Tools for visual scene recognition

Elahe Farahzadeh

2014

Elahe Farahzadeh. (2014). Tools for visual scene recognition. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/59540>

<https://doi.org/10.32657/10356/59540>

TOOLS FOR VISUAL SCENE RECOGNITION

ELAHE FARAHZADEH

School of Computer Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

2014

Acknowledgments

I would like to express my deep and sincere gratitude to my supervisors, Dr. Cham Tat-jen and Dr. Andrzej Sluzek, whose encouragement, supervision and support helped me to develop an understanding of the subject. Their wide knowledge and constructive comments have been great values for me.

I am thankful of **Centre of Computational Intelligence (C2I)** and **Centre for Multimedia and Network Technology (CeMNet)** for providing me resources to follow my research. I would also like to appreciate **School of Computer Engineering (SCE)** and **Nanyang Technological University (NTU)** for offering the scholarship to pursue my Ph.D. candidate study.

I am especially forever grateful to my parents and my brother Arman, for their love, understanding and encouragement. Without them, I would not be the person I am today. Above all I would like to thank my husband Behrouz, for his love and constant support and for unwavering belief in me. I thank him for being always there for me.

Contents

Acknowledgments	i
List of Figures	v
List of Tables	vii
List of Abbreviations	viii
Abstract	x
1 Introduction	1
1.1 Motivation	1
1.2 Topics and Approaches of this thesis	5
1.2.1 Scene Recognition by Semantic Visual Words [1, 2]	6
1.2.2 Local-Global Feature Fusion (LGFF) for Scene Recognition [3] . .	6
1.2.3 Capturing Image Spatial Information in Scene Recognition Sys- tems [4, 5]	6
1.2.4 Scene Datasets used in this dissertation	7
1.3 Outline of Thesis	8
2 Literature Review	11
2.1 Content-based Image Retrieval	11
2.2 Low-level Feature Extraction	13
2.2.1 Feature Detection	13
2.2.2 Feature Description	16
2.3 Similarity Measure	19
2.4 Semantic Gap Reduction	21
2.4.1 Region-based Systems for Semantic Gap Reduction	21
2.4.2 Object Ontology	22

2.4.3	Learning Methods	22
2.4.4	Relevance Feedback (RF)	22
2.4.5	Semantic Templates (ST)	23
2.5	Scene recognition techniques	24
3	Scene Recognition by Semantic Visual Words	28
3.1	Introduction	28
3.2	Related Works	31
3.3	Overview of the Proposed Framework	33
3.3.1	Feature Extraction	33
3.3.2	Term-document Matrix	34
3.4	Concept Space	36
3.4.1	Embedding into Concept Space using Latent Semantic Analysis .	36
3.4.2	Embedding into Concept Space using Probabilistic Latent Semantic Analysis	37
3.5	Experimental Results for Scene Recognition	40
3.5.1	Results and Analysis	40
3.6	Chapter Summary	46
4	Local-Global Feature Fusion for Scene Recognition	48
4.1	Introduction	48
4.2	Related Works	49
4.3	Overview of the proposed framework	51
4.4	Feature Extraction and Representation	52
4.4.1	Capturing Semantic Information by Scene Modeling with Probabilistic Latent Semantic Analysis (pLSA)	52
4.4.2	Capturing Contextual Information	52
4.5	Local-Global Feature Fusion (LGFF)	53
4.5.1	Exemplar-Set Selection	53
4.5.2	Problem Formulation	55
4.5.3	Parameter Learning	56
4.6	Experimental Results	59

4.6.1	Evaluation Results of 15-Scene Dataset	59
4.6.2	Evaluation Results of 67-Indoor Scenes Dataset	62
4.7	Chapter Summary	65
5	Capturing Image Spatial Information in Scene Recognition Systems	66
5.1	Introduction	66
5.2	Spatial Content Capture in Semantic Vocabulary Framework for Scene Recognition	68
5.2.1	Capturing Spatial Information with Spatial Pyramid Matching (SPM)	68
5.2.2	Spatial Content Capture using Global and Region-based Models .	69
5.2.3	Experimental Evaluation	71
5.3	Spatial Content Capture in Local-Global Feature Fusion for Scene Recognition	73
5.3.1	Capturing Spatial Location Information	73
5.3.2	Spatial Local-Global Feature Fusion (Spatial-LGFF)	74
5.3.3	Experimental Evaluation	77
5.3.4	Implementation details and Memory complexity of the Spatial-LGFF method	80
5.4	Chapter Summary	81
6	Conclusions and Future Work	88
6.1	Summary of Contributions	88
6.2	Future Work	89
6.2.1	Scene Recognition by Considering the Contextual Information of Local Patches	90
6.2.2	Determining the Number of Topics in the Latent Semantic Space	90
6.2.3	Modified Exemplar-Set Selection for Local-Global Feature Fusion (LGFF)	91
6.2.4	Adopting other optimization techniques for parameter learning . .	92
	Publications	93
	References	94

List of Figures

1.1	Sample images from the 15-Scene dataset.	8
1.2	Sample images from the 67-Indoor Scenes categories dataset.	9
2.1	SIFT descriptor applied to a patch [6].	19
2.2	Relevance Feedback workflow.	23
3.1	Standard Bag-of-words Framework.	29
3.2	Constructing the semantic visual vocabulary.	33
3.3	$M \times N$ word/term document matrix.	35
3.4	SVD decomposition, word space and concept space.	37
3.5	Graphical view of pLSA model.	38
3.6	pLSA ,word space and concept space.	38
3.7	Performance of proposed method on 15-Scene dataset with different number of topics using LSA.	41
3.8	Performance of proposed method on 15-Scene dataset with different number of topics using pLSA for embedding.	42
3.9	Performance of proposed method on 15-Scene dataset with $K_f = 1000$ and $K_i = 1500, 2000, 3000$, with different number of topics using LSA.	42
3.10	Performance of proposed method on 15-Scene dataset divided into Indoor and Outdoor- classes with different number of topics using LSA.	43
3.11	Comparison of results with the classic framework for different sizes of vocabulary in 15-Scene dataset.	44
3.12	Confusion matrix for the best result achieved in 15-Scene dataset.	45
4.1	Block diagram of learning step in Local-Global Feature Fusion method.	51

4.2	Exemplar-set selection.	54
4.3	Recognition accuracy vs. exemplar images per category in 15-Scene dataset.	60
4.4	Per class recognition rate for Supervised pLSA, Gist and LGFF.	61
4.5	Comparison of scene recognition accuracy for different features (pLSA vs. Gist vs. CENTRIST) and combination of features using (LGFF vs. MKL).	63
4.6	Recognition accuracy vs. exemplar images per category in 67-Indoor Scenes dataset.	63
4.7	Comparison of scene recognition accuracy for different features (pLSA vs. Gist vs. CENTRIST) and combination of features using LGFF.	65
5.1	Example of constructing a three-level pyramid.	70
5.2	Performance of proposed global method using LSA.	71
5.3	Performance of proposed global method using pLSA.	72
5.4	Recognition accuracy using pLSA vs. LSA concept space for initial vocabulary sizes $K_i = 1000, 1500, 2000$ using global method.	82
5.5	Recognition accuracy using pLSA and LSA concept space in different levels of resolution.	83
5.6	Comparing the performance of proposed global method vs. region-based method using LSA.	83
5.7	Block diagram of learning phase in our Feature Fusion method.	84
5.8	Block diagram of Scene Recognition for test input image.	85
5.9	Confusion matrix for the best result in 15-Scene dataset.	86
5.10	Spatial-LGFF Recognition rate vs. Baseline frameworks for 15-Scene dataset.	86
5.11	Spatial-LGFF Recognition rate vs. Baseline frameworks for 67-Indoor Scenes dataset.	87

List of Tables

3.1	Comparison with recently reported results for 15-Scene dataset.	44
	47	
4.1	Comparison with recently reported results for 15-Scene.	60
4.2	Recognition rates for indoor, outdoor categories in 15-Scene dataset. . . .	61
4.3	Comparison with recently reported results for 67-Indoor Scenes.	64
5.1	Comparison of execution time for building semantic visual vocabulary in 15-Scene dataset using LSA/pLSA.	73
5.2	Comparison with recently reported results for 15-Scene.	78
5.3	LGFF vs. MKL execution time for 15-Scene dataset.	79
5.4	Comparison with recently reported results for 67-Indoor Scenes.	79
5.5	Recognition Accuracy per class for 67-Indoor Scenes.	80
5.6	LGFF execution time for 67-Indoor Scenes dataset.	80

List of Abbreviation

BOW	Bag Of Words
CBIR	Content-Based Image Retrieval
CCA	Canonical Correlation Analysis
CENTRIST	CENsus TRansform hISTogram
CT	Census Transform
DFT	Discrete Fourier Transform
DM	Diffusion Map
DoG	Difference of Gaussians
DPM	Deformable Part-based Models
DST	Discriminant Spectral Template
EBR	Edge Boundary Regions
EM	Expectation Maximization
EMD	Earth Mover Distance
FAST	Features From Accelerated Segment Test
HIK	Histogram Intersection Kernel
IBR	Intensity-Based Regions
KLT	Karhunen-Loeve Transform
KMCKK	means with Connectivity Constraint
K-CCA	Kernel-Canonical Correlation Analysis
LGFF	Local-Global Feature Fusion
LDA	Linear Discriminant Analysis
LSA	Latent Semantic Analysis
MKL	Multiple Kernel Learning
MMI	Maximization of Mutual Information
MSER	Maximally Stable Extremal Region
PCA	Principal Component Analysis
pLSA	Probabilistic Latent Semantic Analysis
PMK	Pyramid Match Kernel
PmSVM	Power mean SVM
RBF	Radial Basis Function
RF	Relevance Feedback
ROI	Region-of-Interests
SIFT	Scale-Invariant Feature Transform

SPM	Spatial Pyramid Matching
SPMSM	Spatial Pyramid Matching on the Semantic Manifold
SURF	Speeded Up Robust Features
ST	Semantic Template
SVD	Singular Value Decomposition
SVM	Support Vector Machine
SVT	Semantic Visual Templates
tf-idf	term frequency-inverse document frequency
WDST	Windowed Discriminant Spectral Template

Abstract

Scene recognition is an important step towards a full understanding of an image. This thesis presents novel ideas related to semantic-spatial content capture and local-global feature fusion techniques and applies them for scene recognition. It shows how the proper use of these approaches, without trying to recognize objects in the scene images, can lead to improvement in recognition accuracy for scene classification.

First, we propose a method to build a semantic visual vocabulary. The features are extracted from the image patches and the initial vocabulary is constructed by performing k-means clustering on the extracted features and choosing the cluster centers as the visual words. The feature vectors are quantized based on the initial vocabulary to form a word-image matrix that describes the occurrence of words in the images. The codebooks are then embedded into the concept space by latent semantic models. We demonstrate this embedding using Latent Semantic Analysis (LSA) as well as Probabilistic Latent Semantic Analysis (pLSA). In the proposed space, the distances between words represent the semantic distances, which are used to construct a discriminative and semantically meaningful vocabulary. The main contributions of the first chapter are as follows: 1. Using semantic word space to co-cluster similar words together to form a semantic visual vocabulary. This will improve the results compared to other methods that use document space directly after pLSA embedding. 2. Investigating changes in the number of latent variables. 3. Using LSA embedding when all other vision systems to date only use pLSA. This method has shown promising results on 15-Scene categories when the proposed model extracts one type of visual feature.

Second, since fusing local and global features is beneficial for achieving a promising performance for scene categorization systems [7], we propose a novel Local-Global Feature Fusion (LGFF) method with the capability to fuse latent semantic patches adaptively.

The image local feature space is embedded into the latent semantic space by employing pLSA modeling; afterwards, this semantic space and the global contextual feature space are mapped into a kernel-space. To perform this embedding, the global features and latent variables are relatively weighted for each scene category. The following is a summary of the main contributions of the second chapter: 1. Weighting latent semantic topics based on their discriminative power 2. Defining a novel exemplar-based distance learning 3. Defining a category-dependent map function; the experimental evaluations indicate radical improvements in the *15-Scene* and *67-Indoor Scenes* datasets.

Third, every scene image contains a high value of spatial information. Capturing the spatial positions of visual features plays an essential role in scene categorization; however, the methods proposed in the previous chapters disregard this position information. Inspired by methods that construct pyramid levels over image primitives [8, 9], pyramid matching kernels are employed to measure the dissimilarity scores of image pyramids. We improved the semantic vocabulary framework by considering the position of semantic local patches and their surrounding neighborhood properties using either the global or region-based spatial pyramid method. In the global method, after projecting the image into the concept space, it is divided into finer sub-scenes and then the pyramid match kernels are employed over the proposed space for co-clustering the semantic visual words. The region-based method initially divides the image into sub-scenes and then projects each sub-scene into the concept space. In the Local-Global Feature Fusion framework, by using the global features gist and CENsus TRansform hISTogram (CENTRIST), the global spatial layout of the scene is already captured and, to further improve the results, the image is divided into sub-regions at different levels of resolution. The pyramid matching kernels are then applied over these sub-regions. The representation of these sub-regions is obtained either by applying CENTRIST or a bag of Scale-Invariant Feature Transform (SIFT) visual features. The experimental results outperform most of the best published results for both the *15-Scene* and *67-Indoor Scenes* datasets.

This thesis concludes with a discussion of future directions for extending the proposed works on scene recognition.

Chapter 1

Introduction

1.1 Motivation

The fast growth of digital imaging on the web and the prevailing use of digital cameras, which lead to an enormous number of digital photo albums, have created very large and varied image databases. To manage and manipulate this excessive number of images extensive research has to be done to find effective methods to access the visual information in these databases on the levels of objects and scenes.

Primary image classification and retrieval systems use meta-data such as *tags* and *keywords* but, due to the rapid expansion of image datasets, it is not feasible to manually associate text labels with all of the images (annotation). Furthermore, the personal perspective of who annotates the image may affect the entire process, i.e. during manual annotation some of the image's concepts may be ignored and consequently the *keywords* associated with the image are not very accurate. To address these drawbacks, content-based visual systems were introduced in the 90s; these systems do not use any information other than the visual contents of the images [10].

Scene recognition is one of the most important problems in content-based visual systems. The objective of scene recognition systems is typically to assign a semantic label (such as *library*, *bedroom*, *highway*) to each class of images with regards to their visual content. (these class labels range from specific classes such as *kitchen* to generic classes such as *coast*)

Scene classification systems are motivating, challenging and remain an active research area. It is a difficult task due to the variation of scene structures, scales and illumination.

To define the scene recognition problem, we need to see how it is different from the two closely related areas of image understanding and object classification.

Image understanding is the most ambitious goal of computer vision. It is the highest level of processing and includes knowledge regarding all of the objects within the image and high-level reasoning of their relationships. Image classification is an easier goal than image understanding and can be performed without knowledge of every object within the image; it is enough to detect so-called rigid objects in image. For example, to recognize a park scene it is sufficient to detect the sky, grass and some people in the middle. Sometimes it is not even necessary to go through object detection as low-level image processing is enough for classification. Therefore, scene classification is the first step to pursue to achieve image understanding. [11, 12]

Scene recognition is not considered a derivation of object classification although they are highly related. In fact, for scene classification it is usually necessary to detect rigid objects and their geometric relationships. [11]

To conclude, let us consider we have a scene category called *birthday party*. Object detection plays an important role in scene classification by detecting the rigid objects such as birthday cake and birthday decorations. However, to have a full understanding of the image, the non-rigid objects (human faces, home furniture, etc.) need to be detected to provide more information, including the age of the birthday child, facial expressions, etc. [11]

Scene recognition has many applications in the real world:

Content-based indexing for image retrieval [13]. One of the most important applications of every image classification system, as well as scene recognition, is content-based image indexing which helps with later retrieval. Content-based image retrieval is an important area in computer vision and multimedia computing. Due to the rapid growth of online and personal digital libraries, the huge number of images needs to be organized by grouping and indexing to perform real-time image retrieval, e.g. if a user wants to plan his vacation, travel agencies must be able to provide access to indexed digital archives so he can make any query he needs.

Mobile robotics: navigation and localization systems [14, 15]. Mobile robotics is a fundamental vision problem that includes tasks such as *localization*, *navigation*

and *mapping*. Mobile robotics implementations use various types of sensors including *sonar* and *lasers*. The data received from these sensors is sometimes not accurate because of the complexity of the environment, e.g. outdoor environments. Outdoors, due to the *non-existence of flat walls or narrow corridors* and *having to deal with surface irregularities*, it is very difficult to predict the sensor input. To deal with these challenges, mobile robotics implementations inspired by scene recognition models intuitively use gist [16] or salient features and have shown promising results.

Automatic tour guides using camera phones [17]. Camera phones are portable imaging devices which according to a forecast at *www.capv.com*, about 9 out of 10 (89%) mobile consumers have on their phones. This suggests a great opportunity to develop interactive visual applications on mobile phones, such as touristic guides. These guides could be enhanced to multi-feature platforms where text, audio and visual descriptions are provided. The input image of a specific location (e.g. a popular building or statue) is analyzed by the application after a person points their camera phone toward it. After searching through a database of landmarks and scenes, information regarding the place is accessible by the user.

Automatic mode cameras [18]. Recently scene recognition systems have been used as a built-in feature in cameras to adjust the auto-shoot settings to obtain the best possible shot. For example, Nikon's Scene Recognition System (SRS) [18] uses the camera's built-in image dataset information to configure the *autofocus*, *auto exposure* and *auto white balance controls* after capturing the properties of the photo's subject. These properties are the *position*, *color* and *brightness of the scene* that the subject is located in.

In spite of the intensive research performed and numerous techniques developed for scene classification, the performance of even the most sophisticated recognition system is not satisfactory when it comes to real-world conditions such as changes in light and viewpoint, background occlusion and high intra-class variety. Due to the limitations and drawbacks of the previously mentioned frameworks, novel techniques for image classification should be developed to achieve more reliable and accurate classification systems.

The main difficulty in scene recognition is apparently the semantic gap between the visual information extracted from the scene and its semantic content. A human observer of an image is able to grasp the scene concept in a glance and, hence, images are compared according to their semantics. Unlike human understanding of a scene, the scene classification applications compare images with regards to their extracted visual primitives. Therefore, it is necessary to develop better conceptual methodologies for scene recognition. Region-based systems have been proposed for this mission. These systems exploit the image's local features instead of employing global features from the entire image. The performance of region-based systems justifies the ability of local features to reduce the semantic gap as, the visual objects within the image are better described by extracting the local features.

In this dissertation we propose using local features to develop statistical representations and machine learning techniques which are popular trends in recognition systems. Statistical approaches and machine learning methods have demonstrated remarkably promising results when applied for the purpose of clustering, classification or even developing interactive search frameworks based on user feedback [19, 20].

The advantages of local features in comparison with global features are summarized as follows [21]:

- **Robustness:** local features, due to their size and compactness, can be kept safe from image clutter and occlusion.
- **Repeatability:** two images that share the same object or scene with minor changes in view, lighting or geometric conditions should share many local features.
- **Expressiveness:** since local features are represented by high dimensional descriptors, they carry out more appearance information.
- **Local geometric invariance:** some applications must use features that are invariant to changes in scale, rotation or affine transformation. By using the appropriate local features, it is possible to bring this ability into account.
- **Compactness:** using local features has the advantage of decreasing time and memory usage because the number of features extracted by most detectors is lower than the number of pixels in the image.

As mentioned earlier, scene recognition is still an open problem in computer vision and some of the challenges are outlined below:

High intra-class variety. The intra-class diversity in scene categories is very high. Although scene images within the same category consist of the same set of objects, the number and type of objects vary from one image to another. This high variety makes it very difficult to discriminate scene categories. For example, consider objects inside stores; although the stores may contain similar objects (e.g. *checkout counter, trolleys, display cabinets*), there are many differences which enable us to distinguish one kind of store from another (e.g. *toy store* vs. *clothing store*).

The enormous number of objects within a scene is an inconvenience for scene classification as well because, having many objects in a scene makes it impossible to detect every single object inside the image and perform scene categorization based on the objects, not to mention that obtaining a robust object detection and recognition system is still under research itself.

Illumination, scale and view-point changes. The variants of *illumination, view-point* and *scale*, intuitively reflect the scene representation and, as a result, make scene recognition less accurate.

Illumination, brightness and spectral changes in a scene may occur due to multiple sources of light, the direction of emitted light or reflections from the objects. Changes in the view-point occur when the camera observation point is changed, and this will be reflected in the image's visual structure. Scale variants refer to changes in the size of scene objects, e.g. different sizes of people within a scene.

Inter-class correlations. The inter-class correlation in scene categories, especially for indoor scenes, is very high. In some cases, even a human observer of a scene is not able to distinguish between classes, e.g. *street* vs. *inside city*. This similarity of scene categories is the result of the common objects within scene classes.

1.2 Topics and Approaches of this thesis

The following problems are addressed in this thesis:

1.2.1 Scene Recognition by Semantic Visual Words [1, 2]

In this chapter we propose a novel approach to introduce semantic relations into the *bag-of-words* framework. We use the latent semantic models, such as LSA and pLSA, in order to define semantically-rich features and embed the visual features into a semantic space. The semantic features used in LSA technique are derived from the low-rank approximation of word-document occurrence matrix by SVD. Similarly, by using the pLSA approach, the topic-specific distributions of words can be considered dimensions of a concept space. In the proposed space, the distances between words represent the semantic distances which are used for constructing a discriminative and semantically meaningful vocabulary. We have tested our approach on the *15-Scene* database and have achieved promising results.

1.2.2 Local-Global Feature Fusion (LGFF) for Scene Recognition [3]

In the field of scene recognition, it is usually insufficient to use only one visual feature regardless of how discriminative the feature is. In this chapter we propose an innovative method to integrate global and local feature space into a map function based on a novel distance function. This embedding is performed based on a subset of training images denoted as an exemplar-set. This exemplar-set is composed of images that describe better the scene category's attributes than the other images. The proposed framework learns a weighted combination of local semantic topics as well as global information, where the weights represent the features' contributions in each scene category. An empirical study has been performed on the 15-Scene and 67-Indoor Scenes datasets in order to demonstrate the impact of appropriately incorporating both local and global information for the purpose of scene recognition. The experiments show promising results.

1.2.3 Capturing Image Spatial Information in Scene Recognition Systems [4, 5]

Capturing the spatial location of local features has an essential role in establishing a reliable visual recognition system, especially in scene recognition systems, the scene understanding varies when the position of objects in the scene is changed. In Chapter 3,

we proposed a framework to convey semantic relationships into the *bag-of-words* framework. The proposed method embeds visual words into a new space called the *concept space*. In Section 5.2, we also considered local feature positions to impose spatial information on this framework. To this end, the image is divided into multi-resolution regions and the spatial pyramids are built over the feature distributions of the semantic visual words. This method resulted in improvements in the results up to 5%. In Chapter 4, we proposed a novel framework termed Local-Global Feature Fusion (LGFF) to project the image contextual feature space with the semantic space of local features into a map function. To have a robust scene recognition system, besides the scene’s contextual and semantic information, the spatial locations of the local features need to be captured. Therefore, in Section 5.3 the LGFF method was generalized into a framework that embeds spatial information into the method. An empirical study was performed on two of the most challenging scene datasets, the *15-Scene* categories and *67-Indoor Scenes*, and the promising results of **89.47%** and **45.0%** were achieved respectively.

1.2.4 Scene Datasets used in this dissertation

To demonstrate the efficiency of our feature fusion method, we have evaluated our method on two challenging scene datasets: the *15-Scene* dataset and the *MIT 67-Indoor Scenes*.

15-Scene dataset This popular database consists of both indoor and outdoor classes (see Fig. 1.1).

It contains 13 categories first reported in [22] and 2 more categories were added to it by Lazebnik et al. [9]. Eight of the categories are common with the Oliva and Torralba outdoor scene dataset [16] but in gray scale. In this dataset, the image sizes range from 200 to 400 pixels in width and height; the average size is around 300×250 , the number of images in each category is between 200 to 400. To test the proposed framework using this dataset, we randomly selected 100 images per category as training images and the rest for testing. The results are averaged over five times random splitting of training and testing images.

MIT 67-Indoor Scenes This dataset is a challenging scene dataset (see Fig. 1.2) that contains 67 different indoor scene classes, this dataset is introduced by Quanttoni

Figure 1.1: Sample images from the 15-Scene dataset.

and Torralba [7]. It is divided into five main categories stores (e.g. mall, book store), home (e.g. livingroom, lobby), public spaces (e.g. library, subway), leisure (e.g. bar, gym) and working place (e.g. office, classroom). The scenes are diverse in term of best characterized by their global layout (e.g. corridor) or by their containing objects (e.g. bedroom).

To conduct the experiments on this dataset, we used the exact same partitions as in [7], namely 80 images for training and 20 images for testing, so that the results are directly comparable.

1.3 Outline of Thesis

The subsequent chapters of this thesis are organized as follows. Chapter 2 explains the common concepts of every content-based visual system and reviews the main steps of

Figure 1.2: Sample images from the 67-Indoor Scenes categories dataset.

these systems, i.e. *feature extraction*, *similarity measures* and *semantic gap reduction*, and then overviews recent scene recognition methods.

In Chapter 3 [1, 2], a method is proposed to overcome the shortcomings of the *bag-of-words* framework with regards to capturing the semantic relations of local features. The proposed method projects the visual words into the concept space and, by co-clustering the conceptually close words together, builds a semantic visual vocabulary.

In Chapter 4 [3], we propose a Local-Global Feature Fusion method that locally weights the semantic topics of an image and incorporates the global image representation with this weighted combination. The LGFF method performs this feature fusion by learning a category-dependent adoptive mapping technique.

Since capturing the absolute positions of local features improves the performance in scene recognition, Chapter 5 [4, 5] presents novel techniques to impose spatial information into the semantic vocabulary framework in Chapter 3 and the LGFF method in Chapter 4. The discriminative power is substantially improved.

Finally, we conclude the work in Chapter 6 and give directions for future research.

Chapter 2

Literature Review

Scene classification, which is an important task of image indexing and retrieval, is among the most central problems in the field of content-based image classification. Content-Based Image Retrieval (CBIR) approaches were first proposed in the early 1990s. These systems include a wide range of techniques and applications that work intuitively with the visual content of images instead of working with meta-data. Meta-data refers to manually associated *tags* and *keywords* with the images. Although the objective of CBIR systems and content-based image classification systems is different, many of the CBIR techniques are applicable to object and scene recognition. All of the content-based visual systems follow three main steps:

- (i) Low-level feature extraction
- (ii) Similarity measures
- (iii) Semantic gap reduction

In this section we first describe the CBIR systems, then we review the content-based systems' main steps in detail and lastly we give a comprehensive review of the methods and approaches for scene recognition systems. The parts of this survey regarding CBIR systems are basically according to paper [10].

2.1 Content-based Image Retrieval

The term *content based image retrieval* was used for the first time in 1992 by Hirata and Kato [23] and has since become widely used in the literature. In general, the features for

CBIR systems should be extracted automatically from the image's content and should not be described by any textual keywords; otherwise, the principle of content based image retrieval systems is violated. Therefore, the main difference between primitive retrieval systems and CBIR systems is that CBIR systems work on databases that only include raw data images; this data does not provide us with any description regarding the image semantics before extracting the visual features. Eakins and Graham [24] have identified three different levels of query retrieval for CBIR systems:

Level 1: Query by example (retrieve similar images to the query image). At this level, retrieval is based on low-level features.

Level 2: Query by object (retrieve images that contain a particular object). At this level, semantic parameters interfere since the image features must be used to specify predefined objects.

Level 3: Query by event (retrieve images that contain a specified concept). The retrieval is based on abstract attributes.

In general, the low-level features do not provide (from the human perspective) any abstract concepts or ideas. *Level 1* is exclusively based on these features and therefore there is a semantic gap between this level and *levels 2 and 3*. In fact, these two levels can be referred to as semantic content-based retrieval. Most primitive CBIR systems belong to *level 1*, but recently CBIR systems have been developed at *level 2*. Creating CBIR systems at *level 3* is not very common, although the ideas of these level 3 systems can be found in some domain-specific areas such as newspaper or art retrieval systems. According to Liu et al. [10], *level 2 and 3* systems incorporate three phases: Phase 1, which is low-level feature extraction, Phase 2, which is an image similarity measure, and Phase 3, which is the reduction of semantic gaps between low-level features and image semantics. Phase 1 is the basis of all CBIR systems working at any level; Phases 2 and 3 are used to obtain relations between high-level semantic-based features and low-level appearance-based features. The objective is to fill the semantic gap between level 1 and level 2. A semantic gap may also exist between *levels 2 and 3* (relations between object-based features and very high-level features determining the overall perception of images). It should be noted that high-level features are usually described by keywords (mostly words from natural language) and can be referred to as concepts.

2.2 Low-level Feature Extraction

Low-level feature extraction is the most important part of any content-based visual system and it determines the performance of all subsequent operations. When low-level visual features are extracted from the sub-blocks of the image, we can define a special category of content-based systems referred to as region-based systems. Currently, researchers tend to prefer region-based systems due to their higher accuracy. Global features such as texture, color and shape attributes are sensitive to background changes, occlusion and cluttering; therefore, the use of local features that are less sensitive to background changes and occlusion is encouraged [25].

There are two main steps in extracting visual local features, namely feature detection followed by feature description. The first step extracts the regions of interests. Next, the appropriate feature descriptor is applied to the extracted regions.

2.2.1 Feature Detection

We can group the region extraction methods into the following categories:

1. Region selection by image segmentation.
2. Sliding windows.
3. Using regular grid (dense features).
4. Random Sampling.
5. Interest point detectors

1. **Image Segmentation** [10]

There are plenty of algorithms for automatic image segmentation. The regions extracted by these algorithms are used as regions of interest, i.e. as the input of feature description steps. The primary algorithms of image segmentation are based on curvature [26], energy diffusion [27] and graphs [28]. There are also several approaches focusing on segmentation by color properties which show very promising results for images with homogeneous color regions [29]. Texture-based segmentation of images where texture information is important, such as scene images, is difficult to perform.

Deng and Manjunath [30] have proposed a popular segmentation algorithm termed *JSEG*. In *JSEG*, a quantization process is first done to image colors, then the image pixels are mapped to appropriate classes and finally a spatial segmentation is applied to the performed mapping. The process develops regions that are considered homogeneous color-texture regions. Another approach suitable for segmenting image objects is K-means with Connectivity Constraint (KMCC) [31]. In this method, we use spatial properties to define new centers for the k-means algorithm and apply the component labeling operation. The reason why image segments are usually not used as regions of interest is that extracting meaningful image segments is a very difficult task and requires good knowledge of image contents.

2. Sliding Windows [32]

Sliding windows have been used for object recognition, but face detection is one of the most popular applications. The strategy of sliding windows is to select sub-parts of images at each possible scale and possible location. Sliding windows have the advantage of being less sensitive to the background, but it cannot solve the problem of partial occlusion and cluttering. In addition, they are not completely invariant to geometric transformations. However, the largest disadvantage of sliding windows is that a very large number of features are produced.

3. Dense Features [32]

Dense sampling over a fixed grid has been suggested as a method to overcome the drawbacks of sliding windows [9, 22, 33, 34]. Fixed grids partially address the issue of geometric invariance because of the use of dense sampling over feasible locations and scales. The regular grid of patches can overcome some deformations in the image and has shown very promising results in scene recognition [22].

4. Random Sampling [32]

It is also possible to use randomly sampled patches instead of regular grids of patches. The advantage of random sampling is its flexibility in the number of patches. For example, in [35] random sized sub-windows were used at random positions in the image. This method has also shown satisfying results in scene classification.

5. Interest point detectors [32]

This group extracts local invariant features. According to [32], we can categorize these detectors according to the structure of the feature they extract from the image, i.e. corner detectors, blob detectors or region detectors.

Corner detectors. In this concept the word corner does not refer to real corners; rather, it simply means the points in 2D space with a noticeable curvature that might be invariant under view changes. Corner detectors show promising results for localization within an image's plane. In various surveys and evaluations, the Harris detector [36] has been found to be a very informative detector.

Although the Harris detector itself is just rotation invariant and is not affine or scale-invariant, there are some extensions of this detector that are scale invariant (e.g. Harris-Laplace) and affine-invariant, as suggested in [37]. Corners are usually found in the boundaries of objects (due to intensity changes in those areas), therefore corner extraction works on the non-planar parts of images and backgrounds as well. This may cause the problem of sensitivity to view point and background changes. One suggested way to solve this problem is to use detectors that look for Edge Boundary Regions (EBR) in the images. Then, the non-planar parts are excluded in some cases and a better level of invariance is achieved [38]. The Features from Accelerated Segment Test (FAST) has been proposed in [39, 40] as another feature detector which is built based on SUSAN corner detector. It uses decision trees for evaluating pixel intensities in the suspected corner areas.

Blob detectors. These detectors are widely used in various applications. They can be considered complementary to corner detectors. If used together, they can cover most of the image content properties. In contrast to corner detectors, the localization accuracy of blob detectors is poor. Corner detectors localize features by point coordinates, while for blob detectors localization is based on boundaries. However, the local shapes and scales of images are much better defined by blob detectors. Thus, depending on the application, we can choose the appropriate detector. For example, in object recognition systems, localization is not critical but shape detection is important, thus it is better to apply blob detectors.

The Hessian detector [37] is a blob detector that is derivative-based and not scale or affine invariant, but its scale (Hessian-Laplace) and affine-invariant (Hessian-Affine) [37] extensions have been already proposed. The salient region detector proposed by Kadir and Brady [41] is another kind of blob detector based on intensity probability distribution entropy. Difference-of-Gaussian (DoG) [6] and Speeded-Up Robust Features (SURF) [42] are also categorized under blob detectors although they still consider computational efficiency as a parameter for feature detection. DoG is proposed by Lowe performs Laplacian approximation using scale-space pyramids by applying difference of Gaussian filters. SURF approximates Hessian matrices by integrating images.

Region detectors. These detectors either directly or indirectly extract homogeneous regions in an image. They can also detect blob structures from images, but are not limited to blob features. Thus, a combination of region detectors with corner detectors provides a complementary structural detection of image features. The following are two very popular types of region detectors: the Intensity-Based Regions (IBR) detector [43, 44] and the Maximally Stable Extremal Region (MSER) detector [45]. Both of these methods are scale and affine invariant. As IBR and MSER extract similar structures, they are competing rather than complementary detectors. IBR begins with determined intensity extrema in multiple scales and then it finds arbitrary shaped regions by exploring the image around the extrema. MSER uses the concept of intensity thresholding: it finds connected image components where the pixels within the component have higher (matter of brightness) or lower (matter of darkness) intensity than the outer pixels.

2.2.2 Feature Description

There are several local descriptors that can be applied on the regions found by feature detectors. The simplest descriptor is the vector of image pixels where cross-correlation is used to measure the similarity between two regions. This descriptor is mostly used for pair-wise matching. Using the vector of image pixels as a descriptor has several disadvantages. One of them is high computational complexity due to high dimensionality. Mikolajczyk and Schmid [46] divided feature descriptors into the following categories:

1. Distribution based descriptors.

This type of descriptor is defined by a histogram of pixel intensities distribution, e.g. Johnson and Hebert [47] used pixel location in the neighborhood of interest points. This descriptor works in 3-D space.

2. Non-parametric transformations.

The approach proposed in [48] relies on information about relations and ordering between data. The strong point of this approach is its robustness to illumination changes.

3. Spatial-frequency techniques.

Fourier transform decomposition is used to describe the frequencies of image content. The drawback of this decomposition is that the spatial relations between points are not well defined. Gabor features [49] have been proposed to overcome this drawback, but they also have the problem of high dimensionality with regards to the large number of Gabor filters needed to capture changes in frequency and orientation. Gabor features are mainly used in texture classification.

4. Differential descriptors.

These descriptors approximate a points neighborhood by calculating a series of image derivatives (local jets). Examples of these descriptors are differential invariants, complex filters and steerable filters. Differential invariants combine local jet components to obtain the rotation invariance [50]. The properties of local jets have already been studied in [51].

Schaffalitzky and Zisserman [52] proposed complex filters that use Gaussian derivatives by applying polynomials. Steerable filters were developed in [53]; steer derivatives provide the local jet components in a particular direction (direction of gradient), which makes them invariant to rotations.

5. SIFT descriptor.

The Scale Invariant Feature Transform (SIFT) descriptor was proposed by Lowe [6, 54]. SIFT is one of the strongest and most accurate descriptors due to its scale and orientation invariance. It is also partly invariant to affine distortion, illumination,

3D viewpoint changes and additive noise. The SIFT descriptor is applied to the key points obtained by a scale invariant feature detector. The detector applies Gaussian kernel G to image I at different scales (τ) to produce the scale space image L as:

$$L(x, y, \tau) = G(x, y, \tau) * I(x, y), \quad (2.1)$$

where $G(x, y, \tau) = \frac{1}{2\pi\sigma^2}e^{-(x^2+y^2)/2\sigma^2}$. The DoG (Difference of Gaussian) is computed between each consecutive scale space image as:

$$D(x, y, \tau) = L(x, y, k\tau) - L(x, y, \tau). \quad (2.2)$$

The local maxima for DoG is computed in a neighborhood consisting of spatial and scale space. The candidate keypoints are further pruned considering the contrast and edges. For the remaining keypoints, the gradient magnitude $m(x, y)$ and orientation $\theta(x, y)$ are computed as:

$$m(x, y) = \sqrt{(L(x+1, y) - L(x-1, y))^2 + (L(x, y+1) - L(x, y-1))^2}, \quad (2.3)$$

$$\theta(x, y) = \arctan((L(x+1, y) - L(x-1, y))/(L(x, y+1) - L(x, y-1))). \quad (2.4)$$

A 3D histogram of the gradients and orientations is applied to the regions extracted around the keypoints. The histogram bins of this gradient distribution are weighted by the gradient magnitude.

The region extracted by a detector should be circularly normalized to an axis aligned region. This process is based on finding the dominant orientation of the gradient. Afterward, a grid of 4×4 is defined on this region and each of the grid cells is represented by an 8-bin histogram of gradients. Therefore, the SIFT descriptor is a vector of 128 orientations. The feature description process is illustrated in Figure 2.1. The gradients are defined on the normalized patch of the region (left). Next, a 4×4 grid is defined on the normalized region. A 2×2 sub-array of this grid is shown on the right side of Figure 2.1 (each cell is defined by a vector of its 8 directional gradients).

Figure 2.1: SIFT descriptor applied to a patch [6].

2.3 Similarity Measure

There is a two-level approach to measuring the distance between two images in region-based systems. The first level is to measure the distance between the regions of both images, while the second level is to measure the overall distance between the images. Diversified metrics (distance measures) can be used to measure the distance between two regions. Many popular metrics exist, e.g. Canberra distance, angular distance, Czekanowski coefficient [55], inner product, dice coefficient, cosine coefficient, Jacard coefficient [56], etc., but the most common metric is the Minkowski distance.

If we consider two D -dimensional vectors A and B , i.e. $A = (a_1, a_2, \dots, a_D)$ and $B = (b_1, b_2, \dots, b_D)$, the Minkowski distance is defined as follows:

$$d(A, B) = \left(\sum_{i=1}^D |a_i - b_i|^r \right)^{1/r}, r \geq 1. \quad (2.5)$$

In case $r = 2$, it is equal to the Euclidean distance which is the most popular distance metric. When $r = 1$, it becomes the Manhattan distance.

Another metric is the weighted Minkowski distance which selects the more important dimensions by appropriate weighting. The weighted Minkowski distance is defined as follows :

$$d(A, B) = \left(\sum_{i=1}^D w_i |a_i - b_i|^r \right)^{1/r}, w_i \geq 1. \quad (2.6)$$

Despite its wide usage for measuring similarity of regions, the Minkowski distance is not a perfect metric from the perspective of perceptual similarity. There have been several attempts to resolve this problem [57, 58], but there is still no flexible and generally applicable approach to substitute the Minkowski distance.

For measuring level 2 similarity, there are approaches based on two general methods:

1. One-to-one region match:

According to [59], to find the overall similarity of two images we should find the best region match in the query image for each region of the target image. When each target image region has a match in the query image, the similarity of two images is a sum of the region similarities (additionally we may consider a weight parameter based on the region size).

2. Many-to-many region match:

Many-to-many matching means that each region from the query image can have more than one match in the target image and vice versa. This kind of matching is much more robust to inaccurate segmentation results (if the matched regions are formed by segmentation) and to multiple copies of similar image fragments. There are two metrics that can be used for image matching based on many-to-many region matching, namely the Earth Mover Distance (EMD) and Integrated Region Matching (IRM). Earth Mover Distance [60] has been widely used for measuring image similarity. Originally, EMD was proposed to evaluate the dissimilarity of two multidimensional distributions; the method concentrates on finding the minimal cost to transform one distribution to another based on the linear optimization of the transportation problem (the original transportation problem was to find the minimum cost of turning one pile into another, where the cost is the amount of dirt multiplied by the distance moved). The distance can be used as a similarity measure for region-based systems because it matches perceptual similarity well and it has the ability to handle variable-length distributions (different numbers of regions in the matched images). Another similarity measure that can be used for many-to-many match is Integrated Region Matching (IRM) proposed by Li et al. [61]. The process begins by trying to find a region-to-region match, taking into account the fact that the region of one image can

be matched to several regions of the other image. The strategy is to match the most similar regions first. After finishing the region-to-region matching, the overall similarity of the two images is calculated by a weighted sum of the distances between the matched pairs.

2.4 Semantic Gap Reduction

The basis of any content-based visual system is to reduce the semantic gap between low-level features and high-level concepts.

Jing et al. [62] proposed region-based systems and Relevance Feedback (RF) as the two main methodologies for reducing the semantic gap in content-based systems. Liu et al. [10] have suggested five methodologies (including the RF methodology) that are implemented by the following methods:

- Applying object ontology;
- Applying learning methods;
- Considering relevance feedback;
- Generating semantic templates;
- Using information from two sources, i.e. image visual contents and the available text information.

The last approach is only applicable for web-based applications and, as such, it does not fit in this report. The other methods are explained below. In addition we briefly discuss the advantage of region-based systems for semantic gap reduction.

2.4.1 Region-based Systems for Semantic Gap Reduction

As mentioned earlier, region-based systems use local features instead of global features to capture image properties. The main disadvantage of global features in visual systems is the lack of information regarding the objects within the image. In region-based systems, instead of working at the pixel level and defining descriptors for the whole image, each region is described separately so that we are one level closer to the visual concepts.

2.4.2 Object Ontology

Let us introduce three different levels of image description, i.e. the *feature space*, the *intermediate space* and the *concept space*. As a simple example, consider the concept *sky*. This concept can be explained in *intermediate space* by a light blue (color), a uniform region (texture) and an upper location (spatial attribute). The object ontology technique is based on the quantization of region properties such as color, texture, shape, size and position (i.e. local features) to form nodes (the intermediate space objects). Color and texture are usually the key properties of nodes. There are several CBIR systems that use color properties for defining the object ontology [31, 63, 64]. Texture properties are not very well modeled yet and still require intensive research; therefore, color-based systems are more efficient than texture-based systems.

2.4.3 Learning Methods

Both supervised and unsupervised learning can be used to develop high-level semantic concepts. In supervised learning, we are able to bring some level of semantics to the system by labeling the training images with semantic categories; consequently the test input image is categorized according to the training data. Unsupervised learning can introduce semantic concepts by exploiting the way input images are clustered. Since object and scene recognition are important subjects in image retrieval and image annotation applications, numerous supervised and unsupervised object and scene classification approaches have been proposed for semantic-based image retrieval [65–67].

2.4.4 Relevance Feedback (RF)

This method is an on-line technique that is based on the end-user's intentions to bring some feedback to the system. A typical scenario for RF in a CBIR system is shown in Figure 2.2. We can explain the figure as follows [68]:

- (i) For a given query, there are initial retrieval results.
- (ii) The end-user gives feedback on the degree of satisfaction by choosing the relevant and irrelevant retrieved images.

Figure 2.2: Relevance Feedback workflow.

(iii) A learning algorithm is applied to the CBIR system according to the user feedback.

The last two steps are repeated until all of the user feedback is positive.

2.4.5 Semantic Templates (ST)

Semantic Visual Templates (SVT) bridge the semantic gap between low-level features and high-level semantics. They can be considered as concept models composed from a set of scenes/objects and icons. Chang et al. [69] have introduced several SVTs; each of them models the sample database input images from a personalized view of concepts such as meetings, sunsets and so on so that each template represents a different view of the concept space. The system works in an interactive manner. First, the user proposes a template for a selected concept by weighting the features and icons in that template. Next, this query scenario will pass through the system and the best match template for

the corresponding concept will be trained according to the human feedback. This method seems like a very promising approach, but it is not widely used yet.

2.5 Scene recognition techniques

In this section we review some recently proposed methods in scene recognition [21, 70, 71]. Scene categorization tasks attracted the attention of researchers in the 1990s with applications such as dividing landscapes from city views [72] and indoor scenes from outdoor scenes [73]. These primary scene categorization tasks use the techniques that carry out object detection before performing the actual scene classification. According to the literature regarding human understanding of images, human observers capture a holistic view of the scene in a glance [12]. Subsequent approaches inspired by human perception capture the scene context bypassing individual objects within the image.

A variety of visual features are utilized for scene representation, and the methods applied for scene recognition depend on the way the scene is perceived. The early global feature based systems for scene recognition used mostly texture, color [72, 74, 75] and image derivatives [76]. Since global features are not very well generalized beyond the examples of training set, they are not reliable [77].

The early local feature-based systems used for scene classification easily divided images into sub-regions and extracted low-level visual features from these sub-blocks. Afterwards, the classifiers were applied to each block and eventually the results were collected via a voting approach [73] or a combination of probabilistic classifiers [78]. There are several other methods that initially segmented an image into semantic local regions [79, 80]. The obvious drawback of these methods is their high time complexity due to the segmentation step. The good performance of these methods is dependent to good performance of the segmentation application. Using local features which are invariance to scale and orientation changes (e.g. the SIFT [6, 54] visual feature), leads to a very high performance in scene recognition systems.

One of the robust approaches in scene categorization when local features are applied is *bag-of-words* framework [81]. In this approach, the images are quantized based on a visual vocabulary. The main drawbacks of this method are:

- The semantic relations between features are not considered.
- This method does not include spatial information.

There are some approaches [22, 82, 83] consider each image as a mixture of semantic hidden variables using either pLSA or the Latent Dirichlet Allocation (LDA) [84] model, or build a semantic vocabulary by co-clustering the visual words [1, 85, 86].

There are also methods that capture the spatial information of features, the most effective method so far being the Spatial Pyramid Matching (SPM) by Lazebnik et al. [9]. This method divides the image into regions with several resolution levels and the histogram of visual words is computed over the resulted sub-regions. Frameworks have been developed to incorporate SPM and better classification accuracy was reported [8, 87, 88]. Bosch et al. [8] add position information into their pLSA framework using three different approaches, xy-pLSA, ABS-pLSA and SP-pLSA. In xy-pLSA position information is concatenated to feature vector. ABS-pLSA quantize location information into one of the embedded bins for feature positions and SP-pLSA convey location information by incorporation SPM. Highest results achieved in [8] for SP-pLSA. Kwitt et al. [88] proposed a novel architecture for scene recognition, this architecture named Spatial Pyramid Matching on the Semantic Manifold (SPMSM) and is established by adopting the SPM on a statistical semantic manifold.

To capture an image's contextual layout, the gist global feature [16] was proposed. The gist descriptor performs Discrete Fourier Transform (DFT) to capture an image's spectral information. To compress these spectral signals, a discrete Principle Component Analysis (PCA) called the Karhunen-Loeve Transform (KLT) [70] is applied. The gist descriptor results in satisfactory results for outdoor scene categories. Oliva and Torralba [89] proved that this global descriptor is able to estimate many scene attributes including *naturalness*, *openness*, *roughness* and *ruggedness*. However, this method does not perform well for indoor scene classifications where the degree of *naturalness*, *openness*, *roughness* and *ruggedness* is very low [87].

Wu and Rehg [87] proposed a new descriptor called CENsus TRansform hISTograms (CENTRIST) for scene recognition to capture an image's spatial structure. This descriptor captures the shape of the scene. CENTRIST can be applied to the whole image

and used as a global feature. There is also a spatial representation of CENTRIST called the *sPACT* descriptor which is constructed by applying the CENTRIST to image blocks and construct the spatial pyramids over them. The advantage of CENTRIST (when it is applied as a local descriptor on image blocks) is that, unlike SIFT, CENTRIST is not sensitive to irrelevant textual details [70].

There are approaches that try to deal with scene recognition challenges by imposing high level concepts [7, 90–94]. In [90], the images are represented by a vocabulary of objects called *object-bank*, despite the high computational complexity, the method does not offer much increase in recognition accuracy. Li and Gua [91] improve scene recognition performance by capturing objects’ co-occurrence and their geometric correlations. They build a three levels (super-pixel, object, scene) hierarchical model, the operation in all of the levels is performed automatically. The proposed methods [7, 92–94], model the scene by the regions it contains. In [7] the images are first segmented, after which the ROIs and global information are combined together in a prototype-based model. Since every single parameter in their learning model is based on the prototype images, an excessive number of parameters has to be optimized, which increases computational complexity. Pandey and Lazebnik [92] suggested using Deformable Part-based Models (DPM) with latent SVM [95]. By applying DPM on scene images, salient objects and visual elements of scenes are captured. Afterwards, DPMs and global features are combined together to achieve more promising results. Reconfigurable models [93] preform the scene modeling by considering the scene as a collection of latent parts. The parts (pre-defined regions) are represented by bag-of-words model or spatial bag-of-words. To learn the model parameters, generative and discriminative approaches are used. The generative model employs an expectation-maximization (EM) algorithm and discriminative approach uses latent SVM classifier. In [94] a completely unsupervised method is proposed to build a classifier on discriminative mid-level patches. These are the visual patches that frequently appears in visual words. To find these mid-level patches an iterative approach which alternates between clustering large number of unlabeled visual patches and training linear SVM classifiers is performed. There are methods which label different regions of the scene based on its basic object ontology such as [96, 97] which use Markov Random Field framework for scene parsing.

The proposed scene recognition frameworks can be divided into global feature-based approaches [16, 87] and local feature-based approaches [9, 22, 82, 83, 85, 86]. Global feature-based approaches take the scene image as a whole, extract low-level visual features from either the entire scene or from finer regions, and then build a feature histogram by concatenating them.

Local feature-based approaches model the scene image by the distribution of its low-level features over visual components, e.g. visual words, objects and semantic topics. Local features may be very useful for classifying scene images, but the classification accuracy can drop dramatically when key objects are insufficiently large or if the objects are occluded. Previous literature has demonstrated that exploiting the semantic relationships of local features [22, 98] and their spatial locations [8, 9] can lead to promising results, but they inadvertently affect the classification of indoor scenes where the interclass variation is very high. In order to address this issue, the global or contextual information in the images must be used since it can capture the inherent context.

There are also methods which combine additional information such as texture and color to the features e.g. Multi-spectral SIFT [99] which concatenates SIFT descriptors from R, G and B channels to increase feature efficiency. Other works [100–102] suggested feature fusion by Multiple Kernel Learning [103], to improve the recognition accuracy in scene classification. In these works, various visual features are combined together and a separate kernel is associated for each feature.

When looking at the rest of this thesis, there are three main contributions to improving the performance of scene recognition systems. In the next chapter, the proposed method exploits the semantic relationships of the scenes local features in order to generate a compact codebook that carries semantic information of the scene.

Chapter 3

Scene Recognition by Semantic Visual Words

Parts of this chapter have been published in [1, 2].

3.1 Introduction

The *bag-of-words* (BOW)¹ framework has been shown to be useful in various computer vision applications like object recognition [81], scene recognition [22]. The framework builds a visual vocabulary by vector quantization of raw features extracted from local image patches. The vector quantization essentially involves clustering of the raw features by k -means and choosing a cluster's mean as the codebook or visual word. An unknown image or video is classified by a suitable classifier, according to its histogram of visual words. It is well known that the choice of the number of clusters in k -means clustering determines the discriminative ability of the visual vocabulary that is generated [104]. However, the more important drawback of k -means clustering is that it is based on the appearance of the image or video as represented in the raw features, as opposed to being based on the semantic relations between features. Utilizing the semantics inherent in visual content improves image/video categorization and understanding.

The illustration of the standard *bag-of-work* framework has been shown in Figure 3.1. There have been several attempts to incorporate semantics into the BOW model so that a more discriminative visual vocabulary is realized. Generative methods use latent

¹BoW framework is also names in the literature *Bag-of-Visterns* (BoV, vistern means visual term/word) and *Bag-of-Keypoints* (which implies that interest point detection is applied)

Figure 3.1: Standard Bag-of-words Framework.

variable models like Probabilistic Latent Semantic Analysis (pLSA) [105, 106] and Latent Dirichlet Allocation (LDA) [22, 106] to obtain models for each category and subsequently to fit the query to one of the models in an unsupervised manner. Although these methods are efficient, their unsupervised nature limits their performance. Moreover, the number of topics in these methods is equal to the number of categories. This too limits their efficiency. Discriminative methods which incorporate label information have also been explored. Among the recent methods is the notable work of Liu and Shah which finds a semantic visual vocabulary via Maximization of Mutual Information (MMI) between visual words and images [85] or videos [107]. The algorithm starts with singleton clusters and in each iteration, merges two clusters which result in the minimum loss in mutual information. This procedure continues until a certain threshold in the information loss or in the number of clusters is achieved. This approach is effective in discovering the optimum number of clusters, but the formed clusters do not necessarily represent topics or synonym words which is required for constructing discriminative histograms.

Liu et al. [86] use Diffusion Map (DM) to construct a semantic visual vocabulary. Unlike geodesic distance which is based on the shortest path between points, diffusion distance considers all paths between two points to measure the shortest distance, and, hence, is not sensitive to noise. However, considering connectivity in measuring the semantic distance is not appropriate in the presence of polysemy.² For example assume that word B is a polyseme with two distinct meanings: 1 & 2. If word B is connected to word A based on meaning 1 (they both have the same meaning 1) and also word B is connected to word C based on meaning 2, then words A and C will be connected in the diffusion distance framework, but they convey different meanings. So, diffusion distance does not always represent semantic distance.

Considering these drawbacks, we propose a method for scene recognition based on a semantic visual vocabulary that uses latent aspect models to embed visual words into a rich semantic space which we call the concept space. Using LSA (Latent Semantic Analysis) or its probabilistic version pLSA, the synonym words which convey the same meanings are embedded close to each other so that they can be clustered together into the

²Polysemy is the existence of words which convey different concepts in different documents. For instance in text domain, the word *table* can either be interpreted as *a piece of furniture* or *an arrangement of data*.

same semantic cluster. The distance in the proposed concept space is actually based on the meanings of the words and, thus, it represents the semantic relations. Consequently, the formed histograms based on these semantic clusters are efficient and discriminative for scene recognition.

In contrast with generative methods that do not make use of category labels, our method trains a classifier using the histograms from the training set. Moreover, in our method the number of topics can be changed as opposed to the unsupervised framework where this number is fixed and equal to the number of classes. This will allow us to analyze the semantic relations in more details and to consider as many topics as appropriate. On the other hand, pLSA is able to handle polysemy which is very effective in cases when different categories share the same topics e.g. *office*, *livingroom* and *kitchen* categories include similar visual tokens such as chairs and tables.

3.2 Related Works

In this section, we review some methods that attempt to introduce semantic relations in the BOW framework for object and scene recognition. These approaches can be broadly divided into generative and discriminative methods. Generative methods usually involve hidden variables. These methods [22, 105] try to model each image as a mixture of hidden concepts using either pLSA or LDA. On the other hand, discriminative methods are only based on observed variables. These approaches usually incorporate a classifier. Among these methods, Vogel and Schiele [108] define a set of concept classes (visual words) like *sand*, *sky* and *sea* to label image regions. In this method, image regions are represented by a combination of color and texture features and classified into concept classes. Thus for each image, a concept occurrence vector is constructed and classified for scene retrieval. In labeling the databases containing ambiguous images, this approach claims that obtaining the ground truth for local semantic concepts is easier than for the whole image. However, this approach suffers from the large amount of manual work needed to annotate local regions. Randomized clustering forests have been used by Moosmann et al. [109] for image classification. Ensembles of decision trees are constructed based on the image class labels. Subsequently, visual words are assigned to each leaf. After building the

trees, a bottom-up pruning process is done to reach a threshold of number of leaves to control the codebook size. They use randomized forests for clustering and quantization. Randomized forests have also been used in [110] for object classification and segmentation. In their work decision trees are used directly on image pixels to save time for extracting descriptors. In contrast to [109], which uses forests for clustering only, they use forests for both clustering and classification purposes. Randomized clustering forest is fast, yet discriminative, when compared to conventional k -means clustering. However, it tends to overfit, especially when applied in noisy situations. Moreover the model is complex and it is hard to understand the relation between the predictor variables.

Quelhas et al. [82] have used the pLSA model to extract document-specific distributions of topics in order to represent images. This is followed by a SVM classifier in order to classify scenes. Bosch et al. [83] have also used a similar framework. Liu et al. [86] use diffusion distance to build a semantic dictionary. They construct a graph on the visual words in which the weights between points reflect the similarity. Visual words are represented by pointwise mutual information. By applying the diffusion map, points are embedded into a lower dimensional space in which Euclidean distance is equal to diffusion distance.

The approach proposed in this chapter is a discriminative embedding method that projects visual words into a concept space where the dimensions are discovered in an unsupervised manner by latent topic models. The differences between our framework and similar methods are threefold:

Using word space All of the methods that project images into latent semantic space by pLSA work in new *semantic document space*. This means that, in order to classify a test image, the image is projected into the semantic document space. However, the focus in our framework is the *semantic word space*. The similar visual words in the word space are co-clustered together to form a semantic visual vocabulary. In our framework, in order to classify a test image, it is directly quantized based on the semantic visual vocabulary without embedding the image.

Investigating the changes in the number of topics In the generative frameworks employing pLSA, the number of topics is considered the same as the number of

Figure 3.2: Constructing the semantic visual vocabulary.

categories. This is in contrast to our method which analyzes the changes in the number of topics and empirically fixes the best fit for representing the semantic relations in the scene.

Using LSA embedding Our framework applies both LSA and pLSA semantic embedding. To the best of our knowledge, there is no other method that uses LSA as semantic embedding although, according to our experiments, the results of LSA and pLSA are almost on par. However, the time and memory complexity for LSA are substantially less since it uses a simple Singular Value Decomposition (SVD) compared to pLSA which uses an expensive expectation maximization algorithm. This is the reason that using LSA is really advantageous.

3.3 Overview of the Proposed Framework

Figure 3.2 shows the flowchart for constructing the semantic visual vocabulary via embedding into concept space. We first extract features from patches in the images. The initial vocabulary is constructed by performing k -means clustering on the extracted features and choosing the cluster centers as the codewords. The feature vectors are quantized based on the initial codebook to form the word-image matrix which describes the occurrences of words in images. The codewords are then embedded into the concept space by latent semantic models - we demonstrate the embedding both by LSA as well as by pLSA. Finally, the embedded codewords in the concept space are again clustered using k -means to obtain the desired semantic visual vocabulary.

3.3.1 Feature Extraction

According to the literature [111], using a regular grid with dense sampling is more discriminative for scene recognition than using interest points. Densely sampled regular grid

of features capture all of the rigid objects within the image. Lazebnik [21] argues that the regular grids show promising results in scene recognition first because scene databases usually have very stable contents which should be fully captured by the camera but there is no need to extract geometrically invariant features. Secondly, inside each of the scene classes the images' variety is very high so that the simple local transformation invariance does not have a big influence on the scene classification accuracy. There is also one more important reason: none of the local regions except for dense features on a regular grid can capture the uniform areas like sky, water or road surface, which might be very important in scene recognition.

The features applied in these experiments use the patch of 16×16 size sampled densely with the $M=8$ pixel spacing. The feature descriptor applied on each patch is SIFT [6].

3.3.2 Term-document Matrix

In this section we use the same phrases as used in text mining. However, the ideas explained are applicable to image retrieval task as well. To prevent any ambiguity it should be noted that *document* and *term/word* in text retrieval correspond to *image* and *visual word* in image retrieval, respectively. The feature vectors are quantized based on the initial vocabulary to form the term-document matrix which describes the occurrences of terms in documents.

In the term-document matrix each of the elements (i,j) indicates the occurrence of i^{th} term in j^{th} document. Thus each row in the matrix corresponds to a single term and shows the frequency of this term in different documents. Similarly, each column indicates the relation of each document to different terms. This representation is illustrated in Figure 3.3.

As the number of words in longer documents is higher, the term occurrence for each document is normalized. Thus the term frequency $tf_{i,j}$ is defined as:

$$tf_{ij} = \frac{n_{i,j}}{\sum n_{k,j}}, \quad (3.1)$$

where $n_{i,j}$ is the number of i^{th} term in j^{th} document. The denominator is the total number of occurrences of all terms in j^{th} document.

Figure 3.3: $M \times N$ word/term document matrix.

There is also another representation for distribution of words and documents which is more usable. This representation is called term frequency-inverse document frequency (*tf-idf*). The motivation of using *tf-idf* instead of a simple *term-document* matrix is the fact that some of the words are very frequently used in documents. These words do not contribute important concepts. For instance the word *the* is a commonly used word in English literature. It is not a keyword to distinguish between relevant and non relevant documents and it's number of occurrences is very high. Considering this keyword for classification will decrease the impact of more discriminate keywords which occur rarely. Therefore, a factor named *inverse document frequency* is introduced which gives a proper weight based on the number of occurrences. It is also important to have a factor to measure the overall importance of the i^{th} term. The following formula can be used for this purpose:

$$idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|}. \quad (3.2)$$

In this formula the number of documents is divided by the number of documents which contain the i_{th} term. Finally, the *tf-idf* (*term frequency-inverse document frequency*) is defined as:

$$(tf-idf)_{i,j} = tf_{i,j} \times idf_i. \quad (3.3)$$

According to this formula, a high weight for *tf-idf* will be obtained when a term is occurring frequently in j^{th} document and infrequently in other documents. Also based on this formula, the commonly used words are filtered out.

3.4 Concept Space

We obtain the initial vocabulary by performing k -means clustering on the extracted visual features. This initial codebook forms a reasonable-sized set representing all features, but they are not semantically clustered, i.e. the features in a cluster may convey different concepts. Thus, the formed histograms will not be semantically discriminative for classification. Therefore, we need a space in which semantically related words are adjacent. In order to find such a space, we use latent semantic models that find the underlying latent semantics given the occurrence matrix of word-image. These models are the well known LSA and pLSA, which we briefly review in the following sections. We use *tf-idf* instead of the normal count in the occurrence matrix for a higher efficiency.

3.4.1 Embedding into Concept Space using Latent Semantic Analysis

LSA [112] finds a low-rank approximation for the word-document matrix. Here, document refers to the image. The word-document matrix itself delivers semantics since synonym words appear in similar documents resulting in similarities among their occurrence vectors. However, the original word-document matrix is noisy, sparse and large. Hence, the low-rank approximation to the original matrix is desirable. The consequence of this dimension reduction is that the dimensions relating to synonym words (e.g. *see* and *look* in text domain) are merged. It means the LSA hypothesis is that the synonym terms will have the same direction in the latent semantic space.

Let X be the occurrence matrix whose rows correspond to words and columns correspond to documents. Decomposing of X using SVD, i.e.

$$X = U\Sigma V^T \tag{3.4}$$

gives the orthogonal matrices U and V and the diagonal matrix Σ that contains the singular values of X . By Selecting the L largest singular values and their corresponding

Figure 3.4: SVD decomposition, word space and concept space.

singular vectors, we find the rank- L approximation of X by:

$$X \approx U_L \Sigma_L V_L^T. \quad (3.5)$$

This L -rank is optimal in a sense of l_2 matrix norm. The column vectors of U_L span the concept space of words and the columns of V_L span the concept space of documents, so we can consider $M \times L$ matrix U_L the word space and $L \times N$ matrix V_L the concept space. The SVD decomposition process is illustrated in Figure 3.4. The i^{th} row of X , t_i , describes the i^{th} word. Consequently, the i^{th} row of U_L is the description of the i^{th} word in the concept space with L concepts and we refer to it as $\hat{t}_i$. In fact, each of the L dimensions in the low dimensional vector $\hat{t}_i$ shows the projection of the words along one of the concepts. It is expected that synonym words are close in the concept space.

3.4.2 Embedding into Concept Space using Probabilistic Latent Semantic Analysis

pLSA [98] is the statistical version of LSA which defines a generative model on the data. It is assumed that there is a latent topic variable z_l associated with occurrence of the word w_i in the document d_j . It is expected that the joint probability $P(w_i, d_j, z_l)$ follow the form of the graphical model in Figure 3.5. The observed variables are w_i and d_j while z_l is latent. The probability of observation pair $P(w_i, d_j)$ is:

$$P(w_i, d_j) = P(w_i|d_j)P(d_j). \quad (3.6)$$

Since the occurrences of w_i and d_j are assumed to be independent, we can marginalize over latent topics z_l in order to find the conditional probability $P(w_i|d_j)$, i.e.

$$P(w_i|d_j) = \sum_{l=1}^L P(w_i|z_l)P(z_l|d_j), \quad (3.7)$$

Figure 3.5: Graphical view of pLSA model.

Figure 3.6: pLSA ,word space and concept space.

where $P(z_l|d_j)$ is the probability of occurrence of topic z_l in the document d_j and $P(w_i|z_l)$ is the probability of occurrence of word w_i given the topic z_l . L is the total number of latent topics. Equation 3.7 is a decomposition of the word-document matrix, similar to LSA, but with the condition that the values are normalized to be probability distributions. This equation can be explained using the illustration in Figure 3.6. We fit the model by determining $P(z_l|d_j)$ and $P(w_i|z_l)$ given the observation occurrence matrix. Maximum likelihood estimation of the parameters is performed using Expectation Maximization (EM) algorithm. Assuming a vocabulary of M words and N documents, the likelihood function to be maximized is:

$$\prod_{i=1}^M \prod_{j=1}^N P(w_i|d_j)^{n(w_i,d_j)}, \quad (3.8)$$

where $n(w_i, d_j)$ is the number of words w_i in the document d_j and $P(w_i|d_j)$ is obtained by Equation 3.7.

Expectation Maximization (EM) algorithm: EM algorithm is composed of two main steps:

Expectation (E) step: In this step, posterior probabilities are computed for the latent variables.

$$p(z|d, w) = \frac{P(z)P(d|z)P(w|z)}{\sum_{z' \in Z} P(z')P(d|z')P(w|z')} \quad (3.9)$$

Maximization (M) step: In this step, parameters are updated.

$$P(w|z) \propto \sum_{d \in D} n(d, w)P(z|d, w), \quad (3.10)$$

$$P(d|z) \propto \sum_{w \in W} n(d, w)P(z|d, w), \quad (3.11)$$

$$P(z) \propto \sum_{d \in D} \sum_{w \in W} n(d, w)P(z|d, w). \quad (3.12)$$

The original pLSA algorithm in the unsupervised learning framework tries to categorize the query document given the learned parameters [98]. However, we use the pLSA algorithm only to determine the probabilities $P(w_i|z_l)$. In fact $P(w_i|z_l)$ is equivalent to the l^{th} dimension of $\hat{t}_i$ in the LSA framework. Therefore, using pLSA we obtain the concept space embedded vector $\hat{t}_i$ as:

$$\hat{t}_i = \begin{bmatrix} p(w_i|z_1) & p(w_i|z_2) & \dots & p(w_i|z_L) \end{bmatrix}^T. \quad (3.13)$$

It should be noted here that L , the dimension of the concept space, does not need to be equal to the number of classes; this enables us to define arbitrary latent concepts. In fact the number of semantic topics can be much more than the number of classes. In other words, classes are wider concepts that may include some finer and more detailed concepts which are referred to as topics. For instance “computer” and “pen” are topics related to the class of office. Note that in the unsupervised framework in which pLSA is used (e.g, in [105]), the dimension of the concept space must be equal to the number of classes.

In LSA, each word is projected into a single point in the concept space so that each word can refer to a single meaning only. Instead pLSA is able to capture polysemy. Thus,

given a word w observed in two different documents d_i and d_j , the topics associated with the word in d_i and d_j can be different or, in other words, $\text{argmaxp}(z|d_i, w)$ can be different from $\text{argmaxp}(z|d_j, w)$ [98], [113]. The advantage of LSA compared to pLSA is the faster and easier implementation. LSA needs a simple SVD, while pLSA uses the iterative EM algorithm which is only guaranteed to find a local maximum of the likelihood function [98].

3.5 Experimental Results for Scene Recognition

This section describes the details of the experiments performed on the model presented in Section 3.4 and discusses the obtained results.

The *15-Scene* database is used to evaluate the proposed method. Support Vector Machine (SVM) with Histogram Intersection kernel (HIK) is used as the classifier. HIK [114] is used to measure the similarity between two visual feature histograms H_a and H_b ; it is defined as:

$$K_{HI}(H_a, H_b) = \sum_{i=1}^n \min(H_a(i), H_b(i)). \quad (3.14)$$

Odone et al. [115] imposed that HIK can be used as a kernel with support vector machine since it is positive definite. According to literature, applying HIK as SVM kernel function [9, 115], leads to very promising results.

The size of the initial vocabulary (K_i) is fixed to 1500 during the experiments.

3.5.1 Results and Analysis

One of the advantages of the proposed method is that it allows the number of topics to be varied, in contrast to pLSA using unsupervised framework (where the number of topics is constrained to be the same as the number of classes). Figure 3.7 and Figure 3.8 show how the number of topics L affects the recognition accuracy with LSA and pLSA as the embedding method. The experiments have been performed using three different semantic vocabulary sizes, K_f . As the number of topics is increased above $L = 15$ (which is the number of classes) the recognition rate increases since the increased number of topics enables a better discrimination between classes. However, after around $L = 70$ topics, the recognition accuracy decreases. This is mainly because adding more dimensions to

Figure 3.7: Performance of proposed method on 15-Scene dataset with different number of topics using LSA.

the concept space implies further division into semantic units that are not meaningful. The recognition accuracy has a variance of about 1% – 3% as L varies.

According to the experiments, the number of topics for obtaining the best accuracy for the three different values of K_f is the same. Thus, the number of topics is independent of the final vocabulary size.

More experiments have been conducted on the 15-Scene database using different initial vocabulary sizes. In these experiments, we kept the number of classes fixed and used the same size for the final vocabulary $K_f = 1000$. Three different initial vocabulary sizes ($K_i = 3000, 2000, 1500$) were used to determine the impact of changing the numbers of topics on the accuracy for different initial vocabulary sizes. Figure 3.9 shows that, for all K_i values, the best accuracy is achieved when the number of topics is equal to 70; hence, the number of topics is also independent of the initial vocabulary size. Since the number of topics cannot be affected by changes in the initial or final vocabulary size, we determine that the number of topics is dependent on the number of classes and number of visual concepts in the database.

To further explore this finding, we conducted an experiment to show the impact of changing the number of classes on the number of topics. In each run of this experiment, one of the parameters is kept constant. This was carried out using LSA. Figure 3.10 shows the impact of the number of topics (L) on the recognition accuracy when the

Figure 3.8: Performance of proposed method on 15-Scene dataset with different number of topics using pLSA for embedding.

Figure 3.9: Performance of proposed method on 15-Scene dataset with $K_f = 1000$ and $K_i = 1500, 2000, 3000$, with different number of topics using LSA.

Figure 3.10: Performance of proposed method on 15-Scene dataset divided into Indoor and Outdoor- classes with different number of topics using LSA.

number of visual concepts and the initial vocabulary size are constant. In this case, the entire 15-Scene database is divided into two different classes: indoor (*bedroom, industry, kitchen, living room, store, office*) and outdoor. It is clear that the best accuracy is obtained when the number of topics is $L = 30$, which supports our theory.

To verify the efficiency and discriminative ability of the method in scene classification, we have compared it with the classic BoW framework for different vocabulary sizes. The results are shown in Figure 3.11. The number of topics is chosen to be 70. According to the figure, our method outperforms the classic BOW framework in all cases. This shows the efficiency of the method proposed. Apart from the instabilities in the initial parts of the curves, we can say that the behaviors of LSA and pLSA are consistent. For small vocabulary sizes, pLSA outperforms LSA by a small margin due to its ability to handle polysemy. However, as vocabulary size increases, LSA performs better than pLSA (at approximately $Kf = 900$). Based on the differences between pLSA and LSA embedding, the possible cause of this effect is that the larger vocabulary size brings in more details and compensates for the effect of polysemy. However, pLSA takes into account every possible meaning of a word, even the rare ones, which results in confusion in larger vocabularies, thus reducing the accuracy. Also it should be noted that LSA has always a shorter

Figure 3.11: Comparison of results with the classic framework for different sizes of vocabulary in 15-Scene dataset.

Table 3.1: Comparison with recently reported results for 15-Scene dataset.

Method	Fei-Fei [22]	Quelhas [82]	Bosch [83]	Liu(DM) [86]	Liu(MMI) [85]	Ours
Accuracy(%)	65.2	71.24	73.4	74.9	75.16	79.22

implementation time compared to pLSA. This is due to a time-consuming iterative EM process for pLSA compared to the straightforward SVD in LSA.

The best result achieved on *15-Scene* dataset with our method is **79.22** using pLSA model and the semantic codebook size of 600. The confusion matrix for this accuracy is shown in Figure 3.12.

The recognition rate for outdoor scene classes (e.g. street, suburb, tall building and forest) is much higher than the indoor scenes (e.g. living room, kitchen, office, store). As mentioned in [7], this is because outdoor scenes can be clearly categorized using global features, but for indoor scenes global features are not enough for exact classification. For these categories of scenes, object identification might help in increasing the accuracy for scene recognition.

Table 3.1 summarizes recognition accuracy of our method and some notable related works. As seen from the table, the proposed method is the best. Just as in action recognition experiments, we have not used any spatial information in this method and

Figure 3.12: Confusion matrix for the best result achieved in 15-Scene dataset.

hence, we have not compared the results with approaches that use spatial information e.g. spatial pyramid matching [9]. The work of Fei-Fei and Perona [22] which has used LDA in a generative framework, has a lower performance compared to the methods using a semantic vocabulary but incorporating category labels like Liu (DM) [86] and Liu (MMI) [85]. Also the works of Quelhas et al. [82] and similarly Bosch et al. [83] have lower accuracy compared to the works using co-clustering to obtain semantic vocabulary like Liu (MMI) [107], Liu (DM) [86] and ours. This is mainly because in contrast to former methods, which use a histogram of topics equal to the number of categories, latter methods perform the clustering step in the semantic space to further group the semantically related words together and to construct more discriminative histograms with the actual number of topics.

To confirm the advantage of the semantic visual vocabulary compared to the original visual vocabulary, in Table 3.2 we present the results of applying both vocabularies to four different scene images. For clearer illustration, we have only shown a subset (40) of the visual words. Each color is associated with one visual word. As seen in the tables third column, which is associated with the semantic visual vocabulary, the areas are more uniform, especially in the marked regions by the rectangles in the scene image.

3.6 Chapter Summary

In this chapter, we have proposed a novel approach for using semantic relations in BoW framework. We have used the latent aspect models such as LSA and pLSA to map the visual words into a semantic space. Under the LSA framework this mapping is done by a low-rank decomposition of the word-document occurrence matrix using SVD. Also, the topic-specific distributions of words are considered (using pLSA) as the projections words onto different concepts. The distances in the proposed concept space reveal the semantic relations. Clustering is done in the concept space to capture the semantic structures. We tested our method on 15-Scene categories dataset. Our results confirm that the proposed features are efficient and discriminative for scene recognition. Also our method performs better compare to some similar methods for constructing semantic vocabularies. In the next chapter, the Local-Global Feature Fusion method is proposed to model scene complexities.

Table 3.2: Semantic visual words vs. Original visual words. The first column shows the scene images, the second column illustrates the results of applying the original visual vocabulary to the scene image and the third column illustrates the results of quantizing dense SIFT visual features based on the semantic visual words.

Scene image	Using original vocabulary	Using semantic Vocabulary

Chapter 4

Local-Global Feature Fusion for Scene Recognition

Parts of this chapter have been published in paper [3].

4.1 Introduction

In Chapter 3, we introduced a method to obtain a more discriminative representation of the scene images by quantizing the scene using a semantic codebook. Many proposed scene recognition frameworks have attempted to learn an optimal classifier based on low-level visual features, but these features alone are often not able to offer sufficient discriminative power. In this chapter, we describe how to perform scene recognition by combining global features with weighted semantic topics that are built on top of local features.

There have been several previous attempts at integrating contextual information with local features [7, 100, 101]. Quattoni and Torralba [7] combined image *Regions-of-Interest* (ROIs) and global gist features together in a prototype-based framework, where the ROIs were extracted through segmentation or manual annotation. A well-known method for feature fusion is Multiple Kernel Learning (MKL) [103]. This method was also implemented for scene recognition in [100, 101] using a number of visual features, such as multiple local features and global contextual features. The method involves having a separate kernel associated with each feature, after which a linear combination of these kernels is learned for SVM classification. Although MKL is an effective method for fea-

ture combination, computational cost is often very high. This is especially the case when the kernels are complex since the parameters for every kernel have to be tuned.

In this chapter we propose a novel distance learning method which combine Local-Global features into a single framework, in order to more directly exploit the advantages of both types of features. Opposed to Xiao et al. [101] where many different features are required to obtain higher accuracies in scene recognition, our proposed method uses fewer features and more effective fusion model to achieve robust scene recognition. Since feature extraction is one of the most time consuming steps in scene recognition, our method has advantage in this perspective.

Specifically, image local information is represented by a histogram of local semantic topics, after embedding the visual *bag-of-words* histograms [81] into a Probabilistic Latent Semantic Space (pLSA) [98], an exemplar-based Local-Global Feature Fusion map function is learned on each scene category.

The rest of the chapter is organized as follows. Section 4.2 provides a review of related work on scene recognition. The overview of our proposed model for integrating the visual features is described in Section 4.3. The image visual representations and models which we incorporate into our framework are detailed in Section 4.4. Section 4.5 explains the Local-Global feature fusion framework. The experimental evaluation and discussion are presented in Section 4.6. The final Section concludes the chapter.

4.2 Related Works

Utilizing multiple features in visual recognition systems has proved to increase the recognition accuracy. [7, 100, 101, 116–118]. In [7] for combining local and global features, the images are first segmented, after which the ROIs and global information are combined together in a prototype-based model. In this model the prototype roots and ROIs are represented by gist descriptor and spatial pyramids respectively. Since every single parameter in their learning model is based on the prototype images, a possibly excessive number of parameters has to be optimized, which increases computational complexity.

Canonical Correlation Analysis (CCA) and Kernel-Canonical Correlation Analysis (K -CCA) which is a non-linear extension of CCA are also employed for feature combination. CCA was originally defined for integrating two CCA was originally defined

for integrating two sets of variables [119]. Although it later generalized to work with multiple features. This leads to high complexity due to employing a pairwise correlation maximization for all the projected features. In CCA, the correlation of feature vectors is being considered. CCA applies a joint dimension reduction which project the features into a new space where the canonical correlation between the features is maximized. The feature fusion by canonical correlation is used either to combine various types of features in multimedia systems (e.g, video and text) [120] or to combine visual feature vectors such as local-global visual features [116, 117] or other kinds of visual features [118].

There are other works [100, 101] that carry out feature fusion by Multiple Kernel Learning [103] to improve the recognition accuracy in scene classification. In these works, various visual features are combined together and a separate kernel is used for each feature. These methods are very time consuming and the feature fusion is performed globally, therefore some local information are lost.

The contributions of this work in comparison to the other feature fusion methods are threefold:

Weighting latent semantic topics In contrast to the kernel combination methods that perform feature fusion by combining histogram representations of images without considering the importance of each histogram bin, our method weights each of the semantic topics individually. This differs from the previous pLSA-based methods [22, 82, 83] where all of the semantic topics had the same discriminative power. In our framework, semantic topics are weighted based on their significance level in identifying a particular category.

Exemplar-based distance learning A special subset of training images called the exemplar-set, which includes the more discriminate images in each category, is selected. We use the weighted *exemplar-set* as the basis to assign a scalar measure to each image. The scalar value shows the level of confidence in image x_i belonging to category C_i .

Category-dependent map function Besides the novelty of the proposed distance function, another contribution of this chapter is the category-dependent map function. In this map function, the degree of membership for each exemplar image E_i to the

Figure 4.1: Block diagram of learning step in Local-Global Feature Fusion method.

category C_i can be determined. To this end, the proposed map function embeds the learned distances.

4.3 Overview of the proposed framework

The objective of our scene recognition system is to learn a model which maps images x to scene categories y . The scene model presented in this paper uses a training set $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ to learn the model parameters. The training set T consists of n pairs of image x_i and label y_i , wherein each scene label $y_i \in \{-1, 1\}$ is a binary label indicating whether the image x_i belongs to a scene category or not. Based on a *one-versus-all* strategy, we compute a mapping function $f(x)$ (see Equation 4.6) for each scene category that quantitatively estimates the affinity of a test image x to that scene category. The scene category that best describes image x can simply be determined by the $f(x)$ that returns the largest value.

According to the learning model's block diagram in Figure 4.1 after extracting the visual features, to perform the scene recognition modeling, we retain a subset of training images called the *exemplar-set*: $E = \{E_1, E_2 \dots E_p\}$. The exemplar images are chosen for each category in the dataset. The selection is done such that for each category the most discriminative images are chosen. Exemplar-set selection is described in detail in Section 4.5.1. The exemplar-set is used as a basis to measure the distances. After defining appropriate distance functions, the exemplars again are used to embed the distances into their relevant map functions. Afterwards the functions parameters are tuned by performing a parameter learning.

4.4 Feature Extraction and Representation

In this chapter we use three kinds of low level features: **SIFT** [6], **Gist** [16] and **CENTRIST** [87].

To capture semantic representation of SIFT, initially the local feature topics have to be exploit. Images are divided into 16×16 patches with 50% overlapping both horizontally and vertically. SIFT and CENTRIST descriptors [6, 87] are extracted from each patch.

4.4.1 Capturing Semantic Information by Scene Modeling with Probabilistic Latent Semantic Analysis (pLSA)

The pLSA [98] is a statistical model which employs hidden variables together with a bag-of-words representation of the image. The pLSA models each image as a mixture of topics. It assumes there is an unobserved latent variable $z_k \in \{z_1, z_2, \dots, z_K\}$ which is associated with each observation (w_i, x_j) . The probability of co-occurrence w_i in x_j is represented by $P(w_i, x_j) = P(w_i|x_j)P(x_j)$.

The mathematical expressions to solve this conditional probability and to learn pLSA parameters are represented in Section 3.4.2.

It should be noted that the number of semantic topics K can be much higher than the number of scene categories, unlike in unsupervised frameworks [105]. Consequently what we use in our model is the image representation of $P(z_k|x_j)_{k=1,2,\dots,K}$.

4.4.2 Capturing Contextual Information

In this chapter we used two holistic features that are applied on the whole image to extract the image contextual information.

Gist The gist descriptor is computed by applying 4×4 Gabor filters on 4 scales and in 6 different orientations on spatial blocks of image.

CENTRIST In this visual descriptor, each pixel value is replaced by its Census Transform (CT) [48] value. In Census Transform, the intensity value of center pixel is compared to the values of Surrounded pixels in a 3×3 neighborhood, if the value of neighboring pixels is bigger than the center pixel, bit 0 is assigned to the corresponding location otherwise the value is set to 1. Finally the CT value is obtained

by decoding the generated binary code to a base-10 number. Afterward the CENsus TRansform hISTograms (CENTRIST) are computed from these transformed CT values.

4.5 Local-Global Feature Fusion (LGFF)

To better describe the LGFF method, it is divided into three main blocks, Exemplar-Set selection (Section 4.5.1), problem formulation (Section 4.5.2) and parameter learning (Section 4.5.3).

4.5.1 Exemplar-Set Selection

Similarity-based classification methods compute pairwise distances between the images, in this paper we are proposing a new similarity space which indicates a base-set for distance measurements. This base-set is called an exemplar-set and is obtained by down-sampling of the training set to choose the most robust images of each category. Utilizing the exemplar-set to find distances will decrease time and memory cost due to the reduced number of distance measurements. The proposed SSFF method weights every exemplar based on its relevance to the scene category, therefore, reducing the size of exemplar-set will decrease the complexity because the number of parameters to be learned are decreased. Besides the reduction in computational complexity, use of the exemplar-set also results in improved accuracy, especially in more complex datasets with high intra-class diversity (e.g. 67-Indoor Scenes). Since the corresponding set consists of the most representative images from each category, hence, when sufficient number of exemplars are chosen, it prevents from less-relevant images to affect our learning.

Figure 4.2 shows the diagram of exemplar-set selection. The goal is to select a robust set of exemplars $E = \{E_1, E_2, \dots, E_p\}$ from the scene dataset. This exemplar-set is a composition of image subsets from each category. These subsets consist of images that capture the specific characteristics of their classes more effectively than the other images in the same category $E = \{E_1, E_2, \dots, E_{p_c}\}$.

To this end (see Figure 4.2), for each category C_i an initially smaller subset of training images is randomly chosen and fixed as a basis. The rest of the training images in category

Figure 4.2: Exemplar-set selection.

C_i is then considered as candidates for the exemplar set. A classifier is trained for each of these exemplar candidates by adding this image to the initial basis, and then we determine how well these classifiers perform on the remainder of the training images that are not in the basis. The recognition accuracies for all of the exemplar candidates within category C_i are measured and a selected number of exemplar candidates that resulted in the highest accuracies become the exemplar subset for class C_i . In this paper, each category employs a fixed number of exemplars, denoted by p_c .

4.5.2 Problem Formulation

Formulation of Distance Function

Our distance function $D_i(x)$ measures the contextual-semantic dissimilarity between image x and exemplar image E_i and is defined by

$$D_i(x) = \lambda G_i(x) + \mu Ct_i(x) + L_i(x), \quad (4.1)$$

where $G_i(x)$, $Ct_i(x)$, $L_i(x)$ are component distances that are respectively related to gist, CENTRIST and local semantic features. They are positive and normalized such that $\sum_i G_i(x) = 1$, $\sum_i Ct_i(x) = 1$ and $\sum_i L_i(x) = 1$. This normalization balances the distribution of image x over the whole exemplar-set. The parameters λ and μ control how discriminative the contextual features are for a particular scene category. $G_i(x)$ is the distance function measuring the dissimilarity of gist contextual features between image x and exemplar image E_i using l_2 norm:

$$G_i(x) = \|gist(x) - gist(E_i)\|. \quad (4.2)$$

$Ct_i(x)$ is associated with the distance between the CENTRIST representation of image x and exemplar image E_i using l_2 norm:

$$Ct_i(x) = \|centrist(x) - centrist(E_i)\|. \quad (4.3)$$

$L_i(x)$ measures the dissimilarity of the semantic topics of image x and exemplar image E_i :

$$L_i(x) = \beta [P(z|x) - P(z|E_i)], \quad (4.4)$$

where vector $\beta = [b_1 \ b_2 \ \dots \ b_K]^T$ represents the relevance of all the K semantic topics of image x for predicting that scene category. In (4.4), $P(z|x)$ and $P(z|E_i)$, are vectored probability representations of x and E_i (corresponding to image x and exemplar E_i) by latent topics z , referred to image representation in latent semantic space which is defined in Section 4.4.1.

Therefore (4.4) expands as

$$\begin{aligned} L_i(x) = & b_1(P(z_1|x) - P(z_1|E_i)) + b_2(P(z_2|x) - P(z_2|E_i)) \\ & + \dots + b_K(P(z_K|x) - P(z_K|E_i)). \end{aligned} \quad (4.5)$$

Contextual-semantic distances, given by $D(x) = (D_1(x), D_2(x), \dots, D_p(x))$, are evaluated between the image x and all images in exemplar-set E .

Formulation of Map Function

As mentioned before, each image is mapped to a real number which represent the confidence score of belonging the image x to category y , using the $f(x)$ map function. The defined distance functions are embedded into the map functions, therefore in addition to the proposed feature weights, α_i / α_{si} parameters are learned for each exemplar E_i to measure that exemplar's contribution for each scene category.

The various distance functions computed distance formulation are embedded in a non-linear map function $f(x)$, formulated as:

$$f(x) = \sum_{i=1}^p \alpha_i \exp(-D_i(x)). \quad (4.6)$$

In the proposed mapping, $\alpha_i \in \mathbb{R}$, are parameters that indicate how discriminative each of the exemplars are in that scene category.

4.5.3 Parameter Learning

The objective of our supervised learning is to determine the optimal values for the parameters, so as to best capture the relationship between image features and scene categories. We optimize the parameters $\lambda, \mu, \beta, \alpha$ occurring in (4.1), (4.4), (4.5), (4.6) by minimizing the error loss in estimating the scene category using the contextual-semantic map

function $f(x)$. According to definition of the map function $f(x)$, for semantic-contextual function $f(x)$ the distances are measured using euclidean distance and they are embedded into a RBF kernel.

To solve this optimization problem l_2 -regularized log-loss is employed. During this optimization, non-negativity constraints are applied on λ , μ and β_k parameters to keep the definition of “distances” meaningful.

To measure the error loss, each training image within the training set T is mapped into a real number through map function $f(x)$ defined in (4.6).

We employ the logistic loss function $\phi(m_t)$ for the error-rate measurement:

$$\phi(m_t) = \frac{1}{\ln 2} \ln(1 + e^{-m_t}), \quad (4.7)$$

where

$$m_t = y_t f_{\lambda, \mu, \beta, \alpha}(x_t) \quad (4.8)$$

in which $y_t \in \{-1, 1\}$. Substituting the log-loss with other mainstream loss functions such as hinge-loss and (0-1) loss does not make any significant difference.

The optimal parameter values are obtained by minimizing the standard regularized classification objective, which is consist of the objective loss function with a regularization term R to prevent overfitting:

$$(\lambda, \mu, \beta, \alpha)^* = \arg \min_{(\lambda, \mu, \beta, \alpha)} \sum_{i=1}^T \phi(m_t) + R. \quad (4.9)$$

The regularization term R is defined like:

$$R = C_g \lambda^2 + C_m \mu^2 + C_b \|\beta\|^2 + C_a \|\alpha\|^2, \quad (4.10)$$

where C_g , C_m , C_b , C_a indicate the amount of regularization that limit the size of the parameters.

Finally, by substituting the map function $f(x)$, the objective function is:

$$\begin{aligned} \phi(\lambda, \mu, \beta, \alpha) = \\ R + \frac{1}{\ln 2} \sum_{i=1}^n \ln(1 + e^{-y_t \sum_{i=1}^p \alpha_i \exp(-\lambda G_i(x_t) - \mu C t_i(x_t) - L_i(x_t))}). \end{aligned} \quad (4.11)$$

In this equation, the distance functions $G_i(x)$, $Ct_i(x)$ and $L_i(x)$ are based on (4.2), (4.3) and (4.4), while n and p are the sizes of training set and exemplar-set respectively. To obtain the best parameter values subject to minimize our objective function, an iterative gradient-based optimization is applied. We applied the Broyden-Fletcher-Goldfarb-Shanno (BFGS) [121] algorithm. BFGS is an iterative method that approximates Newton's method. It is categorized as a Quasi-Newton method. Quasi-Newton methods [122] do not need to directly compute the Hessian matrix of second derivatives. The Hessian matrix is evaluated using the rank-one values (or their approximated value) of gradients. In the Newton method, the goal is to minimize the $f(x)$ function through iterations. In each step, the next point is defined by:

$$x_{k+1} = x_k + \alpha_k p_k, \quad (4.12)$$

where α_k is the step size and p_k is the search direction. To find an acceptable step size α_k , a line search is performed $\min_{\alpha_k} f(x + \alpha_k p_k)$, the search direction p_k is defined by:

$$p_k = -H_k^{-1} \nabla f(x_k), \quad (4.13)$$

when the gradient vector is defined as:

$$\nabla f(x) = \left[\frac{\partial}{\partial x} f(x) \right]^T \quad (4.14)$$

and H_k is an approximation of the Hessian matrix ($\nabla^2 f(x)$):

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2}{\partial x_1 \partial x_1} f(x) & \dots & \frac{\partial^2}{\partial x_1 \partial x_n} f(x) \\ \vdots & & \vdots \\ \frac{\partial^2}{\partial x_n \partial x_1} f(x) & \dots & \frac{\partial^2}{\partial x_n \partial x_n} f(x) \end{bmatrix} \quad (4.15)$$

In the BFGS method, the H_k matrix is iteratively updated in each step;

$$H_{k+1} = H_k + U_k + V_k, \quad (4.16)$$

where U_k and V_k are symmetric rank-one matrices with different matrix bases. These iterations are repeated until the criteria are met. The BFGS method is usually used for unconstrained non-linear optimization. Karush-Kuhn-Tucker (KKT) [123, 124] conditions can be applied to impose inequality constraints. More on this will be presented in Section 6.2.4 of Chapter 6.

4.6 Experimental Results

To demonstrate the efficiency of our feature fusion method, we have evaluated our method on two of the most challenging scene datasets: the *15-Scene* dataset and the *MIT 67-Indoor Scenes*. In both datasets, all of the images are rescaled to 256×256 and the codebook size is fixed at 400 during the experiments.

4.6.1 Evaluation Results of 15-Scene Dataset

The first experiment involves finding the optimal size of the exemplar-set. In this experiment we increased the number of exemplar images, starting with 10 images per category. This experiment is performed on validation-set of 10 randomly selected images of each category. As seen in Figure 4.3, the accuracy increases with the number of exemplars until we reach a certain point, after which the accuracy drops due to over-fitting. The number of exemplars at that point is the optimal number. Considering the low variance of the accuracies, there is no reason for fine-tuning of the number of exemplars, hence this number is increased by 10 per step. The optimum number of exemplars for the 15-Scene dataset is 50 images per category. Therefore the number of exemplar images to achieve the best accuracy is 50 exemplar images per category. To train a classifier for each category we sample all training images as positive examples and 25 training images per category as negative examples. We validate our method on the test data using 50 exemplar per category and obtained the recognition accuracy of 82.27%.

To show the efficiency of our method we compared our results with the recently reported results on the dataset. As we see in Table 4.1. LGFF outperforms all the other frameworks. The Wu and Rehg [87] entry in this table is the result of CENTRIST visual feature without building the spatial pyramid histograms over this feature.

In this chapter we do not compare our results with the methods that capture image’s spatial locations. In Chapter 5 the results of LGFF method is investigated by embedding the spatial information into this method.

The benchmark frameworks to this paper are, Bosch et al. [83] and Oliva and Torralba [16] and Wu and Rehg [87]. The images in [83] are considered as mixtures of topics; consequently the document specific distributions of topics are extracted by pLSA model

Figure 4.3: Recognition accuracy vs. exemplar images per category in 15-Scene dataset.

Table 4.1: Comparison with recently reported results for 15-Scene.

Method	Accuracy(%)
LGFF	82.27
Li et al. [90]	80.9
Saghafi et al. [1]	79.22
Liu&Shah [85]	75.16
Liu et al. [86]	74.9
Oliva&Torralba [16]	74.10
Wu&Rehg [87]	73.5
Bosch et al. [83]	73.30
Quelhas et al. [82]	71.24
Fei-Fei&Perona [22]	65.2

to represent the images. This is followed by a SVM classifier with the recognition accuracy achieved in this model much lower than our result. In [16, 87] holistic approaches has been proposed, that capture the scene information bypassing the object-centered approaches and local feature detection. In these approaches global feature gist and CENTRIST are extracted respectively. Although the recognition rate in this holistic method is slightly better than [83] it still is much lower than what we have achieved. The recognition accuracy on each existing category for benchmark framework and ours is shown in Figure 4.4. Our method resulted in higher accuracy for all of the categories except *Suburb*

and *Office* where the gist and centrist methods gave better results, respectively, with the greatest improvement occurring for indoor categories. This is likely because indoor scene categories have high intra-class diversity and multiple objects within their images, necessitating the combination of both local and global features.

Figure 4.4: Per class recognition rate for Supervised pLSA, Gist and LGFF.

We improved average recognition accuracies over the indoor categories (*Bedroom*, *industry*, *Kitchen*, *LivingRoom*, *Store*, *Office*) and outdoor categories (*Suburb*, *Coast*, *OpenCountry*, *street*, *Mountain*, *Highway*, *Forest*, *TallBuilding*) using supervised pLSA, Gist and our LGFF method. Table 4.2 shows the results.

Table 4.2: Recognition rates for indoor, outdoor categories in 15-Scene dataset.

	Supervised pLSA	Gist	CENTRIST	LGFF
Indoor	56.63%	62.41%	56.75%	75.69%
Outdoor	83.56%	82.17%	79.31%	86.66%

To show the superiority of our LGFF method in comparison to the Multiple Kernel Learning feature fusion, we obtained the result of simpleMKL [125] by combining the

same features as with the features used in LGFF method. The results are illustrated in Figure 4.5¹.

The visual features and their associated distance functions and kernels are introduced as follows:

Gist We used Euclidean distance to measure the dissimilarity between a pair of image gist descriptors, and an RBF kernel is associated with this visual feature.

pLSA histograms of Dense Sift As with gist feature, we consider the Euclidean distance and RBF kernel for this representation.

CENTRIST Euclidean distance and RBF kernel is used for this feature.

According to Figure 4.5, combination of pLSA histograms and global features gist, CENTRIST using LGFF performs better compared to MKL. As seen in this figure, LGFF outperforms MKL, moreover LGFF has less time complexity compared to MKL. One run of training and testing for LGFF and simpleMKL, in combining pLSA, gist and CENTRIST take around 12 minutes and 52 minutes respectively. The reason for this is; in LGFF method all the new features will be embedded into $f(x)$ map function, while in MKL, for every new feature added, a new kernel has to be learned which will increase the time complexity especially for complex kernels, hence the computational cost for LGFF is significantly less than MKL.

4.6.2 Evaluation Results of 67-Indoor Scenes Dataset

To find the optimal size of the exemplar-set for the 67-Indoor Scenes dataset we performed a similar experiment which increase the number of exemplar images by 10 images per category in each step started from 10 images per category and stopped when the accuracy began to decline. In this experiment the size of validation-set is random splitting of training set for 10 images per category for validation. shown in Figure 4.6 the optimal number of exemplars is 30 images per category for the 67-Indoor Scenes dataset. Hence the experiments performed on 67-Indoor Scene dataset while number of exemplars are

¹The results in Figure 4.5 is slightly different from the ones reported in the papers [9, 83]. This is due to applying the same experimental settings as with ours e.g. resizing all the images into 256×256 .

Figure 4.5: Comparison of scene recognition accuracy for different features (pLSA vs. Gist vs. CENTRIST) and combination of features using (LGFF vs. MKL).

Figure 4.6: Recognition accuracy vs. exemplar images per category in 67-Indoor Scenes dataset.

fixed to 30/category. To train a classifier for each category we sample all training images as positive examples and 20 training images per category as negative examples. To show the

performance of our feature fusion method, we compared our results with recently reported results on this dataset. As seen in Table 5.4, our method achieved the promising result of **26.7%**.

Table 4.3: Comparison with recently reported results for 67-Indoor Scenes.

Method	Accuracy(%)
LGFF	26.7
Li et al. [90]	37.6
Quattoni&Torralba(Annotation) [7]	26.00
Quattoni&Torralba(Segmentation) [7]	25.04
Wu&Rehg [87]	23.13
Oliva&Torralba [16]	22.0
Quelhas et al. [82]	21.17
Bosch et al. [83]	20

According to this table our method outperform most of the suggested frameworks so far, as has been mentioned before in this chapter we do not compare our results with the methods that embed spatial information into their framework. The object-bank framework suggested by Li et al. [90] is much accurate than our method in 67-Indoor Scenes dataset although our method outperforms object-bank in 15-Scene dataset. This is due to the essence of scene in 67-Indoor Scenes, in this dataset there is numerous objects within the image therefore to discriminate scene classes either the objects (such as object-bank) or location information should be captures. In Chapter 5, we incorporated spatial information into the LGFF framework and achieved very high recognition accuracy among the implemented methods on the 67-Indoor Scenes dataset. Compared to Object-bank and Quattoni and Torralba [7] frameworks, LGFF has much less time complexity. LGFF combines together the low-level features while these two other frameworks need to capture high-level concepts such as objects and Region-of-Interests (ROIs). Quattoni and Torralba [7] method use either annotation or automatic segmentation to capture ROIs, despite high complexity of their method, LGFF still give higher recognition rate. Figure 4.7 shows the significant improvement of recognition accuracy in our feature integration model compares to the baseline frameworks for this model.

Figure 4.7: Comparison of scene recognition accuracy for different features (pLSA vs. Gist vs. CENTRIST) and combination of features using LGFF.

4.7 Chapter Summary

In this chapter an innovative method is proposed to integrate global contextual features together with the semantic information of local features. The features are embedded into a map function based on a novel distance function. A parameter learning model is used to determine the relative importance weights of global information and the latent variables in the latent semantic space. An empirical study has been performed on the 15-Scene and 67-Indoor Scenes datasets in order to demonstrate the impact of appropriately incorporating both local and global information for the purpose of scene recognition, with promising results. The next chapter employs the method proposed in Chapters 3 and 4 to exploit the spatial information of scene partitions.

Chapter 5

Capturing Image Spatial Information in Scene Recognition Systems

Parts of this chapter have been submitted to [4, 5].

5.1 Introduction

In the previous chapters, we proposed approaches to extract semantic information from scenes and combine it with global features to help extract the context of the scene. In the previously proposed methods, spatial information was not considered. In this chapter, we try to make use of spatial information to increase the efficiency.

Spatial arrangements of visual features have a significant impact on image classification systems. One of the earliest attempts to compensate for the lack of spatial information in scene recognition systems was performed by Szummer and Picard [73]. In [73], the images are divided into 4×4 sub-regions and a k -nearest neighbor classifier is applied to each region. Eventually, a voting algorithm is employed to choose the most repeated category-label within the sub-regions as the image semantic label. This method, due to its use of image location information, improves the recognition accuracy significantly. To capture the spatial information of the global features, Oliva and Torralba [16] proposed a Windowed Discriminant Spectral Template (WDST) rather than a global Discriminant Spectral Template (DST) in their gist descriptor. The Windowed DST is used to implicitly capture the spatial layout of the scene. WDST applies a Fourier

transform to the 8×8 grids. They performed an empirical study applying both DST and WDST and found that the Windowed-DST shows radical improvements.

The *bag-of-words* framework is one of the most common strategies when local features are employed for scene recognition. This framework offers an orderless representation for the images and, in this representation, the position of the local features is ignored. Lazebnik et al. [9] developed a spatial version of the Pyramid Match Kernel (PMK) [126] to overcome the lack of spatial information in the *bag-of-words* framework. The Pyramid Match Kernel was proposed by Grauman and Darrell [126] on feature space to find the correspondence between feature sets (such as bag-of-words feature vectors) using a hierarchical quantization. Lazebnik et al. [9] have used Spatial Pyramid Matching on the image space, i.e. instead of dividing the feature space into hierarchical levels, the pyramid levels are built over the image sub-regions. Spatial Pyramid Matching successfully accomplishes the goal of incorporating spatial information into the *bag-of-words* framework. There are frameworks which incorporate spatial pyramid structure and better classification accuracy was reported [8, 87, 88].

In this chapter, we impose spatial information on two of the previous frameworks proposed in Chapter 3 and 4. In Chapter 3, we proposed a framework to embed visual words into the concept space in order to build a semantic visual vocabulary. As mentioned earlier, although *bag-of-words* is very suitable in scene recognition systems, it ignores both the semantic and spatial information of the features. By constructing a semantic vocabulary, we can overcome the lack of semantic information. The next step is to capture the spatial location of local features. In this chapter, we propose two methods (*global* and *region-based*) to capture location information when using a semantic vocabulary. Both methods build spatial pyramids over the image's blocks. The methods differ in that the *Global* method initially forms the semantic vocabulary and then divides the image into spatial sub-regions, while the *Region-based* method forms the semantic vocabulary over each region individually.

Chapter 4 proposed a method to combine semantic local patches with global contextual features by defining a map function for image x . This map function assigns a confidence score to image x that indicates its degree of affinity to the scene category. In

this chapter, we propose a new mapping by introducing a spatial map function that incorporates the features' spatial position together with contextual and semantic information from the image.

The contribution of this chapter is on developing methods to build spatial pyramid kernels over our proposed frameworks (Semantic Visual Vocabulary, Local-Global Feature Fusion).

5.2 Spatial Content Capture in Semantic Vocabulary Framework for Scene Recognition

In this section, the spatial pyramid matching schema is incorporated into the new concept space proposed in Chapter 3.

5.2.1 Capturing Spatial Information with Spatial Pyramid Matching (SPM)

The spatial pyramid matching technique is a simple yet efficient framework. SPM gives a multi-resolution representation of the image by dividing it into increasingly finer sub-regions. It was first proposed by Lazebnik et al. [9] and inspired by the Pyramid Match Kernel (PMK). PMK defines different levels of resolution in the feature space by defining increasingly coarser grids [126]. Although PMK is very precise, it ignores the spatial location of individual features. SPM has been successfully used as an extension of the *bag-of-words* framework. The bag-of-words framework gives an orderless representation of the image, while SPM uses a spatial pyramid representation of the image. The matching score of this pyramid representation is obtained by a weighted combination of histogram intersections at multi-spatial resolutions.

The image local feature f is denoted by $f = (x, y, d)$ where (x, y) shows the feature's location coordinates and d represents the feature descriptor. In the BoW framework, there are K discrete cluster centroids where each centroid represents one of the visual words. Feature f is quantized into one of these centroids with respects to its descriptor, but the location coordinates (x, y) are completely ignored.

On the other hand, in the SPM framework these location coordinates are used to enhance the power of BoW, e.g. in Figure 5.1, the image is divided into $2^l \times 2^l$ uniform

sub-regions at level l of resolution with 2^l evenly sized partitions in each dimension (horizontal or vertical). At higher levels of resolution, the image is divided into finer regions. At resolution l ($0 \leq l \leq L$), feature f is assigned to one of the $2^l \times 2^l$ sub-regions based on its location coordinates, while its descriptor is quantized into one of the K centroids.

Given the two-dimensional vectors X_k and Y_k as the set of coordinates for channel k , $H_{X_k}^l$ and $H_{Y_k}^l$ are our histograms at level l of resolution, where the dimension of this histogram is D . To find the number of matches at level l of channel k , the Histogram Intersection Kernel (HIK) is applied to the H_{l_k} histograms:

$$I(H_{X_k}^l, H_{Y_k}^l) = \sum_{i=1}^D \min(H_{X_k}^l(i), H_{Y_k}^l(i)). \quad (5.1)$$

The matching score in channel m is measured by applying PMK:

$$K^L(X_k, Y_k) = \frac{1}{2^L} I^0 + \sum_{l=0}^L \frac{1}{2^{L-l+1}} I^l. \quad (5.2)$$

The pyramid matching formulation ($K^L(X_k, Y_k)$) shows that the weights are inversely proportional to the size of the sub-regions.

The final match kernel is sum of the match scores from all of the K channels:

$$K^L(X, Y) = \sum_1^K k^L(X_k, Y_k). \quad (5.3)$$

5.2.2 Spatial Content Capture using Global and Region-based Models

To perform the spatial content capture, we suggest two methods: global and region-based. In the *global* method, after projecting the word space into the concept space, the k -means algorithm is applied to co-cluster the synonymous words. The image is divided into increasingly finer blocks. Afterwards, the visual features within each subregion are quantized into histograms of bag-of-words based on the appearance-based vocabulary and then the histogram bins of the synonymous words are merged together.

Figure 5.1: Example of constructing a three-level pyramid.

The final spatial representation based on the semantic vocabulary is obtained after applying the spatial pyramid weighting for each level of resolution.

In the *region-based* method, there is a separate concept space for every spatial block of the image, i.e. the latent semantic models are applied to each sub-region after it is quantized into the bag-of-words histogram according to the appearance-based vocabulary. *K*-means clustering is applied to these new regional word-topic (concept) spaces and, consequently, the histogram bins for the synonymous words in the *bag-of-words* representation are bound together.

We argue that the region-based method is more effective than the global method because the spatial locations are considered while co-clustering the synonym words in the region-based method; this is important because semantic visual words differ according to the visual features located on that partition of the image.

Although the experiments show that the region-based method outperforms the global method, the improvements in the results are not very significant and it is due to the

essence of visual word vocabularies. Since we have only one semantic vocabulary for the whole dataset while building this vocabulary, the location of the features is not important; rather, the most important issue is which part of the image we project by this vocabulary.

5.2.3 Experimental Evaluation

In these experiments the number of topics is fixed to 70 which, according to Chapter 3 results in the highest accuracy.

Figure 5.2: Performance of proposed global method using LSA.

In the first experimental series, we apply the global method to impose spatial information. As shown in Figures 5.2 and 5.3, the results are very promising and the diagrams prove the efficiency of our method. In these figures we study the influence of 3 different initial vocabulary sizes (K_i) on the recognition accuracy while the final semantic vocabulary size (K_f) changes. In Figure 5.2, the LSA method is used as the embedding method. According to this figure, the best accuracy is achieved with $K_i = 1500$ and a final vocabulary size of $K_f = 800$.

Figure 5.3: Performance of proposed global method using pLSA.

We use the pLSA model to project the word-document space into the concept space, the results are illustrated in Figure 5.3, and the best accuracy is achieved at $K_i = 1500$ and $K_f = 600$.

To give a better comparison of the pLSA method vs. the LSA method with different initial vocabulary sizes (K_i) over changes in the final vocabulary size (K_f), the results of LSA vs. pLSA are illustrated in Figure 5.4. As shown, with larger initial vocabulary sizes pLSA gives better results than LSA (i.e. for $K_i = 2000$); however, for $K_i = 1000$ LSA achieves higher accuracy and with $K_i = 1500$ the results are on par.

To implement the proposed global method, we apply pyramid matching on different levels of resolution. The results of the global method in level 1 and 2 of resolution for the LSA and pLSA models are demonstrated in Figure 5.5.

The time complexity of implementing region-based model by applying pLSA embedding is very high, therefore to perform the experiments for this model we just used LSA embedding. The experimental results for the region-based approach using LSA are

Table 5.1: Comparison of execution time for building semantic visual vocabulary in 15-Scene dataset using LSA/pLSA.

Method	Time(min)
LSA	1.16
Spatial LSA (global method)	4.75
Spatial LSA (Region-based method)	17.1
pLSA	61.3
Spatial pLSA (global method)	81.8

demonstrated in Figure 5.6 while the initial vocabulary size is fixed to 1500, $K_i = 1500$. This figure shows the efficiency of the region-based method in using LSA. According to this plot diagram the accuracy increases until we reach $K_f = 800$, then it slightly decreases.

To compare the global method versus the region-based method using LSA model, the results of global method is also illustrated in Figure 5.6. According to this demonstration although the region-based method outperforms the global method, their results are almost on par. However, since it is necessary to project each spatial block separately into a new concept space in the region-based method, the time complexity is very high. Therefore, the difference in time complexity between the global and region-based methods justifies using the global method.

The time execution to build a semantic visual vocabulary using either LSA or pLSA embedding is shown in Table 5.1.

5.3 Spatial Content Capture in Local-Global Feature Fusion for Scene Recognition

In this section we propose a novel distance learning method which combine semantic relationships and the location information of local features with contextual global features. This section is an extension of the LGFF method in Chapter 4.

5.3.1 Capturing Spatial Location Information

Spatial pyramids are constructed over two features SIFT and CENTRIST.

SPM The spatial pyramid of *bag-of-words* is build using the SIFT local features. To construct this spatial pyramid histograms the extracted dense features are quantized into 400 visual words in three levels of resolution ($L=2$ in SPM implementation by Lazebnik et al. [9]). These histograms are computed on the image grids of 1×1 , 2×2 and 4×4 .

sPACT This spatial representation is also based on the spatial pyramid matching schema but the splitting of the image at each level is different. In level 2, the image is divided into 4×4 regions. To prevent artifacts of this non-overlapping division, the division is shifted to form 9 more regions. There are a total of 31 blocks, comprising 25 from level 2, five from level 1 and one from level 0. A CENTRIST descriptor is applied on each of these blocks, followed by PCA to reduce the dimensionality of the features to 40. The Spatial Principal component Analysis of Census Transform Histogram (sPACT) is build by concatenating these block representations.

5.3.2 Spatial Local-Global Feature Fusion (Spatial-LGFF)

To better describe the Spatial-LGFF method, it is divided into two main blocks, learning phase (Section 5.3.2.1) and scene type recognition phase (Section 5.3.2.2).

5.3.2.1 Learning Phase for Spatial-LGFF method

In the learning phase (see Figure 5.7), the holistic contextual features of gist and CENTRIST are extracted from the whole image, while the feature descriptors of SIFT $SF = \{SF_1, SF_2 \dots SF_n\}$ are applied on the local patches, semantic and spatial representation are captured from the SIFT features. Although CENTRIST visual feature is a contextual global feature, it could also be applied on the image patches (denoted by $CF = \{CF_1, CF_2 \dots CF_{31}\}$) after spatial partitioning of the image to build the CENTRIST-based spatial representation.

As mentioned the purpose of Spatial-LGFF method is to project every image x within the C class scene dataset, to C real numbers where each number shows the affinity of the image x to one of the categories. To perform this mapping we use an adoptive category-dependent method, in this method for each C_i category the mapping is consist of learning

two separate map functions $f(x)$ and $f_s(x)$, while the final $h(x)$ map function is summation of normalized $f(x)$ and $f_s(x)$ functions, $h(x) = f(x) + f_s(x)$. These functions ($f(x)$, $f_s(x)$) and their parameters are independent hence the parameters could be learned by defining two separate objective function. The advantage of this learning is the time and memory cost saving and its simple implementation.

The $f(x)$ function is used to map image x based on its contextual and semantic representations. To this end the exemplar-set is used as a basis for measurement, therefore the weighted dissimilarity of image x to exemplar-set is computed using their semantic and contextual features, while the weights indicates the relevance of each exemplar images to C_i category. There are as well parameters which present the amount of contribution for the features meet in this measurement (the importance of gist (λ) and CENTRIST (μ) contextual features also the significance of each latent topic in semantic representation explicitly ($B_1, B_2 \dots B_k$)). Consequently these weights and parameters are assigned during the parameter learning.

On the other hand $f_s(x)$ mapping relies on spatial description of image x . In the parameter learning for this function importance of SIFT spatial structure (γ) and CENTRIST spatial structure (ψ) are balanced. The exemplar images are as well weighted during the learning.

The basic distance metric to measure dissimilarity for $f(x)$ function, is euclidean distance while $f_s(x)$ function utilize Histogram Intersection Kernel (HIK). The different essence of euclidean space and HIK kernel space justifies utilizing two separate map functions, therefore the distances in $f(x)$ are embedded into a Gaussian Radial Basis Function (RBF) kernel while HIK distances in $f_s(x)$ have to be followed by a linear kernel.

Problem Formulation

The spatial distance function $D_{si}(x)$ is defined as follows:

$$D_{si}(x) = \gamma S_i(x) + \psi P_i(x), \quad (5.4)$$

where γ and ψ are the weights on distances between image x and exemplar E_i with respect to the *spatial* representations of SIFT-related SPM features and CENTRIST-related sPACT features.

The function $S_i(x)$ is the distance between image x and exemplar E_i using their SPM representations. The function $P_i(x)$ on the other hand computes the distance between the corresponding sPACT features. $S_i(x)$ and $P_i(x)$ are positive and normalized such that $\sum_i S_i(x) = 1$ and $\sum_i P_i(x) = 1$. As in previous work, the similarity of histograms at each level of resolution in the spatial pyramids representation, whether using SIFT or CENTRIST, is based on histogram intersection. Therefore to obtain the overall HIK distance between two spatial features, the weighted sum of these histogram intersections are computed, as in Pyramid Match Kernel (PMK) [126].

Spatial distances, given by $D_s(x) = (D_{s1}(x), D_{s2}(x), \dots, D_{sp}(x))$, are evaluated between the image x and all images in exemplar-set E . The spatial distance vector $D_s(x)$ on the other hand is embedded into a linear map function $f_s(x)$:

$$f_s(x) = \sum_{i=1}^p \alpha_{si} D_{si}(x), \quad (5.5)$$

where α_{si} as before indicates the discriminativeness of each exemplar E_i .

The final map function for image x is:

$$h(x) = f(x) + f_s(x), \quad (5.6)$$

where the $f(x)$ and $f_s(x)$ functions are standard normalized such that they have zero mean and unit norm. The assigned parameters $\lambda, \mu, \beta, \alpha, \gamma, \psi, \alpha_s$ are estimated during parameter learning as discussed in Section 5.3.2.1.

Parameter Learning

To learn optimum values for γ, ψ, α_s parameters by minimizing the error loss in estimating the scene category using contextual-semantic map function $F_s(x)$.

To measure the error loss, each training image within the training set D is mapped into a real number through map function $F_s(x)$ which is defined by Equation 5.5.

We employed logistic loss function for error-rate measurement, therefore the objective function is like:

$$\phi_s(\gamma, \psi, \alpha_s) = R_s + \frac{1}{\ln 2} \sum_{i=1}^n \ln(1 + e^{-y_t \sum_{i=1}^p \alpha_s i(\gamma S_i(x_t) + \psi P_i(x_t))}), \quad (5.7)$$

where $y_t \in \{-1, 1\}$ and the R_s regularization term is defined as follows:

$$R_s = C_s \gamma^2 + C_p \psi^2. \quad (5.8)$$

5.3.2.2 Scene Type Recognition for Spatial-LGFF method

As shown in Figure 5.8, given a test image, features are extracted to form the semantic and spatial models. The values for $D(x)$ and $D_s(x)$ distance functions are then computed and utilized in the map functions $f(x)$ and $f_s(x)$ to get $h(x)$. These are computed with respect to every image in the overall exemplar-set.

As mentioned previously, there is a unique $h(x)$ function for each scene category. Each such function generates an affinity value for image x to the corresponding scene category. The image is then categorized as the class with the highest $h(x)$ value.

5.3.3 Experimental Evaluation

To demonstrate the efficiency of our feature fusion method, we have evaluated our method on two of the most challenging scene datasets: the *15-Scene* dataset and the *MIT 67-Indoor Scenes*.

5.3.3.1 Evaluation Results on 15-Scene Dataset

According to the experiments on Chapter 4, the optimum number of exemplars for 15-Scene dataset is 50 images per category. Therefore the best accuracy for Spatial-LGFF (**89.47%**) is achieved with 50 exemplar images per category. The Confusion matrix in this condition is illustrated in Figure 5.9.

Table 5.2 shows the effectiveness of our proposed method by comparing the evaluation results with recently published results on this dataset. As seen in this table, our method achieved the state-of-the-art result of **89.47%** on the 15-Scene dataset. This has verified

Table 5.2: Comparison with recently reported results for 15-Scene.

Method	Accuracy(%)
Our FF method	89.47
Xiao et al. [101]	88.1
Wu&Rehg [87]	83.88
Bosch et al. [8]	83.7
Lazebnik et al. [9]	81.4
Li et al. [90]	80.9
Parizi et al. [93]	78.6
Liu&Shah [85]	75.16
Liu et al. [86]	74.9
Oliva&Torralba [16]	74.10
Bosch et al. [83]	73.30
Quelhas et al. [82]	71.24
Fei-Fei&Perona [22]	65.2

that our method performs better than proposed MKL method in [101] despite the fact that our method requires a much smaller number of features.

The recognition accuracies for all of the features that we embedded into our framework is shown in Figure 5.10. Please note that the accuracies shown in this figure are slightly different from the ones reported in the original papers, because in our case, all the images were resized to 256×256 .

In [83] the images are considered as a mixture of semantic topics. If followed by SVM classification, the classification rate is lower than our method. Oliva and Torralba [16] and Wu and Rehg [87] use gist and CENTRIST global features respectively, bypassing object-centered and local information. Although the recognition rates in these holistic methods were higher than those for purely local methods, they were still much lower than our reported results. Capturing the spatial location of local patches in [9] and [87] significantly improved the recognition accuracy for scene recognition. Nevertheless, spatial pyramid models were also less accurate compared to our feature fusion method which comprehensively accounts for semantic, contextual and spatial information within the scene.

The execution time for running the Local-Global Feature Fusion method and the Multiple Kernel Learning method on the 15-Scene dataset is shown in Table 5.3.

Table 5.3: LGFF vs. MKL execution time for 15-Scene dataset.

Method	Time(min)
LGFF	12
Spatial-LGFF	19
MKL	52
Spatial-MKL	75

Table 5.4: Comparison with recently reported results for 67-Indoor Scenes.

Method	Accuracy(%)
Our FF method	45.0
Singh et al. [94]	49.4
wu [127]	47.15
Pandey&Lazebnik [92]	43.1
Parizi et al. [93]	37.93
Li et al. [90]	37.6
Wu&Rehg [87]	36.9
Lazebnik et al. [9]	34.4
Quattoni&Torralba [7]	26.00
Quelhas et al. [82]	21.17
Oliva&Torralba [16]	22.0
Bosch et al. [83]	20

5.3.3.2 Evaluation Results on MIT 67-Indoor Scenes Dataset

The number of exemplars that gives the highest recognition accuracy is 30 images per category. As seen in Table 5.4, our method achieved the promising result of **45.0%**. According to this table our FF method’s result is comparable to the state-of-the-art results in 67-Indoor Scenes.

In [92] Pandey and Lazebnik used the popular Deformable Part-based Model(DPM) [95] for scene recognition and achieved the accuracy of **30.08%**. Subsequently they formed the *color-gist* feature by applying the gist descriptor over the color channels and concatenating the gist vectors together. Eventually the combined use of the color-gist feature together with SIFT spatial pyramids and DPM results achieved an accuracy of 43.1%. The discriminative mid-level patches in [94] achieved **38.1%** accuracy while combining these mid-level patches with color-gist, DPM and SIFT spatial pyramids they obtained **49.4%** recognition accuracy. Wu [127] achieved an effective SVM optimization by proposing

Table 5.5: Recognition Accuracy per class for 67-Indoor Scenes.

	Accuracy		Accuracy		Accuracy		Accuracy
inside airport	10	computer room	44	inside subway	47	pantry	55
art studio	30	concert hall	66	jewelery shop	31	inside pool	75
auditorium	83	corridor	61	kinder garden	30	prison cell	55
bakery	26	deli	5	kitchen	28	restaurant	20
bar	33	dental office	38	laboratory	18	restaurant kitchen	39
bathroom	72	dining room	33	laundromat	22	shoe shop	15
bedroom	23	elevator	76	library	32	stairs case	40
bookstore	15	fast food restaurant	29	living room	20	studio music	63
bowling	75	florist	63	lobby	50	subway	57
buffet	80	game room	15	locker room	38	toy store	27
casino	42	garage	38	mall	50	train station	55
children room	44	green house	95	meeting room	59	tv studio	38
inside church	84	grocery store	23	movie theater	70	video store	27
classroom	55	gym	50	museum	26	waiting room	14
cloister	80	hair salon	23	nursery	60	warehouse	38
closet	83	hospital room	20	office	33	wine cellar	33
clothing store	61	inside bus	78	operating room	26		

Table 5.6: LGFF execution time for 67-Indoor Scenes dataset.

Method	Time(min)
LGFF	254
Spatial-LGFF	291

Power mean SVM (PmSVM) algorithm which generalize many kernels including intersection and χ^2 kernels.

Figure 5.11 shows the significant improvement of recognition accuracy in our feature integration model compares to the visual features we used to incorporate to our model. The accuracy per class for this dataset is illustrated in Table 5.5. The time execution while applying the Local-Global Feature Fusion method on the 67-Indoor Scenes dataset is shown in Table 5.6.

5.3.4 Implementation details and Memory complexity of the Spatial-LGFF method

To perform our experiments using LGFF, we need to keep the distances between the exemplar-set and training images during parameter learning, which requires additional storage. The size of this distance matrix is:

$$\text{Number of Exemplars per category} \times \text{Number of categories} \times \text{Number of training images (n} \times \text{Number of categories),}$$

where n is the number of training images per category. Moreover, to train one versus all classifiers for category C_i , we need to compute the new distances for each category using n positive training images and $n/4$ negative images per category. The corresponding matrix size is:

$$\text{Number of Exemplars per category} \times \text{Number of categories} \times (n + n/4 \times (\text{Number of categories} - 1)).$$

For the 15-Scene dataset the distance matrix size is $50 \times 15 \times 1500$, which takes around 8 MB of memory. Since we have five different features, we need five different distance matrices. The size of the new distance matrix is $50 \times 15 \times 100 + 25 \times 14$ and a memory space of $5 \times 1.5 = 7.5$ MB is required for new distance learning for each category. For the 67-Indoor Scenes dataset, we have $67 \times 30 \times 5360$, which is about 78 MB, and generally $5 \times 78MB$ is required for our distances. A space of $5 \times 20 = 100MB$ is required for new distance learning of each scene category. Based on the time and memory cost, we do not suggest the Local-Global Feature Fusion method for large-scale datasets.

5.4 Chapter Summary

Position information significantly improves scene recognition accuracy [8, 9]. Inspired by this, in this chapter we bring position information into the proposed semantic vocabulary and Local-Global Feature Fusion frameworks. The proposed spatial methods incorporate the pyramid matching technique. In semantic vocabulary, pyramid levels are constructed directly over the latent semantic space using either one concept space for the entire scene or each image sub-scene separately projected into the concept space. In Local-Global Feature Fusion, pyramid levels are formed over the SIFT and CENTRIST low-level features and define a novel spatial map function. The experimental evaluation shows remarkable improvements in both the 15-Scene and 67-Indoor Scenes datasets.

5.4.a: $K_i = 1000$

5.4.b: $K_i = 1500$

5.4.c: $K_i = 2000$

Figure 5.4: Recognition accuracy using pLSA vs. LSA concept space for initial vocabulary sizes $K_i = 1000, 1500, 2000$ using global method.

5.5.a: LSA

5.5.b: pLSA

Figure 5.5: Recognition accuracy using pLSA and LSA concept space in different levels of resolution.

Figure 5.6: Comparing the performance of proposed global method vs. region-based method using LSA.

Figure 5.7: Block diagram of learning phase in our Feature Fusion method.

Figure 5.8: Block diagram of Scene Recognition for test input image.

Figure 5.9: Confusion matrix for the best result in 15-Scene dataset.

Figure 5.10: Spatial-LGFF Recognition rate vs. Baseline frameworks for 15-Scene dataset.

Figure 5.11: Spatial-LGFF Recognition rate vs. Baseline frameworks for 67-Indoor Scenes dataset.

Chapter 6

Conclusions and Future Work

6.1 Summary of Contributions

The central theme of the research done in this thesis is the application of local tools to significantly improve the accuracy of scene recognition.

In Chapter 3 [1, 2], to compensate for the prevision of semantic relations in the *bag-of-words* framework, we propose a novel method in which we project the appearance-based visual feature space into the conceptual feature space using latent semantic models such as LSA and pLSA. These latent semantic models decompose histograms of *bag-of-words* to image space and word space. This chapter concentrates on word space. In the proposed conceptual space, the synonymous visual words are the ones close to each other with regards to their Euclidean distance; hence, to construct a discriminative and semantically meaningful vocabulary, these visual words are co-clustered into one word. This approach achieved very promising results using the *15-Scene* dataset.

Extracting visual content from an image is the most important part of a visual recognition system and has a great impact on the performance. Although we can define robust recognition systems purely using local tools for scene recognition, high intra-class variety within indoor scene classes reduces the accuracy dramatically. Chapter 4 [3] shows that the incorporation of global features into local models improves the discriminative power because global features capture the structure of a scene. Local features are of significant importance for identifying objects within a scene, while the advantage of conceptual global features is that they exploit the scene context through a coarse description of the spatial layout. In our LGFF framework, we integrate global and local feature space into

a novel map function. The LGFF method proposes a new similarity space that indicates a base-set for distance measurements. This base-set is called an exemplar-set and is comprised of the most robust training images from each category. The proposed feature fusion method is performed locally by weighting each of the local semantic patches individually.

In Chapter 5 [4, 5], we discuss the significance of the spatial arrangements of visual features to establish an efficient recognition system; we show that in scene recognition systems the accuracy radically increases when we capture the position of the local features. As mentioned earlier, the main drawbacks of the bag-of-words framework is that it does not capture the spatial locations and semantic relationships of the local features. Chapters 3 and 4 propose approaches to reduce this semantic loss. In Chapter 3, instead of using a classic appearance-based vocabulary, we build a semantic visual vocabulary to capture the semantic relationships of features. To compensate for the lack of spatial information in the proposed method, we suggest two frameworks - one global and one region-based. These frameworks build spatial pyramids over the local features by applying the semantic visual vocabulary. The evaluation results demonstrate a significant increase in accuracy. The Local-Global Feature Fusion (LGFF) method that we proposed in Chapter 4 embeds contextual and semantic information into a novel map function. We generalize the LGFF method to a new framework that captures spatial information by suggesting a spatial map function. The performance of the spatial-LGFF method on two of the most challenging scene datasets, the *15-Scene* categories and *67-Indoor Scenes* datasets, shows promising results.

6.2 Future Work

The research described in this thesis has led to methods for improving recognition accuracy of scene recognition systems. Still, a number of improvements and new directions for scene recognition could be made to overcome the difficulties of the proposed methods. Potential future research based on this thesis includes the following:

6.2.1 Scene Recognition by Considering the Contextual Information of Local Patches

Another way to obtain more discriminative visual words in the vocabulary is to define contextual visual words. Very few publications exist in the area of extracting the contextual¹ information of image's regions of interest. We believe that by extracting the contextual information we can increase the recognition accuracy. The motivation for using contextual information for each of the local patches is based on our belief that the relationships of local patches with other patches in their neighborhood, as well as their structures, can be considered a representation of patches' prototypes; this representation makes the features more discriminative for building visual vocabularies. So far, most of the dictionary-based approaches that exploit contextual information to enrich the image representation, provide contextual information for the framework after building the visual vocabulary; we propose to bring this information into account for building contextual visual words [128]. To this end, we propose a pre-processing step to segment the scene using one of the automatic segmentation algorithms such as Watersheds algorithm [129], normalized cuts [28] or graph cuts [130]. Each Region-of-Interest (ROI) is considered as the context to the local features it contains. The idea of capturing contextual information from image patches slightly contradicts the idea of feature invariance, but when the target application is scene recognition this is no longer an issue. Scene recognition based on particular properties of scene images [21] decreases the importance of geometrically invariant features.

6.2.2 Determining the Number of Topics in the Latent Semantic Space

In projecting the image features into the Latent Semantic Space (concept space) by LSA or pLSA, the number of latent topics is an important factor. As mentioned previously, one of the significant advantages of our framework (compared to former approaches which

¹There are several approaches that suggest the usage of contextual information for object recognition. In these approaches, the context of the object means having a global representation of the image. This representation helps to determine which scene category the object belongs to. In this way, the global context of the image improves the recognition rate by bringing this representation into account. It should be noted that this approach is completely different from the method we have suggested.

also project features to the semantic space) is that we can change the number of image topics to achieve an appropriate number of aspects in latent space, instead of fixing the number of topics to the number of image classes. By increasing the number of topics, we can increase the accuracy. However, we have found that the most appropriate numbers of topics (determined by the number of classes and other factors) should be applied to achieve the best accuracy. These numbers can be obtained experimentally.

We suggest investigating the changes in the number of topics with corresponding changes in the parameters, different feature descriptors, e.g. descriptors of lower dimensionality, the density of image sampling, number of visual concepts in the database (by applying it to a much simpler or more complex database). These investigations will be pursued with the objective of finding a possible approach for estimating the number of topics in the latent space and/or possibly formulating other factors to determine a suitable number of topics for a given problem.

6.2.3 Modified Exemplar-Set Selection for Local-Global Feature Fusion (LGFF)

The proposed method to exploit exemplars for the LGFF framework is very straightforward. Although it leads to promising results, we believe the following approaches for exemplar-set selection will outperform its performance:

- (i) The methods proposed in this thesis do not consider any pre-processing steps such as segmentation, object detection or object localization. Sometimes it is not possible to separate the scenes, e.g. *offices* and *laboratories* have very similar features and are only differentiable by considering the objects inside the images (e.g. computers or lab equipments such as microscopes, etc.) and their locations. Using objects for scene recognition shows improvements in the results [90]. Hence, for the exemplar-set, we suggest the most discriminative images within each category to be chosen based on the co-occurrence and the correlations of their key (rigid) objects.
- (ii) We propose implementing a merge-based exemplar-set selection algorithm inspired by Chang's algorithm [131], which Perner [132] has indicated shows a promising

performance for selecting prototypes in catalogue-based image classifiers. The basic idea is to replace a batch of images with a representative while performing a K -NN classification to assure that this replacement approximately maintains the classification performance. The advantage of implementing this algorithm is that, instead of indicating the appropriate size for the exemplar-set by measuring the final error-rate, the number of exemplars is obtained automatically.

6.2.4 Adopting other optimization techniques for parameter learning

In the Local-Global Feature Fusion (LGFF) and Spatial-LGFF methods, in order to optimize the objective function, the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method, which is an iterative gradient-descent optimization, is employed. We impose the constraints by adding them to the cost-function. Since we have non-negativity constraints in our method, it is worth considering constrained optimization methods, such as Karush-Kuhn-Tucker (KKT) [123, 124] conditions, along with our method. The advantage of the Karush-Kuhn-Tucker method is the simplicity of applying equality and inequality constraints. i.e. for objective function $f(x)$:

$$\begin{aligned} & \min_{x \in R^n} f(x) \\ \text{subject to : } & h_i(x) \leq 0, \quad \forall i = 1, \dots, m, \\ & l_j(x) = 0, \quad \forall j = 1, \dots, r. \end{aligned} \tag{6.1}$$

Publications

- (i) E. Farahzadeh, T-J. Cham, A. Sluzek, "Scene Recognition using Semantic Visual Words, "submitted to *Springer journal of Signal, Image and Video Processing*.
- (ii) E. Farahzadeh, T-J. Cham, W. Li, "Semantic and Spatial Content Capture for Scene Recognition, " Accepted to *Springer collection of New Developments in Robot Vision*, 2014.
- (iii) E. Farahzadeh, T-J. Cham, W. Li, "Incorporating Local and Global Information using a Novel, Distance Function for Scene Recognition," *IEEE Workshop on Robot Vision, Winter Vision Meetings (WVM)*, pp.132-137, 2013.
- (iv) B. Saghafi, E. Farahzadeh, D. Rajan, A. Sluzek, "Embedding Visual Words into Concept Space for Action and Scene Recognition," *British Machine Vision Conference (BMVC)*, pp. 15.1-11, 2010.
- (v) A.Sluzek, D.Yang, M.Paradowski, M.S. Islam, E.Farahzadeh, "Novel Results in Detection, Analysis and Classification of Information in Visual Data," *Information and Management Sciences (IMS)*, 2010.

References

- [1] B. Saghafi, E. Farahzadeh, D. Rajan, and A. Sluzek, “Embedding visual words into concept space for action and scene recognition,” in *British Machine Vision Conference (BMVC)*, 2010.
- [2] A. Sluzek, M. Paradowski, Y. Duanduan, M. Islam, and E. Farahzadeh, “Novel results in detection, analysis and classification of information in visual data,” in *International Conference on Information and Management Sciences*, 2010.
- [3] E. Farahzadeh, T.-J. Cham, and W. Li, “Incorporating local and global information using a novel distance function for scene recognition,” in *IEEE Workshop on Robot Vision, Winter Vision Meetings (WVM)*, 2013.
- [4] E. Farahzadeh, T.-J. Cham, and W. Li, “Semantic and spatial content capture for scene recognition,” in *New Developments in Robot Vision*, Springer, 2014.
- [5] E. Farahzadeh, T.-J. Cham, and A. Sluzek, “Scene recognition using semantic visual words.” Submitted to Springer journal of Signal, Image and Video Processing, November 2013.
- [6] D. Lowe, “Object recognition from local scale-invariant features,” in *IEEE International Conference on Computer Vision (ICCV)*, 1999.
- [7] A. Quattoni and A. Torralba, “Indoor scene recognition,” in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.
- [8] A. Bosch, A. Zisserman, and X. Muoz, “Scene classification using a hybrid Generative/Discriminative approach,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 4, pp. 712–727, 2008.

REFERENCES

- [9] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.
- [10] Y. Liu, D. Zhang, G. Lu, and W. Ma, “A survey of content-based image retrieval with high-level semantics,” *Pattern Recognition.*, vol. 40, no. 1, pp. 262–282, 2007.
- [11] M. R. Boutell, *Exploiting Context for Semantic Scene Classification*. PhD thesis, University of Rochester, 2006.
- [12] I. Biederman, “Recognition-by-Components: A theory of human image understanding,” *Psychological Review*, vol. 94, no. 2, pp. 115–147, 1987.
- [13] A. Vailaya, M. Figueiredo, A. Jain, and H.-J. Zhang, “Image classification for content-based indexing,” *IEEE Transactions on Image Processing*, vol. 10, no. 1, pp. 117 – 130, 2001.
- [14] C.-K. Chang, C. Siagian, and L. Itti, “Mobile robot vision navigation and localization using gist and saliency,” in *IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2010.
- [15] C. Siagian and L. Itti, “Gist: A mobile robotics application of context-based vision in outdoor environment,” in *IEEE International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2005.
- [16] A. Oliva and A. Torralba, “Modeling the shape of the scene: A holistic representation of the spatial envelope,” *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.
- [17] J.-H. Lim, Y. Li, Y. You, and J.-P. Chevallet, “Scene recognition with camera phones for tourist information access,” in *IEEE International Conference on Multimedia and Expo (ICME)*, 2007.
- [18] “Nikon’s scene recognition system (SRS).” <http://www.nikonusa.com/en/Learn-And-Explore/Article/ftlzi4rr/Scene-Recognition-System.html>. Online; Accessed 11-Jan-2013.

REFERENCES

- [19] R. Datta, D. Joshi, J. Li, and J. Z. Wang, “Image retrieval: Ideas, influences, and trends of the new age,” *ACM Computing Surveys*, vol. 40, no. 2, pp. 1–60, 2008.
- [20] N. Vassilieva, “Content-based image retrieval methods,” *Programming and Computer Software*, vol. 35, no. 3, pp. 158–180, 2009.
- [21] S. Lazebnik, *Local, Semi-Local and Global Models for Texture, Object and Scene Recognition*. PhD thesis, University of Illinois at Urbana-Champaign, 2006.
- [22] L. Fei-Fei and P. Perona, “A Bayesian hierarchical model for learning natural scene categories,” in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.
- [23] K. Hirata and T. Kato, “Query by visual example,” in *International Conference on Extending Database Technology*, 1992.
- [24] J. Eakins and M. Graham, *Content-based Image Retrieval*. PhD thesis, Technical Report, University of Northumbria at Newcastle, 1999.
- [25] L. Shao, “An efficient local invariant region detector for image retrieval,” in *Canadian Conference on Computer and Robot Vision*, 2009.
- [26] H. Feng, D. Castanon, and W. Karl, “A curve evolution approach for image segmentation using adaptive flows,” in *IEEE International Conference on Computer Vision (ICCV)*, 2001.
- [27] W. Ma and B. Manjunath, “Edge flow: A framework of boundary detection and image segmentation,” in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 1997.
- [28] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.
- [29] D. Comaniciu and P. Meer, “Robust analysis of feature spaces: color image segmentation,” in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 1997.

- [30] Y. Deng and B. Manjunath, “Unsupervised segmentation of color-texture regions in images and video,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 8, pp. 800–810, 2001.
- [31] V. Mezaris, I. Kompatsiaris, and M. Strintzis, “An ontology approach to object-based image retrieval,” in *IEEE International Conference on Image Processing (ICIP)*, 2003.
- [32] T. Tuytelaars and K. Mikolajczyk, “Local invariant feature detectors: a survey,” *Foundations and Trends in Computer Graphics and Vision*, vol. 3, no. 3, pp. 177–280, 2008.
- [33] J. Winn, A. Criminisi, and T. Minka, “Object categorization by learned universal visual dictionary,” in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005.
- [34] T. Tuytelaars and C. Schmid, “Vector quantizing feature space with a regular lattice,” in *IEEE International Conference on Computer Vision (ICCV)*, 2007.
- [35] R. Maree, P. Geurts, J. Piater, and L. Wehenkel, “Random subwindows for robust image classification,” in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.
- [36] C. Harris and M. Stephens, “A combined corner and edge detection,” in *Alvey Vision Conference*, 1988.
- [37] K. Mikolajczyk and C. Schmid, “Scale & affine invariant interest point detectors,” *International Journal of Computer Vision*, vol. 60, no. 1, pp. 63–86, 2004.
- [38] A. Vedaldi and S. Soatto, “Features for recognition: viewpoint invariance for non-planar scenes,” in *IEEE International Conference on Computer Vision (ICCV)*, 2005.
- [39] E. Rosten and T. Drummond, “Fusing points and lines for high performance tracking,” in *IEEE International Conference on Computer Vision (ICCV)*, 2005.

REFERENCES

- [40] E. Rosten and T. Drummond, “Machine learning for high-speed corner detection,” in *European Conference on Computer Vision (ECCV)*, Springer-Verlag, 2006.
- [41] T. Kadir and M. Brady, “Saliency, scale and image description,” *International Journal of Computer Vision*, vol. 45, no. 2, pp. 83–105, 2001.
- [42] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, “Speeded-up robust features (SURF),” *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.
- [43] T. Tuytelaars and L. Van Gool, “Wide baseline stereo matching based on local, affinely invariant regions,” in *British Machine Vision Conference (BMVC)*, 2000.
- [44] T. Tuytelaars and L. Van Gool, “Matching widely separated views based on affine invariant regions,” *International Journal of Computer Vision*, vol. 59, no. 1, pp. 61–85, 2004.
- [45] J. Matas, O. Chum, U. Martin, and T. Pajdla, “Robust wide baseline stereo from maximally stable extremal regions,” in *British Machine Vision Conference (BMVC)*, 2002.
- [46] K. Mikolajczyk and C. Schmid, “A performance evaluation of local descriptors,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, 2005.
- [47] A. Johnson and M. Hebert, “Recognizing objects by matching oriented points,” in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 1997.
- [48] R. Zabih and J. Wood II, “Non-parametric local transforms for computing visual correspondence,” in *European Conference on Computer Vision (ECCV)*, 1994.
- [49] D. Gabor, “Communication theory and physics,” *Transactions of the IRE Professional Group on Information Theory*, vol. 1, no. 1, pp. 48–59, 1953.

REFERENCES

- [50] L. M. J. Florack, B. M. ter Haar Romeny, J. J. Koenderink, and M. A. Viergever, “General intensity transformations and differential invariants,” *Mathematical Imaging and Vision*, vol. 4, no. 2, pp. 171–187, 1994.
- [51] J. J. Koenderink and A. J. van Doorn, “Representation of local geometry in the visual system,” *Biological Cybernetics*, vol. 55, no. 6, pp. 367–375, 1987.
- [52] F. Schaffalitzky and A. Zisserman, “Multi-view matching for unordered image sets, or How Do I Organize My Holiday Snaps?,” in *European Conference on Computer Vision (ECCV)*, 2002.
- [53] W. Freeman and E. Adelson, “The design and use of steerable filters,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 9, pp. 891–906, 1991.
- [54] D.-G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [55] D. Androutsos, K. Plataniotis, and A. Venetsanopoulos, “Distance measures for color image retrieval,” in *IEEE International Conference on Image Processing (ICIP)*, 1998.
- [56] Z. Chen and B. Zhu, “Some formal analysis of rocchio’s similarity-based relevance feedback algorithm,” *Information Retrieval*, vol. 5, no. 1, pp. 61–86, 2002.
- [57] B. Li, E. Chang, and C.-T. Wu, “DPF - a perceptual distance function for image retrieval,” in *IEEE International Conference on Image Processing (ICIP)*, 2002.
- [58] N. Vasconcelos and A. Lippman, “A multiresolution manifold distance for invariant image similarity,” *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 127–142, 2005.
- [59] S. Ardizzoni, I. Bartolini, and M. Patella, “Windsurf: region-based image retrieval using wavelets,” in *International Workshop on Database and Expert Systems Applications*, 1999.

- [60] Y. Rubner, C. Tomasi, and L. Guibas, “A metric for distributions with applications to image databases,” in *IEEE International Conference on Computer Vision (ICCV)*, 1998.
- [61] J. Li, J. Z. Wang, and G. Wiederhold, “IRM: integrated region matching for image retrieval,” in *ACM International Conference on Multimedia Retrieval (ICMR)*, 2000.
- [62] F. Jing, M. Li, L. Zhang, H.-J. Zhang, and B. Zhang, “Learning in region-based image retrieval,” in *ACM International Conference on Image and Video Retrieval (CIVR)*, 2003.
- [63] S. Kulkarni and B. Verma, “Fuzzy logic based texture queries for CBIR,” in *IEEE International Conference on Computational Intelligence and Multimedia Applications (ICCIMA)*, 2003.
- [64] J. Wang, W.-J. Yang, and R. Acharya, “Color clustering techniques for color-content-based image retrieval from image databases,” in *IEEE International Conference on Multimedia Computing and Systems (ICMCS)*, 1997.
- [65] Y. Li, L. G. Shapiro, and J. A. Bilmes, “A generative/discriminative learning algorithm for image classification,” in *IEEE International Conference on Computer Vision (ICCV)*, 2005.
- [66] Y. Li, J. A. Bilmes, and L. G. Shapiro, “Object class recognition using images of abstract regions,” in *IEEE International Conference on Pattern Recognition (ICPR)*, 2004.
- [67] L. Fei-Fei, R. Fergus, and P. Perona, “Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories,” *Computer Vision and Image Understanding*, vol. 106, no. 1, pp. 59–70, 2007.
- [68] X. S. Zhou and T. S. Huang, “Relevance feedback in image retrieval: A comprehensive review,” *Multimedia Systems*, vol. 8, no. 6, pp. 536–544, 2003.

- [69] S.-F. Chang, W. Chen, and H. Sundaram, “Semantic visual templates: linking visual features to semantics,” in *IEEE International Conference on Image Processing (ICIP)*, 1998.
- [70] J. Wu, *Visual Place Categorization*. PhD thesis, Georgia Institute of Technology, 2009.
- [71] P. Espinace, T. Kollar, A. Soto, and N. Roy, “Indoor scene recognition through object detection,” in *IEEE International Conference on Robotics and Automation (ICRA)*, 2010.
- [72] A. Vailaya, A. Jain, and H. J. Zhang, “On image classification: city vs. landscape,” in *IEEE International Workshop on Content-Based Access of Image and Video Libraries*, 1998.
- [73] M. Szummer and R. Picard, “Indoor-outdoor image classification,” in *IEEE International Workshop on Content-Based Access of Image and Video Database*, 1998.
- [74] I. Ulrich and I. Nourbakhsh, “Appearance-based place recognition for topological localization,” in *IEEE International Conference on Robotics and Automation (ICRA)*, 2000.
- [75] P. Blaer and P. Allen, “Topological mobile robot localization using fast vision techniques,” in *IEEE International Conference on Robotics and Automation (ICRA)*, 2002.
- [76] A. Pronobis, B. Caputo, P. Jensfelt, and H. Christensen, “A discriminative approach to robust visual place recognition,” in *IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2006.
- [77] J. Qin and N. Yung, “Feature fusion within local region using localized maximum-margin learning for scene categorization,” *Pattern Recognition*, vol. 45, no. 4, pp. 1671–1683, 2012.
- [78] S. Paek and S.-F. Chang, “A knowledge engineering approach for image classification based on probabilistic reasoning systems,” in *IEEE International Conference on Multimedia and Expo (ICME)*, 2000.

REFERENCES

- [79] C. Fredembach, M. Schroder, and S. Susstrunk, “Eigenregions for image classification,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 12, pp. 1645 – 1649, 2004.
- [80] A. Mojsilovic, J. Gomes, and B. Rogowitz, “Isee: Perceptual features for image library navigation,” in *SPIE Human Vision and Electronic Imaging*, 2002.
- [81] G. Csurka, C.-R. Dance, L. Fan, J. Willamowski, and C. Bray, “Visual categorization with bags of keypoints,” in *International Workshop on Statistical Learning in Computer Vision, ECCV*, 2004.
- [82] P. Quelhas, F. Monay, J.-m. Odobez, D. Gatica-perez, T. Tuytelaars, and L. Van Gool, “Modeling scenes with local descriptors and latent aspects,” in *IEEE International Conference on Computer Vision (ICCV)*, 2005.
- [83] A. Bosch, A. Zisserman, and X. Muñoz, “Scene classification via pLSA,” in *European Conference on Computer Vision (ECCV)*, 2006.
- [84] D.-M. Blei, A.-Y. Ng, and M.-I. Jordan, “Latent dirichlet allocation,” *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.
- [85] J. Liu and M. Shah, “Scene modeling using co-clustering,” *IEEE International Conference on Computer Vision (ICCV)*, 2007.
- [86] J. Liu, Y. Yang, and M. Shah, “Learning semantic visual vocabularies using diffusion distance,” in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.
- [87] J. Wu and J. M. Rehg, “CENTRIST: A visual descriptor for scene categorization,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1489–1501, 2011.
- [88] R. Kwitt, N. Vasconcelos, and N. Rasiwasia, “Scene recognition on the semantic manifold,” in *European Conference on Computer Vision (ECCV)*, 2012.

- [89] A. Oliva and A. Torralba, “Building the gist of a scene: The role of global image features in recognition,” *Visual Perception, Progress in Brain Research*, vol. 155, pp. 23 – 36, 2006.
- [90] L.-J. Li, H. Su, E. P. Xing, and L. Fei-Fei, “Object bank: A high-level image representation for scene classification and semantic feature sparsification,” in *Neural Information Processing Systems (NIPS)*, 2010.
- [91] X. Li and Y. Guo, “An object co-occurrence assisted hierarchical model for scene understanding,” in *British Computer Vision Conference (BMVC)*, 2012.
- [92] M. Pandey and S. Lazebnik, “Scene recognition and weakly supervised object localization with deformable part-based models,” in *IEEE International Conference on Computer Vision (ICCV)*, 2011.
- [93] S. Parizi, J. Oberlin, and P. Felzenszwalb, “Reconfigurable models for scene recognition,” in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [94] S. Singh, A. Gupta, and A. A. Efros, “Unsupervised discovery of mid-level discriminative patches,” in *European Conference on Computer Vision (ECCV)*, 2012.
- [95] D. M. P. F. Felzenszwalb, R. B. Girshick and D. Ramanan, “Object detection with discriminatively trained part-based models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.
- [96] J. Tighe and S. Lazebnik, “Understanding scenes on many levels,” in *IEEE International Conference on Computer Vision (ICCV)*, 2011.
- [97] C. Liu, J. Yuen, and A. Torralba, “Nonparametric scene parsing via label transfer,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2368–2382, 2011.
- [98] T. Hofmann, “Unsupervised learning by probabilistic latent semantic analysis,” *Machine Learning*, vol. 42, no. 1-2, pp. 177–196, 2001.

- [99] M. Brown and S. Susstrunk, “Multi-spectral SIFT for scene category recognition,” in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.
- [100] I.-H. Jhuo and D. T. Lee, “Boosted multiple kernel learning for scene category recognition,” in *IEEE International Conference on Pattern Recognition (ICPR)*, 2010.
- [101] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba, “SUN database: Large-scale scene recognition from abbey to zoo,” in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.
- [102] G. Patterson and J. Hays, “SUN attribute database: Discovering, annotating, and recognizing scene attributes,” in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [103] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan, “Learning the kernel matrix with semidefinite programming,” *Journal of Machine Learning Research*, vol. 5, pp. 27–72, 2004.
- [104] M. Varma and A. Zisserman, “A statistical approach to texture classification from single images,” *International Journal of Computer Vision*, vol. 62, no. 1, pp. 61–81, 2005.
- [105] J. Sivic, B.-C. Russell, A.-A. Efros, A. Zisserman, and W.-T. Freeman, “Discovering objects and their location in images,” in *IEEE International Conference on Computer Vision (ICCV)*, 2005.
- [106] J.-C. Niebles, H. Wang, and L. Fei-Fei, “Unsupervised learning of human action categories using spatial-temporal words,” *International Journal of Computer Vision*, vol. 79, no. 3, pp. 299–318, 2008.
- [107] J. Liu and M. Shah, “Learning human action via information maximization,” in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

REFERENCES

- [108] J. Vogel and B. Schiele, “Natural scene retrieval based on a semantic modeling step,” in *ACM International Conference on Image and Video Retrieval (CIVR)*, 2004.
- [109] F. Moosmann, B. Triggs, and F. Jurie, “Fast discriminative visual codebooks using randomized clustering forests,” in *Neural Information Processing Systems (NIPS)*, 2006.
- [110] R. C. J. Shotton, M. Johnson, “Semantic texton forests for image categorization and segmentation,” in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.
- [111] E. Nowak, F. Jurie, and B. Triggs, “Sampling strategies for bag-of-features image classification,” in *European Conference on Computer Vision (ECCV)*, 2006.
- [112] S. Deerwester, S.-T. Dumais, G.-W. Furnas, T.-K. Landauer, and R. Harshman, “Indexing by latent semantic analysis,” *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.
- [113] D. Liu and T. Chen, “Unsupervised image categorization and object localization using topic models and correspondences between images,” *IEEE International Conference on Computer Vision (ICCV)*, 2007.
- [114] M. J. Swain and D. H. Ballard, “Color indexing,” *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991.
- [115] F. Odone, A. Barla, and A. Verri, “Building kernels from binary strings for image matching,” *IEEE Transactions on Image Processing*, vol. 14, pp. 169–180, Feb. 2005.
- [116] J. Yang and Z. Xu, “Feature-level fusion of global and local features for finger-vein recognition,” in *IEEE International Conference on Signal Processing (ICSP)*, 2010.
- [117] W. J. Zhao, Y. Fan, “Fusion of global and local feature using KCCA for automatic target recognition,” in *IEEE International Conference on Image and Graphics (ICIG)*, 2009.

REFERENCES

- [118] Q. Sun, S. Zeng, Y. Liu, P. Heng, and D. Xia, “A new method of feature fusion and its application in image recognition,” *Pattern Recognition*, vol. 38, no. 12, pp. 2437–2448, 2005.
- [119] H. Hotelling, “Relations between two sets of variables,” *Biometrika.*, vol. 28, no. 3/4, pp. 321–377, 1936.
- [120] D. R. Hardoon, S. R. Szedmak, and J. R. Shawe-taylor, “Canonical correlation analysis: An overview with application to learning methods,” *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.
- [121] C. G. Broyden, “The convergence of a class of double-rank minimization algorithms 1. general considerations,” *IMA Journal of Applied Mathematics*, vol. 6, no. 1, pp. 76–90, 1970.
- [122] W. Davidon, “Variable metric method for minimization,” Technical report III, Argonne National Laboratories, 1959.
- [123] W. Karush, *Minima of functions of several variables with inequalities as side constraints*. PhD thesis, Masters thesis, Dept. of Mathematics, University of Chicago, 1939.
- [124] H. Kuhn and A. Tucker, “Nonlinear programming,” in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, California, 1951.
- [125] A. Rakotomamonjy and S. Canu, “More efficiency in multiple kernel learning,” in *International Conference on Machine Learning*, 2007.
- [126] K. Grauman and T. Darrell, “The pyramid match kernel: Discriminative classification with sets of image features,” in *IEEE International Conference on Computer Vision (ICCV)*, 2005.
- [127] J. Wu, “Power mean SVM for large scale visual classification,” in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

REFERENCES

- [128] J. Qin and N. Yung, “Scene categorization via contextual visual words,” *Pattern Recognition*, vol. 43, no. 5, pp. 1874–1888, 2010.
- [129] L. Vincent and P. Soille, “Watersheds in digital spaces: an efficient algorithm based on immersion simulations,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 6, pp. 583–598, 1991.
- [130] Y. Boykov and M.-P. Jolly, “Interactive graph cuts for optimal boundary & region segmentation of objects in ND images,” in *IEEE International Conference on Computer Vision (ICCV)*, 2001.
- [131] C. Chang, “Finding prototypes for nearest neighbor classifiers,” *IEEE Transactions on Computers*, vol. 100, no. 11, pp. 1179–1184, 1974.
- [132] P. Perner, “Prototype-based classification,” *Applied Intelligence*, vol. 28, no. 3, pp. 238–246, 2008.