Content and context analysis for mobile landmark recognition

Tao, Chen

2013

Tao, C. (2013). Content and context analysis for mobile landmark recognition. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/54669>

<https://doi.org/10.32657/10356/54669>

NANYANG
TECHNOLOGICAL
UNIVERSITY

Content and Context Analysis for Mobile Landmark Recognition

Tao Chen

School of Electrical and Electronic Engineering

2013

Content and Context Analysis for Mobile Landmark Recognition

Tao Chen

School of Electrical and Electronic Engineering

A thesis submitted to the Nanyang Technological University

in partial fulfillment of the requirement for the degree of

Doctor of Philosophy

2013

Acknowledgements

First and foremost, I would like to express my deep and sincere gratitude and appreciation to my supervisor, Associate Professor Yap Kim Hui, for his constructive advice, invaluable insight, endless patience, continuous encouragement and guidance in the whole period of my PhD study. It makes the thesis possible with his kind supervision and help. He has been a strong and supportive adviser to me throughout my graduate school life in NTU, but he has always given me great freedom to pursue independent work. Especially, when I am working with him as a PhD student and junior researcher, his research enthusiasm and passion in science and technology has inspired me. Not only do I gain the technical knowledge but also the logical reasoning for research work. I am sure what I have gained from him will benefit me for the rest of my life.

I would like to thank my co-supervisor, Associate Professor Chau Lap Pui for his valuable advice and the enlightening discussion to improve my work. His support, guidance and helpful suggestions have served me well. I also would like to thank Associate Professor Chan Kap Luk for his positive suggestions and concerns for my work. Their thoughtful comments and encouragement are highly appreciated.

I would like to acknowledge Nanyang Technological University for its financial and facility support during the entire period of my PhD candidature. I also would like to thank all the members and staff of the Media Technology Laboratory for providing me with a comfortable and pleasant working environment for doing my research.

Last but not least, I would like to thank all my family members for their love and care all these years. A PhD degree does not mean the end. Instead, it is the beginning of my future research life.

Table of Contents

Acknowledgements.....	i
Table of Contents.....	ii
Summary.....	vi
List of Figures.....	ix
List of Tables.. ..	xiii
List of Abbreviations	xv
Chapter 1 Introduction	1
1.1 Motivation	1
1.2 Overview	2
1.2.1 Comparison of Mobile and Desktop-based Landmark Recognition.....	2
1.2.2 Challenges of Mobile-based Landmark Recognition	4
1.3 Contributions of This Thesis	6
1.4 Organization	9
Chapter 2 Background and Literature Review	12
2.1 Related Work.....	12
2.2 Content Analysis.....	14
2.2.1 Global Features	15
2.2.2 Local Features.....	16
2.3 Classification Algorithms	23
2.4 Context Analysis.....	27
2.5 Content and Context Integration.....	28
2.6 Scalable Landmark Recognition.....	28
2.6.1 Scalable Vocabulary Tree	29
2.6.2 Comparison of SVM and SVT.....	30

2.7	Summary.....	31
Chapter 3	Integrated Content and Context Analysis for Mobile Landmark Recognition.	32
3.1	Introduction	32
3.2	Framework Overview	33
3.3	Content Analysis.....	34
3.3.1	Spatial Pyramid Decomposition	36
3.3.2	Pyramid Cell Saliency Estimation	37
3.3.3	Spatial Pyramid Kernel	43
3.4	Integration of Context and Content Analysis	44
3.5	Experimental Evaluation	48
3.5.1	Experimental Setup.....	48
3.5.2	Performance Evaluation of Proposed Content Analysis	50
3.5.3	Performance Evaluation of Context Analysis	54
3.5.4	Performance Comparison of Content, Context and Their Integration	57
3.5.5	Experiments on the Oxford Buildings Dataset	59
3.6	Summary.....	60
Chapter 4	Discriminative Learning BoW for Mobile Landmark Recognition	61
4.1	Introduction	61
4.2	Related Work.....	62
4.3	Framework Overview	64
4.4	Discriminative BoW	66
4.4.1	Iterative Patch Learning	66
4.4.2	BoW Histogram Generation	68
4.5	Fuzzy SVM Training	70
4.5.1	Image Signature Weight Estimation	70

4.5.2	FSVM Training.....	73
4.6	Content and Context Integration Analysis.....	75
4.7	Experimental Evaluation	75
4.7.1	Illustration of the Proposed Patch Learning Method	76
4.7.2	Performance Evaluation on NTU50Landmark Database	76
4.7.3	Performance Evaluation on Oxford Database.....	80
4.7.4	Performance Evaluation of Content-Context Integration	80
4.7.5	Evaluation of The Combined Methods from Chapters 3 and 4	81
4.8	Summary.....	82
Chapter 5	Bag-of- Phrase for Mobile Landmark Recognition.....	83
5.1	Introduction	83
5.2	Related Work.....	84
5.3	Framework Overview	86
5.4	DVP Selection	87
5.4.1	Word-level Semantic Similarity Metric.....	90
5.4.2	Phrase-level Semantic Similarity Metric	95
5.4.3	Visual Phrase Ranking.....	96
5.5	Soft BoP Encoding	98
5.6	Experimental Results.....	100
5.6.1	Comparison of BoW, BoP and their Combination	100
5.6.2	Comparison with Other Phrase Selection Methods	101
5.6.3	Comparison with Hard BoP Quantization	102
5.6.4	Comparison with Other BoP Methods.....	103
5.6.5	Evaluation of Content-Context Integration.....	104
5.6.6	Performance Evaluation on the Oxford Database.....	105

5.6.7	Evaluation of The Combined Methods from Chapters 3, 4 and 5	105
5.7	Chapter Summary	106
Chapter 6	Context-Aware Vocabulary Learning for Mobile Landmark Recognition	107
6.1	Introduction	107
6.2	Related Work.....	108
6.3	Framework Overview	110
6.4	Context-aware Vocabulary Learning.....	111
6.4.1	DSVT Construction	113
6.4.2	ImageRank Analysis	115
6.4.3	DCV Learning.....	120
6.4.4	DSVT-based Context-Content Integration	125
6.5	Experimental Evaluation	126
6.5.1	Experimental Setup.....	126
6.5.2	Performance Comparison with Other Approaches	126
6.5.3	Evaluation of the DCV Compression	128
6.5.4	Evaluation of Recognition Time.....	130
6.6	Experimental Results on San Francisco Landmark Dataset.....	131
6.7	Performance Comparison of SVM and SVT	133
6.8	Chapter Summary	136
Chapter 7	Conclusions and Recommendations	137
7.1	Conclusions	137
7.2	Recommendations for Further Research	139
	List of Author's Publications	140
	References.....	142

Summary

In recent years, the use of mobile/cellular phones has increased greatly. Over 80 percent of the global population has become mobile cellular subscribers by 2012. Today, more than half of the mobile phones in use have camera features. Benefiting from the built-in cameras along with the advancement in network technologies, various mobile device-based applications that take advantage of the interactive features offered by the cellular phones have been designed. Amongst them, mobile landmark recognition that uses camera phone to capture a landmark and determine its related information such as its name, history and activities, is becoming increasingly popular. In view of this, this thesis focuses on mobile device-based landmark recognition and proposes to address this issue using context-aware content analysis techniques.

In Chapter 3, a new content and context integration technique for mobile landmark recognition is proposed. A bags-of-words (BoW) framework that involves a new spatial pyramid decomposition scheme is developed to perform the content analysis. The location obtained through built-in Global Positioning System (GPS) of mobile device and direction information obtained through built-in digital compass of mobile device are incorporated into the context analysis, which is then integrated with the content analysis for mobile landmark recognition.

In Chapter 4, a new BoW approach based on discriminative learning of patches, images and codewords is proposed for landmark recognition. An iterative learning approach based on a differential Gaussian mixture model (DGMM) is developed to estimate the discriminative information of each image patch. This information is then incorporated into vector quantization to generate a BoW histogram. An image signature weighting method is developed to score each image in representing its landmark category, which is then used to train a discriminative classifier through a fuzzy support vector machine (SVM). The context

information such as GPS and direction is finally integrated with the proposed content analysis to speed up the recognition time and improve the recognition performance.

In Chapter 5, different from the BoW-based methods above, a new soft bag-of-phrase (BoP) approach based on category-dependent phrase selection is proposed for mobile landmark recognition. In this chapter, the number of visual words in each phrase is chosen as two, which is named as second-order phrase. Two contributions are made in the proposed approach: (i) a discriminative selection approach that takes advantage of the word-level and phrase-level semantic similarity is developed to select the important phrases from a large number of candidates, and form the descriptive BoP dictionary, (ii) a soft encoding technique is developed to generate a BoP histogram for each image, which reduces the amount of information loss induced by conventional BoP quantization.

In Chapter 6, unlike the above methods that adopt SVM as the recognition technique, a fast landmark recognition approach based on scalable vocabulary tree (SVT) is proposed. The method constructs direction-dependent SVTs (DSVTs) for image quantization, and learns a discriminative compact vocabulary (DCV) to encode the query image. Direction information is first considered to supervise image feature clustering to construct DSVTs. Location information is then incorporated into the DCV learning algorithm, to select the discriminative codewords of the DSVT to form the DCV. An ImageRank technique and an iterative codeword selection algorithm are developed for DCV learning. Inverted indexed files are constructed for the codewords in the vocabulary, which can greatly improve the recognition efficiency.

We validate the proposed algorithms and techniques on several landmark databases, including the NTU50Landmark database created by ourselves, the Oxford building dataset, and the San Francisco landmark database. The experimental results on these landmark databases consistently show the effectiveness of the proposed methods for mobile landmark

recognition. Furthermore, the comparison with other state-of-the-art techniques indicates that the proposed algorithms achieve better performance in terms of the recognition accuracy and computational time.

List of Figures

Figure 1.1 An illustration of a mobile landmark recognition system	1
Figure 1.2 The overview of a mobile landmark recognition system	4
Figure 1.3 Mobile landmark recognition framework in this thesis.....	7
Figure 2.1 Feature representation in landmark recognition	15
Figure 2.2 An illustration of four patch selection methods, (a) keypoint-based [17], (b) random smpling-based, (c) uniform sampling-based, (d) segmentation-based [124].	17
Figure 2.3 Illustration of matching SIFT keypoints extracted from two images.....	18
Figure 2.4 Comparisionof SIFT (left) and SURF (right) in scale space sampling	19
Figure 2.5 Classification algorithms for landmark recognition	23
Figure 2.6 Illustration of SVT.....	30
Figure 3.1 Snapshots of the implemented mobile landmark recognition system in NTU	33
Figure 3.2 Overview of the proposed framework	34
Figure 3.3 Overview of the proposed BoW framework	35
Figure 3.4 New spatial pyramid decomposition strategy.....	36
Figure 3.5 Application of the 1×3 and 3×1 decomposition strategy to sample landmark images	37

Figure 3.6 Illustration of cell's saliency estimation, (a) and (d) are histograms of cells c_1 and c_2 respectively, (b) and (e) are codeword discrimination for landmark category n , (c) and (f) are the estimated saliency score for cells c_1 and c_2 respectively	42
Figure 3.7 Visual viewcone in context analysis.....	46
Figure 3.8 Sample images (left column) and geospatial distribution (right column) of the 50 landmarks.....	49
Figure 3.9 Sample images in the NTU50Landmark database under different conditions.....	50
Figure 3.10 Recognition accuracy of different content analysis methods	51
Figure 3.11 Illustrative examples of the proposed cell saliency estimation	53
Figure 3.12 Comparison between the feature pooling method [126] and the proposed method.....	54
Figure 4.1 Proposed framework for mobile landmark recognition.....	65
Figure 4.2 Selected sample landmark images demonstrating various scenes.....	66
Figure 4.3 Illustration of conventional histogram (second row) and developed histogram (third row).	70
Figure 4.4 Illustration of learned patch discriminative information in some landmark images	76
Figure 4.5 Experiment results of the proposed method	77
Figure 4.6 Experiment results of the combined methods from Chapters 3 and 4.....	82
Figure 5.1 Overview of the proposed BoP framework	87

Figure 5.2 (a) Training image, (b) Second-order feature set, (c) Second-order phrase after feature mapping, (d) Second-order phrase after duplicate removal.....	88
Figure 5.3 Visual phrases detected in the foregrounds and backgrounds areas for the same landmark category under different capturing conditions	89
Figure 5.4 (a) Sample images from 6 landmark categories, (b) Similar category distribution induced by words 1, 2 and 3, (c) different latent topic distribution induced by words 1, 2 and 3.....	91
Figure 5.5 Performance comparison of developed soft BoP encoding method versus hard BoP method.....	103
Figure 5.6 Performance comparison of proposed BoP method versus other BoW and BoP methods	104
Figure 5.7 Experiment results of the combined method from Chapters 3-5	106
Figure 6.1 The proposed mobile landmark recognition architecture	111
Figure 6.2 An overview of the proposed context-aware DCV learning process	112
Figure 6.3 (a) Training images that are divided into various direction clusters, (b) DSVT in each cluster.....	114
Figure 6.4 Ranked images using the proposed ImageRank method	120
Figure 6.5 Comparison of the proposed DCV learning approach with other typical approaches.....	127

Figure 6.6 Comparison with and without using context analysis for DCV learning	128
Figure 6.7 Comparison of the proposed DCV learning method with other methods	129
Figure 6.8 Illustration of the BoW histogram compression effect using the DCV	130
Figure 6.9 Experiment results of different methods on San Francisco landmark dataset.....	132
Figure 6.10 Illustration of two categories of landmark images for which the proposed methods do not perform well. Images (a) and (b) are from the “Yunnan garden” category, whereas images (c) and (d) are from the “Yunnan pavilion” category.....	136

List of Tables

Table 2.1 Comparison of features	23
Table 3.1 Performance comparison using different local descriptors.....	53
Table 3.2 Recognition accuracy of context analysis.....	56
Table 3.3 Experimental results of varying the radius in the viewcone analysis	57
Table 3.4 Performance comparison for content analysis, context analysis and their integration	58
Table 3.5 Experimental results on the Oxford dataset	60
Table 4.1 Recognition results of the proposed discriminative learning in comparison with other methods.....	78
Table 4.2 Comparison of different methods using different kernels	79
Table 4.3 Comparison of proposed methods versus geometric verification methods	79
Table 4.4 Experimental results on the Oxford building dataset.....	80
Table 4.5 Experimental results of content-context integration	81
Table 5.1 Performance comparison of BoW, proposed BoP, and their combination	101
Table 5.2 Performance comparison of proposed DVP selection method with other methods	101
Table 5.3 Performance comparison of various distance metric methods	102
Table 5.4 Experimental results of content-context integration	104
Table 5.5 Experimental results on the Oxford building dataset.....	105

Table 6.1 Average recognition time (s) after histogram compression using proposed method	131
Table 6.2 Average recognition time (s) using different methods	133
Table 6.3 Comparison of the proposed methods on the NTU50Landmark database	134

List of Abbreviations

BoP	Bag-of-Phrases
BoW	Bag-of-Words
CBOW	Contextual BoW
CHOG	Compressed Histogram of Gradients
DCV	Discriminative Compact Vocabulary
DGMM	Differential Gaussian Mixture Model
DoG	Difference of Gaussian
DSVT	Direction-dependent Scalable Vocabulary Tree
DVP	Discriminative Visual Phrase
DVW	Descriptive Visual Words
EM	Expectation Maximization
FIM	Frequent Itemset Mining
FSVM	Fuzzy Support Vector Machine
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
KL divergence	Kullback Leibler divergence
KNN	K Nearest Neighbor
LDV	Location Discriminative Vocabulary
LSA	Latent Semantic Analysis
NTU	Nanyang Technological University
PCA	Principal Component Analysis
PDRM	Probability Density Response Map
PLSA	Probabilistic Latent Semantic Analysis
RANSAC	Random Sample Consensus

RBF	Radial Basis Function
RFID	Radio Frequency Identification
SIFT	Scale Invariant Feature Transform
SPK	Spatial Pyramid Kernel
SURF	Speeded Up Robust Feature
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TSVQ	Tree Structured Vector Quantization
U-SURF	Upright SURF

Chapter 1 Introduction

1.1 Motivation

In recent years, the usage of mobile cellular devices has been growing rapidly. According to an estimate in [1], there are around 6 billion mobile cellular subscribers globally by 2012, penetrating more than 80 percent of the world population. Nowadays, more than half of the mobile phones in use have built-in camera features.

The advantages of built-in cameras and internet connectivity make it increasingly appealing for mobile phone users to snap pictures of objects and then obtain relevant information about the captured objects, with the aid of content and context analysis techniques. This is particularly useful and important in mobile object recognition or navigation. e.g., mobile shopping (obtaining information about a product), personalized mobile service (purchasing ticket), on-site event detection (acquiring information on the spot about an on-going unknown event), and mobile city guide (what is the captured landmark). In fact, mobile phone navigation is one of the fastest-growing segments in personal navigation market [2].

Figure 1.1 An illustration of a mobile landmark recognition system

In this thesis, we focus on the application of mobile landmark recognition. The application has been found to be useful for mobile users or travelers, for instance, as virtual city guide or navigation. An illustration of a typical mobile landmark recognition system is shown in Figure 1.1. First, a tourist sees a landmark (in this case, the Sydney Opera House) that he/she is interested in and would like to find out more information about it. He/she can use his/her mobile phone to capture an image of the landmark, and sends the captured photo to the server. After the server performs recognition of the captured image, it sends the information of the landmark to the mobile user, e.g., its name, history and activities, etc. After the tourist obtains the background information of the landmark, he/she can then make further information retrieval about the landmark according to his/her own needs, e.g., after knowing the landmark is Sydney Opera House, he/she can browse for the scheduled programs, the price and even sitting arrangement for the performance in the evening.

1.2 Overview

Existing landmark recognition works can be categorized into two classes, one is based on desktop environment, which performs landmark recognition solely on a personal computer (PC) or desktop laptop. The other is based on mobile device environment, which performs landmark recognition with the aid of a mobile device, such as a mobile camera phone. In view of this, the following sections will first introduce some differences between mobile and desktop/PC landmark recognition, and then discuss the unique challenges encountered in the mobile landmark recognition.

1.2.1 Comparison of Mobile and Desktop-based Landmark Recognition

Different landmark recognition systems have been developed in recent years [3]-[13]. They mainly focus on the development of new algorithms for content analysis and landmark

recognition. For instance, they extract various visual features from the image and classify these features into a particular landmark category. Much effort has been devoted into the improvement of feature representation and classification techniques. However, it is noted that these systems are not centered on mobile devices but rather on desktop/PC platforms, which underutilize the unique features of mobile devices.

As mobile devices become more popular, various researchers have started to redirect their efforts into mobile-based landmark recognition. As a result, some works that focus on mobile landmark recognition have appeared recently [15]-[24], [57]-[65]. These systems offer various improvements over traditional desktop-based recognition. In particular, they have integrated the context information such as the location or user tags provided by the mobile devices to assist the content analysis for landmark recognition [15], [18]-[24], [57]-[65]. The location or user tags information can serve as a pre-filter to shortlist a limited number of landmark candidates for further content analysis. The recognition time is thus greatly reduced.

As an observation from these mobile and nonmobile landmark recognition works, it is noted that a unique feature of mobile-based recognition is that the context information such as the location information (e.g. captured by Global Positioning System (GPS) or WiFi localization technology), direction information (e.g. obtained from built-in digital compass), and domain information (e.g. interaction between user and information server) are available to enhance their analysis and processing. In many practical cases, the only context information available in desktop/PC-based landmark recognition is that the captured image is a landmark. Therefore, mobile-based landmark recognition is generally considered as an integration of content and context analysis, whereas the desktop landmark recognition is usually considered as a content analysis alone-based system. It is noted that both mobile and

nonmobile landmark recognition require extraction of content-based image features and training of classifiers using these features.

The overview of a mobile landmark recognition system is given in Figure 1.2. It consists of the following components: content analysis (extracting features from the image), context information extraction (e.g. location through GPS), and recognition (identifying the captured landmark). In Chapter 2, we will present the state-of-art feature representation and classification algorithms used in both mobile and nonmobile landmark recognition systems.

Figure 1.2 The overview of a mobile landmark recognition system

1.2.2 Challenges of Mobile-based Landmark Recognition

Mobile devices differ from the desktop environments in several aspects. The inherent challenges/limitations of mobile devices should be considered when designing a mobile landmark recognition system. These include: (i) Limited processing power and battery life. Compared with PC or a laptop, mobile devices still lack sufficient processing power onboard to perform computationally intensive operations, such as content analysis and classification. (ii) Limited bandwidth of mobile network. The wireless network for mobile devices at times has limited and varying bandwidth. This inhibits the interaction between mobile users and the network. Overall, the constrained processing power and network

bandwidth of mobile devices may cause the landmark recognition to be slow, which is inconsistent with mobile users' fast response-time requirements.

In consideration of these challenges and the unique features of mobile devices, a basic framework that is designed for mobile landmark recognition is adopted in this thesis as follows: (i) Considering the hardware limitation of the mobile devices, a client-server structure is adopted to shift the computational load of visual analysis from the mobile client to a high-performance server. The mobile client captures an image and sends it (or the extracted features from the image) to the server. The server matches the image with the database and classifies it to its category. (ii) Considering the availability of the context information provided by the mobile devices, the GPS location and direction information captured by the mobile phone is integrated with visual analysis for recognition. With the aid of the context information, the system only needs to search for the shortlisted candidate landmarks for recognition. This greatly reduces the image recognition time, and satisfies the mobile users' fast response-time requirement. (iii) Considering the high computational cost in processing the large number of local features in an image, bag-of-words (BoW) method [45] is chosen as the basic content analysis framework since it can quantize the huge number of features in an image into a codeword histogram.

Compared with local patch-based image representation [18] for landmark recognition, the BoW method demonstrates better recognition accuracy and speed. However, there still exist several shortcomings for the conventional BoW method in current mobile landmark recognition:

- 1) Conventional BoW method does not take or only include limited codeword discriminative information for different landmark categories into consideration. Different codewords in fact have different discriminative capability for each landmark category. Such information should be learned and utilized for mobile landmark recognition.

2) Many conventional BoW methods consider the local patches/regions in an image as equally important when quantizing them into a particular codeword. This is inconsistent with the fact that the patches in the foreground area are usually more important than that in the background area, and thus should be emphasized.

3) Many BoW methods assume the codewords in an image are independent of each other, and ignore their co-occurrence relationship in a local image neighborhood. The frequently co-occurring words in a local neighborhood can be treated as an entity to characterize the image. The image representation induced by the co-occurring words is more descriptive than that induced by one single word. The co-occurrence of visual words should be studied.

4) Many BoW methods do not consider the context (direction and location) information for word vocabulary learning. The context information not only can be used to reduce the search space for the query image, but also can be used to supervise the codeword learning. However, the effort devoted into the study of employing the context information for codeword learning is still limited.

1.3 Contributions of This Thesis

In view of the unique features of mobile devices and the limitations of conventional BoW methods, we propose four new mobile landmark recognition methods, where some of them adopt the support vector machines (SVMs) as the classifier, and the rest uses the scalable vocabulary trees (SVTs) for classification.

In order to better illustrate the contributions, a general landmark recognition framework that encompasses various techniques proposed in different chapters is given in Figure 1.3. The developed framework consists of three major steps: (i) image discriminative information learning, (ii) image representation and (iii) image training and classification. In

the first step, the discriminative information learning at codeword level, patch level and phrase level are proposed in Chapters 3, 4 and 5, respectively. In the second step, the image representation using spatial pyramid BoW, discriminative BoW, and soft BoP are developed in Chapters 3, 4 and 5 respectively. Finally in the third step, the SVM and SVT-based image training and classification methods are developed in Chapters 3-5, and 6 respectively.

Figure 1.3 Mobile landmark recognition framework in this thesis

From the figure it can be seen that the image discriminative information learning plays an important role in the developed framework. For patch-level discriminative learning, a Differential Gaussian mixture model (DGMM) is developed to identify the patches that are descriptive to a specific landmark category. For codeword discriminative learning, the location and direction information are used to supervise the codeword generation. Latent semantic analysis (LSA) is adopted to reduce the computational cost while maintaining the discriminative information among different landmark categories. The image discriminative information is also considered for codeword selection. For phrase-level learning, a discriminative divergence metric in latent space is developed to measure the similarity between pairwise phrases. A phrase ranking algorithm that leverages the idea of PageRank technique is developed for discriminative phrase selection. The details of the contributions are given as follows.

1) An effective content and context integration method.

For content analysis, a new BoW framework is developed to perform the content analysis. It involves development of a new spatial pyramid decomposition and cell saliency estimation scheme to achieve effective representation of landmark images. The method also incorporates a new spatial pyramid kernel which employs codeword discrimination for classification. During the context analysis, different strategies involving formation of viewcones which incorporate the location information (obtained through built-in GPS of mobile device) and direction information (obtained through built-in digital compass of mobile device) are developed. A new integration of content and context analysis is proposed, which integrates the soft likelihood scores generated by the content- and context-based analysis.

2) A discriminative BoW content analysis method.

The main features of the proposed discriminative BoW approach consists of two parts, which include: (i) An iterative learning approach based on a differential Gaussian mixture model (DGMM) is developed to estimate the discriminative information of each image patch. This information is then incorporated into vector quantization to generate a continuous BoW histogram. (ii) An image signature weighting method based on the Mahalanobis distances between the image to the nearest negative category and image to the target category is developed, which is then used to train a discriminative classifier through a fuzzy SVM that consists of a Laplacian RBF kernel that considers codebook compactness for recognition.

3) A bags-of-phrase (BoP) content analysis approach.

In this method, the candidate visual phrases defined as adjacent pairwise codewords are

first generated for each category. The number of visual words in each phrase is chosen as two. A discriminative visual phrase (DVP) selection approach is developed in latent semantic space, which exploits the probabilistic latent semantic analysis (PLSA) and Bayesian estimation to derive a discriminative symmetric Kuller-Leibler (KL) divergence metric for word-level distance computing. After extending it to phrase-level distance metric, a PhraseRank technique is developed for DVP selection. A new soft BoP encoding method is developed to reduce information loss caused by the conventional BoP quantization method and generate a histogram for each image. The histograms are used to train an SVM classifier for landmark recognition.

4) A context-aware discriminative compact vocabulary (DCV) learning approach.

In this work, a direction-dependent scalable vocabulary tree (DSVT) is constructed for image quantization and a discriminative compact vocabulary (DCV) is learned to compress the image descriptor for fast mobile landmark recognition. Firstly, the DSVTs are constructed by using the direction information to supervise the image feature clustering. Then, a context-aware DCV learning algorithm is proposed to select the discriminative codewords of the DSVT to form the DCV. An ImageRank technique based on Latent Semantic Analysis (LSA) is developed to estimate the importance of each image. An iterative codeword selection algorithm that incorporates the learned image weights is then developed to generate the DCV, which is used to compress the image's original BoW histogram into a compact vector. This vector is further matched with the database images that are indexed using inverted file system for fast recognition.

1.4 Organization

The main focus of this thesis is on various content and context analysis algorithms that are

designed for mobile landmark recognition. This thesis is organized as follows,

In Chapter 1, an introduction to mobile landmark recognition is presented. The motivation, limitation and unique features of mobile landmark recognition are discussed. The major contributions of this thesis are given.

In Chapter 2, a background and literature review for mobile landmark recognition is presented. The techniques and algorithms used in the state-of-the-art mobile landmark recognition systems, which include content analysis, context analysis, and content-context integration are discussed.

In Chapter 3, an effective content and context integration technique for mobile landmark recognition is presented. The chapter proposes a new BoW framework for content analysis and the fusion of location and direction for context analysis. These information are then integrated to perform mobile landmark recognition. The SVM is used as the classifier.

In Chapter 4, a BoW content analysis approach based on discriminative learning of images, patches and codewords is presented. The discriminative BoW approach is integrated with GPS location and direction information to achieve better recognition speed and accuracy. The fuzzy SVM is used as the classifier.

In Chapter 5, a new soft BoP content analysis approach is presented for mobile landmark recognition. Unlike the BoW-based methods in previous chapters, the frequently co-occurring words are used for image representation. The content-context integration approach as in previous chapters is adopted. The SVM is used as the classifier.

In Chapter 6, unlike the previous chapters that adopt SVM as the classification technique, a fast mobile landmark recognition approach based on SVT is presented. The SVT-based method employs both the location and direction information for vocabulary learning, and achieves faster recognition speed than previous SVM-based methods, with satisfactory recognition performance.

Finally in Chapter 7, a conclusion for the thesis is given and several possible directions for future research are provided.

Chapter 2 Background and Literature Review

This chapter provides a review on the background and literatures of the state-of-the-art mobile landmark recognition, which includes an introduction to related work, and the techniques and algorithms used for content analysis, context analysis and content-context integration.

2.1 Related Work

As discussed in Chapter 1, mobile landmark recognition needs to take the unique features of mobile devices into consideration, including (i) limited computing power and battery capacity, (ii) mobile user's fast response-time requirement, and (iii) the availability of context information such as location and direction information.

In view of the issues (i) and (ii), a client-server structure is adopted in many mobile recognition works [17]-[18], [22], [57]-[60], [62]-[64], [67], [115]-[117]. Using this architecture, the computational load is shifted from the mobile client to high performance servers, and the powerful computing capability of the servers can speed up the recognition process. Among these works, some of them capture an image at the client end (mobile device), and send the query image to the server. The server extracts features from the image, and then matches these features with the offline constructed database for recognition. However, one important concern for the above process is that the size of the query image may be large. This will cause the image transmission time over the wireless network to be long and prone to interruption when the network condition is poor. Considering this, some works [58], [64], [116], [117] have endeavored to extract the image features directly on the

mobile device and then transform these features into a compact descriptor for transmission. By sending the compact descriptor instead of the original image, the transmission time can be greatly reduced. For example, in [116], local features (SURF or SIFT) are first extracted from the query image, followed by an SVT [112] to quantize these features into a BoW histogram, which is then encoded and transmitted to the remote server for recognition. In [58], a compressed histogram of gradients (CHoG [117]) is adopted to replace SIFT or SURF as the image descriptors for transmission. An effective vocabulary learning method for mobile landmark recognition has been proposed in [64] recently. The authors consider the GPS information and propose a location discriminative vocabulary encoding approach. The original BoW histogram of the query image is compressed into a compact descriptor using the learned vocabulary. The quantitative evaluation shows that transmitting the compact descriptor [64] instead of the original BoW histogram such as in [59] achieves higher recognition rate and faster speed.

In view of the issue (iii), a few works that use GPS for mobile scene/landmark recognition [17]-[18], [22] have been developed recently. In [17]-[18], content analysis integrated with location information is used to recognize the outdoor and urban scenes. Informative SIFT [17] and discriminative patch [18] content analysis are adopted for similarity matching. A mobile tourist information prototype has been developed in [22]. It uses the GPS information and 3D model to match the query image. Besides, some context-aware systems that focus on large-scale desktop PC-based landmark recognition have appeared in [69]-[75]. A world-scale landmark database has been constructed from GPS tagged photos and online tour web pages in [69]. Unsupervised graph clustering and image matching algorithm have been used to model each landmark category. Recognition is achieved by matching the query image against the model using the nearest neighbor (NN) principle. An effective vocabulary tree has been developed in [70]. It uses the most

informative features of each landmark to select the vocabulary and makes it more suitable for large-scale dataset. In [71]-[75], GPS is first used to tag the large-scale database images for pruning purpose, visual analysis or random sample consensus (RANSAC) geometric algorithms are then used to localize the query image. However, the one-to-one matching RANSAC algorithm is computationally intensive.

In addition to these works that focus on landmark recognition, there are also some related works on image retrieval or annotation [76]-[78]. In [76], an image retrieval method is developed by using BoW algorithm as a filtering step, followed by a RANSAC algorithm [79] for spatial re-ranking. An unsupervised mining approach for community photo collections is presented in [77]. Based on geo-spatial information and visual content analysis, a database of mined objects and events that are labeled with a Wikipedia link is constructed. Similarly, a database is collected in [78] from photo sharing website worldwide. Context information such as textual tags and geographical location, and content analysis including BoW and term frequency-inverse document frequency (tf-idf) are utilized to develop an object-level auto-annotation system for the landmark images.

In summary, the above-mentioned works mainly use content analysis, context analysis, or content-context integration methods for mobile landmark recognition. A detailed review to these techniques will be given in the following sections.

2.2 Content Analysis

Finding a good feature representation is crucial for landmark recognition. The visual features (content) used for landmark recognition can be broadly categorized into global features and local features. Global features characterize the overall properties of an image. Local features aim to represent the content of an image using local information extracted from the salient regions or patches within the images.

It is important to note that careful selection of appropriate features is crucial for mobile landmark recognition. This is because landmark images possess unique characteristics that need to be taken into consideration. For example, when taking photos of a landmark, images captured by different users using different mobile devices may experience changes in illumination, viewpoint, scale, rotation and other effects such as blurring, background clutter, and shadowing. Therefore, a thorough review of the features used in content analysis of landmark images, their advantages and disadvantages will be discussed in the following section. A summary of the state-of-the-art features used in mobile landmark recognition is given in Figure 2.1.

Figure 2.1 Feature representation in landmark recognition

2.2.1 Global Features

The commonly used global features for landmark recognition include color [3], shape [14], and texture [7]. Color is one of the simplest features for landmarks. However, color is generally sensitive to changes in illumination and contrast. As mobile cameras tend to take

relatively noisy images, using color features in mobile landmark recognition needs careful consideration. To alleviate the impact of color and illumination distortion on landmark recognition, different photometric color-invariant features have been studied and compared in [5]. The GIST descriptor is developed in [14] for global scene understanding. It uses a set of perceptual dimensions (naturalness, openness, roughness, expansion and ruggedness) to represent the dominant spatial shape of a scene. Similarly, a wavelet transform is performed on the images to extract its texture features in [7] for place recognition. Overall, the shape and texture features are based on how humans perceive images and focus on the similarity of contrast, coarseness, regularity, directionality and line likeness.

Global features are the earliest features for landmark recognition due to their low computational cost and simplicity. However, they can only describe the global statistical properties of the image and ignore the regions of interest. Therefore, they are usually used in conjunction with some local features for landmark recognition.

2.2.2 Local Features

Local features are widely used in landmark recognition due to their capability to describe the properties of regions-of-interest. It usually consists of two stages: (i) detection of salient patches (regions), and (ii) extraction of descriptors to characterize the appearance of the local regions.

Based on patch/region selection/determination strategy, local feature representation and extraction can be categorized into four broad categories: (i) keypoint-based, (ii) random sampling-based, (iii) uniform sampling-based, and (iv) segmentation-based. Figure 2.2 is an illustration of the four patch detection methods. The leftmost image is keypoint-based patch selection; middle left and right images are based on random and uniform patch selection respectively; rightmost is the segmentation-based selection.

Figure 2.2 An illustration of four patch selection methods, (a) keypoint-based [17], (b) random sampling-based, (c) uniform sampling-based, (d) segmentation-based [126].

From the figure, we can see that keypoint-based approaches extract the descriptors of the patches centered on the keypoints. The random and uniform patch selection strategies sample the patches either randomly or uniformly from the image, and then the local descriptors of these patches are extracted. The segmentation-based techniques perform image segmentation before the descriptors of the homogeneous regions are extracted. Details on these feature extraction and representation strategies will be discussed in the following sections.

A. Keypoint-based Approaches

Keypoint-based features are widely used in landmark recognition due to their robustness to changes in illumination, scale, occlusion and background clutters. This class of methods comprise two steps: (i) keypoints of interest are detected at different scale spaces, and (ii) local descriptors are extracted from the patches (regions) centered on these keypoints. Commonly used keypoint-based descriptors include scale-invariant feature transform (SIFT) [27], speeded-up robust feature (SURF) [56], compressed histogram of gradients (CHoG) [110] and DAISY [111]. The following sections will give a concise description of them.

The SIFT descriptor developed by David Lowe [27] is one of the most widely used local features in the state-of-art mobile landmark recognition systems [16]-[17], [58]-[59], [64]. It

first selects keypoints that are invariant to image scale and rotation using difference of Gaussian (DoG) scale space. The locations of the selected keypoints are defined as the optima of the result of DoG functions, which are applied in scale-space to a series of smoothed and down-sampled images. The candidate points with low contrast and edge response points are discarded.

Next, descriptors are extracted by computing orientation histograms in the local regions around the keypoints. Dominant orientations are assigned to the keypoints to ensure that they are more stable for image matching and recognition. The descriptors are robust to changes in illumination, viewpoint, occlusion and background clutters. An example of using SIFT descriptors for matching two images is illustrated in Figure 2.3. The two images are taken on the same landmark but from different viewpoints. The lines connecting the two images indicate the SIFT points that are matched between the two images.

Figure 2.3 Illustration of matching SIFT keypoints extracted from two images

The conventional SIFT involves detection of a large number of salient keypoints, and extraction of 128-dimensional feature vector centered on each of these keypoints. In view of this, several SIFT-based variants have been developed to reduce the computational cost by decreasing the number of keypoints and dimension of the feature vector. These include similar SIFT keypoints clustering in [16], informative SIFT keypoint selection (I-SIFT) [17], and principal component analysis (PCA)-SIFT for dimension reduction [36].

Recently, another feature descriptor called speeded-up robust feature (SURF) has been proposed in [56]. SURF has faster speed in the feature extraction step. The key steps of SURF comprise: keypoint detection, scale space and feature descriptor representation. Due to the similarity of SURF and SIFT in the keypoint extraction step, details of SURF method will not be reiterated here. However, it is worthwhile to notice that the difference of SURF from SIFT is the sampling method in scale space representation. In SIFT, the scale spaces are represented as an image pyramid, i.e. the images are repeatedly convolved with Gaussian functions with different scales, and then sub-sampled to reach the higher level of the pyramid. DoG is performed by subtracting the neighboring pyramid levels. On the contrary, SURF adopts a reverse operation as shown in Figure 2.4, where the scale space is analyzed by up-scaling the filter size rather than down-scaling the image size. Through this, a faster computational speed is achieved.

Figure 2.4 Comparison of SIFT (left) and SURF (right) in scale space sampling

The compressed histogram of gradients (CHoG) descriptor [110] is similar to SIFT which uses the HoG as the local descriptor. The difference is that CHoG represents gradient histograms as tree structures which can be efficiently compressed. Therefore, it has low bit-rate and complexity, and has significant speed-up in the image matching stage. Similar to

CHoG and SIFT, the DAISY descriptor developed in [111] also adopts histograms of gradients. The difference is that it uses a Gaussian weighting and circularly symmetrical kernel. This gives the fast speed for dense computation.

Besides the HoG-based descriptors above, there also exist other local descriptors, e.g., Fourier Transform (FT)-based features [13] and Harris corner-based features [38]. The FT-based method in [13] uses keypoint detectors with PCA to extract patch features. It is reported to be robust toward background variation, occlusion and planar rotation. The multi-scale oriented patches (MOPS) method in [38] detects points-of-interest using Harris corner detector. These points are then extended to patches by Taylor expansion at each image pyramid level. However, it is noted that MOPS performs poorly when there are shadows or reflections in the images.

Using local descriptors of image patches centered on keypoints in an image, a visual entity (scene or object) is characterized as several parts. These approaches utilize the most informative regions in an image for landmark recognition. It is robust toward occlusion and background clutters as it relies on local patches centered on informative keypoints rather than the global features. Therefore it is attractive for mobile landmark recognition.

B. Dense Sampling Approaches

The dense sampling patch-based features usually can be divided into two main approaches: random and uniform sampling. As they bear close resemblance in processing and also achieve comparable performance, most literatures do not treat them separately. In this thesis, they are presented and discussed together under the same section.

Based on the scales of the sampled patches, dense sampling approaches can further be categorized into two groups: (i) single-scale patches, which mean all the patch samples taken from the images are of the same size [6], [10], [30] and (ii) multi-scale patches, where

multiresolution patches are sampled from the images [18], [28], [32]. This will alleviate the difficulty of different landmark image resolutions as local feature descriptors extracted from these multiresolution patches will provide a broader spectrum of descriptors corresponding to different scales in the images. Compared with single-scale patches, multi-scale patches are more robust toward changes in scale. This is a clear advantage as the landmark can be potentially captured from different distances, resulting in different scales and sizes of the landmark in the captured images.

After local patches are sampled from the images uniformly or randomly, visual features are extracted from these patches as their descriptors. There are a few techniques on how these patch features can be used to represent the image. Amongst them are two representative techniques called (i) the bags-of-words (BoW) method [45] and (ii) the informative/discriminative patch method [17]-[18]. The BoW method first performs vector quantization of the patch descriptors from all the classes to form a codebook. The dense patches from the image are then represented as a histogram of the codewords [10]. The second framework strives to find the most informative/discriminative patches amongst all the dense patches to represent the image [17]-[18]. In general, the BoW framework requires lower computational time as the image descriptor is in the form of a codeword histogram rather than a lot of dense patches. In view of the fast response-time requirement, BoW framework is a suitable technique for mobile landmark recognition.

C. Segmentation-based Approaches

This class of methods uses image segmentation techniques to locate the homogeneous regions in the image and then extracts descriptors from these regions. Based on how the image segmentation is performed, it generally consists of two steps: (i) partitioning different regions based on visual features, and (ii) modeling the features of different regions using

region descriptors. A statistical model-based segmentation method is proposed in [12]. It first partitions the image into a collection of local regions called “themes”, and then uses unsupervised learning to represent these “themes” as codewords. Finally the distribution of these themes is modeled by Bayesian hierarchical model. A contour-based analysis combined with region-based analysis is integrated to partition color images into disjoint meaningful regions in [39]. It extracts homogenous regions by grouping adjacent pixels based on color histogram and wavelet coefficients. Description length of these regions are then evaluated and clustered to form a larger region to minimize it. Finally semantically meaningful objects are extracted. The main disadvantage of segmentation-based techniques is the unreliable segmentation results due to the complex scene found in many landmark images.

D. Summary

In summary, it is seen that the dense sampling patch strategy followed by keypoint-based descriptor to describe each patch is attractive for mobile landmark recognition. To validate this, a comparative experiment is conducted to compare the performance of SIFT, SURF and dense sampling SIFT for mobile landmark recognition in [57], which shows that dense sampling SIFT achieves the best performance, followed by sparse SURF and SIFT. According to the analysis above, we can draw the following conclusions on content analysis for landmark recognition: (i) Local features perform better than global features. This observation is supported by statistical hypothesis testing as in [49]; (ii) SIFT has superior performance when compared with many other local descriptors; (iii) The dense sampling patch approach offers better performance than the sparse keypoint-based and segmentation-based techniques. A summary is given in Table 2.1.

Table 2.1 Comparison of features

Feature		speed	robustness	Recognition performance
Global		Fast	Poor	Poor
Local	Dense SIFT-based	Average	Good	Good
	Keypoint-based	Slow	Good	Good
	Segmentation-based	Slow	Average	Average

2.3 Classification Algorithms

The classification methods for mobile landmark recognition can be broadly categorized into two groups: discriminative-based classification and generative-based classification as shown in Figure 2.5. Discriminative classification usually constructs decision boundaries to discriminate between different classes of images. In contrast, generative classification tries to model the image data using density distribution.

Figure 2.5 Classification algorithms for landmark recognition

For discriminative classification, two of the most popular techniques are support vector machine (SVM) [18], [29], [31] and Adaboost [13], [34]-[35]. SVM classifier is based on supervised learning. It constructs a separating hyperplane as the decision boundary using the support vectors from the training sets. The principals can be briefed as: Given a set of

training samples $S = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^p, y_i \in \{-1, 1\}\}_{i=1}^n$, SVM tries to construct the optimal separating hyperplane by solving the following problem:

$$\begin{aligned} & \text{Minimize } \frac{1}{2} \|\mathbf{w}\|^2 + R \sum_{i=1}^n \varepsilon_i \\ & \text{Subject to } y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \varepsilon_i, \\ & \varepsilon_i \geq 0, i = 1, 2, \dots, n \end{aligned} \quad (2.1)$$

where $\mathbf{w}$ is the normal vector of the hyperplane, and b is the bias that is a scalar. ε_i is the slack variable related to classification errors. R is a regularization parameter that controls the balance between margin maximization and classification errors. The optimization problem of SVM can be transformed into dual problem by constructing the Lagrangian:

$$\begin{aligned} & \text{Maximize } \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \\ & \text{Subject to } \sum_{i=1}^n y_i \alpha_i = 0, 0 \leq \alpha_i \leq R, i = 1, 2, \dots, n \end{aligned} \quad (2.2)$$

where α_i is the Lagrange multiplier, $K(\mathbf{x}_i, \mathbf{x}_j)$ is the kernel function, e.g., radial basic function,

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp \left\{ -\frac{\|\mathbf{x}_i - \mathbf{x}_j\|}{2\sigma^2} \right\} \quad (2.3)$$

where $\|\mathbf{x}_i - \mathbf{x}_j\|$ is the norm-2 distance between $\mathbf{x}_i, \mathbf{x}_j$. After that, the corresponding decision boundary is formulated as,

$$f(\mathbf{x}) = \sum_{i=1}^n c_i \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b \quad (2.4)$$

where $\mathbf{x}_i$ is the support vector.

Adaboost algorithm is a generic leaning method for classification. It constructs a strong classifier through the combination of weak classifiers. Adaboost is sensitive to noisy data and outliers. The principal can be briefed as follows.

Given a set of training samples $S = \{(\mathbf{x}_i, y_i) | \mathbf{x}_i \in \mathbb{R}^p, y_i \in \{-1, 1\}\}_{i=1}^n$, and number of iterations T , initialize the distribution for each sample $D_1(i) = \frac{1}{n}$, $i = 1, 2, \dots, n$.

For iteration $t = 1, \dots, T$,

- 1) Find the classifier h_t from the family of weak classifiers that minimizes the error with respect to D_t :

$$h_t = \arg \min_{h_j} \varepsilon_j \quad (2.5)$$

where $\varepsilon_j = \sum_{i=1}^n D_t(i) [y_i \neq h_j(\mathbf{x}_i)]$, and corresponds to the weighted classification error.

- 2) Pre-condition: $\varepsilon_t < 0.5$, otherwise stop.
- 3) Choose the weighting factor $\alpha_t = \frac{1}{2} \ln \frac{1 - \varepsilon_t}{\varepsilon_t}$.
- 4) Update $D_{t+1} = \frac{D_t \exp(-\alpha_t y_i h_t(\mathbf{x}_i))}{Z_t}$.

where Z_t is a normalization factor to make sure that D_{t+1} is a probability distribution.

Finally, the classifier is

$$H(\mathbf{x}) = \text{sign} \left(\sum_{t=1}^T \alpha_t h_t(\mathbf{x}) \right) \quad (2.6)$$

Discriminative-based classification constructs flexible decision boundaries and allows learning the discriminative features of images and background. Combined with properly extracted image features, discriminative-based classification can perform well in the presence of clutter, viewpoint changes, and partial occlusions.

For generative classification, some of the most commonly used models are Gaussian mixture model (GMM) [5][11], Bayesian model [12], and hidden Markov model (HMM) [25]-[26].

A GMM-based method is proposed to recognize scene types in [11]. Local regions of an image are first modeled as a GMM distribution, and analyzed using principle component analysis (PCA). Expectation maximization (EM) is then used to estimate the model parameters. The GMM formulation is expressed as follows:

$$P(\mathbf{x}) = \sum_{j=1}^J \pi_j \cdot P(\mathbf{x} | c_j) \quad (2.7)$$

$$P(\mathbf{x} | c_j) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_j|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_j)^T \Sigma_j^{-1} (\mathbf{x} - \mu_j)\right) \quad (2.8)$$

where π_j is the mixing proportion of the GMM component c_j , μ_j is the mean vector of c_j , Σ_j is the covariance matrix of c_j , and d is the dimension of the feature vector $\mathbf{x}$.

A Bayesian hierarchical model to learn and recognize natural scene categories is proposed in [12], [46]. Local image features are first represented as codewords, then Bayes model uses a probabilistic framework to learn the distribution of these codewords. A Bayesian posterior probability formula is given by:

$$P(\theta | \mathbf{X}) = \frac{P(\mathbf{X} | \theta) P(\theta)}{P(\mathbf{X})} \quad (2.9)$$

where $\mathbf{X}$ is the observed data, θ is the parameter, $P(\mathbf{X} | \theta)$ is the likelihood or evidence for data $\mathbf{X}$ given the parameter θ , and $P(\theta | \mathbf{X})$ is the calculated posterior probability.

Hidden Markov model (HMM) has been used to recognize places and object in [25]-[26] and shown to be effective, especially suitable for noisy images with temporal variations.

Generative model can be learned incrementally and used to deal with missing data. As it is robust toward misclassified points and noise, the model can handle significant intra-class variation of object appearance. However, this model tends to produce false positives, especially for objects that share high visual similarity.

Discriminative classification and generative classification are both effective learning-based methods. Each of these classification methods has its own limitations. Discriminative

classification lacks the elegance of generative model and the relationships between variables are implicit. Generative classification, on the other hand, is more sensitive to outliers.

2.4 Context Analysis

In mobile landmark recognition, various forms of context information can be exploited to facilitate recognition. Location is one of the most important context information for mobile landmark recognition. If the location information is available, the search scope of the captured landmark can be reduced significantly, which in turn will greatly improve the recognition accuracy and speed. Location information can be obtained through GPS, a widely used location-sensing technology, which is supported in most mobile phones nowadays. However, at times the GPS signals may be weak or unavailable, such as when a person is indoors. This will reduce the accuracy of the GPS location information or even render it inapplicable. Hence, other alternatives such as WiFi and/or ultra-wideband (UWB) localization may be used to pinpoint the location information.

Direction is also an important context for mobile landmark recognition. In recent years, mobile devices equipped with digital compass are becoming popular, e.g. HTC Dream and Apple iPhone. Couple the direction information (e.g. obtained from the digital compass/gyroscope) with the location information (e.g. obtained from GPS), the system can reduce the search space of the captured image to those landmarks located within the viewcone of the mobile device. However, it is worth noting that landmark recognition based on the context information alone can be unreliable as the GPS drift can be up to 100m in some dense built-up areas.

2.5 Content and Context Integration

As mentioned before, landmark recognition via context or content analysis alone is still less than ideal. Therefore, it would be desirable to integrate content and context information to achieve better performance. The main idea is to integrate the strengths of content and context information while alleviating their respective difficulties. Using context information, we can narrow down the search space of the categories significantly. Together with content information, better recognition performance can be achieved by exploiting the full potential of all information.

Recently, there exist some works on mobile landmark recognition using content and context integration. A mobile location recognition system is developed in [15]. It first identifies the relevant keywords related to the query image from the database using content analysis. Then it searches a larger number of images from the Web to identify the location. A Photo-to-Search system [19] has been developed which uses photos captured by mobile phones for image query. It uses multimodal queries, e.g., photo, key phrases and user input, to perform image matching and retrieve results similar and related to the query image. A mobile landmark recognition system called Snap2Tell has been developed as tourist information guide in [33]. The GPS information is used to shortlist several candidate landmarks, which is followed by content analysis for landmark recognition. The Google Goggles which is an image recognition system is developed by Google Inc [66]. The Google Goggles' search is carried out using the images taken from the camera phones.

2.6 Scalable Landmark Recognition

Various content and context analysis methods have been presented in the previous sections. However, they do not consider the scalability issue. Considering this, the scalable landmark

recognition problem is discussed in this section, and two types of mobile landmark recognition methods based on SVM and scalable vocabulary tree (SVT) are discussed.

2.6.1 Scalable Vocabulary Tree

Among the discriminative and generative models discussed in Section 2.3, SVM is widely used in landmark or object recognition due to its superior performance over other methods [3], [4], [6], [16], [18], [26]. However, one issue of SVM is that it has quadratic computational complexity, and this renders it not easy to scale up to a large number of landmark categories. Scalable vocabulary tree (SVT), which is developed in [112] and has been shown to have good scalability for large volume of image search, is used in this thesis for mobile landmark recognition. There are two phases for SVT to be used for landmark recognition: the offline learning and online testing. In the offline phase, the vocabulary tree is generated by performing hierarchical k -means clustering algorithm iteratively on a set of local descriptors, until a predefined leaf nodes number is reached. First, an initial k -means algorithm is run on the training features to yield k cluster centers. The same process is then recursively applied to each new group to generate new centers. Each cluster center and the k -means iteration number can be treated as the branch and depth of the tree, respectively. Figure 2.6 is an illustration of a simple SVT with the depth equals to 2 and the branch equals to 3. In the figure, each node of the tree denotes a cluster center. In the online phase, each descriptor of the image is propagated down the tree, by comparing with the cluster centers at each depth level and choosing the closest one. The computation is efficient since the number of comparison at each level depends on the number of branch which is usually not too large.

Figure 2.6 Illustration of SVT

When the SVT is used for large-scale landmark recognition, an inverted indexed files structure is first constructed as in [58], to improve the computational efficiency. The inverted file maintains two lists: a sorted array of images indicating the database images that have visited that node (codeword), and simultaneously a corresponding array of counts indicating the visit number. When performing a query for an image, those sorted database images can be quickly scored by traversing only the nodes that are visited by the query descriptors.

2.6.2 Comparison of SVM and SVT

Through the analysis in previous sections, it is found that both SVM and SVT have some advantages for mobile landmark recognition. For SVM, it constructs a strong classifier through a supervised and deterministic learning process. Its advantage lies in the good classification capability, which is realized through learning an optimal separating hyperplane between different classes. For SVT, it constructs a hierarchical vocabulary tree that can quickly quantize an image and match it with the database. Its advantage is the good scalability for large-scale image recognition, which is realized through an inverted file system. In view of this, both methods are adopted to perform landmark recognition in this thesis, where the SVM-based method is adopted in Chapters 3, 4 and 5, and the SVT-based method is adopted in Chapter 6.

In order to utilize SVM for mobile landmark recognition, the context information including the GPS location and direction is used to reduce the search space so that the SVM computational time is manageable. This can alleviate the scalability problem of SVM. Therefore, SVM is more suitable to be used in conjunction with context information such as GPS to constrain the number of landmarks to a small number.

However, when the database is increased to a large volume, SVM may not be suitable since the context information (GPS location and direction) of some images in the database may be not available. As a result, the image recognition time will be long. In this case, SVT can be utilized to solve the scalability issue. The SVT constructs an inverted file system for the large number of database images to achieve fast recognition. Therefore, SVT is more suitable to perform large-scale landmark recognition, where the GPS or direction information may not be available.

2.7 Summary

This chapter provides a comprehensive review of various feature extraction and representation techniques (e.g. different types of keypoint-based features, patches-based descriptors) and various category classifiers for mobile landmark recognition. For the concept classification, SVM and SVT both are explored and found to be suitable for performing mobile landmark recognition.

Chapter 3 Integrated Content and Context Analysis for Mobile Landmark Recognition

3.1 Introduction

This chapter presents a new approach for mobile landmark recognition based on integrated content and context analysis. Conventional scene/landmark recognition methods such as [47][67] focus mainly on desktop/PC platform, where content analysis alone is used to perform landmark recognition. These systems, however, do not take unique features of mobile devices into consideration, e.g., limited computational power and fast response-time requirement of mobile users. On the other hand, most existing context-aware mobile landmark recognition methods such as [71]-[75] mainly rely on GPS location information for context analysis. In view of this, this chapter proposes a new content and context integration technique for mobile landmark recognition. A BoW framework is developed to perform content analysis. It involves development of a new spatial pyramid decomposition and cell saliency estimation approach to achieve effective representation of landmark images. The technique also incorporates a new spatial pyramid kernel which employs codeword discrimination for classification. During the context analysis, different strategies involving formation of viewcones which incorporate the location information (obtained through built-in GPS of mobile device) and direction information (obtained through built-in digital compass of mobile device) are developed. A new integration of content and context analysis is proposed, which greatly enhances the recognition performance of the system, as demonstrated through various experimental results in later sections.

The remainder of the chapter is organized as follows. An overview of proposed mobile landmark recognition system is presented in Section 3.2. The content analysis and its integration with context analysis are discussed in Sections 3.3 and 3.4, respectively. Experimental results and discussions are given in Section 3.5. Section 3.6 concludes the chapter with a summary of our findings.

3.2 Framework Overview

Similar to the work [17]-[18], a client-server structure is adopted in this chapter to enable the computational load to be shifted from the mobile client to the server. The screenshots of the implemented system in Nanyang Technological University (NTU) using the proposed method is given in Figure 3.1. In the figure, the mobile client captures an image, and sends it to the server. The server then uses the developed algorithm to identify the captured landmark, and sends the recognized landmark and its related information back to the client.

Figure 3.1 Snapshots of the implemented mobile landmark recognition system in NTU

The overview of the proposed mobile landmark recognition system is given in Figure 3.2. It comprises the offline learning and online querying stages. In the offline learning

stage, the local features of the training images are extracted and represented using an efficient BoW framework. A new spatial pyramid decomposition technique is developed to partition the image into multiple cells. A saliency estimation scheme based on visual activities and codeword discrimination is then used to evaluate the discriminative power of these cells. Finally, a multi-class SVM employs the new spatial pyramid kernel (SPK) to train the landmark classifiers. During online querying stage, the content analysis is first performed. It is then integrated with context analysis (fusion of location and direction information), and forward to the trained SVM classifier for landmark recognition.

Figure 3.2 Overview of the proposed framework

3.3 Content Analysis

BoW method has been used in many recent state-of-the-art scene/landmark recognition systems [10], [45]-[48] due to its good recognition performance and fast processing speed. However, early BoW frameworks do not take the spatial layout information of the landmark images into consideration. This underutilizes the information available as landmarks are usually located near the center of the images, and backgrounds such as sky and grass are located near the top and bottom of the image, respectively. In view of this, spatial pyramid matching scheme has been developed in [47], [128]. In [47], the method partitions an image

into multiresolution pyramid cells and then represents each cell in the pyramid using BoW. In [128], it extends the method in [47] by generalizing vector quantization to sparse coding followed by spatial max pooling, and proposes a linear spatial pyramid kernel based on SIFT sparse codes. This remarkably reduces the complexity of SVMs both in training and testing stages. Further, a receptive field learning method for feature pooling is proposed in [129]. It first produces a large number of receptive field candidates in the images, and then trains a classifier with structured sparsity to select a sparse subset of all the features. An efficient algorithm based on incremental feature selection is proposed for fast learning. However, these spatial methods consider all codewords to be equally important when representing landmark images. In other words, they do not consider discriminative power of different codewords in representing different landmark categories. Further, they also do not consider the importance of different cells in the images when fusing the BoW feature. This is inconsistent with the observation that image regions/patches with more visual saliency should be given more emphasis when considering its visual representation.

Figure 3.3 Overview of the proposed BoW framework

In view of these, this chapter proposes a new BoW framework which is summarized in Figure 3.3. The framework includes the following contributions: (i) it develops a new spatial pyramid decomposition scheme that can represent landmark images more effectively, (ii) it develops a pyramid cell saliency estimation scheme that determines the importance (saliency) of different cells at different pyramid levels, and integrate this information into the BoW model. The saliency information is estimated based on two factors: visual activities and codeword discrimination, (iii) it develops a new spatial pyramid kernel in training of SVM classifiers, which employs codeword discrimination among different landmark categories.

3.3.1 Spatial Pyramid Decomposition

Current spatial pyramid schemes [47] adopt a three-level decomposition with 1×1 , 2×2 , and 4×4 cell partitioning strategy. Even though the 1×1 and 4×4 cell structures are able to capture the overall picture and detailed spatial orientation of the image reasonably well, the 2×2 partitioning at level 1 is unable to represent the landmark image effectively. The reason is that landmarks are usually located near the center of the image, and backgrounds such as sky and lawn are usually located near the top and bottom of the image, respectively. The 2×2 cell structure tends to partition the central landmark into 4 quadrants, thereby diminishing the effectiveness of individual cell histogram. Therefore, this section will introduce an alternative pyramid decomposition strategy, as illustrated in Figure 3.4.

Figure 3.4 New spatial pyramid decomposition strategy

From the figure, it can be seen that the new decomposition scheme employs a 1×1 , 1×3 , 3×1 , and 4×4 cell decomposition strategy. As opposed to the conventional level-1 2×2 cell structure, the proposed decomposition involves a combination of 1×3 and 3×1 cell structures to better partition the image into the central region-of-interest and the background areas. Figure 3.5 provides various illustrations of this decomposition scheme when applied to the landmark images. From the figures, it can be seen that landmarks are often located in the central third of the 1×3 and 3×1 cells. Experimental results in later section will demonstrate the effectiveness of this strategy in performing landmark recognition.

Figure 3.5 Application of the 1×3 and 3×1 decomposition strategy to sample landmark images

3.3.2 Pyramid Cell Saliency Estimation

The decomposition scheme in previous section partitions the landmark image into multi-scale cells that contain different amount of saliency information. Cells that contain more saliency are generally more useful for classification, since stronger saliency indicates more discriminative or relevant information is contained in the cell. Therefore, a pyramid cell saliency estimation scheme is developed in this section to estimate the saliency of different pyramid cells, so that this information can be incorporated into the training of classifiers. The scheme involves development of two techniques based on cell visual saliency and

codeword discrimination among different landmark categories. To simplify discussion and notation in the following sections, we will denote a typical j -th cell at the l -th pyramid level in an image I as c_j^l .

A. Cell Visual Saliency

The visual saliency of cell c_j^l depends on two factors: (i) the visual activities that are an indicator of the information contained in the cell and (ii) the position of the cell as landmarks are more likely to be located near the center of the image. As visual activities can be reflected by the number of SIFT keypoints in a cell, more SIFT keypoints imply that the particular cell should be given more emphasis when considering its representation. Taking this into consideration, the saliency of the cell due to the number of SIFT keypoints can be estimated using:

$$z(c_j^l) = \frac{1}{1 + \exp(-15 \frac{t}{T})}, \quad l = 0, 1, 2 \quad (3.1)$$

where t is the number of SIFT keypoints inside cell c_j^l , and T is the total number of SIFT keypoints in the image. Sigmoid function is used as large t value will cause $z(c_j^l)$ to be large, indicating that the cell with more SIFT keypoints is visually more salient in representing the image.

It is observed that landmarks are typically located near the central region of images. Therefore, cells that are closer to the image center should be assigned larger weight. The saliency of the cell due to the location can be estimated using:

$$q(c_j^l) = \frac{1}{1 + \exp(0.5 \|\mathbf{g} - \mathbf{p}\|)} \quad (3.2)$$

where $\mathbf{g}$ and $\mathbf{p}$ represent the positions of the cell center and image center, respectively. Finally, the visual saliency of cell c_j^l can be estimated by combining the saliency due to visual activity and cell location to give:

$$\omega(c_j^l) = z(c_j^l)q(c_j^l) \quad (3.3)$$

It is worth noting that the assumption that more SIFT features in a cell imply higher importance is heuristic at times, since some high-frequency clutters may also have large number of SIFT keypoints. Therefore, a codeword discrimination learning method is proposed in following sections to complement the cell discrimination information.

B. Codeword Discrimination

The cell visual saliency estimation in the previous section evaluates the cell's importance based on only a single image. It does not utilize the discriminative information available from other images. Therefore, this section will propose a cell saliency estimation approach based on codeword discrimination. The method will first determine the discriminative power of different codewords for different landmarks. Using this discriminative information, the method estimates and adjusts the importance of various pyramid cells. Algorithm 3.1 summarizes the proposed method.

Algorithm 3.1 Codeword discriminative learning

1. Utilize the method in Section 3.3.1 to partition the images from N landmarks into multiple cells $c_j^l, j=1, 2, \dots, J$, where $J = 1, 6$ and 16 for $l = 0, 1$ and 2 respectively.
2. For each c_j^l , perform multiscale patch sampling, and use SIFT to describe each patch.
3. Generating the codebook consisting of M codewords through performing the k -means clustering [50] algorithm on the set of dense SIFT features.

4. Construct an occurrence matrix $\mathbf{S}$ of size $M \times N$ as below:

$$\begin{array}{ccccccc}
 & \text{category 1} & \dots & \text{category } n & \dots & \text{category } N \\
 \text{codeword 1} & S(1,1) & \dots & S(1,n) & \dots & S(1,N) \\
 \dots & \dots & & & & \dots \\
 \text{codeword } m & S(m,1) & \dots & S(m,n) & \dots & S(m,N) \\
 \dots & \dots & & \dots & & \dots \\
 \text{codeword } M & S(M,1) & \dots & S(M,n) & \dots & S(M,N)
 \end{array}$$

where each item $S(m, n)$ denotes the number of the n -th landmark's local patches that are quantized as the m -th word.

5. Normalize each column of the matrix $\mathbf{S}$ to unit length, and generate $\mathbf{P}$:

$$P(m,n) = \frac{S(m,n)}{\sum_{m=1}^M S(m,n)}, \quad m=1, \dots, M, \quad n=1, \dots, N \quad (3.4)$$

6. Normalize each row of the matrix $\mathbf{P}$ to unit length, and yield $\mathbf{D}$ which reflects the codeword discriminative information:

$$D(m,n) = \frac{P(m,n)}{\sum_{\substack{\bar{n}=1, \\ \bar{n} \neq n}}^N P(m,\bar{n})}, \quad m=1, \dots, M, \quad n=1, \dots, N \quad (3.5)$$

7. For each landmark category $n=1, \dots, N$, compute the saliency score of cell c_j^l in the image I using codeword discrimination matrix $\mathbf{D}$:

$$\rho(c_j^l) = \sum_{m=1}^M \mathbf{h}_{I^n}^m(c_j^l) D(m,n) \quad (3.6)$$

where $\rho(c_j^l)$ is the calculated saliency score of cell c_j^l . $\mathbf{h}_{I^n}^m(c_j^l)$ is the BoW histogram of cell c_j^l , namely, the number of occurrences that codeword m appears in cell c_j^l . In the

following discussion, we will abbreviate $\mathbf{h}_{I^n}^m(c_j^l)$ as $\mathbf{h}_{I^n}^m$ to simplify the notation.

From the Algorithm 3.1 it can be seen that, firstly, the k -means algorithm is utilized to cluster the dense features extracted from the training images into M clusters, with each

cluster represented by its centroid. Each centroid is used as a codeword and their union forms the codebook. In order to learn the discriminative information of each codeword for each landmark category, a matrix $\mathbf{S}$ is formed to capture the relationships between various codewords and landmark categories, where the matrix entry $S(m, n)$ denotes the number of the n -th landmark's local patches that are quantized as the m -th codeword. The intuitive interpretation for the occurrence matrix $S(m, n)$ is given in below. For codeword $m = 1, 2, \dots, M$, where M is the codebook size, if

$$S(m, n) > S(m, \bar{n}), \forall \bar{n} \neq n \quad (3.7)$$

namely, more SIFT descriptors (dense patches) from landmark category n belong to the m -th codeword than SIFT descriptors (patches) from the rest landmark categories, codeword m is then claimed to have more representative capability for the n -th landmark than for the other landmarks. In practice, however, as different landmark categories may have different number of training images for codebook construction, we have to normalize the matrix $S(m, n)$ column-wise to obtain $P(m, n)$. Next, the algorithm considers $P(m, n)$ and $P(m, \bar{n})$ for the current target (positive) category n and all other (negative) categories $\bar{n}$, respectively. It then estimates the discrimination power $D(m, n)$ of each codeword m for each category n using (3.5). Intuitively, larger $D(m, n)$ value indicates that codeword m has larger discriminative power for the n -th landmark than for the other landmarks $\bar{n}$.

Based on the matrix $D(m, n)$ for codeword $m = 1, \dots, M$, and landmark category $n = 1, \dots, N$, we calculate the cross-correlation between the cell's BoW histogram $\mathbf{h}_{I^n}^m$ and the n -th column of the $D(m, n)$ in (3.6) to estimate the cell's importance in representing landmark category n . In short, if $\mathbf{h}_{I^n}^m$ is highly correlated with the n -th column of the $D(m, n)$, this implies that the particular cell's histogram contains more discriminative power to represent landmark category n , and hence should be given more weight in representing

landmark n . Figure 3.6 demonstrates the application of (3.6) in estimating the saliency of BoW histograms for two different cells c_1 and c_2 . From the figure, it can be seen that the histogram distribution of cell c_1 exhibits stronger similarity (more correlated) with the codeword discrimination column of category n than the histogram of cell c_2 . Therefore, cell c_1 will be assigned a larger saliency score when used to train the classifier for the n -th landmark category.

Figure 3.6 Illustration of cell's saliency estimation, (a) and (d) are histograms of cells c_1 and c_2 respectively, (b) and (e) are codeword discrimination for landmark category n , (c) and (f) are the estimated saliency score for cells c_1 and c_2 respectively

Based on the previous discussion on cell visual saliency and codeword discrimination, the final saliency index γ_n^l for cell c_j^l to represent landmark category n is obtained as:

$$\gamma_n^l = w^l \left[(1 - \lambda) \omega(c_j^l) + \lambda \rho(c_j^l) \right] \quad (3.8)$$

where λ is the regularization parameter that controls the relative contribution between the cell visual saliency and codeword discrimination, and w^l is a level-dependent weight used to emphasize the cells with finer resolution. The idea that cells with finer resolution should

be given more emphasis is consistent with studies in [47]. The weight for the three levels, w^l ($l = 0, 1, 2$) is set to 0.25, 0.5, and 1, respectively. The histogram $\mathbf{h}_{I^n}^m$ of cell c_j^l is then weighted with its saliency index γ_n^l to obtain a weighted histogram $\mathbf{h}_{I^n}^m$

$$\mathbf{h}_{I^n}^m = \gamma_n^l \mathbf{h}_{I^n}^m \quad (3.9)$$

Based on the decomposition scheme in Section 3.3.1, we will obtain multiple weighted histograms $\mathbf{h}_{I^n}^m(c_j^l)$, $j=1, 2, \dots, J$; $l=0, 1, 2$. These weighted cell histograms are then concatenated to form a long vector to represent image I^n from n -th landmark category.

3.3.3 Spatial Pyramid Kernel

Most current spatial pyramid kernel [47] treats all the codewords with equal emphasis when calculating the distance between two cells' histograms. Suppose the weighted histograms of cell c_j^l in two images I^n and $I^{\bar{n}}$ are $\mathbf{h}_{I^n}^m$ and $\mathbf{h}_{I^{\bar{n}}}^m$, then the kernel K_j^l that is used to evaluate the two cells' similarity can be expressed as:

$$K_j^l(I^n, I^{\bar{n}}) = \sum_{m=1}^M k_j^l(m) \quad (3.10)$$

where

$$k_j^l(m) = f\left(\mathbf{h}_{I^n}^m, \mathbf{h}_{I^{\bar{n}}}^m\right), m = 1, 2, \dots, M \quad (3.11)$$

and f is an element-by-element histogram similarity function.

From (3.10), it can be seen that the current pyramid kernel K_j^l is just a simple summation of $k_j^l(m)$. It treats $k_j^l(m)$ from $m = 1$ to M equally. However, as discussed in the Section 3.3.2, different codewords in fact have different discriminative capability when representing different landmark categories n . The histogram distance $k_j^l(m)$ calculated

using more discriminative codewords should be given more emphasis. Based on this principle, the kernel function K_j^l for two cells is updated as follows:

$$K_j^l(I^n, I^{\bar{n}}) = \sum_{m=1}^M D(m, n) k_j^l(m) \quad (3.12)$$

From (3.12), it can be observed that the new kernel between two cells becomes a weighted summation of $k_j^l(m)$ in accordance with the codeword discrimination $D(m, n)$. Finally, the new spatial pyramid kernel K for two images $(I^n, I^{\bar{n}})$ is given as follows:

$$K(I^n, I^{\bar{n}}) = \sum_{l=0}^L \sum_{j=1}^J w^l K_j^l(I^n, I^{\bar{n}}) \quad (3.13)$$

$$K_j^l(I^n, I^{\bar{n}}) = \sum_{m=1}^M D(m, n) k_j^l(m) \quad (3.14)$$

$$k_j^l(m) = \frac{\left(\mathbf{h}_{I^n}^m - \mathbf{h}_{I^{\bar{n}}}^m \right)^2}{\mathbf{h}_{I^n}^m + \mathbf{h}_{I^{\bar{n}}}^m}, \quad m = 1, 2, \dots, M \quad (3.15)$$

where L is the number of image levels, J is the number of cells at image level l , and M is the codebook size. $\mathbf{h}_{I^n}^m$ is the weighted BoW histogram of the cell c_j^l in image I^n . We use χ^2 function in (3.15) to evaluate the similarity between two histograms. The new spatial pyramid kernel described from (3.13)-(3.15) is then used to train SVM classifiers to perform mobile landmark recognition.

3.4 Integration of Context and Content Analysis

Conventional context analysis usually employs location information for mobile landmark recognition [17]-[18]. Location analysis infers landmark based on the distance between the user's position and the landmarks. However, the location data such as the GPS readings may contain significant errors, especially in the dense built-up areas. This will reduce the recognition performance of mobile landmark recognition if location information alone is used. In view of this, this chapter proposes to complement the location information with

direction information. Today, increasingly more mobile devices are equipped with digital compass, such as the Google Android HTC Dream and iPhone. These mobile devices will provide direction information to enhance the performance of mobile landmark recognition.

The context analysis developed in this chapter involves the formation of visual viewcones based on the location and direction information. Using the user's position and camera's pointing direction, a viewcone can be drawn which corresponds to the camera phone's field-of-view as illustrated in Figure 3.7. Three suggested strategies are given as follows:

- (I) Nearest-to-center strategy: selecting the landmark category nearest to the center of the viewcone as the recognized landmark. In Figure 3.7, landmark 2 will be chosen based on this strategy.
- (II) Wide viewcone nearest-to-user strategy: selecting the landmark category nearest to the user's location from a wide viewcone (± 25 degrees). In Figure 3.7, landmark 1 will be chosen based on this strategy.
- (III) Narrow viewcone nearest-to-user strategy: selecting the landmark category nearest to the user's location from a narrow viewcone (± 10 degrees). In Figure 3.7, landmark 3 will be chosen based on this strategy.

Figure 3.7 Visual viewcone in context analysis

This chapter proposes an integration of context and content analysis to perform mobile landmark recognition. The integration approach is summarized in Algorithm 3.2.

Algorithm 3.2 Content-context integration for mobile landmark recognition

1. A filtering step is first adopted to shortlist a number of candidate landmarks L_i , $i=1, 2, \dots, K$ using the developed context-based viewcone strategy.
2. For each candidate landmark L_i , a context-based soft likelihood score $P_1(i)$ is estimated as follows:

$$P_1(i) = \cos(\theta_i) \cdot \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x_i^2 + y_i^2)\right) \quad (3.16)$$

where x_i and y_i are the latitudinal and longitudinal distance between the user and the candidate landmark L_i , σ is a scaling parameter, and θ_i is the angle between the candidate landmark and the center direction of the user's view cone.

3. For each candidate landmark L_i , a content-based soft likelihood score $P_2(i)$ is also estimated using the content analysis method as follows:

$$P_2(i) = \frac{\exp(s(i))}{\sum_{j=1}^K \exp(s(j))} \quad (3.17)$$

where $s(i)$ is the output of the i -th SVM landmark classifier for the query image.

Softmax function [18] is used to normalize the obtained value between $[0, 1]$.

4. The final likelihood score is then calculated to determine the recognized landmark I as follows:

$$\begin{aligned} I &= \arg \max_i (P_1(i) \cdot P_2(i)) \\ &= \arg \max_i \left(\cos(\theta_i) \cdot \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x_i^2 + y_i^2)\right) \cdot \frac{\exp(s(i))}{\sum_{j=1}^K \exp(s(j))} \right) \end{aligned} \quad (3.18)$$

The integration technique in Algorithm 3.2 is explained as follows. First, the context viewcone analysis is used to select a number of candidate landmarks. This will reduce the subsequent content analysis time in order to meet the fast response-time requirement of mobile users. Next, the context soft likelihoods are estimated. It considers the following two factors: (i) distance between the user and the candidate landmark, and (ii) the angle between the candidate landmark and the center direction of the user's viewcone. Intuitively, landmark that is closer to the user and nearer to the center of the viewcone will have higher degree of likelihood. The Gaussian and cosine functions are used to incorporate the distance and direction information, respectively. Next, the captured image is passed through the SVM classifiers of each candidate landmark to calculate the content analysis scores. Finally, the context- and content-based scores are combined to estimate the final likelihood for each candidate landmark. Landmark with the largest score will be selected as the recognized landmark.

3.5 Experimental Evaluation

3.5.1 Experimental Setup

To address the issue of limited processing power of mobile devices while achieving good recognition performance in fast response-time requirement, a client-server structure is adopted in the experiments. We use Google Android (HTC Dream G1) that has built-in GPS and digital compass as the mobile client, and a high performance computer as the backend sever. It has the following hardware/software specifications: Intel Core (TM) 2, Quad Processor (2.25GHz), 4G RAM, MS Windows XP operating system, Java for system interface, and Matlab R2008a for algorithm implementation.

To evaluate the performance of the proposed method, a landmark database called NTU50Landmark is constructed. The database covers a large variety/complexity of real-world conditions that can be used in the training of the landmark classifiers. The NTU50Landmark database consists of 3622 training images and 534 testing images for 50 landmark categories from the campus in Nanyang Technological University (NTU). The selected landmarks are well-recognized buildings, structures, and places-of-interests in NTU such as Chinese Heritage Center, Yunnan Monument and Pavilion, Auditorium Sculpture, etc. They are chosen based on feedbacks from various students and visitors that these are places-of-interests in NTU. The definition of a landmark is taken as a building, structure or place-of-interest that is unique or distinctive. The landmark images are captured using camera phones equipped with GPS receiver and digital compass. All the test images contain the location and direction information of the camera phone when the images are taken. Sample images of these landmarks are given in Figure 3.8 (left column). The geospatial distribution of the landmarks is given in Figure 3.8 (right column), and they cover an area of more than 2 km^2 . From the figure, it can be seen that the landmarks spread across the whole campus of NTU, with certain areas having a higher concentration of

landmarks. The numbers of the training and testing images for each landmark are mostly uniform. There are on average 70 images for training and 10 images for testing for each landmark.

Figure 3.8 Sample images (left column) and geospatial distribution (right column) of the 50 landmarks

In order to cover a large variety of capturing environments for the training images, Different image acquisition conditions are used when capturing the images. These include different camera settings (contrast, resolution), different capturing time (morning, noon, afternoon), illumination and weather changes (sunny, cloudy), changes in viewpoints (side, front, left, right, full, part), changes in scale (far, near) and occlusion. Figure 3.9 illustrates several sample images that are captured under different conditions.

Figure 3.9 Sample images in the NTU50Landmark database under different conditions

The following sections will demonstrate the performance of the methods under three different conditions: (i) content analysis alone, (ii) context analysis alone, and (iii) integration of content and context analysis. The performance metrics are given in terms of recognition accuracy and processing time. Recognition accuracy is defined as the ratio of number of correctly classified images to the total number of testing images, while the processing time includes content analysis, context analysis and classification time on the server.

3.5.2 Performance Evaluation of Proposed Content Analysis

To demonstrate the efficiency of different components in Section 3.3, the arrangements for different modules are outlined as follows: (i) Improvement 1: new spatial pyramid decomposition using BoW, (ii) Improvement 2: Improvement 1 + new pyramid cell saliency estimation, and (iii) Improvement 3: Improvement 2 + new spatial pyramid kernel. 3622 images from NTU50Landmark database are used for training and the rest 534 images are used for testing. The regularization parameter $\lambda = 0.7$ is determined empirically through experiments. The standard BoW method in [45] and spatial pyramid kernel with BoW

(SPK-BoW) in [47] are chosen as the baseline for comparison. Multi-scale patch sampling strategy (16×16 , 32×32 and 64×64 overlapping patches) is adopted for dense SIFT feature extraction in an image. These features are quantized to form a BoW histogram based on the codebook. One-versus-All SVM is then used to train the classifiers. The average recognition accuracy for these methods is given in Figure 3.10, under codebook sizes varying between 150 and 400. The error bars are given in terms of 95% confidence interval.

Figure 3.10 Recognition accuracy of different content analysis methods

Legends

SPK-BoW: spatial pyramid kernel in bags-of-words method [47]

Improvement1: new spatial pyramid decomposition with BoW

Improvement2: Improvement1 + pyramid cell saliency estimation

Improvement3: Improvement2 + new spatial pyramid kernel

From the figure, it is observed that the proposed methods (Improvements 1, 2 and 3) consistently outperform the benchmark BoW and SPK-BoW methods under various codebook sizes. The overall content analysis (Improvement 3) can achieve a recognition rate of around 93% for codebook size of 400. It can be seen that each component in the proposed content analysis (spatial pyramid decomposition, pyramid cell saliency estimation, and spatial pyramid kernel) can offer 2-3% improvement in the recognition rate, rendering an overall improvement around 6-9% over the SPK-BoW in [47] and 10-20% over the baseline BoW [45]. The improvements are due to the following reasons: (i) it estimates the saliency of different pyramid cells so that this information can be incorporated into the training of classifiers. This is consistent with the observation that image regions/patches with more visual saliency should be given more emphasis when considering its visual representation and (ii) the proposed spatial pyramid representation can represent landmark images more effectively by involving a combination of 1×3 and 3×1 cell structures as opposed to the conventional decomposition. This can better partition the image into the central region-of-interest and the background areas.

Figure 3.11 illustrates some visual examples where the proposed weighted method outperforms the nonweighted approaches. Three selected images and their estimated cell saliency are given in the figure. Brighter areas imply the cells contain more discriminative information. From the figure, it can be seen that the proposed method can provide more emphasis to those cells corresponding to landmarks. Therefore, the method is able to offer improvement over the nonweighted approaches where all cells are given equal weights.

Figure 3.11 Illustrative examples of the proposed cell saliency estimation

Further, we have also used different local descriptors including sparse SIFT [27], dense SIFT [12], SURF and Upright-SURF (U-SURF) [56] in the proposed method for comparison. The results are given in Table 3.1. From the table, it is observed that the dense SIFT descriptor offers the best performance, followed by U-SURF, SURF, and sparse SIFT. The reasons why dense SIFT has outperformed the other local descriptors is because the dense SIFT descriptors are obtained from overlapping regular patches in the image without rotation. As the landmarks images are usually acquired in upright position with little rotation, the dense SIFT descriptors turn out to be more effective than sparse SIFT and SURF that consider orientation information in landmark recognition.

Table 3.1 Performance comparison using different local descriptors

Local features	Sparse SIFT	SURF	U-SURF	Dense SIFT
Recognition rate (%)	86.5	89.3	91.6	93.1

Finally, the comparison experiments using the feature pooling method in [128] and the proposed spatial pyramid BoW method in this chapter have been conducted. For feature

pooling, the linear spatial pyramid matching method based on sparse coding and max pooling in [128] is used for comparison. Sparse coding selects salient codewords in the codebook to encode SIFT descriptors into sparse codes. These codes are then pooled from different spatial locations over different pyramid levels to generate a new image representation, which is then used by linear SVM for recognition. The experimental results are given in Figure 3.12. From the figure, it can be seen that the proposed method outperforms the feature pooling method by about 1-2%. The reason why the proposed method is able to perform well is because: (i) it adopts a new spatial pyramid decomposition method which can represent landmark images more effectively, (ii) it considers the codeword discriminative information among different landmark categories for spatial pyramid cell weighting, and (iii) it develops a new spatial pyramid kernel based on chi-square distance for SVM training. These are the improvements of the proposed method when compared with the feature pooling methods in [128].

Figure 3.12 Comparison between the feature pooling method [128] and the proposed method

3.5.3 Performance Evaluation of Context Analysis

This section will show the experimental evaluation of three proposed context analysis techniques highlighted in Section 3.4. We have also implemented the context analysis using GPS alone for comparison. It selects the landmark nearest to the user as the recognized

landmark. In order to demonstrate the performance of the techniques in various areas with different GPS errors, three sets of experiments are conducted: (a) dataset with small GPS errors (typically less than 10m) which consists of 120 test images captured mainly in the open sky areas, (b) dataset with medium-large GPS errors (typically between 50m and 100m) which consists of 110 test images captured mainly in the dense buildup areas, and (c) the complete dataset inclusive of (a) and (b) which consists of 534 test images captured in all environments. The results are given in Table 3.2, from which it can be seen that (i) the three proposed viewcone approaches consistently outperform the GPS alone method. This shows that the viewcone approaches based on fusion of GPS and direction are more effective than the GPS alone method. (ii) Image dataset with small GPS errors obtains higher recognition rate than that with medium-large GPS errors. This shows that medium-large GPS errors can reduce the recognition accuracy significantly. (iii) The narrow viewcone nearest-to-user strategy produces the best recognition accuracy. This is because the strategy utilizes the observation that mobile users tend to capture the landmark that is close to the user and near the center of the viewcone. (iv) The context analysis is faster than content analysis while its recognition accuracy is unsatisfactory. This is due to the ambiguity associated with multiple candidate landmarks in the viewcone. Therefore, context analysis needs to be integrated with content analysis to achieve good recognition accuracy.

When using the content-context integration analysis for mobile landmark recognition, the viewcone analysis experiences the limitation that if the coverage of the generated viewcone is not large enough, it will not be able to include the correct landmark in the candidate set. As illustrated in Figure 3.7, the coverage of the viewcone is determined by its view angle and radius. As the field-of-view for a camera phone is usually ± 30 degrees, the angle of the viewcone is thus set as 60 degrees. Then, a set of experiments are conducted by varying the radius to form a suitable viewcone to check whether the correct landmark is

included in the shortlisted candidates. The results are shown in Table 3.3. The ratio of the number of test images with viewcones containing the correct landmark to the number of all test images is used as the evaluation criteria. From the table, it can be seen that when the radius is increased to 200 meters, the generated viewcone will include all the correct landmarks for all the test images. When the viewcone radius becomes larger, the result remains the same but the computational complexity will be higher as more landmark candidates will be selected. Therefore, considering both the computational complexity and recognition performance, the radius and angle of the viewcone are set as 200m and 60 degrees respectively.

Table 3.2 Recognition accuracy of context analysis

(a) Image dataset with small GPS errors

Context analysis	Recognition rate (%)
Nearest-to-center strategy	91.7
Wide viewcone nearest-to-user strategy	90.0
Narrow viewcone nearest-to-user strategy	92.5
GPS alone	86.7

(b) Image dataset with medium-large GPS errors

Context analysis	Recognition rate (%)
Nearest-to-center strategy	66.4
Wide viewcone nearest-to-user strategy	67.3
Narrow viewcone nearest-to-user strategy	70.9
GPS alone	54.5

(c) Complete image dataset

Context analysis	Recognition rate (%)
Nearest-to-center strategy	73.0
Wide viewcone nearest-to-user strategy	74.0
Narrow viewcone nearest-to-user strategy	77.2
GPS alone	69.5

Table 3.3 Experimental results of varying the radius in the viewcone analysis

The result is defined as the ratio of the number of test images whose viewcones contain the correct landmark to the total number of the test images.

Viewcone radius	50m	100m	150m	200m
Result	0.80	0.93	0.96	1

3.5.4 Performance Comparison of Content, Context and Their Integration

Based on the previous discussion in Sections 3.5.2 and 3.5.3, it is clear that using context or content analysis alone is not able to achieve satisfactory results as context analysis alone is susceptible to GPS errors whereas content analysis alone is computationally intensive as it needs to compare the query image with all landmark categories in the database. In view of this, this section will demonstrate the performance comparison of content analysis, context analysis, and integration of content-context analysis. The content analysis used is the overall Improvement 3 used in Section 3.5.2 and the context analysis used is the narrow viewcone nearest-to-user strategy used in Section 3.5.3. The content-context integration in Section 3.4 is run. The candidate number will depend on how many landmarks are located in the viewcone. These candidates are then used for subsequent visual recognition. The results are given in Table 3.4. From the table it can be seen that the integration technique achieves better recognition rate than the context and content analysis alone for all three datasets. In particular, for the whole landmark dataset, the integration technique achieves an improvement of 19.4% and 3.5% over the context and content analysis, respectively. This shows that the integration method is able to complement the advantages of content analysis and context analysis to improve the recognition rate of the system.

In terms of the computational time, it is noted that context analysis alone requires the least recognition time, albeit with a much lower recognition accuracy. The proposed integration technique, on the other hand, offers the best compromise as it reduces the computational time of the content analysis by approximately 50%, while achieving much

superior recognition rate than the context analysis alone. The reason why the integration method is much faster than the content analysis method is because using the context analysis can reduce the search scope of landmark categories from a large number (50 categories in the experiments) to only a few candidates in the viewcone. In short, the proposed content-context integration approach utilizes the advantages of content and context analysis respectively, e.g., the high recognition accuracy of content analysis and the fast processing of context analysis to achieve excellent performance. This integration produces a good recognition rate of 96.6% while maintaining a reasonably low computational time of around 5 seconds.

Table 3.4 Performance comparison for content analysis, context analysis and their integration

Method	Dataset	Recognition rate (%)	Average Recognition Time (s)
Content analysis alone	Small GPS errors (Open sky areas)	94.2	9.8
	Medium-large GPS errors (Buildup areas)	92.7	10.1
	Whole dataset	93.1	10.2
Context analysis alone	Small GPS errors (Open sky areas)	92.5	0.0034
	Medium-large GPS errors (Buildup areas)	70.9	0.0035
	Whole dataset	77.2	0.0039
Proposed integration of context and content analysis	Small GPS errors (Open sky areas)	98.3	5.2
	Medium-large GPS errors (Buildup areas)	95.4	5.0
	Whole dataset	96.6	5.1

Next, we will consider the image query time that is defined as the total processing time between the moment the user submits an image query on the mobile device to the moment he/she receives the recognition result. Based on the experiments, the transmission time through wireless LAN network is 0.6 second for an image of dimension 240×320 pixels. The time spent on data wrapping/formatting and function call is 1.7 second. The recognition time on the server which includes content analysis, context analysis and classification time

for mobile landmark recognition is 5.1s second on average. Therefore, the final image query time based on content and context integration is on average $0.6+1.7+5.1=7.4$ seconds. It is worth noting that the proposed back-end algorithm is currently implemented mainly in Matlab. The total recognition time can be further reduced if it is implemented in compiling language such as C/JAVA with code optimization.

3.5.5 Experiments on the Oxford Buildings Dataset

In order to demonstrate the effectiveness of the proposed method on other datasets, we have used a second dataset to validate the proposed algorithms. We have downloaded the Oxford building dataset from [83], and used it for the experiments. The dataset consists of 5062 images collected from Flickr for Oxford landmarks. It has been manually annotated for 11 different landmarks. Each landmark category contains 5 queries, giving rise to a total of 55 queries over which the landmark recognition system can be evaluated. During the training phase, the one-versus-all strategy is used to train an SVM for each landmark category. During the testing phase, the landmark category that gives the highest SVM matching score for the query image is recognized as the correct landmark. It is noted that the Oxford dataset does not contain the location or direction information, we can only use them to evaluate the effectiveness of the proposed content analysis method. The proposed approach is applied and compared with other benchmark methods. The codebook size is set as 300. The kernel SVM is used in both the proposed method and the spatial pyramid method [47]. For the proposed spatial pyramid kernel, the codeword discriminative information is used to train the SVM for each category. This is in contrast with the pyramid kernel in [47] where the codeword discriminative information is not considered. The results are given in the Table 3.5. It can be seen that the proposed method has a good recognition accuracy of 81.8%, which is 9.1% higher than Lazebnik's method in [47]. This shows that the proposed method

can obtain good performance in a different benchmark dataset, not only in the constructed NTU landmark database. Further, the reason that the recognition accuracy (81.8%) on the Oxford database is lower than the recognition accuracy (93%) on the NTU50Landmark database is mainly due to the difference on the training dataset collection process. The NTU database is collected with the training purpose in mind, e.g., different perspective views, illuminations, scales, etc. While for the Oxford training set, the images are collected mainly from the Flickr website, which does not contain comprehensive image capturing conditions.

Table 3.5 Experimental results on the Oxford dataset

Method	Csurka's method [45]	Lazebnik's method [47]	The proposed method
Accuracy (%)	54.5	72.7	81.8

3.6 Summary

This chapter presents a content-context integration approach for mobile landmark recognition. The key contributions of this chapter include (i) development of a new BoW content analysis method that considers the codeword discriminative information, and (ii) proposal of an effective content and context integration approach, which takes the GPS location and direction information into consideration. Experimental results based on the NTU50Landmark database and Oxford dataset show that the proposed methods can achieve good recognition performance in mobile landmark recognition.

Chapter 4 Discriminative Learning BoW for Mobile Landmark Recognition

4.1 Introduction

In Chapter 3, a content and context integration approach is developed for mobile landmark recognition. The content analysis improves the conventional BoW method, and achieves good recognition performance. However, the improved BoW method is still insufficient since no discriminative information of the landmark images is learned. For instance, they consider the patches/regions that are quantized to a particular codeword as equally important when generating the BoW histogram. Although some works that start to consider the discriminative information of the image patches have appeared, they mainly focus on selecting the representative patches for training, and discard the rest. This binary hard selection approach results in underutilization of the information available, as some discarded patches may still contain useful discriminative information. Further, not all the selected patches will contribute equally to the learning process.

In view of this, this chapter presents a new discriminative BoW approach for mobile landmark recognition. The main contribution of the method is that the representative and discriminative information of the landmark is learned at three levels: patches, images and codewords. The patch discriminative information for each landmark is first learned and incorporated through vector quantization to generate continuous BoW histograms. Coupled with the learned representative information of the images and codewords, these histograms are used to train an ensemble of classifiers using fuzzy SVM. Finally, the proposed discriminative BoW approach is integrated with the GPS location and direction information through the method in Chapter 3 to perform mobile landmark recognition. Experimental

results on two different datasets show that the proposed method is effective in mobile landmark recognition.

The remainder of this chapter is organized as follows. A literature review of the patch discrimination leaning works is presented in Section 4.2. The framework overview of the discriminative BoW approach is given in Section 4.3. The patch discrimination learning and continuous BoW histogram generation are discussed in Section 4.4. The fuzzy training framework including the image signature weighting, codebook compactness determination and fuzzy SVM training is discussed in Section 4.5. The content and context integration analysis is given in Section 4.6. Experimental results and discussions are given in Section 4.7. Section 4.8 concludes the chapter with a summary of our findings.

4.2 Related Work

According to study on the recent works [13], [18], [60], [65], [69]-[70], [85]-[86], it is noted that the BoW-based methods have become the mainstream in landmark recognition. However, a shortcoming of these BoW methods [6], [10], [12], [33], [47] is that they consider all the local patches/regions obtained from the image as equally important for training. They do not discern different importance of these patches in recognizing the landmarks. For example, some local patches in the image may contain background information such as the sky and grass that is common to many different landmark categories. Such patches, however, are not discriminative for recognition, and hence their importance should be reduced when generating the BoW histogram. On the other hand, patches consisting of landmark should be emphasized.

Considering this, some researchers have studied the discriminative information of various image patches in recent years [18], [88]-[90]. A probability density response map (PDRM) has been proposed for each image in [88]. PDRM estimates the likelihood of local

patches being produced by each semantic class. However, this method requires each image in the database to be partitioned into different subimages, and labeled accordingly. This is laborious and time intensive when the database is large. A method for automatic detection of image-specific and spatially-localized groups of features is developed in [89]. The features which are confusing are suppressed, and then combined with BoW to improve the place recognition performance. A shortcoming of this method is that finding the confusing features is time-consuming as it requires one-to-one image matching. A patch saliency estimation based on low-level contextual information has been proposed in [90]. It adopts the bags-of-features framework to capture the co-occurrence and spatial information between codewords. A shortcoming of this method is that it only uses the contextual information within the same landmark category to represent the image and ignores the discriminative information across other categories. A landmark recognition system called Snap2Tell has been developed by Lim et al. in [18]. They use a non-parametric density estimator to evaluate each patch's importance. However, the method is sensitive to the window size of the estimator and it has high computational complexity, hence diminishing its effectiveness for mobile landmark recognition. In addition, some techniques that try to find representative images to model each landmark have been developed in [91], [92]. These methods are unsuitable for mobile applications as the one-to-one image geometric verification requires high computational cost and the global descriptor used for clustering lack a detailed characterization for the image.

In summary, a common shortcoming of the above-mentioned methods is that they mainly focus on selecting representative patches for training, and discard the rest. This binary hard decision approach in patch selection results in underutilization of the information available, as some discarded patches may still contain useful discriminative information. In addition, they do not have any effective image-level discriminative learning

for recognition, which plays an important role in landmark recognition since different images usually have different significance.

In view of this, we propose a new BoW approach for mobile landmark recognition based on discriminative learning of patches, images and codewords in this chapter. The main contributions of this chapter are: (i) An iterative learning approach based on a differential Gaussian mixture model (DGMM) is developed to estimate the discriminative information of each image patch. This information is then incorporated into vector quantization to generate a continuous histogram. (ii) An image signature weighting method is developed to evaluate the effectiveness of each image in representing its landmark category, which is then used to train a discriminative classifier through a fuzzy SVM that consists of a Laplacian RBF kernel that considers codebook compactness. The context information such as GPS and direction is finally integrated with content analysis to speed up the recognition time and improve the recognition performance. The experiment results using two landmark databases demonstrate the effectiveness of the proposed method.

4.3 Framework Overview

The overview of the proposed content analysis is given in Figure 4.1. It consists of three parts: patch-level continuous BoW histogram generation, image-level signature weight estimation, and fuzzy SVM training with LRBF kernel that considers codebook compactness. A DGMM is learned for each landmark category based on local features extracted from the dense patches in the images. An iterative learning approach is developed to filter out the unimportant patches while retaining the discriminative ones. This is achieved through the log-likelihood ratio of the GMM output between the positive and negative categories. This score is then used to determine whether to select it for next iteration of DGMM learning. A continuous BoW histogram is generated by incorporating

the patches' discriminative information into the quantization process. Next, the generated histograms are used to estimate the image signature weight to measure the effectiveness of each image in representing its landmark category. The training images are used to train fuzzy SVM classifiers. The classifier uses an LRBF kernel that integrates codebook compactness information. During online testing, the BoW histogram of each image is generated and forwarded to the trained FSVMs for classification. The following sections will provide a detailed description of the proposed approach.

Figure 4.1 Proposed framework for mobile landmark recognition

4.4 Discriminative BoW

4.4.1 Iterative Patch Learning

In most landmark images, there exist various kinds of background patches (e.g., grass, trees, sky, ground, etc.) that are common across many different landmark categories, as shown in Figure 4.2. The backgrounds are less discriminative for recognition and thus should be given less emphasis when performing training and classification. In view of this, this section develops a DGMM to estimate the discriminative information of patches for each landmark category. The reason to adopt GMM to model the distribution of image patches in each landmark category is based on the observation that patches from the same scene usually share similar visual features and tend to be clustered together, while patches from different scenes usually differ in appearance and tend to form different clusters. As opposed to the conventional GMM modeling method in [93], the DGMM is more effective as the model is iteratively refined through continuously filtering out the unimportant patches while retaining the discriminative ones. The importance of the patches is mainly estimated based on the differential between the target and other negative categories.

Figure 4.2 Selected sample landmark images demonstrating various scenes

In order to construct the DGMM for each category, a multiscale decomposition strategy is first applied to the training images to obtain an ensemble of dense patches. Three different scales are adopted as 16×16 , 32×32 and 64×64 (pixels). SIFT descriptor is used as

each patch's feature representation. SIFT descriptor is shown to be robust to changes in illumination, viewpoint, and clutter in [27]. Gaussian function is used to normalize the features in each dimension within $[0, 1]$. Principal Component Analysis (PCA) is then used to reduce the feature's dimension while maintaining discriminative information across different categories, similar to [94]. Next, the Expectation Maximization (EM) algorithm is employed to estimate the parameters of each category's GMM, which is used to model the distribution of the patches in each category ($n = 1, 2, \dots, N$, where N is the total number of landmark categories):

$$P(\mathbf{f}) = \sum_{i=1}^k \pi_i P(\mathbf{f}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \quad (4.1)$$

where

$$P(\mathbf{f}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{\frac{d}{2}} \det(\boldsymbol{\Sigma}_i)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{f} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{f} - \boldsymbol{\mu}_i)\right) \quad (4.2)$$

where $\mathbf{f}$ is the patch's descriptor, π_i , $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are the mixing proportion, mean and covariance of the i -th Gaussian function respectively, d is the dimension of the descriptor, and k is the total number of Gaussian components which is set as 8 through experiment. From our study, it is observed that the system provides relatively stable performance when the number of components ranges from 7 to 15. $P(\mathbf{f})$ refers to the GMM output of the current target category. After learning the GMM for each category, the DGMM output ω for each patch is given by:

$$\begin{aligned} \omega &= \log \frac{P(\mathbf{f})}{\bar{P}(\mathbf{f})} \\ &= \log P(\mathbf{f}) - \log \bar{P}(\mathbf{f}) \\ &= \log \left(\sum_{i=1}^k \pi_i P(\mathbf{f}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \right) - \frac{1}{N-1} \sum_{\substack{j=1, \\ j \neq n}}^N \log \left(\sum_{i=1}^k \pi_i^j P(\mathbf{f}; \boldsymbol{\mu}_i^j, \boldsymbol{\Sigma}_i^j) \right); \end{aligned} \quad (4.3)$$

where $\bar{P}(\mathbf{f})$ refers to the mean value of the GMM outputs for the negative categories. Here we use the union of all the other categories as the negative categories.

The DGMM is developed due to the observation that backgrounds such as the sky and grass are common in different landmark categories while the foreground landmark is specific to a category. As a result, background patches will have moderate GMM scores both in the target and negative categories. The differential operation will cause output of these background patches to be small. On the other hand, the landmark patches that are distinctive to a specific category will have large differential outputs. The discriminative patches with DGMM scores greater than a defined threshold are selected for next iteration of DGMM training. The process is repeated until the final DGMM scores for all the selected patches are above the threshold or a maximum number of iterations is reached. We have set the threshold to be 0.4 since it has been experimentally shown to produce good performance. Finally after constructing the DGMM for each category, the discriminative score ω of each patch is obtained by (4.3).

4.4.2 BoW Histogram Generation

The BoW histogram is generated by incorporating the discriminative information of the patches into the quantization process. The visual codebook $\mathbf{w}_m$ ($m = 1, 2, \dots, M$; where M is the size of the codebook) is constructed using hierarchical K -means clustering method given in [42]. The BoW histogram generation process is summarized in Algorithm 4.1.

Algorithm 4.1 BoW histogram generation

1. For each landmark image in the database, perform multiscale decomposition as in Section 4.4.1 and extract the SIFT descriptors $\mathbf{f}_j$ from the dense patches, and learn their

discriminative scores $\omega_j (j = 1, 2, \dots, N_p; N_p \text{ is the number of patches})$ using (4.3).

2. Integrate the discriminative scores of the patches with the feature quantization to generate the BoW histogram as follows:

$$\mathbf{h}(i) = \sum_{j=1}^{N_p} \omega_j \mathbf{I}_i(\mathbf{f}_j), \quad i = 1, 2, \dots, M \quad (4.4)$$

where $\mathbf{h}$ is the obtained BoW histogram, $\mathbf{I}_i(\mathbf{f}_j)$ is the identity function to determine whether $\mathbf{f}_j$ belongs to $\mathbf{w}_i$ and is defined as:

$$\mathbf{I}_i(\mathbf{f}_j) = \begin{cases} 1, & \text{if } i = \arg \min_m \|\mathbf{f}_j - \mathbf{w}_m\|, \quad m = 1, 2, \dots, M \\ 0, & \text{otherwise} \end{cases} \quad (4.5)$$

and $\|\mathbf{f}_j - \mathbf{w}_m\|$ is the L2 norm between $\mathbf{f}_j$ and visual word $\mathbf{w}_m$.

From the algorithm it can be seen that the method accumulates the discriminative scores of the patches quantized as each codeword to compute the continuous value of each index in the histogram. In such a way, the constructed continuous histogram will be more robust to the background/clutter in the image since the background patches will have smaller discriminative scores, and thus contribute less to the histogram generation.

Figure 4.3 illustrates the difference between the conventional and proposed BoW histograms. From the figure, it can be seen that the proposed continuous histogram generation in the third row assigns continuous weights ranging between 0 and 1 of different patches to each codeword. For the patches that are quantized to the same codeword, they may be assigned different relevance scores based on how much discriminative information they contain. The scores of these patches are calculated using a DGMM output given in (4.3). In general, it is observed that larger scores (stronger emphasis) are assigned to patches that are quantized to discriminative landmark codewords (in black) while lower scores (weaker emphasis) are assigned to patches that are quantized to the background codewords

(in green and grey). This is as opposed to the conventional BoW histogram in the second row that assigns discrete value of 1 to all patches that are quantized to each codeword. The advantages of employing the discriminative BoW in representing the landmark image can be observed since its continuous-valued histogram shows stronger components/indices for codewords corresponding to the foreground landmark.

Figure 4.3 Illustration of conventional histogram (second row) and developed histogram (third row).

4.5 Fuzzy SVM Training

4.5.1 Image Signature Weight Estimation

In order to cover a large variety/complexity of real-world conditions that can be used in the training of the landmark classifiers, various images captured from different viewpoints, clutter and occlusion are collected for training. As some training images may be less representative of the landmarks, this suggests that different training images should be given different emphasis/weights during the training. For example, images with significant occlusion may not be representative of the landmark categories, and hence should be de-

emphasized during the training. Therefore, an image signature weighting method is proposed to estimate how representative an image is for a category. The method can be described as follows. For each image in the target landmark category n , the method in Section 4.4.1 is used to learn its dense patches' discriminative scores $\omega_j (j = 1, 2 \dots, N_p)$ and generate the histogram $\mathbf{h}$. The image signature weight is then estimated based on two factors:

The first factor is the local patch's discriminative information, which sums up the discriminative scores of all the patches in the image, denoted as:

$$P(e|n) = \frac{1}{N_p} \sum_{j=1}^{N_p} \omega_j = \frac{1}{M} \sum_{i=1}^M \mathbf{h}(i) \quad (4.6)$$

where $P(e|n)$ is the score to evaluate the image's effectiveness to represent category n .

The second factor is the overall image histogram's within-class and between-class similarity, which calculates the ratio of the Mahalanobis distances between the image to the nearest negative category and image to the target category. The ratio provides an indication of the representative capability of the image for its category. This is based on the observation that the images with high ratio are more discriminative for its category. The calculation is given as:

$$P(o|n) = \frac{\min_{\substack{j=1,2,\dots,N \\ j \neq n}} ((\mathbf{h} - \boldsymbol{\mu}_j) \mathbf{A} (\mathbf{h} - \boldsymbol{\mu}_j)^{-1})}{(\mathbf{h} - \boldsymbol{\mu}_n) \mathbf{A} (\mathbf{h} - \boldsymbol{\mu}_n)^{-1}} \quad (4.7)$$

where $P(o|n)$ is the representation score, $\boldsymbol{\mu}_n$ and $\boldsymbol{\mu}_j$ correspond to the means of the target and negative category's histograms, and $\mathbf{A}$ is the matrix to be learned from the labeled histograms. In order to learn a good Mahalanobis distance matrix $\mathbf{A}$ which makes the histograms within the same category as close as possible and those across different categories as far away as possible, an objective function is proposed as:

$$\mathbf{A}^* = \min_{\mathbf{A}} f(\mathbf{A}) = \min_{\mathbf{A}} \sum_{i=1}^N (\mathbf{H}_{(i)} - \boldsymbol{\mu}_i) \mathbf{A} (\mathbf{H}_{(i)} - \boldsymbol{\mu}_i)^{-1} - \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^N (\mathbf{H}_{(i)} - \boldsymbol{\mu}_j) \mathbf{A} (\mathbf{H}_{(i)} - \boldsymbol{\mu}_j)^{-1} \quad (4.8)$$

where $\mathbf{A}$ is Positive Semi-Definite and $\mathbf{H}_{(i)}$ denotes all the histograms in the i -th category.

The first term is the total variation of the histograms within the same categories, while the second term refers to the variation of the histograms across different categories. To optimize this function in (4.8) with respect to $\mathbf{A}$, we use the gradient descent method:

$$\mathbf{A}_{t+1} = \mathbf{A}_t - \alpha \nabla f(\mathbf{A}_t) \quad (4.9)$$

where α is the learning parameter, and ∇ is the gradient. After obtaining $\mathbf{A}_{t+1}$, the eigenvalues λ_i and eigenvectors $\mathbf{v}_i$ are calculated using eigen decomposition, and only the positive eigenvalues are retained to reconstruct $\mathbf{A}_{t+1}$ to ensure that it is positive semi-definite:

$$\mathbf{A}_{t+1} = \sum_i \max(\lambda_i, 0) \mathbf{v}_i \mathbf{v}_i^T \quad (4.10)$$

The reason to include the first factor in calculating the image signature weight is because an image that has more discriminative patches should naturally be given more emphasis during training. For the second factor, it is based on the observation that an image that is closer to the target category in the feature space and farther away from the negative category will be more representative of its category. In (4.7), the numerator indicates the distance of the image to the nearest negative category, while the denominator indicates the distance of the image to the target category. Finally, the two scores in (4.8) and (4.9) are combined to give an overall estimate of the image signature weight using:

$$s = P(e|n) \cdot P(o|n) \quad (4.11)$$

4.5.2 FSVM Training

In the previous sections, we generate a continuous BoW histogram and learn the signature weight for each image. In order to take the relative importance of the training images into consideration, a fuzzy SVM [43] instead of traditional SVM is adopted for classifier training. In this work, the developed fuzzy SVM is extended by incorporating the codebook compactness information into the kernel design. Compared with the binary SVM classifiers that are used in existing landmark recognition [17]-[18], FSVM can take the relative importance of the training samples into consideration and provide more emphasis to the training images with larger fidelity. Suppose a set of training images are given: $S = \{y_i, \mathbf{h}_i, s_i\}, i = 1, 2, \dots, N_t$ where $\mathbf{h}_i$ is the generated BoW histogram for each image, y_i is its label, $y_i \in \{+1, -1\}$, s_i is the image signature weight in (4.11) and N_t is the number of training images. FSVM first maps the input data into a high-dimensional feature space through a mapping function $\mathbf{z} = \varphi(\mathbf{h})$. The optimal separating hyperplane can then be obtained by solving the following optimization problem:

$$\begin{aligned} & \text{Minimize } \frac{1}{2} \|\mathbf{w}\|^2 + R \sum_{i=1}^{N_t} s_i \varepsilon_i \\ & \text{Subject to } y_i (\langle \mathbf{w}, \mathbf{z}_i \rangle + b) \geq 1 - \varepsilon_i, \\ & \varepsilon_i \geq 0, i = 1, 2, \dots, I \end{aligned} \tag{4.12}$$

where $\mathbf{w}$ is the normal vector of the hyperplane, and b is the bias that is a scalar. ε_i is the slack variable related to classification errors. R is a regularization parameter that controls the balance between margin maximization and classification errors. It is worth noting that the error term ε_i is scaled by the membership value s_i from (4.11), which in fact reflects the importance of the patch. The higher the value, the more impact it will have on the FSVM training. The optimization problem of FSVM can be transformed into dual problem by constructing the Lagrangian:

$$\text{Maximize } \sum_{i=1}^{N_t} \alpha_i - \frac{1}{2} \sum_{i=1}^{N_t} \sum_{j=1}^{N_t} \alpha_i \alpha_j y_i y_j \mathbf{z}_i \mathbf{z}_j$$

$$\text{Subject to } \sum_{i=1}^{N_t} y_i \alpha_i = 0, 0 \leq \alpha_i \leq s_i R, i = 1, 2, \dots, N_t \quad (4.13)$$

where α_i is the Lagrange multiplier. In order to determine the mapping function φ , Laplacian RBF (LRBF) is adopted as the matching kernel since it has been shown to have superior performance in histogram-based image categorization [80]. The codebook compactness is then analyzed and incorporated into the LRBF kernel. Codebook compactness refers to the compactness of each codeword in the cluster. It is defined as follows:

$$c(m) = \frac{1}{N_m} \sum_{j=1}^{N_m} \|\mathbf{f}_j - \mathbf{w}_m\|, m = 1, 2, \dots, M \quad (4.14)$$

where $\|\mathbf{f}_j - \mathbf{w}_m\|$ is the L2 norm between each patch descriptor $\mathbf{f}_j$ and codeword $\mathbf{w}_m$. N_m is the number of the patches that are clustered to word $\mathbf{w}_m$, $c(m)$ is the compactness score for codeword m . From (4.14), it can be seen that the compactness represents how tight the descriptors distributed around the codeword. Smaller value of $c(m)$ means codeword m has stronger compactness and thus is more representative of the landmarks. The value of $c(m)$ is further normalized between 0 and 1 by the sigmoid function, and used for codeword weighting. After this, the weight is incorporated into the histogram matching kernel for the FSVM training, which is summarized as follows. Suppose $\mathbf{h}_1$ and $\mathbf{h}_2$ represent the histograms of two images, $\mathbf{D}$ is a diagonal matrix that can be denoted as $\mathbf{D} = \text{diag}\left(\frac{c(1)}{h(1)}, \frac{c(2)}{h(2)}, \frac{c(m)}{h(m)}, \dots, \frac{c(M)}{h(M)}\right)$ where $h(m)$ is the m -th component of $\mathbf{h}_1 + \mathbf{h}_2$. The coefficient $h(m)$ is used to normalize the distance between the two histograms in (4.15). The final matching kernel K for the two histograms can be defined as:

$$K(\mathbf{h}_1, \mathbf{h}_2) = \exp\left\{-(\mathbf{h}_1 - \mathbf{h}_2)^T \mathbf{D}(\mathbf{h}_1 - \mathbf{h}_2)\right\} \quad (4.15)$$

From (4.15) it is noted that the new LRBF kernel provides stronger emphasis to more compact codewords that have stronger representative power while providing weaker emphasis to the less compact codewords that have weaker representative power. This causes the kernel to be more discriminative in learning the histogram distance. In order to take the spatial layout information of the image into consideration, spatial pyramid matching method in [47] is also incorporated into our work, which calculates a continuous BoW histogram for each spatial grid at each image level and then concatenates them into a long descriptor. During the recognition stage, the BoW histogram is generated and forwarded to the trained classifiers for landmark recognition.

4.6 Content and Context Integration Analysis

The integration framework adopts the same approach as in Chapter 3, which involves a two-stage process of context analysis followed by content analysis. In the first stage, the context analysis employs a fusion of location and direction information to obtain a viewcone. The viewcone is the field-of-view of the camera phones (typically around ± 30 degrees), with a radius of 200 meters from the user. Landmark categories that fall within the viewcone will be shortlisted as the candidate landmarks. In the second stage, the proposed discriminative BoW approach as explained before will perform landmark recognition from the shortlisted landmark categories. This integration framework will therefore improve the recognition rate of the landmark while reducing the computational time.

4.7 Experimental Evaluation

The NTU50Landmark campus database and Oxford building dataset as described in Chapter 3 are used for experimental evaluation. The experiment results are as follows.

4.7.1 Illustration of the Proposed Patch Learning Method

In this section we demonstrate the performance of the proposed patch discriminative learning in landmark recognition. The images are first decomposed into dense patches. Each patch is then assigned a discriminative score using (4.3). Figure 4.4 illustrates some examples of learned scores for patches in landmark images. Higher luminance value indicates that the patch has larger score and is more important in learning. From the figure it is seen that patches with larger scores tend to coincide with the landmark areas. This shows that the proposed patch learning is effective in identifying the discriminative patches in the landmark images.

Figure 4.4 Illustration of learned patch discriminative information in some landmark images

4.7.2 Performance Evaluation on NTU50Landmark Database

In this section, we use the SPK-BoW method in [47] as the baseline for comparison, and then combine SPK-BoW with the following respective components to evaluate their individual performance gain: (i) SPK with the continuous BoW histogram, (ii) SPK with the developed FSVM, and (iii) SPK with the codebook consistency, and (iv) SPK with all the components (i)-(iii). The results are given in Figure 4.5.

Figure 4.5 Experiment results of the proposed method

From the figure, it is observed that each of the three components can provide clear performance improvement when compared with the benchmark BoW and SPK-BoW methods for codebook size varying between 150 and 400. We follow the SPK method in [47] and keep the codebook size to be within 400. This is because that when combining with spatial pyramid method, a codebook size of 400 is enough to achieve good performance. From the experiments it is also seen that the results reach steady values when codebook size is greater than 400. It is observed that the continuous BoW histogram generated by patch-level learning offers a performance gain of 4-6%, the fuzzy SVM training using image-level soft scores offers a performance gain of 2-3%, and the SVM kernel using the codeword-level compactness offers a performance gain of 0.5-2%. It achieves a 6-11% improvement over the SPK-BoW in [47] and 15-23% over the baseline BoW [45]. When all the three components are used, the method can achieve a recognition rate of 94.4% for codebook size

of 400. The improvements are due to the following reasons: (i) the patch-level discriminative learning can differentiate importance of different patches and incorporate this information to generate a continuous BoW histogram, (ii) the image-level signature weight estimation can discover the effectiveness of different training images to represent their categories so that FSVM can utilize this information to train the classifier, and (iii) the codebook compactness is analyzed and incorporated into the spatial pyramid kernel to make the histogram matching more robust.

Next we compare the proposed patch discrimination learning with other patch learning methods. Two representative patch learning methods, namely D. Parikh's in [90] and J. Knopp's in [89] are implemented for comparison. The codebook size is set as 400. The results are given in Table 4.1. From the table, it can be seen that the proposed method achieves the best performance, and outperforms the Parikh's method [90] and J. Knopp's method [89] by 5.6% and 3.4% respectively.

Table 4.1 Recognition results of the proposed discriminative learning in comparison with other methods

Patch discriminative learning method	Recognition accuracy (%)
Parikh's method [90]	88.8
Knopp's method [89]	91.0
Proposed patch learning method	94.4

In order to demonstrate the performance of the proposed method and compare it with the other methods using different kernels, we have conducted additional experiments using the proposed kernel, the chi-squared kernel, and the histogram intersection kernel. The results are given as Table 4.2. From the table, it can be seen that: (i) the proposed method consistently outperforms the other two methods for different kernels, and (ii) the proposed kernel presents better performance than the other two kernels for each method. This shows the effectiveness of the proposed method and kernel for landmark recognition.

Table 4.2 Comparison of different methods using different kernels

Methods Kernels	Proposed method	BoW [45]	SPK-BoW [47]
Proposed kernel	94.4	79.4	88.0
Chi squared kernel	91.6	78.5	87.3
Histogram intersection kernel	88.6	75.1	84.1

We have also compared the proposed method with two state-of-the-art geometric verification methods: (i) kd-tree indexed full patch representation with RANSAC-based geometric verification as in [87], and (ii) kd-tree indexed BoW representation with RANSAC-based geometric verification for image re-ranking, similar to [76]. The top returned image is used to determine which category the query image belongs to. Kd-tree is used to speed up the matching process in [87] and [76]. The results are shown in Table 4.3. From the table, it can be seen that the proposed method achieves the highest recognition rate, and offers an improvement of 21% and 9.2% over the methods in [87] and [76], respectively. The recognition time for the three methods is comparable. It is also noted that even though the kd-tree searching is fast, the RANSAC-based verification process has a high computational cost which contributes significantly to the recognition time in [87] and [76].

Table 4.3 Comparison of proposed methods versus geometric verification methods

Method	Recognition accuracy (%)	Average recognition time (s)
Full patch representation with RANSAC geometric verification [87]	73.4	7.1
BoW representation with RANSAC geometric verification [76]	85.2	8.1
Proposed method	94.4	6.8

Finally, we also conducted some experiments to compare the performance of the proposed codebook compactness method with the conventional TF-IDF approach [95], and

found that the word compactness method can offer a 0.8% improvement in recognition rate when compared with the TF-IDF method.

4.7.3 Performance Evaluation on Oxford Database

In this experiment, the proposed method is applied and compared with several other representative landmark recognition methods in [90] and [89] on the Oxford dataset. For the proposed method, the discriminative BoW histogram is used as the image representation. The developed LRBF kernel considering the codeword compactness is used to train the SVM. For other methods, the patch saliency learning methods in [89][90] are respectively used to generate the BoW histograms for image representation, and the LRBF kernel without considering the codeword compactness is used to train the SVM. The results are given in the Table 4.4. From the table it can be seen that, the proposed method achieves a good recognition accuracy of 85.5%, which is 9.1% and 5.5% higher than the method in [90] and [89] respectively.

Table 4.4 Experimental results on the Oxford building dataset

Method	Recognition accuracy (%)
Parikh's method [90]	76.4
Knopp's method [89]	80.0
Proposed method	85.5

4.7.4 Performance Evaluation of Content-Context Integration

This section integrates context information captured by the mobile phone with the proposed content analysis for landmark recognition. The NTU50Landmark database is selected for experiment. The content-context integration approach as discussed in Section 4.6 is adopted for recognition. Table 4.5 gives the experimental results. From the table it can be seen that

methods integrating content and context analysis obtain better performance and results in much faster recognition time than method using content analysis alone. This is because the analysis of the context information can reduce the search space of landmark categories from a large number to a few category candidates.

Table 4.5 Experimental results of content-context integration

Method	Recognition rate (%)	Average total recognition time (s)
Proposed content analysis	94.4	6.8
Context: location Content: proposed content analysis	95.5	3.8
Context: location and direction Content: proposed content analysis	97.4	3.2

4.7.5 Evaluation of The Combined Methods from Chapters 3 and 4

In this section, the cell-based technique in Chapter 3 and the patch-based BoW technique in this chapter are combined using the same spatial representation for landmark recognition. First, a landmark image is partitioned into multiple pyramid cells and their saliency scores are computed using the spatial pyramid method proposed in Chapter 3. Then a discriminative BoW histogram is generated in each cell using the DGMM-based method in this chapter, which is further weighted by the cell saliency score. A signature weight is calculated for each image. Finally, the BoW representation coupled with the signature weight for each image is forwarded to the fuzzy SVM for training and classification. The experimental results are reported in Figure 4.6. It can be seen that the combined method presents better performance than each method by itself. The reason why the combined method outperforms each individual method is that the combined method considers both the spatial layout information and the patch discriminative information in landmark recognition.

Figure 4.6 Experimental results of the combined methods from Chapters 3 and 4

4.8 Summary

This chapter presents a new discriminative BoW learning technique for mobile landmark recognition. The key contribution of this chapter is to develop an iterative DGMM patch discrimination learning and generate a continuous BoW histogram for each landmark image. The image representative capability is also evaluated so that FSVM can take this information into consideration to train a classifier for each category. Experimental results on the NTU50 Landmark database and Oxford dataset show that the proposed method can achieve good recognition performance in mobile landmark recognition.

Chapter 5 Bag-of-Phrase for Mobile Landmark Recognition

5.1 Introduction

In the previous chapters, two effective content analysis methods based on BoW are presented. The developed methods learn the landmark discriminative information and then integrate it with the BoW technique for mobile landmark recognition. The experimental results show the effectiveness of the proposed methods. However, an issue of BoW is that it ignores the co-occurrence relationship between neighboring visual words in an image. The frequently co-occurring words form a “visual phrase” that can be exploited for landmark recognition. Although some works that focus on visual phrase learning have appeared, they mainly construct a generalized phrase dictionary from all categories for recognition, which lacks descriptive capability for a specific category. Further, these works adopt hard assignment of numerous feature sets to a limited number of phrases, which causes some useful feature sets to be discarded, and yields information loss.

In view of this, this chapter proposes a new bag-of-phrase (BoP) approach for mobile landmark recognition based on category-dependent visual phrase learning. The candidate phrases defined as adjacent pairwise codewords are first generated for each category. The important candidates are then selected through a proposed discriminative visual phrase (DVP) selection approach to form the BoP dictionary. Finally, a soft encoding method is developed to quantize each image into a BoP histogram.

The remainder of this chapter is organized as follows. A literature review of the BoP related works for landmark recognition is presented in Section 5.2. The overview of the BoP framework is given in Section 5.3. The detailed techniques, including visual phrase

candidate generation, DVP selection, and soft BoP encoding are discussed in Sections 5.4 and 5.5. Experiment results and discussions are given in Section 5.6. Section 5.7 concludes the chapter with a summary of our findings.

5.2 Related Work

Recently, some works that learn a set of visual phrases for image representation have appeared in [80], [96]-[102]. The “visual phrase” is defined as a set of frequently co-occurring visual words within a neighborhood image region. A simple method to discover visual phrase containing two member words is proposed in [101], where the appearance number of each word is first determined, and the adjacent word pairs formed by frequent words are then determined to find the visual phrases. Some methods that mine visual phrases containing 2, 3, and 4 words are proposed in [96]-[98]. These works first use the K-nearest neighbor (KNN) method to shortlist some phrase candidates, and then adopt frequent itemset mining (FIM) [102] to discover important phrases. A visual phrase significance metric is proposed in [97], followed by pattern summarization to cluster the phrases that represent the same object to reduce the intraclass variation. A method to extract phrases from the support regions is proposed in [99]. An information bottleneck principle is then used to cluster those phrases into a “visual synset”. Instead of using the KNN method, the work [98] extracts the phrase candidates within each segmented region. After that, FIM is used to mine the important phrases. The segmentation-based method is not widely used as it is time-consuming and does not work well when the images contain complex and ambiguous objects that are hard to segment.

An observation from the works above is that the phrases are first extracted in a local neighborhood, followed by learning the semantic similarity for refinement. Alternatively, the works [80], [100] do not constrain the extracted phrases in a spatial neighborhood. They

consider extracting the phrases that are composed of semantically similar words and spatially neighboring words separately. In [80], visual words that represent the same semantic object or are spatially neighboring in an image are grouped into phrases respectively. A contextual BoW histogram is then generated by mapping the image feature sets to the generated phrases. A graph mining technique is used in [100] to mine the semantically similar words which often co-occur together in an image.

In summary, the visual phrase methods above are learned from various object categories and their objective is to maximize the generalization ability of the phrase dictionary across different categories, which obtain good performance for image recognition. However, a shortcoming of these works is that their phrase dictionary lacks descriptive capability for a specific object class and thus cannot distinguish images from different categories well. Although some efforts have been made to learn the descriptive visual phrases for each category such as [101], they are insufficient since they only consider the frequency information to select the discriminative phrases, which ignores three factors: (1) the negative impact of the unimportant phrase candidates in the background areas, (2) the pairwise word semantic similarity in a phrase, which refers to the similarity of two words in representing the same semantic object, and (3) the pairwise phrase semantic similarity (relevance), which refers to the similarity of two visually different phrases in representing the same semantic object. Further, they all use the traditional hard quantization method to generate the phrase histogram. The hard assignment causes some useful feature sets to be discarded and generates a sparse histogram, since some phrases may not find sufficient matches within the feature sets.

In view of this, this chapter proposes a new bag-of-phrase (BoP) approach for mobile landmark recognition based on category-dependent phrase selection. In this work, the number of visual words in each phrase is chosen as two, which is named as second-order

phrase. Two contributions are made in the proposed approach: (i) a discriminative selection approach that takes advantage of the word-level and phrase-level semantic similarity is developed to select the important phrases from a large number of candidates, and form the descriptive BoP dictionary, (ii) a soft encoding technique is developed to generate a BoP histogram for each image, which reduces the amount of information loss induced by conventional BoP quantization.

5.3 Framework Overview

The overview of the proposed BoP method is shown in Figure 5.1. It consists of three steps: visual phrase candidate generation, discriminative phrase selection and soft BoP histogram generation. In the first step, visual features (e.g., SIFT [27]) are first extracted from the images and clustered to generate a word codebook. For each detected keypoint (Harris corner detectors [103]), its K nearest neighbors (KNN) [104] are determined in a local neighborhood. The center keypoint and each of its K neighbors (K is set as 3 in this work) form a second-order feature set, whose member features are then mapped to the codewords to create a second-order phrase. The randomly generated noisy phrases are removed through a statistical hypothesis testing in [97], which compares the likelihood ratio of the phrase generated by a particular hidden pattern versus random pattern to yield a set of preliminary phrase candidates.

In the second step, a DVP selection approach is proposed to create the BoP dictionary. The method develops a latent-space discriminative divergence metric to calculate the pairwise words distance. The distance (divergence) between pairwise words refers to the similarity of the two words in representing the same semantic category (object). Larger distance between two words indicates that they share less similarity in representing the same object. This word-level divergence metric is then extended to phrase-level metric.

Finally, a PhraseRank technique that employs the word-level and phrase-level semantic metric is developed to select the important phrases to form the BoP dictionary.

In the third step, a soft BoP histogram is generated to represent each image using the developed encoding method. It models the soft assignment of feature sets to visual phrases, which reduces the information loss in conventional quantization. The soft BoP histograms are used to train an SVM classifier.

Figure 5.1 Overview of the proposed BoP framework

5.4 DVP Selection

In this section, we give a detailed description on the DVP selection for each landmark category. As illustrated in Figure 5.2(a), a spatial circle neighborhood is first determined around each detected keypoint in the training image. The radius of the neighborhood is chosen to be four times the scale of the center keypoint, which is the same as in [101], to achieve a good trade-off between computational efficiency and recognition performance.

Secondly, the center and its neighboring keypoints determined by *KNN* in the neighborhood form a second-order feature set g as shown in Figure 5.2(b). These feature sets are then mapped to the word codebook $\Omega = \{w_1, \dots, w_M\}$ where M is the codebook size, to create the second-order phrase candidates as shown in Figure 5.2(c). The repeated candidates such as (w_4, w_6) and (w_6, w_4) due to different permutation are merged into one phrase as in Figure 5.2(d). The method in [97] is used to solve the problem of over-counting frequent feature sets. The neighborhood is used to constrain the words in each phrase candidate so that they are spatially close to each other. After that, a certain number of visual phrase candidates for each category are generated.

Figure 5.2 (a) Training image, (b) Second-order feature set, (c) Second-order phrase after feature mapping, (d) Second-order phrase after duplicate removal

After generating the preliminary phrase candidates for each category, a DVP selection method is proposed to select the important candidates to form the BoP dictionary. There are two reasons to perform discriminative phrase selection. One is that the number of the phrase candidates from all categories is large. For instance, there exist 300^2 possible duplet phrases for a 300-sized BoW codebook. This poses a heavy load to mobile landmark recognition which has limited processing power and fast response-time requirement. Another is that some unimportant phrases such as the background phrases can have negative impact on

recognition, and thus should be removed. Considering this, a discriminative visual phrase (DVP) selection approach is proposed to retain the phrases that are discriminative in distinguishing its category from other categories, and to remove those that are common across all categories.

Figure 5.3 Visual phrases detected in the foregrounds and backgrounds areas for the same landmark category under different capturing conditions

Conventional methods, such as [101], evaluate the discriminative capability of a phrase based on its occurrence frequency in different categories. This is insufficient for landmark recognition applications. For instance, when some background phrases have large occurrence frequencies, they have high chances to be selected as DVP. On the other hand, when some foreground phrases have low occurrence frequency, they are likely to be removed. Take Figure 5.3 as an example, it is seen that the background phrase (e, f) occurs 3 times in the images of the landmark, while the foreground phrases (x, y) , (m, n) and (c, d) that represent the same landmark area each occur only once, due to the changes of scales and perspectives. Therefore the phrase (e, f) may be selected as DVP, while (x, y) , (m, n) and (c, d) may be removed from the DVP candidates.

In order to resolve this problem, a new evaluation criterion is proposed to measure the semantic similarity (or conversely distance) between pairwise phrases, which can enhance the importance of the foreground phrases with low occurrence frequencies. The semantic similarity is defined as the similarity degree between two different phrases in representing

the same object, where larger similarity value indicates that the two phrases are more likely to represent the same object, and vice versa. Using this criterion, the foreground phrases (x, y) , (m, n) and (c, d) that represent the same object in Figure 5.3 will be emphasized since they have large pairwise semantic similarity value.

In the following sections, we first propose a word-level semantic metric by developing discriminative symmetric Kullback-Leibler (KL) divergence based on Probabilistic Latent Semantic Analysis (PLSA). This word-level metric is then extended to phrase-level metric, after which a visual phrase ranking algorithm that employs the word-level and phrase-level semantic metric is developed to select the DVPs for each category. These DVPs are then combined together to form the BoP dictionary.

5.4.1 Word-level Semantic Similarity Metric

Conventional methods mainly measure the semantic distance between pairwise words based on information gain criterion [105]. Specifically, the KL divergence [106] between distributions of categories induced by words is widely adopted for its effectiveness. The category distribution conditioned on the words can be interpreted as how much the word votes for each of the categories whenever it occurs. The work [80] utilizes such distribution and further improves the KL divergence to be symmetric. However, one condition for this method to work well is that the images from different landmark categories should not contain similar semantic concepts, which in fact is hard to guarantee. This is illustrated in Figure 5.4. Figure 5.4(a) illustrates some sample images from different landmark categories that are composed of various buildings. Figure 5.4(b) illustrates the category distributions induced by three different words based on the experiments, where the horizontal axis denotes the landmark category, and the vertical axis denotes the occurrence probability of each word across all the categories.

Figure 5.4 (a) Sample images from 6 landmark categories, (b) Similar category distribution induced by words 1, 2 and 3, (c) different latent topic distribution induced by words 1, 2 and 3.

As observed from the experiments, the local features clustered to the word 1 are mainly from the window area, and thus word 1 is claimed to be related with the semantic concept “window”. Similarly, word 2 and 3 are found to be related with the concept “greenery” and “ground”, respectively. From Figure 5.4(b), it is seen that the occurrence probabilities of the words 1, 2 and 3 in the six categories are similar. In such case, the three different words in the original domain are difficult to be distinguished, since they have no obvious distribution differences across various categories. This can be attributed to the fact that the images of the six categories contain similar semantic concepts: “window”, “ground” and “greenery”.

To overcome the above-mentioned problem, we propose to model a latent topic distribution induced by the word for pairwise word KL divergence measurement. The

distribution $P(z|w)$ can be interpreted as how much the word w votes for each of the latent topics z . The core intuition behind this measurement is that visual words that represent the same semantic object have more similar distributions over the latent topics rather than over the categories. The reason is that the latent topic is often closely related with a specific semantic concept (e.g., the “window” concept in Figure 5.4) while a category may contain multiple semantic concepts (e.g., each image in Figure 5.4(a) contains concepts “window”, “greenness” and “ground”). In this work, we use PLSA to infer the latent topics. PLSA is a statistical technique for the analysis of co-occurrence data, which has applications in information retrieval and machine learning [108]. It is based on a mixture decomposition derived from a latent class model. The latent class serves as a bridge between the images and the words. For initialization, we randomly select 20% of the training images from each category to form the dataset D , and represent each image by a BoW histogram. The following notations are defined for later use: $P(d)$ denotes the probability of observing a particular image d in the dataset, $P(w|z)$ denotes the conditional probability of a specific word w given the unobserved latent topic z , and $P(z|d)$ denotes an image specific probability distribution over the latent variable space. Based on these definitions, the PLSA model can be generated using three steps: (i) Choose an image d with probability $P(d)$, (ii) pick a latent topic z with probability $P(z|d)$, and (iii) generate a word w with probability $P(w|z)$. By processing all images in this way, we obtain a co-occurrence matrix $n(w, d)$. The joint distribution of (w, d) can then be formulated as:

$$P(w, d) = P(d)P(w|d) = P(d) \sum_{z \in Z} P(w|z)P(z|d) = \sum_{z \in Z} P(w|z)P(z)P(d|z) \quad (5.1)$$

where Z is all the topics in latent semantic space. Following the likelihood principle, the parameters $P(z)$, $P(w|z)$ and $P(d|z)$ can be determined by maximizing the following log-likelihood function:

$$\begin{aligned}
 L = \log P(D, W) &= \sum_{d \in D} \sum_{w \in W} n(w, d) \log P(w, d) \\
 s.t. \sum_{z \in Z} P(z) &= 1, \sum_{w \in W} P(w | z) = 1, \sum_{d \in D} P(d | z) = 1
 \end{aligned} \tag{5.2}$$

The model is fitted using the Expectation Maximization (EM) algorithm, as described in [108]. After obtaining $P(z)$, $P(w | z)$ and $P(d | z)$, Bayesian estimation is first utilized to infer the latent topic distribution induced by individual word $P(z | w)$ and the word occurrence probability $P(w)$ as follows,

$$P(w) = \sum_{z \in Z} P(w | z) P(z) \tag{5.3}$$

$$P(z | w) = \frac{P(z, w)}{P(w)} = \frac{P(w | z) P(z)}{\sum_{z \in Z} P(w | z) P(z)} \tag{5.4}$$

Figure 5.4(c) illustrates the generated latent topic distributions induced by three words based on (5.4), which shows that the latent topic distributions for 3 semantically different words have obvious differences, and can be easily distinguished.

Secondly, the discriminative capability of the inferred latent topics z is estimated and further incorporated into the KL divergence metric. For the topics that are closely related with the landmark area in an image, they are more representative for the landmark image and should be emphasized, and vice versa. Here we use the conditional entropy H of training image set D given the latent topic to determine the discriminative capability of each topic. This can be formulated as,

$$\begin{aligned}
 H(D | z \in Z) &= P(z) \sum_{d \in D} P(d | z) \log\left(\frac{1}{P(d | z)}\right) \\
 &= -P(z) \sum_{d \in D} P(d | z) \log(P(d | z)) \\
 &= -\sum_{d \in D} P(d, z) \log(P(d | z))
 \end{aligned} \tag{5.5}$$

From (5.5) it is seen that the conditional entropy H qualifies the remaining uncertainty of D given a latent topic z . In other words, larger H means that z cannot give a predictable result

for D and thus has less discriminative capability. Based on this, a Gaussian function is used to normalize the value of $H(D|z)$ within $[0, 1]$ to produce the discrimination weight $\omega(z)$ as follows,

$$\omega(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}H^2(D|z)} \quad (5.6)$$

The weight $\omega(z)$ reflects the discriminative capability of the latent topic z . After obtaining $\omega(z)$ and the latent topic distribution $P(z|w)$ conditioned on a word, a discriminative KL divergence metric is developed to calculate the distance between two latent topic distributions $P(z|w_i)$, $P(z|w_j)$. This distance measures the semantic similarity between two words w_i, w_j in latent space, and is defined as,

$$\begin{aligned} d(w_i, w_j) &= KL(P(z|w_i) \| P(z|w_j)) \\ &= \sum_{z \in Z} \omega(z) P(z|w_i) \log \frac{P(z|w_i)}{P(z|w_j)} \end{aligned} \quad (5.7)$$

where $d(w_i, w_j)$ is the calculated semantic distance between two words. From (5.7) it is seen that the developed KL divergence $KL(\cdot \| \cdot)$ considers the discriminative capability of each latent topic $\omega(z)$, and uses the weighted summation of all the latent topics as the word-level distance. However, an issue of this metric is that it is asymmetric. To ensure that it is symmetric, we further enhance it by considering the KL divergence of $P(z|w_j)$ with respect to $P(z|w_i)$. By integrating (5.3)-(5.7), the final word-level semantic distance d in latent space is developed as,

$$d(w_i, w_j) = \frac{P(w_i)}{P(w_i) + P(w_j)} \cdot KL(P(z|w_i) \| P(z|w_j)) + \frac{P(w_j)}{P(w_i) + P(w_j)} \cdot KL(P(z|w_j) \| P(z|w_i)) \quad (5.8)$$

where $KL(\cdot \| \cdot)$ is defined the same as (5.7).

5.4.2 Phrase-level Semantic Similarity Metric

As demonstrated in Figure 5.3, the same landmark object may be represented by different phrases due to different image capturing conditions. This causes some foreground phrases to have low occurrence frequency and be removed from the DVP candidates. To overcome this problem, a phrase-level semantic similarity metric is developed in this section. Using this metric, the less frequently occurred foreground phrases that have more semantic similarity with other phrases will be emphasized and selected as DVPs, since they are considered to be representative for their category.

Based on the principle in developing the pairwise words semantic distance metric, we extend the developed discriminative KL divergence from word-level to phrase-level by replacing the word w with phrase p in (5.8). Then the semantic distance between pairwise phrases p_i and p_j is formulated as,

$$d(p_i, p_j) = \frac{P(p_i)}{P(p_i) + P(p_j)} \cdot KL(P(z | p_i) \| P(z | p_j)) + \frac{P(p_j)}{P(p_i) + P(p_j)} \cdot KL(P(z | p_j) \| P(z | p_i)) \quad (5.9)$$

where $KL(\cdot \| \cdot)$ is defined the same as in (5.7), $P(p_i)$ is the occurrence probability of phrase p_i over all latent topics, $P(z | p_i)$ is the latent topic distribution induced by the phrase p_i . Similar to (5.3)-(5.4), these distributions can be derived by the Bayesian estimation as,

$$P(p_i | z) = \sum_{d \in D} P(p_i | d) P(d | z) \quad (5.10)$$

$$P(p_i) = \sum_{z \in Z} P(p_i | z) P(z) = \sum_{z \in Z} \sum_{d \in D} P(p_i | d) P(d | z) P(z) \quad (5.11)$$

$$P(z | p_i) = \frac{P(p_i | z) P(z)}{\sum_{z \in Z} P(p_i | z) P(z)} = \frac{P(z) \sum_{d \in D} P(p_i | d) P(d | z)}{\sum_{z \in Z} \sum_{d \in D} P(p_i | d) P(d | z) P(z)} \quad (5.12)$$

where $P(p_i | d)$ is the occurrence probability of phrase p_i in the image d which can be observed. After generating $d(w_i, w_j)$, $d(p_i, p_j)$ in (5.8) and (5.9), a Gaussian function similar

to (5.6) is used to normalize their value to be within 0 and 1, to give the final word-level and phrase-level semantic similarity scores that are denoted as $s(w_i, w_j)$, $s(p_i, p_j)$, respectively,

$$\begin{aligned} s(w_i, w_j) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}d^2(w_i, w_j)} \\ s(p_i, p_j) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}d^2(p_i, p_j)} \end{aligned} \quad (5.13)$$

5.4.3 Visual Phrase Ranking

In this section, the DVP discovery can be formalized as a visual phrase ranking problem which leverages the idea of word ranking in [101]. The ranking algorithm iteratively estimates the phrase's rank order by multiplying it with a matrix. The highly ranked phrases from the candidates are selected as the DVP. In the PhraseRank algorithm, two factors are considered when determining whether a phrase $p_i = (w_{ia}, w_{ib})$ is discriminative for its category: (i) the intraphrase semantic similarity, which refers to the pairwise words semantic similarity $s(w_{ia}, w_{ib})$ in a phrase. Larger $s(w_{ia}, w_{ib})$ means that the two words in the phrase have more semantic similarity, and are likely to represent the same semantic object. This phrase is thus considered to have more inherent semantic consistency. (ii) the interphrase semantic similarity, which refers to the semantic similarity $s(p_i, p_j)$ between the phrase p_i and other phrases p_j in its category. Larger similarity between the phrase and other phrases means that this phrase has more representative capability in its category. Based on this, a matrix $\mathbf{R}$ is built as follows,

$$\begin{aligned} R(i, i) &= \rho(p_i) \cdot s(w_{ia}, w_{ib}) \\ R(i, j) &= \rho(p_i) \cdot s(p_i, p_j) \end{aligned} \quad (5.14)$$

where $\rho(p_i) = P(p_i) / \bar{P}(p_i)$ is a weighting factor, and denotes the phrase occurrence frequency ratio between the positive category ($P(p_i)$) and negative categories ($\bar{P}(p_i)$),

which is similar to [101]. This is based on the fact that a phrase that occurs more frequently in the positive category while less frequently in the negative categories is usually more discriminative. From (5.14) it is seen that the diagonal element $R(i, i)$ records the weighted intraphrase semantic similarity, and the non-diagonal element $R(i, j)$ records the weighted interphrase semantic similarity. The diagonal and non-diagonal elements are respectively normalized to unit length by their summation. Algorithm 5.1 summarizes the proposed PhraseRank algorithm. From the Algorithm, it can be seen that the rank-updating iterations are performed by multiplying the matrix with the rank vector. During each iteration, the phrase candidates with more inherent semantic consistency (reflected by $R(i, i)$) and more similarity with highly ranked phrases (reflected through $R(i, j)$) will be emphasized. Finally the phrases with high ranking values in each category are selected as DVPs.

Algorithm 5.1 Visual phrase ranking algorithm

Input: matrix $\mathbf{R}$ for each category, maximum iteration number $l_{\max}$.

Initialize: the rank order for the candidate phrases in each category as equal: $rank_{old} = 1$.

While iteration number $loop < l_{\max}$

 Multiply $\mathbf{R}$ with previous rank vector to obtain new rank: $rank_{new} = \mathbf{R} \cdot rank_{old}$

if $\|rank_{new} - rank_{old}\| > \varepsilon$, where $\|\cdot\|$ is the norm-2 distance

$loop++$,

$rank_{old} \leftarrow rank_{new}$

else

break,

end

Output: the phrases with high ranking values in $rank$ as the DVPs.

5.5 Soft BoP Encoding

After generating the DVPs for each category, we combine them together to form the BoP dictionary to encode detected local feature sets in each image. Current BoP quantization methods [96]-[100] describe an image as a bag of discrete phrases selected from the BoP dictionary, where the frequency distribution of phrases in an image allows classification. A limitation of this hard quantization model is that some phrases in the BoP dictionary at times are not able to find exact matches with the feature sets in the image. As a result, these phrases render the generated BoP histogram sparse, and at the same time, the feature sets that cannot be mapped to the dictionary have to be discarded, which causes information loss. In order to overcome this problem, we propose a soft BoP encoding method. The principle is detailed as:

First for each element of second-order feature set $g=(f_1, f_2)$, two soft values $c_1(j)$ and $c_2(j)$, $j = 1, \dots, M$, for M codewords are generated using the soft assignment as follows,

$$c_i(j) = \frac{\exp(-\frac{1}{2}\|w_j, f_i\|)}{\sum_{j=1}^M \exp(-\frac{1}{2}\|w_j, f_i\|)}, \quad i = 1, 2; j = 1, \dots, M \quad (5.15)$$

where $\|w_j, f_i\|$ is the norm-2 distance between word w_j and feature f_i , $c_i(j)$ corresponds to the soft assignment value of $f_i \rightarrow w_j$, and the exponential function is used to smooth the data.

Then for each phrase $p=(w_m, w_n)$ in the BoP dictionary, a value $h(p)$ which denotes the occurrence number of phrase p in the image is generated by soft assignment of (f_1, f_2) to (w_m, w_n) ,

$$h(p) = \min_{\substack{f_1 \rightarrow w_m \\ f_2 \rightarrow w_n}} \{c_1(m), c_2(n)\}, \quad \text{where } p = (w_m, w_n) \quad (5.16)$$

where $c_1(m)$, $c_2(n)$ has been calculated by (5.15). From (5.16) it is seen that the smaller value of $c_1(m)$ and $c_2(n)$ is chosen to calculate $h(p)$. The reason to choose the smaller value is that it can be understood as the co-occurrence number of the two words in the image. However, an issue of (5.16) is that it only considers the match order of $f_1 \rightarrow w_m, f_2 \rightarrow w_n$, and ignores the match order of $f_1 \rightarrow w_n, f_2 \rightarrow w_m$. In view of this, we further enhance it by considering both match orders, and select the order that produces larger value to calculate $h(p)$,

$$h(p) = \max_r h_r(p), r = 1, 2; \quad (5.17)$$

where r is the match order, $r = 1$ corresponds to $f_1 \rightarrow w_m, f_2 \rightarrow w_n$, $r = 2$ corresponds to $f_1 \rightarrow w_n, f_2 \rightarrow w_m$. The larger value indicates that the feature set g finds a better match with the phrase p .

Finally, all the detected local feature sets in the image are mapped to the BoP dictionary using the developed soft encoding method and a BoP histogram $\mathbf{h}$ is generated.

As an observation, BoW can be regarded as reflecting the coarse spatial distribution of a single word in the image, and BoP can be regarded as reflecting the fine spatial distribution of two co-occurring words in the image. In order to exploit the respective advantages of BoW and BoP, we combine the BoW and BoP histograms together for recognition. The pyramid matching method in [47] is adopted. The χ^2 function is used as the matching kernel due to its good performance for histogram matching. The new pyramid matching kernel K for the two images X and Y is defined as:

$$K(X, Y) = \alpha K_{BoW}(X, Y) + (1 - \alpha) K_{BoP}(X, Y) \quad (5.18)$$

where K_{BoW} and K_{BoP} are the chi-square kernel distance calculated for BoW and BoP histograms respectively, α is used to balance the contributions of K_{BoW} and K_{BoP} , and is set

as 0.25 by cross validation. The chi-square kernel is well-known and can be found in different references such as [57].

5.6 Experimental Results

In this section, several experiments are conducted to demonstrate the performance of the proposed BoP approach for mobile landmark recognition. For the database, the NTU50landmark database and Oxford building dataset as described in previous chapters are used for experimental evaluation. For PLSA, the number of latent topics is chosen as equal to the number of landmark categories in the dataset.

5.6.1 Comparison of BoW, BoP and their Combination

We investigate the performance of BoW, proposed BoP and their combination for recognition experiments. The sizes of the BoW codebook and BoP dictionary vary from hundreds to thousands. The results are shown in Table 5.1, from which it is seen that the performance of BoW and BoP combination for landmark recognition is better than either BoW or BoP alone for all codebook sizes. Specifically, a highest recognition rate of 95.3% is obtained when using the combination of 1200-sized BoW codebook and 1800-sized BoP dictionary. This is because BoW and BoP can complement each other by describing not only the image's single word distribution but also pairwise words co-occurrence distribution. Furthermore, it is noted that the proposed BoP method consistently outperforms the conventional BoW method for various codebook/dictionary sizes. The reason is that phrases have better descriptive capability than words in characterizing the features' spatial distribution in an image.

Table 5.1 Performance comparison of BoW, proposed BoP, and their combination

Size of BoW codebook/ BoP dictionary	BoW	Proposed BoP	Combined BoW and BoP
300/300	75.1	82.6	84.8
300/600	75.1	86.3	88.8
300/900	75.1	89.9	91.6
800/900	82.4	89.9	91.9
800/1200	82.4	91.6	93.1
1200/1500	85.4	92.1	93.8
1200/1800	85.4	93.1	95.3
1500/2100	85.4	93.4	95.3

5.6.2 Comparison with Other Phrase Selection Methods

In this section, we compare the proposed DVP selection methods as described in Section 5.4 with several other phrase selection methods in [80], [100], [101] for landmark recognition. To achieve a fair comparison, we use the conventional hard BoP quantization as in [100] to generate the phrase histogram. The size of the phrase dictionary is set as 1800 which has been shown to obtain good performance as in Table 5.1. The comparison results are reported in Table 5.2, which show that the proposed DVP selection method obtains the best performance of 89.3% and outperforms the methods in [80], [100], [101] by 2.2%, 7.7% and 3.3% respectively.

Table 5.2 Performance comparison of proposed DVP selection method with other methods

Method	Contextual BoW method in [80]	Graph-based method in [100]	Occurrence frequency-based phrase method in [101]	Proposed DVP method
Recognition accuracy (%)	87.1	81.6	86.0	89.3

Further, we compare the developed symmetric KL divergence metric with other commonly used distance metrics such as the Euclidean distance, KL divergence [106], symmetric KL (SKL) divergence [80] and Earth Mover's Distance (EMD) [109] for landmark recognition. We use the proposed DVP selection method and hard BoP

quantization for experiments. The results are shown in Table 5.3, and it is seen that the developed discriminative symmetric KL divergence metric obtains the best performance among all the distance metrics. This can be attributed to the fact that the developed metric calculates the KL divergence between pairwise words/phrases in latent semantic space through the PLSA, while the other metrics such as [80], [109] are based on the BoW (word-image) space. The latent space is more concrete in representing the semantic object than the BoW space. Further, the developed metric considers the discriminative capability of various latent topics when calculating the divergence, while the other works do not consider this.

Table 5.3 Performance comparison of various distance metric methods

Various distance metrics	Euclidean	KL divergence [106]	Symmetric KL divergence [80]	EMD [109]	Developed metric
Recognition accuracy (%)	84.8	86.5	88.4	85.6	89.3

5.6.3 Comparison with Hard BoP Quantization

The performance comparison between the proposed soft BoP encoding and conventional hard BoP quantization methods are performed under different sizes of phrase dictionary. The proposed DVP selection method is used for dictionary generation. The results are shown in Figure 5.5. From the figure it can be seen that the developed soft BoP encoding method consistently outperforms the conventional hard BoP method for all dictionary sizes. The rationale is that the soft BoP encoding method can produce less quantization errors by soft assignment of local feature sets to visual phrases, while the hard BoP quantization method causes quantization error between the numerous feature sets and limited number of phrases. Further, the performance gap between the two methods decreases when the dictionary size increases. This is because the quantization errors of local feature sets to visual phrases for the conventional method are greatly reduced when more phrases are selected into the BoP dictionary.

Figure 5.5 Performance comparison of developed soft BoP encoding method versus hard BoP method

5.6.4 Comparison with Other BoP Methods

In this section we conduct a group of experiments to compare the overall proposed BoP approach with the standard BoW in [45] and two other BoP methods in [80], [101], under different dictionary sizes. The developed soft BoP encoding method is adopted to generate the phrase histogram for all these works to achieve a fair comparison. The results are shown in Figure 5.6. From the figure it is seen that the proposed BoP method generates the best performance for all dictionary sizes, followed by the two visual phrase methods [80], [101] with inferior performance, and lastly the baseline BoW in [45]. The reason that the BoP methods outperform the BoW method is that the phrase in the BoP dictionary is composed of spatially neighboring words, and it has stronger descriptive capability for the image than each individual word. Furthermore, the superiority of the proposed BoP method over the other two visual phrase methods shows the effectiveness of the proposed DVP selection algorithm and soft BoP encoding technique.

Figure 5.6 Performance comparison of proposed BoP method versus other BoW and BoP methods

5.6.5 Evaluation of Content-Context Integration

We conduct some experiments in this section by integrating the context information provided by the mobile phone with the proposed content analysis for landmark recognition. The content-context integration approach in Chapter 3 is adopted for recognition. Table 5.4 gives the experiment results. From the table it can be seen that methods integrating content and context analysis obtain better performance than method using content analysis alone.

Table 5.4 Experimental results of content-context integration

Method	Recognition rate (%)	Average Recognition time (s)
Proposed content analysis	95.3	8.5
Context: location	96.5	5.1
Content: proposed content analysis		
Context: location and direction	98.3	4.7
Content: proposed content analysis		

5.6.6 Performance Evaluation on the Oxford Database

In this experiment, the proposed method is applied and compared with other methods [80][101] on the Oxford dataset. For the proposed method, the soft BoP is used as the image representation. For other methods, the contextual BoW in [80] and the descriptive phrase in [101] are used for image representation, respectively. All the methods adopt the Chi-square distance kernel to train the SVM. The results in Table 5.5 show that the proposed method achieves good accuracy of 89.1%, which is 5.5%, and 7.3% higher than the method in [80] and [101] respectively.

Table 5.5 Experimental results on the Oxford building dataset

Method	Recognition accuracy (%)
Li's method [80]	83.6
Zhang's method [101]	81.8
Proposed method	89.1

5.6.7 Evaluation of The Combined Methods from Chapters 3, 4 and 5

In this section, the BoW methods proposed in Chapters 3-4 are combined with the BoP method in this chapter for landmark recognition. Firstly, a landmark image is partitioned into multiple pyramid cells and their saliency scores are computed using the method in Chapter 3. Then for each pyramid cell, a discriminative BoW histogram and a soft BoP histogram are generated, using the DGMM method in Chapter 4 and the BoP method in this chapter, respectively. Finally, the BoW and BoP histograms weighted by the cell saliency score are combined to represent the image, and forwarded to the SVM for training and classification. The matching function in this chapter is used to calculate the pairwise image distance at each pyramid level. The spatial pyramid kernel in Chapter 3 is used to sum up the distances that are calculated at various pyramid levels. The experimental results using the combined BoW and BoP methods are reported in Figure 5.7. It can be seen that the combined method

outperforms the individual methods in Chapters 3-5. The reason is that the combined method considers both the word and phrase distribution information in the pyramid cells for image representation.

Figure 5.7 Experimental results of the combined method from Chapters 3-5

5.7 Chapter Summary

In this chapter, an effective BoP technique is developed for mobile landmark recognition based on discriminative visual phrase selection. The key contributions of this chapter include (i) The proposal of a DVP selection approach in latent semantic space, which exploits the PLSA and Bayesian analysis to derive a discriminative symmetric KL divergence metric for word-level distance computing. After extending it to phrase-level distance metric, a PhraseRank technique is developed for DVP selection. (ii) The development of a new soft BoP encoding method to reduce the information loss caused by the conventional BoP quantization method. Experimental results on the NTU50 Landmark database and Oxford dataset show that the proposed method can achieve good recognition performance in mobile landmark recognition.

Chapter 6 Context-Aware Vocabulary Learning for Mobile Landmark Recognition

6.1 Introduction

In previous chapters, several content analysis techniques based on BoW and BoP are developed to be integrated with the context analysis (location and direction information) for mobile landmark recognition. The experimental results on the NTU50Landmark and Oxford building datasets show the effectiveness of these methods. An issue for these methods is that they adopt SVM to train the classifier models for landmark recognition. The SVM is suitable to be used in conjunction with context information such as GPS to constrain the number of landmarks to a small number, so that the computational time is manageable. However, when the database is scaled up and the context information of some images may be unavailable, SVM will have long recognition time since it does not have good scalability.

In order to better address the scalability issue and speed up the landmark recognition process, a new direction-dependent scalable vocabulary tree (DSVT)-based mobile landmark recognition framework, which exploits both the context (direction and location information) and content analysis for vocabulary learning is proposed in this chapter. The SVT has good scalability for image search in large-scale databases, and can achieve fast recognition time. The vocabulary learning generates a compact set of discriminative codewords for image representation, which reduces the dimension of the image's BoW histogram while still retaining its discriminative power. In this chapter, we will focus on learning a discriminative compact vocabulary (DCV) based on the SVT. Many state-of-the-art methods use content analysis alone for vocabulary learning, which underutilize the

context information provided by mobile devices such as the location and direction data. Although some works start to consider the images' location information for vocabulary learning, the location alone is insufficient since GPS data has significant errors in dense built-up areas. The context analysis techniques that use GPS to shortlist the geographically nearby landmark candidates for subsequent image matching are at times inadequate. In view of this, this chapter proposes to employ both direction and location information to learn a DCV for mobile landmark recognition. Direction information is first considered to supervise image feature clustering to construct DSVTs. Location information is then incorporated into the proposed DCV learning algorithm to select the discriminative codewords of the DSVT to form the DCV. An ImageRank technique and an iterative codeword selection algorithm are developed for DCV learning.

The rest of the chapter is organized as follows. An introduction to the related work is given in Section 6.2. A general overview of the proposed DCV mobile landmark recognition system is presented in Section 6.3. The context-aware DCV learning approach is discussed in detail in Section 6.4, which includes the DSVT construction, the ImageRank technique, and the iterative codeword selection algorithm. A quantitative evaluation that includes the experimental results and discussions is given in Sections 6.5-6.7. Section 6.8 concludes this chapter.

6.2 Related Work

Recently, a lot of works have shown that learning a discriminative vocabulary to encode the query image is effective in mobile landmark recognition, e.g., [59], [64]. In fact, various vocabulary (codebook) learning works have appeared recently in [57], [60], [101], [118]-[121]. A category-related codeword discrimination learning method based on the word occurrence ratio of positive to negative categories is proposed in [57] and [101]. In [60], the

discriminative capability of each codeword for each category is learned through an information gain method. In [118], a fast hierarchical merging algorithm is proposed to merge a large number of words into a small number of discriminative words, which can reduce the memory cost. In [119], a discriminative codebook learning algorithm is developed which iteratively feeds back the SVM classifier outputs to the codeword generation, until the final codebook achieves the strongest discriminative power. A multi-resolution codeword selection approach is proposed in [120], which tries to select the discriminative words from the multi-resolution local patches using a boosting process. In [121], a multi-output linear function is proposed to model the relationship between the codeword selected data matrix and the indicator matrix. The most discriminative words are defined as those leading to minimal fitting error.

The common purpose of these works is to construct a compact vocabulary by maximally preserving the separability of the object classes. This can improve the computational efficiency and reduce memory cost. However, an issue of these works is that their vocabulary learning methods are mainly based on the content analysis alone, and do not consider the context information (e.g., the GPS data) provided by the mobile device. Even though some works such as [64] have started to integrate the context with content analysis for vocabulary learning, the context information is still limited to GPS alone, which can have significant errors (up to 100 meters) in dense built-up areas, and this causes the context analysis using GPS alone to be unreliable at times. Furthermore, the GPS and content integration technique is preliminary, which simply uses the location information to shortlist the geographically nearby landmark candidates for subsequent image matching.

In view of this, this chapter proposes to exploit both location and direction information to learn a DCV for mobile landmark recognition. To the best of our knowledge, this is the first work to employ direction information for DCV learning. This is possible since many

smart phones are now equipped with digital compass, e.g., Blackberry 9220, Nokia Lumia 610, iPhone 4, etc. This renders the direction information readily available. In the near future it is believed that more mobile phones will have such function. Two contributions are made in this chapter. Firstly, DSVTs are constructed by using the direction information to supervise the image feature clustering process. The leaf nodes (codewords) of the DSVT are treated as the preliminary DCV candidates. Secondly, a context-aware DCV learning algorithm is developed to select the discriminative codewords of the DSVT to form the DCV. An ImageRank technique based on Latent Semantic Analysis (LSA) is developed to estimate the importance of each image. An iterative codeword selection algorithm that incorporates the image weights is then proposed to generate the DCV, which is used to encode the query image for landmark recognition.

6.3 Framework Overview

Due to the unique characteristics of mobile devices as discussed before, a client-server architecture is adopted in this work. Through incorporating the learned DCV into the structure, an effective mobile landmark recognition prototype is proposed in Figure 6.1. It is seen that at the client end, a mobile user first captures a landmark image that is tagged with both location and direction data. Local features (e.g., SIFT) are then extracted from the image and quantized into a BoW histogram using the DSVT that is stored in the mobile device beforehand. This high dimensional (the leaf nodes in the DSVT usually range from several thousands to hundreds of thousands) BoW histogram is compressed into a compact vector by retaining only the codewords that belong to the DCV. The compact vector is finally encoded using arithmetic or other coding techniques and transmitted to the server over the wireless network (e.g., WiFi, 3G, etc.). Compared with the original BoW histogram, the compact vector has a smaller memory size and lower bit rate and can be transmitted

over a narrow bandwidth network. This is important for mobile landmark recognition which is often hindered by poor wireless network connections. At the backend server, the received packet is first decoded into the compact vector. The database images that are indexed using inverted files are then scored according to their histogram similarity with the query image. The inverted file system is constructed similar to [58], which can increase the recognition speed significantly.

Figure 6.1 The proposed mobile landmark recognition architecture

6.4 Context-aware Vocabulary Learning

In this section, we will give a detailed description of the context-aware DCV learning, which plays a key role in the proposed landmark recognition system. As shown in Figure 6.2, the proposed DCV learning approach mainly consists of three steps, which are described as follows:

Figure 6.2 An overview of the proposed context-aware DCV learning process

(i) **DSVT construction**, which first extracts local features and context data from the training images, and then utilizes the direction information to supervise the clustering of local features into a DSVT. The leaf nodes (codewords) of the DSVT are treated as the DCV candidates and used to quantize each image into a BoW histogram.

(ii) **ImageRank analysis**, which ranks the images' importance according to their discriminative capability. The reason for this step is based on the observation that, discriminative images usually have a higher potential to contain the desired important codewords than other images, and thus should be given more emphasis in learning the DCV. In the developed ImageRank technique, a term-document matrix is constructed using the

BoW histograms of the training images. SVD is then performed on these histograms for latent semantic analysis, which can speed up the process.

(iii) **Codeword selection**, which first develops a vocabulary discriminative metric based on image intra-class and inter-class similarity, and then proposes an iterative codeword selection algorithm based on this metric to form the DCV. The algorithm iteratively selects the most discriminative codeword from the original codebook into the DCV, until the termination criteria or a predefined DCV size is met.

6.4.1 DSVT Construction

Recently, the SVT model [112] has been used in different landmark recognition works [58], [64], [116], due to its effectiveness in handling scalable visual search. However, an issue of these works is that they do not consider any context information when constructing the SVT. Therefore in this work, direction information is considered to supervise the image feature clustering to construct the DSVT. The rationale is that the consistency of the codewords for a landmark will depend on the capturing angles of the landmark, and direction data from smartphones is generally accurate and reliable. The DSVT is constructed as follows. First, as illustrated in Figure 6.3(a), the training images are divided into various clusters according to their associated tagged direction data, e.g., images with directions in 0-30 degrees form a cluster, 30-60 degrees form the next cluster, and so on. This process is repeated until all the training images are divided into their nearest clusters. Totally there will be $\frac{360^\circ}{30^\circ} = 12$ direction clusters. The reason for adopting 30 degrees as the cluster range is that the image descriptors used in this work are relatively robust to a perspective change of 30 degrees, e.g., SIFT or SURF. Then, hierarchical k -means algorithm is performed on the local features extracted from the images in each direction cluster to construct an SVT as in Figure 6.3(b), after which a set of leaf nodes are generated and used as the DCV candidates. When a query

image is given, its local features are first extracted and quantized into a histogram through the DSVT. The histogram indicates the codewords visit number on the tree nodes. This histogram is then compared with the histograms of the database images to score them. The local features are taken as visual bag-of-words in the DSVT matching. For an H -depth, B -branch SVT, it will produce B^H leaf nodes (codewords) [64].

Figure 6.3 (a) Training images that are divided into various direction clusters, (b) DSVT in each cluster.

There are two benefits when using direction to supervise image clustering for DSVT construction: (i) DSVT generated codewords have more discriminative capability in differentiating the landmark images in the same cluster. The rationale is that the same landmark's images in the cluster are often visually closer to each other due to similar viewpoints. As a result, the generated codewords from these images are also visually consistent and have better discriminative capability, (ii) the search space in the database is greatly reduced when performing similarity matching for the query image. Unlike previous works such as [62], [64] that use GPS alone for image filtering, the direction and location information are combined to select the landmark candidates in this work, which can provide better results than using location alone. The database images in the nearest direction cluster

with the query are first determined, followed by GPS to select the geographically nearby landmarks for subsequent image matching.

However, these DSVT derived codewords cannot be directly used as the final vocabulary to encode the query image, since the total number of codewords in each direction cluster may be large, and this may cause the dimension of the BoW histogram to be high. This is unsuitable for mobile recognition since the mobile network usually has limited bandwidth. Furthermore, some words that are commonly observed in many landmarks may have weak discriminative capability in the direction cluster. Considering this, we designed a vocabulary learning method to select those discriminative codewords to form the DCV.

6.4.2 ImageRank Analysis

As mentioned earlier, ImageRank analysis is performed to emphasize the discriminative images since they have greater potential to contain words that belong to the DCV, and vice versa. The flowchart of the proposed ImageRank algorithm is shown in Figure 6.2. It can be seen that a term-document (in our case it is word-image) matrix is first constructed using the BoW histograms of the training images. LSA is then performed on the matrix for image discriminative analysis. The reason for adopting LSA is that: (i) the original BoW histogram is high dimensional, and this causes image analysis on the BoW level to have high computational cost. By transforming the BoW histogram to latent topic representation through selecting the topics with large singular values, the dimension of the histogram can be greatly reduced. The singular values in fact reflect the topics' importance in discriminating various landmark images. (ii) Landmark images usually consist of basic semantic units, e.g., windows, glasses, walls, and doors, etc. Therefore, it is reasonable to assume that some semantic topics exist among these images. The latent topic under the

word-image connection can be discovered via a semantic model. The LSA, which has been shown to be useful for exploring the relationships between a set of documents and the terms they contain by producing a set of latent concepts related to the documents and terms [122], can be adopted for our use. Based on LSA, suppose a word-image matrix $\mathbf{W}_{M \times N} = [\mathbf{w}_1 \ \dots \ \mathbf{w}_N]$ is given, where M and N denote the number of codewords and images respectively, each column $\mathbf{w}_n$ denotes the M -dimensional BoW histogram for the n -th image, and is normalized to unit length. The SVD can be used to decompose the matrix into the product of three matrices:

$$\begin{aligned} \mathbf{W}_{M \times N} &= \mathbf{U}_{M \times M} \mathbf{\Sigma}_{M \times N} \mathbf{V}_{N \times N}^T \\ \Rightarrow [\mathbf{w}_1 \ \dots \ \mathbf{w}_N] &= [\mathbf{u}_1 \ \dots \ \mathbf{u}_M] \cdot \begin{bmatrix} \sigma_{1,1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{N,N} \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0_{M,N} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_N^T \end{bmatrix} \end{aligned} \quad (6.1)$$

where $\mathbf{U}_{M \times M}$ can be understood as the orthogonal word-topic matrix of size $M \times M$, with each column $\mathbf{u}_m$ an M -dimensional vector, and $\mathbf{V}_{N \times N}$ can be understood as the orthogonal image-topic matrix of size $N \times N$, with each column $\mathbf{v}_n$ an N -dimensional vector. $\mathbf{\Sigma}_{M \times N}$ is an $M \times N$ rectangular diagonal matrix with all the diagonal elements σ larger than 0 and sorted in decreasing order. In order to maintain the real data structure of $\mathbf{W}$ while ignoring the unimportant details, only the top Z ($Z < M$, $Z < N$) largest diagonal elements (corresponding to the most important topics) in $\mathbf{\Sigma}$ are retained to generate a compact matrix $\bar{\mathbf{\Sigma}}$. In the meanwhile, for the matrix $\mathbf{U}$ and $\mathbf{V}$, only the columns corresponding to the selected latent topics are kept to produce $\bar{\mathbf{U}}$ and $\bar{\mathbf{V}}$, respectively. Based on this, equation (6.1) can be transformed as,

$$\begin{aligned}
 \overline{\mathbf{W}}_{M \times N} &= \begin{bmatrix} \overline{\mathbf{w}}_1 & \cdots & \overline{\mathbf{w}}_N \end{bmatrix} \\
 &= \begin{bmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_Z \end{bmatrix} \cdot \begin{bmatrix} \sigma_{1,1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_{Z,Z} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_Z^T \end{bmatrix} \\
 &= \overline{\mathbf{U}}_{M \times Z} \overline{\Sigma}_{Z \times Z} \overline{\mathbf{V}}_{N \times Z}^T
 \end{aligned} \tag{6.2}$$

The dimensionality reduction operation for a new image is then achieved through performing the following mathematical transforms on (6.2),

$$\begin{aligned}
 \overline{\mathbf{W}}^T &= \left(\overline{\mathbf{U}} \overline{\Sigma} \overline{\mathbf{V}}^T \right)^T \Rightarrow \overline{\mathbf{W}}^T = \overline{\mathbf{V}} \overline{\Sigma} \overline{\mathbf{U}}^T \\
 &\Rightarrow \overline{\mathbf{W}}^T \overline{\mathbf{U}} \overline{\Sigma}^{-1} = \overline{\mathbf{V}} \overline{\Sigma} \overline{\mathbf{U}}^T \overline{\mathbf{U}} \overline{\Sigma}^{-1} \\
 &\Rightarrow \overline{\mathbf{V}} = \overline{\mathbf{W}}^T \overline{\mathbf{U}} \overline{\Sigma}^{-1}
 \end{aligned} \tag{6.3}$$

Based on (6.3), given a query image's original BoW histogram $\mathbf{w}$, its latent topic representation $\mathbf{v}$ is formulated as,

$$\mathbf{v} = \mathbf{w}^T \overline{\mathbf{U}} \overline{\Sigma}^{-1} \tag{6.4}$$

In order to determine Z , we follow the method in [123] and find that best results are obtained when Z is equivalent to the top 30% of the total diagonal elements in Σ , which means that the dimension of the new image's latent topic representation is reduced to only 30% of the original BoW histogram. The computational cost thus can be greatly reduced. Next, in order to learn the discriminative capability of each image for its category, we take two factors into consideration. The first factor is the images' pairwise cosine similarity. The images having more pairwise similarity with other images of the same landmark are generally considered more representative. For a pair of images i and j , their cosine similarity can be calculated using latent topic representation, and denoted as $S(i, j)$,

$$S(i, j) = \|\mathbf{v}_i, \mathbf{v}_j\|_{\cos} = \frac{\mathbf{v}_i^T \mathbf{v}_j}{\|\mathbf{v}_i\| \cdot \|\mathbf{v}_j\|} \tag{6.5}$$

where $\|\mathbf{v}_i\|$ is the norm-2 value of $\mathbf{v}_i$.

The second factor is the images' within-class to between-class similarity ratio. The

image that has more similarity with its own category's images while less similarity with other categories' images is considered to be more representative for its category. The ratio $S(i, i)$ for an image i is formulated as,

$$S(i, i) = \frac{\sum_{j=1, j \neq i}^{N_t} \|\mathbf{v}_i, \mathbf{v}_j\|_{cos}}{\sum_{j=1}^{N_o} \|\mathbf{v}_i, \mathbf{v}_j\|_{cos} \mathbf{I}_R(\mathbf{v}_i, \mathbf{v}_j)} \quad (6.6)$$

where the numerator and denominator represent the image's within-class and between-class similarity respectively, N_t and N_o denote the number of the images in the target and negative landmark categories in the direction cluster, respectively, $\mathbf{I}_R(\cdot)$ is an indicator function used to select the geographically nearby images, which are within a radius of R from the target landmark for similarity analysis. R is set as 200m based on our field study. $\mathbf{I}_R(\cdot)$ is defined as:

$$\mathbf{I}_R(\mathbf{v}_i, \mathbf{v}_j) = \begin{cases} 1, & \text{if } \|\mathbf{v}_i, \mathbf{v}_j\|_{Geo} < R \\ 0, & \text{otherwise} \end{cases} \quad (6.7)$$

where $\|\mathbf{v}_i, \mathbf{v}_j\|_{Geo}$ is the geographical distance between the negative image j and the target landmark that the image i belongs to, which is calculated the same as in [124].

Finally, inspired by the VisualRank idea in [125], an ImageRank technique is developed here. The rationale is that the ImageRank method can assign different rank orders to different images according to their discriminative capability. The orders reflect the importance of different images in its category. The method works as follows. A matrix $\mathbf{S}$ is first built. The diagonal element records the within-class to between-class similarity ratio of each image, and can be represented by $S(i, i)$ as calculated in (6.6) and further normalized by their diagonal summation to unit length. The non-diagonal element records the similarity between pairwise images, and can be represented by $S(i, j)$ as calculated in (6.5) and further normalized by their columnwise summation to unit length. Iterations are then carried out by

multiplying this matrix with an initialized rank order of these images. During each iteration, the images with higher within-class to between-class similarity ratio and more similarity with other images will be highly ranked, while the images with lower within-class to between-class similarity ratio and less similarity with other images will be lowly ranked. Iterations will stop when the image orders converge or a maximum iteration number is reached. Inspired by [123], the image rank order $\mathbf{r}$ is iteratively estimated as:

$$\mathbf{r} = d\mathbf{S}\mathbf{r} + (1-d)\mathbf{p} \quad (6.8)$$

where $\mathbf{p}$ is an $N_t \times 1$ distracting vector for random walk behavior, with each element equivalent to $\frac{1}{N_t}$, similar to [125]. d is a constant damping factor, and is usually chosen to be equal to 0.8 as in [123]. $\mathbf{r}$ is an $N_t \times 1$ rank order vector for the N_t images in the target category. The ImageRank technique is summarized in Algorithm 6.1.

Algorithm 6.1 ImageRank algorithm

Input: A set of BoW histograms $\mathbf{w}_n$ for the training images of each landmark category in the direction cluster, SVD decomposed matrices: $\bar{\mathbf{U}}, \bar{\Sigma}$.

Output: Learned image rank order $\mathbf{r}$ for each category

- 1 Transform the original BoW histogram to latent topic-based representation according to (6.4).
- 2 Construct the matrix $\mathbf{S}$ using (6.5) and (6.6).
- 3 Learn the image rank order $\mathbf{r}$ iteratively as:

While $\{ \|\mathbf{r}^{(i)} - \mathbf{r}^{(i-1)}\| < \tau \text{ and } i < i_{\max} \}$ **do**

Update the rank order $\mathbf{r}$ using (6.8);

Iteration number $i++$;

End
- 4 Output the final learned image rank order $\mathbf{r}$.

Figure 6.4 illustrates several landmark images that are ranked according to their learned discrimination weights, in a decreasing order from left to right. From the figure it can be seen that the leftmost images present more discriminative information of the captured landmark, while the rightmost image present less discriminative information due to the clutters, occlusions, etc. This shows the effectiveness of the developed ImageRank method.

Figure 6.4 Ranked images using the proposed ImageRank method

6.4.3 DCV Learning

In this section we first formulate the DCV learning problem in the BoW space, and then solve it in latent semantic space. The DCV learning in fact is a codeword selection process, which selects the most discriminative codewords from the original codebook into the newly generated DCV. Based on this, we first create a binary matrix $\mathbf{B}$ of size $K \times M$ to record whether a word in the original codebook is selected, with each column $m = 1, 2, \dots, M$ (M is the original codebook size) denotes the word index in the original codebook, each row $k = 1, 2, \dots, K$ (K is the DCV size, $K < M$) denotes the selected word index in the generated DCV, and each item $B(k, m)$ denotes that whether the m -th word in the original codebook is selected (“1” indicates yes, “0” indicates no) and further mapped as the k -th word in the

new DCV. Using this matrix, an original high-dimensional BoW histogram $\mathbf{w}$ generated by the DSVT can be compressed into a low-dimensional compact vector $\mathbf{c}$ for transmission, which is calculated by,

$$\mathbf{c}_{K \times 1} = \mathbf{B}_{K \times M} \mathbf{w}_{M \times 1} \quad (6.9)$$

where each row of the matrix $\mathbf{B}$ must and only contain one non-zero element to select an item of $\mathbf{w}$ into $\mathbf{c}$.

From the above, it is seen that the DCV learning can be transformed into finding an optimal word selection matrix $\mathbf{B}$ to determine which codewords in the original codebook are selected into the DCV. In order to learn the optimal $\mathbf{B}$, a vocabulary discrimination metric based on intra-class and inter-class image similarity is developed. An iterative codeword selection algorithm is then proposed to find the optimal $\mathbf{B}$ that can maximize the developed vocabulary discrimination metric.

A. Vocabulary Discrimination Metric

The developed vocabulary discrimination metric consists of two parts: image intra-class similarity, which refers to how close and similar the compressed histograms of the images within the same category, as indicated in the numerator in (6.10), and image inter-class similarity, which is defined as how different the compressed histograms of the images between various landmark categories, as indicated in the denominator in (6.10). With respect to the word selection matrix $\mathbf{B}$, the proposed metric is defined as,

$$D(\mathbf{B}) = \frac{\sum_{l=1}^L \sum_{m=1}^{N_l} r(m) \sum_{n=1, n \neq m}^{N_l} \|\mathbf{B}\mathbf{w}_m, \mathbf{B}\mathbf{w}_n\|_{\cos} \sin \theta}{\sum_{l=1}^L \sum_{m=1}^{N_l} r(m) \sum_{n=1}^{N_m} \|\mathbf{B}\mathbf{w}_m, \mathbf{B}\mathbf{w}_n\|_{\cos} \mathbf{I}_R(\mathbf{w}_m, \mathbf{w}_n)} \quad (6.10)$$

where D is the vocabulary discrimination value using $\mathbf{B}$, L is the number of all landmark categories in the direction cluster, N_l is the number of the images in the landmark category l ,

N_m is the number of negative landmark images for the image m , which is determined according to the indicator function $\mathbf{I}_R(\cdot)$ as defined in (6.7). $r(m)$ is the image rank value as calculated in (6.8) which is used to emphasize the more discriminative images and weaken the less discriminative ones. $\|\mathbf{B}\mathbf{w}_m, \mathbf{B}\mathbf{w}_n\|_{\cos}$ is the cosine similarity between compressed histograms of the m -th and n -th images using $\mathbf{B}$. θ denotes the angle between capturing directions of the two images, and is always smaller than 30 degrees since the range of the direction cluster is 30 degrees. $\sin(\theta)$ is a weighting factor, and used to emphasize the words that spread across a large perspective view around the landmark. The reason is that if a word can be found across a large view angle around the landmark, then it has good representative capability for this landmark and will be emphasized. On the contrary, if a word can be found only within a narrow view angle around the landmark, then it has limited representative capability as reflected by small $\sin(\theta)$ value.

From (6.10) it is seen that the proposed vocabulary discrimination metric is an image similarity ratio between intra-classes and inter-classes, and determined by the word selection matrix $\mathbf{B}$. Based on this, the optimal $\mathbf{B}$ can be defined as that leading to maximum intra-class image similarity and minimum inter-class image similarity, that is, maximizing the discrimination value D . This is formulated as:

$$\begin{aligned}
 \mathbf{B}_{optimal} &= \arg \max_{\mathbf{B}} D(\mathbf{B}) \\
 &= \arg \max_{\mathbf{B}} \left(\frac{\sum_{l=1}^L \sum_{m=1}^{N_l} r(m) \left(\sum_{n=1, n \neq m}^{N_l} \|\mathbf{B}\mathbf{w}_m, \mathbf{B}\mathbf{w}_n\|_{\cos} \sin \theta \right)}{\sum_{l=1}^L \sum_{m=1}^{N_l} r(m) \left(\sum_{n=1}^{N_m} \|\mathbf{B}\mathbf{w}_m, \mathbf{B}\mathbf{w}_n\|_{\cos} \mathbf{I}_R(\mathbf{w}_m, \mathbf{w}_n) \right)} \right) \quad (6.11)
 \end{aligned}$$

where $\mathbf{B}_{optimal}$ is the learned optimal word selection matrix. In order to speed up the learning process of $\mathbf{B}$, the LSA, which is used by the ImageRank technique in Section 6.4.2, is

adopted here to transform the original function (6.11) from BoW space into latent semantic space for solving problem, as follows,

$$\begin{aligned} \mathbf{B}_{optimal} &= \arg \max_{\mathbf{B}} D(\mathbf{B}) \\ &= \arg \max_{\mathbf{B}} \left(\frac{\sum_{l=1}^L \sum_{m=1}^{N_l} r(m) \left(\sum_{n=1, n \neq m}^{N_l} \| (\mathbf{B}\mathbf{w}_m)^T (\mathbf{B}\bar{\mathbf{U}}) \bar{\Sigma}^{-1}, (\mathbf{B}\mathbf{w}_n)^T (\mathbf{B}\bar{\mathbf{U}}) \bar{\Sigma}^{-1} \|_{cos} \sin(\theta) \right)}{\sum_{l=1}^L \sum_{m=1}^{N_l} r(m) \left(\sum_{n=1}^{N_m} \| (\mathbf{B}\mathbf{w}_m)^T (\mathbf{B}\bar{\mathbf{U}}) \bar{\Sigma}^{-1}, (\mathbf{B}\mathbf{w}_n)^T (\mathbf{B}\bar{\mathbf{U}}) \bar{\Sigma}^{-1} \|_{cos} \mathbf{I}_R(\mathbf{w}_m, \mathbf{w}_n) \right)} \right) \end{aligned} \quad (6.12)$$

B. Iterative Codeword Selection

In order to find the optimal $\mathbf{B}$ in (6.12), an iterative codeword selection algorithm is proposed. The algorithm iteratively selects the word that produces the largest discrimination value into the DCV. The following notations are defined for use: i denotes the iteration count which is also the selected word index in the generated DCV, since during each iteration only one word is selected; j denotes the word index in the original codebook, $\mathbf{B}^{i-1}$ denotes the learned word selection matrix during the $(i-1)$ -th iteration, $\mathbf{y}^j = [0, \dots, pos(j), \dots, 0]_{M \times 1}$ denotes that the j -th visual word in the original codebook that will be selected into the DCV, $\mathbf{u}^i = [0, \dots, pos(i), \dots, 0]_{K \times 1}$ denotes that the selected word is mapped as the i -th word c_i in the generated DCV. Based on these definitions, we use $\mathbf{u}^i (\mathbf{y}^j)^T$ to indicate that the j -th word in the original codebook is selected and mapped as the i -th word in the new DCV during the i -th iteration. The word selection matrix $\mathbf{B}^{i-1}$ is updated to $\mathbf{B}^i$ by following operation:

$$\mathbf{B}^i = \mathbf{B}^{i-1} + \mathbf{u}^i (\mathbf{y}^j)^T, \quad (6.13)$$

From (6.13) it is seen that the key step is to determine which word (indexed by j) in the original codebook is selected as c_i in the new DCV during the i -th iteration. This is solved by,

$$\begin{aligned} c_i &= \arg \max_j D(\mathbf{B}^i) \\ &= \arg \max_j D(\mathbf{B}^{i-1} + \mathbf{u}^i (\mathbf{y}^j)^T), j = 1, \dots, M. \end{aligned} \quad (6.14)$$

where the discrimination value of D with respect to the word selection matrix $\mathbf{B}$ is calculated through the LSA as follows,

$$D(\mathbf{B}) = \frac{\sum_{l=1}^L \sum_{m=1}^{N_l} r(m) \left(\sum_{n=1, n \neq m}^{N_l} \| (\mathbf{B} \mathbf{w}_m)^T (\mathbf{B} \bar{\mathbf{U}})^{-1} \bar{\Sigma}^{-1}, (\mathbf{B} \mathbf{w}_n)^T (\mathbf{B} \bar{\mathbf{U}})^{-1} \bar{\Sigma}^{-1} \|_{\cos} \sin(\theta) \right)}{\sum_{l=1}^L \sum_{m=1}^{N_l} r(m) \left(\sum_{n=1}^{N_m} \| (\mathbf{B} \mathbf{w}_m)^T (\mathbf{B} \bar{\mathbf{U}})^{-1} \bar{\Sigma}^{-1}, (\mathbf{B} \mathbf{w}_n)^T (\mathbf{B} \bar{\mathbf{U}})^{-1} \bar{\Sigma}^{-1} \|_{\cos} \mathbf{I}_R(\mathbf{w}_m, \mathbf{w}_n) \right)} \quad (6.15)$$

After determining the selected word c_i using (6.14), $\mathbf{B}$ is then updated using (6.13). The word selection process above will be terminated when the maximum number of selected words K is reached or the discrimination value D with respect to $\mathbf{B}$ starts to decrease. Algorithm 6.2 summarizes the proposed iterative word selection process.

Algorithm 6.2 Iterative word selection

Input: BoW histograms for all the landmark images, SVD decomposed matrix $\bar{\mathbf{U}}, \bar{\Sigma}$, DCV size: K

Output: Optimal word selection matrix $\mathbf{B}_{optimal}$

For $i = 1, 2, \dots, K$

- 1 Select the most discriminative codeword c_i using (6.14).
- 2 Update the word selection matrix $\mathbf{B}^i$ using (6.13).
- 3 Calculate the discriminative value D^i with respect to current $\mathbf{B}^i$ using (6.15).
- 4 If $D^i - D^{i-1} < \varepsilon$, then stop, where ε is a predefined threshold.
- 5 Iteration number $i++$.

End

6.4.4 DSVT-based Context-Content Integration

This section details the image recognition process based on the integration of direction, location and content analysis. At the client end, a landmark image is first captured by the mobile device as the query, which is tagged with both location and direction data. Then the closest direction cluster with the query image is determined and the DSVT for that cluster is used to quantize the image into a BoW histogram, as described in [112]. Finally, the learned word selection matrix $\mathbf{B}$ is used to compress this high-dimensional BoW histogram into a compact vector through (6.9), which is then encoded through the arithmetic or other coding techniques to be transmitted to the server.

At the server end, the DSVT addresses the scalable search issue through an inverted indexed file system, which is similar to the file structure in [58], [64]. The inverted file system indexed the training images to the codewords, and maintains two lists: (i) a sorted array of images indicating the database images that have visited that codeword, and (ii) a corresponding array of counts indicating the visit number. When performing a query for an image, only the images of the landmarks that lie (i) in the selected direction cluster and (ii) within a 200m geographical distance away from the query image (as discussed in Section 6.4.2), are selected for similarity matching. These images can be quickly scored by traversing only the codewords in the DCV that are visited by the query descriptors. In cases when the query image's direction is near to the boundary of the two direction clusters, e.g., within ± 5 degrees, we choose both the two nearest clusters for consideration. The highest scored image is taken as the recognized landmark for the query.

6.5 Experimental Evaluation

6.5.1 Experimental Setup

To the best of our knowledge, there are few landmark datasets that are tagged with direction data. Therefore in order to evaluate the effect of both direction and location information, we use the NTU50Landmark database for experimental evaluation.

6.5.2 Performance Comparison with Other Approaches

In this experiment, dense SIFT descriptors that are shown to be effective in describing landmark images are extracted from multiscale patches in an image, followed by hierarchical k-means clustering to generate the DSVT. The mobile visual search method in [58] is used as the baseline for comparison. Several mobile landmark recognition methods are also implemented for comparison, including the information gain-based word learning in [70], the descriptive word selection in [101], the latent visual context word learning in [123] and the location discriminative vocabulary learning in [64]. Figure 6.5 gives the recognition rates of these methods for different DSVT depths with H ranging from 3 to 6, and $B = 10$. It is seen that the proposed method consistently outperforms the other methods for H ranging from 3 to 6, and can achieve a maximum recognition rate of around 95% when H is beyond 4. There are two reasons for the superiority of the proposed approach over other methods. Firstly, both direction and location information are considered to learn the DCV in this work while only location information is used for vocabulary learning in [64]. The direction data is reliable, and using its integration with GPS for DCV learning can provide better results for landmark recognition. Secondly, the proposed DCV learning considers the discriminative information of the images, whereas the conventional methods such as [70], [101] do not consider this. It is also noted that when the DSVT depth goes

beyond 4, the performance improvement of all the methods become less obvious. This is because a DSVT with $H = 4$ is large enough for the size of this landmark database.

Figure 6.5 Comparison of the proposed DCV learning approach with other typical approaches

Furthermore, a comparison with and without using the context analysis for DCV learning is conducted, which includes the (i) content-based DCV learning (using the ImageRank and iterative word selection algorithm), which does not include either the direction or location information (e.g., using conventional SVT to replace the DSVT, location and direction analysis in (6.6)(6.10) is removed), (ii) direction and content-based DCV learning, which uses DSVT and content analysis as in (i), and (iii) the proposed overall approach for DCV learning. The results are given in Figure 6.6. It can be seen that the overall approach using direction, location and content integration presents the best performance for each DSVT depth, followed by direction and content analysis, and lastly the content alone. This shows that direction and location both play important roles in context-aware DCV learning approach.

Figure 6.6 Comparison with and without using context analysis for DCV learning

6.5.3 Evaluation of the DCV Compression

In this section, the proposed DCV learning approach and the location discriminative vocabulary (LDV) method in [64] are compared to investigate their performance for vocabulary compression, which is meaningful for mobile landmark recognition since smaller DCV size indicates less transmission cost and faster recognition speed. The LDV method in [64] is chosen for comparison due to its better performance when compared with other state-of-art methods as shown in Figure 6.5. The DSVT depth is chosen as 4, which produces 10^4 codewords in each direction cluster. Since the DCV is generated by selecting a certain number of discriminative words, we will show the experimental results of the DCV with various numbers of words, which is given in Figure 6.7. From the figure it is seen that the proposed method based on integration of direction, location and content analysis achieves the best recognition performance for various DCV sizes. It is also noted that when using both direction and location information for DCV learning, the DCV size can be compressed to 50% of the original number of the words, with the best performance achieving around 95% accuracy. This shows the effectiveness of the proposed method for

vocabulary compression. Further, the proposed method using content analysis alone consistently outperforms the method in [64] for all compressed DCV sizes. This shows the effectiveness of the proposed ImageRank and iterative word selection algorithms. Finally, when the DCV size goes beyond a certain threshold, the performance of all the methods tend to decrease gradually. This is because some words have poor discriminative capability such as the background words. If these words are included into the DCV for recognition, they will have a negative impact on the recognition result. Furthermore, when the DCV size is equal to 10^4 , which means no word selection process is involved, the recognition rates for all these methods become the same since they are using the same set of vocabulary.

Figure 6.7 Comparison of the proposed DCV learning method with other methods

Furthermore, as discussed before a key issue in mobile recognition is the limited bandwidth or at times poor wireless network condition, it is thus important to reduce the size of the transmitted image descriptors. Based on this, the effect of BoW histogram compression using the learned DCV is shown in Figure 6.8, where the vertical axis denotes

the dimension ratio between the compressed vector and the original BoW histogram, and the horizontal axis denotes the DSVT depth. From the figure it is seen that (i) the compressed BoW histogram (vector) using DCV accounts for only a small proportion of the original BoW histogram, due to the small size of the compressed DCV. (ii) Some words in the BoW histogram may have zero counts. When these zero elements are removed, the histogram can be further compressed. (iii) When the DSVT depth becomes greater, the compression effect of the proposed method becomes more obvious. This is because when the DSVT depth becomes larger, more words that have poor discriminative capability will be removed using the proposed DCV learning.

Figure 6.8 Illustration of the BoW histogram compression effect using the DCV

6.5.4 Evaluation of Recognition Time

Table 6.1 shows the average recognition time of the proposed approach. From the table it is seen that the DCV method takes less recognition time than the original BoW method. Specifically, when the DSVT depth $H = 4$, an average recognition time of 0.81s is achieved which satisfies mobile users' fast response-time requirement. Furthermore, it is noted that when the DSVT depth H increases from 3 to 6, the average recognition time for the

proposed method increases slightly, from 0.72s to 0.99s. This shows that the proposed method achieves good computational efficiency.

Table 6.1 Average recognition time (s) after histogram compression using proposed method

DSVT depth Method	H = 3	H = 4	H = 5	H = 6
Original BoW histogram	1.12	1.31	1.45	1.53
DCV-based compressed vector	0.99	1.22	1.29	1.35
Compressed vector followed by zero elements removal	0.72	0.81	0.92	0.99

6.6 Experimental Results on San Francisco Landmark Dataset

In order to demonstrate the effectiveness of the proposed method on a large-scale dataset, we have used the large-scale San Francisco landmark dataset from [126] for experiments. The San Francisco dataset contains a database of 1.7 million images of buildings in San Francisco (1.06M Perspective Central Images (PCI) and 638K Perspective Frontal Images (PFI)) with ground truth labels, geotags, and calibration data, as well as a query set of 803 cell phone images taken with a variety of different camera phones by various people over several months. We select the 1.06M PCIs for experiments, and exclude the PFIs since the PFIs require the query image to be rectified using vanishing points for matching, which is beyond the scope of this work. The LDV method in [64] is selected for comparison as discussed before. As the San Francisco dataset contains only GPS data, we use two methods in this experiment: one uses content analysis alone as discussed in Section 6.5.2, and another uses the content and GPS integration analysis. For content and GPS integration, in the training phase, the conventional SVT is constructed to obtain preliminary codewords, the ImageRank and codeword selection algorithm (the $\sin\theta$ factor in (6.10) is removed in this case) is used to generate the DCV. In the testing phase, GPS data is first used to

shortlist geographically nearby landmark candidates, the proposed method in Section 6.4.4 (using SVT to replace DSVT) is further used for refinement.

Figure 6.9 gives the recognition rates of these methods for different SVT depth H ranging from 3 to 6, and $B = 10$. From the figure it can be seen that when the SVT depth reaches 6, which means a million leaf codewords are generated, the proposed content and GPS integration analysis achieves the highest recognition rate among all the methods (around 89%). This shows that the proposed method can obtain good performance in a different large-scale benchmark dataset.

Figure 6.9 Experiment results of different methods on San Francisco landmark dataset

The recognition time using the proposed approach on the San Francisco landmark database is also investigated, as shown in Table 6.2. The SVT depth is set as 6. From the table, it is seen that the proposed content and GPS integration method has the best average recognition time. This shows the good scalability of the proposed SVT-based method for mobile landmark recognition.

Table 6.2 Average recognition time (s) using different methods

Methods	Proposed method		LDV in [64]	
	Content analysis alone	Content and GPS integration	Content analysis alone	Content and GPS integration
Average recognition time (s)	1.6	1.2	1.9	1.5

6.7 Performance Comparison of SVM and SVT

In this section, the SVM and SVT-based landmark recognition methods are discussed. A performance comparison between these two methods is conducted, in terms of recognition accuracy and time. They are summarized as follows.

Chapters 3, 4 and 5 adopt the SVM as the recognition method. Since SVM does not have good scalability, the context information including the GPS and direction data is used to reduce the search space so that the SVM computation is manageable. The procedure is given as follows. During the online querying phase, a number of landmark candidates are first selected from the viewcone that is formed through context analysis. The query image is then forwarded to the SVM classifiers corresponding to the shortlisted landmark candidates for recognition. Through this, the image query time is greatly reduced. This is because when performing mobile landmark recognition, we only need to perform content-based classification on a small number of candidate landmarks that are obtained through viewcone of context analysis.

This Chapter 6 adopts the SVT as the recognition method. In order to obtain good recognition results, the context information is involved for SVT learning and matching. During the SVT learning phase, direction and location information are exploited to create DSVTs. The discriminative codewords of the DSVT are then selected to compress the image into a compact BoW histogram. An inverted file structure is constructed for the

DSVT to speed up the recognition process. During the querying phase, the testing image propagates down the DSVT, and scores each database image through the inverted files.

In order to better understand the advantages and disadvantages of these methods, a comparison is given in Table 6.3, where the recognition rate and time of these methods on the NTU50Landmark database are given. From the table it is seen that when using the context (direction and location) and content integration, the SVM classification obtains better recognition performance than the SVT method. Specifically, the proposed BoP and BoW combination in Chapter 5 achieves the highest recognition rate of 98.3%, which outperforms the SVT method by 3%. The reason is that SVM focuses on data classification while SVT focuses on data clustering. SVM learning is a supervised and deterministic process while SVT is not. Further, there is no decision boundary learning process during the SVT construction, which is important for differentiating two classes of images. However, it is noted that the SVT method has faster recognition time of less than 1s when compared with the SVM methods whose recognition time ranges from 3-6 seconds.

Table 6.3 Comparison of the proposed methods on the NTU50Landmark database

Method		Recognition rate (%)	Average recognition time (s)
SVM	Integrated content and context Analysis (Chapter 3)	96.6	5.1
	Discriminative BoW (Chapter 4)	97.4	3.2
	Combination of BoW and BoP (Chapter 5)	98.3	4.7
SVT	Context-aware DCV (Chapter 6)	95.3	0.8

Therefore, the SVM method has higher recognition accuracy but is comparatively slower. Therefore, it is more suitable to be used in conjunction with context information

such as GPS to constrain the number of landmarks to a small number. The SVT method has faster computational efficiency but comparatively lower recognition accuracy. Therefore, it is more suitable to be used in a large-scale landmark dataset, where the GPS information may not be available.

Finally, an illustration of two categories of images from “Yunnan garden” and “Yunnan pavilion” is given in Figure 6.10, for which the performance of the proposed methods in Chapters 3, 4 and 6 is less than satisfactory. The reason is that several challenges exist when the images of these categories are used for landmark recognition experiments. These include: (i) the two categories of images in Figure 6.10 contain a lot of backgrounds but little foreground regions, which results in the generated BoW representation to be less descriptive of the foreground landmark. Meanwhile, the images of the “Yunnan garden” category present diverse visual appearance, which further increases the difficulty to perform recognition on these images. (ii) Both the two categories of images do not have clear foreground/background structure, and the major areas such as lawn, trees and ground in the two categories of images share much similarity. This renders the generated BoW representations for these images to be similar, and difficult to differentiate. Therefore, considering that the proposed methods mainly rely on the foregrounds for landmark recognition (e.g., the discriminative BoW method in Chapter 4 uses the foreground information to learn the DGMM model for each category), the performance of the proposed methods when dealing with categories in Figure 6.10 is less than satisfactory. It is noted that the challenges described above are also encountered in many other scene/landmark recognition works, and are currently being studied by researchers.

Figure 6.10 Illustration of two categories of landmark images for which the proposed methods do not perform well. Images (a) and (b) are from the “Yunnan garden” category, whereas images (c) and (d) are from the “Yunnan pavilion” category.

6.8 Chapter Summary

In this chapter, a new DCV learning approach for mobile landmark recognition is proposed by exploiting the context information, especially the direction information for vocabulary learning. The key contributions of this chapter include (i) incorporation of direction information into the SVT construction to generate DSVT for image quantization, (ii) proposal of an effective ImageRank technique and iterative codeword selection algorithm for DCV learning. Experimental results on the NTU50Landmark and San Francisco database show that the proposed method achieves both high recognition rate and fast recognition time in mobile landmark recognition. Furthermore, the SVM-based mobile landmark recognition methods in previous chapters are compared with the SVT-based method in this chapter, and the results show that the SVM has comparatively higher recognition accuracy while the SVT has comparatively faster recognition time.

Chapter 7 Conclusions and Recommendations

7.1 Conclusions

In this thesis, several context-aware content analysis approaches are proposed for mobile landmark recognition. The direction and location information captured by the mobile device are explored, and integrated with BoW and BoP content analysis for recognition. The SVM and SVT are used for classification respectively. Three landmark datasets including the NTU50Landmark database, the Oxford building database and the San Francisco landmark database are used for experimental evaluation. The results show the effectiveness of the proposed mobile landmark recognition methods. Some conclusions are given as follows.

Firstly, the proposed content and context integration approach in Chapter 3 achieves better recognition performance and faster speed than either the content or context analysis alone for mobile landmark recognition. The context analysis involving fusion of location and direction information reduces the search space, and filters out the geographically faraway landmark candidates. The content analysis including the BoW and BoP methods can then refine the results to find the right landmark category for the query image. The experimental results on the datasets demonstrate the effectiveness of the proposed content-context integration approach for mobile landmark recognition.

Secondly, the proposed discriminative BoW learning approach in Chapter 4 achieves better recognition performance than the conventional BoW method for mobile landmark recognition. The proposed method learns the importance of different local patches in an image and incorporates this information to generate a continuous BoW histogram. The image discriminative information is also estimated, which discovers the effectiveness of

different training images to represent their categories. Fuzzy SVM utilizes the image discriminative information to train a classifier. Finally, the codebook compactness is analyzed and incorporated into the spatial pyramid kernel to make the histogram matching more robust. The experimental results on the NTU50Landmark and Oxford building datasets show the effectiveness of the proposed method.

Thirdly, the BoP content analysis method proposed in Chapter 5 achieves better recognition performance than the BoW method for mobile landmark recognition. Furthermore, a combination of the BoW and BoP achieves higher recognition accuracy than either BoW or BoP alone. The proposed BoP method finds the discriminative visual phrases for each category and combines them as a dictionary to encode an image. The visual phrases have better descriptive capability than single words in characterizing the features' spatial distribution in an image, and thus achieve better recognition performance. The BoW and BoP combination complements each other by describing not only the single word distribution, but also the pairwise words co-occurrence distribution in an image, and thus achieves better performance than either BoW or BoP alone. The experimental results using the NTU50Landmark and Oxford building datasets demonstrate the effectiveness of the proposed BoP method.

Finally, the proposed DSVT-based method in Chapter 6 achieves better scalability and faster speed than the SVM-based methods for large-scale image recognition. The DSVT method employs direction and location information to construct an SVT and selects the discriminative codewords to form the DCV. The learned DCV compresses the BoW histogram of each image into a compact vector, which is then matched with database images through an inverted file system for recognition. The experimental results on the NTU50Landmark and the San Francisco datasets show the good scalability of the SVT-based mobile landmark recognition.

7.2 Recommendations for Further Research

This thesis studies mobile landmark recognition application, which transforms a mobile device into a portable information terminal that can perform landmark recognition for the mobile users. The techniques developed can be extended to the following future directions:

(1) Mobile visual search in other domain applications.

Mobile visual search allows the user to take a snapshot of an object using the mobile device and find out the relevant information about the object. Besides the landmark as described in this thesis, the captured objects can also be a book cover, CD/DVD cover, or a product logo. Some challenging issues including the geometric deformations, photometric distortions, lighting geometry (direction and intensity of light source) and illuminant color (spectral power distribution) which substantially affect the quality of the image will need to be further studied in these applications.

(2) Mobile RFID/2D-barcode recognition

Mobile RFID/2D-barcode recognition is a sensor-based technique that uses camera phone to capture the RFID or 2D-barcode tag of an object, and further find the information about the object. The technique can be integrated with the computer vision solution to perform mobile landmark recognition. However, the sensing techniques also encounter limitations. First, they are essentially short-range solutions, requiring the user to be located near the structure containing the tags or barcodes. Second, they come at the considerable cost of equipping mobile phones with RFID tag readers and installing tags on structures near the landmarks. Third, they require the user to know where the tags and barcodes are. Therefore, a more practical system would be based on long-range computer vision (as are those we have discussed in this thesis) and complemented by tag-sensing techniques.

List of Author's Publications

Journal Papers

[J1] **T. Chen**, K.-H. Yap, and L.-P. Chau, "Integrated Content and Context Analysis for Mobile Landmark Recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, pp. 1476 - 1486, 2011.

[J2] **T. Chen**, K.-H. Yap, "Context-Aware Discriminative Vocabulary Learning for Mobile Landmark Recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, in press, 2013.

[J3] **T. Chen**, K.-H. Yap, "A Discriminative BoW Learning Technique for Mobile Landmark Recognition," *IEEE Transactions on Systems, Man and Cybernetics, Part B*, in press, 2013.

[J4] K.-H. Yap, **T. Chen**, Z. Li, and K. Wu, "A Comparative Study of Mobile-based Landmark Recognition," *IEEE Intelligent Systems Magazine*, vol.25, no.1, pp. 48-57, 2010.

[J5] **T. Chen**, K.-H. Yap, D.J. Zhang, "Discriminative Soft Bag-of-Visual Phrase for Mobile Landmark Recognition," *IEEE Transactions on Multimedia (Under Revision)*.

Conference Papers

[C1] **T. Chen**, K.-H. Yap, D.J. Zhang, "Discriminative Bag-of-Visual Phrase Learning for Landmark Recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, Mar.2012.

[C2] **T. Chen**, K.-H. Yap, and L.-P. Chau, "Content and Context Information Fusion for Mobile Landmark Recognition," *International Conference on Information, Communications and Signal Processing*, Singapore, 2011.

[C3] **T. Chen**, K.-H. Yap, L.-P. Chau, "From Universal Bag-of-Words to Adaptive Bag-of-Phrase for Scene Recognition," *IEEE International Conference on Image Processing*, pp. 841-884, Brussels, Belgium, Sep. 2011.

[C4] **T. Chen**, K.-H. Yap, L.-P. Chau, "A Discriminative Learning Approach for Mobile Landmark Recognition," *IEEE International Conference on Image Processing*, pp. 217-220, Brussels, Belgium, Sep. 2011.

[C5] **T. Chen**, Z. Li, K.-H. Yap, K. Wu, and L.-P. Chau, "A Multi-scale Learning Approach for Landmark Recognition Using Mobile Devices," *International Conference on Information, Communications and Signal Processing*, Macau, 2009.

[C6] **T. Chen**, K. Wu, K.-H. Yap, Z. Li, and F. S. Tsai, "A Survey on Mobile Landmark Recognition for Information Retrieval Systems," *MDM International Workshop on Mobile Media Retrieval*, Taipei, Taiwan, May 2009.

References

- [1] <http://globalgoodgroup.com/blog/2012/04/16/future-direction-of-mobile-applications/>
- [2] <http://motoring.asiaone.com/Motoring/Owners/Motor%2BTech/Story/A1Story20080919-88750.html>
- [3] O. Linde, T. Lindeberg, "Object recognition using composed receptive field histograms of higher dimensionality," *IEEE Proc. Image processing and pattern recognition (ICPR)*, vol.2, pp.1-6, 2004.
- [4] A. Qamra, Edward, Y. Chang, "Scalable landmark recognition using EXTENT," *Multimedia tools and Applications*, vol.38, pp.187-208,2008
- [5] G. J. Burghouts and J. M. Geusebroek, "Performance evaluation of local color invariants," *Computer Vision and Image Understanding*, vol. 113, pp. 48–62, 2009.
- [6] B. Yamauchi and P. Langley, "Place recognition in dynamic environments," *Journal of Robotic Systems*, vol. 14, pp. 107-120, 1997.
- [7] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin. "Context-based vision system for place and object recognition," *IEEE International Conference on Computer Vision*, vol. 1, pp. 273-280, Oct. 2003.
- [8] F. Scalzo, J. H. Piater, "Adaptive patch features for object class recognition with learned hierarchical models," *IEEE Conference on Computer Vision and Pattern recognition*, pp.1-8, 2007.
- [9] P. Quelhas et al., "Modeling scenes with local descriptors and latent aspects," *IEEE Conference on Computer Vision*, vol.1, pp.883-890, 2005.
- [10] A. Bosch, A. Zisserman, and X. Munoz, "Scene classification using a hybrid generative/discriminative approach," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 4, no. 3, pp. 712-726, Apr. 2008.
- [11] R. Fergus, P. Perona, A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," *IEEE International Conference on Computer Vision and Pattern Recognition*, vol.2, pp.264-271, 2003.
- [12] F. F. Li and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," *IEEE International Conference on Computer Vision and Pattern Recognition*, vol.2, pp.524-531, June 2005.

- [13] G. Dudek and D. Jugessur, "Robust place recognition using local appearance based methods," *Proc. IEEE International Conference on Robotics and Automation*, San Francisco, CA, vol. 2, pp. 1030-1035, Apr. 2000.
- [14] J. Hays and A. A. Efros, "Scene completion using millions of photographs," *ACM Transactions on Graphics*, vol. 26(3), pp.87-94, 2007.
- [15] T. Yeh, K. Tollmar, and T. Darrell, "Searching the web with mobile images for location recognition," *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 76-81, 2004.
- [16] Y. Li, Joo and H. Lim, "Outdoor place recognition using compact local descriptors and multiple queries with user verification," *International Conference on Multimedia*, Augsburg, Germany, pp. 549-552, Sept. 2007.
- [17] G. Fritz, C. Seifert, and L. Paletta, "A mobile vision system for urban detection with informative local descriptors," *Proc. IEEE International Conference on Computer Vision Systems*, pp. 30-35, Jan. 2006.
- [18] J. Lim, Y. Li, and Y. You, "Scene recognition with camera phones for tourist information access," *International Conference on Multimedia*, Beijing, pp.100-103, July 2007.
- [19] X. Fan, X. Xie, Z. W. Li, M. J. Li, and W. Y. Ma, "Photo-to-Search: Using multimodal queries to search the web from mobile devices," *International Workshop on Multimedia Information Retrieval in conjunction with ACM Multimedia*, Singapore, pp. 143-150, Nov. 2005.
- [20] http://www.esa.int/SPECIALS/TTP2/SEMK2B3XQEF_1.html
- [21] I. Haritaoglu, "Infoscope: Link from real world to digital information space," *UbiComp 2001: Ubiquitous Computing: Third International Conference Atlanta, Georgia, USA*, volume 2201/2001 of Lecture Notes in Computer Science, pp. 247–255. Springer-Verlag Heidelberg, September 30 - October 2 2001.
- [22] W. Mai, G. Dodds, and C. Tweed, "A pda-based system for recognizing buildings from user-supplied images," *Mobile and Ubiquitous Information Access: Mobile HCI 2003 International Workshop*, volume 2954 / 2004, pp. 143–157. Lecture Notes in Computer Science, Springer-Verlag Heidelberg, Sept 2003.

- [23] <http://www.windowsfordevices.com/news/NS7065484804.htm>
- [24] P. Föckler, T. Zeidler, B. Brombach, E. Bruns, and O. Bimber, "PhoneGuide: museum guidance supported by on-device object recognition on mobile phones," *International Conference on Mobile and Ubiquitous Multimedia*, vol. 154, pp. 3-10, 2005.
- [25] O. Aycard, J. F. Mari, F. Charpillet, "Second order hidden Markov models for place recognition: new results," *Tools with Artificial Intelligence*, Taiwan, pp.408-415, 1998.
- [26] O. Aycard, E Charpillet, D. Fohr, and J.F Mari., "Place learning and recognition using hidden markov models," *IEEE International Conference on Intelligent Robots and systems*, pp.1741-1746,1997
- [27] D.G. Lowe. "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91-110, 2004.
- [28] F. Scalzo and J. H. Piater, "Adaptive patch features for object class recognition with learned hierarchical models," *IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 1-8, June 2007.
- [29] C. Seifert, L. Paletta, A. Jeitler, E. Hödl, J.. Andreu, P. Luley and A. Almer, "Visual object detection for mobile road sign inventory," *Lecture Notes in Computer Science, MobileHCI 2004*, pp. 491-495, Springer, 2004.
- [30] C. Cheng, D. L. Abidi, "Object-based place recognition and loop closing with jigsaw puzzle image segmentation algorithm," *ICRA*, Pasadena, CA, pp. 557-562, 2008
- [31] A. Pronobis, B. Caputo, P. Jensfelt, and H. I. Christensen, "A discriminative approach to robust visual place recognition," *IEEE International Conference on Intelligent Robots and Systems*, pp.3829-3836, 2006.
- [32] P. Piro, S. Anthoine, E. Debreuve, and M. Barlaud "Image retrieval via Kullback-Leibler divergence of patches of multiscale coefficients in the KNN framework," *CBMI*, pp. 230-235, London, UK, 2008

- [33] J. P. Chevallet, J. H. Lim, and M. K. Leong, "Object identification and retrieval from efficient image matching: Snap2Tell with the STOIC dataset," *Information Processing and Management*, vol. 43, pp. 515-530, 2007.
- [34] T. Chin, H. Goh, J. Lim, "Boosting descriptors condensed from video sequences for place recognition," *CVPRW apos; 08. IEEE Computer Society Conference on Computer Vision and Pattern Recognition apos*, pp.1-8, 2008.
- [35] A. Pinz, A. Fussenegger, M. Auer, "Generic object recognition with boosting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.28, pp.416-431, 2006
- [36] Y. Ke and R. Sukthankar, "PCA-SIFT: a more distinctive representation for local image descriptors," *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 506-513, June 2004.
- [37] J. Wu, J. M. Rehg, "Where am I: Place instance and category recognition using spatial PACT," *IEEE International Conference on Computer Vision and Pattern Recognition*, pp.1-8, 2008.
- [38] M. Brown, R. Szeliski, and S. Winder, "Multi-image matching using multi-scale oriented patches," *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, vol.1, pp. 510-517, June 2005.
- [39] K. Hirata, E. Kasutani and Y.Hara, "On image segmentation for object-based image retrieval," *ICPR*, vol.3, pp.1031-1034, 2002.
- [40] A. E. Abdel-Hakim and A. A. Farag, "CSIFT: A SIFT descriptor with color invariant characteristics," *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 1978-1983, 2006.
- [41] Dinov, Ivo D, "Expectation maximization and mixture modeling tutorial," *California Digital Library, Statistics Online Computational Resource*, UC Los Angeles, December 9, 2008.
- [42] S. Chiu, "Fuzzy model identification based on cluster estimation," *Journal of Intelligent & Fuzzy Systems*, vol. 2, no. 3, pp. 267-278, 2007.

- [43] C.-F. Lin, S.-D. Wang, "Fuzzy support vector machines," *IEEE Transactions on Neural Networks*, vol.13, no. 2, pp.464-471, 2002.
- [44] O. L. Mangasarian, David R. Musicant, "Lagrangian support vector machines," *Journal of Machine Learning Research*, vol.1, pp.161-177, 2001.
- [45] G. Csurka, C. Dance, L.X. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," *ECCV Workshop on Statistical Learning in Computer Vision*, pp.1-22, 2004.
- [46] F. F. Li, F. Rob, T. Antonio, "Recognizing and learning object categories," *IEEE Conference on Computer Vision and Pattern Recognition Short Course*, 2007.
- [47] S. Lazebnik, C. Schmid, J.Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol.2, pp.2169-2178, 2006.
- [48] K. Grauman and T. Darrell, "Pyramid match kernels: Discriminative classification with sets of image features," *International Conference on Computer Vision*, vol.2, pp.1458-1465, 2005.
- [49] K. -H. Yap, T. Chen, Z. Li, K. Wu, "A comparative study of mobile-based landmark recognition techniques," *IEEE Intelligent System*, vol. 25, no. 1, pp. 48-57, 2010.
- [50] T. Kanungo, D. M. Mount, N.S. Netanyahu, et al., "An efficient k-Means clustering algorithm: analysis and implementation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.24, no.7, pp.881-892, 2002.
- [51] <http://www3.ntu.edu.sg/home/wukui/Landmark/>.
- [52] W. Mai, G. Dodds, and C. Tweed, "A pda-based system for recognizing buildings from user-supplied images," *Mobile and Ubiquitous Information Access: Mobile HCI International Workshop*, vol. 2954, pp. 143–157, *Lecture Notes in Computer Science*, Springer-Verlag Heidelberg, Sept 2003.
- [53] B. F. Peterson, "Learning to see creatively: design, color & composition in photography," *Amphoto Press*, 2003.

- [54] C. Reigber, G. Balmino, P. Schwintzer, etc., “A high-quality global gravity field model from CHAMP GPS tracking data and accelerometry,” *Geophysical Research Letter*, vol. 29, no. 14, pp.371-374, 2002.
- [55] O. Montenbruck, E. Gill, “Ionospheric correction for GPS tracking of LEO satellites,” *The Journal of Navigation*, vol. 55, pp. 293–304, 2002.
- [56] H. Bay, et al., "SURF: Speeded Up Robust Features," in *Computer Vision – ECCV*, vol. 3951, pp. 404-417, 2006.
- [57] T. Chen, K. -H. Yap, L. -P. Chau, “Integrated content and context analysis for mobile landmark recognition,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, pp. 1476 – 1486, 2011.
- [58] B. Girod, V. Chandrasekhar, D. M .Chen, “Mobile visual search,” *IEEE Signal Processing Magazine*, pp. 61-76, Jul. 2011.
- [59] S. Tsai, D. Chen, G. Takacs et al., “Location coding for mobile image retrieval,” *International Mobile Multimedia Communications Conference*, article number. 8, 2010.
- [60] D. M. Chen, G. Baatz, K. Koser et al., “City-scale landmark identification on mobile devices,” *IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 737-744, Jun. 2011.
- [61] R.-R. Ji, X. Xie, H. Yao, W.-Y. Ma, “Mining city landmarks from blogs by graph modeling,” *ACM International Conference on Multimedia*, pp. 105–114, 2009.
- [62] W. Zhang, J. Kosecka, “Image based localization in urban environments,” *Proceedings of the Third International Symposium on 3D Data Processing, Visualization, and Transmission*, pp. 33–40, 2006.
- [63] J. -A. Lee, K. -C. Yow, A. Sluzek, “Image-based information guide on mobile devices,” *Proceedings of the 4th International Symposium on Advances in Visual Computing*, pp.346–355, 2008.
- [64] R. -R. Ji, L. -Y. Duan, J. Chen et al., “Location discriminative vocabulary coding for mobile landmark search,” *International Journal of Computer Vision*, pp. 1-25, Jul. 2011.
- [65] H. James and A. A. Efros, “IM2GPS: estimating geographic information from a single image,” *IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 1-8, 2008.

- [66] <http://www.google.com/mobile/goggles/#text>
- [67] J. S. Hare and P. H. Lewis, "Content-based image retrieval using a mobile device as a novel interface," *Proc. of SPIE Storage and Retrieval Methods and Applications for Multimedia*, vol. 5682, pp. 64-75, San Jose, USA, Jan.2005.
- [68] G. Fritz, C. Seifert, and L. Paletta, "A mobile vision system for urban detection with informative local descriptors," *IEEE International Conference on Computer Vision Systems*, pp. 30-35, 2006.
- [69] Y. T. Zheng, M. Zhao, Y. Song, et al., "Tour the world: building a web-scale landmark recognition engine", *IEEE Conference on Computer Vision and Pattern Recognition*, pp.1085-1092, 2009.
- [70] G. Schindler, M. Brown and R. Szeliski, "City-scale location recognition," *IEEE International Conference on Computer Vision and Pattern Recognition*, pp.1-7, 2007
- [71] T. Quack, H. Bay, L. V. Gool, "Object recognition for the internet of things," *International Conference on The Internet of Things*, pp.230-246, 2008.
- [72] D. Crandall, L. Backstrom, D. Huttenlocher, et al., "Mapping the world's photos," *Proceedings of the 18th International Conference on World Wide Web*, pp.761-770, 2009.
- [73] S. Gammeter, T. Quack, D. Tingdahl, et al., "Size does matter: improving object recognition and 3D reconstruction with cross-media analysis of image," *ECCV*, pp.734-747, 2010.
- [74] T. Chen, K.-H. Yap, L.-P. Chau, "A Discriminative Learning Approach for Mobile Landmark Recognition," *IEEE International Conference on Image Processing*, pp. 217-220, Sep. 2011.
- [75] A. R. Zamir, M. Shah, "Accurate image localization based on Google maps street view," *European Conference on Computer Vision*, pp.255-268, 2010.
- [76] J. Philbin, O. Chum, M. Isard, et al., "Object retrieval with large vocabularies and fast spatial matching," *IEEE International Conference on Computer Vision*, pp.1-8, 2007.
- [77] T. Quack, B. Leibe, L. V. Gool, "World-scale mining of objects and events from community photo collections," *Proc. Of the International Conference on Content-based Image and Video Retrieval*, pp. 47-56, 2008.

- [78] S. Gammeter, L. Bossard, T. Quack, L. V. Gool, "I know what you did last summer: object-level auto-annotation of holiday snaps," *IEEE International Conference on Computer Vision*, pp.614-621, 2009.
- [79] R. Hartley, A. Zisserman, *Multiple View Geometry in Computer Vision* (2nd ed.), Cambridge University Press, 2003.
- [80] T. Li, T. Mei, I.-S. Kweon, et al., "Contextual bags-of-words for visual categorization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol.21, no.4, pp.381-392, 2010.
- [81] L. Wu, Y. Hu, and M. Li, et al., "Scale-invariant visual language modeling for object categorization," *IEEE Transactions on Multimedia*, vol. 11, no. 2, pp. 286-294, 2009.
- [82] J. Liu and M. Shah, "Scene modeling using co-clustering," *Proceedings of the IEEE International Conference on Computer Vision*, pp.1-7, 2007.
- [83] <http://www.robots.ox.ac.uk/~vgg/data/oxbuildings/index.html>
- [84] B. Yamauchi and P. Langley, "Place recognition in dynamic environments," *Journal of Robotic Systems*, vol. 14, pp. 107-120, 1997.
- [85] M. Cummins, P. Newman, "Highly scalable appearance-only SLAM - FAB-MAP 2.0," *International Journal of Robotics Research*, vol.30, no.9, pp. 1100-1123, 2010.
- [86] E. Nowak, F. Jurie, Bill Triggs, "Sampling strategies for bag-of-features image classification," *European Conference on Computer Vision*, pp. 490-503, 2006.
- [87] M. Aly, M. Munich, P. Perona, "Indexing in large scale image collections: scaling properties and benchmark," *IEEE Workshop on Applications of Computer Vision*, pp.418-425, 2011.
- [88] L. Lu, K. Toyama, G. D. Hagar, "A two level approach for scene recognition," *IEEE International Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 688-695, 2005.
- [89] J. Knopp, J. Sivic, T. Pajdla, "Avoiding confusing features in place recognition," *European Conference on Computer Vision*, pp.748-761, 2010.
- [90] D. Parikh, C. L. Zitnick, T. Chen, "Determining patch saliency using low-level context," *Proceedings of the 10th European Conference on Computer Vision*, pp.446-459, 2008.

- [91] X. W. Li, C. C. Wu, C. Zach et al., "Modeling and recognition of landmark image collections using iconic scene graphs," *European Conference on Computer Vision*, pp.427-440, 2008.
- [92] L. S. Kennedy, M. Naaman, "Generating diverse and representative image search results for landmarks," *International Conference on World Wide Web*, pp.297-306, 2008.
- [93] T. Deselaers, G. Heigold, H. Ney, "Object classification by fusing SVMs and Gaussian mixtures," *Pattern Recognition*, vol. 43, pp. 2476-2484. 2010.
- [94] J. C. Gemert, C. J. Veenman, A. W. M. Smeulders, "Visual word ambiguity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, 2010.
- [95] J. Sivic, A. Zisserman, "Video google: A text retrieval approach to object matching in videos," *Proc. Intl. Conf. on Computer Vision*, pp. 1470-1477, 2003.
- [96] Q. Zheng, W. Wang, and W. Gao, "Effective and efficient object-based image retrieval using visual phrases," *ACM International Conference on Multimedia (ACM MM)*, pp.77-80, 2006.
- [97] J. Yuan, Y. Wu, M. Yang, "Discovery of collocation patterns: from visual words to visual phrases," *Proceedings of IEEE Computer Vision and Pattern Recognition*, pp.1-8, 2007.
- [98] Y. Song, Q. Tian, M. Wang et al., "Multiple instance learning using visual phrases for object classification," *IEEE International Conference on Multimedia and Expo*, pp. 649-654, 2010.
- [99] Y.-T. Zheng, M. Zhao, S.-Y. Neo, et al., "Visual synset: Towards a higher-level visual representation," *IEEE International Conference on Computer Vision and Pattern Recognition*, pp.1-8, 2008.
- [100] J. B. Yeh, C. H. Wu, "Extraction of robust visual phrases using graph mining for image retrieval," *IEEE International Conference on Multimedia and Expo*, pp. 3681-3684, 2010.
- [101] S. L. Zhang, Q. Tian, G. Hua, et al., "Descriptive visual words and visual phrases for image applications," *ACM International Conference on Multimedia*, pp.75-84, 2009.
- [102] J. Han, H. Cheng, D. Xin, and X. Yan. "Frequent pattern mining: current status and future directions," *Data Mining and Knowledge Discovery*, pp.55-86, 2007.

- [103] J. -B. Ryu, C.-G. Lee, H.-H. Park, "Formulas for Harris corner detector," *Electronics letters*, vol. 47, no.3, pp.180-181, 2011.
- [104] V. Garcia, E. Debreuve, M. Barlaud, "Fast k nearest neighbor search using GPU," *IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 1-6, 2008.
- [105] K. J. Hintz, "A measure of the information gain attributable to cueing," *IEEE Transactions on Systems, Man, and Cybernetics*, vol.2, pp.434-442, 1991.
- [106] S. Kullback, "The kullback-leibler distance," *The American Statistician*, vol. 41, pp.340-341, 1987.
- [107] A. Bosch, A. Zisserman, X. Munoz, "Scene classification using a hybrid generative/discriminative approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp.712-727, 2008.
- [108] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning*, vol. 41, no. 2, pp. 177-196, 2001.
- [109] Y. Rubner, C. Tomasi and L. J. Guibas, "The Earth Mover's Distance as a metric for image retrieval," *International Journal of Computer Vision*, pp. 99-121, 2000.
- [110] V. Chandrasekhar, G. Takacs, D. Che et al., "Compressed histogram of gradients: a low bitrate descriptor," *International Journal on Computer Vision*, vol. 94, no. 5, pp.384-399, May 2011.
- [111] E. tola, V. lepetit, P. fua, "Daisy: an efficient dense descriptor applied to wide baseline stereo," *IEEE Trans. On Pattern Analysis and Machine Intelligence*, vol.32, no.5, pp.815-830, 2010.
- [112] D. Nister, H. Stewenius, "Scalable recognition with a vocabulary tree," *IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 2161-2168, 2006.
- [113] A. Irschara, C. Zach, J. Frahm, et al., "From structure from-motion point clouds to fast location recognition," *IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 2599–2606, 2009.
- [114] E. Kalogerakis, O. Vesselova, J. Hays et al., "Image sequence geolocation with human travel priors," *IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2009.

- [115] D. Liu, M. Scott, R. Ji, et al., “Location sensitive indexing for image-based advertising,” *ACM International Conference on Multimedia*, pp. 793–796, 2009.
- [116] D. Chen, S. Tsai, V. Chandrasekhar, “Tree histogram coding for mobile image matching,” *Data Compression Conference*, pp. 143–152, 2009.
- [117] V. Chandrasekhar, G. Takacs, D. Chen et al., “CHoG: Compressed histogram of gradients a low bit-rate feature descriptor,” *IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 2504–2511, 2009.
- [118] L. Wang, L. P. Zhou, C. H. Shen, “A fast algorithm for creating a compact and discriminative visual codebook,” *European Conference on Computer Vision*, vol. 5305, pp. 719-732, 2008.
- [119] L. Yang, R. Jin, R. Sukthankar, F. Jurie, “Unifying discriminative visual codebook generation with classifier training for object category recognition,” *IEEE International Conference on Computer Vision and Pattern Recognition*, pp.1-8, 2008.
- [120] L. Wang, “Toward a discriminative codebook: Codeword selection across multi-resolution,” *IEEE International Conference on Computer Vision and Pattern Recognition*, pp.1-8, 2007
- [121] L. J. Zhang, C. Chen, J. J. Bu et al., “Discriminative codeword selection for image representation,” *ACM International Conference on Multimedia*, pp. 173-182, 2010.
- [122] S. Deerwester, S. T. Dumais, R. Harshman, “Indexing by latent semantic analysis,” *Journal of the American Society for Information Science*, vol. 41(6), pp. 391–407, 1990.
- [123] W. G. Zhou, Q. Tian, Y. Lu et al., “Latent visual context learning for web image applications,” *Pattern Recognition*, vol. 44, pp. 2263-2273, 2010.
- [124] <http://www.meridianworlddata.com/Distance-Calculation.asp>
- [125] Y. Jing, S. Baluja, “VisualRank: applying PageRank to large-scale image search,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp.1877–1890, 2008.
- [126] D. M. Chen, G. Baatz, K. Koeser, et al., “City-scale landmark identification on mobile devices,” *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.737-744, 2011.

- [127] J. Q. Chen, T. Pappas et al., “Adaptive perceptual color-texture image segmentation,” *IEEE Transaction on Image Processing*, vol. 14, pp.1524-1536, 2005.
- [128] J. C. Yang, K. Yu, Y.H. Gong et al., “Linear spatial pyramid matching using sparse coding for image classification,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp.1794-1801, 2009.
- [129] Y. Jia, C. Huang, T. Darrell, “Beyond spatial pyramids: Receptive field learning for pooled image features,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp.3370-3377, 2012.