniel Leung and Shawn Newsam. "Exploring Geotagged Images for Land-Use Classification." In *Proceedings of the ACM International Conference on Multimedia: Workshop on Geotagging and Its Applications in Multimedia*, pp. 3–8, 2012.

[Low99] D. G. Lowe. "Object recognition from local scale-invariant features." In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2, pp. 1150–1157, 1999.

[Low04] David G. Lowe. "Distinctive Image Features from Scale-Invariant Keypoints." *International Journal of Computer Vision*, **60**(2):91–110, 2004.

[LSX10] L. Li, H. Su, E. Xing, and Li Fei-Fei. "Object Bank: A High-Level Image

Representation for Scene Classification and Semantic Feature Sparsification." In *Neural Information Processing Systems (NIPS)*, pp. 1378–1386, Canada, 2010.

[MKM08] Emily Moxley, Jim Kleban, and B. S. Manjunath. "SpiritTagger: A geo-aware tag suggestion tool mined from Flickr." In *Proceedings of the ACM International Conference on Multimedia Information Retrieval*, pp. 24–30, 2008.

[MOV98] B. S. Manjunath, J. R. Ohm, Vinod V. Vasudevan, and Akio Yamada. "Color and Texture Descriptors." *IEEE Transactions on Circuits and Systems for Video Technology*, **11**:703–715, 1998.

[NL13] Shawn Newsam and Daniel Leung. "Georeferenced Social Multimedia as Volunteered Geographic Information." In Shaowen Wang and Michael F. Goodchild, editors, *CyberGIS: Fostering a New Wave of Geospatial Discovery and Innovation*. Springer, Dordrecht, Netherlands, 2013.

[NYG05] Mor Naaman, Ron B. Yeh, Hector Garcia-Molina, and Andreas Paepcke. "Leveraging context to resolve identity in photo albums." In *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 178–187, 2005.

[OT01] Aude Oliva and Antonio Torralba. "Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope." *International Journal of Computer Vision*, **42**(3):145–175, 2001.

[PHS06] J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, editors. *Toward Category-Level Object Recognition*, volume 4170 of *LNCS*. Springer, 2006.

[QLV08] Till Quack, Bastian Leibe, and Luc Van Gool. "World-scale mining of objects and events from community photo collections." In *Proceedings of the International Conference on Content-based Image and Video Retrieval*, pp. 47–56, 2008.

[Sta65] *Standard Land Use Coding Manual*. Urban Renewal Administration, Housing and Home Finance Agency and Bureau of Public Roads, Dept. of Commerce, 1965.

[SWG06] Cees G. M. Snoek, Marcel Worring, Jan C. van Gemert, Jan-Mark Geusebroek, and Arnold W. M. Smeulders. "The challenge problem for automated detection of 101 semantic concepts in multimedia." In *Proceedings of the ACM International Conference on Multimedia*, pp. 421–430, 2006.

[TMF04] A. Torralba, K. P. Murphy, and W. T. Freeman. "Sharing features: Efficient boosting procedures for multiclass object detection." In *Proceedings of the*

IEEE International Conference on Computer Vision and Pattern Recognition, pp. 762–769, 2004.

[Tob70] W. Tobler. "A computer movie simulating urban growth in the Detroit region." *Economic Geography*, **46**(2):234–240, 1970.

[VJ01] Paul Viola and Michael Jones. "Rapid Object Detection using a Boosted Cascade of Simple Features." In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 511–518, 2001.

[WMN00] P. Wu, B. S. Manjunath, S. Newsam, and H. D. Shin. "A texture descriptor for browsing and image retrieval." *Journal of Signal Processing: Image Communication*, **16**(1):33–43, 2000.

[YCK07] Akira Yanagawa, Shih-Fu Chang, Lyndon Kennedy, and Winston Hsu. "Columbia University's Baseline Detectors for 374 LSCOM Semantic Visual Concepts." Technical report, Columbia University ADVENT #222-2006-8, 2007.

[YYQ09] Keiji Yanai, Keita Yaegashi, and Bingyu Qiu. "Detecting cultural differences using consumer-generated geotagged photos." In *Proceedings of the International Workshop on Location and the Web*, 2009.

[ZWN10] Shiai Zhu, Gang Wang, Chong-Wah Ngo, and Yu-Gang Jiang. "On the sampling of Web images for learning visual concept classifiers." In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 50–57, 2010.

[ZZS09] Yan-Tao Zheng, Ming Zhao, Yang Song, H. Adam, U. Buddemeier, A. Bissacco, F. Brucher, Tat-Seng Chua, and H. Neven. "Tour the world: Building a Web-scale landmark recognition engine." In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 1085–1092, 2009.