UNIVERSITY OF CALIFORNIA, MERCED

# Image Retrieval, Classification and Object Recognition Using Local Invariant Features in High Resolution Remote Sensing Imagery

A dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Electrical Engineering and Computer Science

by

Yang Yang

Committee in Charge:

Professor Shawn Newsam, Chair

Professor Qinghua Guo

Professor Ming-Hsuan Yang

December 2012

Image Retrieval, Classification and Object Recognition Using Local Invariant Features

in High Resolution Remote Sensing Imagery

The Dissertation of Yang Yang 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

2012

# Curriculum Vitæ

## Yang Yang

### Education

| | |
|---|---|
| 2003 | Bachelor of Engineering<br>Department of Control Engineering<br>Tsinghua University, Beijing, China |
| 2012 | Doctor of Philosophy<br>Electrical Engineering and Computer Science<br>University of California, Merced |

### Selected Publications

Y. Yang and S. Newsam: "Geographic image retrieval using local invariant features," *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 2012.

Y. Yang and S. Newsam: "Estimating the spatial extents of geospatial objects using hierarchical models," In *IEEE Workshop on Applications of Computer Vision (WACV)*, 2012.

Y. Yang and S. Newsam: "Spatial pyramid co-occurrence for image classification," In *IEEE International Conference on Computer Vision (ICCV)*, 2011.

Y. Yang and S. Newsam: “Bag-of-visual-words and spatial extensions for land-use classification,” In *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACMGIS)*, 2010.

S. Newsam and Y. Yang: “Integrating gazetteers and remote sensed imagery,” In *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACMGIS)*, 2008.

Y. Yang and S. Newsam: “Comparing SIFT descriptors and Gabor texture features for classification of remote sensed imagery,” In *IEEE International Conference on Image Processing (ICIP)*, 2008.

S. Newsam and Y. Yang: “Comparing global and interest point descriptors for similarity retrieval in remote sensed imagery,” In *ACM International Symposium on Advances in Geographic Information Systems (ACMGIS)*, 2007.

S. Newsam and Y. Yang: “Geographic image retrieval using interest point descriptors,” In *International Symposium on Visual Computing (ISVC)*, 2007.

# Abstract

## Image Retrieval, Classification and Object Recognition Using Local Invariant Features in High Resolution Remote Sensing Imagery

Yang Yang

High resolution remote sensed image data continues to become more accessible. One consequence of this is that novel geographic information system are playing an increasingly important role not only for academia, but also for daily human business and life. Nevertheless, to automate the understanding of the exponentially growing geographic image data repositories remains by-and-large an unsolved problem. In this dissertation, we put forward efforts to tackle the most important and comprehensive problems in understanding the remotely sensed image data: image retrieval, classification and object recognition.

In the interest of high resolution overhead images, we adapt and extend techniques that have been highly developed in generic vision tasks. We investigate the applications of low-level local descriptors to the remote sensed image analysis. In particular, we evaluate how local invariant descriptors perform compared to proven global texture as well as color features for similarity retrieval. We further investigate how different similarity measurements and sizes of the set of interest points used to represent images influence the retrieval. In addition, we extend our work to image classification using bag-of-visual-words models. Moreover, we explore the potential for increased synergy between two complementary data sources: gazetteers and overhead imagery. We explore ways in which these two data sources can be integrated to more fully automate geographic data management. In particular, we propose a hieararchial model to estimate the spatial extents of archived geospatial objects from gazetteers such that their spatial representations can be extended from a single latitude/longtidue pair to a bounding box.

# Acknowledgements

First, I would like to acknowledge all the funding that has supported me to conduct the work in this dissertation. This includes NSF grant IIS-0917069, a Department of Energy Early Career Scientist and Engineer/PECASE award, and NSF travel grant IIS-1049534.

Second, I would like to thank my advisory committee for their advice through the work in this dissertation. I would like to thank Professor Qinghua Guo for his comments and suggestions of this dissertation from a different academic background and perspective.

I would like to thank Professor Ming-Hsuan Yang for encouraging me to be diligent and hardworking throughout my study, for encouraging me to think critically and independently during his class, and for offering wonderful classes that expanded my horizons.

I would like to thank my advisor Professor Shawn Newsam, who have always been supportive and understanding, for giving me the opportunities to be exposed to and explore interesting research problems, as well as for his insight, guidance, and help since the first day I joined this young campus. None of the work in this dissertation would be possible without his support.

Third, I would like to thank all my fellow graduate students and friends for their help during my spell at Merced. I want to thank Lun for his help in the very early days that I came to this country. Life would have been much harder without him being around. I want to thank Ling for discussions about research and help with my homework. I want to thank Zhe and Jimei for their help. I also want to thank the Chiao family for opening their home to international students like me. They really made me feel at home.

Fourth, I would like to thank my wife Celine for her constant love. It was really heartwarming that on the days that I was busy, she took care of everything, and in the nights, she waited up for me. That made me feel rewarded and gave me the strength to continue.

Last, I would like to dedicate my work to my uncle T. H. Li. I regret that I did not get the chance to see goodbye to him but I know he will always be with me.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Background

Remote sensing imagery has been of interest to researchers in geography, the environmental sciences, and the military, and continues to accumulate at an increasing rate. Higher resolution imagery is available at more affordable or sometime no cost. Exciting new geographic information platforms, such as Google Earth and Microsoft Virtual Earth, have become accessible to the public at no charge, allowing more and more people to access these images. However, these systems only allow users to view the raw image data. A much richer interaction would be enabled by the integration of automated techniques for annotating the image content. Services such as classification, similarity retrieval, and spatial data mining would not only satisfy known demands but would also spawn new unthought-of applications.

The increased spatial resolution and coverage of overhead imagery from satellites and aircraft provides novel opportunities for advancing the field of remote sensing image analysis, particularly with regard to automated image understanding. A greater range

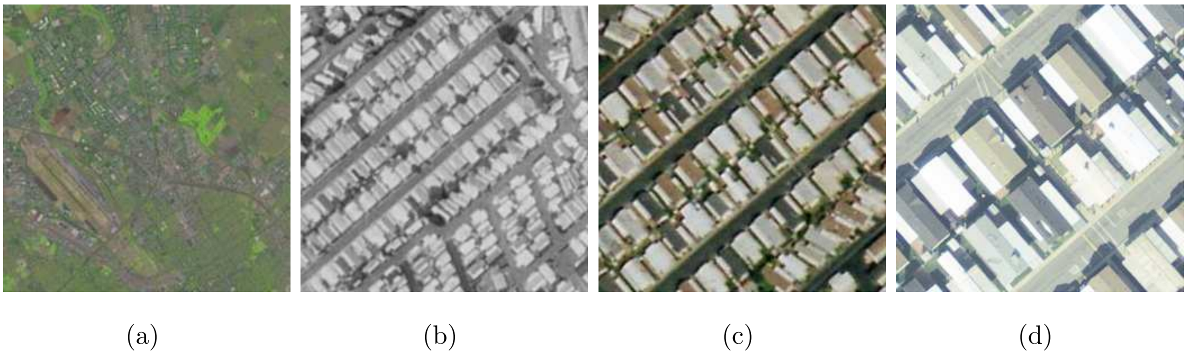


**Figure 1.1:** Images with resolutions of (a) 30m, (b) 1m, (c) 2ft (approximately 60cm), (d) and 1ft (approximately 30cm). The increased resolution of newer imagery supports analysis methods that were not possible before such as approaches based on local features which characterize individual objects and their components instead of patterns.

of objects and spatial patterns can be observed than ever before due to the increased resolution. Figures 1.1(a) through 1.1(d) show images with spatial resolutions of 30m, 1m, 2ft (approximately 60cm), and 1ft (approximately 30cm). The image in Figure 1.1(a) is from Landsat V which was launched in 1999. The images in figures 1.1(b) and 1.1(c) are aerial images with approximately the same resolutions as IKONOS which was launched in 2000 and Quickbird which was launched in 2001. Imagery with the resolution of the aerial image in 1.1(d) or even higher is now available for large geographic regions.

## 1.2 Related Work

Automated remote sensing image analysis remains by-and-large an unsolved problem. There has been significant effort over the last several decades in using low-level image descriptors, such as spectral, shape and texture features, to make sense of the raw image data. For example, textures features have been used to classify land motifs [5] and to perform similarity retrieval [6][7][8]. Unsalan et al. [9] take advantage of the shapes of rectangular buildings by detecting straight lines to classify urban and rural regions. Similar work on using building detectors to classify urban and rurual regions can be found in [10][11]. Gamba et al. [12] use edge information to detect urban areas. Benediktsson et al. [13] use morphological transformations to detect similar distributions of pixel intensities in order to find houses, but the technique is applied to only a few classes. Zhong et al. [14][15] introduce a conditional random field (CRF) framework and incorporate multilevel structural information to detect urban areas. These works represent some of the recent results in automated remote sensing image analysis. While there has been noted successes for specific problems, ample opportunities remain.

### 1.2.1 Local Invariant Features

Most of the work above has two key steps: first, compute a representation of the image; and second, analyze this representation. The representations, which are usually called features, should capture the salient characteristics of the image. Features can be generally categorized into global features and local features. Global features capture the overall representation of an image using a single vector. The simplest global feature is a vector composed of all the pixel values. Other global features include a histogram representing the distribution of the pixel intensities of the image, or the outputs of a bank of filters applied to the whole image. Local features, by contrast, are extracted from localized regions in the image. They capture the local structure of the image and thereby are potentially richer and more discriminative than global features.

The recent emergence of local features, also termed interest point descriptors, has revitalized several research areas in computer vision and have shown to be effective for a range of computer vision problems. They have been successfully applied to a number of problems including image stereo pair matching, object recognition and categorization, robot localization, panorama construction, and image retrieval. Excellent comparisons of different interest point detectors and descriptors can be found in [16] and [17] respectively.

### 1.2.2 Desirable Properties

There are generally two steps to using local invariant features for image analysis. First, a *detection* step identifies interesting locations in the image usually according to some measure of saliency. These are termed interest points. Second, a *descriptor* is computed for each of the image patches centered at the interest points. The following describes the properties of the detection and descriptor that contribute to the effectiveness of local invariant features.

#### Local

The local property of the features makes their use robust to two common challenges in image analysis. First, they do not require the challenging preprocessing step of segmentation. The descriptors are not calculated for image regions corresponding to objects or parts of objects but instead for image patches at salient locations. Second, since objects are not considered as a whole, the features provide robustness against occlusion. They have been shown to reliably detect objects in cluttered scenes even when only portions of the objects are visible. Note that occlusion includes the case where part of an object is hidden as well as the case where the object is cropped by the edge of the image.

### Invariance

Local image analysis has a long history including corner and edge detection [18]. However, the success of the more recent approaches to local analysis is largely due to the invariance of the detection and descriptors to geometric and photometric image transformations. Note that it makes sense to discuss the invariance of both the detector and descriptor. An invariant detector will identify the same locations independent of a particular transformation. An invariant descriptor will remain the same. Often, the detection step estimates the transformation parameters necessary to normalize the image patch (to a canonical orientation and scale for example) so that the descriptor itself need not be completely invariant. Geometric image transformations result from changes in viewing geometry and include translation, Euclidean (translation and rotation), similarity (translation, rotation, and uniform scaling), affine (translation, rotation, non-uniform scaling, and shear), and projective, the most general linear transformation in which parallel lines are not guaranteed to remain parallel. While affine invariant detectors have been developed [19], we choose a detector that is invariant up to similarity transformations only for two reasons. First, remote sensing imagery is acquired at a relatively fixed viewpoint (overhead) which limits the amount of non-uniform scaling and shearing. Second, affine invariant detectors have been shown to perform worse than similarity invariant descriptors when the transformation is restricted to translation, rotation, and uniform scaling [19].

Invariance to translation and scale is typically accomplished through scale-space analysis with automatic scale selection [20]. Invariance to rotation is typically accomplished by estimating the dominant orientation of the gradient of a scale-normalized image patch. We construct the evaluation dataset in the experiments below to contain regions and objects that occur at arbitrary and varying orientations as is generally the case in overhead imagery.

Photometric image transformations result from variations in illumination intensity and direction. Photometric invariance is typically obtained in both the detection and descriptor by simply modelling the transformations as being linear and relying on changes in intensity rather absolute values. Utilizing intensity gradients accounts for the possible non-zero offset in the linear model and normalizing these gradients accounts for the possible non-unitary slope. We construct the dataset used in the experiments below to contain images acquired under a range of different illumination conditions and from a number of different optical sensors.

### Robust Yet Distinctive

The features should be robust to other image transformations for which they are not designed to be invariant through explicit modelling. The detection and descriptor should not be greatly affected by modest image noise, image blur, discretization, compression artifacts, etc. Yet, for the features to be useful, the detection should be sufficiently sensitive to the underlying image signal and the descriptor sufficiently distinctive. Comprehensive evaluation [17] has shown that local invariant features achieve this balance. The evaluation dataset below contains images of varying quality.

#### Density

While detection is image dependent, it typically results in a large number of features. This density of features is important for robustness against occlusion as well as against missed and false detections. Of course, the large number of features that result from natural images present representation and computational challenges. The histograms of quantized descriptors used in this work have shown to be an effective and efficient method to help mitigate the associated costs.

#### Efficient

The extraction of local invariant features can be made computationally very efficient. This is important when processing large collections of images, such as is common in geographic image analysis, as well as for real-time applications. Real-time object detection using local features has been demonstrated in prototype systems [21] as well as in commercial products such as the SnapTell camera-phone recognition application [22].

### 1.2.3 Limitations

One of the main drawbacks of local features is that they do not provide information about local scales and shapes. Since the local features are extracted from points or small patches, they do not characterize the shape or contour of local regions. Segmentation or region based features can be regarded as a complimentary approach. However, they are sensitive to noise and are generally unstable.

Another limitation of local features is that they usually lack semantic meaning. Some local features may represent meaningful parts of the objects in the image, but in general the extraction of local features is a bottom-up approach and thus local features can only be seen as low level features. Combined with top-down approaches, local features may result in intermediate level representations which are semantically meaningful.

### 1.2.4 Computing the Similarity Between Images

One of the most important tasks in computer vison, as well as in the analysis of remote sensing image data, is computing the similarity between images. With global descriptors, each image is represented by a single vector, and therefore standard distance metrics such as the Euclidean distance can be used to measure the similarity between images. With local features, computing similarity is not as straightforward. Local features allow a more versatile description of an image than the global features, but are more difficult to compare. For example, different images may contain different numbers of features but may be perceptually very similar; the consistency of locations of features extracted from different images may not imply their correspondence in the feature space; and due to variable scale and orientation, features may be duplicated. In this case, standard metrics that are used to compare global features may not be suitable for local features.

Two widely-used methods for comparing images using interest point features are 1) voting strategies and 2) matching based on histograms of quantized features. Mikolajczyk et al. [23] introduced a voting algorithm to search for the most similar image in a database to a query based on interest point features. If the distance between two features falls below a threshold, a vote is added to the corresponding database image. The database image with the highest number of votes is considered to be the most similar one. Similar work can be found in [24]. This method is computationally expensive, and considers each feature independently. In contrast, histogram matching based on quantized features takes into account the joint distribution of the local features. It applies vector quantization to the features to generate a so called bag of words, and then represents the image by the frequency counts of those words. This strategy can be found in [25][26]. The key issues for this method are the codebook design, namely the number and the selection of the words, and the histogram similarity measure. A fundamental contribution of our work is an extensive exploration of the effects of these design parameters [27]. To our knowledge, ours is the most extensive such exploration to date and while the focus is on remote sensing images, our results should generalize to other types of images.

## 1.3 Motivation

Perhaps the most popular local invariant feature is Lowe's SIFT (Scale-Invariant Feature Transform) features [28] which are essentially histograms of gradient information at salient locations. A number of other local features have been propsed however. Belongie et al. [29] developed a local descriptor termed shape context that is a histogram of edge information with respect to a salient location. Kadir et al. [30] developed an algorithm that can find salient locations based on entropy theory. Mikolajczyk et al. [19] developed the Harris-Affine detector that is modification of the Harris detector that is scale and affine invariant. We utilize SIFT features in our work. We expect our results, however, to hold for other local invariant features.

Due to their discrmination ability, robustness to changes in scale, orientation, and illumination, SIFT features have been successfully applied to a range of computer vision problems, including object recognition [28], video tracking [31], gesture recognition [32], and 3D modeling [33] [34]. Schmid et al. [17] has shown that SIFT features outperform a number of other local features in matching under different image transformations. However, until recently, the resolution of remote sensing imagery was not high enough to support local features and the analysis focused on "stuff" rather than things. This changed as very high resolution imagery became available, and techniques that have been applied to generic vision problems have become appropriate. Our investigation on local invariant features for analyzing high resolution remote sensing imagery is thus timely since it only recently became possible.

While local features have potential for analyzing high resolution remote sensing imagery, there has not been much work in this area. Porway et al. [35] used edge features, color histograms, and SIFT features to represent several classes of objects, and combined those features in a hierarchical and contextual model to learn different scenes. However, they mainly focus on urban areas, especially commercial areas, thus their approach can only learn objects such as trees, cars, roofs, roads, and parking lots. Sirmacek et al. [36] modeled each house in an image as a subgraph in which SIFT features are taken as vertices. Recognition is then formulated as a subgraph matching and graph cuts problem. Their images are extracted only from uniform residential areas and their method is mainly based on the affinity of the appearances of different houses. In reality, the appearance of houses may vary a lot, thus their strategy may become restrictive and unrealistic.

## 1.4 Overview of the Dissertation

This thesis focuses on the automatic understanding of remote sensing imagery, demonstrating how local invariant features are effective for the analysis of remotely sensed images in three major aspects: image retrieval, classification, and object detection. The following chapter shows the application of local features to image retrieval and investigates how different settings and similarity metrics can affect the retrieval performance. Chapter 3 empirically evaluates the employment of local features for classifying different land-use/land-cover (LULC) classes and investigates how spatial information can improve the overcall classification. Chapter 4 presents a framework in which non-image data and image data can fuse together to detect complex geospatial objects. The last chapter concludes this dissertation.

## 1.5 Summary of Contributions

The contribution of this dissertation can be summarized as follows:

- The first study of local invariant features for content-based geographic image retrieval, in particular showing their superiority over standard features such as color and texture.

- The most thorough investigation of the effects of different design parameters of quantized local invariant features for any image analysis problem, not just overhead imagery.

- A first-of-its-kind 21 land-use/land-cover dataset made publicly available to other researchers. We anticipate this will serve as a standardized dataset for evaluating different techniques, which has largely been lacking in the field of analyzing remote sensing imagery.

- The proposed spatial pyramid co-occurrence approach which characterizes both the absolute and relative spatial layout of images.

- The investigation of the potential synergy between gazetteers and overhead imagery for estimating the spatial extents of complex geospatial object and a novel framework which integrates both non-image and image data for solving this problem.

- A hierarchical model that represents the appearance of complex geospatial objects using latent land-use/land-cover classes.

# Chapter 2

# Image Retrieval

The work presented in this chapter was published as peer-reviewed full conference papers at the International Symposium on Visual Computing in 2007 [37] and the ACM International Symposium on Advances in Geographic Information Systems in 2007 [38], and as a journal paper in the IEEE Transactions on Geoscience and Remote Sensing in 2012 [27].

## 2.1 Background

Content-based image retrieval (CBIR) has been an active research area in computer vision for over a decade with IBM's Query by Image Content (QBIC) system from 1995 [39] being one of the earliest successes. A variety of image descriptors have been investigated including color, shape, texture, spatial configurations, and others. A survey of CBIR is available in [40].

Similarity based image retrieval has been proposed as an automated method for managing and interacting with the the growing collections of remote sensing imagery.

As in other domains, a variety of descriptors have been investigated including spectral [41, 42], shape [43], texture [6–8, 44, 45], and combinations such as multi-spectral texture [46]. While the most effective descriptor is problem dependent, texture features have enjoyed success since, unlike spectral features, they incorporate spatial information which is clearly important for remote sensing imagery but avoid the difficult pre-processing step of segmentation needed to extract shape features.

In this chapter, we investigate the application of interest points, or local invariant features, to remote sensing image retrieval. Interest point descriptors have enjoyed surprising success for a range of traditional computer vision problems. The application of interest point detectors and descriptors to image retrieval has focused primarily on *retrieving images of the same object or scene under different conditions.* Examples include finding additional appearances of a given object in scenes or shots in a video [26], finding images of 3D objects acquired from different viewpoints [47–49] or against different backgrounds [50], finding images belonging to distinct, homogeneous semantic categories [48, 51], finding frames of the same scene in a video [52], and finding images of the same indoor scene for localization [53]. There has been little application to finding *similar* images or image regions. Further, there has been little research into applying interest point detectors and descriptors to the problem of similarity retrieval in large collections of high resolution remote sensing imagery.

In this chapter, our investigation is done in the context of similarity retrieval. In particular, we compare interest point descriptors to global texture features, which have been shown to be particularly effective for remote sensing image retrieval, as well as color histogram features. Similarity retrieval is not only an interesting application but also serves as an excellent platform for evaluating the overall descriptiveness of a descriptor. Many of the findings will likely inform other image analysis. Indeed, computing image similarity is fundamental to kernel based methods such as non-linear support vector machines.

## 2.2 Image Features

### 2.2.1 Global Features

We consider Gabor texture features as the global features. Texture features, and in particular Gabor texture features, have proven to be effective for performing content-based similarity retrieval in remote sensing imagery [6–8, 44–46]. The MPEG-7 Multimedia Content Description Interface [54] standardized Gabor texture features after they were shown to outperform other texture features for similarity retrieval. One of the evaluation datasets used in the competitive standardization process consisted of remote sensing imagery.

Gabor texture analysis is accomplished by applying a bank of scale and orientation selective Gabor filters to an image. Gabor functions are Gaussians functions modulated by a sinusoid. Two dimensional spatial filters based on Gabor functions can be made orientation and scale selective by controlling this modulation. While the choice of the number of orientations and scales is application dependent, experimentation has shown that a bank of filters tuned to combinations of five scales, at octave intervals, and six orientations, at 30-degree intervals, is sufficient for the analysis of remote sensing imagery.

A Gabor texture feature vector is formed from the filter outputs as follows [55]. Applying a bank of Gabor filters with $R$ orientations and $S$ scales to an image results in a total of $R$x$S$ filtered images:

$$f'_{11}\left(x,y\right),\ldots,f'_{RS}\left(x,y\right)\ . \tag{2.1}$$

A single global feature vector for the original image is formed by computing the first and second moments of the filtered images. That is, a $2RS$ dimension feature vector, $h_{GABOR}$, is formed as

$$h_{GABOR}=\left[\mu_{11},\sigma_{11},\mu_{12},\sigma_{12},\ldots,\mu_{1S},\sigma_{1S},\ldots,\mu_{RS},\sigma_{RS}\right]\ , \tag{2.2}$$

where $\mu_{rs}$ and $\sigma_{rs}$ are the mean and standard deviation of $f'_{rs}\left(x,y\right)$. Finally, to normalize for differences in range, each of the $2RS$ components can be scaled to have a mean of zero and a standard deviation of one across the entire dataset.

### 2.2.2 Local Invariant Features

We choose David Lowe's SIFT [28, 56] as the local feature. SIFT-based descriptors have been shown to be robust to image rotation and scale, and to be capable of matching images with geometric distortion and varied illumination. An extensive comparison with other local descriptors found that SIFT-based descriptors performed the best in an image matching task [17]. Like most interest point based analysis, there are two components to SIFT-based analysis. First, a detection step locates points that are identifiable from different views. This process ideally locates the same regions in an object or scene regardless of viewpoint, illumination, etc. Second, these locations are described by a descriptor that is distinctive yet also invariant to viewpoint, illumination, etc. In short, SIFT-based analysis focuses on image patches that can be found and matched under different image acquisition conditions.

The detection step is designed to find image regions that are salient not only spatially but also across different scales. Candidate locations are initially selected from local extrema in Difference of Gaussian (DoG) filtered images in scale space. The DoG images are derived by subtracting two Gaussian blurred images with different $\sigma$

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma) \ . \tag{2.3}$$

where $L(x, y, \sigma)$ is the image convolved with a Gaussian kernel with standard deviation $\sigma$, and $k$ represents the different sampling intervals in scale space. Each point in the three dimensional DoG scale space is compared with its eight spatial neighbors at the same scale, and with its nine neighbors at adjacent higher and lower scales. The local maximum or minimum are further screened for minimum contrast and poor localization along elongated edges. The last step of the detection process uses a histogram of gradient directions sampled around the interest point to estimate its orientation. This orientation is used to align the descriptor to make it it rotation invariant. We refer to the standard approach described above as saliency-based feature extraction.

A feature descriptor is then extracted from the image patch centered at each interest point. The size of this patch is determined by the scale of the corresponding extremum in the DoG scale space. This makes the descriptor scale invariant. The feature descriptor consists of histograms of gradient directions computed from a 4x4 spatial grid. The interest point orientation estimate described above is used to align the gradient directions to make the descriptor rotation invariant. The gradient directions are quantized into eight bins so the final feature vector has dimension 128 (4x4x8). This histogram-of-gradients descriptor can be roughly thought of a summary of the edge information in the image patch centered at the interest point.

An alternative way to extract SIFT descriptors is to use a fixed grid. This approached is refered as grid-based feature extraction. It is often all called dense sampling as it typically results in a larger number of descriptors since interest points are not detected in non-salient regions (of uniform intensity for example).

### 2.2.3 Color Features

Color histogram features are computed in three color spaces: red green blue (RGB), hue lightness saturation (HLS), and CIE Lab. Each dimension is quantized into eight bins for a total histogram feature length of 512. The histograms are normalized to sum to one (L1 norm equal to one). Given a color image, this results in three different color histogram features: $f_{RGB}$, $f_{HLS}$, and $f_{Lab}$.

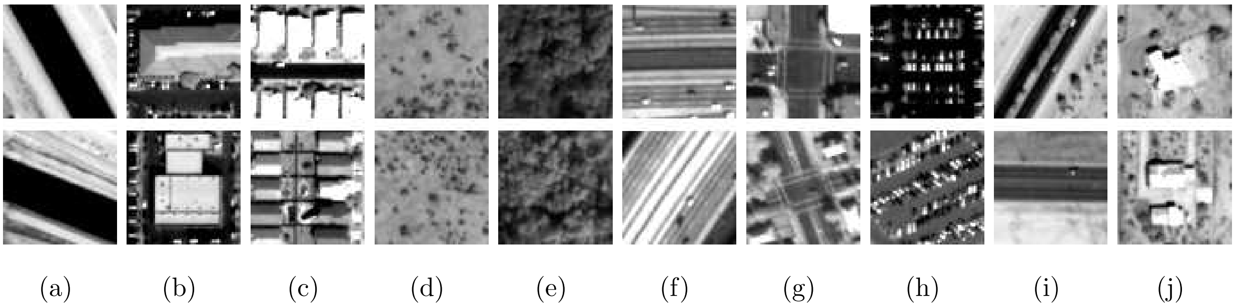


**Figure 2.1:** Two examples from each of the groundtruth classes in the IKONOS dataset. (a) Aqueduct. (b) Commercial. (c) Dense residential. (d) Desert chaparral. (e) Forest. (f) Freeway. (g) Intersection. (h) Parking lot. (i) Road. (j) Rural residential.

## 2.3 Groundtruth Datasets

Two groundtruth datasets were created to conduct quantitative analyses: the IKONOS dataset and the USGS dataset.

### 2.3.1 IKONOS Dataset

The first groundtruth dataset was created from a collection of IKONOS 1-m panchromatic satellite images of the United States. Ten sets of 100 64-*by*-64 pixel grayscale images were manually extracted from 22 large IKONOS images covering major cities in the US for the following LULC classes: aqueduct, commercial, dense residential, desert chaparral, forest, freeway, intersection, parking lot, road, and rural residential. Figure 2.1 shows two examples from each of these ten classes.

### 2.3.2 USGS Dataset

The second groundtruth dataset consists of images of 21 LULC classes selected from aerial orthoimagery with a pixel resolution of one foot. Large images were downloaded from the United States Geological Survey (USGS) National Map of the following US regions: Birmingham, Boston, Buffalo, Columbus, Dallas, Harrisburg, Houston, Jacksonville, Las Vegas, Los Angeles, Miami, Napa, New York, Reno, San Diego, Santa Barbara, Seattle, Tampa, Tucson, and Ventura. 100 images measuring $256 \times 256$ pixels were manually selected for each of the following 21 classes: agricultural, airplane, baseball diamond, beach, buildings, chaparral[1], dense residential, forest, freeway, golf course, harbor, intersection, medium density residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis courts. These classes were selected because they contain a variety of spatial patterns, some homogeneous with respect to texture, some homogeneous with respect to color, others not homogeneous at all, and thus represent a rich dataset for our investigation.

Five samples of each class are shown in figure 2.2. The images downloaded from the National Map are in the RGB colorspace. Both RGB and grayscale versions of the 2100 groundtruth images are used where $Gray = 0.299 * R + 0.587 * G + 0.114 * B$.

A significant benefit of using aerial orthoimagery from the USGS National Map is that the data is already in the public domain. Thus, our 21 class LULC dataset is the largest dataset of its kind that can be made publicly available to other researchers. The dataset

[1] This class might be more appropriately labelled "desert scrub".

is available free of charge on our research group's website. Our server logs indicate this dataset has been downloaded over 100 times.

## 2.4 Experiments

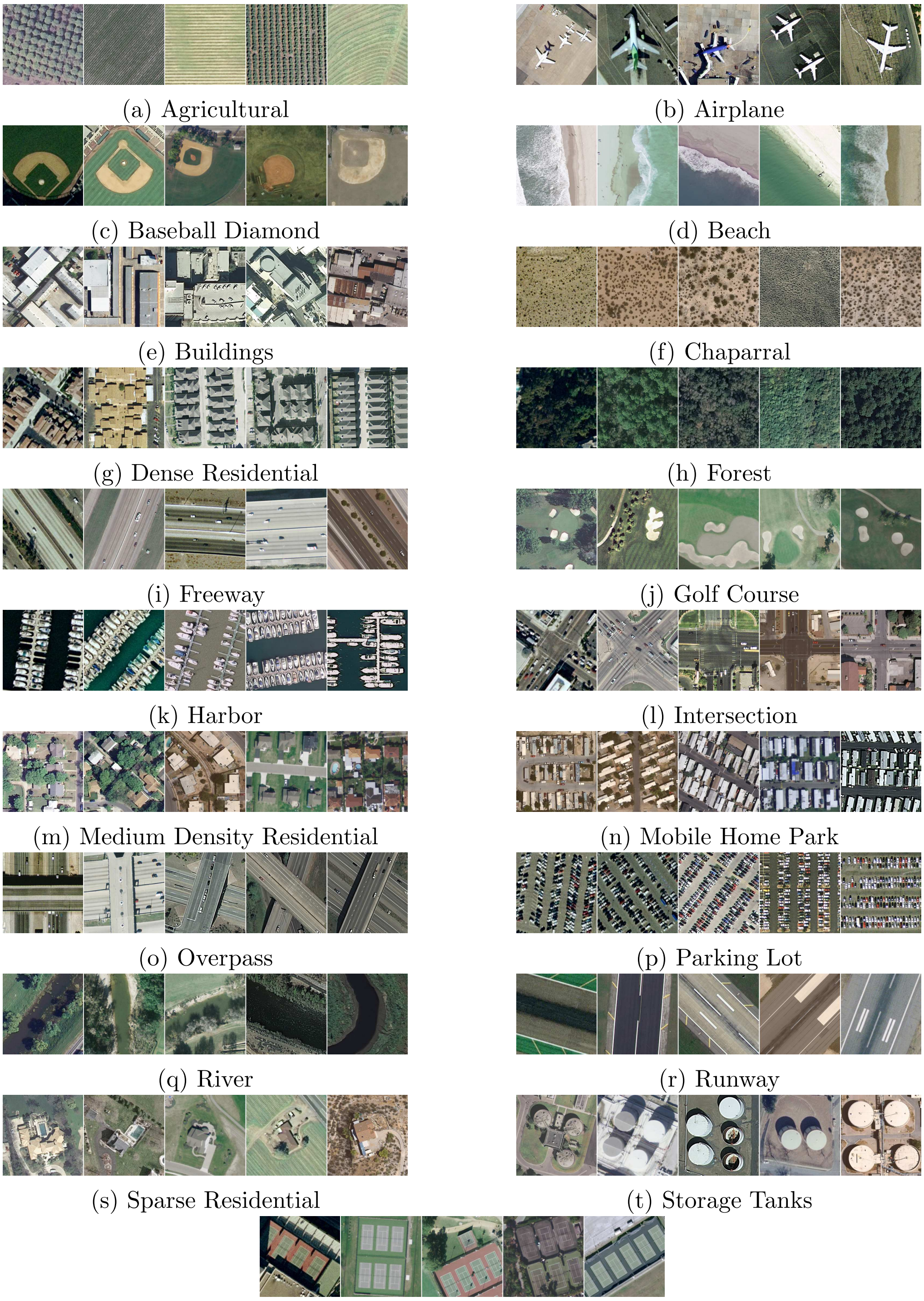
### 2.4.1 Similarity Retrieval

The metrics to measure the performance of different features in the similarity retrieval are as follows. Let $T$ be a collection of $M$ images; let $h^m$ be the feature vector extracted from image $m$, where $m \in 1, \ldots, M$; let $d(\cdot, \cdot)$ be a distance function defined on the feature space; and let $h^{query}$ be the feature vector corresponding to a given query image. Then, the image in $T$ most similar to the query image is the one whose feature vector minimizes the distance to the query's feature vector:

$$m^* = \underset{1 \leq m \leq M}{\arg\min}\, d(h^{query}, h^m) \ . \tag{2.4}$$

Likewise, the $k$ most similar images are those that result in the $k$ smallest distances when compared to the query image. Retrieving the $k$ most similar items is commonly referred to as a $k$-nearest neighbor ($k$NN) query.

Given a groundtruth dataset, there are a number of ways to evaluate retrieval performance. One common method is to plot the precision of the retrieved set for different values of $k$. Precision is defined as the percent of the retrieved set that is correct and can be computed as the ratio of the number of true positives to the size of the retrieved set.



(a) Agricultural

(b) Airplane

(c) Baseball Diamond

(d) Beach

(e) Buildings

(f) Chaparral

(g) Dense Residential

(h) Forest

(i) Freeway

(j) Golf Course

(k) Harbor

(l) Intersection

(m) Medium Density Residential

(n) Mobile Home Park

(o) Overpass

(p) Parking Lot

(q) River

(r) Runway

(s) Sparse Residential

(t) Storage Tanks

(u) Tennis Courts

**Figure 2.2:** The groundtruth dataset contains 100 images from each of 21 LULC classes. Five samples from each class are shown above.

It is straight forward and meaningful to compute and compare the average precision for a set of queries when the groundtruth sizes are the same. (It is not straight forward to do this for precision-recall curves.)

Plotting precision versus the size of the retrieved set provides a graphical evaluation of performance. A single measure of performance that not only considers that the groundtruth items are in the top retrievals but also their ordering can be computed as follows [54]. Consider a query $q$ with a groundtruth size of $NG(q)$. The $Rank(k)$ of the $k$th groundtruth item is defined as the position at which it is retrieved. A number $K(q) \geq NG(q)$ is chosen so that items with a higher rank are given a constant penalty

$$Rank(k) = \begin{cases} Rank(k), & \text{if } Rank(k) \leq K(q) \\ \\ 1.25K(q), & \text{if } Rank(k) > K(q) \end{cases} . \qquad (2.5)$$

$K(q)$ is commonly chosen to be $2NG(q)$. The *Average Rank* (AVR) for a single query $q$ is then computed as

$$AVR(q) = \frac{1}{NG(q)} \sum_{k=1}^{NG(k)} Rank(k) \; . \qquad (2.6)$$

To eliminate influences of different $NG(q)$, the *Normalized Modified Retrieval Rank* (NMRR)

$$NMRR(q) = \frac{AVR(q) - 0.5[1 + NG(q)]}{1.25K(q) - 0.5[1 + NG(q)]} \qquad (2.7)$$

is computed. $NMRR(q)$ takes values between zero (indicating whole groundtruth found) and one (indicating nothing found) irrespective of the size of the groundtruth for query

$q$, $NG(q)$. Finally, the *Average Normalized Retrieval Rate* (ANMRR) can be computed for a set $NQ$ of queries

$$ANMRR = \frac{1}{NQ} \sum_{q=1}^{NQ} NMRR(q) \ . \tag{2.8}$$

### 2.4.2 Comparing Global and SIFT Features on the IKONOS Dataset

#### Dissimilarity Measures

**Gabor Texture Features** The (dis)similarity between two images is measured by computing the Euclidean distance between their texture features

$$d(h1, h2) = \|h1 - h2\|_2 \ = \sqrt{\sum_{i=1}^{2RS} (h1_i - h2_i)^2} \ . \tag{2.9}$$

This results in an orientation (and scale) sensitive similarity measure. Orientation invariant similarity is possible by using the modified distance function

$$d_{RI}(h1, h2) = \min_{r \in R} \|h1_{<r>} - h2\|_2 \tag{2.10}$$

where $h_{<r>}$ represents $h$ circularly shifted by $r$ orientations:

$$\begin{aligned} h_{<r>} &= [(h_{r1}, h_{r2}, \cdots, h_{rS}), (h_{(r+1)1}, h_{(r+1)2}, \cdots, h_{(r+1)S}), \\ &\quad \cdots, (h_{R1}, h_{R2}, \cdots, h_{RS}), (h_{11}, h_{12}, \cdots, h_{1S}), \\ &\quad \cdots, (h_{(r-1)1}, h_{(r-1)2}, \cdots, h_{(r-1)S})] \,. \end{aligned} \tag{2.11}$$

Parentheses have been added for clarity. Conceptually, this distance function computes the best match between rotated versions of the images without repeating the feature extraction. The granularity of the rotations is of course limited by the filter bank construction.

**SIFT Features** Rather than work with the full 128 dimension SIFT feature vectors, we adopt a standard approach, termed bag-of-visual-words (BOVW) [26], to summarize the descriptors without regard to where they appear in an image. The analogy to representing text documents as word count frequencies is made possible by quantizing the 128 dimension SIFT descriptors. We clustered a large sampling of the features and labelled the full SIFT feature set with the label of the closest cluster center. Representing the features using the cluster labels has been shown to be effective in other image retrieval tasks [26]. The clustering was performed using the standard k-means algorithm.

The final interest point descriptor used to compute the similarity between two images is composed of the frequency counts of the labelled SIFT feature vectors. That is, $h_{INT}$

for an image, is

$$h_{INT} = [t_0, t_1, \ldots, t_{k-1}] \ , \tag{2.12}$$

where $t_i$ is number of occurrences of SIFT features with label $i$ in the image. $h_{INT}$ is similar to a term vector in document retrieval. The cosine distance has shown to be effective for comparing documents represented by term vectors [57] so we use it here to compute the similarity between images:

$$d(h1, h2) = \frac{\sum\limits_{i=0}^{k-1} h1_i h2_i}{\sqrt{\sum\limits_{i=0}^{k-1} h1_i^2 \sum\limits_{i=0}^{k-1} h2_i^2}} \ . \tag{2.13}$$

The cosine distance measure ranges from zero (no match) to one (perfect match). To make it compatible with the distance function used for comparing the global Gabor texture features, for which zero is a perfect match, we use one minus the cosine distance to perform similarity retrieval using interest point descriptors.

### Datasets

The IKONOS groundtruth dataset is used to conduct the quantitative analyses and perform the comparisons. In addition, a separate dataset was created for the qualitative analyses. Two large IKONOS images of the Phoenix and Los Angeles areas were partitioned into non-overlapping 64-*by*-64 pixel tiles. The Phoenix image measures 21,248-*by*-11,328 pixels for a total of 58,764 tiles and the Los Angeles image measures 10,560-*by*-

**Figure 2.3:** Image patches corresponding to two of the 50 clusters used to label the SIFT features. The top row shows a cluster that has captured corner-like patches. The bottom row shows a cluster that has captured grid-like patches.

10,624 pixels for a total of 27,390 tiles (86,154 tiles in total). The images have a resolution of 1m and are grayscale, which is the same as the IKONOS groundtruth dataset.

### Feature Extraction

For the qualitative analyses, a single Gabor texture feature was extracted from each tile from the two large IKONOS images using a filterbank tuned to $R = 6$ orientations and $S = 5$ scales. The interest points descriptors were extracted and assigned to the tiles as follows. First, interest points and SIFT features were extracted from the complete images, resulting in 4,880,415 features for the Phoenix image and 2,406,787 features for the Los Angeles image. 100,000 features were sampled from the combined set and clustered into 50 clusters using k-means clustering. Figure 2.3 shows sample image patches for two of the 50 clusters. Each of the 7,287,202 features in the large images was labelled based on the clustering results and assigned to the tile containing the interest point location. Thus, the 86,154 tiles contained 84.6 labelled SIFT features on average. Finally, a single interest point descriptor consisting of the label counts was assigned to each tile.

For the quantitative analyses, a single Gabor texture feature was extracted for each of the 1,000 groundtruth images of the IKONOS dataset, again using a filterbank tuned to

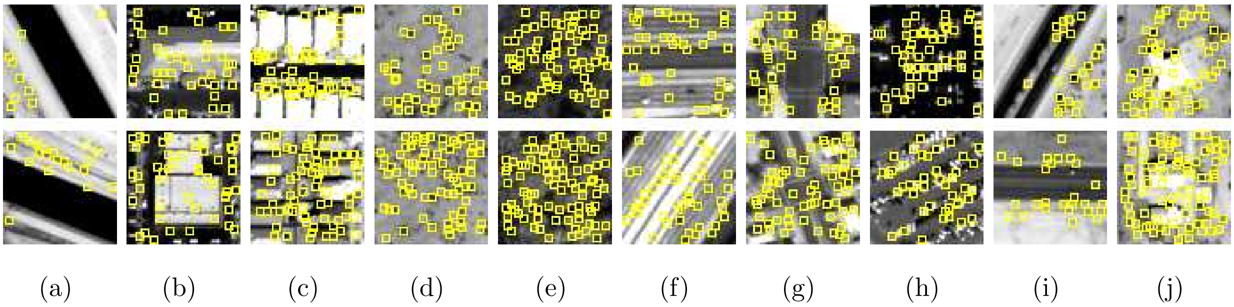


**Figure 2.4:** The interest point locations for the images in figure 2.1.

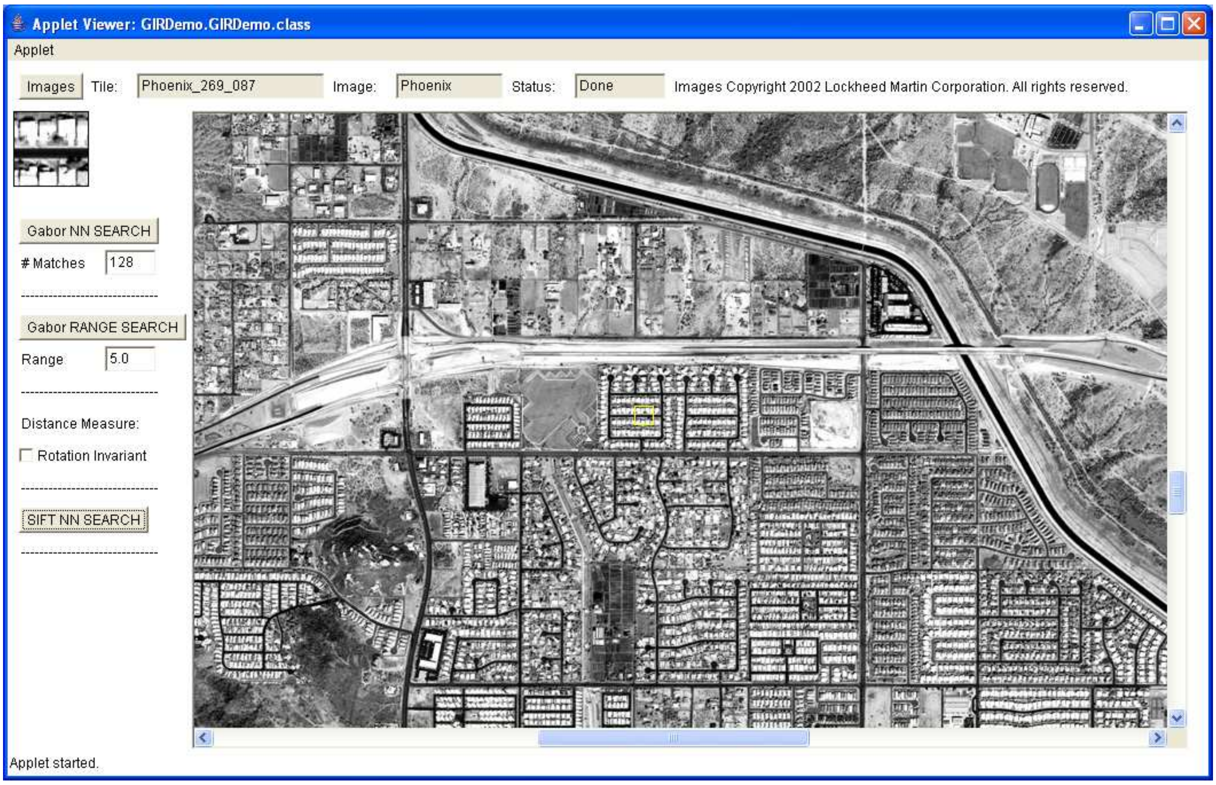
$R = 6$ orientations and $S = 5$ scales. Interest points and SIFT features were also extracted from each image and labelled using the clustering from the larger dataset above (thus the clustering and labeling was not tuned to the groundtruth dataset). A single interest point descriptor consisting of the label counts was assigned to each image. The images here contained an average of 59.1 labelled features (fewer than above since the SIFT features were extracted from the small images placing an upper bound on the scale of the interest points). Figure 2.4 shows the locations of the detected interest points for the sample images in figure 2.1. It is worth pointing out the different feature extraction times for the groundtruth dataset. It took approximately 51 seconds to extract and label the interest points and approximately 353 seconds to extract the Gabor texture features for the 1,000 images in the groundtruth dataset (on a typical desktop workstation). While the extraction software was not optimized and the timing measurements were not scientific, we believe this order-of-magnitude difference between the two features is to be expected. Efficient extraction is a noted strength of SIFT features.

### Qualitative Analysis

A Geographic Image Retrieval (GIR) demonstration application was used for the qualitative analysis. The GIR demo allows a user to navigate large IKONOS images and select 64-*by*-64 pixel tiles as query images. The user can then perform a $k$-nearest neighbor query using either the interest points or the global Gabor texture features. The most similar $k$ tiles in the result set is displayed in order of decreasing similarity. This demo turns out to be a valuable tool for evaluating the descriptive power of a feature. Figure 2.5 shows a screen capture of the GIR demo in which the user has selected a tile from a dense residential region in the center of the displayed IKONOS image of Phoenix. The user is now ready to perform a 128-nearest neighbor query in the 86,154 tile Los Angeles and Phoenix image dataset. Figure 2.6 shows the top 32 retrievals in order of decreasing similarity for this query tile for each of the three approaches.

### Quantitative Analysis

The quantitative analysis involved a comprehensive set of similarity retrievals using each of the 1,000 images in the groundtruth dataset as a query. Precision was computed for each query as a function of retrieved set size from 1 to 1,000. These precision values were then averaged over the 100 queries from each of the ten groundtruth classes. This was performed three times: 1) for the interest point descriptors; 2) for the global Gabor texture features using the standard orientation sensitive distance measure; and 3) for the global Gabor texture features using the modified rotation invariant (RI) distance



Applet Viewer: GIRDemo.GIRDemo.class

Tile: Phoenix_269_087

Image: Phoenix

Status: Done

Images Copyright 2002 Lockheed Martin Corporation. All rights reserved.

# Matches 128

Range 5.0

Distance Measure:

☐ Rotation Invariant

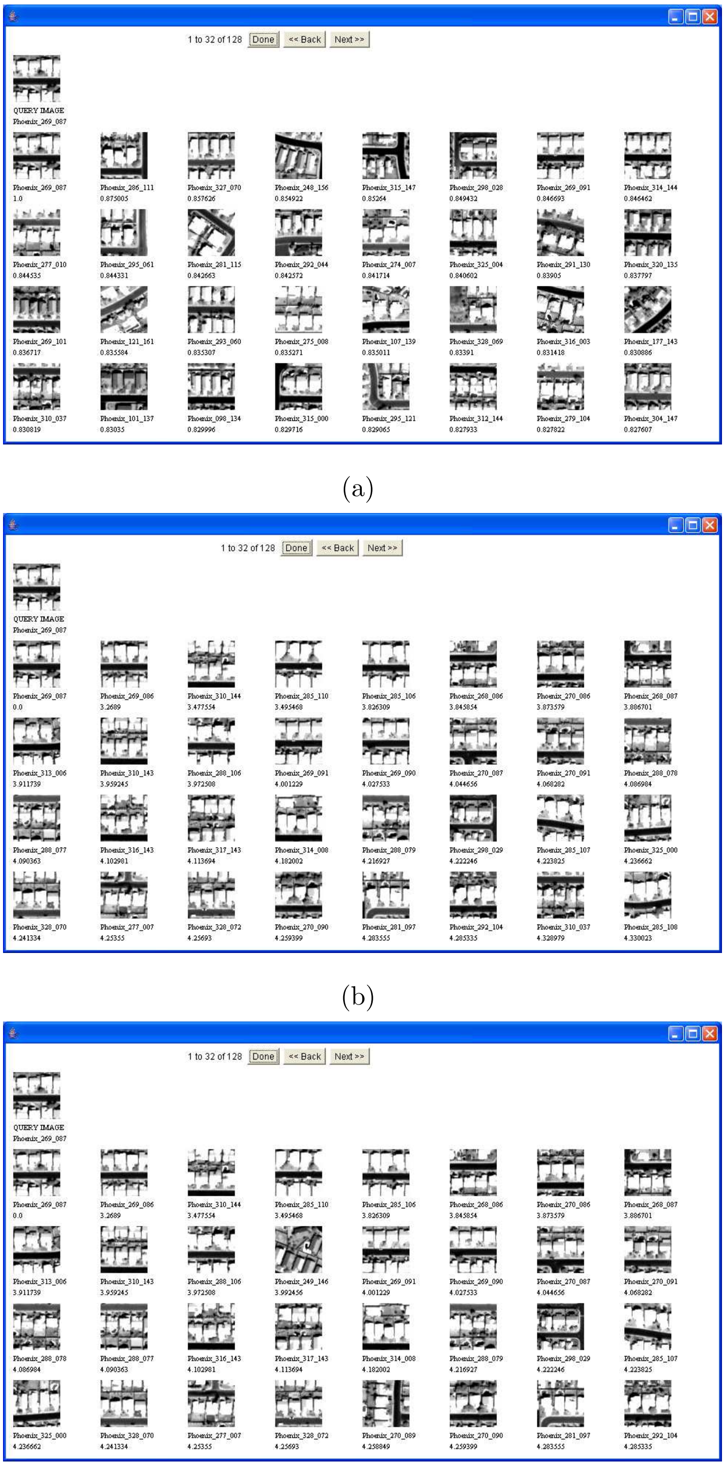
Applet started.

**Figure 2.5:** The Geographic Image Retrieval demo which allows users to perform similarity retrieval in remote sensing imagery.

measure. Figure 2.7 shows the averaged precision curves for the ground truth datasets. The optimal case is also plotted for comparison.

The *Average Normalized Modified Retrieval Rate* (ANMRR) described in section 2.4.1 was also computed for each of the three similarity retrieval methods, for each of the ten groundtruth classes. Table 2.1 shows these values which range from zero for all the groundtruth items retrieved in a result set the size of the groundtruth to one for none of the groundtruth items retrieved.

Again, the interest point descriptors were more computationally efficient, this time in terms of how long it took to perform all 1,000 queries. On average, using the interest point descriptors took only two seconds, using the global Gabor texture features took 12



QUERY IMAGE
Phoenix_269_087

Phoenix_269_087
1.0

Phoenix_286_111
0.875005

Phoenix_327_070
0.857626

Phoenix_248_156
0.854922

Phoenix_315_147
0.85264

Phoenix_298_028
0.849432

Phoenix_269_091
0.846693

Phoenix_314_144
0.846462

Phoenix_277_010
0.844535

Phoenix_295_061
0.844331

Phoenix_281_115
0.842663

Phoenix_292_044
0.842572

Phoenix_274_007
0.841714

Phoenix_325_004
0.840602

Phoenix_291_130
0.83905

Phoenix_320_135
0.837797

Phoenix_269_101
0.836717

Phoenix_121_161
0.835584

Phoenix_293_060
0.835307

Phoenix_275_008
0.835271

Phoenix_107_139
0.835011

Phoenix_328_069
0.83391

Phoenix_316_003
0.831418

Phoenix_177_143
0.830886

Phoenix_310_037
0.830819

Phoenix_101_137
0.83035

Phoenix_098_134
0.829996

Phoenix_315_000
0.829716

Phoenix_295_121
0.829065

Phoenix_312_144
0.827933

Phoenix_279_104
0.827822

Phoenix_304_147
0.827607

(a)

QUERY IMAGE
Phoenix_269_087

Phoenix_269_087
0.0

Phoenix_269_086
3.2689

Phoenix_310_144
3.477554

Phoenix_285_110
3.495468

Phoenix_285_106
3.826309

Phoenix_268_086
3.845854

Phoenix_270_086
3.873579

Phoenix_268_087
3.886701

Phoenix_313_006
3.911739

Phoenix_310_143
3.959245

Phoenix_288_106
3.972508

Phoenix_269_091
4.001229

Phoenix_269_090
4.027533

Phoenix_270_087
4.044656

Phoenix_270_091
4.068282

Phoenix_288_078
4.086984

Phoenix_288_077
4.090363

Phoenix_316_143
4.102981

Phoenix_317_143
4.113694

Phoenix_314_008
4.182002

Phoenix_288_079
4.216927

Phoenix_298_029
4.222246

Phoenix_285_107
4.223825

Phoenix_325_000
4.236662

Phoenix_328_070
4.241334

Phoenix_277_007
4.25355

Phoenix_328_072
4.25693

Phoenix_270_090
4.259399

Phoenix_281_097
4.283555

Phoenix_292_104
4.285335

Phoenix_310_037
4.328979

Phoenix_285_108
4.330023

(b)

QUERY IMAGE
Phoenix_269_087

Phoenix_269_087
0.0

Phoenix_269_086
3.2689

Phoenix_310_144
3.477554

Phoenix_285_110
3.495468

Phoenix_285_106
3.826309

Phoenix_268_086
3.845854

Phoenix_270_086
3.873579

Phoenix_268_087
3.886701

Phoenix_313_006
3.911739

Phoenix_310_143
3.959245

Phoenix_288_106
3.972508

Phoenix_249_146
3.992456

Phoenix_269_091
4.001229

Phoenix_269_090
4.027533

Phoenix_270_087
4.044656

Phoenix_270_091
4.068282

Phoenix_288_078
4.086984

Phoenix_288_077
4.090363

Phoenix_316_143
4.102981

Phoenix_317_143
4.113694

Phoenix_314_008
4.182002

Phoenix_288_079
4.216927

Phoenix_298_029
4.222246

Phoenix_285_107
4.223825

Phoenix_325_000
4.236662

Phoenix_328_070
4.241334

Phoenix_277_007
4.25355

Phoenix_328_072
4.25693

Phoenix_270_089
4.258849

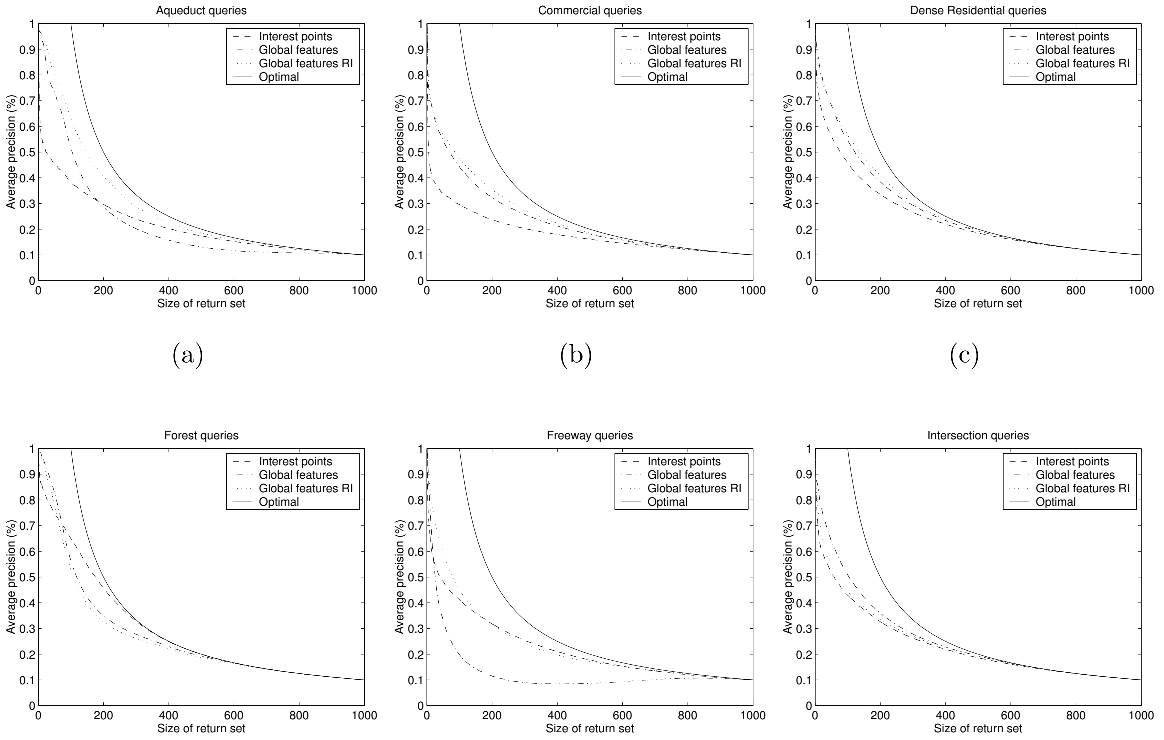
Phoenix_270_090
4.259399

Phoenix_281_097
4.283555

Phoenix_292_104
4.285335

(c)

**Figure 2.6:** Examples of similarity retrieval using the GIR demo. The query tile (top left) and the top 32 retrieved images in order of decreasing similarity for (a) interest point descriptors, (b) global Gabor texture features, and (c) global Gabor texture features using the rotation invariant similarity measure.



(a)

(b)

(c)

(d)

(e)

(f)

**Figure 2.7:** Precision as a function return set size for the three similarity retrieval methods for the groundtruth classes. (RI=rotation invariant) (a) Aqueduct. (b) Commercial. (c) Dense residential. (d) Forest. (e) Freeway. (g) Intersection. Not shown are desert chaparral (methods perform comparably), parking lot (curves are similar to forest), road (curves are similar to aqueduct), and rural residential (curves are similar to commercial).

seconds, and using the texture features with the rotation invariant distance measure took 60 seconds. This variation could be critical for supporting interactive similarity retrieval.

### Discussion

The qualitative analysis provided by the GIR demo showed the interest point descriptors support effective similarity retrieval. Retrieval results for the interest points, such as the example in figure 2.6, are rotation invariant, and when compared to the global features, are less sensitive to differences in scale. This makes sense because the interest

**Table 2.1:** *Average Normalized Modified Retrieval Rate* (ANMRR). Lower value is better.

| **Groundtruth** | **Interest pts** | **Global** | **Global RI** |
|---|---|---|---|
| Aqueduct | 0.494 | 0.417 | 0.243 |
| Commercial | 0.604 | 0.432 | 0.385 |
| Dense residential | 0.413 | 0.314 | 0.280 |
| Desert chaparral | 0.023 | 0.015 | 0.020 |
| Forest | 0.188 | 0.327 | 0.368 |
| Freeway | 0.458 | 0.761 | 0.430 |
| Intersection | 0.438 | 0.358 | 0.420 |
| Parking lot | 0.358 | 0.502 | 0.460 |
| Road | 0.637 | 0.623 | 0.485 |
| Rural residential | 0.463 | 0.413 | 0.454 |
| Average | 0.408 | 0.416 | 0.354 |

points are normalized for scale during the detection step. We also observed that they are more robust to variation in the spatial configurations of the groundtruth classes. Notice that the retrieved set for the interest point descriptors in figure 2.6 exhibits greater variability in the arrangement of the houses and streets than the retrieved sets for the global texture features.

None of the approaches was shown to clearly outperform the others in the quantitative analysis. Both the precision curves and the ANMRR values indicate that different descriptors are better for different groundtruth classes. The following general observations can be made from the precision curves in figure 2.7. The interest points descriptors have difficulty with the aqueduct, commercial, and road classes (the precision curves for the road class are not shown but are very similar in shape to those for the aqueduct class). These classes tend to be very structured which presents a challenge for the interest point descriptors. The interest point descriptors perform the best for the forest and parking lot classes (the curves for parking lot are similar to those for forest). Again, these classes exhibit less structure–forest is a stochastic rather than a regular pattern, and the parking lots vary in how full they are. The rotation invariance of the interest point descriptors makes them perform comparable to the rotation invariant global texture approach for the freeway class. Finally, the rotation sensitive global texture approach performs the best for the intersection and rural residential classes (the curves for rural residential are similar to those for forest). Due to the nature of the IKONOS images, these classes tend to be similarly oriented thus providing an advantage to an approach that exploits this.

The ANMRR values in table 2.1 are in agreement with these observations. The ANMRR averaged over all groundtruth classes indicates the rotation invariant global texture approach performs the best overall, followed by the interest points, and last is the rotation sensitive global texture features.

This work represented an initial investigation into using interest point descriptors for content-based analysis of remote sensing imagery. These features were shown to perform comparably to proven approaches to similarity retrieval, such as texture.

### 2.4.3 Further Investigation on Image Retrieval Using SIFT Features

As an extension of the work from the previous section, several methods using interest point descriptors to perform similarity retrieval are further investigated. We compare the results of using quantized versus full-length descriptors, of using different descriptor-to-descriptor distance measures, and of using different methods for comparing the sets of descriptors representing the images.

### Similarity Measures Using Full Descriptors

This section describes methods for computing the similarity between two images represented by sets of full interest point descriptors. First, we describe the comparison of single descriptors and then extend this to sets of descriptors.

**Comparing Single Descriptors** SIFT descriptors are represented by 128 dimension feature vectors. We use standard Euclidean distance to compute the similarity between two SIFT descriptors. Let $h_1$ and $h_2$ be the feature vectors representing two SIFT descriptors. The Euclidean distance between these features is then computed as

$$d_{Euc}(h_1, h_2) = \sqrt{(h_1 - h_2)^T (h_1 - h_2)}\ . \tag{2.14}$$

We also consider using the Mahalanobis distance to compare single descriptors. The Mahalanobis distance is equivalent to the Euclidean distance computed in a transformed feature space in which the dimensions (feature components) have uniform scale and are uncorrelated. The Mahalanobis distance between two feature vectors is computed as

$$d_{Mah}(h_1, h_2) = \sqrt{(h_1 - h_2)^T \Sigma^{-1} (h_1 - h_2)} \tag{2.15}$$

where $\Sigma$ is the covariance matrix of the feature vectors.

**Comparing Sets of Descriptors** Since images are represented by multiple interest point descriptors, we need a method to compute the similarity between sets of descriptors. We formulate this as a bipartite graph matching problem between a query and target graph in which the vertices are the descriptors and the edges are the distances between descriptors computed using either the Euclidean or Mahalanobis distance.

We consider two different methods for making the graph assignments. In the first method, we assign each query vertex to the target vertex with the minimum distance, allowing many-to-one matches. Let the query image contain the set of $m$ descriptors $H_q = \{h_{q1}, ..., h_{qm}\}$ and the target image contain the set of $n$ descriptors $H_t = \{h_{t1}, ..., h_{tn}\}$. Then, we define the *minimum distance measure* between the query and target image to be

$$D_{min}(Q, T) = \frac{1}{m} \sum_{i=1}^{m} dmin(h_{qi}, T) \tag{2.16}$$

where

$$dmin(h_{qi}, T) = \min_{1 \leq j \leq n} d(h_{qi}, h_{tj}) \tag{2.17}$$

and $d(\cdot, \cdot)$ is either the Euclidean or Mahalanobis distance. The factor of $1/m$ normalizes for the size of the query descriptor set.

We also consider the optimal complete (perfect) assignment between query and target vertices. In this assignment we allow a query vertex to be assigned to at most one target vertex. In the case where there are fewer target than query vertices, we allow some of the query vertices to remain unassigned. We define the *complete distance measure* between the query and target image to be

$$D_{comp}(Q, T) = \min_f \sum_{i=1}^{m} d(h_{qi}, h_{tf(i)}) \tag{2.18}$$

where $f(\cdot)$ is an assignment which provides a one-to-one mapping from $(1, ..., m)$ to $(1, ..., n)$. Again, $d(\cdot, \cdot)$ is either the Euclidean or Mahalanobis distance. In the case where $m > n$, we allow $m - n$ values not to be mapped and not contribute to the distance summation. We find the optimal mapping using the Hungarian algorithm [58] which runs in polynomial time in $m$ and $n$. Finally, we normalize for the number of descriptors by dividing the distance by $\min(m, n)$.

### Similarity Measures Using Quantized Descriptors

As an alternate to using the full 128 dimension descriptors, quantized features are also used. The 128 dimension descriptors were quantized using the $k$-means algorithm and labelled using both the Euclidean and Mahalanobis distance measures for comparisons. The quantized descriptor is defined by equation 2.12 and the similarity between a query image and a target image is defined by equation 2.13.

### Dataset

The IKONOS groundtruth dataset is used for evaluation. Each groundtruth image is represented by the following:

- A set of full interest point descriptors.

- Quantized feature counts based on clustering using Euclidean distance.
- Quantized feature counts based on clustering using Mahalanobis distance.

### Results

The retrieval performance of the different representations and similarity measures is evaluated by performing a comprehensive set of $k$-nearest neighbor similarity searches using each of the 1,000 images in the groundtruth dataset as a query. In particular, the following six methods were compared:

1. Quantized descriptors based on Euclidean clustering. Cosine distance.
2. Quantized descriptors based on Mahalanobis clustering. Cosine distance.
3. Full descriptors. Minimum distance measure using Euclidean distance.
4. Full descriptors. Minimum distance measure using Mahalanobis distance.
5. Full descriptors. Complete distance measure using Euclidean distance.
6. Full descriptors. Complete distance measure using Mahalanobis distance.

These methods will be referred to by number in the rest of this section.

Similarity retrieval using the quantized descriptors was compared for cluster counts $c$ ranging from 10 to 1000. The clustering was performed on 100,000 points selected at random from the large IKONOS images that was used for qualitative analysis in section 2.4.2. We computed the average ANMRR over the ten groundtruth classes. This was done ten times for each value of $c$ since the clustering process is not deterministic. Figure

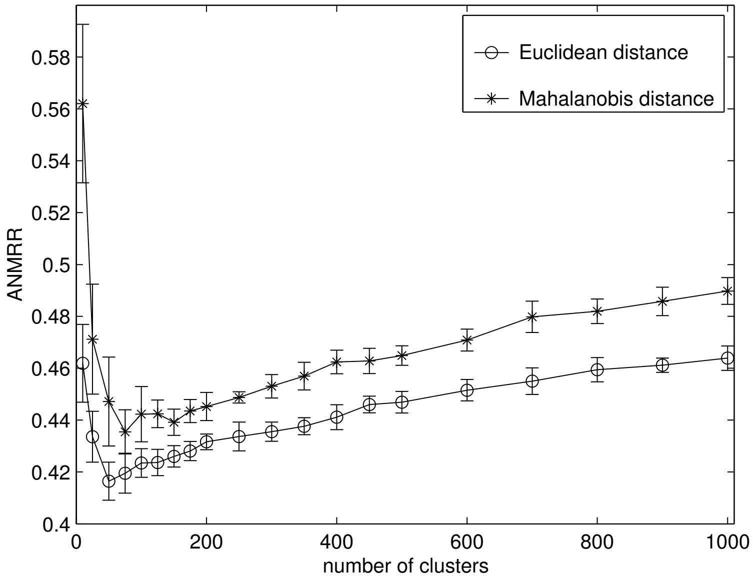


**Figure 2.8:** Retrieval performance of descriptors quantized using $k$-means clustering for different numbers of clusters $c$. Shown for clustering with Euclidean and Mahalanobis distances. Image-to-image similarity is computed using the cosine distance measure.

2.8 shows the ANMRR values for different numbers of clusters. Error bars show the first standard deviation computed over the ten trials for each $c$. Again, ANMRR values range from zero for all the groundtruth items retrieved in a result set the size of the groundtruth to one for none of the groundtruth items retrieved.

We make two conclusions from the results in Figure 2.8. One, that it is better to quantize the descriptors using Euclidean $k$-means clustering; and two, that the optimal number of clusters is 50. We use this optimal configuration in the remaining comparisons.

Figure 2.9 plots precision (the percent of correct retrievals) versus result set size for the different methods. These values are the average over all 1,000 queries. Quantized descriptors are shown to outperform full descriptors for all result set sizes. The minimum distance measure is shown to outperform the complete distance measure for comparing sets of full descriptors. Finally, as above, Euclidean distance is shown to outperform

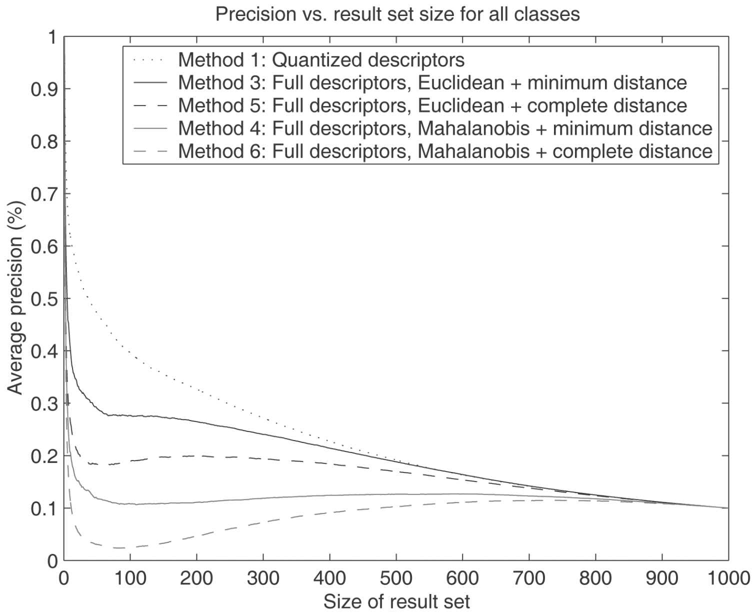


**Figure 2.9:** Retrieval performance in terms of precision versus size of result set.

Mahalanobis distance, this time when used for full descriptor-to-descriptor comparison. Table 2.2 lists the ANMRR values for the specific image categories. The values are the average over all 100 queries in each category. These results confirm that the quantized descriptors outperform the full descriptors on average. It is interesting to note, however, that no single method performs best for all categories.

**Table 2.2:** *Average Normalized Modified Retrieval Rate* (ANMRR). Lower value is better.

| **Ground-truth** | **Method 1** | **Method 3** | **Method 4** | **Method 5** | **Method 6** |
|---|---|---|---|---|---|
| Aqueduct | **0.488** | 0.655 | 0.573 | 0.621 | 0.577 |
| Commercial | **0.575** | 0.668 | 0.703 | 0.761 | 0.896 |
| Dense residential | 0.432 | **0.412** | 0.795 | 0.670 | 0.959 |
| Desert chaparral | 0.015 | **0.002** | 0.062 | 0.003 | 0.493 |
| Forest | 0.166 | **0.131** | 0.764 | 0.338 | 0.940 |
| Freeway | 0.497 | 0.384 | **0.290** | 0.401 | 0.307 |
| Intersection | **0.420** | 0.435 | 0.672 | 0.675 | 0.953 |
| Parking lot | 0.314 | 0.361 | 0.526 | **0.301** | 0.617 |
| Road | 0.680 | 0.494 | **0.417** | 0.660 | 0.680 |
| Rural residential | **0.460** | 0.592 | 0.833 | 0.706 | 0.943 |
| Average | **0.405** | 0.413 | 0.563 | 0.514 | 0.736 |

Finally, it is worth comparing the computational complexity of the different methods. On average, the 1,000 queries took approximately 2 seconds using the quantized descriptors, approximately 10 hours using the minimum distance measure for sets of full descriptors, and approximately 14 hours using the complete distance measure for sets of full descriptors. This significant difference results from the combinatorial expansion of comparing sets of descriptors and the cost of full descriptor-to-descriptor comparisons in the 128 dimension feature space. Conversely, comparing two images using quantized features only requires a single cosine distance computation. These timings were measured on a typical workstation. No distinction is made between using Euclidean and Mahalanobis distances since the latter is implemented by transforming the feature vectors before performing the queries.

### Discussion

We reach the following conclusions based on the results above. Similarity retrieval using quantized interest point descriptors is more effective and significantly more efficient than using full descriptors. This is true regardless of how the sets of full descriptors for the images are matched–minimum or complete–and how the individual descriptors are compared–Euclidean or Mahalanobis. This finding is initially a bit surprising. One might expect the loss of information from quantizing the descriptors to reduce performance. However, it seems that a binary comparison between quantized descriptors is more effective than an exact (Euclidean or Mahalanobis) comparison between full descriptors. The cosine distance can be viewed as comparing sets of descriptors in which individual descriptors are matched if they are quantized to the same cluster. The exact distance between descriptors does not matter, only that they are in some sense closer to each other than they are to other descriptors. This actually agrees with how interest point descriptors are used to determine correspondences between stereo pairs [56]. It is not the exact distance between a pair of descriptors that is used to assign a point in one image to a point in another but the ratio of this distance to that of the next closest point.

We showed that the optimal number of clusters used to quantize the descriptors seems to be around 50. This is lower than we expected. Other researchers [26] found that a much larger number of clusters, on the order of thousands, performed better for matching objects in videos. While our application is different it would be interesting to investigate this further. This finding is significant as a coarser quantization supports higher scalability since it results in reduced feature representation and faster similarity comparison.

We found that using the Euclidean distance to compare descriptors is better than the Mahalanobis distance. This is true for using $k$-means clustering to construct the quantization space. It is also true for computing individual descriptor-to-descriptor distances when comparing sets of full descriptors. This results from the distribution of the descriptors in the 128 dimension space. This again differs from the findings of other researchers [26] who used the Mahalanobis distance to cluster descriptors. It is not clear, however, if the Euclidean distance was considered or if it was just assumed that removing correlations and scale would improve the quantization induced by the clustering.

We discovered that when comparing sets of full descriptors, it is better to allow many-to-one matches; that is, the minimum distance measure outperformed the complete distance measure. This agrees conceptually with the superior performance of the quantized descriptors. The cosine distance used to compare quantized descriptors "allows" multiple matches.

Finally, we found that no method performed the best for all image classes. This requires additional investigation perhaps with a simpler, more homogeneous groundtruth dataset. Preliminary observations suggest that some methods are better at discriminating visually similar classes than others. In particular, the Mahalanobis distance measure seems better than the Euclidean distance measure at distinguishing the aqueduct, freeway and road classes which are very similar visually.

### 2.4.4 Image Retrieval on the USGS Dataset

In the previous section, the performance of local features for image retrieval is evaluated on the IKONOS dataset. The IKONOS dataset limits our evaluation in that: 1) the images are copyrighted and thus cannot be made available to other researchers; 2) the images are limited to certain areas and thus only a few LULC classes are present; and 3) the images were taken more than 10 years ago and thus the resolution is limited to only 1-m. The UGSS dataset can overcome these limitations in that: 1) it is publicly accessible data; 2) it allows for broader coverage and thus provides more diverse LULC classes and more images; and 3) the ground resolution is 1-foot, about one third of the IKONOS images. Due to these merits, in this section an extensive evaluation of local invariant features for image retrieval of LULC classes is performed using the USGS groundtruth dataset. We report on the effects of a number of design parameters on a bag-of-visual-words representation including saliency- versus grid-based local feature extraction, the size of the visual codebook, the clustering algorithm used to create the codebook, and the dissimilarity measure used to compare the BOVW representations. We also perform comparisons with standard features such as color and texture.

### Image Features

To perform a quantitative evaluation of retrieval performance using the USGS dataset, four image features are considered: simple statistics, global texture, SIFT, and color histogram features. The global texture and color histogram features are described in section 2.2.

**Simple Statistics Features** A two dimensional feature vector is computed for each image consisting of the mean and standard deviation of the grayscale values:

$$f_{SS} = (\mu, \sigma) \ .$$

This is referred to as the simple statistics feature and serves as a baseline for the experiments.

**SIFT Features** 128 dimensional local invariant descriptors are extracted for each groundtruth image using the SIFT descriptor algorithm. Two extraction modes are considered: saliency-based extraction using the SIFT detector and grid-based feature extraction. Saliency-based extraction results in a mean of 668 descriptors for each 256 by 256 image over all classes. The runway class tends to have the fewest descriptors per image with a mean of 218 and the forest class has the most with a mean of 1117. Figure 2.10 indicates the per class means.

Grid-based feature extraction is performed using three different grid spacings, 4-pixel, 8-pixel, and 16-pixel, which result in 3721, 961, and 256 features per image respectively.

The SIFT descriptors are quantized using codebooks resulting from applying $k$-means clustering to a large number of SIFT descriptors sampled at random from the large aerial images from which we created the groundtruth dataset. The 2100 images in the evaluation dataset represent less than four percent of the total area in the large images. So, for all practical purposes, there is no overlap between the images used to created the codebooks and the evaluation dataset. The codebooks are thus not specific to the particular images in the evaluation dataset. Further, the codebooks are not specific to the 21 classes (as they are based on SIFT features randomly sampled from the diverse large aerial images) and we would expect them to generalize to additional classes.

Codebooks are created using $k$-means for:

- A wide range of different numbers of clusters ($k$): 10, 25, 50, 75, 100, 125, 150, 175, 200, 250, 300, 350, 400, 450, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 15000, 20000.

- Different sized sets of randomly sampled points: one hundred thousand and one million.

- Different distance measures: Euclidean and Mahalanobis.

Each clustering is performed 10 times using a different set of randomly sampled points.

SIFT histogram features are calculated for each groundtruth image by using a codebook to quantize the SIFT descriptors extracted from the image. The histogram features thus range in length from 10 to 20000 components. Three versions of the histogram features are considered: 1) unnormalized SIFT histogram features which simply contain the codeword counts; 2) L1 normalized SIFT histogram features where the components are normalized to sum to one; and 3) L2 normalized SIFT histogram features where the components are normalized so the feature vectors have length one.

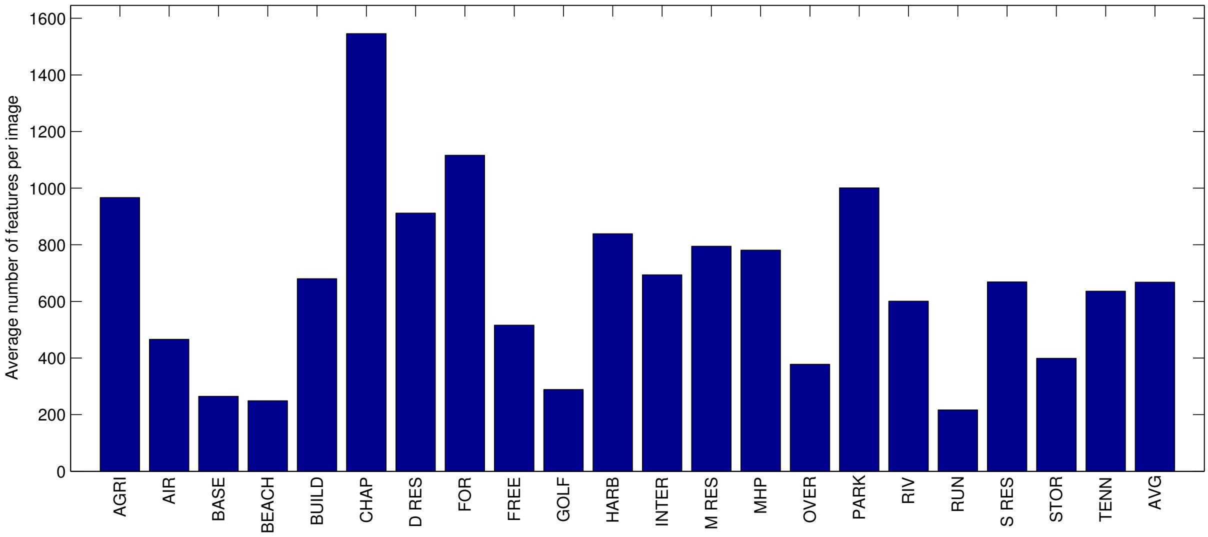


**Figure 2.10:** The average number of features per class for saliency-based local feature extraction.

### Dissimilarity Measures

Each image in the groundtruth dataset is represented by a multidimensional feature vector. This is either a two-dimensional simple statistics feature, a 60-dimensional texture feature, a 512-dimensional color histogram feature, or a $k$-dimensional SIFT histogram feature where $k$ is the size of the codebook used to quantize the SIFT descriptors. The dissimilarity measure used to compare two images depends on the type of feature. (While a few of the measures below are technically similarity measures, we refer to them as dissimilarity measures for consistency. A similarity measure can be treated as a dissimilarity measure for retrieval by simply reversing the ranking of the retrieved set.)

**Simple Statistics** The dissimilarity between two images with simple statistics features $f1$ and $f2$ is computed using the L2 or Euclidean distance:

$$d_{SS}(f1, f2) = \|f1 - f2\|_2 \; = \sqrt{(\mu 1 - \mu 2)^2 + (\sigma 1 - \sigma 2)^2} \; .$$

**Texture** The dissimilarity between two images with texture features is also computed using the L2 distance as defined in section 2.4.2. It includes both orientation (and scale) sensitive and orientation invariant dissimilarity measures.

**Color Histogram** A number of different histogram distance measures are used to compute the dissimilarity between pairs of images with respect to color histogram features, including: Bhattacharyya, chi-square, correlation, cosine, inner product, intersection, L1, L2, and Earth Mover's Distance. For two images with color histogram features $f1$ and

$f2$ of dimension $d$, the first eight of these are as follows[1]:

$$
\begin{aligned}
d_{Bhattacharyya} &= \sqrt{1 - \sum_{i=1}^{d} \frac{\sqrt{f1_i f2_i}}{\sqrt{\sum_{j=1}^{d} f1_j \sum_{k=1}^{d} f2_k}}}\ ; \\
d_{chi-square} &= \sum_{i=1}^{d} \frac{(f1_i - f2_i)^2}{f1_i + f2_i}\ ; \\
d_{correlation} &= \frac{\sum_{i=1}^{d} f1'_i f2'_i}{\sqrt{\sum_{j=1}^{d} f1'_j f2'_j}} \\
&\quad \text{where } f' = f - \frac{1}{d}\sum_{i=1}^{d} f\ ; \\
d_{cosine} &= \frac{\sum_{i=1}^{d} f1_i f2_i}{\sqrt{\sum_{j=1}^{d} f1_j^2 \sum_{k=1}^{d} f2_k^2}}\ ; \\
d_{innerproduct} &= \sum_{i=1}^{d} f1_i f2_i\ ; \\
d_{intersection} &= \sum_{i=1}^{d} \min(f1_i, f2_i)\ ; \\
d_{L1} &= \|f1 - f2\|_1 = \sum_{i=1}^{d} |f1_i - f2_i|\ ;\ \text{and} \\
d_{L2} &= \|f1 - f2\|_2 = \sqrt{\sum_{i=1}^{d} (f1_i - f2_i)^2}\ .
\end{aligned}
$$

The Earth Mover's Distance (EMD) [59] measures the distance between two distributions, in our case histograms, by viewing the distributions as "piles of dirt" and computing the cost of turning one pile into another. The cost is the amount of dirt times the distance it is moved. We consider two cases. One, the default, in which the distance between histogram

[1] Note: for clarity and simplicity, we herein list all the distances measures. However, it is obvious some of these measures will be equal when certain conditions are met. For example, when the histograms are L2 normalized, the cosine distance and the inner product distance will be the same. It is expected that some of the results in the experiments will carry the same value when the distance measures are mathematically equal.

bins is simply the Euclidean distance between the bin indices (the color histograms are three dimensional). In the other case, a cost-matrix indicates the actual distance between histogram bins in color space.

**SIFT Histogram** The same histogram distance measures are used to compare the SIFT histogram features. Only the cost-matrix version of the EMD is used as the bin indices of the SIFT histogram features provide no information on the relations between the bins. The cost-matrix is computed as the Euclidean distances between the 128 dimensional centroids corresponding to the bins.

### Results

Table 2.3 summarizes the best results for the different features considered in this study in terms of ANMRR values averaged over all 21 classes. The local invariant features are shown to perform better than the standard features. The best performance for the texture features results from using the RI measure to compare unnormalized features. The best performance for the color descriptors results from using the EMD cost matrix measure to compare HLS histograms. And, the best performance for the local descriptors results from using the L1 measure to compare L1 normalized histograms based on a codebook of 15000 words created using $k$-means clustering with the Euclidean distance and descriptors extracted using the saliency-based method. Figure 2.11 shows the precision-recall curves corresponding to these configurations. (Precision is the fraction of correct retrievals and recall is the fraction of groundtruth items retrieved for a given result set.) Precision and

**Table 2.3:** Summary of best results as measured using ANMRR values averaged over the 21 classes. See section 2.4.4 for the optimal configurations that produced these results.

| **Feature** | **ANMRR** | **Time (sec.)** |
|---|---|---|
| Simple Statistics | 0.8079 | 0.3300 |
| Texture | 0.6304 | 40.04 |
| Color Histogram | 0.7351 | 1.654E+05 |
| Local Features | *0.5914* | 193.3 |

recall are calculated as the number of retrievals is varied from 1 to 2100 and the plots are the average taken over all 2100 queries. The saliency-based local features result in higher precision for all but the lowest recall levels.

Table 2.3 also indicates how long the similarity retrievals took in seconds as an empirical comparison of the computational complexity of the different descriptors and their distance measures. The table shows the number of seconds required to perform 2100 queries in which the pairwise distance between a query and 2100 target images is computed and then used to order the result sets. These are approximate timings on a standard desktop machine and are provided for comparison purposes.

The remainder of this section describes the performance of the specific features in more detail.

**Global Texture Features** Table 2.4 shows the ANMRR values and timings of different texture feature configurations. The rotation invariant distance measure performs better than the orientation selective one. This makes sense because the image classes do not have a preferred orientation: either they do not have a distinct orientation–e.g., chaparral–or, if they do, it is not consistent–e.g., beach. The rotation invariant measure does take considerably longer as expected due to its increased computational complexity.

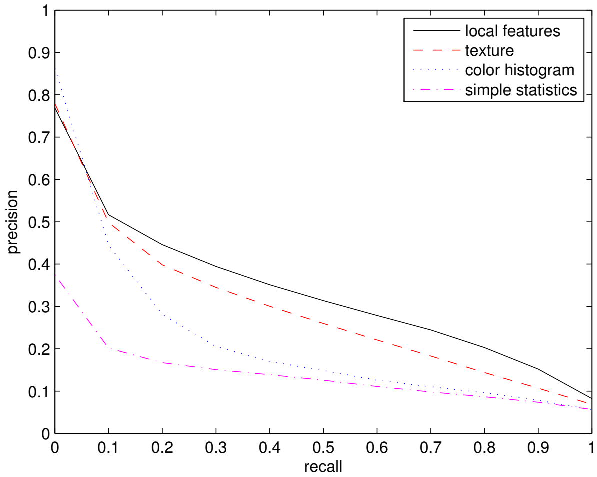


**Figure 2.11:** Precision-recall curves for the different features. Saliency-based local invariant features result in higher precision for all but the lowest recall levels.

The unnormalized texture features perform better than the L2 normalized ones. This is true for both distance measures. This is an interesting result which, to the best of our knowledge, has not been investigated or reported before. Previous applications of texture features based on Gabor filters [55] usually perform L2 normalization to account for different dynamic ranges between the feature components. However, the design of the filterbanks [8] includes scaling factors to compensate for the different regions of support and so our results indicate that further normalization suppresses discriminative frequency information and results in decreased performance.

Figure 2.13 compares the per-class performance of the optimal texture feature configuration with the optimal configurations of the other features (corresponding to table 2.3). Ignoring the baseline simple statistics, the texture features perform the best on three and the worst on one of the classes. They perform well on the beach and golf course classes but poorly on the chaparral class.

**Table 2.4:** Performance of Texture Features.

| **Features** | **Dissimilarity Measure** | **ANMRR** | **Time (sec.)** |
|---|---|---|---|
| Unnormalized | Orientation Selective | 0.6957 | 0.8500 |
| Normalized | Orientation Selective | 0.7036 | 0.8600 |
| Unnormalized | Rotation Invariant | *0.6304* | 40.04 |
| Normalized | Rotation Invariant | 0.6555 | 40.47 |

**Table 2.5:** Performance of Color Histogram Features.

| | **HLS** | | **Lab** | | **RGB** | |
|---|---|---|---|---|---|---|
| **Dissimilarity Measure** | **ANMRR** | **Time (sec.)** | **ANMRR** | **Time (sec.)** | **ANMRR** | **Time (sec.)** |
| Bhattacharyya | 0.7490 | 37.15 | 0.7433 | 37.29 | 0.7446 | 34.77 |
| Chi-Square | 0.7447 | 186.7 | *0.7405* | 134.5 | *0.7402* | 149.2 |
| Correlation | 0.7769 | 15.80 | 0.7702 | 15.68 | 0.7711 | 15.64 |
| Cosine | 0.7765 | 15.49 | 0.7700 | 15.49 | 0.7711 | 15.50 |
| EMD | 0.7371 | 1.650E+05 | 0.7414 | 5.820E+03 | 0.7453 | 1.417E+04 |
| EMD Cost Matrix | *0.7351* | 1.654E+05 | 0.7414 | 5.813E+03 | 0.7453 | 1.445E+04 |
| Inner Product | 0.8005 | 6.920 | 0.8335 | 6.880 | 0.8016 | 6.850 |
| Intersection | 0.7468 | 13.99 | 0.7455 | 14.06 | 0.7438 | 14.01 |
| L1 | 0.7468 | 5.56 | 0.7455 | 5.56 | 0.7438 | 5.57 |
| L2 | 0.7894 | 6.14 | 0.7648 | 6.13 | 0.7789 | 6.13 |

Figure 2.14 shows the “confusion matrices” for the different features. The rows indicate the query class and the columns the target classes. For each query image, we record the fraction of images in the top 100 retrievals that are in each of the 21 target classes. These values are then averaged over all 100 query images and displayed in the confusion matrices using a grayscale colorbar. For example, a value of 0.72 in row $X$ and column $Y$ indicates that on average, 72% of the top 100 retrievals had class $Y$ when class $X$ is the query image. The confusion matrix in figure 2.14(b) indicates the forest, river, and three residential classes appear very similar to the chaparral class with respect to the texture features.

**Color Histogram** Table 2.5 lists the performance of the color histogram features computed in the three different color spaces and compared using different dissimilarity measures. While the performance varies with the particular color space and measure, it is overall much worse than the texture features (compare with table 2.4) and only marginally better than the baseline simple statistics. This is not unexpected as many of the classes are spectrally similar and differ mostly spatially. The best results use the EMD distance with a cost matrix applied in the HLS color space although the increased computational complexity is evident in the timing.

Interestingly, there is no best color space. The HLS color space is optimal for the two variants of the EMD and the inner product dissimilarity measures; the Lab color space is optimal for the Bhattacharyya, correlation, cosine, and L2 dissimilarity measures; and the RGB color space is optimal for the chi-square, intersection, and L1 dissimilarity measures.

Likewise, there also is no best dissimilarity measure. The correlation, cosine, and inner product measures generally perform poorly. The chi-square measure is optimal for the Lab and RGB color spaces, and second only to the EMD measures for the HLS color space, and thus could be considered the best measure overall. The L1 measure is computationally efficient while nearly optimal. (While the intersection measure is equal to the L1 measure for L1 normalized histograms, the L1 measure is computationally more efficient since it avoids computing the minimum between feature components.) The chi-square and L1 measures thus provide a performance-efficiency tradeoff for similarity retrieval using color histogram features.

Figure 2.13 indicates the optimal color histogram configuration does not perform the best on any and performs the worst on 14 of the classes (again, ignoring the baseline simple statistics). Color histogram features perform poorly on the agricultural, freeway, and runway classes. The confusion matrix in figure 2.14(c) indicates that, with respect to color histogram features, the agricultural class appears similar to the golf course class, the freeway class appears similar to the intersection and overpass classes, and the runway class appears similar to the airport and freeway classes. This makes sense based on the sample images in figure 2.2.

**Local Features** This section first presents some of the more general observations on the performance of the local invariant features, such as whether the codebooks should be constructed using $k$-means clustering based on the Euclidean or Mahalanobis distance. It then focuses on details such as saliency- versus grid-based feature extraction, the effect of the codebook size and the choice of dissimilarity measure. Finally, the local features are compared to the other features considered in this study.

An exhaustive set of experiments is performed using all possible combinations of codebook construction, feature normalization, and dissimilarity measures. Codebooks were constructed through $k$-means clustering using either one hundred thousand or one million randomly sampled SIFT descriptors using either the Euclidean or Mahalanobis distance, and for codebook sizes ranging from 10 to 20000. Ten codebooks were created in each case by randomly sampling different sets of SIFT descriptors. Local feature histograms with dimension equal to the the size of the codebooks were computed for each image based on the unnormalized counts of the quantized features. Histograms of L1 and L2 normalized counts were also computed. Finally, dissimilarity comparison was computed using the Bhattacharyya, chi-square, correlation, cosine, inner product, intersection, L1, L2, and EMD cost matrix measures.

**Clustering Using the Euclidean vs. Mahalanobis Distance** The experiments indicate that codebooks constructed through $k$-means clustering using the Euclidean distance perform better than those constructed using the Mahalanobis distance. The Euclidean distance codebooks result in a 1-2% increase in performance on average independent of all other settings: sample size, codebook size, feature normalization, and dissimilarity measure. Thus, it appears that the correlations between dimensions in the 128 dimensional SIFT feature space as well as the difference in scales along the dimensions is important when using $k$-means clustering to construct the codebooks.

**Clustering One Hundred Thousand vs. One Million Points** Codebooks constructed by applying $k$-means clustering to a million randomly sampled SIFT descriptors similarly perform better than those constructed using only one hundred thousand descriptors. The increase is again around 1-2% on average independent of other settings and can be as high as 5%. Thus, the additional computation of applying $k$-means to larger sample sets of points is worthwhile especially since this is a preprocessing step which does not impact the cost of retrieval.

**Saliency-Based vs. Dense Extraction** The remainder of the results in this section assume codebooks constructed by applying $k$-means clustering using the Euclidean distance to one million randomly sampled SIFT descriptors.

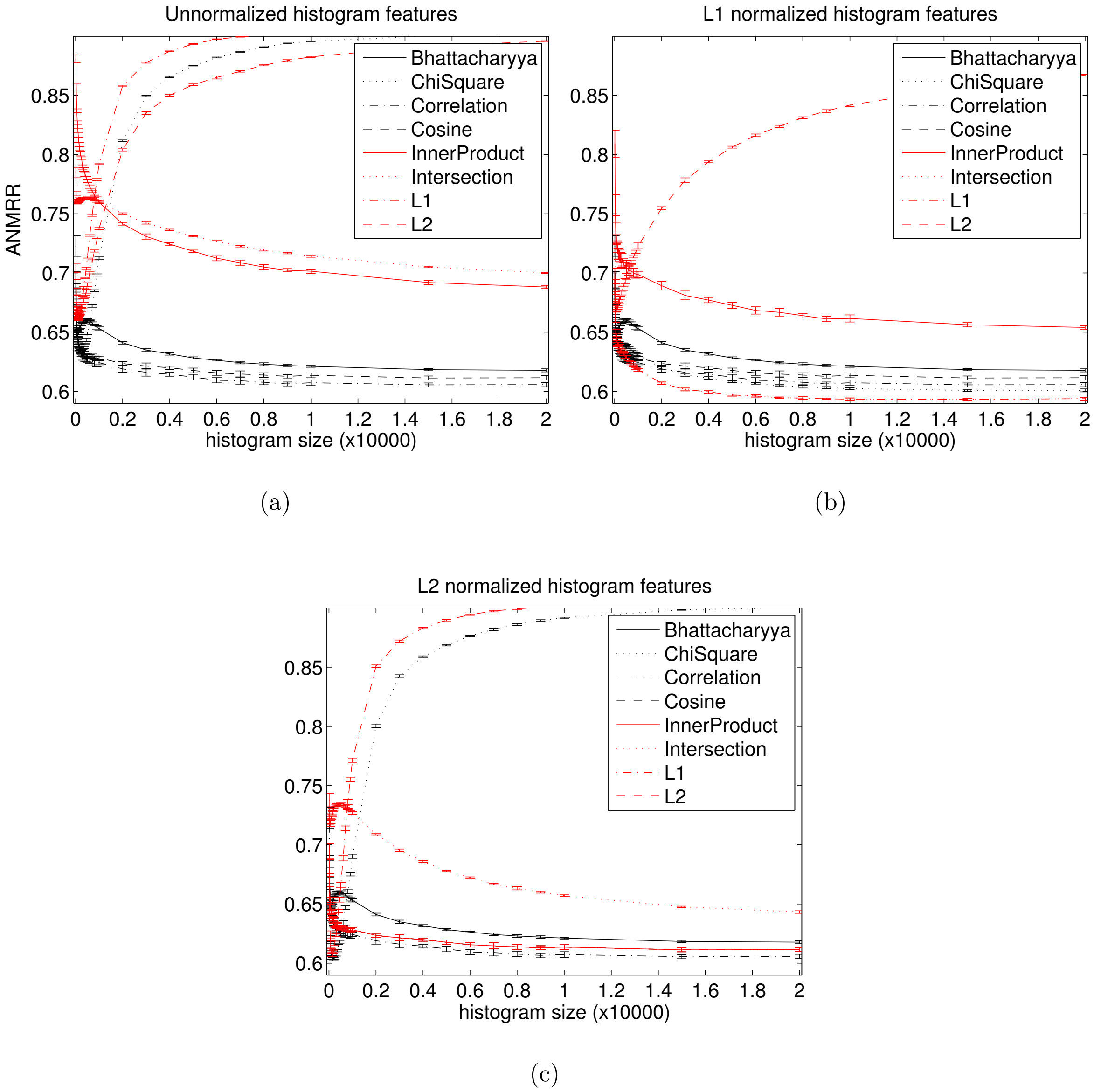
The experiments indicate that local invariant features extracted from salient image locations outperform on average those extracted on a grid. Table 2.6 compares the perfor-

**Table 2.6:** ANMRR values for saliency- and grid-based local feature extraction. Values reported correspond to the optimal codebook size and normalization scheme (not listed) for each dissimilarity measure.

| | **Feature extraction** | | | |
|---|---|---|---|---|
| **Dissimilarity Measure** | **Saliency** | **4-pixel grid** | **8-pixel grid** | **16-pixel grid** |
| Bhattacharyya | 0.6178 ± 1.254E-3 | 0.6326 ± 2.218E-3 | 0.6461 ± 1.919E-3 | 0.6688 ± 2.808E-3 |
| Chi-Square | 0.6009 ± 1.209E-3 | 0.6201 ± 2.240E-3 | 0.6262 ± 3.138E-3 | 0.6604 ± 6.505E-3 |
| Correlation | 0.6055 ± 1.635E-3 | 0.6430 ± 2.687E-3 | 0.6679 ± 3.806E-3 | 0.6964 ± 5.931E-3 |
| Cosine | 0.6113 ± 1.665E-3 | 0.6416 ± 2.045E-3 | 0.6605 ± 2.103E-3 | 0.6888 ± 1.783E-3 |
| Inner Product | 0.6113 ± 1.665E-3 | 0.6423 ± 1.814E-3 | 0.6605 ± 2.103E-3 | 0.6888 ± 1.783E-3 |
| Intersection | 0.5933 ± 9.076E-4 | 0.6201 ± 2.298E-3 | 0.6310 ± 1.998E-3 | 0.6647 ± 2.157E-3 |
| L1 | 0.5933 ± 9.076E-4 | 0.6201 ± 2.278E-3 | 0.6310 ± 1.986E-3 | 0.6644 ± 2.171E-3 |
| L2 | 0.6113 ± 1.665E-3 | 0.6423 ± 1.814E-3 | 0.6605 ± 2.103E-3 | 0.6888 ± 1.783E-3 |

mance of saliency- to grid-based feature extraction for different grid spacings. Saliency-based extraction is shown to be optimal for all dissimilarity measures. This reconfirms the benefit of extracting features at locations based on the image content that was the original motivation behind the SIFT and similar detectors. The SIFT detector is designed to identify the same object components regardless of where they appear in the image and thus is not affected by possible misalignment problems that result from using a fixed grid. Further, saliency-based extraction only considers image locations where there is meaningful image information. We note that other researchers have found the opposite, that grid-based extraction is optimal. However, this was for image classification and not retrieval and was not for geographic images. Per-class comparisons between the optimal saliency- and grid-based configuration are shown in figure 2.13. The optimal grid-based configuration was found to be using the L1 dissimilarity measure to compare unnormalized histograms created using 20000 codewords extracted from a 4-pixel grid. The best ANMRR for this configuration was 0.6179.

### Codebook Size and Dissimilarity Measure



(a)

(b)

(c)

**Figure 2.12:** The effect of codebook size on retrieval performance for different dissimilarity measures. Results are shown for (a) unnormalized histogram features, (b) L1 normalized histogram features, and (c) L2 normalized histogram features. These results are for saliency-based feature extraction.

Figure 2.12 summarizes the performance of the different dissimilarity measures by plotting ANMRR values for codebook sizes ranging from 10 to 20000. These results correspond to saliency-based feature extraction–the plots have similar shapes for grid-based feature extraction but are shifted up (worse ANMRR). Three plots are shown. Figure 2.12(a) shows the results for unnormalized histogram features, figure 2.12(b) for L1 normalized histogram features, and figure 2.12(c) for L2 normalized histogram features. Error bars indicate the standard deviation of the ANMRR values over the ten different codebooks (again, corresponding to clustering different random sets of SIFT descriptors). The results corresponding to the EMD cost matrix measure were significantly worse than any other measure and are thus not included in the detailed analysis.

The best results correspond to using the L1 measure to compare L1 normalized features for larger codebook sizes. Much more can be observed from these plots however. Interestingly, the effect of codebook size on performance generally falls into two categories. Either the performance improves with increasing size, in most cases in a monotonic fashion until gently peaking for large codebooks of size around 15000; or the performance improves sharply for small codebook sizes but then decreases steadily for larger sizes. The behavior of a specific dissimilarity measure depends on the feature normalization. These findings are significant because they show that *the range of optimal codebook sizes depends greatly on the choice of normalization and measure.* Larger codebooks tend to result in increased performance. However, there are some important cases where a narrow range of small codebook sizes is nearly optimal which is important since retrieval and storage costs are often a concern.

The best performance for unnormalized features results from the correlation measure applied to a codebook of size 15000 (see figure 2.12(a)). This performance is not significant since it is still worse than the best results corresponding to normalized features and correlation is one of the more computationally expensive measures. The performance of the L1, L2, and chi-square measures does peak for small codebook sizes but not sufficiently to make the application of these measures to unnormalized features a competitive configuration.

The best performance for L1 normalized features results from the L1 or, equivalently, intersection measure applied to a codebook of size 15000 (see figure 2.12(b)). The performance of the L2 distance does peak for small codebook sizes but again not sufficiently.

The best performance for L2 normalized features results from the chi-square measure applied to a codebook of size only 150 (see figure 2.12(c)). The chi-square performance peaks for a narrow range of small codebook sizes. The next best performance is the correlation measure applied to a codebook of size 15000. Also noteworthy is the L1 measure which also peaks for small codebook sizes and whose optimal performance for a codebook of size 150 is only slightly worse than the optimal chi-square and correlation results but is significantly more efficient.

Table 2.7 shows the optimal performance for each of the dissimilarity measures. Since each measure/codebook-size/feature-normalization combination is evaluated using ten codebooks corresponding to different random sets of SIFT descriptors, this table reports both the results from the best codebook in the column labelled “ANMRR (best)” as well as the average and standard deviation over the ten codebooks in the column labelled

**Table 2.7:** Results for different dissimilarity measures for histograms of quantized local features. These results are for saliency-based feature extraction.

| **Dissimilarity Measure** | **ANMRR (best)** | **# Codewords** | **Time (sec.)** | **ANMRR (overall)** | **Normalization** |
|---|---|---|---|---|---|
| Bhattacharyya | 0.6161 | 20000 | 1518 | 0.6178 ± 1.254E-3 | Unnormalized |
| Chi-Square | 0.5989 | 20000 | 835.1 | 0.6009 ± 1.209E-3 | L1 |
| Correlation | 0.6018 | 15000 | 439.5 | 0.6055 ± 1.635E-3 | Unnormalized |
| Cosine | 0.6070 | 15000 | 432.3 | 0.6113 ± 1.665E-3 | Unnormalized |
| Inner Product | 0.6070 | 15000 | 201.3 | 0.6113 ± 1.665E-3 | L2 |
| Intersection | *0.5914* | 15000 | 377.8 | 0.5933 ± 9.076E-4 | L1 |
| L1 | *0.5914* | 15000 | 193.3 | 0.5933 ± 9.076E-4 | L1 |
| L2 | 0.6070 | 15000 | 197.4 | 0.6113 ± 1.665E-3 | L2 |
| Chi-Square | 0.6014 | 150 | 13.43 | 0.6034 ± 1.510E-3 | L2 |
| L1 | 0.6045 | 150 | 1.770 | 0.6068 ± 1.288E-3 | L2 |

"ANMRR (overall)". The ranking of the dissimilarity measures is the same for both. Also shown in the last two rows is the results for the chi-square and L1 measures for small codebooks of size 150 of L2 normalized histogram features. These configurations perform slightly worse than the optimal configuration resulting in performance reductions of 1.7% and 2.2% respectively. However, they are significantly more efficient, requiring histogram features that are two orders of magnitude smaller and retrieval times that are one and two orders of magnitude faster respectively. They are still more effective and more efficient than the optimal configurations of the texture and color histogram features shown in table 2.3, and thus represent an excellent efficiency-performance trade off.

Figure 2.13 indicates the optimal local feature configuration for sparse-based extraction performs the best on eight and the worst on five of the classes. This configuration performs particularly well on the chaparral, harbor, and parking classes when compared to the other methods. The optimal local feature configuration for grid-based extraction performs the best on ten and the worst on one of the classes (but still performs worse than sparse-based extraction when averaged across all classes). This configuration performs particularly well on the forest class when compared to the other methods.

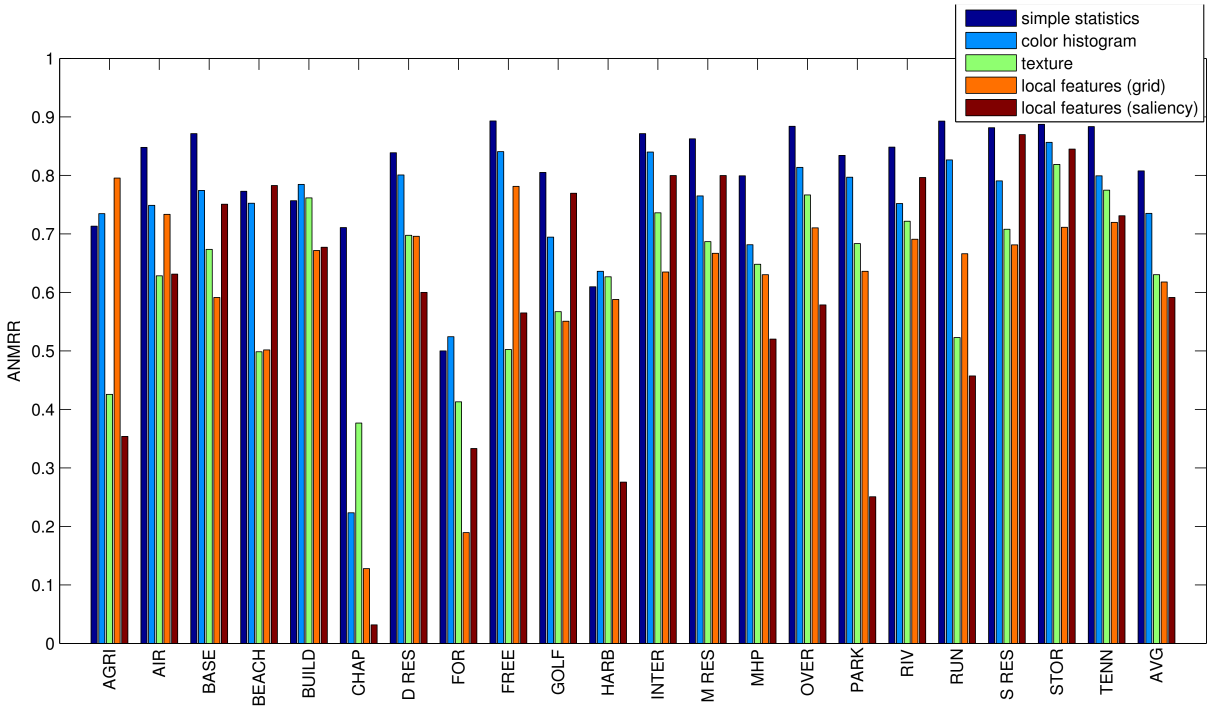
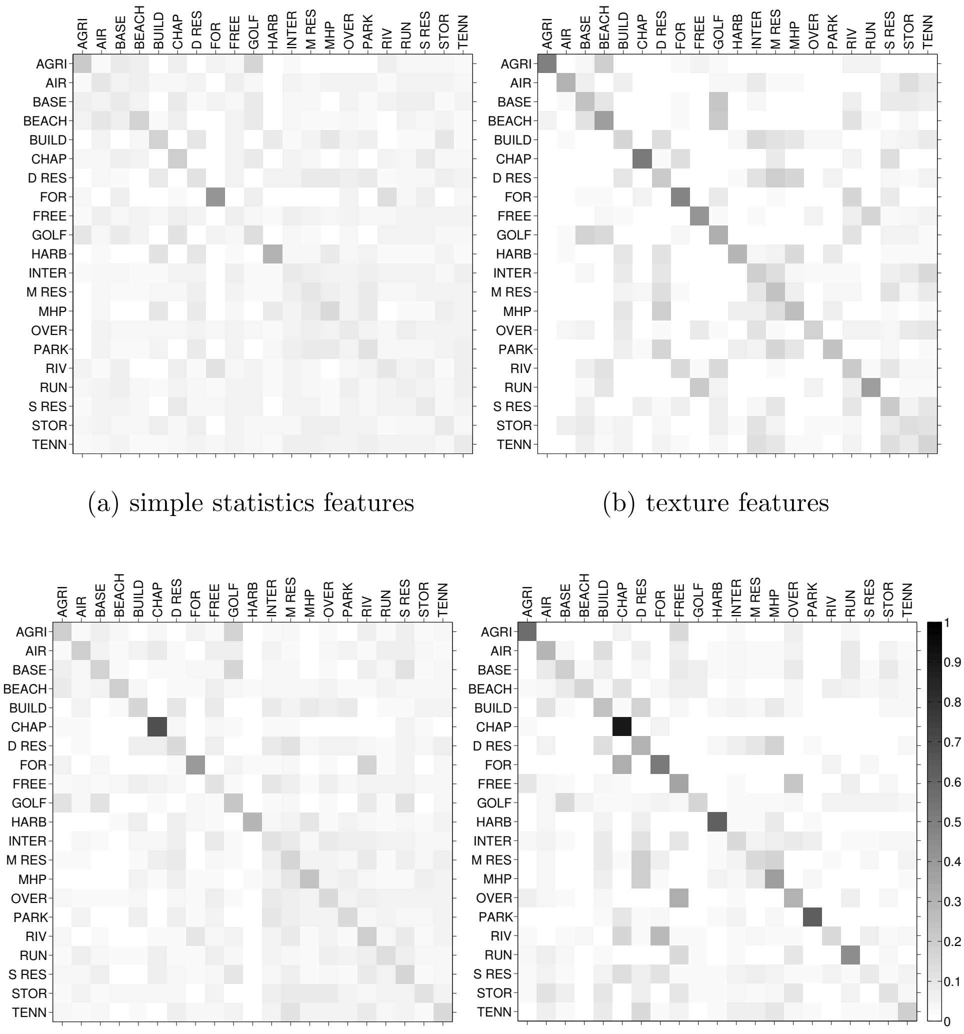


**Figure 2.13:** The per class performance corresponding to the optimal feature configurations.

Figure 2.15 presents select retrieval results corresponding to the optimal local feature configuration for sparse-based extraction. The results are shown for queries from 11 classes ordered by decreasing performance based on average ANMRR.

## Discussion

It is worthwhile discussing the performance of the local features in the context of the desirable properties described in section 1.2.2. The experimental results indicate the saliency-based local features perform well on the chaparral, dense residential, harbor, mobile home park, and parking classes. These classes are characterized by repeated occurrences of a specific object–i.e., parking lots consist of cars parked in varying spatial arrangements–which leads to hypothesize the following. The *local* property enables the



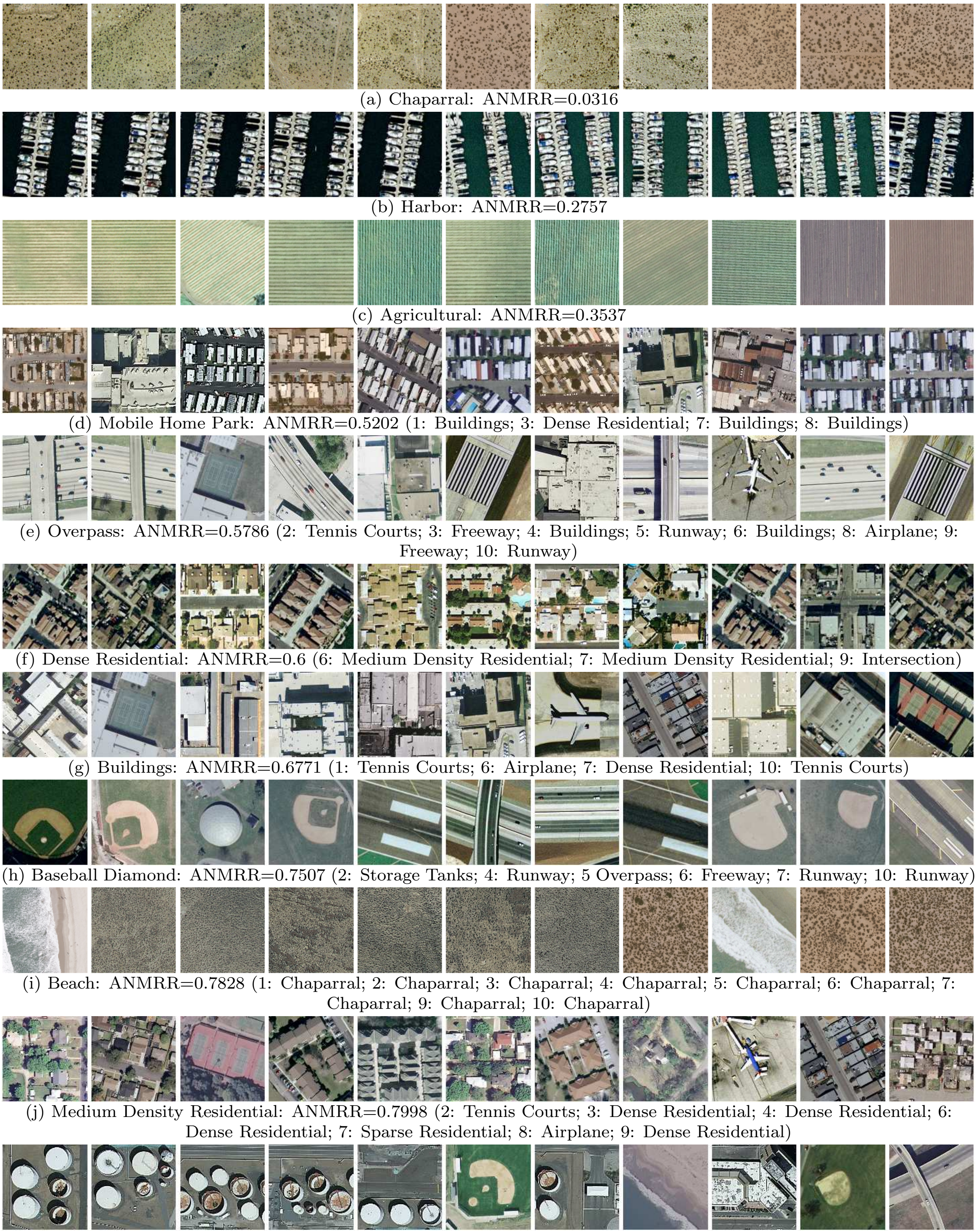
(a) simple statistics features

(b) texture features

(c) color histogram features

(d) saliency-based local features

**Figure 2.14:** Confusion matrices corresponding to (a) simple statistics features, (b) texture features, (c) color histogram features, and (d) saliency-based local features. The rows indicate the query class and the columns the target classes. For each query image, we record the fraction of images in the top 100 retrievals that are in each of the 21 target classes. These values are then averaged over all 100 query images.



(a) Chaparral: ANMRR=0.0316

(b) Harbor: ANMRR=0.2757

(c) Agricultural: ANMRR=0.3537

(d) Mobile Home Park: ANMRR=0.5202 (1: Buildings; 3: Dense Residential; 7: Buildings; 8: Buildings)

(e) Overpass: ANMRR=0.5786 (2: Tennis Courts; 3: Freeway; 4: Buildings; 5: Runway; 6: Buildings; 8: Airplane; 9: Freeway; 10: Runway)

(f) Dense Residential: ANMRR=0.6 (6: Medium Density Residential; 7: Medium Density Residential; 9: Intersection)

(g) Buildings: ANMRR=0.6771 (1: Tennis Courts; 6: Airplane; 7: Dense Residential; 10: Tennis Courts)

(h) Baseball Diamond: ANMRR=0.7507 (2: Storage Tanks; 4: Runway; 5 Overpass; 6: Freeway; 7: Runway; 10: Runway)

(i) Beach: ANMRR=0.7828 (1: Chaparral; 2: Chaparral; 3: Chaparral; 4: Chaparral; 5: Chaparral; 6: Chaparral; 7: Chaparral; 9: Chaparral; 10: Chaparral)

(j) Medium Density Residential: ANMRR=0.7998 (2: Tennis Courts; 3: Dense Residential; 4: Dense Residential; 6: Dense Residential; 7: Sparse Residential; 8: Airplane; 9: Dense Residential)

(k) Storage Tanks: ANMRR=0.8451 (5: Baseball Diamond; 7: Beach; 8: Buildings; 9: Baseball Diamond; 10: Overpass)

**Figure 2.15:** Sample retrievals for different classes corresponding to the optimal local feature configuration for sparse-based extraction. The classes are ordered in increasing average ANMRR. The leftmost image is the query image in each case and the remaining images are the top ten retrievals in decreasing order of similarity. The captions indicate the average ANMRR for the classes as well as the rank and class of the incorrect retrievals for the specific query.

features to detect the individual object instances as opposed to characterizing the gestalt or overall essence of an image. The *invariance* property enables the objects to be detected regardless of where or in what orientation they appear in an image. It also enables them to be detected independent of photometric variations such as illumination intensity and density as well as differences in color. The features are shown to be *robust* to other image variations such as the amount of the noise in the images and the quality of the different camera systems in terms of visual acuity (see, for example, the variation in the images of the parking lot class in figure 2.2). The *density* of the features allows them to be robust against occlusion caused by shadows, trees, etc. It also allows them to be robust against missed detections which is important for detecting the large number of object occurrences. The local features are *efficient* which is import for realtime application or use in large image datasets. The appropriate pairing of small codebooks and dissimilarity measures resulted in retrieval performance times that were an order of magnitude or faster than competitive features.

The saliency-based local features perform poorly on the baseball diamond, beach, golf, and runway classes. Saliency-based feature extraction results in relatively few features in these classes as indicated in figure 2.10 as they tend to contain large uniform regions. The sparseness of the resulting BOVW histograms reduces their discrimination ability.

The texture features perform well on the beach, baseball diamond, golf course, intersection, river, and the sparse and medium density residential classes. The grid-based local features also perform well on these classes. This correlation is likely due to the fact that extracting SIFT descriptors–which are after all summaries of local edge information–on a regular grid is similar to applying Gabor filters which can also be considered local edge detectors.

Finally, it is interesting that the saliency-based local features perform better than the texture features on the classes which upon first inspection appear more “texture” like. For example, the rows of cars in the parking lots or the boats in the harbor result in the kinds of regular patterns suitable for representation by frequency based texture features. However, upon closer examination, these patterns do vary at the microscopic level in that the elements are often missing, such as empty boat slips, or are arranged at different angles with respect to each other such as the parked cars. The patterns also vary at the macroscopic level in that the rows do not necessarily have the same spacing from one image to another. Saliency-based local features are more invariant to these micro- and macroscopic variations since they instead detect the individual objects or parts thereof. These irregularities can also explain why the grid-based local features perform poorly on these classes.

### Conclusion

An investigation into local invariant features for overhead image retrieval is presented. It is demonstrated that local invariant features are more effective than standard features such as color and texture for image retrieval of LULC classes in high resolution aerial imagery. We also quantitatively analyzed the effects of a number of design parameters on a BOVW representation including saliency- versus grid-based feature extraction, the size of the visual codebook, the clustering algorithm used to create the codebook, and the dissimilarity measure used to compare the BOVW representations. Such a study is timely given the increased interest by the remote sensing community in using local features for image analysis. While the focus is on image retrieval, the insights on the effects of the design parameters is informative for other applications such as detection and classification.

## 2.5 Summary

In this chapter, we compare interest point descriptors to global texture features as well as color histogram features in the context of similarity retrieval. We have shown that similarity retrieval serves as an excellent platform for evaluating the overall descriptiveness of a descriptor.

Our comparison confirms that local invariant features show great promise for remote sensing image analysis and outperform proven global texture features as well as color features in image similarity retrieval. This finding paves the way for our further investigation of the application of local invariant features to image classification and object detection in remote sensing imagery. These are considered in the following chapters of this dissertation.

# Chapter 3

# Land-use/Land-cover Classification

In the previous chapter, we explored content-based image retrieval of remote sensing imagery using local invariant features. In this chapter, we turn to image classification. A richer set of LULC classes are observable in satellite imagery than ever before due to the increased sub-meter resolution. Individual objects, such as cars and houses, are now recognizable. This chapter considers local invariant features for labelling LULC classes characterized by identifiable objects. We provide a series of experiments to empirically evaluate the effectiveness of local features on two groundtruth datasets. We investigate BOVW approaches to LULC classification and in addition, investigate how spatial information can improve the BOVW model. We first investigate the spatial pyramid matching kernel in the LULC classification which characterizes the absolute spatial arrangement of local features, and then propose a spatial co-occurrence kernel which takes into account the relative spatial arrangement of the local features. And finally, we propose a spatial pyramid co-occurrence kernel which captures both the absolute and relative spatial arrangement of the local features.

The work in this chapter was published as peer-reviewed full conference papers at the IEEE International Conference on Image Processing in 2008 [60], the SIGSPATIAL International Conference on Advances in Geographic Information Systems in 2010 [61], and the IEEE International Conference on Computer Vision in 2011 [62].

## 3.1 Comparing SIFT Descriptors and Gabor Texture Features

The first experiment is to compare SIFT features against the state-of-the-art Gabor texture features for classifying complex LULC classes using the IKONOS groundtruth dataset described in section 2.3.1.

### 3.1.1 Image Features

The SIFT and Gabor features are described in section 2.2. Again, 128 dimension local SIFT descriptors were extracted from the images. The local SIFT descriptors were quantized by assigning the label of the closest cluster center. The frequency counts of these labels form the global SIFT descriptors. Previous work in section 2.4.3 showed that 50 clusters was optimal for content-based image retrieval using quantized SIFT descriptors on the IKONOS dataset. Our global SIFT features thus have dimension 50.

Gabor texture features were extracted from the images using a filterbank tuned to five scales and six orientations. The 30 dimension feature vectors, one at each pixel location, form the local Gabor texture features. The mean and standard deviation of each filter output form the 60 dimension global Gabor texture features (see equation 2.2).

Each groundtruth image is thus represented by:

- A set of 128 dimension SIFT descriptors. Each image has 59.1 descriptors on average.
- A 50 dimension global SIFT descriptor.
- A set of 30 dimension local Gabor texture features, one at each pixel location.
- A 60 dimension global Gabor texture feature.

### 3.1.2 Classification Methods

**Maximum A Posteriori**

Image classification based on local features is performed using maximum a posteriori (MAP) classifiers. An image with the set of local features, $\mathbf{x}$, is assigned to class $c^*$ where

$$c^* = \underset{1 \leq c \leq C}{\arg\max}\, P\left(c|\mathbf{x}\right) \ . \tag{3.1}$$

The feature distributions of the classes are modelled by Gaussian mixtures so that the posterior probabilities, $p(c|\mathbf{x})$, are computed using Bayes' rule where the class-conditioned probabilities, $p(\mathbf{x}|c)$, are

$$p\left(\mathbf{x}|c\right) = \sum_{j=1}^{J} P\left(j|c\right) p\left(\mathbf{x}|j,c\right) \ . \tag{3.2}$$

The class- and mixture-conditioned probabilities for a single feature vector are

$$p\left(x|j,c\right) = \frac{1}{\left(2\pi\right)^{d/2}\left|\Sigma_j\right|^{1/2}}\, \mathrm{e}^{-\frac{1}{2}(x-\mu_j)^T\Sigma_j(x-\mu_j)} \tag{3.3}$$

where $\mu_j$ is the mean vector and $\Sigma_j$ is the covariance matrix of the $j^{th}$ mixture for class $c$. The local features are considered to be independent so the joint probability of a set of features is computed as the product of the individual probabilities. The Gaussian mixture model (GMM) parameters, $\mu_j$ and $\Sigma_j$, are learned from a training set using the expectation-maximization (EM) algorithm [63].

Design decisions for the MAP classifiers include the number of mixtures in the GMMs and the form of the covariance matrices. We investigate these in the experiments below.

### Support Vector Machines

The global features are classified using support vector machines (SVMs). When applied to classification, SVMs seek the optimal separating hyperplane between two classes, typically in a higher dimensional space than the original features. In our multi-class problem, we use a "one-against-one" strategy wherein a binary classifier is trained for each pair of classes. Unknown samples are classified using a majority voting strategy among the binary classifiers. We use the LIBSVM package [64] in the experiments below.

### 3.1.3 Results

The feature and classifier combinations are evaluated by ten-fold cross validation. The IKONOS groundtruth dataset is split into ten partitions each containing ten images from each of the ten classes. Ten rounds of training and testing are performed in which nine partitions are used for training and the remaining partition is used for testing. Each round uses a different partition for testing. A single classification rate is computed indicating the percent of the 1,000 images that are assigned to the correct class.

The MAP classifiers use the local features. A separate classifier is trained for each class. All of the local SIFT descriptors for an image are used in training and testing. Due to the large number of local Gabor texture features–4,096 for a 64-*by*-64 pixel image–only a random sampling of 100 features per image is used. Using a larger number of samples did not have a significant effect on the classification rates.

We first investigated the number and shapes of the Gaussians in the mixture models. We found that the classification rates did not vary significantly between spherical Gaussians (diagonal covariance matrix with the same value at each entry), elliptical Gaussians with axes aligned with the dimensions of the feature space (diagonal covariance matrix with possibly different values), and elliptical Gaussians with axes at any orientation (covariance matrix with possibly non-zero off diagonal entries). Since the minimum description length principle favors fewer parameters, spherical Gaussians are used in the remainder of the results. Training via the EM algorithm is also significantly faster in the spherical case.

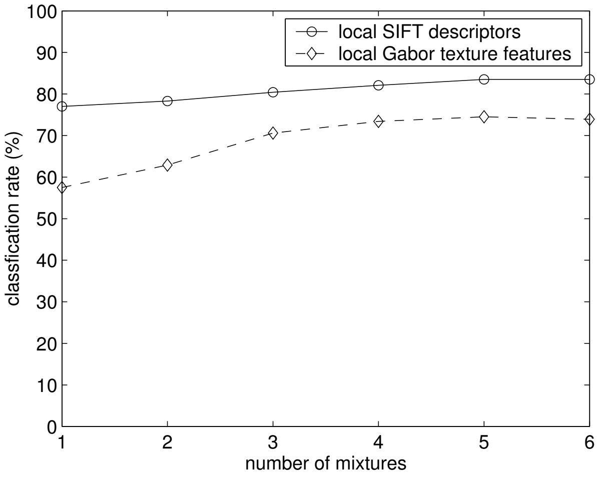


**Figure 3.1:** MAP classification rate versus number of mixtures.

Figure 3.1 plots the MAP classification rate versus the number of mixtures in the GMMs. The rate peaks at around five mixtures for both features. We therefore use five mixtures in the remainder of the results.

The SVM multi-class classifiers use the global features. One classifier is trained and tested during each round of the ten-fold cross validation. We use a linear kernel. Initial investigation into using other kernels, such as polynomial and radial bases function produced similar classification rates.

Table 3.1 shows the classification rates for four different combinations of features and classifiers: 1) local SIFT descriptors classified using MAP classifiers; 2) local Gabor texture features classified using MAP classifiers; 3) global SIFT descriptors classified using SVMs; and 4) global Gabor texture features classified using SVMs.

**Table 3.1:** Classification rates for different feature and classifier combinations.

| | **SIFT** | **Gabor** |
|---|---|---|
| **MAP** | 84.5% | 73.9% |
| **SVM** | 76.2% | 89.8% |

### 3.1.4 Discussion

The groundtruth dataset used in the experiments contains substantial within class variability, as illustrated in Figure 2.1, so classification rates nearing 90% are significant. A recent retrospective on satellite image classification reports that the average classification rate over the last 15 years is only 76.19% with a standard deviation of 15.59% [65]. Of course, the classification rate depends on the difficulty of the problem so talking about an average rate across problems is not that meaningful. Nonetheless, some of the feature and classifier combinations presented above have rates at the top of this distribution, a result underscored by the fact that only spatial information is used. Incorporating spectral information would certainly improve the results.

The best results are achieved by Gabor texture features and SVM classification. However, the next best combination, SIFT descriptors and MAP classification, perform comparably so that the order of magnitude difference in feature extraction speed could be a deciding factor especially for realtime application. These two best results are also achieved with features that are fundamentally different representations. A global Gabor texture feature is associated with a well-defined region, usually rectangular in shape. The local SIFT descriptors can represent more general regions. While both are applied here to classifying entire images, they could enable different techniques when applied to classifying regions within larger images.

## 3.2 Bag-of-visual-words Model and Spatial Extensions

The experimental results on the IKONOS groundtruth datasets build an important foundation for the work in this section. We investigate BOVW approaches to LULC classification on the USGS dataset. We also consider two spatial extensions, the established spatial pyramid match kernel which considers the absolute spatial arrangement of the local image features, as well as a spatial co-occurrence kernel we developed that considers the relative arrangement. These extensions are motivated by the importance of spatial structure in geographic data.

### 3.2.1 Bag-of-visual-words Model

The BOVW approach usually quantizes local invariant image descriptors using a visual dictionary typically constructed through a clustering algorithm. The set of visual words is then used to represent an image regardless of their spatial arrangement similar to how documents can be represented as an unordered set of words in text analysis. The quantization of the often high-dimensional local descriptors provides two important benefits: it provides further invariance to photometric image transformations, and it allows compact representation of the image such as through a histogram of visual word counts and/or efficient indexing through inverted files. The size of the visual dictionary used to quantize the descriptors controls the tradeoff between invariance/efficiency and discriminability.

A BOVW representation can be used in kernel based learning algorithms, such as nonlinear support vector machines, by computing the intersection between histograms. Given $BOVW^1$ and $BOVW^2$ corresponding to two images, the BOVW intersection kernel is computed as:

$$K_{BOVW}(BOVW^1, BOVW^2) = \sum_{m=1}^{M} \min\big(BOVW^1(m), BOVW^2(m)\big). \tag{3.4}$$

The intersection kernel is a Mercer kernel which guarantees an optimal solution to kernel-based algorithms based on convex optimization such as nonlinear support vector machines.

### 3.2.2 Spatial Extensions to BOVW

The BOVW approach does not consider the spatial locations of the visual words in an image (just as a BOW approach in text analysis does not consider where words appear in a document). Our main goal is to explore spatial extensions to the BOVW approach for LULC classification. We are motivated by the obvious fact that spatial structure is important for geographic data and its analysis. As Walter Tobler stated in his so-called first law of geography in the early 1970's, all things are related, but nearby things are more related than distant things [66]. Since our objective is LULC classification in high-resolution overhead imagery of the earth's surface, this law motivates us to consider the spatial distribution of the visual words. In particular, we consider two extensions of the BOVW representation: the spatial pyramid match kernel and the spatial co-occurrence kernel. The spatial pyramid match kernel has shown to be successful for object and scene recognition in standard (non-overhead) imagery. It considers the *absolute* spatial arrangement of the visual words. By contrast, the spatial co-occurrence kernel, which we developed, considers the *relative* spatial arrangement.

### Spatial Pyramid Match Kernel

The spatial pyramid match kernel (SPMK) was introduced by Lazebnik et al. in 2006 [1]. It is motivated by earlier work termed pyramid matching by Grauman and Darrell [67] on finding approximate correspondences between sets of points in high-dimensional feature spaces. The fundamental idea behind pyramid matching is to partition the feature space into a sequence of increasingly coarser grids and then compute a weighted sum over the number of matches that occur at each level of resolution. Two points are considered to match if they fall into the same grid cell and matched points at finer resolutions are given more weight than those at coarser resolutions. The SPMK applies this approach in the two-dimensional image space instead of a feature space; that is, it finds approximate spatial correspondence between sets of visual words in two images.

More specifically, suppose an image is partitioned into a sequence of spatial grids at resolutions $0, \ldots, L$ such that the grid at level $l$ has $2^l$ cells along each dimension for a total of $D = 4^l$ cells. Let $H_l^1$ and $H_l^2$ be the histograms of visual words of two images at resolution $l$ so that $H_l^1(i,m)$ and $H_l^2(i,m)$ are the counts of visual word $m$ contained in grid cell $i$. Then, the number of matches at level $l$ is computed as the histogram intersection:

$$I\left(H_l^1, H_l^2\right) = \sum_{i=1}^{D} \sum_{m=1}^{M} \min\left(H_l^1\left(i,m\right), H_l^2\left(i,m\right)\right). \tag{3.5}$$

Abbreviate $I\left(H_l^1, H_l^2\right)$ to $I_l$. Since the number of matches at level $l$ includes all matches at the finer level $l+1$, the number of new matches found at level $l$ is $I_l - I_{l+1}$ for $l = 0, \ldots, L-1$. Further, the weight associated with level $l$ is set to $\frac{1}{2^{L-l}}$ which is inversely proportional to the cell size and thus penalizes matches found in larger cells. Finally, the spatial pyramid match kernel for two images is given by:

$$\begin{aligned} K_L &= I_L + \sum_{l=0}^{L-1} \frac{1}{2^{L-l}} \left(I_l - I_{l+1}\right) && (3.6) \\ &= \frac{1}{2^L} I_0 + \sum_{l=1}^{L} \frac{1}{2^{L-l+1}} I_l. && (3.7) \end{aligned}$$

The SPMK is a Mercer kernel. The SPMK is summarized in Figure 3.2.

### Spatial Co-occurrence Kernel

We take further motivation from early work on pixel-level characterization of LULC classes in overhead imagery from Haralick et al.'s seminal work [68] on gray level co-occurrence matrices (GLCM) and the set of 14 derived textural features which represents some of the earliest work on image texture. A GLCM provides a straightforward way to characterize the spatial dependence of pixel values in an image. We extend this to the spatial dependence of the visual words.

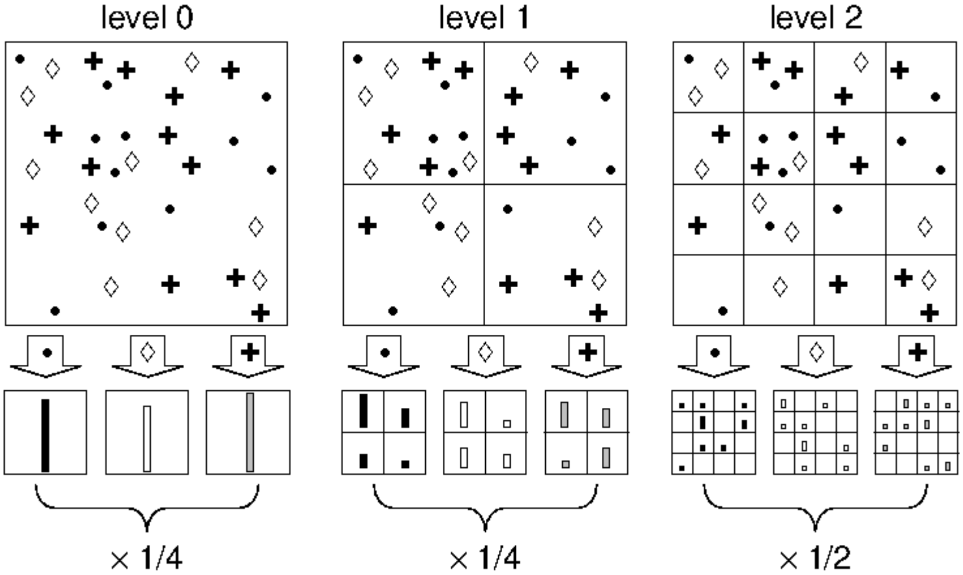


**Figure 3.2:** Toy example of a three-level spatial pyramid (adapted from [1]). The image has three visual words and is divided at three different levels of resolution. For each level, the number of words in each grid cell is counted. Finally, the spatial histogram is weighted according to equation 3.7.

More formally, given an image $I$ containing a set of $n$ visual words $c_i \in C$ at pixel locations $(x_i, y_i)$ and a binary spatial predicate $\rho$ where $c_i \rho c_j \in \{0, 1\}$, we define the visual word co-occurrence matrix (VWCM) as

$$VWCM_{\rho}(u, v) = \|(c_i, c_j)|\ (c_i = u) \wedge (c_j = v) \wedge (c_i \rho c_j)\|. \tag{3.8}$$

That is, the VWCM is a count of the number of times two visual words satisfy the spatial predicate. The choice of the predicate $\rho$ determines the nature of the spatial dependencies. While this framework provides the flexibility for variety of dependencies, we focus on proximity and, given a distance $r$, define $\rho$ to be true if the two words appear within $r$ pixels of each other:

$$c_i \rho_{prox} c_j = \begin{cases} 1, & \text{if } \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \leq r; \\ 0, & \text{otherwise.} \end{cases} \tag{3.9}$$

The VWCMs computed thus represent the number of times pairs of words appear near to each other.

A fundamental challenge to using GLCMs or our proposed counterparts, VWCMs, is their size. For example, given a visual dictionary of size $M$, the VWCM has dimension $M \times M$. Even though a symmetric predicate such proximity results in a symmetric co-occurrence matrix, its size is still quadratic with respect to the dictionary, containing $M(M+1)/2$ entries. Haralick et al. therefore defined a set of 14 scalar quantities to summarize the GLCMs. We initially also summarized our VWCMs through six commonly used scalar quantities–entropy, maximum probability, correlation, contrast, energy, and homogeneity–but this did not prove to be effective for characterizing the spatial dependencies between visual words. We thus use the full co-occurrence matrix (up to its symmetry if applicable) and instead investigate smaller dictionaries.

Given two visual co-occurrence matrices $VWCM^1_\rho$ and $VWCM^2_\rho$ corresponding to images $I^1$ and $I^2$, we now compute the spatial co-occurrence kernel (SCK) as the intersection between the matrices

$$K_{SCK}(VWCM^1_\rho, VWCM^2_\rho) = \sum_{u,v \in C} \min(VWCM^1_\rho(u,v), VWCM^2_\rho(u,v)). \tag{3.10}$$

To account for differences between images in the number of pairs of codewords satisfying the spatial predicate, the matrices are normalized to have an L1 norm of one. Note that the SCK, as an intersection of two multidimensional counts, is also Mercer kernel and thus still guarantees an optimal solution in the learning stage of non-linear support vector machines.

### SCK Combined With BOVW

While the proposed SCK can be used by itself, it can also serve as a spatial extension to the non-spatial BOVW representation. Specifically, given histograms $BOVW^1$ and $BOVW^2$ corresponding to two images, we compute the combined kernel as the sum of SCK and the intersection of the histograms

$$\begin{aligned} &K_{SCK+BOVW}(\{VWCM_\rho^1, BOVW^1\}, \{VWCM_\rho^2, BOVW^2\}) \\ &\qquad = K_{SCK}(VWCM_\rho^1, VWCM_\rho^2) + K_{BOVW}(BOVW^1, BOVW^2). \end{aligned} \tag{3.11}$$

Note that the visual dictionary used for the spatial co-occurrence matrices need not be the same as that used for the BOVW representation. We explore the effect of using different dictionaries in the experiments below. Again, since this combined kernel is a (positively weighted) sum of two Mercer kernels, it is itself a Mercer kernel. While it is possible to weight the spatial and non-spatial components of the combined kernel differently, we have so far not considered this and leave it for future work.

### 3.2.3 Experiments

We use the USGS groundtruth dataset described in section 2.3.2 for the empirical evaluation. We compare the BOVW representations with color histograms and Gabor texture features, which are described in section 2.2. Visual dictionaries of varying size are constructed by applying $k$-means clustering to over a million SIFT features randomly sampled from images disjoint from the groundtruth images. These dictionaries are then used to label SIFT features extracted from the 2,100 groundtruth images.

The approaches are compared by performing multi-class classification. Classifiers are trained on a subset of the groundtruth images and then applied to the remaining images. The classification rate is simply the percentage of the held-out images that are labelled correctly.

The premise behind our experiments is that once a classifier has been trained on a labelled dataset, it could be used to classify novel image regions. One of the benefits of a local feature based approach is that the regions need not be constrained to rectagonal or other regular shapes. Note that an image region need not be assigned one of the class labels and can instead be assigned a null-label if the classifier provides a confidence or similar score. Such is the case for support vector machines.

We use SVMs to perform the classification. Multi-class classification is implemented using a set of binary classifiers and taking the majority vote. Non-linear SVMs incorporating the kernels described above are trained using grid-search for model selection. For the histogram intersection type kernels–BOVW, SPMK, SCK, BOVW+SCK, and color histogram–the only parameter is $C$, the penalty parameter of the error term. The RBF kernel used for homogeneous texture contains an addition width parameter $\gamma$. Five-fold cross-validation is performed in which the groundtruth dataset is randomly split into five equal sets. The classifier is then trained on four of the sets and evaluated on the held-out set. The classification rate is the average over the five evaluations. Most results are presented as the average rate over all 21 classes. The SVMs were implemented using the LIBSVM package [64].

We compare a variety of difference configurations for the BOVW approaches:

- Different sized visual dictionaries for the BOVW and SPMK approaches: 10, 25, 50, 75, 100, 125, 150, 175, 200, 250, 300, 400, 500, 1000, and 5000.
- Different numbers of pyramid levels for the SPMK approach: 1 (essentially standard BOVW), 2, 3, and 4.
- Different sized visual dictionaries used to compute the co-occurrence matrices for the SCK approach: 10, 50, and 100.
- Different sized radii for the spatial predicate used to compute the co-occurrence matrices for the SCK approach: 20, 50, 100, and 150 pixels.

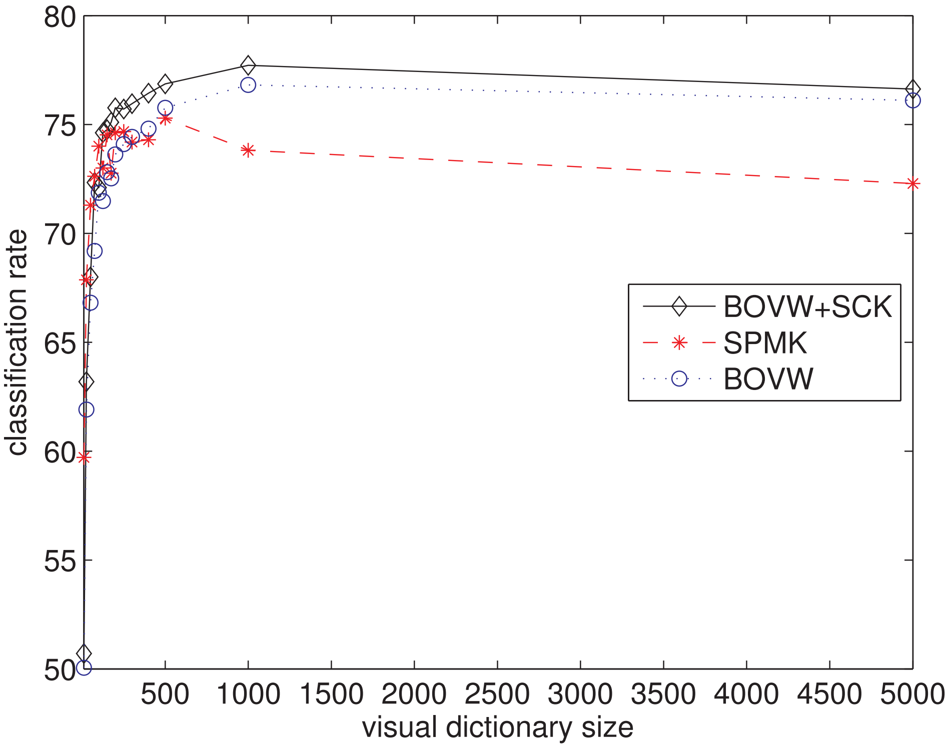


**Figure 3.3:** Comparison of BOVW, SPMK, and BOVW+SCK for different visual dictionary sizes. The size of visual dictionary used to derive the co-occurrence matrices for the SCK is as follows: 10 when the BOVW dictionary has size 10 or 25; 50 when the BOVW dictionary has size 50 or 75; and 100 otherwise. The radius used to derive the co-occurrences matrices is fixed at 150.

### 3.2.4 Results

#### Overall

Table 3.2 shows the best average classification rate across all 21 classes for the different approaches. The best rates result from the following settings: for BOVW, a dictionary size of 1000; for SPMK, a dictionary size of 500 and a three level pyramid; for SCK, a co-occurrence dictionary size of 100 and a radius of 150; and for BOVW+SCK, a BOVW dictionary size of 1000, a co-occurrence dictionary of size 100, and a radius of 150. Overall, these results are impressive given that chance classification for a 21 class problem is only

4.76%. Interestingly, the average rate is similar for all the approaches with perhaps color histograms computed in the CIE Lab colorspace as the one outlier. Section 3.2.4 below compares the per-class rates which exhibit more variation between approaches.

Color histograms computed in the HLS colorspace perform the best overall, achieving a rate of 81.19%. This was somewhat unexpected because several of the classes exhibit significant inter-image color variation. But, the color and BOVW approaches are orthogonal in that the interest point descriptors are extracted using only the luminance channel so that a combined approach is likely to perform even better. This is possible future work.

The results from the BOVW approaches–BOVW, SPMK, and BOVW+SCK–for different sized dictionaries are compared visually in Figure 3.3 and in tabular form in table 3.3. The SPMK performs best for smaller dictionaries with 150 visual words or less. Smaller dictionaries correspond to a coarser quantization of the interest point feature space–that is, each visual word is less discriminative–so that the spatial arrangement of words is more important. But, as the dictionary size increases, the absolute spatial representation of SPMK actually leads to decreased performance over the non-spatial BOVW approach.

Significantly, BOVW+SCK outperforms BOVW for all dictionary sizes indicating that the relative spatial representation of SCK is complementary to the non-spatial information of BOVW. Therefore, a fundamental conclusion is that the SCK extension improves the BOVW approach for LULC classification. This is particulary true for smaller dictionary sizes which is significant from a computational viewpoint since the increased performance provided by the extension is almost equal to that which results from an order-of-magnitude increase in dictionary size for the non-spatial BOVW approach.

### BOVW

As shown in table 3.3, a larger dictionary results in improved performance for the non-spatial BOVW up to around 1000 words. Very small dictionaries do not provide sufficient discrimination even though they might appeal from a computational and storage viewpoint. That performance decreases for very large dictionaries likely indicates that the visual words are too discriminative; that is, they are no longer robust to image perturbations caused by noise, blurring, discretization, etc. This decreases the likelihood that similar image patches are labelled as the same word.

### SPMK

Figure 3.4 and table 3.4 provide further insight into the SPMK. Our results here confirm the findings of the originators of the method in that a spatial pyramid consisting of three levels tends to be optimum [1]. This remains true for dictionaries up to size 250 after which a single level pyramid which is the same as the non-spatial BOVW performs best. This indicates that the absolute spatial configuration of highly discriminative visual words is not effective for distinguishing the LULC classes.

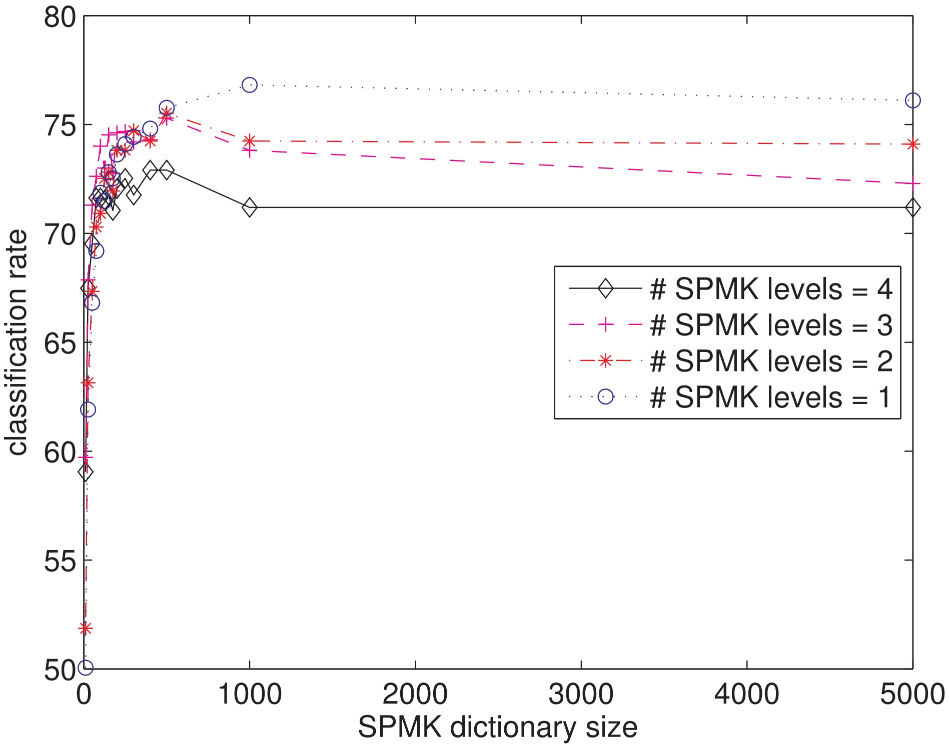


**Figure 3.4:** The effect of the number of levels used in the SPMK method.

### SCK

Figure 3.5 shows the effect of the size of the dictionary and the radius of the spatial predicate used to compute the co-occurrence matrices for the SCK. The optimal configuration is a dictionary of size 100 and a radius of 150 pixels. We did not try dictionaries larger than 100 since the SCK representation grows quadratically with the number of visual words but the plots indicate that a slightly larger dictionary should increase performance. The results for the different radii indicate that longer range spatial interactions are significant for distinguishing the LULC classes. In particular, since our dataset has a ground resolution of one foot per pixel, the co-occurrence of visual words as far apart as 100 feet is discriminating. There seems to be little improvement past 100 feet though.

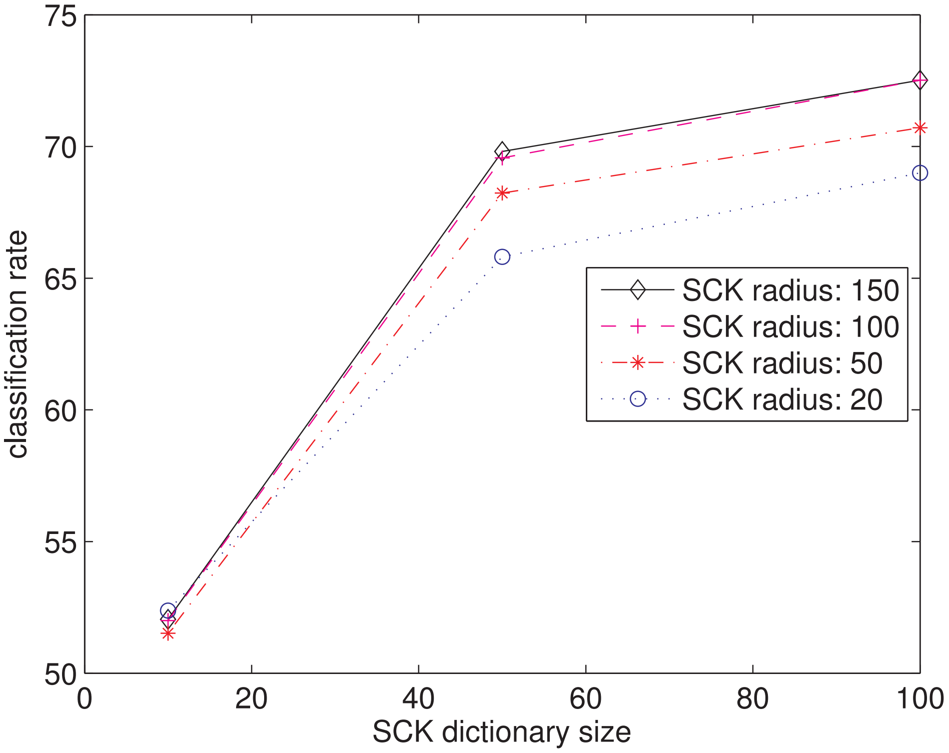


**Figure 3.5:** The effect of co-occurrence radius and dictionary size on the SCK method.

The SCK outperforms the BOVW for dictionaries of sizes 10, 50, and 100 for radii of 100 or 150 pixels (see table 3.3 for the BOVW values). It is unlikely, however, that this trend would continue for larger dictionaries (and it would be computationally and storage intensive) which motivates combining the BOVW and SCK approaches, possibly using different sized dictionaries for each.

### BOVW+SCK

Section 3.2.4 above already showed that extending BOVW with SCK results in improved performance for all BOVW dictionary sizes. We now examine the effects of the SCK co-occurrence dictionary size and predicate radius on the combined method. Figure 3.6 and table 3.5 indicate that larger co-occurrence dictionary sizes result in improved

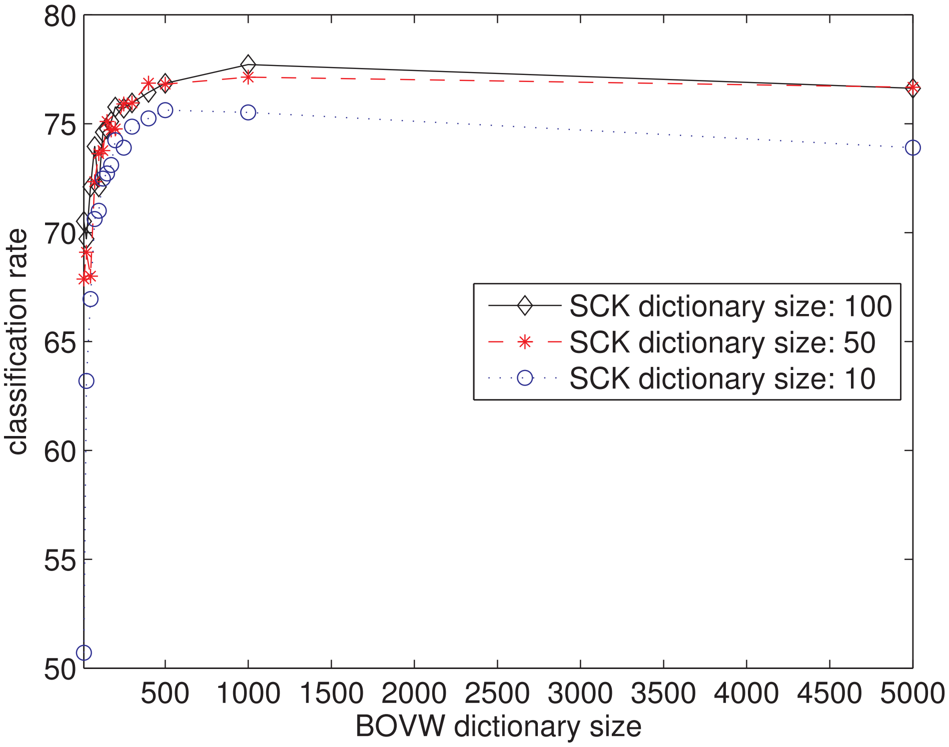


**Figure 3.6:** The effect of co-occurrence dictionary size on the BOVW+SCK method. The radius used to derive co-occurrence matrices is fixed at 150.

performance although there is no clear winner between 50 and 100 visual words. Figure 3.7 and table 3.6 indicate that a larger spatial predicate radius results in improved performance again with little difference between radii of 100 and 150 pixels. All these observations are consistent with those of the SCK only method (see section 3.2.4 above) with the slight difference that a co-occurrence dictionary size of 100 does not consistently result in improved performance over a size of 50.

### Per-Class Classification Rates

So far, we have only considered average classification rates over all classes. Figure 3.8 compares the per-class rates for the best configurations of the different methods. Not only is there significant variation between classes but there is also variation between methods

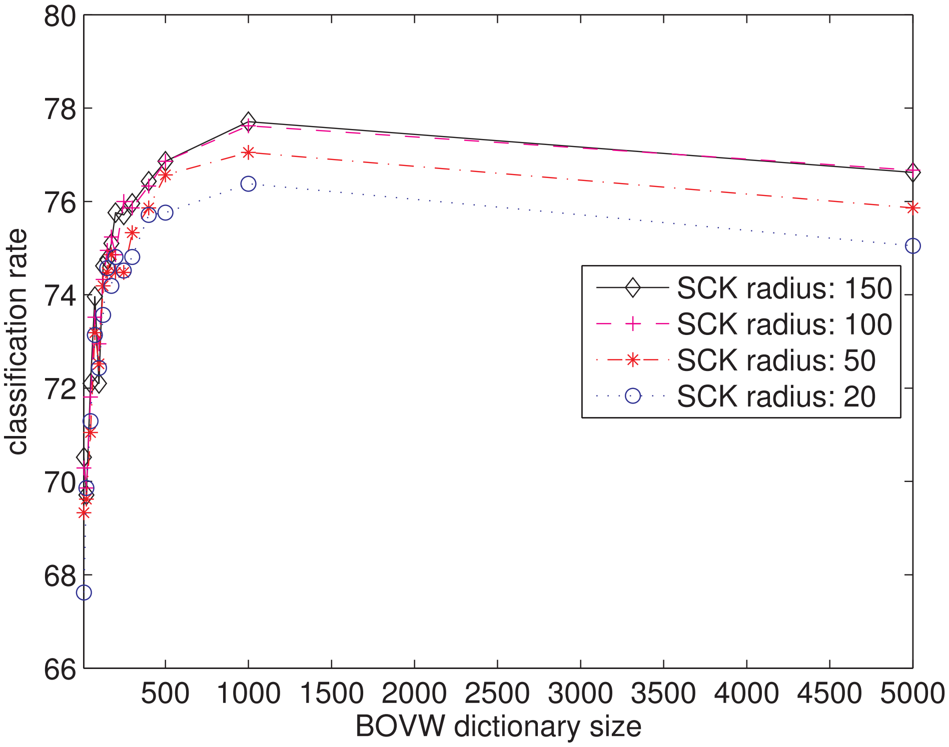


**Figure 3.7:** The effect of co-occurrence radius on the BOVW+SCK method. The size of the co-occurrence dictionary is fixed at 100.

within a class even though the methods do comparably when averaged over all classes. We summarize our observations as follows.

**Easiest Classes** Chaparral, harbor, and parking lot, and to a certain extent forest, are the classes with the highest classification rates. These classes tend to be very homogeneous with respect to both color and texture. The variation between cars does result in color histogram features performing slightly worse than the other approaches on the parking lot class.

**Most Difficult Classes** Storage tanks and tennis courts, and to a degree the three residential classes, baseball diamond, and intersection, are the classes with the lowest classification rates. Indeed, these are the classes which have the most complex spatial arrangements as well as large inter-image variation. It is for many of these classes that color histograms outperform the other approaches since, as global features, they are invariant to spatial arrangement.

**BOVW** BOVW proves to be a robust “middle-of-the-pack” approach, never significantly outperforming nor underperforming the other techniques.

**SPMK** SPMK performs better than the other visual word approaches on the beach, building, runway, and tennis courts classes. This is due to the absolute spatial arrangement of these classes being important. Even though the coastline may be oriented differently in the beach images, for any particular orientation, the sand and the surf will be in the same image regions. The same is true for the runway images. SPMK performs the worst of all methods on the freeway class. While this class is similar to runway, this result is likely due to the different locations that the vehicles occur as well as the variation in the shoulders, medians, and road widths.

**SCK** SCK performs the best of all techniques on the forest and intersection classes. These are the classes for which relative spatial arrangement is important. It additionally performs better than SPMK on the agricultural, freeway and river classes, and when compared with SPMK also restricted to dictionary of size 100 (results not shown), this list also includes buildings, golf course, harbor, parking lot, and runway. These are the classes for which relative spatial arrangement is more important than absolute arrangement.

**BOVW+SCK** Extending the non-spatial BOVW approach using SCK maintains the robustness of BOVW while improving results for 12 of the 21 classes. If the common underlying BOVW component is restricted to a dictionary size of 100, BOVW+SCK outperforms BOVW for 16 of the 21 classes. This improvement is most significant for the beach and intersection classes. The SCK extension does result in a notable decrease in performance for the baseball diamond class (although this decrease is marginal for the smaller BOVW dictionary). This again supports one of the fundamental conclusions that the SCK extension improves the BOVW approach for LULC classification.

**Color** As mentioned above, the performance of color histograms extracted in the HLS colorspace was somewhat unexpected. HLS histograms perform significantly better than the other methods on the baseball diamond, golf course, medium density residential, river, sparse residential, and storage tanks classes. This advantage over methods which only consider luminance is a result of the large intra-class homogeneity with respect to color. This advantage is compounded when HLS histograms are compared to the spatial methods since many of these classes have complex spatial arrangements often with large intra-class variation.

**Texture** Texture performs the best on the agricultural, airplane, freeway, and runway classes. These results are consistent with our previous use of homogeneous texture descriptors based on the outputs of Gabor filters for analyzing remote sensing imagery [45].

### 3.2.5 Conclusion

We evaluated BOVW and spatial extensions for LULC classification on the USGS dataset. While the BOVW-based approaches do not perform better overall than the best standard approach, they represent a robust alternative that is more effective for

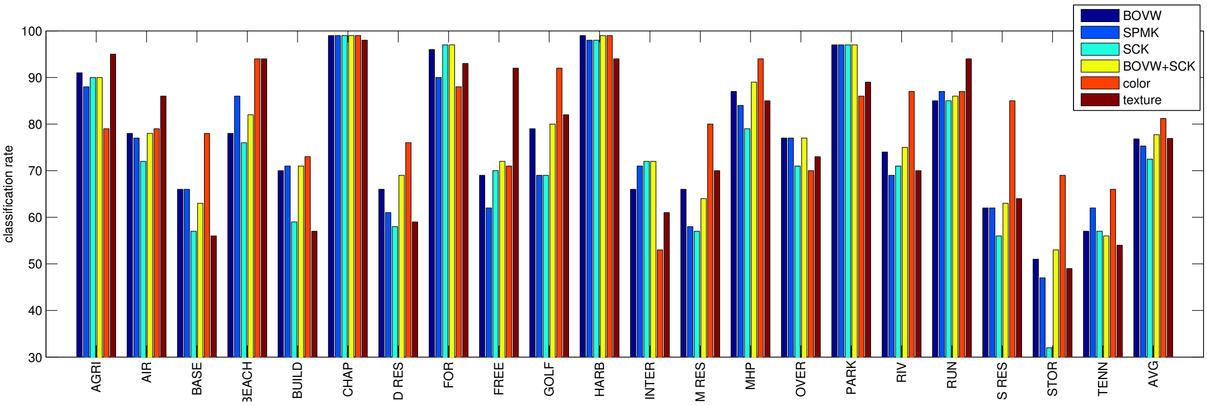


**Figure 3.8:** Per-class classification rates corresponding to the optimal configuration for each method. BOVW uses a dictionary size of 1000. SPMK uses a dictionary size of 500 with three levels. SCK uses a co-occurrence dictionary size of 100 and radius of 150. BOVW+SCK uses a BOVW dictionary size of 1000 and a co-occurrence dictionary size of 100 and radius of 150. The color histograms are computed in the HLS colorspace.

**Table 3.2:** Best classification rates for different approaches. See text for details.

| | **BOVW** | **SPMK** | **SCK** | **BOVW+SCK** | **Color–RGB** | **Color–HLS** | **Color–Lab** | **Texture** |
|---|---|---|---|---|---|---|---|---|
| **Rate** | 76.81 | 75.29 | 72.52 | 77.71 | 76.71 | 81.19 | 66.43 | 76.91 |

certain classes. We also proposed a spatial extension termed spatial co-occurrence kernel and showed that it consistently improves upon a BOVW baseline. Potential extensions of this work may include further investigation into which classes interest point based approaches are the most appropriate for and integrating interest point and color analysis since they are complementary.

**Table 3.3:** Classification rates for BOVW, SPMK, and BOVW+SCK for different visual dictionary sizes. The size of visual dictionary used to derive the co-occurrence matrices for the SCK is as follows: 10 when the BOVW dictionary has size 10 or 25; 50 when the BOVW dictionary has size 50 or 75; and 100 otherwise. The radius used to derive the co-occurrences matrices is fixed at 150.

| | **10** | **25** | **50** | **75** | **100** | **125** | **150** | **175** | **200** | **250** | **300** | **400** | **500** | **1000** | **5000** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **BOVW** | 50.05 | 61.91 | 66.81 | 69.19 | 71.86 | 71.48 | 72.81 | 72.52 | 73.62 | 74.10 | 74.43 | 74.81 | 75.76 | 76.81 | 76.10 |
| **SPMK** | 59.71 | 67.86 | 71.29 | 72.62 | 74.00 | 73.00 | 74.52 | 72.76 | 74.62 | 74.67 | 74.19 | 74.29 | 75.29 | 73.81 | 72.29 |
| **BOVW+SCK** | 50.71 | 63.19 | 68.00 | 72.33 | 72.10 | 74.62 | 74.76 | 75.10 | 75.76 | 75.71 | 75.95 | 76.43 | 76.86 | 77.71 | 76.62 |

**Table 3.4:** The effect of the number of levels used in the SPMK method.

| | **10** | **25** | **50** | **75** | **100** | **125** | **150** | **175** | **200** | **250** | **300** | **400** | **500** | **1000** | **5000** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 50.05 | 61.91 | 66.81 | 69.19 | 71.86 | 71.48 | 72.81 | 72.52 | 73.62 | 74.10 | 74.43 | 74.81 | 75.76 | 76.81 | 76.10 |
| **2** | 51.86 | 63.14 | 67.33 | 70.29 | 70.91 | 72.52 | 72.81 | 71.95 | 73.81 | 73.81 | 74.71 | 74.24 | 75.52 | 74.24 | 74.10 |
| **3** | 59.71 | 67.86 | 71.29 | 72.62 | 74.00 | 73.00 | 74.52 | 72.76 | 74.62 | 74.67 | 74.19 | 74.29 | 75.29 | 73.81 | 72.29 |
| **4** | 59.05 | 67.48 | 69.52 | 71.62 | 71.62 | 71.52 | 71.76 | 71.05 | 72.05 | 72.52 | 71.76 | 72.90 | 72.90 | 71.19 | 71.19 |

**Table 3.5:** The effect of co-occurrence dictionary size on the BOVW+SCK method. The rows correspond to the size of the dictionary used to derive the co-occurrence matrices. The columns correspond to the size of the dictionary for the BOVW component. The radius used to derive the co-occurrence matrices is fixed at 150.

| | **10** | **25** | **50** | **75** | **100** | **125** | **150** | **175** | **200** | **250** | **300** | **400** | **500** | **1000** | **5000** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **10** | 50.71 | 63.19 | 66.95 | 70.62 | 71.00 | 72.48 | 72.71 | 73.10 | 74.24 | 73.90 | 74.86 | 75.24 | 75.62 | 75.52 | 73.90 |
| **50** | 67.86 | 69.10 | 68.00 | 72.33 | 73.62 | 73.76 | 75.10 | 74.71 | 74.76 | 75.90 | 75.95 | 76.86 | 76.81 | 77.14 | 76.67 |
| **100** | 70.52 | 69.71 | 72.10 | 73.95 | 72.10 | 74.61 | 74.76 | 75.10 | 75.76 | 75.71 | 75.95 | 76.43 | 76.85 | 77.71 | 76.62 |

**Table 3.6:** The effect of co-occurrence radius on the BOVW+SCK method. The rows correspond to the radius used to derive the co-occurrence matrices. The columns correspond to the size of the dictionary for the BOVW component. The size of the co-occurrence dictionary is fixed at 100.

| | **10** | **25** | **50** | **75** | **100** | **125** | **150** | **175** | **200** | **250** | **300** | **400** | **500** | **1000** | **5000** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **20** | 67.62 | 69.86 | 71.29 | 73.14 | 72.43 | 73.57 | 74.57 | 74.19 | 74.81 | 74.52 | 74.81 | 75.71 | 75.76 | 76.38 | 75.05 |
| **50** | 69.33 | 69.62 | 71.05 | 73.19 | 72.52 | 74.19 | 74.48 | 74.86 | 74.48 | 74.48 | 75.33 | 75.86 | 76.57 | 77.05 | 75.86 |
| **100** | 70.29 | 69.86 | 71.81 | 73.52 | 72.95 | 74.33 | 74.95 | 75.24 | 74.86 | 76.00 | 75.86 | 76.33 | 76.86 | 77.62 | 76.67 |
| **150** | 70.52 | 69.71 | 72.10 | 73.96 | 72.10 | 74.62 | 74.76 | 75.10 | 75.76 | 75.71 | 75.96 | 76.43 | 76.86 | 77.71 | 76.62 |

## 3.3 Spatial Pyramid Co-occurrence

This section describes a novel image representation termed *spatial pyramid co-occurrence* which characterizes both the relative and absolute spatial arrangement of the local features. Specifically, the co-occurrences of visual words–quantized local invariant features–are computed with respect to spatial predicates over a hierarchical spatial partitioning of an image. The local co-occurrences combined with the global partitioning allows the proposed approach to capture both the relative and absolute layout of an image.

We are motivated again by the problem of analyzing overhead imagery. This imagery generally does not have an absolute reference frame and thus the relative spatial arrangement of the image elements often becomes the key discriminating feature. We evaluate our approach using the USGS groundtruth dataset. Our approach is shown to result in higher classification rates on the USGS dataset than a non-spatial bag-of-visual-words approach as well as a popular approach for characterizing the absolute spatial arrangement of visual words, the spatial pyramid representation.

### 3.3.1 Related Work

The major drawback of the BOVW model is that it discards all the spatial information of local features. The spatial pyramid representation [1] which characterizes the absolute location of the visual words was one of the first works to address the lack of spatial information in the BOVW representation.

Saverese et al. [69] propose a model which instead characterizes the relative locations. Motivated by earlier work on using correlograms of quantized colors for indexing and classifying images [70], they use correlograms of visual words to model the spatial correlations between quantized local descriptors. The correlograms are three dimensional structures which in essence record the number of times two visual words appear at a particular distance from each other. Correlogram elements corresponding to a particular pair of words are quantized to form correlations. Finally, images are represented as histograms of correlations and classified using nearest neighbor search against exemplar images. One challenge of this approach is that the quantization of correlograms to correlations can discard the identities of associated visual word pairs and thus may diminish the discriminability of the local image features.

Ling and Soatto [3] also characterize the relative locations of visual words. Their proximity distribution representation is a three dimensional structure which records the number of times a visual word appears within a particular number of nearest neighbors of another word. It thus captures the distances between words based on ranking and not absolute units. A corresponding proximity distribution kernel is used for classification in a SVMs framework. However, since proximity kernels are applied to the whole image, distinctive local spatial distributions of visual words may be overshadowed by global distributions.

Liu et al. [71] extend the BOVW framework by calculating spatial histograms where the co-occurrences of local features are calculated in circular regions of varying distances. However, the spatial histograms are only extracted for select visual words determined through an additional feature selection algorithm. Also, the spatial histograms are generated by averaging the counts of co-occurrences throughout the entire image and thus may also fail to capture distinctive local spatial arrangements.

Researchers have also proposed shape based approaches [72][73][74] to characterizing the spatial arrangement of visual words. Object boundaries play an important role here. They rely heavily on the shapes while do not take into account the appearances of objects. Yuan et al [75] select less ambiguous co-occurrence of visual words but require exhaustive search due to spatial overlaps.

Our proposed spatial pyramid co-occurrence differs from the above approaches in the following ways:

- It characterizes both the absolute and relative spatial layout of an image.
- It can characterize a greater variety of local spatial arrangements through the underlying spatial predicate. For example, combined proximity and orientation predicates can capture the general spatial distribution of visual words as well as the shape of local regions.
- The approach is simple in that it does not require learning a complex model.
- The representation can be easily combined with other representations such as a non-spatial bag-of-visual-words. And, since the representations are fused late, visual dictionaries of different sizes can be used for the spatial (co-occurrence) and non-spatial components of the combined representation. This allows the non-spatial component to leverage the increased discriminability of the larger dictionary while

limiting the computational costs associated with storing and comparing the co-occurrence structures.

### 3.3.2 Methods

#### Spatial Co-occurrence Revisited and Combining Multiple Spatial Predicates

The spatial pyramid co-occurrence builds upon the spatial co-occurrence described in section 3.2.2. Given two visual co-occurrence matrices $VWCM1_{\rho}$ and $VWCM2_{\rho}$ corresponding to two images, the SCK is defined by equation 3.10.

Multiple binary spatial predicates can be combined as follows. Given co-occurrence matrices $VWCM1_{\rho_A}(u,v)$ and $VWCM2_{\rho_A}(u,v)$ corresponding to predicate $\rho_A$ for two images, and co-occurrence matrices $VWCM1_{\rho_B}(u,v)$ and $VWCM2_{\rho_B}(u,v)$ corresponding to predicate $\rho_B$ for the same two images, a single SCK is computed as the sum of the individual SCKs

$$K_{SCK_{\rho_A+\rho_B}} = K_{SCK_{\rho_A}}(VWCM1_{\rho_A}, VWCM2_{\rho_A}) + K_{SCK_{\rho_B}}(VWCM1_{\rho_B}, VWCM2_{\rho_B}). \tag{3.12}$$

#### Spatial Pyramid Co-occurrence Kernel

Again, as described in section 3.2.2, suppose an image is partitioned into a sequence of spatial grids at resolutions $0, \ldots, L$ such that the grid at level $l$ has $2^l$ cells along each dimension for a total of $D = 4^l$ cells. The spatial co-occurrence of visual words is then computed separately for each cell in the multiresolution grid. Specifically, given a binary spatial predicate $\rho$, compute

$$VWCM^{l}_{\rho}(k,u,v) = \|(c_i,c_j)|\ (c_i = u) \wedge (c_j = v) \wedge (c_i \rho c_j)\|$$

where the visual words $c_i$ are restricted to those in grid cell $k$ at pyramid level $l$.

A spatial pyramid co-occurrence kernel (SPCK) corresponding to the spatial pyramid co-occurrences for two images $VWCM1_{\rho}$ and $VWCM2_{\rho}$ is then computed as

$$K_{SPCK}(VWCM1_{\rho}, VWCM2_{\rho}) = \sum_{l=0}^{L} w_l \sum_{k=1}^{D} \sum_{u,v \in M} \min(VWCM1^{l}_{\rho}(k,u,v), VWCM2^{l}_{\rho}(k,u,v))$$

where the weights $w_l$ are chosen so that the sum of intersections has the same maximum achievable value for each level; e.g., $w_l = 1/4^l$. As a sum of intersections, the SPCK is a Mercer kernel.

Note that the spatial pyramid co-occurrence representation captures both the absolute and relative spatial arrangements of the visual words. The pyramid decomposition characterizes the absolute locations through the hierarchical gridding of the image and the VLCMs characterize the relative arrangements within the individual grid cells.

Multiple binary spatial predicates can again be combined by summing the SPCKs corresponding to the individual predicates.

### Extended SPCK

The SPCK and the non-spatial BOVW representations are complementary and so it is natural to consider combining them. We thus form an extended SPCK representation, termed SPCK+, as the sum of the individual kernels:

$$\begin{aligned} &K_{SPCK+}(\{VWCM1_\rho, BOVW1\}, \{VWCM2_\rho, BOVW2\}) = \\ &\qquad\qquad K_{SPCK}(VWCM1_\rho, VWCM2_\rho) + K_{BOVW}(BOVW1, BOVW2). \end{aligned}$$

Note that since the spatial and non-spatial components of our representation are fused late, *the visual dictionary used to derive the spatial co-occurrence matrices need not be the same as that used to derive the BOVW histograms.* Indeed, the experiments below show that smaller co-occurrence dictionaries are preferable for SPCK+ as the spatial predicates become more specialized. This helps reduce the computational complexity of the proposed approach.

Since SPCK and SPMK are also complementary in how they characterize spatial dependencies, we also consider a second extended SPCK representation, termed SPCK++, as the sum of the SPCK and SPMK kernels:

$$K_{SPCK++} = K_{SPCK} + K_{SPMK}. \tag{3.13}$$

### Computational Complexity

We compare the computational costs of BOVW, SMPK, and SPCK in terms of the sizes of the different representations and the operations required to evaluate the kernels.

For a dictionary of size $M$, the BOVW representation has size $M$ and evaluating the BOVW kernel requires $M$ min computations (plus $M-1$ additions). For the same sized dictionary, an SPMK representation with levels $0,\ldots,L$ has size

$$S_{SPMK} = \sum_{l=0}^{L}\sum_{k=1}^{l^4} M$$

and evaluating the SPMK kernel requires the same number of min computations. For $L=2$, $S_{SPMK} = 21M$.

A VLCM corresponding to a co-occurrence dictionary of size $N$ has $N^2$ entries (this reduces to $N(N+1)/2$ unique entries for symmetric spatial predicates such as those used in the experiments below). So, a SPCK representation with levels $0,\ldots,L$ has size

$$S_{SPCK} = \sum_{l=0}^{L}\sum_{k=1}^{l^4} N^2$$

and evaluating the SPCK kernel requires the same number of min computations. For $L=2$, $S_{SPCK} = 21N^2$. In the case where $N \leq \sqrt{M}$, the computational complexity of SPCK in terms of storage and kernel-evaluation is $O(M)$, the same as for BOVW and SPMK. This remains true when combining multiple spatial predicates. This is significant with respect to the finding in the experiments below that greatly reduced co-occurrence dictionaries are sufficient or even optimal for the extended SPCK representations.

### 3.3.3 Experiments

We evaluate the spatial pyramid co-occurrence representation using SIFT features on three datasets: 1) the USGS groundtruth dataset, 2) the publicly available Graz-01 object class evaluation dataset, and 3) the publicly available 15-Scene evaluation dataset.

**Spatial Predicates**

We consider two types of spatial predicates: *proximity* predicates which characterize the distance between pairs of visual words, and *orientation* predicates which characterize the relative orientations of pairs of visual words.

**Proximity** The first predicate is the proximity predicate defined by equation 3.9. We believe nearby things are more informative than distant things and use the VWCM corresponding to $\rho_{prox}$ indicates the number of times pairs of codewords appear within $r$ pixels of each other in a given image or region. Figure 3.9(a) shows an example of where $\rho_{prox}$ evaluates to $F$ for two words.

**Orientation** The SIFT detector provides the orientation of the interest points used to derive the visual words. We postulate that these orientations are indicative of the local shape of image regions and thus derive orientation predicates $\rho_{orien}$ which consider the relative orientations of pairs of visual words.

Given visual words $c_i$ and $c_j$ with (absolute) orientations $\theta_i$ and $\theta_j$ with respect to some canonical direction such as the $x$-axis, we define a pair of orientation predicates, one which evaluates to true when the visual words are in-phase (pointing in the same direction) and another which evaluates to true when the visual words are out-of-phase (pointing in opposite directions):

$$c_i \rho_{orien2_1} c_j = \begin{cases} T, & \text{if } cos(\theta_i - \theta_j) \geq 0; \\ F, & \text{otherwise} \end{cases}$$

and

$$c_i \rho_{orien2_2} c_j = \begin{cases} T, & \text{if } cos(\theta_i - \theta_j) < 0; \\ F, & \text{otherwise} \end{cases}$$

where $-\pi < \theta_i, \theta_j \leq \pi$. Figure 3.9(b) shows an example of where $\rho_{orien2_1}$ evaluates to $T$ and $\rho_{orien2_2}$ evaluates to $F$ for two words.

We also define a set of four orientation predicates $\rho_{orien4_{1,\ldots,4}}$ which partition the phase space into four bins. That is, the four predicates separately evaluate to true for $\{\sqrt{2}/2 \leq cos(\theta_i - \theta_j)\}$, $\{0 \leq cos(\theta_i - \theta_j) < \sqrt{2}/2\}$, $\{-\sqrt{2}/2 \leq cos(\theta_i - \theta_j) < 0\}$, and $\{cos(\theta_i - \theta_j) < -\sqrt{2}/2\}$.

We characterize the relative instead of absolute orientation of pairs of visual words since overhead imagery generally does not have an absolute reference frame.

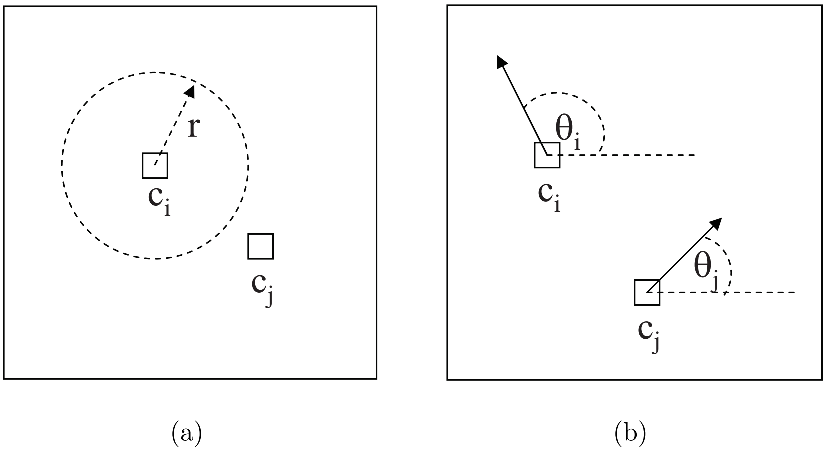


**Figure 3.9:** We consider spatial predicates which characterize (a) the distance between pairs of visual words, and (b) the relative orientation of pairs of visual words.

### Experiments On the USGS Dataset

We use SVMs to perform the classification in the same way as in section 3.2.3. We also use the same visual dictionaries derived in section 3.2.3.

We compare the following approaches:

- The "baseline" non-spatial BOVW kernel.
- The spatial pyramid match kernel [1]
- The proposed spatial pyramid co-occurrence kernel (SPCK) (sec. 3.3.2).
- The extended SPCK+ and SPCK++ representations (sec. 3.3.2).

We also compare the following configurations:

- A proximity predicate alone. This is referred to as SP1 below. We consider distances of $r$ = 20, 50, 100, and 150 pixels.

- A proximity predicate combined with orientation predicates corresponding to a two-bin phase space. This is referred to as SP2 below.
- A proximity predicate combined with orientation predicates corresponding to a four-bin phase space. This is referred to as SP3 below.
- Visual dictionary sizes of 10, 50, and 100 for the co-occurrence component of the SPCK.
- Different numbers of pyramid levels in the SPCK.

### Results On the USGS Dataset

Table 3.7 compares the best classification rates of the different approaches for the USGS dataset. A visual dictionary size of 100 is used for the BOVW and SPMK approaches as well as the BOVW and SPMK extensions to SPCK. Visual dictionary sizes of 100, 10, and 50 are used for the co-occurrence components of SPCK, SPCK+, and SPCK++. Combined proximity plus 4-bin orientation predicates (SP3) are used for SPCK, SPCK+, and SPCK++.

**Table 3.7:** Classification rates for the USGS dataset. See text for details.

| **BOVW** | **SPMK [1]** | **SPCK** | **SPCK+** | **SPCK++** |
|---|---|---|---|---|
| 71.86 | 74.00 | 73.14 | 76.05 | 77.38 |

The USGS dataset is challenging and so the improvement that the proposed SPCK+ and SPCK++ provide over SPMK is significant. In particular, *SPCK++ improves performance over SPMK by more than what SPMK itself improves over the non-spatial*

*BOVW*. And, SPCK+ provides about the same improvement. Note that since SMPK includes BOVW by construction, SPCK+ is a suitable comparison since it is simply SPMK combined with BOVW.

We pick a relatively small BOVW and SPMK dictionary size of 100 for the sake of comparison. SPCK+ and SPCK++ provide a similar improvement over BOVW and SPMK for larger dictionary sizes.

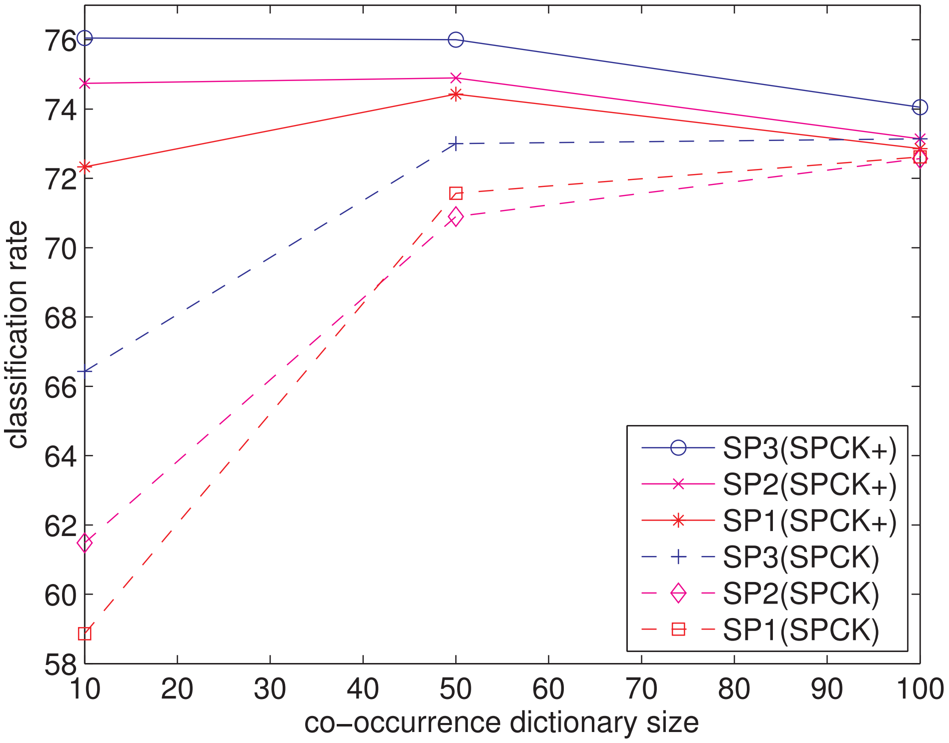


**Figure 3.10:** The effect of co-occurrence dictionary size on SPCK and SPCK+.

**Co-occurrence Dictionary Size** Table 3.8 and Figure 3.10 show the effect of co-occurrence dictionary size on SPCK and SPCK+. The significant result here is that *smaller co-occurrence dictionaries become sufficient or even optimal as the SPCK+ spatial predicates become more specialized.* In particular, a co-occurrence dictionary of just

**Table 3.8:** The effect of co-occurrence dictionary size (rows) on SPCK and SPCK+. Results are shown for the three different spatial predicate configurations (SP1=proximity only, SP2=proximity+2-bin orientation, SP3=proximity+4-bin orientation), and for the baseline SPCK and extended SPCK+.

| | SP1 | | SP2 | | SP3 | |
|---|---|---|---|---|---|---|
| | **SPCK** | **SPCK+** | **SPCK** | **SPCK+** | **SPCK** | **SPCK+** |
| **10** | 58.86 | 72.33 | 61.48 | 74.76 | 66.43 | 76.05 |
| **50** | 71.57 | 74.43 | 70.90 | 74.90 | 73.00 | 76.00 |
| **100** | 72.62 | 72.86 | 72.57 | 73.14 | 73.14 | 74.05 |

**Table 3.9:** The effect of the spatial predicate proximity distance (rows) on SPCK. Results are shown for different spatial predicate configurations and different co-occurrence dictionary sizes (columns).

| | SP1 | | | SP2 | | | SP3 | | |
|---|---|---|---|---|---|---|---|---|---|
| | **10** | **50** | **100** | **10** | **50** | **100** | **10** | **50** | **100** |
| **20** | 58.00 | 66.52 | 67.19 | 60.67 | 66.38 | 66.38 | 62.19 | 65.71 | 64.76 |
| **50** | 58.76 | 69.24 | 70.81 | 60.43 | 69.76 | 69.81 | 65.57 | 70.14 | 69.14 |
| **100** | 58.62 | 70.76 | 72.00 | 61.38 | 70.48 | 72.38 | 66.29 | 72.62 | 72.43 |
| **150** | 58.86 | 71.57 | 72.62 | 61.48 | 70.90 | 72.57 | 66.43 | 73.00 | 73.14 |

10 codewords provides better SP3 (SPCK+) performance than one with 50 or 100 codewords. This reduces the computational complexity of SPCK+ to be of the same order as SPMK.

**Proximity Distance** Table 3.9 and Figure 3.11 show the effect of the spatial predicate proximity distance ($r$ in Eq. 3.9) on SPCK. Results are shown for the three different spatial predicate configurations as well as for different co-occurrence dictionary sizes. The clear trend is that larger distances improve performance. This indicates that even long range spatial interactions between visual words is important for characterizing the USGS classes.

**Pyramid Levels** Table 3.10 and Figure 3.12 shows the effect of the number of pyramid levels on SPCK. Results are shown for just level 0, just level 1, just level 2, and for

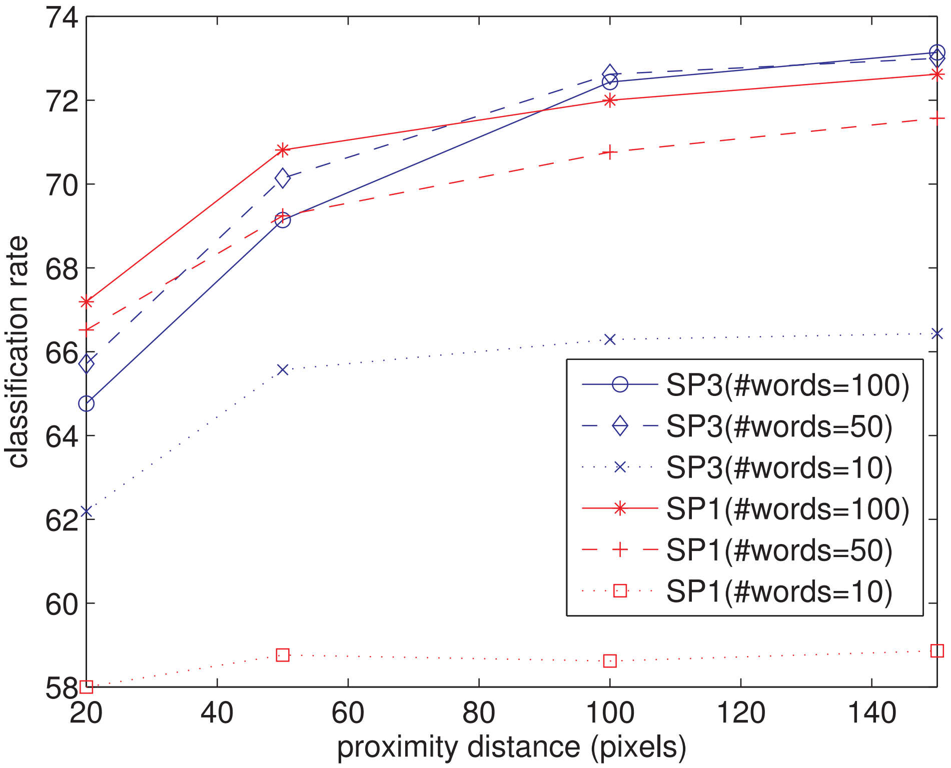


**Figure 3.11:** The effect of the spatial predicate proximity distance on SPCK. Results are shown for two different spatial predicate configurations as well as for different co-occurrence dictionary sizes.

all three levels combined. While combining all three levels usually performs best, the interesting trend is that the order of the individual levels depends on the size of the co-occurrence dictionary. In particular, level 0 performs best for a co-occurrence dictionary of size 100 while level 2 performs best for a dictionary of size 10. We will investigate this further in future work.

### Results On Generic Image Datasets

Even though our primary objective is analyzing overhead imagery, we also apply our approach to datasets of generic images. We demonstrate that our approach performs

**Table 3.10:** The effect of the number of pyramid levels on SPCK. The rows indicate just level 0, just level 1, just level 2, and all three levels combined. The columns indicate different spatial predicate configurations and co-occurrence dictionary sizes.

| | **SP1** | | | **SP2** | | | **SP3** | | |
|---|---|---|---|---|---|---|---|---|---|
| | **10** | **50** | **100** | **10** | **50** | **100** | **10** | **50** | **100** |
| **0** | 52.05 | 69.81 | 72.52 | 55.10 | 70.19 | 72.52 | 60.57 | 72.57 | 74.19 |
| **1** | 51.81 | 66.05 | 68.00 | 52.86 | 65.86 | 67.57 | 58.23 | 67.52 | 68.33 |
| **2** | 55.01 | 66.05 | 66.00 | 58.52 | 63.52 | 62.00 | 61.19 | 62.48 | 59.00 |
| **0+1+2** | 58.86 | 71.57 | 72.62 | 61.48 | 70.90 | 72.57 | 66.43 | 73.00 | 73.14 |

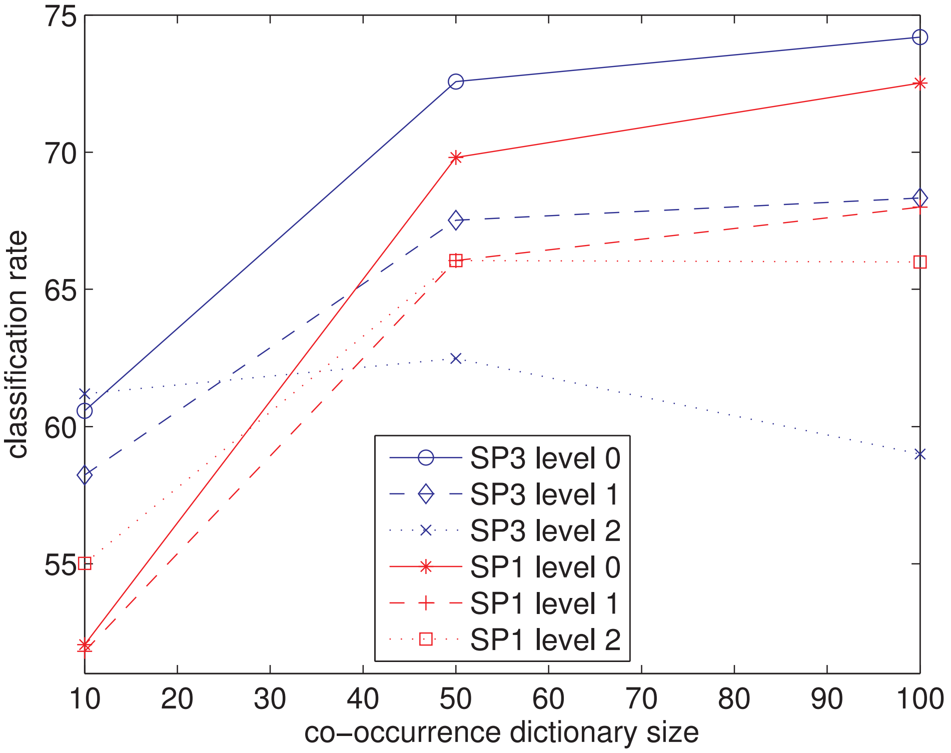


**Figure 3.12:** The performance of the individual pyramid levels on SPCK. Results are shown for two different spatial predicate configurations as well as for different co-occurrence dictionary sizes.

competitively on the Graz object class evaluation dataset and the 15-Scene evaluation dataset.

**Graz-01 Dataset** We also apply our approach to the publicly available dataset Graz-01 [2]. This dataset contains 373 images of category bike, 460 images of category person, and 270 background images as category "counter-class". All the images measure 640x480 pixels and the objects come in different scales, poses, and orientations. This dataset is challenging due to high intra-class variation and have been broadly used as an evaluation dataset in the computer vision community. We evaluate our approach using the same experimental set up as in [2]. In particular, our training set contains 100 positive images (bike or person) and 100 negative images from the other two categories, where half are from the background and half are from the other object category. Our test set consists of 100 images with a similar distribution to the training set. We report equal error rates averaged over ten runs.

Table 3.11 compares our technique with other approaches that characterize the spatial arrangement of visual words, namely the Boosting+SIFT approach of Opelt et al. [2], the SPMK approach of Lazebnik et al. [1], the proximity distribution kernel (PDK) approach of Ling and Soatto [3], and the naive Bayes nearest neighbor (NBNN) approach of Boiman et al. [4] (while NBNN is not a spatial based approach, we include it here for completeness). Our SPCK+ is shown to perform comparably to the other approaches.

**15-Scene Dataset** Finally, we apply our approach to the publicly available 15-Scene dataset [1]. This dataset contains a total of 4485 images in 15 categories varying from

**Table 3.11:** Evaluation using the Graz-01 dataset. Comparison of the proposed SPCK+ approach with Boosting+SIFT [2], SPMK [1], PDK [3], and NBNN [4].

| | **[2]** | **SPMK [1]** | **PDK [3]** | **NBNN [4]** | **SPCK+** |
|---|---|---|---|---|---|
| Bike | 86.5 | 86.3±2.5 | 90.2±2.6 | 90.0±4.3 | 91.0±4.8 |
| Person | 80.8 | 82.3±3.1 | 87.0±3.8 | 87.0±4.6 | 87.2±3.8 |

**Table 3.12:** Results on the 15-Scene dataset.

| | **SPMK [1]** | **SPCK++** |
|---|---|---|
| Classification Rate | 81.40±0.50 | 82.51±0.43 |

indoor scenes such as store, bedroom, and kitchen, to outdoor scenes such as coast, city, and forest. Each category has between 200 to 400 images and each image measures approximately 300x300 pixels. Following the same experiment setup as [1], we randomly pick 100 images per category for training and use the rest for testing. Table 3.12 compares our results with those of SPMK. We see again that SPCK++ performs comparably to SPMK.

The images in the 15 Scene dataset tend to be strongly aligned so that local spatial arrangement tends to be less important than global layout. The proposed approach thus results in only a modest improvement over SPMK (and does not beat the best published results) on this dataset since it is designed to distinguish between image classes that possibly differ only in their relative spatial arrangements such as the USGS dataset above. The global alignment of the 15 Scene dataset is a much stronger signal for discriminating between classes than relative spatial arrangement. It is for these same reasons that SPCK is not appropriate for strongly aligned object class datasets such as Caltech-101.

### 3.3.4 Conclusion

Our experimental results showed that the spatial pyramid co-occurrence representation captures both the absolute and relative spatial arrangements of visual words and can characterize a wide variety of spatial relationships through the choice of the underlying spatial predicates.

We performed an empirical evaluation using the USGS dataset and the spatial pyramid co-occurrence representation was shown to perform better on this dataset than a non-spatial bag-of-visual-words approach as well as a popular approach for characterizing the absolute spatial arrangement of visual words. And, while our primary objective is analyzing overhead imagery, we also demonstrated that the approach performs competitively on the Graz-01 object class dataset as well as the 15-Scene dataset.

## 3.4 Summary

In this chapter, we provided a series of experiments to empirically evaluate the effectiveness of local invariant features for labelling LULC classes characterized by identifiable objects in overhead imagery, in comparsion to texture and color features. We investigated BOVW approaches to LULC classification as well as how the incorporation of spatial arrangements of local invariant features can improve the standard BOVW model.

# Chapter 4

# Geospatial Object Detection

In the previous two chapters, we have been focused on content-based image retrieval and LULC classificaion using local invariant features. In this chapter, we turn to object detection in overhead imagery. In particular, our goal is to estimate the spatial extents of complex geospatial object such as high schools and golf courses. We first explore the potential synergy between gazetteers and overhead imagery for solving this problem and then propose a novel framework that uses readily available high resolution overhead imagery to estimate the boundaries of known object instances in order to update the gazetteers. We perform an empirical evaluation of our approach using a manually labelled dataset of geospatial objects.

The work in this chapter was published as peer-reviewed full conference papers at the SIGSPATIAL International Conference on Advances in Geographic Information Systems in 2008 [76] and the IEEE Workshop on Applications of Computer Vision in 2012 [77].

## 4.1 Background

Thanks to advances in technology, our ability to capture and store remote sensing imagery continues to improve. Unfortunately, our ability to analyze this imagery has not scaled proportionally and so automated methods are needed to realize the true value of this data. Fully automated image understanding based only on the image content–the pixel values–is likely to remain an unsolved problem. Even the most effective system we know of, the human visual system, relies heavily on experience, context, and other information external to an image. One approach that is proving particularly promising for making progress on automated image understanding is to *leverage non-image data associated with the images*. The nature of this meta-data varies not only in the richness of its description but also in how directly it is related to the image content. Extremes range from a well-annotated set of images in which keywords are associated with specific image regions, to the co-occurrence of large sections of text and multiple images in the same document without explicit correspondences, such as on a Web page. Our goal therefore, is to investigate how different modalities of geographic data can be integrated to automate information discovery that would not be possible through any one modality alone.

## 4.2 Related Work

### 4.2.1 Multimodality Integration in Multimedia Data

Multimodal data integration has been successfully applied to other information discovery problems particularly in multimedia analysis. Researchers have exploited connections between image and non-image data such as image annotations or video transcripts to improve image understanding and other challenging tasks.

Content based image retrieval (CBIR) is limited by the semantic gap between the low-level image features, such as color and texture, and the kind of higher-level annotation needed to support meaningful search. Researchers have explored various ways to overcome this by incorporating other sources of information. CBIR systems for the World Wide Web (WWW) often query associated text, such as an image's ALT tag or text near the image, in addition to the low-level image features. The CORTINA system of Quack et al. [78] combines text and image context to search over 3 million images on the WWW. Other systems exploit the structure of the WWW, such as Newsam et al.'s category based image retrieval system [79] which makes use of the hierarchical structure of the WWW pages containing the images in addition to associated text. Other approaches enable keyword search by tackling the more difficult problem of assigning keywords to images. Barnard and Forsyth in [80] learn the statistical associations between image regions and words in a large corpus of annotated images. The learned associations can then be used to annotate new images. (Interestingly, this association also enables the complementary task of auto-illustration.)

Computer vision researchers have exploited various forms of meta-data associated with image collections to learn visual object models. Berg et al. in [81] data mine a large collection of captioned images of faces from online news sources to train a recognition system for commonly occurring people. Barnard et al. in [82] develop an object recognizer using 10,000 images of works of art along with associated free text which varies greatly from physical description to interpretation and mood. And, Li et al. in [83] turn the search paradigm around by using search results from the Google image search engine to learn visual models for a variety of object categories.

Closer to the work presented in this chapter is the idea of using spatial and temporal meta-data to organize personal photo collections. Researchers have explored ways in which access to large collections of snapshots can be improved by classifying and grouping photographs using not only the time stamps of individual photographs but also the temporal spacing between photographs [84]. The advent of hand held global positioning systems (GPS) and even on-camera GPS now allows photographs to be organized geospatially. While research continues in this area, such as Yahoo Research's ZoneTag project [85, 86], many of the online photo sharing portals, such as Flickr [87], already offer this capability.

### 4.2.2 Multimodality Integration in Geographic Data

Researchers working in the geographic information sciences have proposed a number of ways to leverage non-image data sources to improve remote sensing image understanding. Using satellite or aerial imagery to maintain road networks has always held great appeal but automatically extracting roads is a challenging task. An obvious way to improve road extraction, at least for known roads, is to use existing vectorized road networks as seeds [88–90]. Researchers have also incorporated other information to improve road extraction, such as using digital surface models to account for gaps between road segments due to shadows [91]. Automated building extraction is another appealing use of remote sensing imagery. Agouris et al. [92] propose a SpatioTemporal Gazetteer that incorporates aerial imagery as well as existing vector datasets of extracted outlines and thematic datasets (building blueprints, building usage records) to automatically detect changes to the spatial footprints of buildings using template matching.

There has been research effort on using non-image GIS data to recognize a broader variety of object classes but largely without implementation or experimental results. Bailloeul et al. in [93] describe the theoretical aspects of a system which uses a priori knowledge in form of outdated urban maps to control contour- and region-based segmentation of new imagery. And, Baltsavias in [94] discusses a wide range of ways in which “knowledge” can improve image analysis but, as he indicates, his definition of non-image information is extremely broad and includes concepts such as rules, models, and context, in additional to specific GIS data. (He notes that very few of the works he surveyed use priori knowledge in the form of maps, GIS or other geodatabases.) Walter and Fritsch in [95] do provide results from using GIS data to automatically derive appearance models which are then applied to imagery to verify the GIS data but their application is at the level of land use classification.

Note that the discussion above, and the context of the work in this chapter, concerns techniques for integrating multiple data sources, one of which is unprocessed image data. This is different from approaches which assume the image analysis has already been carried out such as the interesting recent work by Michalowski et al. [96] on combining street vector data, phone-book records and building outlines to map postal addresses.

## 4.3 Motivation

We contend that there is significant opportunity for exploiting connections between image and non-image geographic data due to 1) location being simple yet powerful key for associating widely varying data modalities; and 2) the growing available of data annotated with location information either explicitly or implicitly.

We focus on combining high-resolution aerial imagery with gazetteers. Our work is motivated by the fact that current gazetteers, geographic dictionaries of what-is-where on the surface of the Earth, are deficient in that *the spatial extents of the archived objects are limited to a single point, a latitude/longitude pair.* While the systems include provisions for storing at least a bounding box representation, this information has simply never been acquired or computed. As the development team of the University of California at Santa Barbara Alexandria Digital Library (ADL) gazetteer points out [97], "for a digital library application, the spatial extent of the feature, either approximately with a bounding box or more accurately with a polygonal representation, is better, but there are no large sets of gazetteer data with spatial extents." They go on to state that "establishing the standards that will enable the sharing of gazetteer data will help harvest data from many sources, but ultimately deriving spatial locations and extents from digital mapping products and other sources automatically will be needed." The long-term goal of this work is to do just as the ADL gazetteer development teams proposes, leverage readily available high resolution overhead imagery to estimate the spatial extents of known object instances with minimal user supervision.

## 4.4 Integrating Gazetteers and Remote Sensing Imagery

We first explore the potential for synergy between gazetteers and overhead imagery. We propose a framework for using prior knowledge in the form of gazetteer records to learn appearance models for a variety of geospatial objects in an unsupervised fashion. We then propose how to use appearance models to improve the spatial extent associated with geospatial objects in a gazetteer.

### 4.4.1 Data Sources

This section describes the two data sources being considered for integration, gazetteers and high-resolution remote sensing imagery.

### Gazetteers

A gazetteer is a geographic directory. It contains records indicating what-is-where on the surface of the earth. The what varies greatly both in terms of the class of object as well as its physical characteristics, such as its spatial extent. While the minimal set of fields is a name and a point location, gazetteers typically include both proper names (e.g., San Francisco Internal Airport) and object classification (e.g., Airport), more complex spatial extents, such as a bounding box or higher-order polygon, and relations to other records (e.g., part of San Francisco County). Gazetteers are not a new concept but advances in information technology have enabled them to become more extensive, thanks to automated aggregation, and more accessible, thanks to the Internet, than ever before. One of the gazetteers we consider is the Alexandria Digital Library (ADL) gazetteer [98] which is part of the ADL project at the University of California at Santa Barbara. The ADL gazetteer is an exemplar of modern gazetteers. It was created by aggregating the United States Geological Survey's *Geographic Names Information System* (GNIS) and the National Imagery and Mapping Agency's *Geographic Names Processing System* (GNPS) [97]. It has an online browser-based map interface for interactive querying. More important for our work is the ADL Gazetteer Service Protocol which supports remote query and response functions using standard HTTP XML-formatted requests.

The ADL gazetteer contains almost six million records and thus represents one of the most extensive and diverse collections of its kind. It covers the entire world although due to the nature of its original sources, its coverage is more complete for the US. It catalogs over 200 different classes of geographic features. Further, a feature class thesaurus maps almost 1,000 non-preferred terms to these primary classes. The feature classes are organized hierarchically under the following six root classes: administrative areas, hydrographic features, land parcels, manmade features, physiographic features, and regions. Table 4.1 contains a list of the primary feature classes and their total counts.

While the ADL gazetteer is an impressive collection of geographic data, it has several shortcomings. Some of these could potentially be compensated for by advances in automated image analysis such as the techniques we propose in this chapter. At the moment, the spatial extent of the records is limited to a single point, a longitude/latitude pair. While the system includes provisions for at least a bounding box representation, this information was not present in the original sources and post-ingest manual specification is prohibitively expensive.

### Remote Sensing Imagery

We believe that remote sensing image analysis is poised to undergo a paradigm shift thanks to the growing availability of wide area coverage submeter pixel resolution data. The domain is faced with the prospects and challenges of object level analysis on a scale not possible before. The higher resolution imagery greatly increases the variety of objects that are now observable, at least according to theoretical bounds such as Shannon's sampling theorem. This aligns the problem more closely with the well researched computer vision challenge of generic object recognition and allows it to leverage the many advances made over the past several decades in that area. In particular, given the work from the

**Table 4.1:** Object classes indexed by the ADL gazetteer along with counts.

| Class | Count | Class | Count | Class | Count |
|---|---:|---|---:|---|---:|
| administrative areas | 2,126,610 | cemeteries | 64,535 | dunes | 5,270 |
| military areas | 813 | disposal sites | 247 | flats | 4,722 |
| parks | 20,408 | fisheries | 45 | gaps | 15,762 |
| political areas | 32,623 | fortifications | 2,471 | isthmuses | 79 |
| countries | 165 | historical sites | 66,228 | karst areas | 113 |
| countries, 1st order div | 3,328 | archaeological sites | 1,654 | ledges | 644 |
| countries, 2nd order div | 14,602 | hydrographic structures | 123,991 | mesas | 5,529 |
| countries, 3rd order div | 13,115 | breakwaters | 101 | mineral deposit areas | 667 |
| countries, 4th order div | 1,330 | canals | 21,482 | moraines | 39 |
| populated places | 2,000,821 | dam sites | 45,828 | mountains | 362,194 |
| cities | 273 | harbors | 5,250 | mountain ranges | 526 |
| capitals | 271 | levees | 726 | mountain summits | 21,401 |
| reference locations | 42,597 | marinas | 45 | ridges | 21,916 |
| reserves | 4,704 | offshore platforms | 58 | natural rock formations | 2,154 |
| tribal areas | 4,183 | piers | 403 | arches (natural formation) | 446 |
| hydrographic features | 636,5 64 | reservoirs | 49,976 | plains | 7,171 |
| bays | 34,898 | waterworks | 82 | plateaus | 1,165 |
| fjords | 1,656 | landmarks | 545 | playas | 4 |
| channels | 13,874 | mine sites | 24,070 | reefs | 8,270 |
| deltas | 66 | monuments | 8,560 | seafloor features | 2,112 |
| drainage basins | 208 | oil fields | 4,834 | continental margins | 140 |
| estuaries | 424 | recreational facilities | 7,526 | ocean trenches | 252 |
| floodplains | 4 | amusement parks | 18 | seamounts | 1,025 |
| gulfs | 420 | camps | 3,744 | submarine canyons | 590 |
| guts | 932 | performance sites | 312 | tectonic features | 476 |
| ice masses | 3,569 | sports facilities | 1,948 | earthquake features | 345 |
| glacier features | 3,492 | storage structures | 11,969 | faults | 127 |
| lakes | 94,758 | telecom features | 12,240 | fracture zones | 123 |
| seas | 273 | towers | 14,423 | folds (geologic) | 4 |
| oceans | 42 | transportation features | 77,933 | anticlines | 4 |
| ocean currents | 24 | airport features | 24,814 | valleys | 36,815 |
| streams | 480,921 | heliports | 3,802 | canyons | 18,706 |
| rivers | 15,907 | seaplane bases | 451 | volcanic features | 2,262 |
| bends (river) | 1,637 | aqueducts | 101 | lava fields | 391 |
| rapids | 3,095 | bridges | 4,709 | volcanoes | 1,819 |
| waterfalls | 4,945 | locks | 259 | regions | 120,132 |
| springs (hydrographic) | 3,016 | parking sites | 5 | biogeographic regions | 44,782 |
| thermal features | 407 | pipelines | 186 | barren lands | 11 |
| land parcels | 12,424 | railroad features | 34,029 | deserts | 588 |
| manmade features | 858,145 | roadways | 578 | forests | 19,476 |
| agricultural sites | 174,912 | trails | 7,582 | petrified forests | 2 |
| buildings | 243,448 | tunnels | 761 | woods | 871 |
| capitol buildings | 29 | wells | 71,680 | grasslands | 4,419 |
| commercial sites | 2,711 | windmills | 171 | habitats | 56 |
| industrial sites | 1,465 | physiographic features | 575,964 | oases | 362 |
| power generation sites | 26 | alluvial fans | 66 | shrublands | 1,900 |
| court houses | 664 | arroyos | 45,359 | snow regions | 77 |
| institutional sites | 189,02 | badlands | 17 | tundras | 6 |
| correctional facilities | 32 | banks (hydrographic) | 2,014 | wetlands | 17,887 |
| educational facilities | 14,885 | bars (physiographic) | 12,167 | coastal zones | 165 |
| medical facilities | 5,558 | basins | 8,957 | economic regions | 8 |
| religious facilities | 78,26 3 | storage basins | 20 | firebreaks | 8 |
| library buildings | 2,134 | beaches | 3,584 | land regions | 70,094 |
| museum buildings | 886 | bights | 590 | continents | 3 |
| post office buildings | 11,16 | capes | 19,707 | islands | 70,088 |
| research facilities | 339 | caves | 966 | map regions | 2,311 |
| data collection facilities | 14 | cirques | 196 | map quadrangle regions | 2,311 |
| residential sites | 22,719 | cliffs | 5,557 | research areas | 195 |
| housing areas | 21,756 | craters | 435 | uncategorized | 20,719 |

previous two chapters, it can leverage local invariant descriptors, which have proven effective at modelling the appearances of diverse categories of objects viewed under widely varying conditions.

Meter (e.g. IKONOS) and submeter (e.g. Quickbird) pixel resolution satellite imagery has been available commercially now for nearly a decade. GeoEye-1, which launched in August of 2008, provides 0.41 meter/pixel imagery. This is in addition to the growing collection of aerial imagery which boasts even higher resolution. Automated image analysis is needed to realize the full potential of these information rich data sources. We believe the integration of gazetteers and remote sensing imagery represents a promising step in this direction.

### 4.4.2 Geospatial Object Modelling

This section describes the framework that uses prior knowledge in the form of gazetteer records to learn appearance models in an unsupervised fashion. First, we use the ADL gazetteer to locate and extract image regions corresponding to 13 different classes of geospatial objects in a large collection of IKONOS satellite imagery. We then extract high-dimensional image features from these regions and show they are clustered in the feature space, a potentially sufficient condition for learning object models. We then investigate classification strategies to further demonstrate that we are able to learn appearance models of geospatial objects in an unsupervised fashion through integration of a gazetteer and remote sensing imagery.

### Data

Thanks to a generous grant from Lockheed Martin Corporation, we have access to a set of 16 georeferenced IKONOS satellite images which cover over 3,000 square kilometers of the continental US. While this imagery consists of one meter panchromatic band and four meter multispectral bands, we only utilize the panchromatic imagery. These IKONOS images are mostly of metropolitan areas and include coverage of Phoenix, Los Angeles, New York City, San Diego, San Jose, Washington DC, and El Paso.

We selected a subset of object classes cataloged by the ADL gazetteer (see Table 4.1 for the full list) using the following criteria. First, the spatial extent of the objects must be limited enough for them to be contained within a single IKONOS image. This rules out larger objects such as counties. Second, the objects must be visible in the imagery. This obvious constraint rules out subterranean objects such as caves. Third, the visual characteristics of the object taken as a whole must be distinctive. This allows different object types to share some visual features but requires that there should be some way of discriminating between them visually. Finally, since most of our imagery is of urban/suburban regions, we focused on manmade objects. This selection process resulted in the 13 geospatial object classes listed in Table 4.2.

We then used the ADL Gazetteer Service Protocol to identify instances of these 13 geospatial object classes in our IKONOS imagery. This was achieved by issuing XML formatted queries requesting all instances of a particular object class contained within the spatial footprint of each of the images separately. Table 4.2 lists the total number of instances identified in all 16 images. Table 4.2 also indicates the class ID assignments that will be used in subsequent tables.

**Table 4.2:** Appearance models were learned for the following geospatial object classes. The column labelled count refers to the total number of examples in the IKONOS imagery as found by querying the ADL gazetteer.

| ID | Object Class | Count |
|---|---|---|
| 1 | Airport Features | 24 |
| 2 | Cemeteries | 40 |
| 3 | Country Clubs | 12 |
| 4 | Educational Facilities | 694 |
| 5 | Golf Courses | 23 |
| 6 | Harbors | 6 |
| 7 | Heliports | 50 |
| 8 | Medical Facilities | 70 |
| 9 | Mobile Home Parks | 88 |
| 10 | Parks | 274 |
| 11 | Railroad Features | 9 |
| 12 | Religious Facilities | 126 |
| 13 | Shopping Centers | 147 |

As mentioned earlier, a shortcoming of the ADL gazetteer (and most gazetteers) is that it specifies the spatial extents of the objects as only a single point location. Therefore, even though we can use the gazetteer to identify object instances in the IKONOS imagery, we know very little about which region in the image actually corresponds to the objects. Without additional information, the best we can do is extract a square subimage centered at the gazetteer specified location and assume the object is contained somewhere in this subimage. At the moment, the size of these subimages is set manually and varies from 300 by 300 pixels for heliports to 1000 by 1000 pixels for airports with most being 600 by 600 pixels. A better estimate of the size of the subimages for different object classes could be obtained using a dataset with spatial extents. This would reduce the amount of background (non-object) regions in the subimages and therefore improve the appearance modelling.

Figure 4.1 contains two sample subimages from four object classes: educational facilities, golf courses, mobile home parks, and shopping centers. Again, these subimages were extracted in a completely automated fashion by cropping a fixed sized region from an IKONOS image centered at the point location specified by the ADL gazetteer. These are some of the better samples in that 1) the point locations are actually within the objects; 2) the ratio of the object to background is quite high; and 3) the objects are good visual examples of their classes. Overall, the subimages form a fairly noisy dataset. Frequently, the ADL specified point location is not within the object due to inaccuracies of the original gazetteer data sources and/or the georeferencing process. We can compensate for this by extracting larger regions but this results in a lower object to background ratio. There can also be significant variance visually between the samples for a particular class. The less-than-perfect nature of the subimage dataset make the object modelling task an interesting challenge.

### Image Features

We use the standard SIFT interest point detector and the standard BOVW representation. We use the cosine distance measure to compute the similarity between images.

We first investigate whether the frequency features belonging to the different object classes form clusters in the space induced by the cosine distance. We compare the average within cluster feature distance to the average between cluster feature distance. The results

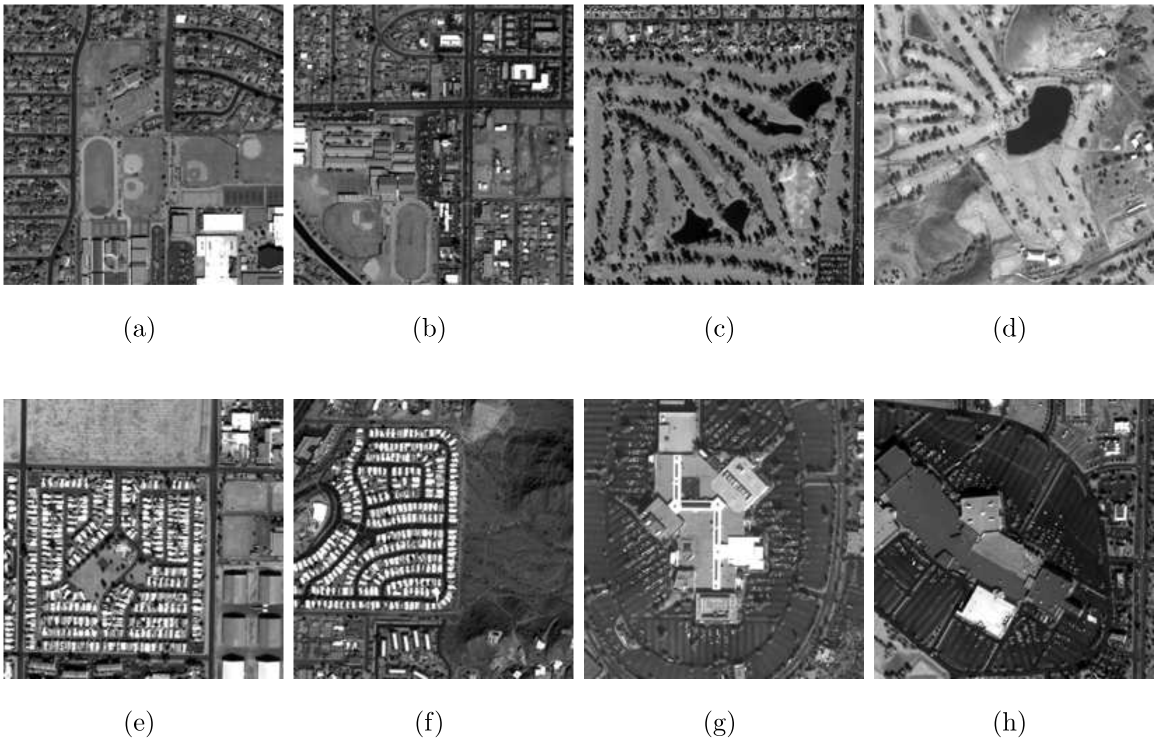


**Figure 4.1:** Sample subimages of different object classes extracted in an unsupervised fashion from IKONOS imagery using the ADL gazetteer. Two samples of each of the following classes are shown in scan order: educational facilities, golf courses, mobile home parks, and shopping centers.

of this analysis are shown in Table 4.3. We observe that the features for most classes are more similar to themselves than to the features for other classes.

### Classification

We used a classification framework to further show that we are learning discriminative appearance models for the 13 object classes. We computed average classification rates using a leave-one-out approach in which we train classifiers using all the features for an object type except one selected at random which is later used as the test feature.

**Table 4.3:** The features belonging to different object classes form clusters in the feature space. The value at row $i$, column $j$ is the average distance of the features for object class (ID) $i$ to the features for object class $j$. The higher the value, the more similar the features. Note that the features of most object classes are more similar to themselves than to the features of other object classes. The entries on the diagonal are emphasized for readability.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | **0.854** | 0.837 | 0.813 | 0.847 | 0.814 | 0.859 | 0.835 | 0.851 | 0.848 | 0.839 | 0.878 | 0.839 | 0.870 |
| 2 | 0.837 | **0.926** | 0.902 | 0.922 | 0.896 | 0.896 | 0.853 | 0.916 | 0.896 | 0.913 | 0.920 | 0.919 | 0.907 |
| 3 | 0.813 | 0.902 | **0.891** | 0.896 | 0.884 | 0.873 | 0.836 | 0.892 | 0.873 | 0.889 | 0.894 | 0.895 | 0.885 |
| 4 | 0.847 | 0.922 | 0.896 | **0.938** | 0.878 | 0.912 | 0.865 | 0.932 | 0.917 | 0.924 | 0.933 | 0.938 | 0.929 |
| 5 | 0.814 | 0.896 | 0.884 | 0.878 | **0.891** | 0.863 | 0.830 | 0.876 | 0.854 | 0.875 | 0.885 | 0.875 | 0.869 |
| 6 | 0.859 | 0.896 | 0.873 | 0.912 | 0.863 | **0.938** | 0.869 | 0.915 | 0.882 | 0.903 | 0.925 | 0.924 | 0.921 |
| 7 | 0.835 | 0.853 | 0.836 | 0.865 | 0.830 | 0.869 | **0.841** | 0.869 | 0.861 | 0.857 | 0.882 | 0.864 | 0.880 |
| 8 | 0.851 | 0.916 | 0.892 | 0.932 | 0.876 | 0.915 | 0.869 | **0.928** | 0.912 | 0.919 | 0.931 | 0.934 | 0.928 |
| 9 | 0.848 | 0.896 | 0.873 | 0.917 | 0.854 | 0.882 | 0.861 | 0.912 | **0.928** | 0.902 | 0.916 | 0.911 | 0.920 |
| 10 | 0.839 | 0.913 | 0.889 | 0.924 | 0.875 | 0.903 | 0.857 | 0.919 | 0.902 | **0.912** | 0.922 | 0.925 | 0.915 |
| 11 | 0.878 | 0.920 | 0.894 | 0.933 | 0.885 | 0.925 | 0.882 | 0.931 | 0.916 | 0.922 | **0.949** | 0.931 | 0.937 |
| 12 | 0.839 | 0.919 | 0.895 | 0.938 | 0.875 | 0.924 | 0.864 | 0.934 | 0.911 | 0.925 | 0.931 | **0.948** | 0.930 |
| 13 | 0.870 | 0.907 | 0.885 | 0.929 | 0.869 | 0.921 | 0.880 | 0.928 | 0.920 | 0.915 | 0.937 | 0.930 | **0.938** |

We explored two types of classifiers, a simple nearest centroid classifier and a maximum likelihood classifier.

The nearest centroid classifier assigns a test feature to the object class with the closest centroid as computed using the cosine distance. This centroid is simply the average of the training features for an object type. A separate classifier is trained for each class. Table 4.4 shows the confusion matrix for 1,500-fold cross validation. The classification rates for different object classes vary significantly from a low of 0.02 for parks to 0.69 for mobile home parks, with several above 0.50. Overall, the fact that we are doing much better than chance (0.08) confirms we are learning discriminative appearance models, at least for most of the classes.

We also performed classification using a maximum likelihood classifier to better incorporate the spread of the features for an object class. We modelled the distribution of the features for each class using a von Mises-Fisher (vMF) probability density function.

**Table 4.4:** The confusion matrix of classification using a closest centroid classifier. 1,500-fold cross validation is performed using a leave-one-out approach. The value at row $i$, column $j$ indicates the average classification rate that a sample of object class $i$ is classified as class $j$. The entries on the diagonal are emphasized for readability. These results confirm that we are learning discriminative appearance models for most object classes.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | **0.58** | 0.04 | 0.00 | 0.09 | 0.04 | 0.03 | 0.00 | 0.00 | 0.05 | 0.04 | 0.08 | 0.00 | 0.04 |
| 2 | 0.00 | **0.34** | 0.02 | 0.03 | 0.19 | 0.03 | 0.00 | 0.02 | 0.06 | 0.05 | 0.02 | 0.20 | 0.04 |
| 3 | 0.00 | 0.00 | **0.16** | 0.00 | 0.34 | 0.00 | 0.00 | 0.00 | 0.26 | 0.00 | 0.00 | 0.08 | 0.15 |
| 4 | 0.04 | 0.08 | 0.00 | **0.27** | 0.01 | 0.03 | 0.01 | 0.02 | 0.15 | 0.03 | 0.06 | 0.25 | 0.05 |
| 5 | 0.00 | 0.08 | 0.04 | 0.00 | **0.56** | 0.04 | 0.00 | 0.00 | 0.00 | 0.13 | 0.00 | 0.03 | 0.12 |
| 6 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.67** | 0.16 | 0.00 | 0.00 | 0.00 | 0.00 | 0.17 | 0.00 |
| 7 | 0.13 | 0.00 | 0.00 | 0.04 | 0.10 | 0.13 | **0.33** | 0.02 | 0.06 | 0.03 | 0.06 | 0.06 | 0.06 |
| 8 | 0.06 | 0.09 | 0.01 | 0.11 | 0.03 | 0.09 | 0.05 | **0.14** | 0.10 | 0.03 | 0.02 | 0.20 | 0.08 |
| 9 | 0.01 | 0.05 | 0.01 | 0.07 | 0.00 | 0.03 | 0.03 | 0.02 | **0.69** | 0.01 | 0.01 | 0.00 | 0.08 |
| 10 | 0.02 | 0.16 | 0.01 | 0.17 | 0.08 | 0.10 | 0.02 | 0.04 | 0.08 | **0.02** | 0.07 | 0.18 | 0.05 |
| 11 | 0.21 | 0.00 | 0.00 | 0.00 | 0.11 | 0.00 | 0.00 | 0.11 | 0.00 | 0.11 | **0.22** | 0.12 | 0.11 |
| 12 | 0.00 | 0.05 | 0.01 | 0.08 | 0.02 | 0.18 | 0.01 | 0.03 | 0.00 | 0.03 | 0.05 | **0.54** | 0.01 |
| 13 | 0.07 | 0.02 | 0.00 | 0.06 | 0.01 | 0.06 | 0.08 | 0.04 | 0.22 | 0.01 | 0.06 | 0.11 | **0.27** |

A vMF density can be thought of as a normal density for points on a unit hyper-sphere which is appropriate for our features since they are directional (thus compared using the cosine distance). The probability density function of a vMF distribution has the following form

$$p(x|\mu, \kappa) = c_d(\kappa) e^{\kappa \mu^T x} \ , \tag{4.1}$$

where $d$ is the dimension, and $\mu$, the distribution "mean," and $\kappa$, the distribution spread, are the parameters. We again use a leave-one-out approach in which a separate classifier is learned for each class by estimating $\mu$ and $\kappa$ from the set of features with one left out. The left out samples are then classified as the object class which maximizes their likelihood:

$$c^* = \arg \max_{c \in allclasses} p(x_{sample}|\mu_c, \kappa_c) \ . \tag{4.2}$$

The results of performing a 1,500-fold cross validation are shown in Table 4.5. Again, though the classification rates vary significantly between classes, several of the rates are above 0.5 confirming we are learning discriminative appearance models.

**Table 4.5:** The confusion matrix of classification using a maximum likelihood classifier. The features of each object class are modelled using a von Mises-Fisher distribution. 1,500-fold cross validation is performed using a leave-one-out approach. The value at row $i$, column $j$ indicates the average classification rate that a sample of object class $i$ is classified as class $j$. The entries on the diagonal are emphasized for readability. These results confirm that we are learning discriminative appearance models for most object classes.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | **0.63** | 0.04 | 0.00 | 0.00 | 0.04 | 0.00 | 0.02 | 0.22 | 0.04 | 0.01 | 0.00 | 0.00 | 0.00 |
| 2 | 0.00 | **0.45** | 0.00 | 0.00 | 0.14 | 0.00 | 0.01 | 0.21 | 0.05 | 0.13 | 0.00 | 0.00 | 0.00 |
| 3 | 0.00 | 0.17 | **0.01** | 0.00 | 0.36 | 0.00 | 0.00 | 0.08 | 0.33 | 0.05 | 0.00 | 0.00 | 0.00 |
| 4 | 0.04 | 0.12 | 0.00 | **0.00** | 0.01 | 0.00 | 0.03 | 0.50 | 0.16 | 0.13 | 0.00 | 0.00 | 0.00 |
| 5 | 0.00 | 0.13 | 0.06 | 0.00 | **0.49** | 0.00 | 0.00 | 0.18 | 0.05 | 0.09 | 0.00 | 0.00 | 0.00 |
| 6 | 0.21 | 0.00 | 0.00 | 0.00 | 0.00 | **0.00** | 0.29 | 0.34 | 0.00 | 0.16 | 0.00 | 0.00 | 0.00 |
| 7 | 0.30 | 0.02 | 0.02 | 0.00 | 0.07 | 0.00 | **0.35** | 0.11 | 0.07 | 0.06 | 0.00 | 0.00 | 0.00 |
| 8 | 0.05 | 0.13 | 0.02 | 0.00 | 0.01 | 0.00 | 0.04 | **0.52** | 0.17 | 0.06 | 0.00 | 0.00 | 0.00 |
| 9 | 0.01 | 0.06 | 0.00 | 0.00 | 0.00 | 0.00 | 0.09 | 0.06 | **0.76** | 0.02 | 0.00 | 0.00 | 0.00 |
| 10 | 0.02 | 0.19 | 0.01 | 0.00 | 0.05 | 0.00 | 0.07 | 0.45 | 0.10 | **0.10** | 0.00 | 0.00 | 0.00 |
| 11 | 0.23 | 0.19 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.51 | 0.00 | 0.05 | **0.00** | 0.00 | 0.00 |
| 12 | 0.00 | 0.09 | 0.00 | 0.00 | 0.01 | 0.00 | 0.07 | 0.68 | 0.00 | 0.15 | 0.00 | **0.00** | 0.00 |
| 13 | 0.08 | 0.03 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.44 | 0.33 | 0.03 | 0.00 | 0.00 | **0.00** |

### Discussion

This section described a fully automated technique for learning visual appearance models for geospatial object classes by integrating the ADL gazetteer and high-resolution IKONOS imagery. Preliminary exploration of the class feature distributions and on training both a nearest centroid and a maximum likelihood classifier indicates we are learning discriminative models even though the training images are extremely noisy.

We find the two classifiers perform quite differently. In particular, the maximum likelihood classifier, which accounts for the variation in feature spread between classes, completely miss-classifies educational facilities, harbors, railroad features, religious facilities, and shopping centers. These objects instead are predominantly classified as medical facilities. Further investigation here is needed although we observed that the estimated von Mises-Fisher distributions for these classes are more peaked (the value of $\kappa$ is larger) and thus classes with wider distributions, such as medical facilities, can "sweep-up" noisy instances.

### 4.4.3 Conclusion

In section 4.4, we explored the potential for increased synergy between gazetteers and remote sensing imagery. We described two specific ways in which these complementary data sources can be integrated to more fully automate geographic data management. We presented a novel approach to using gazetteers as a source of semi-supervised training data for appearance based modelling of geospatial objects.

## 4.5 Estimating the Spatial Extents of Geospatial Objects Using Hierarchical Models

The previous section explores how to use gazetteers to learn appearance models of geospatial objects in an unsupervised fashion. In this section, we propose a hierarchical model to represent the appearance of geospatial objects and thus to estimate the spatial extents when learned. We learn the model in both supervised and semi-supervised fashions and show the potential of our approach to estimate the bounding boxes of archived geospatial objects in the gazetteer using limited labelled data.

### 4.5.1 Approach Overview

An overview of the proposed approach is shown in Figure 4.2. First, gazetteers are queried for object instances, for example all high schools in a geographic region such as a city. The point locations of these objects are then used to retrieve high resolution images from online repositories. The spatial extents of the objects are either fully or partially labelled in a small subset of the images to form a training set which is used to learn the object models. The model is used to estimate the spatial extents of the objects in the target images. Finally, the gazetteers are updated with the spatial extents of the originally queried objects.

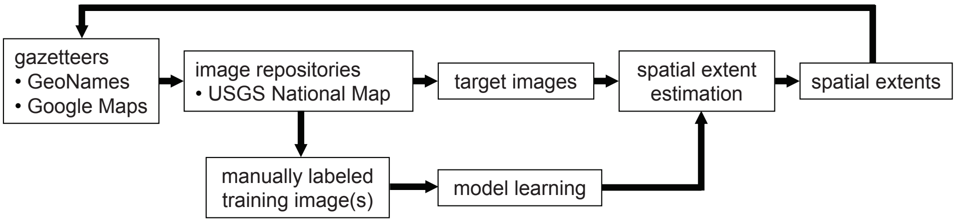


**Figure 4.2:** An overview of our proposed approach for estimating the spatial extents of geospatial objects.

### 4.5.2 Data Sources

#### Gazetteers

We utilize two gazetteers in the experiment. First GeoNames[1], an online worldwide gazetteer compiled from several dozen sources including other gazetteers such as the USGS Geographic Names Information System. It contains over 8 million features (objects) categorized into nine top-level classes which are further subcategorized into 645 feature codes. All the data is accessible free of charge through a number of webservices as well as a daily database export. The GeoNames web interface allows fuzzy search using geographic names, locations, features codes, and feature classes. Queries to GeoNames return a single latitude/longitude point as the spatial extent of an object.

We also treat Google Maps as a gazetteer in that it allows us to perform location-based searches for geospatial objects such as Costco shopping centers. We further use the Google Maps Geocoding API[2] to translate the street addresses provided by Google Maps into latitude/longitude points.

[1] http://www.geonames.org

[2] http://code.google.com/apis/maps/documentation/geocoding

### Image Repositories

We limit our study area to the US and use the USGS National Map Seamless Data Server[1] interface to automatically download imagery. This interface accepts spatial queries for a range of data collections including High Resolution Orthoimagery of major US urban areas at 3-inch, 6-inch, 1-foot, and 2.5-foot spatial resolutions, and the US Department of Agriculture (USDA) National Agriculture Imagery Program imagery of the conterminous United States at 1-meter or 2-meter spatial resolutions.

Images are retrieved from the National Map using a simple rectangular query region specified by its bounding latitude and longitude values. In our case, the single latitude/longitude point from the gazetteer serves as the center of a region whose size is chosen to ensure that the retrieved image contains the target object. This size is chosen empirically in the experiments below based on the observed sizes of sample objects. A single size is picked for each object type and then fixed for all the retrievals. Note that the gazetteer point does not always fall inside the object due to data collection, geo-registration, or other errors.

## 4.5.3 Hierarchical Models

The three levels of the hierarchical model are shown in Figure 4.3. We now describe each of the levels in detail.

[1] http://seamless.usgs.gov

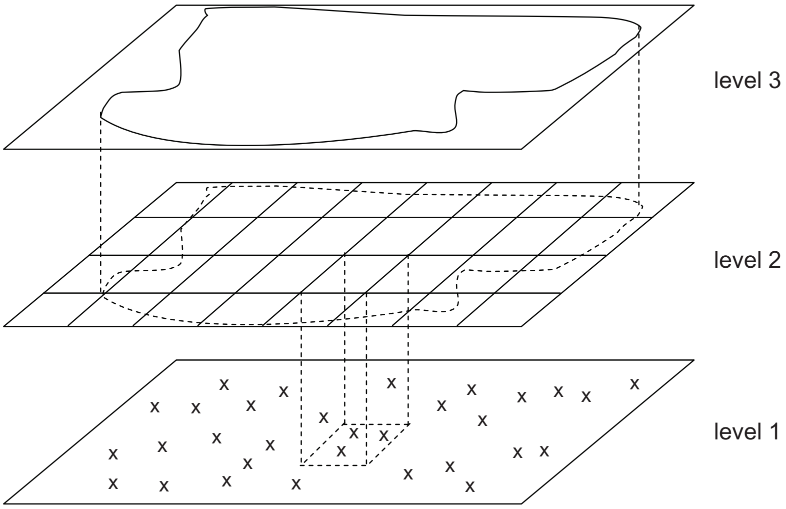


**Figure 4.3:** The three levels of our hierarchical model. Level 1 represents the object using quantized SIFT features shown here as x's. BOVW histograms are computed for image tiles and SVM classifiers are used to assign LULC labels to the tiles in level 2. The distribution of the LULC classes in level 3 constitutes the final object model.

### Level 1 - Local Invariant Features

We use local invariant features to characterize the objects at the lowest level of the hierarchy. We use SIFT features as our local invariant feature detector and descriptor. We adopt a standard BOVW approach and construct a visual dictionary by performing $k$-means clustering on a large number of SIFT features (from a dataset different from that used to train the object models). We use 256x256 pixel tiles in all the experiments and the BOVW histogram is normalized to have unit L1 norm to account for the difference in the number of interest points between tiles.

### Level 2 - Latent LULC Classes

An intermediate, latent level bridges the gap between the low-level local invariant features and the high-level objects. Specifically, LULC labels are assigned to image tiles using support vector machines (SVMs).

We leverage our work on the USGS LULC classification from the previous chapter. We use a one-against-all strategy to perform multi-class SVM classification. We also use the probabilistic output option of the LIBSVM package [64]. Specifically, for each tile $i$ in an image, we compute the probability distribution over the $M$ LULC classes as

$$P(tile_i) = [p_1, p_2, ..., p_M] \ ,$$

where $p_m$ corresponds to the probability that tile $i$ is assigned to the $m$th class by the SVM classifiers. The SVM classifiers take as input the BOVW histograms from level 1. We normalize $P(tile_i)$ so that $\sum^M p_m = 1$.

In order to reduce the effect of tile (mis)alignment, we perform the LULC labeling on tiles which overlap by 50 percent. Thus, each 128x128 pixel *block* appears in four 256x256 pixel tiles. We apply a smoothing mechanism to the LULC class distribution at the block level

$$P(block_j) = \frac{1}{4} \sum P(tile_i) \ , \tag{4.3}$$

where the sum is taken over the four tiles in which block $j$ appears.

To summarize, our final representation at level 2 in the hierarchy is a probability distribution $P(block_j)$ over $M$ LULC classes for each 128x128 pixel block $j$.

### Level 3 - Object Model

The top level of our representation also models the objects as probability distributions over LULC classes. The distributions corresponding to different object types can be easily learned from one or more training samples. Given $N$ training samples encompassing a set of $\mathbb{U}$ blocks labelled at level 2, we compute

$$P(object) = \frac{1}{|\mathbb{U}|} \sum_{block_j \in \mathbb{U}} P(block_j) \ , \tag{4.4}$$

where $P(block_j)$ is computed using equation 4.3 and $|\mathbb{U}|$ is the cardinality of $\mathbb{U}$.

## 4.5.4 Spatial Extent Estimation

The primary goal is to estimate the spatial extent of known object instances. Again, in the context of this problem, the gazetteer provides a single latitude/longitude point for the object. This point is used to download a target image $T$ large enough to encompass the object. An object model $P(object)$ is then used to estimate the spatial extent of the object as follows.

First, we extract and quantize SIFT features from the target image using the same visual dictionary as in level 1 of the object model. We then compute the BOVW histograms for overlapping 256x256 pixel tiles and the multi-class SVM classifiers are used to to compute the LULC class distributions for each of the tiles. The LULC class distributions are then computed for each 128x128 pixel block using equation 4.3.

The problem now reduces to determining the contiguous set of blocks that are most similar to the object model. We simplify this search by 1) scoring overlapping square windows each containing a fixed number of blocks, 2) applying a threshold to the scores, and 3) computing the final spatial extent as the union of the selected windows.

Specifically, we slide a square window of size $w$x$w$ blocks over the image in increments of one block. For each window location, we compute the probability distribution of the window over the LULC classes:

$$P(window) = \frac{1}{w^2} \sum_{block_j \in window} P(block_j) \ , \tag{4.5}$$

where $P(block_j)$ is computed using equation 4.3. We then compute the similarity between the window and the object model $D(window, object)$ using the intersection measure

$$D(window, object) = \sum^{M} min(P(window)[m], P(object)[m])) \ , \tag{4.6}$$

where $[m]$ indicates the $m$th component and $M$ is the number of LULC classes. If $D(window, object)$ is above a threshold $\theta$, we label all the blocks in the window as belonging to the target object. (We discuss the setting of $\theta$ below.) Finally, after each window location has been visited, we compute the spatial extent of the object as the union of all the selected blocks. The key is to learn the object model and two learning regimes are employed: fully supervised and semi-supervised. In the fully supervised mode, the object model is learned using fully manually labelled groundtruth images in which boundaries of geospatial objectes are delienated. In the semi-supervised mode, the learning process of the object model makes use of not only labelled groundtruth data, but also weakly labelled data. We consider the image data retrieved automatically using existing gazetteers as weakly labelled data as a whole, in that the current gazetteers represent the spatial extent of the archive geospatial object using only a single latitude/longitude point and we expect at least part of the image data to contain positive examples of an object. As for each individual image, it is acutally unlabelled since the object boundary is not delienated, if the manual groundtruth is not given. To include the unlabelled data in the training set we assume that unlabelled examples which are similar to labelled ones should have a similar label. Being able to incorporate unlabelled data is important as it allows for taking full advantage of the automatically generated data using gazetteers. For the semi-supervised learning, we use a recent bipartite ranking function which can incorporate unlabelled data into the learning phase [99]. The goal in bipartite ranking is to learn a scoring function $H$ which assigns higher scores to relevant instances than to irrelevant ones. In our problem, we consider regions within the boundaries of the geospatial objects as being relevant and regions belonging to the background as being irrelevant. Our goal then is to mark as relevant those regions within the groundtruth spatial extent of a test image.

### 4.5.5 Experimental Results

We evaluate our approach using an groundtruth dataset consisting of four object types: high schools, golf courses, mobile home parks, and Costco shopping centers.

#### Dataset

We use the first stage of the framework in Figure 4.2 to identify the locations of target objects and their corresponding images. The GeoNames gazetteer is used to identify 44 high schools, 27 golf courses, and 23 mobile home parks, and Google Maps is used to identify 18 Costco shopping centers. The National Map Seamless Data Server is then used to download 1-foot resolution orthoimagery using a large query region to ensure the images contain the target objects. The images are in the RGB colorspace.

A groundtruth dataset is created by manually delineating the target objects using a polygon representation. We also compute the rectangular, axis aligned bounding boxes of the target objects using the polygonal boundaries.

SIFT features are extracted from each of the images and quantized using a visual dictionary consisting of 100 visual words. A BOVW histogram is computed for overlapping 256x256 pixel tiles.

Tile-level LULC distributions are computed using a set of SVMs corresponding to 18 LULC classes: agricultural, airplane, baseball diamond, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium density residential, mobile home park, overpass, parking lot, runway, sparse residential, and tennis courts.

Finally, block-level LULC distributions are computed using equation 4.3 and object-level distributions are computed using equation 4.4.

### Evaluation

We quantitatively evaluate the results by computing two values. First, how much of the true spatial extent is selected, and second, how much of the estimated spatial extent does not belong to the true spatial extent. We want the first to be large and the second to be small. These values are equivalent to true positive and false positive rates.

Given a target image with groundtruth spatial extent $L_{true}$, and estimated spatial extent $L_{est}$ corresponding to a specific setting of the relevancy threshold $\theta$, we compute precision as the fraction of the estimated region that actually belongs to the true spatial extent:

$$precision = \frac{|L_{est} \bigcap L_{true}|}{|L_{est}|} \tag{4.7}$$

We compute recall as the fraction of the true spatial extent that appears in the estimated region:

$$recall = \frac{|L_{est} \bigcap L_{true}|}{|L_{true}|}\ , \tag{4.8}$$

In these equations, $|\cdot|$ indicates the area of a region in pixels and $\bigcap$ indicates set intersection. As usual, precision and recall range from 0 to 1.

We consider two cases of our problem. First, where the groundtruth spatial extent is a polygon. In this case, $L_{true}$ is the set of 128x128 pixel blocks contained within the groundtruth *polygon* (a block is considered to be inside a polygon if the majority of its area is) and $L_{est}$ is the set of blocks computed in section 4.5.4. We also consider the case where the groundtruth spatial extent is a *bounding box* (derived from a the groundtruth polygon). In this case, $L_{est}$ is the bounding box encompassing the set of blocks computed in section 4.5.4.

**Experiments**

We compare two different learning regimes. First, the fully supervised case where the object models learned using strongly labelled groundtruth data only; i.e., images in which the spatial extent of an object has been manually delineated. And, second, the semi-supervised case where the object model is learned using a combination of strongly and weakly labelled training data. The weakly labelled data are images in which the object has not been delineated. The significance here is that such weakly labelled training data can be automatically generated using existing gazetteers even though they represent the spatial extent using only a single latitude/longitude point. This point can be used to retrieve imagery from the National Map or other image repository which should contain the object roughly centered. The query region is chosen to be larger than the typical size of the particular object type.

To know how much the improvement can be provided by the weakly labelled data, the strongly labelled data consists of *only one manually labelled image* in both learning regimes. The LULC distribution of the groundtruth region as computed using equation

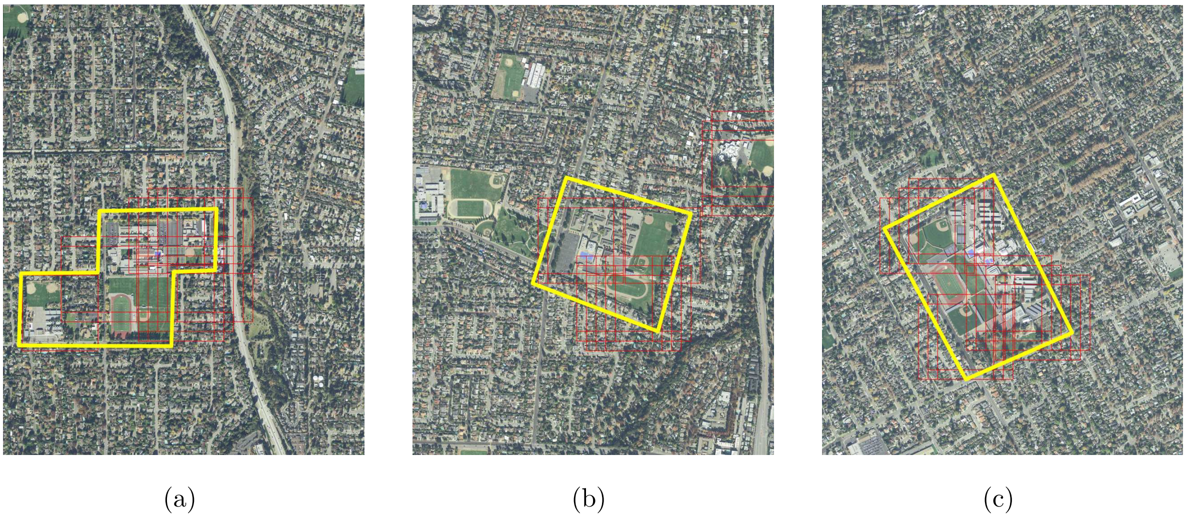
4.4 is the single relevant example. The LULC distributions over windows outside the object region are the irrelevant examples.

The unlabelled examples in the semi-supervised learning regime are the LULC distributions over windows from a set of weakly labelled images. We equally weight the labelled and unlabelled data in the learning as described in [99].

We evaluate performance using cross-validation. Each image in the groundtruth dataset is taken separately as the labelled image. The rest of the images are separated equally into unlabelled training data and test data. Both learning regimes see the single labelled image. The semi-supervised regime also sees the unlabelled training data. The ranking function is then applied to each of the test images separately. For each test image, a set of precision-recall values are computed as the relevancy threshold $\theta$ is varied. The final set of precision-recall values is computed by averaging over all trials in the cross-validation (one for each image in the groundtruth dataset). We also compute a single average precision value from this final set.

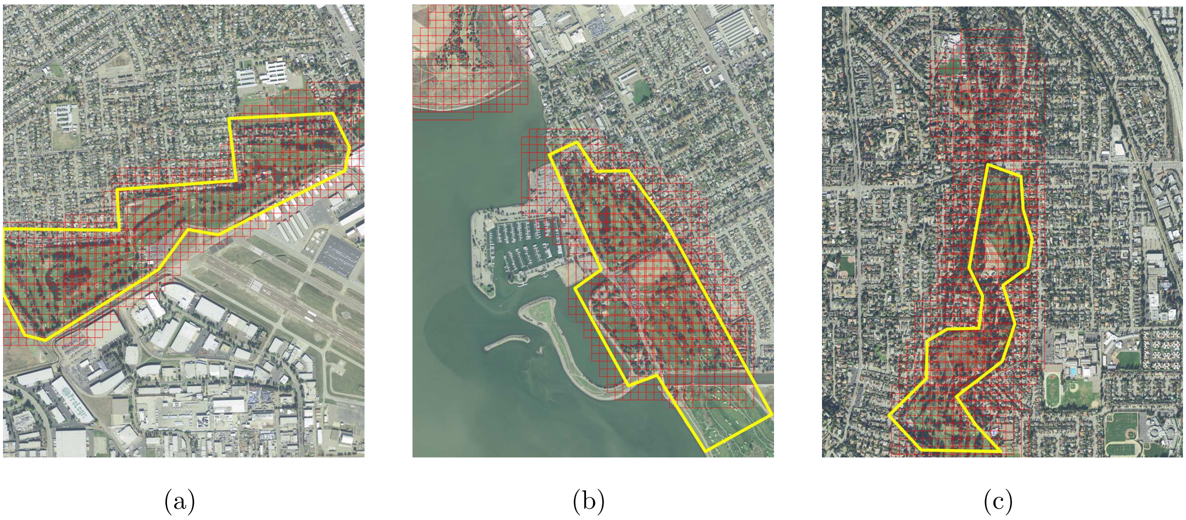
### Results

**Qualitative Results Using Supervised Learning** Results for three samples of each object type are shown in figures 4.4-4.7. In these results, the yellow polygons indicate the groundtruth spatial extents and the union of the red regions indicate the estimated spatial extents for the empirically chosen threshold value $\theta$. Overall, our object model is shown to be effective especially given that only one groundtruth image is available during the training process.



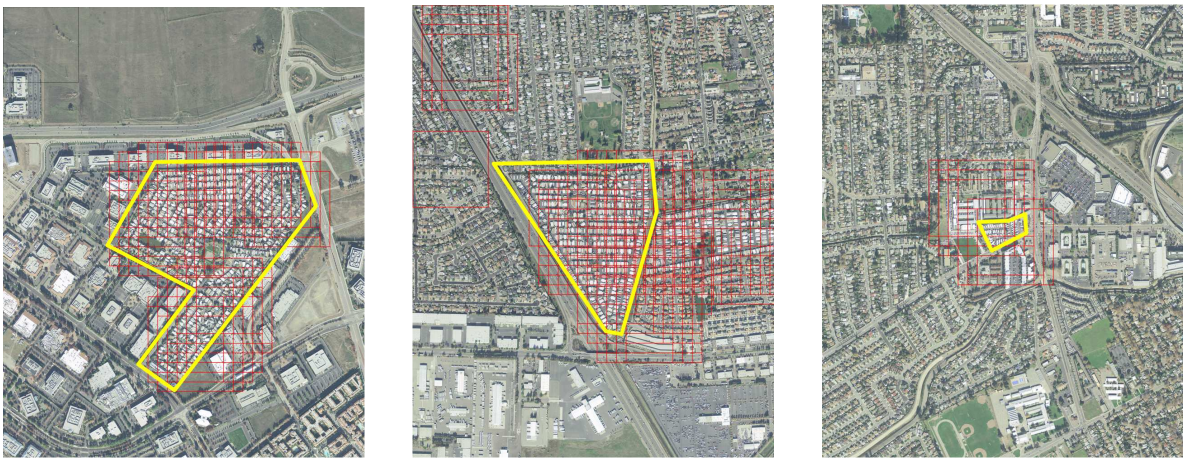
(a) (b) (c)

**Figure 4.4:** Three sample images of high schools. The yellow polygons indicate the manually delineated groundtruth spatial extents. The union of the red regions indicate the estimated spatial extents using a single labelled training image in a fully supervised learning framework.



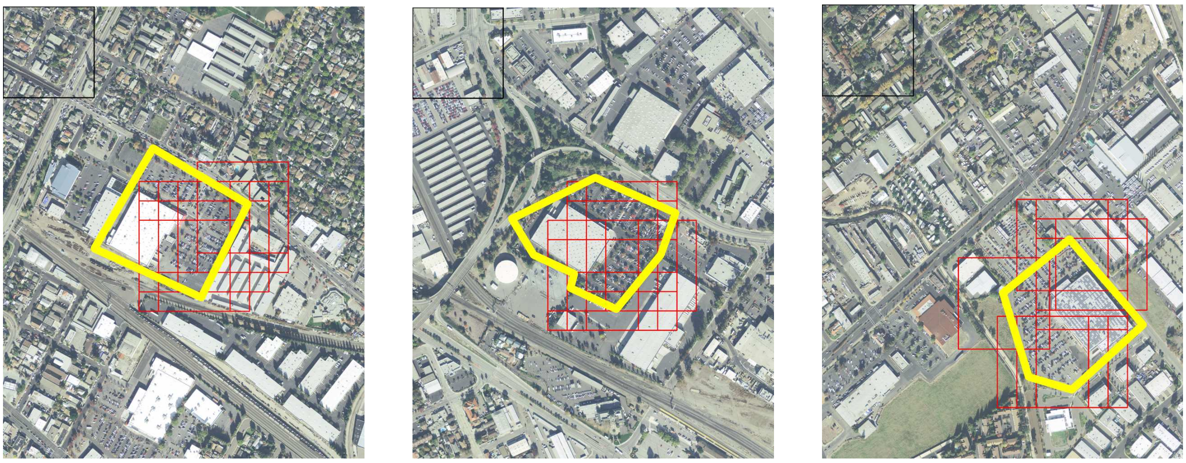
(a) (b) (c)

**Figure 4.5:** Three sample images of golf courses. The yellow polygons indicate the manually delineated groundtruth spatial extents. The union of the red regions indicate the estimated spatial extents using a single labelled training image in a fully supervised learning framework.



(a) (b) (c)

**Figure 4.6:** Three sample images of mobile home parks. The yellow polygons indicate the manually delineated groundtruth spatial extents. The union of the red regions indicate the estimated spatial extents using a single labelled training image in a fully supervised learning framework.



(a) (b) (c)

**Figure 4.7:** Three sample images of Costco shopping centers. The yellow polygons indicate the manually delineated groundtruth spatial extents. The union of the red regions indicate the estimated spatial extents using a single labelled training image in a fully supervised learning framework.

### Qualitative Comparision between Supervised and Semi-supervised Learning

Figures 4.9 through 4.12 shows the results for one instance of each object class. The left panel in each of these figures indicates the manually delineated groundtruth, the middle panel shows the results of the fully supervised case, and the right panel shows the results of the semi-supervised case. The rectangles in the middle and right panels are the bounding boxes of the windows detected for an empirically chosen relevancy threshold. (The same threshold value is used for the fully and semi-supervised cases.) These bounding boxes could be used to revised the gazetteer entries for these objects which currently specify the middle of the image as the spatial extent. Clearly, the bounding boxes computed using the proposed semi-supervised approach are a more accurate estimate of the spatial extents of the objects than those computed using the fully supervised approach.

### Quantitative Comparision between Supervised and Semi-supervised Learning

Figure 4.8 shows the precision-recall curves for the four object types. Our results show that incorporating the weakly labelled training samples provided by the gazetteers improves the object appearance models and thus the the spatial extent estimation. The results for the fully supervised learning regime are shown using blue x's. The results for the semi-supervised regime are shown using red squares. The proposed semi-supervised regime results in higher precision at almost all values of recall, the exceptions being at very high recall values where there is not much difference.

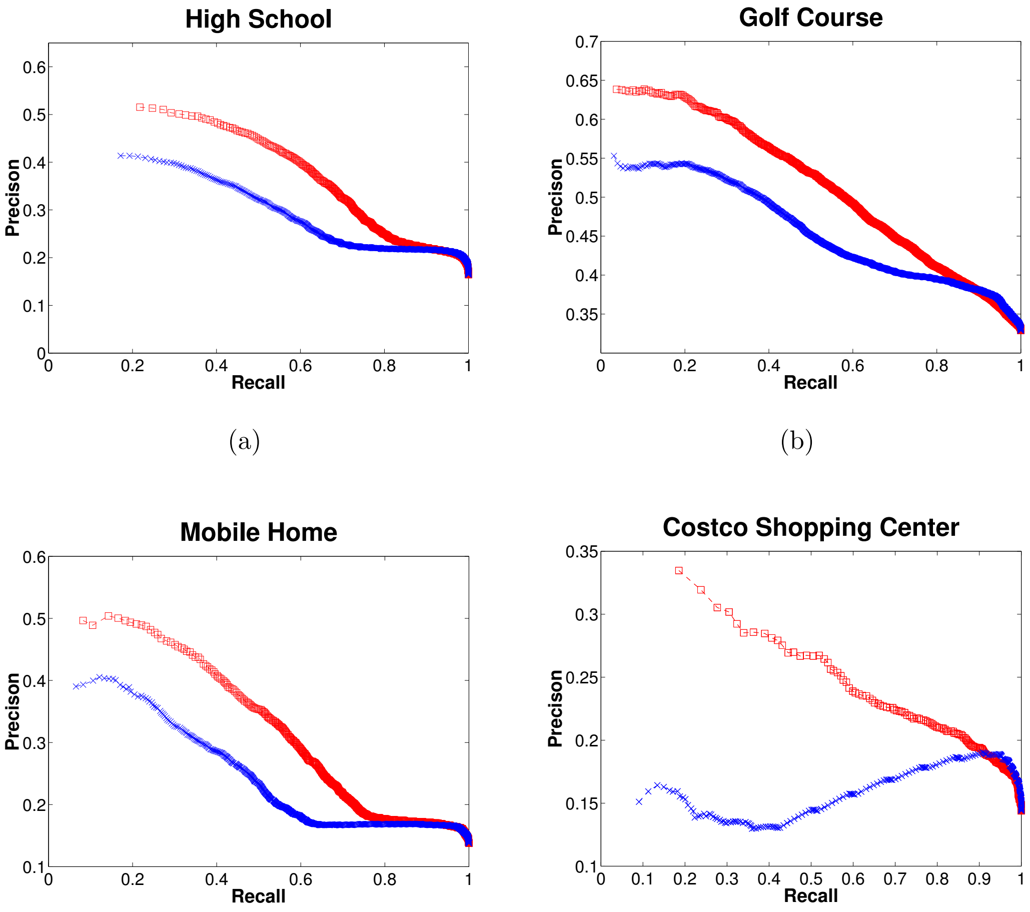
**Table 4.6:** Average precision values for the two learning regimes.

| Learning | HS | GC | MHP | Costco |
|---|---|---|---|---|
| Fully Supervised | 0.316 | 0.460 | 0.260 | 0.190 |
| Semi-Supervised | 0.401 | 0.518 | 0.340 | 0.260 |

The average precision for the four object types are listed in Table 4.6. These results again demonstrate that the semi-supervised regime improves over the fully supervised one.

### Discussion

The most salient aspect of the hierarchical model is the latent intermediate level. First, by characterizing the LULC classes that constitute an object, it allows our approach to bridge the gap between the low-level features and the high-level objects. Second, it allows us to model complex objects which are composed of multiple LULC classes. And, finally, its effectiveness is due as much to it modeling the LULC classes that do not appear in an object as those that do. In particular, the large number of LULC classes allows the windowing step to readily reject background regions that have high proportions of classes which do not appear in the object. Such discrimination would not be possible using binary single-class LULC classifiers (and then again, would only be applicable to homogeneous objects). The results above show the potential of use our object model to estimate the spatial extents of geospatial objects. While not perfect, this level of accuracy for the spatial extent is a big improvement over the single latitude/longitude point currently present in gazetteers. The results also demonstrate that by incorporating



(a)

(b)

(c)

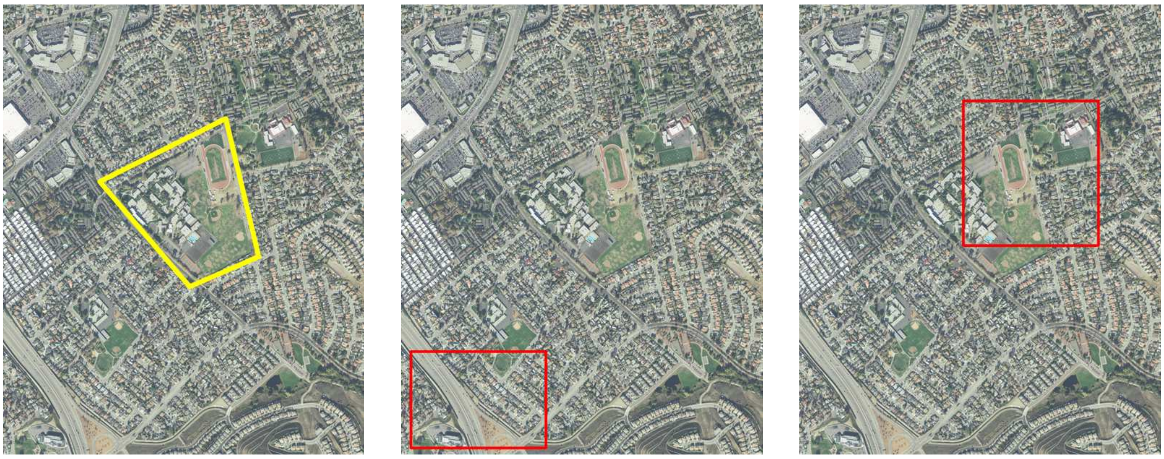
(d)

**Figure 4.8:** Precision-recall curves for the four object types. The results for the fully supervised learning regime are shown using blue x's. The results for the proposed semi-supervised regime are shown using red squares.

weakly labelled training data the standard supervised framework can be improved upon. This is significant for the integration of gazetteer and image data.

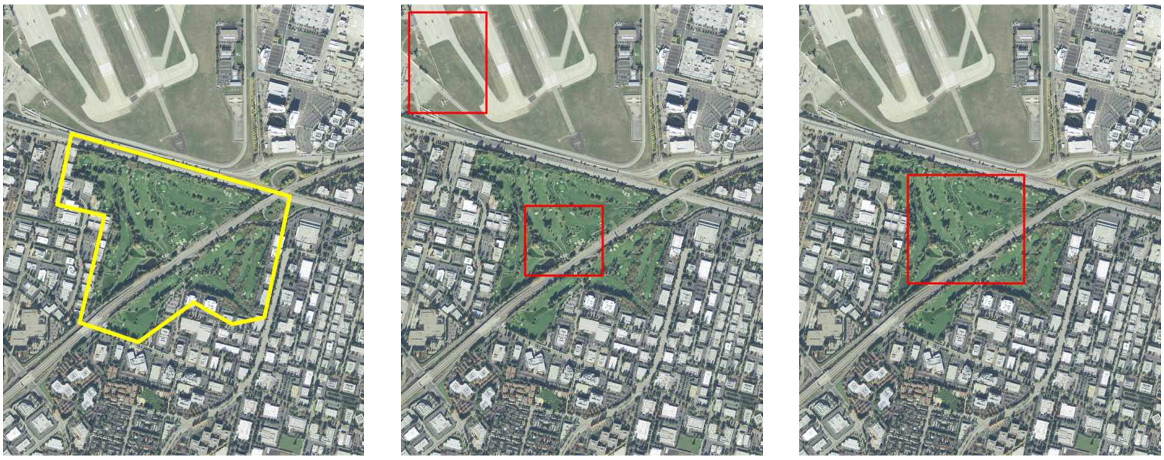
## 4.6 Conclusion

In this chapter, we have explored the potential synergy between gazetteers and overhead imagery and proposed a framework to integrate them together. Further, we have proposed a hierarchical model that can use image data generated from the latitude/longitude point locations from the gazetteers to estimate the spatial extents of geospatial objects and thus to update the gazetteers with a bounding box presentation.



(a) (b) (c)

**Figure 4.9:** Results for an instance of the high school class. (a) The manually delineated groundtruth. (b) The bounding box as detected using the fully supervised appraoch. (c) The bounding box as detected using the proposed semi-supervised approach.



(a) (b) (c)

**Figure 4.10:** Results for an instance of the golf course class. (a) The manually delineated groundtruth. (b) The bounding box as detected using the fully supervised approach (here, two sets of connect windows are detected and thus two bounding boxes are generated). (c) The bounding box as detected using the proposed semi-supervised approach.

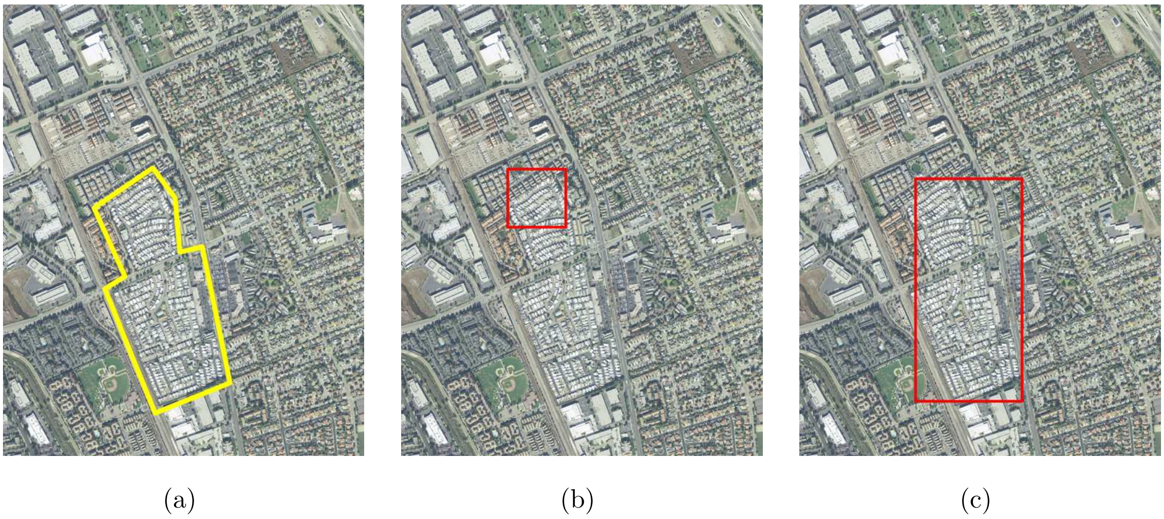

(a) (b) (c)

**Figure 4.11:** Results for an instance of the mobile home park class. (a) The manually delineated groundtruth. (b) The bounding box as detected using the fully supervised approach. (c) The bounding box as detected using the proposed semi-supervised approach.

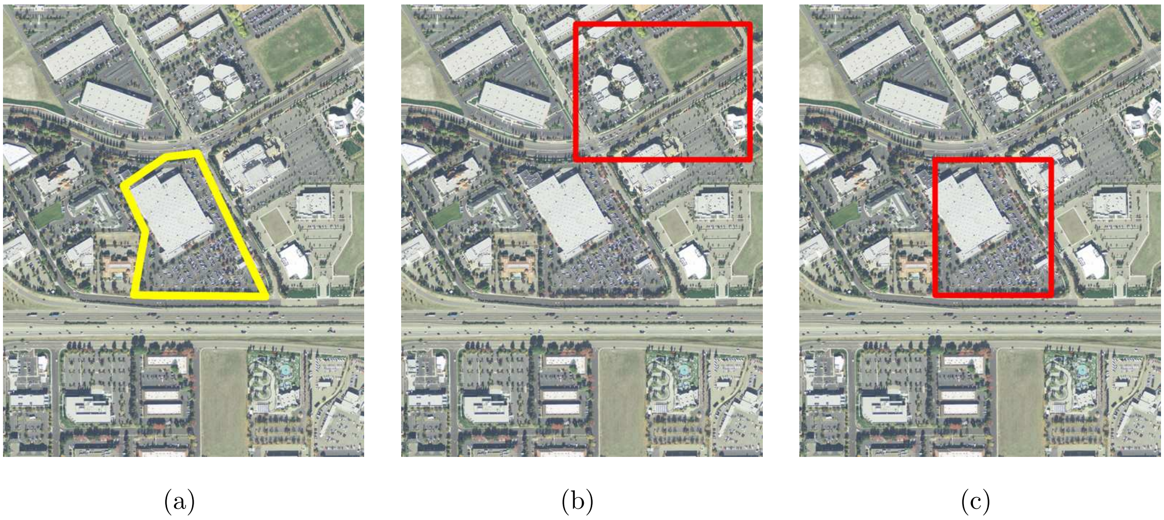

(a) (b) (c)

**Figure 4.12:** Results for an instance of the Costco shopping center class. (a) The manually delineated groundtruth. (b) The bounding box as detected using the fully supervised approach. (c) The bounding box as detected using the proposed semi-supervised approach.

# Chapter 5

# Conclusion

This dissertation has explored the use of local invariant features in the analysis of remote sensed imagery in three aspects: image retrieval, LULC classification, and detection of geospatial objects. It has been demonstrated that the local features can meet the challenges of analyzing higher resolution, cheaper-to-get, and easier-to-access overhead images that has become available over the last decade.

We have performed an extensive evaluation of local invariant features for image retrieval of LULC classes. We have empirically evaluated the effects of a number of design parameters on a BOVW representation including feature extraction, the size of the visual codebook, and the dissimilarity measure used to compare the BOVW representations. We have also performed comparisons with standard features such as color and texture.

The second problem this work addresses is using local invariant features to label LULC classes. We have applied the BOVM approach to LULC classification and proposed to capture the geometric aspect of images by taking into account the spatial arrangement of local features which is usually neglected by the standard BOVM approach. The proposed SPCK approach was motivated by the fact that overhead images generally do not have an absolute reference frame and thus the relative spatial layout may be key in discriminating among different LULC classes. To this end, we have proposed to compute the co-occurrences of visual words with respect to spatial predicates over a hierarchical partitioning of an image such that both the absolute and relative layout of the image can be captured.

One of the fundametal contributions of this work is the construction of a first-of-its-kind LULC groundtruth dataset. The dataset offers 21 classes and 100 images of each class, the largest of its kind. The dataset consists of royalty free images and has been made publicly available to other researchers through our lab's website. We anticipate this will serve as a standardized dataset for analyzing different techniques in the remote sensing community.

In addition, we have also explored the potential synergy between non-image and image geospatial data. We have proposed a novel framework to fuse together gazetteers and overhead images such that the information can flow both ways: on one hand, the images of archived instances in the gazeteers retrieved around the latitude/longitude point locations provided by the gazetteers can be treated as either strongly or weakly labelled samples through which the appearance model of the archived complex geospatial objects can be learned; on the other hand, the trained appearance model can be used to estimate the spatial extents of geospatial objects of the same type so that the point representation in the gazetteers can be replaced by a bounding box. This helps enable fully automatic geographic information management and integration.

# Bibliography

[1] Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: IEEE Conference on Computer Vision and Pattern Recognition. Volume 2. (2006) 2169–2178

[2] Opelt, A., Fussenegger, M., Pinz, A., Auer, P.: Weak hypotheses and boosting for generic object detection and recognition. In: European Conference on Computer Vision. Volume 3022. (2004) 71–84

[3] Ling, H., Soatto, S.: Proximity distribution kernels for geometric context in category recognition. In: IEEE International Conference on Computer Vision. (2007) 1–8

[4] Boiman, O., Shechtman, E., Irani, M.: In defense of nearest-neighbor based image classification. In: IEEE Conference on Computer Vision and Pattern Recognition. (2008) 1 –8

[5] Bhagavathy, S., Manjunath, B.: Modeling and detection of geospatial objects using texture motifs. IEEE Transactions on Geoscience and Remote Sensing **44** (2006) 3706–3715

[6] Hongyu, Y., Bicheng, L., Wen, C.: Remote sensing imagery retrieval based-on Gabor texture feature classification. In: International Conference on Signal Processing. (2004) 733–736

[7] Li, Y., Bretschneider, T.: Semantics-based satellite image retrieval using low-level features. In: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium. Volume 7. (2004) 4406–4409

[8] Manjunath, B.S., Ma, W.Y.: Texture features for browsing and retrieval of image data. IEEE Transactions on Pattern Analysis and Machine Intelligence **18** (1996) 837–842

[9] Unsalan, C., Boyer, K.: Classifying land development in high-resolution panchromatic satellite images using straight-line statistics. IEEE Transactions on Geoscience and Remote Sensing **42** (2004) 907–919

[10] Unsalan, C., Boyer, K.: A theoretical and experimental investigation of graph theoretical measures for land development in satellite imagery. IEEE Transactions on Pattern Analysis and Machine Intelligence **27** (2005) 575–589

[11] Kim, T., Muller, J.P.: Development of a graph-based approach for building detection. Image and Vision Computing **17** (1999) 3 – 14

[12] Gamba, P., Dell' Acqua, F., Lisini, G., Trianni, G.: Improved VHR urban area mapping exploiting object boundaries. IEEE Transactions on Geoscience and Remote Sensing **45** (2007) 2676–2682

[13] Benediktsson, J., Pesaresi, M., Amason, K.: Classification and feature extraction for remote sensing images from urban areas based on morphological transformations. IEEE Transactions on Geoscience and Remote Sensing **41** (2003) 1940–1949

[14] Zhong, P., Wang, R.: Using combination of statistical models and multilevel structural information for detecting urban areas from a single gray-level image. IEEE Transactions on Geoscience and Remote Sensing **45** (2007) 1469–1482

[15] Zhong, P., Wang, R.: A multiple conditional random fields ensemble model for urban area detection in remote sensing optical images. IEEE Transactions on Geoscience and Remote Sensing **45** (2007) 3978–3988

[16] Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Gool, L.V.: A comparison of affine region detectors. International Journal of Computer Vision **65** (2005) 43–72

[17] Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence **27** (2005) 1615–1630

[18] Harris, C., Stephens, M.: A combined corner and edge detector. In: Proceedings of The Fourth Alvey Vision Conference. (1988) 147–151

[19] Mikolajczyk, K., Schmid, C.: Scale and affine invariant interest point detectors. International Journal of Computer Vision **60** (2004) 63–86

[20] Lindeberg, T.: Feature detection with automatic scale selection. International Journal of Computer Vision **30** (1998) 79–116

[21] Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: IEEE Conference on Computer Vision and Pattern Recognition. Volume 2. (2006) 2161–2168

[22] Snaptell: (Snaptell - visual product search) http://snaptell.com/.

[23] Mikolajczyk, K., Schmid, C.: Indexing based on scale invariant interest points. In: IEEE International Conference on Computer Vision. Volume 1. (2001) 525–531

[24] Schaffalitzky, F., Zisserman, A.: Automated scene matching in movies. In: Proceedings of the International Conference on Image and Video Retrieval, Springer-Verlag (2002) 186–197

[25] Lazebnik, S., Schmid, C., Ponce, J.: A sparse texture representation using affine-invariant regions. In: IEEE Conference on Computer Vision and Pattern Recognition. Volume 2. (2003) 319–324

[26] Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: IEEE International Conference on Computer Vision. Volume 2. (2003) 1470–1477

[27] Yang, Y., Newsam, S.: Geographic image retrieval using local invariant features. IEEE Transactions on Geoscience and Remote Sensing **PP** (2012) 1 –15

[28] Lowe, D.G.: Object recognition from local scale-invariant features. In: IEEE International Conference on Computer Vision. Volume 2. (1999) 1150–1157

[29] Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence **24** (2002) 509–522

[30] Kadir, T., Brady, M.: Saliency, scale and image description. International Journal of Computer Vision **45** (2001) 83–105

[31] Zhou, H., Yuan, Y., Shi, C.: Object tracking using SIFT features and mean shift. Computer Vision and Image Understanding **113** (2008) 345–352

[32] Agarwal, A., Triggs, B.: A local basis representation for estimating human pose from cluttered images. In: Proceedings of the Asian Conference on Computer Vision. Volume 1. (2006) 50–59

[33] Ohbuchi, R., Osada, K., Furuya, T., Banno, T.: Salient local visual features for shape-based 3d model retrieval. In: IEEE International Conference on Shape Modeling and Applications. (2008) 93–102

[34] Wu, C., Clipp, B., Li, X., Frahm, J.M., Pollefeys, M.: 3D model matching with viewpoint-invariant patches (vip). In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. (2008) 1–8

[35] Porway, J., Wang, K., Yao, B., Zhu, S.C.: A hierarchical and contextual model for aerial image understanding. In: IEEE Conference on Computer Vision and Pattern Recognition. (2008) 1–8

[36] Sirmacek, B., Unsalan, C.: Urban-area and building detection using SIFT keypoints and graph theory. IEEE Transactions on Geoscience and Remote Sensing **47** (2009) 1156–1167

[37] Newsam, S., Yang, Y.: Geographic image retrieval using interest point descriptors. In: Proceedings of the International Conference on Advances in Visual Computing. Volume 2. (2007) 275–286

[38] Newsam, S., Yang, Y.: Comparing global and interest point descriptors for similarity retrieval in remote sensed imagery. In: Proceedings of the Annual ACM International Symposium on Advances in Geographic Information Systems. (2007) 9:1–9:8

[39] Ashley, J., Flickner, M., Hafner, J., Lee, D., Niblack, W., Petkovic, D.: The query by image content (QBIC) system. In: Proceedings of the ACM SIGMOD international conference on Management of data. (1995) 475

[40] Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. In: Penn State University Technical Report CSE 06-009. (2006)

[41] Bretschneider, T., Cavet, R., Kao, O.: Retrieval of remotely sensed imagery using spectral information content. In: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium. (2002) 2253–2255

[42] Bretschneider, T., Kao, O.: A retrieval system for remotely sensed imagery. In: International Conference on Imaging Science, Systems, and Technology. Volume 2. (2002) 439–445

[43] Ma, A., Sethi, I.K.: Local shape association based retrieval of infrared satellite images. In: IEEE International Symposium on Multimedia. (2005) 551–557

[44] Li, C.S., Castelli, V.: Deriving texture feature set for content-based retrieval of satellite image database. In: IEEE International Conference on Image Processing. Volume 1. (1997) 576–579

[45] Newsam, S., Wang, L., Bhagavathy, S., Manjunath, B.S.: Using texture to analyze and manage large collections of remote sensed image and video data. Journal of Applied Optics: Information Processing **43** (2004) 210–217

[46] Newsam, S., Kamath, C.: Retrieval using texture features in high resolution multispectral satellite imagery. In: SPIE Defense and Security Symposium, Data Mining and Knowledge Discovery: Theory, Tools, and Technology VI. (2004)

[47] Schmid, C., Mohr, R.: Local grayvalue invariants for image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence **19** (1997) 530–535

[48] Tian, Q., Sebe, N., Lew, M., Loupias, E., Huang, T.: Content-based image retrieval using wavelet-based salient points. In: SPIE International Symposium on Electronic Imaging, Storage and Retrieval for Media Databases. (2001)

[49] Wang, J., Zha, H., Cipolla, R.: Combining interest points and edges for content-based image retrieval. In: IEEE International Conference on Image Processing. (2005) 1256–1259

[50] Zhang, H., Rahmani, R., Cholleti, S.R., Goldman, S.A.: Local image representations using pruned salient points with applications to CBIR. In: Proceedings of the Annual ACM International Conference on Multimedia. (2006) 287–296

[51] Hsu, C.T., Shih, M.C.: Content-based image retrieval by interest points matching and geometric hashing. In: SPIE Photonics Asia Conference. Volume 4925. (2002) 80–90

[52] Wolf, C., Kropatsch, W., Bischof, H., Jolion, J.M.: Content based image retrieval using interest points and texture features. In: International Conference on Pattern Recognition. Volume 4. (2000) 234–237

[53] Ledwich, L., Williams, S.: Reduced SIFT features for image retrieval and indoor localisation. In: Australasian Conference on Robotics and Automation. (2004)

[54] Manjunath, B.S., Salembier, P., Sikora, T., eds.: Introduction to MPEG7: Multimedia Content Description Interface. John Wiley & Sons (2002)

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

[56] Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60** (2004) 91–110

[57] Hand, D., Mannila, H., Smyth, P.: Principles of Data Mining. The MIT Press (2001)

[58] Kuhn, H.W.: The Hungarian Method for the assignment problem. Naval Research Logistic Quarterly **2** (1955) 83–97

[59] Rubner, Y., Tomasi, C., Guibas, L.: A metric for distributions with applications to image databases. In: IEEE International Conference on Computer Vision. (1998) 59–66

[60] Yang, Y., Newsam, S.: Comparing SIFT descriptors and Gabor texture features for classification of remote sensed imagery. In: IEEE International Conference on Image Processing. (2008) 1852 –1855

[61] Yang, Y., Newsam, S.: Bag-of-visual-words and spatial extensions for land-use classification. In: Proceedings of the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. (2010) 270–279

[62] Yang, Y., Newsam, S.: Spatial pyramid co-occurrence for image classification. In: IEEE International Conference on Computer Vision. (2011) 1465 –1472

[63] Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood estimation from incomplete data via the EM algorithm. Journal of the Royal Statistical Society **39** (1977) 1–38

[64] Chang, C.C., Lin, C.J.: (LIBSVM–a library for support vector machines) http://www.csie.ntu.edu.tw/ cjlin/libsvm/.

[65] Wilkinson, G.G.: Results and implications of a study of fifteen years of satellite image classification experiments. IEEE Transactions on Geoscience and Remote Sensing **43** (2005) 433–440

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

[67] Grauman, K., Darrell, T.: The pyramid match kernel: Discriminative classification with sets of image features. In: IEEE International Conference on Computer Vision. Volume 2. (2005) 1458–1465

[68] Haralick, R.M., Shanmugam, K., Dinstein, I.: Texture features for image classification. IEEE Transactions on Systems, Man, and Cybernetics **3** (1973) 610–621

[69] Savarese, S., Winn, J., Criminisi, A.: Discriminative object class models of appearance and shape by correlatons. In: IEEE Conference on Computer Vision and Pattern Recognition. Volume 2. (2006) 2033–2040

[70] Huang, J., Kumar, S.R., Mitra, M., Zhu, W.J., Zabih, R.: Image indexing using color correlograms. In: IEEE Conference on Computer Vision and Pattern Recognition. (1997) 762–768

[71] Liu, D., Hua, G., Viola, P., Chen, T.: Integrated feature selection and higher-order spatial feature extraction for object categorization. In: IEEE Conference on Computer Vision and Pattern Recognition. (2008) 1–8

[72] Berg, A., Berg, T., Malik, J.: Shape matching and object recognition using low distortion correspondences. In: IEEE Conference on Computer Vision and Pattern Recognition. Volume 1. (2005) 26–33

[73] Leordeanu, M., Hebert, M., Sukthankar, R.: Beyond local appearance: Category recognition from pairwise interactions of simple features. In: IEEE Conference on Computer Vision and Pattern Recognition. (2007) 1–8

[74] Bosch, A., Zisserman, A., Munoz, X.: Representing shape with a spatial pyramid kernel. In: Proceedings of the ACM International Conference on Image and Video Retrieval. (2007) 401–408

[75] Yuan, J., Wu, Y., Yang, M.: Discovery of collocation patterns: from visual words to visual phrases. In: IEEE Conference on Computer Vision and Pattern Recognition. (2007) 1–8

[76] Newsam, S., Yang, Y.: Integrating gazetteers and remote sensed imagery. In: Proceedings of the ACM SIGSPATIAL international conference on Advances in geographic information systems. (2008) 26:1–26:10

[77] Yang, Y., Newsam, S.: Estimating the spatial extents of geospatial objects using hierarchical models. In: IEEE Workshop on Applications of Computer Vision. (2012) 305 –312

[78] Quack, T., Mönich, U., Thiele, L., Manjunath, B.S.: Cortina: a system for large-scale, content-based web image retrieval. In: Proceedings of the Annual ACM International Conference on Multimedia. (2004) 508–511

[79] Newsam, S., Sumengen, B., Manjunath, B.: Category-based image retrieval. In: IEEE International Conference on Image Processing. Volume 3. (2001) 596–599

[80] Barnard, K., Forsyth, D.: Learning the semantics of words and pictures. In: IEEE International Conference on Computer Vision. Volume 2. (2001) 408–415

[81] Berg, T., Berg, A., Edwards, J., Maire, M., White, R., Teh, Y.W., Learned-Miller, E., Forsyth, D.: Names and faces in the news. In: IEEE Conference on Computer Vision and Pattern Recognition. Volume 2. (2004) 848–854

[82] Barnard, K., Duygulu, P., Forsyth, D.: Clustering art. In: IEEE Conference on Computer Vision and Pattern Recognition. Volume 2. (2001) 434–441

[83] Li, L.J., Wang, G., Fei-Fei, L.: Optimol: automatic Online Picture collecTion via Incremental MOdel Learning. In: IEEE Conference on Computer Vision and Pattern Recognition. (2007) 1–8

[84] Cooper, M., Foote, J., Girgensohn, A., Wilcox, L.: Temporal event clustering for digital photo collections. In: Proceedings of the Annual ACM International Conference on Multimedia. (2003) 364–373

[85] Naaman, M.: Eyes on the world. Computer **39** (2006) 108–111

[86] ZoneTag: (Zonetag research prototype) http://zonetag.research.yahoo.com.

[87] Flickr: (Flickr photo sharing) http://www.flickr.com.

[88] Zhang, C.: Towards an operational system for automated updating of road databases by integration of imagery and geodata. ISPRS Journal of Photogrammetry and Remote Sensing **58** (2004) 166–186

[89] Agouris, P., Gyftakis, S., Stefanidis, A.: Using a fuzzy supervisor for object extraction within an integrated geospatial environment. International Archives of Photogrammetry and Remote Sensing **32** (1998) 191–195

[90] Doucette, P., Agouris, P., Musavi, M., Stefanidis, A.: Automated extraction of linear features from aerial imagery using Kohonen learning and GIS data. In: Integrated Spatial Databases. Volume 1737. (1999) 20–33

[91] Baumgartner, A., Eckstein, W., Mayer, H., Heipke, C., Ebner, H.: Context-supported road extraction. Automatic Extraction of Man-Made Objects from Aerial and Space Images **2** (1997) 299–308

[92] Agouris, P., Beard, K., Mountrakis, G., Stefanidis, A.: Capturing and modeling geographic object change: A spatiotemporal gazetteer framework. Photogrammetric Engineering & Remote Sensing **66** (2000) 1241–1250

[93] Bailloeul, T., Duan, J., Prinet, V., Serra, B.: Urban digital map updating from satellite high resolution images using GIS data as a priori knowledge. In: Proceedings of the GRSS/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas. (2003) 283–287

[94] Baltsavias, E.P.: Object extraction and revision by image analysis using existing geodata and knowledge: current status and steps towards operational systems. ISPRS Journal of Photogrammetry and Remote Sensing **58** (2004) 129–151

[95] Walter, V., Fritsch, D.: Automatic verification of GIS data using high resolution multispectral data. International Archives of Photogrammetry and Remote Sensing **32** (1998) 485–490

[96] Michalowski, M., Knoblock, C.A., Bayer, K., Choueiry, B.Y.: Exploiting automatically inferred constraint-models for building identification in satellite imagery. In: Proceedings of the Annual ACM International Symposium on Advances in Geographic Information Systems. (2007) 35–42

[97] Hill, L.L., Frew, J., Zheng, Q.: Geographic names: The implementation of a gazetteer in a georeferenced digital library. D-Lib **5** (1999)

[98] Alexandria Digital Library Gazetteer. 1999- . Santa Barbara CA: (Map and Imagery Lab, Davidson Library, University of California, Santa Barbara. Copyright UC Regents.) http://www.alexandria.ucsb.edu/gazetteer.

[99] Amini, M.R., Truong, T.V., Goutte, C.: A boosting algorithm for learning bipartite ranking functions with partially labeled data. In: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval. (2008) 99–106