TAMPERE UNIVERSITY OF TECHNOLOGY

Wei Liu

SCENE IMAGE CLASSIFICATION AND RETRIEVAL

Master of Science Thesis

Examiners: Prof. Moncef Gabbouj
and Prof. Serkan Kiranyaz

Examiners and topic approved in the Faculty of Computing and Electrical Engineering Council meeting on 7 December 2011.

PREFACE

This research was sponsored by Alma project and conducted within the MUVIS group in the Department of Signal Processing, Tampere University of Technology, Finland.

First and foremost, a few words of appreciation to my supervisors are in order. I am unequivocally grateful to Professor Serkan Kiranyaz and Professor Moncef Gabbouj for entrusting me with this project. Their continuous encouragement and support have hardened my motivation in the research process. More important, I was offered significant creative freedom which is elemental in the formation of this thesis.

Additionally, I really appreciate the quiet and soothing working environment that all members of the MUVIS group manage to build. Specifically, I would like to thank Jenni Pulkkinen and Stefan Uhlmann for their presentations on particle swarm optimization (PSO) and collective network of binary classifiers (CNBC). And during the initial training process, Guanqun Cao has provided valuable technical support on lab equipment and his knowledge on computer vision is also highly appreciated.

Last but not least, I hope my fellow colleagues Weiyi “Tommi” Xie and Xiaobi “Bree” Xu enjoy the short but productive research assistantship as much as I did since the experience wouldn’t be as such without them.

Tianjin, April 2012.

Wei Liu

DingZiGu, Hongqiao District
300130 Tianjin, China

ABSTRACT

TAMPERE UNIVERSITY OF TECHNOLOGY

Degree Programme in Information Technology

Liu, Wei: Scene Image Classification and Retrieval

Master of Science Thesis, 94 pages, 2 Appendix pages

April 2012

Major subject: Signal Processing

Examiner: Professor Moncef Gabbouj and Professor Serkan Kiranyaz

Keywords: image feature, spatial envelope, gist, scene classification, scene retrieval

Scene image classification and retrieval not only have a great impact on scene image management, but also they can offer immeasurable assistance to other computer vision problems, such as image completion, human activity analysis, object recognition etc. Intuitively scene identification is correlated to recognition of objects or image regions, which prompts the notion to apply local features to scene categorization applications. Even though the adoption of local features in these tasks has yielded promising results, a global perception on scene images is also well-conditioned in cognitive science studies. Since the global description of a scene imposes less computational burden, it is favoured by some scholars despite its less discriminative capacity. Recent studies on global scene descriptors have even yielded classification performance that rivals results obtained by local approaches.

The primary objective of this work is to tackle two of the limitations of existing global scene features: representation ineffectiveness and computational complexity. The thesis proposes two global scene features that seek to represent finer scene structures and reduce the dimensionality of feature vectors. Experimental results show that the proposed scene features exceed the performance of existing methods.

The thesis is roughly divided into two parts. The first three chapters give an overview on the topic of scene image classification and retrieval methods, with a special attention to the most effective global scene features. In chapter 4, a novel scene descriptor, called ARP-GIST, is proposed and evaluated against the existing methods to show its ability to detect finer scene structures. In chapter 5, a low-dimensional scene feature, GIST-LBP, is proposed. In conjunction with a block ranking approach, the GIST-LBP feature is tested on a standard scene dataset to demonstrate its state-of-the-art performance.

CONTENTS

List of Tables

List of Figures

Abbreviations and Acronyms

1-D	One Dimensional
3-D	Three Dimensional
AI	Artificial Intelligence
ANMRR	Averaged Normalized Modified Retrieval Rate
ARP	Angular Radial Partitioning
BoW	Bag of Words
CBIR	Content Based Image Retrieval
CENTRIST	Census Transform Histogram
CIE	Commission Internationale de L'éclairage - International Commission on Illumination
CT	Census Transform
DFT	Discrete Fourier Transform
EBR	Edge Based Regions
EBSR	Entropy Based Salient Region detector
IBR	Intensity Based Regions
KLT	Karhunen-Loeve Transform
KNN	K Nearest Neighbour classification
LBP	Local Binary Pattern
LDA	Latent Dirichlet Allocation
MPEG	Moving Picture Experts Group
MSER	Maximally Stable Extremal Regions
NCA	Neighbourhood Components Analysis
PCA	Principal Component Analysis
RBF	Radial Basis Function
SBIR	Sketch Based Image Retrieval
SE	Spatial Envelope
SIFT	Scale-Invariant Feature Transform
SVD	Single Value Decomposition
SVM	Support Vector Machine
VQ	Vector Quantization
WFT	Windowed Fourier Transform

1. INTRODUCTION

In the field of computer vision, image classification and retrieval have been two active research themes for the past few decades since in a general sense, the objective of computer vision research is to emulate the human visual system and make sense out of visual objects. And for us human, the rationalization of our visual world is to recognize or classify, i.e., when one is searching for a container to pour wine, the recognition of glasses is made possible through the power of visual perception in correspondence with analytical abilities of our brains.

An image retrieval or classification system usually requires the assistance of several components, including image analysis or feature extraction module, dimensionality reduction or feature selection unit, and/or machine learning system. Figure 1.1 shows the generalization of such a system. It is evident that each component stems from a specialized area of research: a.) the feature extraction component is a direct application of advances in image, or more generally, signal processing. It deals with efficient representation and analysis of a given image and transforming it into a digital form that can be conveniently further processed by a computer algorithm; b.) the output from the initial feature extraction phase can be very high dimensional and not all dimensions in the feature carry the same weight with respect to retrieval or classification. Therefore, proper feature selection scheme is needed to reduce the dimensions of feature vectors for computational simplicity and accentuate the features that are most useful for the task. Generally speaking, feature selection algorithms can be divided into two categories—supervised and unsupervised. In a supervised setting, such as Neighbourhood Components Analysis (NCA) [47], class labels are taken into consideration in the selection process so that the final features not only account for a small portion of dimensions of the original feature, but also present good clustering properties in the transformed feature space. In contrast, an unsupervised feature selection algorithm does not involve prior knowledge of class labels. Rather, its primary concern is to compact most information carried by the original data into a few variables so that this limited subset of variables can reconstruct the original data with a small error. The simplest and most widely used form of unsupervised feature selection techniques is Principal Component Analysis (PCA) which will be presented in detail as it is an essential part of implementation for this thesis work; c.) once the low-level features are obtained for each image, they are either fed through a machine learning algorithm or simply ranked with respect to a specified distance measure. In the former case, the primary concern is to label data points using a trained classifier, obtained through a specific machine learning algorithm, including K Nearest Neighbour classification (KNN) [25], Support Vector Machine

(SVM) [12], among others. Once the system is fully trained, it is able to extract low-level features from a new image, obtain its feature vector and categorize the image into one (single-class) of several predetermined classes. On the other hand, if the primary objective of the system is to rank images with respect to their similarity to a query, a simple distance measure is usually enough to achieve such purpose. The retrieval result is therefore the ranked images based on the feature.

Scene image classification and retrieval can be seen as a special instance of such a system, and the general framework shown in Figure 1.1 is also applicable to this thesis work. This thesis will focus mainly on feature extraction methods, as they are the key contributions of this work which are summarized as follows:

- Angular Radial Partitioning GIST descriptor (ARP-GIST) [78]. A new image feature is proposed for scene classification. As an extension to the original GIST descriptor [99], the proposed scene feature outperforms other popular features or frameworks on standard datasets and it achieves a desirable balance between classification accuracy and computational efficiency.
- GIST-LBP (GIST and Local Binary Pattern [96]) feature [77]. This novel scene feature is based on two of the most discriminative texture features and it leverages the advantages of both. In conjunction with PCA, the proposed feature tabulates only 320 dimensions, ensuring efficient retrieval performance.
- Block ranking scheme [77]. This novel feature processing scheme is designed to further reduce the dimensions of feature vectors, based on the observation that not all regions in one image carry the same weight for accurate retrieval. Through the use of the proposed algorithm, the feature dimensions can be halved and meanwhile retrieval performance can be further boosted.

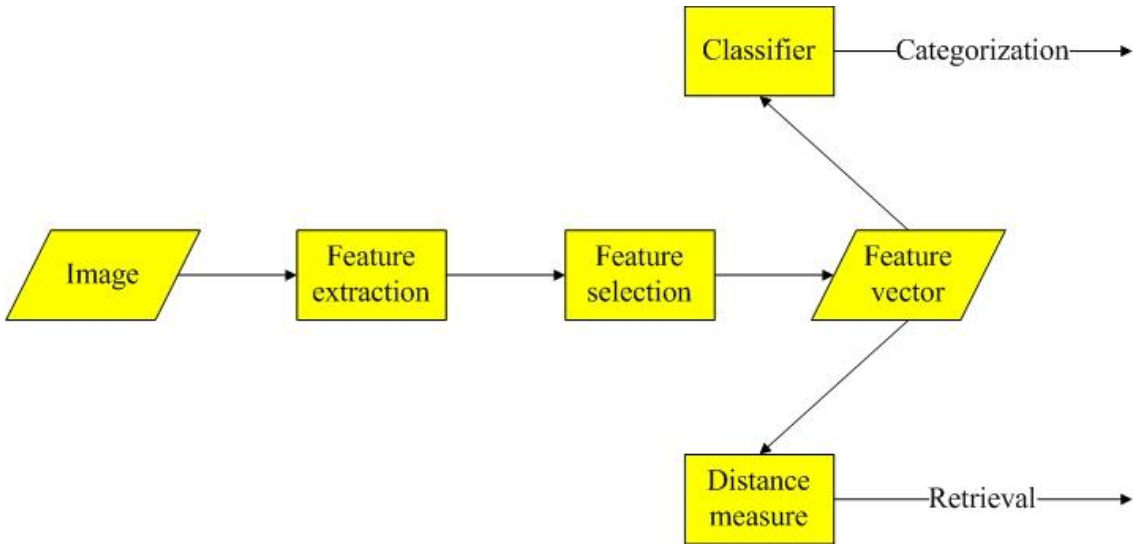


Figure 1.1. *The basic components of an image retrieval or classification system and their roles in the process.*

1.1. Definition and categorization

There is no official definition as to what constitutes a scene image. But the general consensus is that a scene image is empirically defined by the objects a scene contains or the foreground/background regions inside a scene. For example, as shown in Figure 1.2, a coast scene is usually characterized by several regions that depict the ocean, the sky and sometimes the beach; in a highway scene, a very long extending road is always featured as the predominant structure of the image; in a purely man-made scene such as an office, however, the semantic category is defined by many objects shown in the image, including, but not limited to, computers, desks, books and etc. It should be noted that the exhibition of objects, foreground or background regions is not identical within each category. Rather, a certain degree of intra-class variations is observed. It is not uncommon to observe a cruise boat that occupies a relatively significant part of a coast scene just as it is not unusual to spot small mountain ridges along the coastline in images from the same category. Despite such slight variations within each class in natural scenes, these images are often very consistent in visual appearance—again in a coast scene, the sky often occupies the upper portion of the image and the ocean often the middle or lower part. This strictly enforced structural layout—some researchers call this the logic of a scene—is highly articulated in natural scenes and can be very useful in scene retrieval or classification. The man-made scenes, however, may present even more versatility due to their generally unpredictable nature: a desk is almost always present in an office scene, but other items, such as desktops, laptops, telephones, file cabins, portfolios, books, bookshelves and etc., are rarely fixtures in all office scenes. Thanks to such diversity, the logic of a man-made scene is hardly straightforward, making it harder to ensure scene matching.

Figure 1.2. *Sample scene images depicting scenes from coast, highway and office category.*

It should be pointed out that the scene images used in this study are not randomly selected. Instead, they resemble significantly to the works of professional photographers. The camera settings for all images are essentially similar, including focal length, camera

height, depth of field and etc. The camera is almost always set up to be parallel to horizontal level and the field of view largely corresponds to the semantic label of the image and sometimes the majority of semantic information of the scene is concentrated in the field of view. Images from the same semantic category share similar spatial layout—ocean, trees, sky and etc., and these similar structures often present the same scale information.

1.2. Areas of application

The primary reason why this topic has prompted tremendous amount of research effort and publications is due to the fact that in computer vision, the major objective of research is to duplicate or simulate the functions of human visual system and scene recognition is one of such functions. For this purpose, the manner in which the human visual system processes scene images and extracts representative features has been intensively studied both in the area of computer vision and cognitive science [2], [43], [56], [108]. These studies concentrate mostly on low-level image features since these features are elemental in a simple scene classification system [99]. In addition, these features can also be further processed for efficiency concerns [132] or be integrated with other types of features [35] prior to classification or retrieval. In the application of place recognition, low-level scene features are an essential part of the recognition system, delineating the spatial structures of a given scene to complement salient regions in the image [120]. With the advent of concrete improvement in scene image analysis, new advances have spurred novel research interests or aided other studies indirectly by providing information on structural attributes of a scene. In object recognition studies, some researchers have applied scene features to their recognition systems as a way of providing scene context so that objects can be more accurately localized and recognized [34], [113]. In a similar fashion, scene matching is also employed in human activity analysis [58], on the assumption that similar actions tend to take place in similar surroundings. For example, people usually play tennis in a tennis court, so the scene features of a tennis court can be used as the first indicator of this activity. Scene matching is also applicable to computer graphics, in the area of scene completion and image geolocation. In the first pass of scene completion process, Hays et al. [55] have adopted scene features as a criterion to search for semantically similar images from the database according to the examplar image with a missing region. In an image geolocation inference task [54], [65], the semantics of a scene can be of significant importance when it is integrated with other low-level features for predicting the likelihood of the geolocation of a test image.

As evidenced by these applications, scene classification and retrieval studies are the backbone of several research themes in the field of computer vision and graphics. It is obvious that performance improvement in scene matching can better assist the advancement of other research projects.

1.3. Scene abstraction levels and datasets

Oliva et al. [99] follow the studies in psychology [111] and define scene recognition at three levels of abstraction, namely the subordinate level, the basic level and the superordinate level. The objective of scene recognition varies among these three levels, which has a direct impact on the categorization and manual annotation methods of scene image datasets.

At the subordinate level, scene matching is defined with respect to the matching of objects or other types of local structures. The *Blobworld* framework proposed by Carson et al. [17], [18] is an example of such level of abstraction. Due to the local nature of this level of recognition, scene images are often segmented into several small regions that are described specifically by texture and colour. Scene matching proceeds as a process of matching at object level. More specifically, this approach requires the search for similar scenes in terms of configuration and region similarities. Thus, at this level of recognition, the semantics of the whole scene may not be as important as the identities of local details and the retrieved scene images may not belong to the same semantic category. Therefore, there is usually no need for precise manual annotation.

The basic level of recognition requires a coarser and more holistic identification of scene images than the subordinate level. At this level, recognition of local objects is replaced by identification of the function of the scene. To this end, a more global description of scene images is typically used for scene recognition and precedence should be given to the rough configuration of different macro-regions of the scene instead of detailed local structures. Since at this level scene recognition demands the comprehension of the semantics of a scene, images are often categorized according to global meanings: coast, mountain, street, tall building etc. As this corresponds to the requirement of Alma project, the image datasets will be selected according to this level of abstraction.

At the superordinate level, even the basic level of semantics of a scene is ignored. This high level of abstraction only concerns the scene structures at the coarsest resolution. In a sense, it corresponds to the human perception on a scene at a reasonable distance. At such level of abstraction, the variations of spatial properties within each category are more pronounced than those at the other levels. For example, some scholars have explored the research theme of natural/man-made scene separation [60], [71], [79], [114]; while others have dedicated their research work to distinguishing between indoor and outdoor scenes [33], [135], [130], [128], [51], [68], [83], [66], [118], [102].

It is obvious that the level of abstraction for scene identification is highly correlated to the set-up of scene categories for experimentation. And both the approach for manual annotation of scene images and the specification of level of scene recognition clearly influence the type of scene features used for scene classification tasks. Since this thesis is only interested in the basic level of scene recognition, it is only fitting to annotate scene images according to their semantic categories. And it is also important to utilize scene features that extract from scene images the basic level of information that can be easily translated to the description of their global structures.

In order to test the effectiveness and generalization of any proposed algorithm, it is important to conduct carefully devised experiments on publically available datasets so that explicit comparison with other algorithms can be made and objective evaluation of the proposed algorithm can be performed. Following the same spirit, the proposed scene features and algorithms are tested on two publically available scene image datasets in which scene images are categorized according to their semantic meanings. The two scene datasets are:

1.3.1. MIT spatial envelope dataset

The MIT spatial envelope (SE) dataset [99] is the testbed for the original GIST descriptor [99]. It is specifically devised to evaluate the effectiveness of SE properties for scene image recognition and classification. The dataset consists of eight outdoor scene categories: *coast*, *forest*, *highway*, *inside city* (perspective view of urban area), *mountain*, *open country* (perspective view of rural area), *street* and *tall building*. There are 2688 colour images in total with around 300 in each category. All images in the dataset have the same resolution of 256×256. Figure 1.3 shows some sample scene images from this dataset, with one from each category.

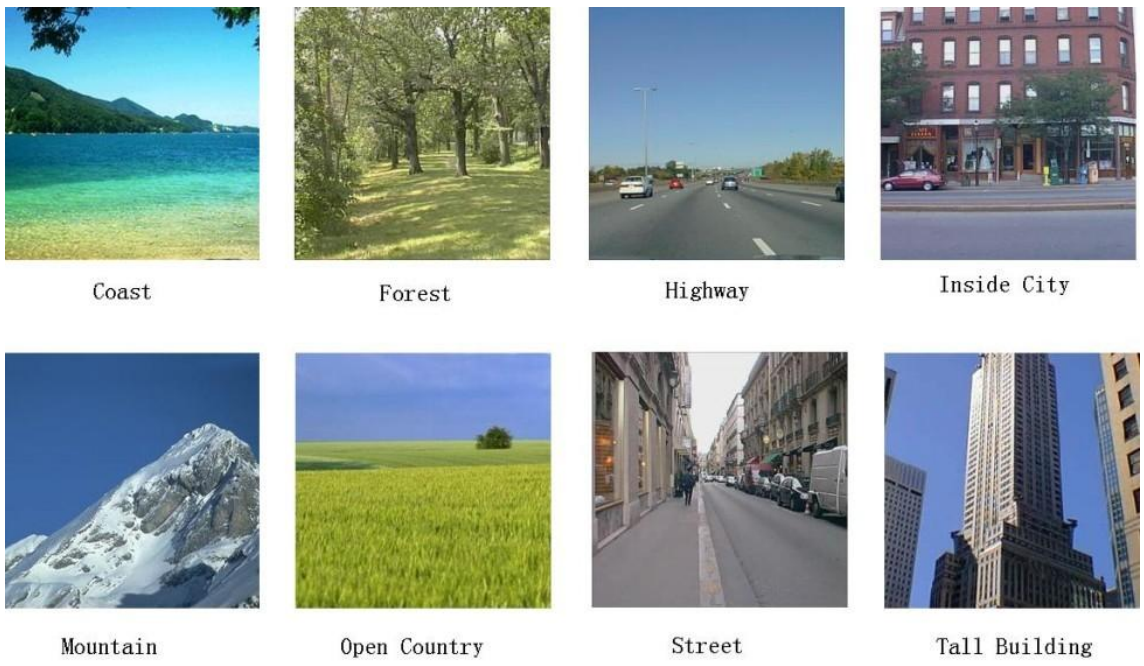


Figure 1.3. *Sample scene images from the Spatial Envelope dataset, with one image from each category.*

Since the primary object of the GIST descriptor is to capture the intrinsic structural characteristics of real-world scenes at the basic level, the minor details in most scene images are not accentuated. In other words, these images are selected such that the resolution of minor objects or structures, when compared to that of the whole scene, is relatively insignificant.

1.3.2. UIUC 15-category dataset

The UIUC 15-category dataset [40] is an extension to the MIT SE dataset. It contains not only all the outdoor scenes shown in Figure 1.3 (all of the MIT SE dataset images are presented in gray-scale), but also some additional indoor and outdoor categories: *bedroom*, *industrial*, *kitchen*, *living room*, *office*, *store* and *suburb*. Most images from these additional categories are also presented in gray-scale. Aspect ratio of these additional images varies within each category and among categories also. And so does the resolution. One image from each of the additional categories is shown in Figure 1.4.

With the inclusion of some indoor scenes, this dataset poses more challenges for scene classification algorithms. Among them is the accommodation for flexibility of spatial positions, since in indoor scenes, significant structures may not always conform to a specific area, e.g., the position of the desk in an office scene can vary greatly from image to image.

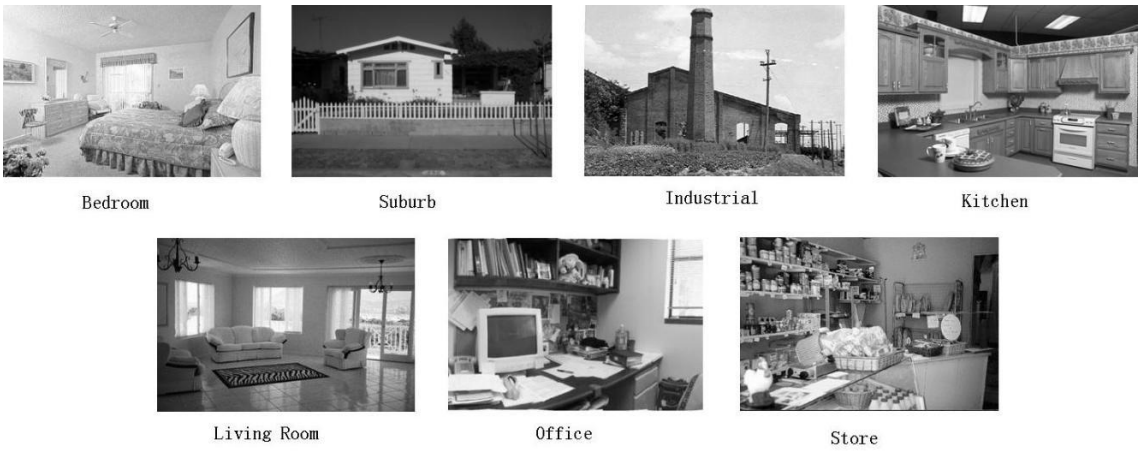


Figure 1.4. *Additional scene categories from the UIUC 15-category dataset, with one image from each category.*

2. SCENE FEATURES AND APPROACHES

The first concern of any computer vision application is image representation—how to delineate an image in a numerical manner so that a computer or other electronic devices can understand efficiently the essence or the semantics of the given image in one or a few aspects, such as colour, shape and etc. Any image representation approach that captures the features of an image is called an image descriptor. And any image descriptor should satisfy the following requirements: effective description and efficient computation.

Effective description requires that the image descriptor used is appropriate for the computer vision task. As each image usually presents several distinctive properties—gray scale property, colour cues, colour layout, texture characteristics, edge information, spatial structures, it is paramount to select the right feature in correspondence with the application requirement. For instance, if the primary objective of a content-based image retrieval (CBIR) system is to allow users to search the image database for forest images or images that contain colourful flowers, colour histogram in the simplest form can be effective enough as the image descriptor. And when it comes to a library of gray scale images, colour cues are out of the question completely. Therefore, the selection of image features can be largely ad-hoc and it is almost impossible to settle on one or a combination of a few image features without any knowledge on application requirements or image properties.

On the other hand, the computational cost of the system should be taken into consideration, especially when there is specific requirement on the amount of resources the system is allowed to consume or how promptly the system is supposed to respond. In early face recognition applications, researchers used only the gray scale (and properly aligned) images as features for face identification [136], [137]. With normal face images measuring to 100×100 in size, the dimensions of face features are in the magnitude of 10 thousand, which would impose compromising computational burden on any face recognition system without feature selection or dimensionality reduction measures. Another reason for such reluctance to use gray scale images directly as features is that gray scale feature is barely robust to luminance, scale, shape or intra-class variations. The pixel-wise matching requirement sets the strictest spatial and numerical template for semantic conformity, which is rarely the case for most CBIR applications. In a typical image library, images that share similarity defined by the goal of a retrieval application usually present non-negligible variations in shape, colour, spatial layout or other image attributes. Therefore, it is a common practice in image processing applications to use image descriptors as a feature extraction measure to capture the properties of images

that are essential for the task. With the summarizing capacity of image descriptors, the image features extracted usually amount to a mere thousand or less. And it is relatively easier to quantize such a feature vector while keeping the essential information intact. Observing images through the help of image descriptors also allow a certain degree of robustness to shape, colour, illumination or other types of variations as image descriptors hardly confine images to pixel-by-pixel matching, providing flexibility in the retrieval process.

Besides detailed elaboration on image features, especially those that pertaining to scene images, this section also covers a concise summary of the prevailing approaches in scene classification and retrieval. In terms of how scene images are perceived, scene features can be categorized as global—processed as a whole or in a few blocks—or local features—images are processed at a local level where only tiny image patches are the basic description units. Generally speaking, scene images (or any other type of images) can be classified using a discriminative or a generative model. In a broad sense, a discriminative model only deals with the estimation of the posterior probability of a certain theme, such as the likelihood of an image depicting a coast theme when the feature of the image is given; whereas for the generative counterpart, the primary concern is to model the likelihood probability as well as the prior probability, which is to estimate the probability of presenting a certain feature given a coast theme image and the possibility of observing the coast theme among all images in the library. More details will be covered in the second half of this chapter.

2.1. Scene feature—colour, shape or texture?

Colour is probably the most widely used image feature in simple image processing or elaborative computer vision applications, either as a standalone descriptor or an integral part of several features. The popularity stems from the rich colour cues observed from real-world pictures. The chromatic property is particularly helpful when the colour information coincides with the semantics or other retrieval/matching criteria of images in the library. In other words, similar images should be distributed closely together in the colour space and general margin among different clusters should be significant enough for reasonable separation. This condition is mostly pronounced in a natural setting, such as a garden or park scene that is most likely to be dominated by the colour of green. While in an artificial scene, chromatic information is more unpredictable and does not necessarily follow a pattern that can be captured by a colour feature. It should be noted that such strict criterion only applies to the situation where perfect retrieval is a de facto requirement in a relatively small image library. In fact, in commercial applications, perfect retrieval is rarely a primary consideration. Especially when the library contains a large volume of images with an ample quantity from each label, colour descriptors are normally capable of returning similar images in the first few nearest neighbours.

The chromatic property has been reasonably employed in the MPEG-7 standard [86] which provides a comprehensive standardized set of tools to give users access to multi-

media content. Formally known as “Multimedia Content Description Interface”, the MPEG-7 standard incorporates numerous advances in multimedia processing available at the time to allow effective and efficient management over an ever-growing supply of multimedia resources. As part of a development by the “Moving Picture Experts Group”, MPEG-7 does not aim, however, to offer users a specialized approach to any multimedia applications. Instead, the standard serves to provide a general solution without any particular application in mind. Aiming to describe as many attributes of multimedia content as possible with efficiency being a primary concern, MPEG-7 has integrated several colour descriptors because of their low dimensionality. These colour descriptors support a number of colour spaces, namely the RGB colour space, YCbCr colour space, HSV colour space, HMMD colour space, linear transformation matrix with reference to RGB and gray scale monochrome colour space. In terms of colour histograms, the MPEG-7 standard not only offers them in several colour spaces, but they can be either uniformly quantized or nonlinearly clustered to form dominant colours. Besides straight-forward colour histograms and standard dominant colour descriptors, MPEG-7 also takes into account the scalability of regular colour histograms. By applying the Haar function to colour histogram in HSV colour space, the scalable colour descriptor is a binary representation of chromatic features using the coefficients of the Haar wavelet transformation. The number of bins selected and the accuracy of binary representation can offer great scalability to this image descriptor. Alternatively, spatiality can also be integrated into chromatic cues, which is the exact intention of the colour layout descriptor in MPEG-7 standard. Instead of counting chromatic cues in each pixel, the colour layout descriptor generates average colour in each block, in conjunction with transforming these blocks into frequency domain, largely compresses the feature dimensions and allows compact representation of colour and spatiality.

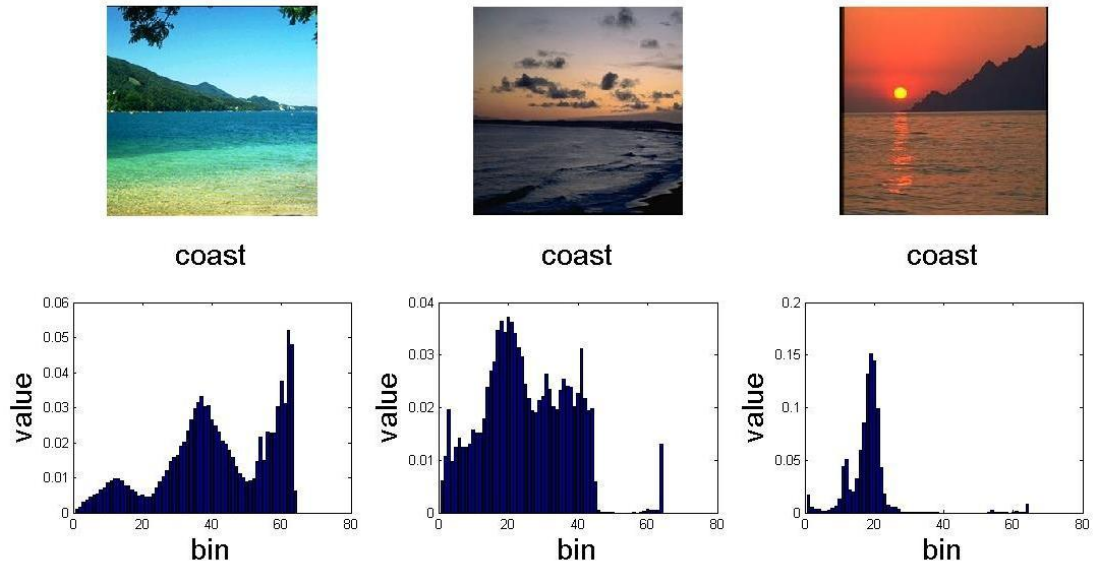


Figure 2.1. *Coast scenes that vary significantly in colour space and their corresponding histograms.*

Although colour features present numerous merits when it comes to image description, such as compactness of image representation, scalability of feature dimensions, invariance to rotation, image size, aspect ratio among others, colour feature of any kind would fail to capture the intricate structures within a scene, especially when semantically similar scenes are captured in different natural settings, resulting in a wide range of colour palettes. It should be noted that in such occasions where the sole purpose is to retrieve semantically similar scenes (preferably in similar chromatic range) in a voluminous scene library, colour features can be used as a wrapper solution due to their efficiency. In this thesis, however, where scene classification is a major concern and there are only a limited number of images in each scene category, a more descriptive feature is needed to discriminate characteristics among different scene structures. The deficiency in discriminative power of colour histogram is illustrated in Figure 2.1. The first row of images are three coast scenes that vary significantly in the colour space, which is evident from their corresponding colour histograms presented on the bottom row. These histograms are derived from the RGB colour space with 4 quantization levels in each chromatic channel, resulting in a 64-bin configuration. Since there is hardly any observable pattern among these histograms and the distances (histogram intersection [126]) between the histogram of the leftmost coast scene and those of the other two are too tangible to ignore, it is almost impossible to use colour histogram as the scene feature to perform classification or retrieval, since scene images that share category membership do not form a cluster in the colour space. In fact, if the first coast scene is used as a query image, a large number of semantically irrelevant images will be ranked higher than those other two coast scenes while using histogram intersection as a distance measure. Some of the top ranked scenes are shown in Figure 2.2, which presents scenes from the mountain, tall building and high way category. None of these top-ranked scenes are visually or semantically similar to the query.

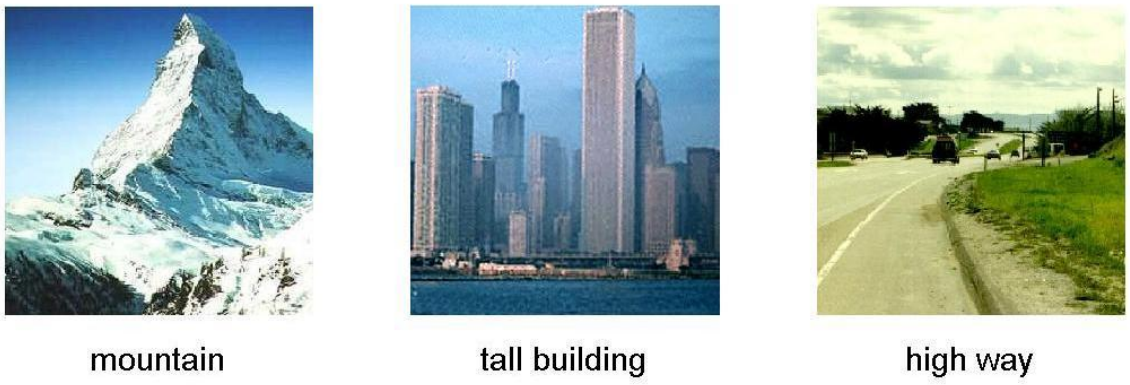


Figure 2.2. *Top-ranked scenes with respect to the coast scene in the previous image as the query image using colour indexing.*

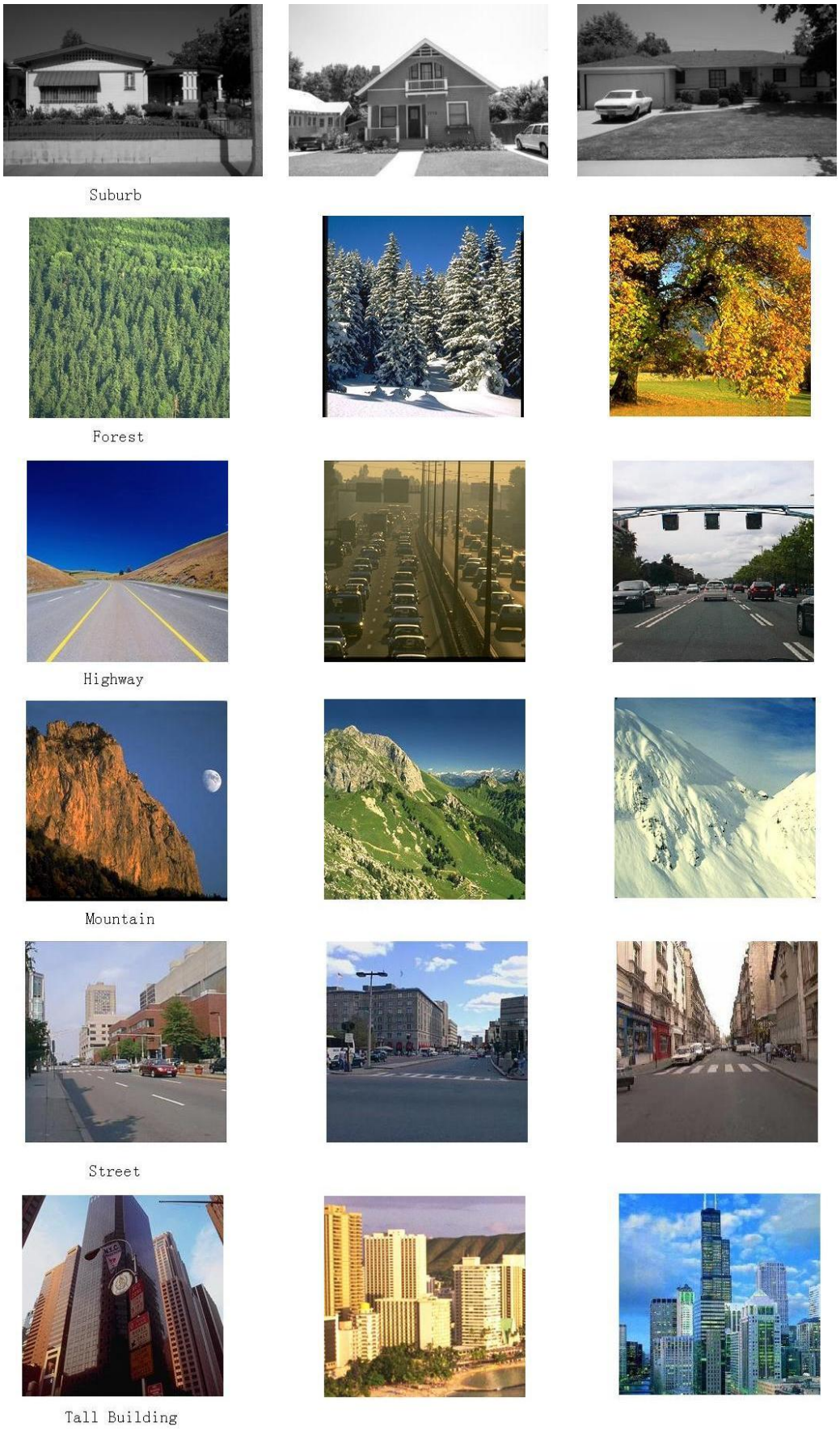


Figure 2.3. *Images from different categories that vary in colour despite presenting similar semantic meaning.*

Through closer observation, however, it is encouraging to notice that the texture of each scene carries most of its semantic information. In other words, the texture of scenes from the same category tends to agree with each other to a noteworthy degree. It is fairly easy to observe from Figure 2.1 that the texture or the quality of material from all three coast scenes shares the same attributes: the watery ocean, the sparsely cloudy sky and the long-range ocean bank. Indeed, this textural property is highly conspicuous in clean and mostly natural scene images. Generally, in a forest scene, the texture of trees dominates the image while the colour of forest scenes can vary from light yellow to dark green; in a suburb scene, the mixed texture of man-made structure (the house) and the lawn and trees characterizes its distinguishable features; even though highway scenes depict strictly man-made structures, they usually exhibit relatively clean contours that are set against the backdrop of the sky, resulting in a unique spatial layout and textural pattern; similar to forest images, mountain scenes show more chromatic variations than most natural environments, which results from the materials—plants or snow—that cover the mountain ridges. The ruggedness of the mountain ranges, however, overwhelms other non-textural properties of these images and can be extracted with texture analysis methods. Of course, some purely man-made scenes also exhibit unique intra-class textural properties: the existence of skyscrapers largely dominates the scene of tall building, which makes it highly distinguishable from other scenes; similar to a highway scene, the presence of a road will exhibit a unique textural pattern and the presence of buildings can be detected by texture descriptors to characterize streets against highway scenes. These properties are illustrated in Figure 2.3.

However, most man-made scenes, especially those that depict indoor environments, do not follow strict textural patterns. More often than not, the semantics is defined by the objects scenes contain. As indoor scenes are usually functional with respect to the functionalities of the items within the scenes, proper recognition of scene objects seems heuristically important to indoor scene understanding (Figure 2.4). For example, a room is considered a bedroom if and only if the presence of a bed is detected, no matter how suggestive the recognition of closets or other items may be; a kitchen scene is usually characterized by the combination of a number of objects that might be found in a real-world kitchen, such as the sink, the refrigerator, the stove, the dining table and etc., all of which vary in shape and colour in different settings; living room and store scenes are even harder to define, since there is little agreement within each class in terms of functional objects. This is one of the reasons why shape descriptors are not widely employed for scene understanding.

Another reason for discarding shape descriptors is that prior to shape representation, image segmentation is explicitly required and should be applied to an excellent extent. But due to unresolved difficulties in current segmentation algorithms, such condition is hard to meet for a few reasons: first, most image segmentation approaches are involuntarily ad-hoc [94], so it is exceptionally difficult to find one algorithm that can uniformly produce satisfactory results in all categories of scenes; second, as mentioned earlier, the semantic information of natural scenes is often conveyed in terms of fore-

ground or background in the scene and simply describing and representing the shape is not nearly enough to extract its meaning; third, most scene images do not include depth information—obtainable mostly through depth acquisition or stereoscopic estimation—which is necessary for scene segmentation [121].

To sum up, the most viable candidate for scene image retrieval or classification seems to be texture analysis. Despite its relative deficiency in discriminative power when it comes to indoor scenes (Figure 2.4), it can still render state-of-the-art performance in terms of classification accuracy and retrieval rate, which will be shown in the following sections and chapters.

Figure 2.4. *Man-made indoor scenes that are characterized by the objects or items they contain.*

2.2. The GIST descriptor

Officially known as the *Spatial Envelope* (SE) representation, the GIST descriptor [100] is the first global scene feature that has yielded promising categorization result with a low dimensional feature vector.

Traditionally, scene pictures have been perceived as a complex combinatorial configuration of different objects or regions, and therefore, prior knowledge or recognition of objects plays an important role in subsequent scene understanding. In contrast, Oliva et al. propose to view a scene picture as an individual object, a much similar idea of perceiving objects and regions as an integral part of a scene without image segmentation. This level of abstraction is well founded on the research advancement in psychology, cognitive science and behavioural analysis that study the mechanism of human perception on scene images. The GIST descriptor is expected to model several fundamental properties of a scene, namely *naturalness, openness, roughness, expansion* and *ruggedness*, extracting dominant spatial structures that are essential to understanding the semantics of scene pictures and therefore provide significant ground work for accurate scene categorization.

In the original proposal, the GIST feature is a computational model that seeks to project scene images into a multidimensional spatial property space, in which scenes that belong to the same category are clustered closely together. Using spectral information and coarse spatial layout, this scene-centred approach computes the energy spectrum of scene images which share conspicuous similarities if they are drawn from the same category. Based on this theoretical foundation, Oliva et al. [99] propose to build the gist of a scene in a multi-scale and multi-orientation manner. This low-dimensional representation captures only essential spatial structures and coarse localization of a scene, offering a compact summarization of SE properties.

It should be noted that the term *Spatial Envelope* representation is used interchangeably with GIST. Both terms stand for the concept of transforming real-world scenes into a number of abstractions that describe the relationship between a composite set of surfaces or boundaries.

2.2.1. Theoretical justifications

Early computational vision theories on scene recognition pay special attention on local details such as contours and edges which are subsequently reconstructed and holistically combined to render classification decision [6], [87]. In this case, however, the level of recognition demanded is not simply understanding the functions of environments, but rather the extraction of meaningful information about the 3-D structure and useful attributes of the surfaces of a scene, namely shapes, locations and etc. In this school of studies, researchers propose to dissect images into a number of regions, objects or shapes, and the combination of which can be employed to form the high-level decision layer. Following this conception, several studies have focused on extracting the low-level or mid-level representations of these regions [16], objects [5] or shapes, using

simple features such as colour, orientation, texture, etc. for local description. An organized identification of these local features can then be further processed to render a final recognition decision.

The scope of this thesis, on the other hand, extends only to the identification of the semantic category of a given scene, which does not necessarily require local or object-level recognition, according to early computational vision research on scene recognition [10], [105]. On such level of recognition, a holistic perception of the scene can provide ample information for categorical inference since the rough structures and coarse spatial layout and localization carry the most semantic meanings with respect to its function. This school of computer vision research is generally in favour of a scene-centred representation of real-world environmental pictures and such notion has been well conditioned and justified through experimental studies in computational vision, psychology and cognitive science.

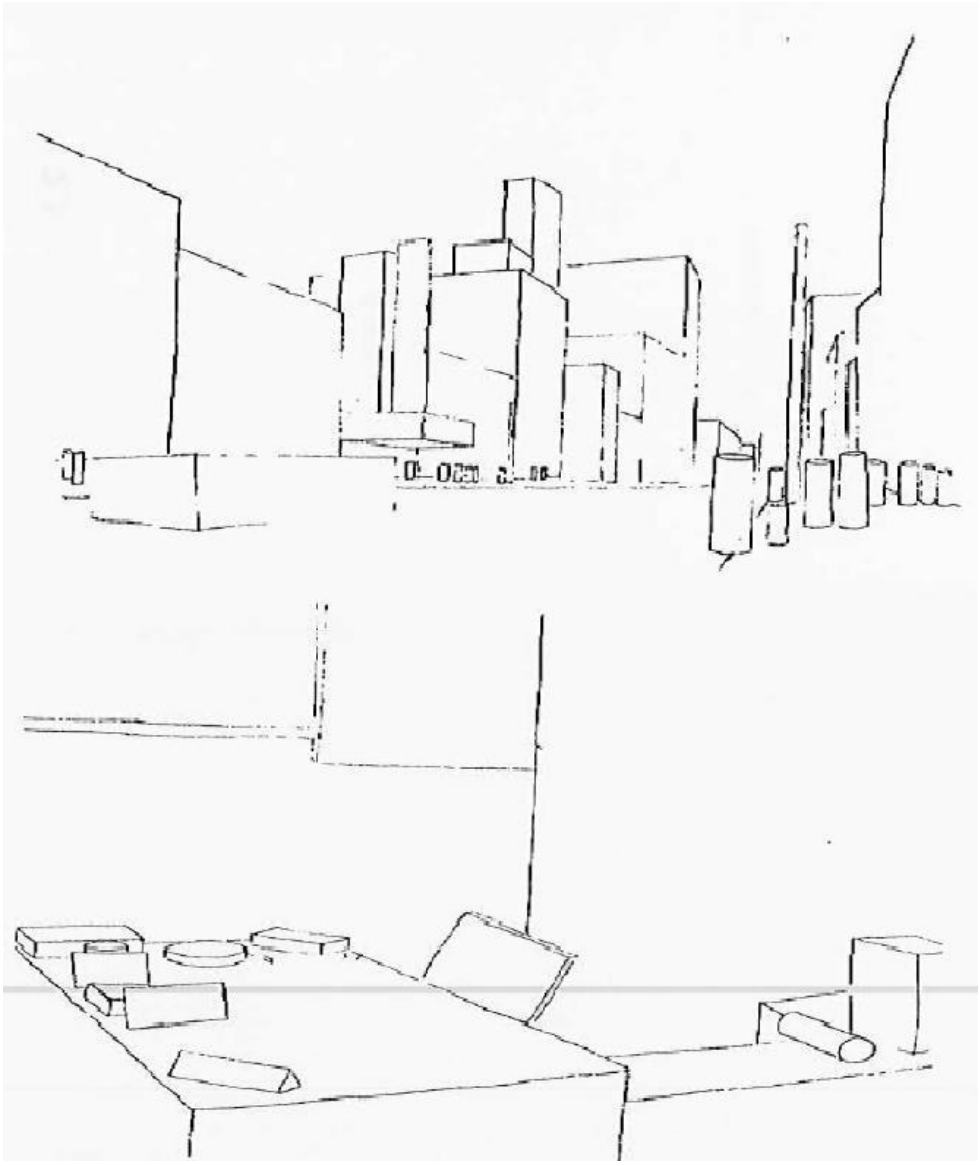


Figure 2.5. *Office and street scenes drawn by Robert Messanolle. None of the objects in these scenes is idenfiable, in isolation, as anything other than geon. [10]*

There has been research in the early age of image understanding to suggest that image processing through human visual system can grasp the semantics and the raw configuration with only a glance of the picture despite its complexity [105]. In such accelerated perception, visual memory of objects and their locations is relegated to non-essential information [57], [109]. Thorpe et al. [131] have found that human visual system can collect enough information for scene understanding in a mere 20 ms, which shows how competent human visual capacity is. Research on the effect of image manipulation further strengthens the claim that object identification only provides trivial knowledge on scene categorization for human perception. There is adequate evidence to suggest that even when images are purposefully blurred to the point where the identities of scene objects are beyond comprehension, people can still manage to understand the semantic category of presented scenes [98], [117]. In fact, it is reported that mere 4 to 8-cycles/image can be sufficient for labelling unknown scenes without dwelling on the given pictures. Biederman has made an interesting demonstration with geons—simple shapes that represent the most basic geometric forms. Figure 2.5 shows two scenes that depict a street and an office setting. The clear spatial layout or ordered arrangement of geons strongly indicates the semantic categories of these two scenes while the actual identities of the objects are unrecognizable for human perception. This example unequivocally shows the important role global structures play in accurate scene categorization.

Figure 2.6. *”Illustration of the effect of a coarse layout (at a resolution of 8 cycles/image) on scene identification and object recognition. Despite the lack of local details in the left blurred scene, viewers are confident in describing the spatial layout of a street. However, the high-resolution image reveals that the buildings are in fact furniture. This misinterpretation is not an error of the visual system. Instead, it illustrates the strength of the global spatial layout in constraining the identities of the local image structures.” [99]*

It has long been theorized that the human vision system prioritize reception and processing of global structures and spatial layout. Through a few experiments, Navon comes to the conclusion that global feature extraction is given precedence over local detailed measurements by human visual system, which supposedly serves a few purposes, such as efficient usage of processing resources to concentrate on low resolution information [95]. Navon's claim has since been vindicated by other researchers [69], leading to a generalized belief that within a glance, the spatial relationship between basic elements inside a scene is prioritized over the recognition of local details in scene categorization tasks. In particular, this is especially true when it comes to a busy scene presenting a significant amount of details, where memory use and processing time should be economized for human perception [70].

Oliva et al. [99] further illustrates this point with a cognitive experiment designed to show the mechanism of human perception on scene pictures within the time of a glance. In this experiment, human subjects are presented with two images, shown in Figure 2.6. During the first phase, viewers are asked to describe the first image which is intentionally blurred so that no local details are distinguishable for object identification, and render a semantic category of the scene. Unsurprisingly, the viewers unanimously describe the image as depicting a street scene with a high level of confidence and consistently venture to identify the blurred regions and objects as buildings, cars and the sky. When the details of the blurred regions are revealed, viewers are surprised to find that the perceived buildings in the first image are actually cabinets which are clearly transplanted from a kitchen scene. Nevertheless, the second picture as a whole unequivocally exhibit features of a street scene, with dominating regions of street, the sky and the cars perched on the street lending further clues to the perception. Despite the 30% of the intentionally manipulated image exhibiting features of an indoor scene, this experiment provides concrete evidence that when it comes to rapid scene recognition, human perception relies more on the global structures, the holistic arrangement of objects or regions and coarse spatial layout than specific details of local measurement. And this evidence prompts studies to concentrate on scene-centred approaches rather than object-centred theories.

There have been numerous studies following the concept of scene-centred approach that relies on the global configurations of a scene, in an attempt to minimize recourse consumption and processing time while mitigate the necessity for image segmentation. Depending on the complexity of the application, the proposed methods' levels of sophistication also vary. In [128], [143], the studies have been concentrating on differentiation between indoor and outdoor scenes or natural and man-made structures by applying simple low-level features extraction techniques. Similar to the GIST descriptor, Rogowitz et al. [110] study the similarity between human perception and computational vision using low-level features such as colour histograms and a multidimensional framework of colour, contrast and orientation-selective attributes to order images along different semantic axes. Due to the excellent performance achieved by the GIST features, Torralba et al. [133] propose to apply the same type of concept to depth estima-

tion using spectral signature. These studies suggest that contrary to traditional belief, global features provide highly discriminative information to mediate semantic categorization for scene images. Even though human perception does not completely ignore local details and object identities in a scene, it is the arrangement of objects and spatial layout that convey the essence of the scene in a rapid scene classification task. The correlation between global attributes and semantic category substantiates the concept of scene-centred descriptors.

2.2.2. The shape of a scene

Shape has been a powerful indicator of functional category of an object and thus has been greatly studied for object recognition. For example, different types of shape descriptors are widely incorporated into the MPEG-7 standard for object identification or retrieval. These descriptors include region shape, contour shape and shape 3D, attempting to capture the intrinsic characteristics of an object. In early scene matching research, the shape of a scene might refer to the orderless accumulation of different shapes of different objects within a scene. However, as elaborated in the previous section, in a scene-centred approach, the necessity of describing the detailed contours or regions of objects should be mitigated according to research on human perception. Instead, Oliva et al. [100] propose to perceive each individual scene as a single object and argue that this holistic perception is essential to extracting the "gist" of scene images. Similar to an object shape, the shape of a scene can convey its most discriminative attributes. In other words, the shape of a scene largely determines its semantic category and can be highly reliable for scene identification. As opposed to the shape of an object, however, the shape of a scene is a rather abstract concept. Even though it is a uniform perception on scenes, it carries several distinctive properties of scene pictures, which is difficult to visualize.

Figure 2.7 shows scene images from eight different categories and their corresponding surface appearance. Scene pictures on the first row may justify the necessity of object or region recognition before rending any decision on their categorical labels. Nevertheless, a rough showing of their surface appearance, obtained by transforming the intensity of each image pixel into height of the surface graph after low-passing the original scene pictures for noise attenuation, indicates a rather abstract property of the scenes—roughness. One advantage of using this level of abstraction to redefine scene images is that scenes that are from the same category share the same attribute in abstract terms. And the surface appearance or roughness of a scene can be perceived as one aspect of the shape of a scene.

It should be noted that the GIST feature is not a hierarchical processing algorithm of scene images with each stage corresponding to one abstract property of the scene. Rather, the feature extraction process is a simultaneous procedure—all perceptual properties of a scene are extracted at once, which ensures the computational efficiency of the operation. After feature extraction, these abstract properties of a scene are engrained in

the GIST feature to form a holistic description of the scene which is called the SE representation.

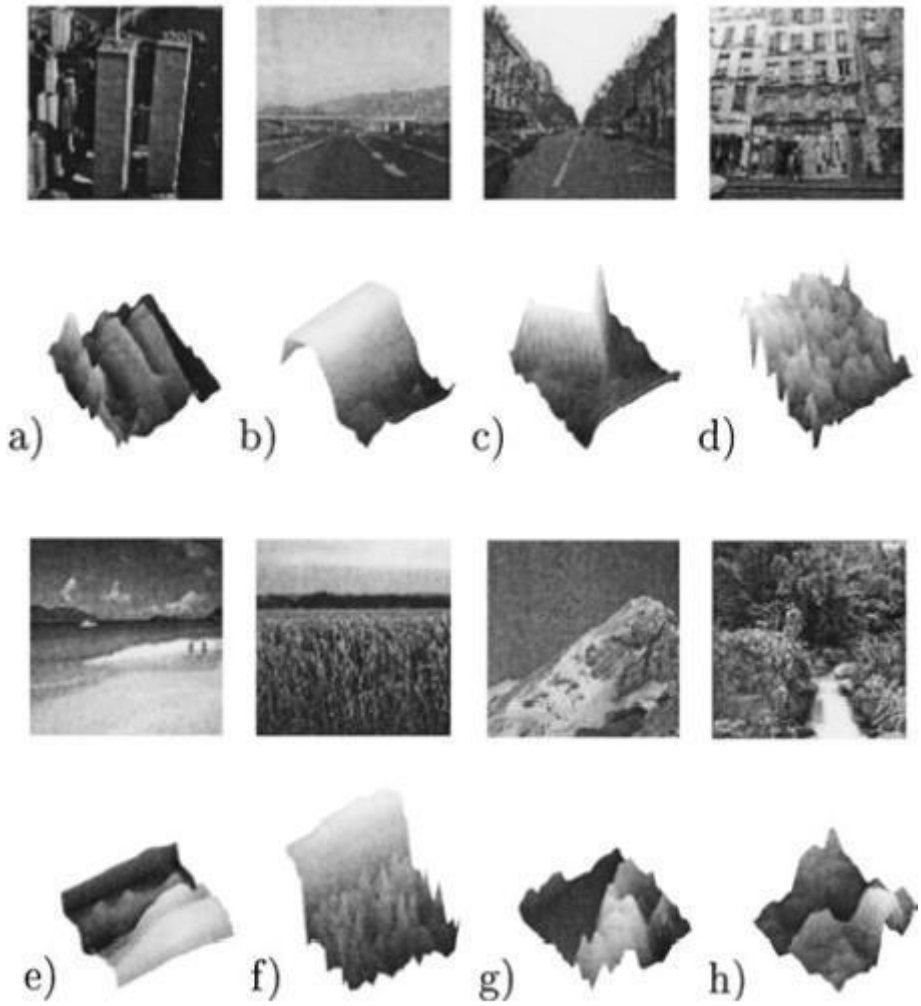


Figure 2.7. *"Scenes with different spatial envelopes and their surface representation, where the height level corresponds to the intensity at each pixel (images were low-passed): a) skyscrapers, b) an highway, c) a perspective street, d) view on a flat building, e) a beach, f) a field, g) a mountain and e) a forest. The surface shows the information really available after projection of the 3D scene onto the camera. Several aspects of the 3D scene have a direct transposition onto the 2D properties (e.f., roughness)." [100]*

2.2.3. Spatial envelope

The term *Spatial Envelope* (SE) is used extensively in architecture or interior design. The term may refer to the exterior of a building when it is mentioned in the context of architectural design. It also signifies the internal boundaries within an indoor environment, which is the case of interior decoration.

Oliva et al. [100] adopt the term in an attempt to capture the relationship between different boundaries within a scene and the intrinsic properties of these boundaries or surfaces. For example, in an outdoor scene such as coast, the SE is described as clear-textured sky connecting to the water surface of the ocean where occasionally certain concaves are observed (boat or cruise ship); a highway scene is perceived somewhat similar to a coast scene with the sky directly bordering the hard surface of pavement which is characterized with near vertical lines; in a man-made scene such as tall building, extending vertical lines are connected to a common ground with small textured regions that represent windows and other exterior designs.

2.2.4. Spatial categories

As discussed in the first chapter, there are numerous ways of scene categorization. And in this thesis, we adopt the semantics of each scene as its label, such as coast, mountain, street and etc., a much similar configuration to the one used by Tversky et al. in [141]. This may seem like a good strategy for managing scene images, but such way of categorization does not highlight different structural compositions in scenes and therefore fails to translate into the properties of SE representation. To this end, there is a need to explore spatial categories—categories of SE properties—so that each semantic category of a scene can be defined by a combination of SE properties.

There have been numerous studies that attempt to determine the most discriminative spatial properties for texture analysis [2], [56], [108], [129]. Notably, Rao et al. [108] have identified a few perceptual properties that are the most powerful in texture differentiation, namely repetitiveness, contrast and granularity. However, these studies were mainly concentrated on texture images or textural regions within an image. The exploration of translating the perceptual properties of texture images into scene images was not initiated until the introduction of GIST features. Nevertheless, some scholars have treaded the territory of using abstract concept to describe real-world scenes and uncovering the relationship between scene category and types of connectivity among blobs [98], [117].

In an attempt to discover the useful dimensions of spatial categories that characterize the SE properties of real-world scenes, the initiators of the original GIST features devised an experiment in which 17 human subjects were allocated the task of categorizing 81 pictures according to a set of rules. The most preceding guideline was, the object identities or other details of local measurement should be ignored during the categorization process, and so should be the holistic, underlying meaning of the scenes, such as field, mountain, street etc. The global criterion regarding image splitting was as simple as spatial structure. The experiment proceeded in three stages. These three stages entailed hierarchical dividing images into a few subgroups. In each stage, each group was split into two subdivisions, resulting in eight subgroups in total after all three stages were completed.

All participants were required to specify the criteria they used for splitting the images after each stage of the experiment in their own words. And these words were trans-

lated into perceptual properties, namely *naturalness, openness, perspective, size, diagonal plane, depth, symmetry* and *contrast*, which are summarized in Table 2.1.

Table 2.1. *Perceptual properties of the human vision system generated from cognitive experiment. The results are percentage numbers. Each colomn represents the tally of one stage of the experiment and the counts are independent for each stage. The total number of times a certain criterion used is listed in the last column. [100]*

Property	S1	S2	S3	Total
Naturalness	65	12	0	77
Openness	6	53	24	83
Perspective	6	18	29	53
Size	0	0	47	47
Diagonal	0	12	29	41
Depth	18	12	29	59
Symmetry	0	0	29	29
Contrast	0	0	18	18

As shown in Table 2.1, *naturalness* and *openness* are the first two dimensions of perceptual criteria that were used to split real-world scene images. Interestingly, about 65% of human subjects regard *naturalness* to be the priority consideration for spatial categorization (in the first stage), which means that the human visual system gives precedence to differentiation of man-made structures and natural landscape scenes. *Openness*—the most used criterion used in the second stage—signifies the extent of confinement of a scene. That is, whether a scene is dominated by an open area or it is enclosed by walls or natural materials. During the third stage, the viewers seemed to pay special attention to the *size* of the scene. It should be noted that the *size* in this context does not refer to the dimensions of the image. Instead, it correlates with the size of objects, items or regions within the image. In a sense, the term *size* used in this experiment refers mostly to the scale of an image. The criterion *diagonal plane* stands for the rough contours that expand diagonally upwards or downwards, which is the most notable characteristics in the scene dominated with mountain ranges or rocks. The property *depth* employed by human subjects, however, lacks consistency of meaning. Some observers attribute the depth of a scene as distance of the scene with respect to the focal point; some consider the distance of objects within a scene as the primary indicator of scene depth; other viewers correlate the depth of a scene with its degree of expansion. These inconsistencies are well explained by Oliva et al. in [100]. But a noteworthy finding is that the most important criteria for object identification are not frequently selected for scene categorization. Specifically, the properties of symmetry and contrast only account for 29% and 18% of the times respectively. Such discovery corroborates the findings by Sanocki et al. [115]

2.2.5. Spatial envelope properties

According to the experiment presented above, Oliva et al. summarize the perceptual attributes viewers used in five SE properties—*naturalness, openness, roughness, expansion* and *ruggedness*, which are defined as following:

- *Degree of Naturalness*: The *degree of naturalness* is an important indicator that separates man-made structures and natural habitats. It is observed that most man-made scenes are characterized primarily by horizontal and vertical lines throughout the image. In contrast, the contours of natural landscapes are more flexible and stretch along continuously changing directions. So in the case of a scene that exhibits both natural environment and man-made structures, the *degree of naturalness* provides a measurement of the dominance of either characteristic.
- *Degree of Openness*: This property mainly refers to the extent of how enclosed a scene is. In some scene categories, the open areas dominate the texture of all images, such as coast scenes, open country scenes; whereas other scenes are characterized by confining structures that enclose most parts of the images. For example, in a bedroom scene, the primary objects—bed and closets—are enclosed by the bedroom walls; and in a forest scene, the trees are usually clustered together and propagate throughout the scene, edging the open area—the sky—to a small portion in the image.
- *Degree of Roughness*: In a sense, the *degree of roughness* represents what granularity refers to in texture analysis. It depends on the configuration of the most basic elements in a scene, especially the complexity of their configuration. In addition, it also accounts for the relations between these basic elements and their ability to form more complex structures. It should be pointed out that the *degree of roughness* is highly correlated to the spatial scale in which the elements are measured. If a scene presents a *high degree of roughness*, it means that the general configuration of the scene is rather complex. And the finer the spatial scale is, the more structural details can be uncovered.
- *Degree of Expansion*: This particular SE property refers to how extensive the lines in a scene appear to be. For example, in a highway scene, the vertical lines seem to extend further and further until they are out of sight. In a two dimensional image, this phenomenon is observed in terms of degree of convergence of horizontal or vertical lines. If all the major lines in a scene have a tendency to converge to a certain point, the viewer will acknowledge that a significant distance is covered inside the image. On the other hand, if most lines of a certain scene propagate throughout the scene in parallel, the scene is said to have a *low degree of expansion*.
- *Degree of Ruggedness*: The *degree of ruggedness* stands for how the horizon of a scene emerges to the viewer. If a scene is said to present a *high degree of rug-*

gedness, its horizontal line should be either ambiguous or nonexistent at all. To the contrary, scenes with a *low degree of ruggedness* usually exhibit a clean-cut horizon. For example, in a mountain scene, the mountain base and the horizon are so inseparable that the horizontal line is nearly invisible; whereas in a street scene, building are usually perfectly perched on the horizon, which suggests street scenes exhibit a *low degree of ruggedness*.

These five SE properties are used together to model the abstract and perceptual appearances of real-world scenes. In the following section, it is shown that scene images that share the same attributes tend to present similar structures and belong in the same semantic category

2.2.6. Computational model

In order to represent scene pictures without taking into account local details, images can be transformed into frequency domain using discrete Fourier transform (DFT):

$$\begin{aligned} I(u,v) &= \sum_{x,y=0}^{N-1} i(x,y)h(x,y)e^{-j2\pi(ux+vy)} \\ &= A(u,v)e^{j\Phi(u,v)} \end{aligned} \tag{1}$$

where $i(x,y)$ represents the intensity of the scene image at each pixel location, u,v denote the frequency variables in the Fourier domain, $h(x,y)$ is the Hanning window that handles boundary problems and $I(u,v)$ is the DFT coefficients. The DFT of an image can be perceived in another manner: in the frequency domain, the transform $I(u,v)$ can be decomposed into the magnitude of DFT coefficients $A(u,v)$ (also called the amplitude spectrum) and its phase function $\Phi(u,v)$.

Since the amplitude spectrum of an image does not reveal any spatial information, it provides general analysis on its global structures without specifying the identities of local objects. Such analysis includes the direction, roughness and length and width of contours in a scene. In addition, the energy spectrum of an image—the squared magnitudes of DFT coefficients—describes the energy distribution with respect to the spatial frequency variables. In a similar fashion, the energy spectrums of real-world scenes carry the structural imprints of the whole image. Several studies have found that the global structural information encoded in the energy spectrum is useful in image classification [48], [50], [101], [128], [142], [143]. The phase function of Fourier transform, on the other hand, concerns image properties at a local level with certain information on object positions.

Even though the DFT of an image offers excellent perspective on its global configuration, this coarse representation does not offer any insight on the interplay of different structures within a scene which, as evidenced by previous discussions, correlates with the semantic category of scene images. Several studies have shown that relationship between distinctive structures within an image can assist image retrieval and classifica-

tion in general [17], [18], [31], [76], [134]. To this end, the spectral layout of an image can be modelled using windowed Fourier transform (WFT) in the following form:

$$\begin{aligned} I(x,y,u,v) &= \sum_{x',y'=0}^{N-1} i(x',y')h_r(x'-x,y'-y)e^{-j2\pi(ux'+vy')} \\ &= A(x,y,u,v)e^{j\Phi(x,y,u,v)} \end{aligned} \tag{2}$$

where $h_r(x,y)$ denotes a Hamming window whose circular support is r. Similarly, the structural configuration and their localized information are encapsulated in the energy spectrum of WFT, also known as a spectrogram. In this form of representation, the localized information can be coarse or detailed, depending on the size of the Hamming window. For the purpose of estimating spatial envelope properties of large structures in scene images, Oliva et al. propose to compute WFT around a 32-pixel radius neighbourhood, resulting in 8 × 8 spatial locations.

In order to reduce the dimensions of the energy spectrum or the spectrogram, a good measure of feature selection algorithm should be selected to ease the overall computation. In [100], Principal Component Analysis (PCA) is used to project the image features into orthogonal bases so that the energy spectrum or spectrogram can be decorrelated. In this new space, the orthogonal functions that account for the most variance are kept as the principal components of the image features. In this dimensionality reduction algorithm, the Karhunen-Loeve Transform (KLT) is used for orthogonal projection. Equation (3) and (4) show the KLT decomposition of the energy spectrum $A(u,v)^2$ and spectrogram $A(x,y,u,v)^2$ respectively.

$$A(u,v)^2 \simeq \sum_{i=1}^{N_G} v_i \psi_i(u,v) \tag{3}$$

$$A(x,y,u,v)^2 \simeq \sum_{i=1}^{N_L} w_i \Psi_i(x,y,u,v) \tag{4}$$

where $\psi_i(u,v)$ are the orthogonal functions of the energy spectrum and $\Psi_i(x,y,u,v)$ are the basis functions of the spectrogram. N_G and N_L denote the number of KL functions used for the energy spectrum and the spectrogram respectively. They also represent the final dimensions of reduced image features. v_i and w_i represent the decorrelated coefficients of the KL functions for the energy spectrum and the spectrogram respectively. They can be obtained from the following equations:

$$\begin{aligned} v_i &= \langle A^2, \psi_i \rangle \\ &= \iint A(u,v)^2 \psi_i(u,v) du dv \end{aligned} \tag{5}$$

$$w_i = \sum_x \sum_y \iint A(x,y,u,v)^2 \times \Psi_i(x,y,u,v) du dv \tag{6}$$

v_i and w_i are the SE representations of scene images and can be considered as the final image features after dimensionality reduction. v_i represents the global structures of a scene that is similar to the concept of scene shape, since there is no spatial information

on any detail, object or region. This level of scene features estimates the SE properties of the whole scene without any evaluation of localized information. Figure 2.8 shows the first eight principal components of real-world scenes. In contrast, w_i stands for the structural details of different regions of a scene. This spectrogram can be perceived as the scene layout of energy spectra for different parts of the scene and therefore captures the relationship between large neighbouring structures. In other words, the spectrogram coefficients w_i give a general description of perceptual properties for different regions. An illustration of spectrogram is shown in Figure 2.9.

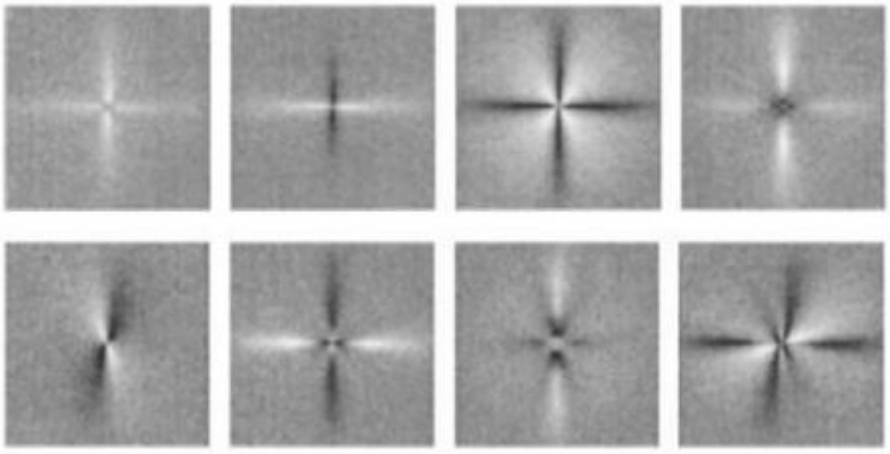


Figure 2.8. *"The first eight principle components for energy spectra of real-world scenes." The zero frequencies for u and v are shifted to the centre of the image. [100]*

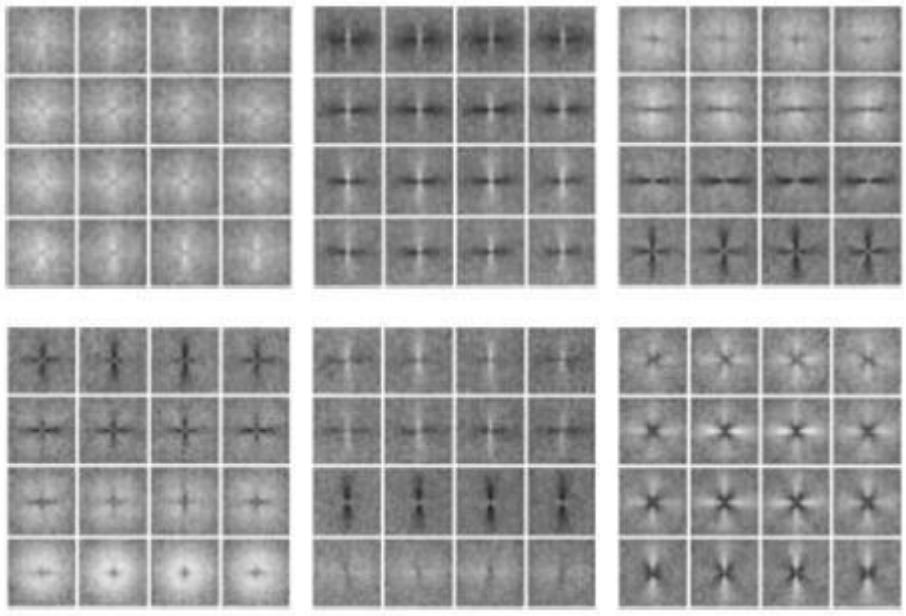


Figure 2.9. *"The first six principal components for the spectrogram of real-world scenes. The spectrogram is sampled at 4×4 spatial location for a better visualization. Each subimage corresponds to the local energy spectrum at the corresponding spatial location." [100]*

The use of energy spectrum or spectrogram in natural image analysis has been experimented in several studies [43], [44], [145]. And the transformation from spatial domain to frequency domain has been compared to the human cognitive system [3], [43]. In addition, the relationship between the energy spectrum and the semantic label of a scene has also been studied and it is proven that certain energy spectra have a strong implication regarding the appearance of a scene [4], [101], [127].

To capture the distinctive characteristics of energy spectra for real-world scenes, Oliva et al. propose to use their mean. The averaged energy spectrum can be perceived as the spectral signature for all the scenes from the same class. The scene categories include *tall buildings*, *highways*, *city centers*, *city close-up views*, *coasts*, *mountains* and *forests*, representing scenes from both man-made and natural environments with hundreds of images from each category. Through the averaging operation, it is shown that most scenes from the same category share the same spectral structures.

Similar to the studies by van der Schaaf et al. [145], Oliva et al. try to extract the spectral signature of all the images from the same class by using the following approximation function:

$$E[A(f,\theta)^2 \mid S] \simeq \Gamma_s(\theta) / f^{-\alpha_s(\theta)} \tag{7}$$

where $E[A(f,\theta)^2 \mid S]$ is the expected value of the power spectrum $A(f,\theta)^2$ given the semantic category S of all the images. It should be noted that equation (7) is expressed in a polar coordination system. For each orientation θ, the averaged energy spectrum is linearly fitted on logarithmic units to obtain the functions $\Gamma_s(\theta)$ and $\alpha_s(\theta)$ for each category S. An example of linear fitting is shown in Figure 2.10. In the figure, the averaged spectrum is linearly fitted at three orientations for three scene categories, namely *coastlines*, *buildings* and *forests*.

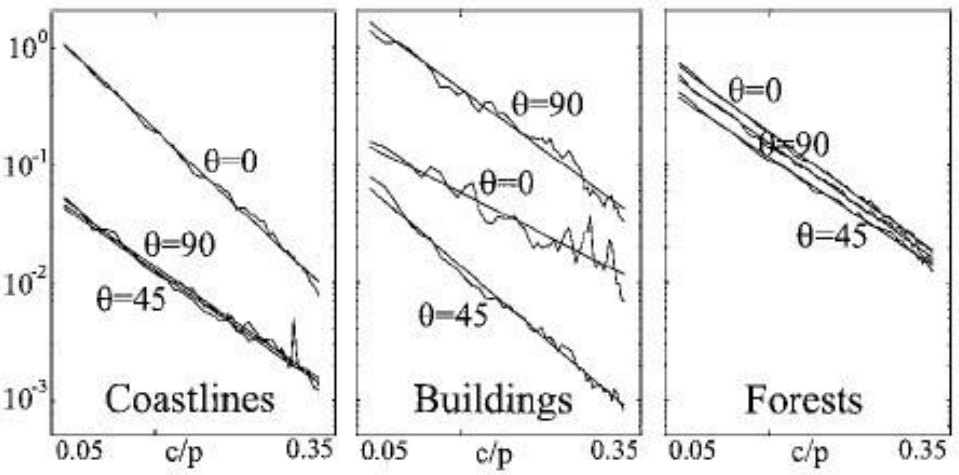


Figure 2.10. *Linear fitting of the averaged power spectrum at three orientation for three scene categories. [100]*

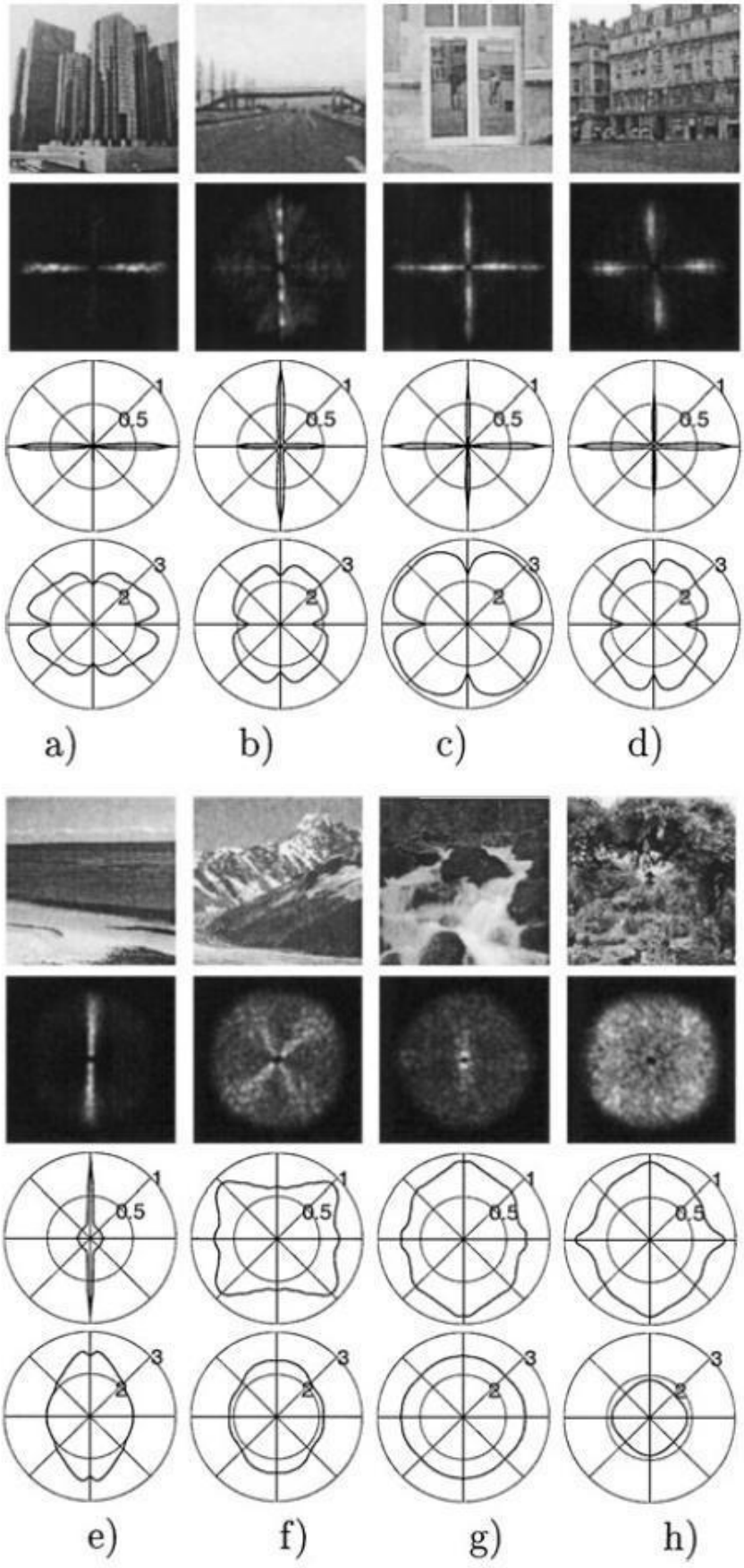


Figure 2.11. *Scene images and their spectral signatures. [100]*

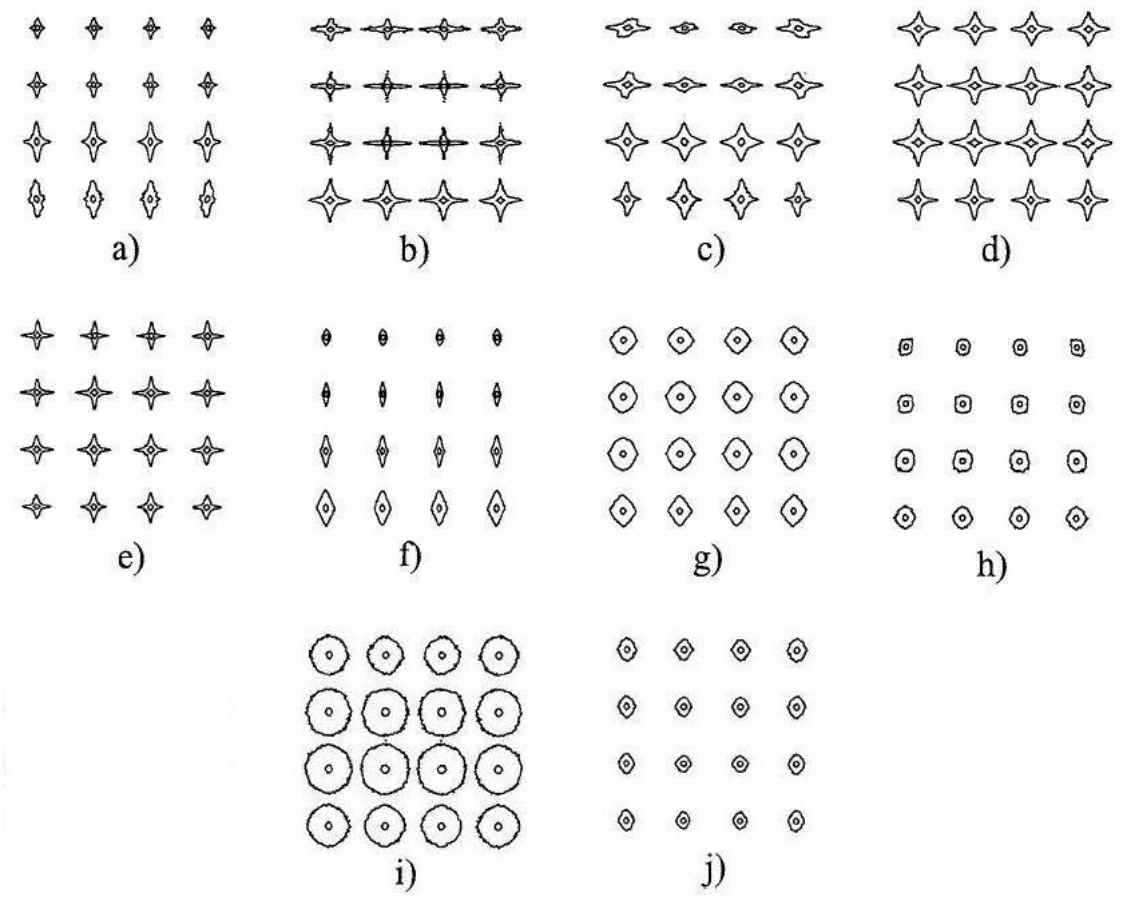


Figure 2.12. *Example spectrogram signatures from ten different scene types. [100]*

Figure 2.11 shows the relationship between scene images and their spectral signature. The first line shows a few sample images from eight scene categories: *tall building*, *highway*, *inside city*, *street*, *coast*, *mountain*, *open country* and *forest*. The power spectra for these eight images are illustrated on the second row. With the energy spectra of hundreds of images from each category, the spectral signatures for all categories are linearly fitted according to equation(7). In particular, function $\Gamma_s(\theta)$ is shown on the third row and it summarizes the major orientations exhibited in a scene category. On the last row, the function $\alpha_s(\theta)$ represents the complexity of the scene category.

Similarly, the average spectrograms for ten different types of scenes are shown in Figure 2.12. The first five represent spectrograms from man-made scenes whereas the other half illustrates the signature of spectrograms for natural environment. In the man-made scenes, a) is the general signature of an open scene; b) represents scenes with dominant vertical lines; c) depicts the average spectrogram of *street* scenes; d) shows the characteristics of *tall buildings*; e) provides the signature for *inside city* scenes. In the natural scenes, f) and g) illustrates the average spectrogram for open and enclosed scenes respectively; h) shows the localized spectral signature of *mountain* scenes; i) depicts the general characteristics of *forest* scenes and j) provides some insight on how the average spectrogram of *stream* scenes looks like.

It is quite conspicuous that the SE properties are encapsulated in the two spectral features of scene images. But these spectral representations do not enumerate each individual perceptual characteristic, namely *naturalness*, *openness* and etc. In addition, it is impossible for human subjects to extract high-level semantics from these spectral features, although they do carry abstract information regarding the global structures of a scene and their relationships. In order to generate a high-level description of scene images, Oliva et al. [100] propose to estimate the five SE attributes from the spectral features using linear regression as a way of bridging the gap between the low-level feature space and the high-level low-dimensional semantic space. It should be noted, however, that the estimation of SE properties is only necessary when these attributes are explicitly demanded by a human user, e.g., when the user specifically wants to order images in the dataset according to the degree of *naturalness*. In real-world applications, the spectral signatures of scene images are sufficient as scene features for scene retrieval or classification.

2.2.7. Building the gist of a scene

As mentioned in previous sections, a spectrogram carries localized information of structural properties. Such information contains all five SE properties of a scene. At each spatial location, the spectral features represent the structural characteristics of the scene patch and the entire spectrogram models the global distribution of different perceptual attributes. Specifically, each spectrogram of a scene encapsulates both orientation patterns at different spatial locations and energy values with respect to different spatial frequencies. In other words, the descriptive capacity of spectrograms depends predominantly on the information of spatial scale and orientation from real-world scenes.

In light of such reasoning, there exists other ways to extract SE properties for scene images. Oliva et al. propose in [99] to estimate the SE properties by means of a multiscale-orientation analysis of scene images. By passing a scene through a collective set of multiscale-oriented filters, the energy value is extracted at each pixel location in each spatial scale at each orientation. Such feature extraction approach decorrelates perceptual attributes into multiple filtered images, making it easier to measure the textual difference between scenes. The resulting feature is called the gist of a scene, (hence the name of the feature,) and can be implemented in the following steps:

① The first step concerns image preprocessing with the goal of attenuating illumination variation and heightening strong scene structures. At high spatial frequency levels, the finer details of small objects are easily observed, but the presence of noise can be inconvenient to structural analysis and should be effectively filtered out [88], [119]. However, at low spatial frequencies, the contrast of the image is more observable [95], [125], although low resolution sometimes induces ineffective structural extraction. In light of such dilemma, Oliva et al. propose to extract details and heighten significant scene structures by subtracting a low-passed blurred image from the original scene, in which the Gaussian filter G

is defined in equation (9). The variance of G in the frequency domain is determined by the frequency cycle c which usually takes the value from 1-8 (equation (8), where ln represents the natural logarithm operation). The filtered image f is then normalized pixel-wise against local contrast with equation (11). Figure 2.13 shows the preprocessing effects on an original scene image.

$$\sigma = c / \ln(2) \tag{8}$$

$$G = e^{-\frac{u^2+v^2}{\sigma^2}} \tag{9}$$

$$f = I - I \otimes g \tag{10}$$

$$output = f / (0.2 + | f \otimes g |^2) \tag{11}$$

original image whitened image contrast-normalized image

Figure 2.13. *A sample coast image and the outputs from the preprocessing stage. The whitened image is only an intermediate product. The image on the right is the final output.*

② The second stage involves a cascade of filters that analyze the spatial frequencies and orientation energy values of a scene. Per the discussion from previous sections, the objective of scene attribute analysis is to extract its gist (SE properties). The computational model for such analysis relies largely on spectrogram (WFT) of a scene image. In the field of image processing, WFT is often presented or implemented as a jet of Gabor filters, the promise of which has been compared to simple receptive fields in cat striate cortex [59]. The Gabor filters perform texture analysis on images in a multiscale-orientation manner. The most important parameters of the filters are the number of scales (also known as spatial frequencies N_s) and the number of orientations for each scale (N_o), resulting in $N_s \times N_o$ filter banks in total. It should be noted that N_o can vary from scale to scale, but it is more straight-forward to fix the value of N_o for all scales. At the second stage of the GIST feature extraction, the Gabor filters in the frequency domain are set according to the following equation:

$$G(\rho,\theta) = e^{-10\cdot 0.35(\frac{\rho}{N\cdot(0.3/1.85^{s-1})}-1)^2 - 2\pi(16\cdot {N_o}^2/32^2)(\theta+\frac{\pi(o-1)}{N_o})^2} \tag{12}$$

where ρ and θ denote the radius and angle in the polar coordination system, N represents the width or height of image resolution, assuming the aspect ratio is 1:1, s and o stand for the scale and orientation respectively. In this transfer function, the angle θ is scaled to take value between $-\pi$ and π. A direct visualization of the Gabor filters when computed in 4 scales ($N_s = 4$) with 8 orientations ($N_o = 8$) for each scale is shown in Figure 2.14. The Gabor filter outputs, resulting from the product between the image in the frequency domain and the transfer function shown in equation (12), are $N_s \times N_o$ images. And since these images carry information on orientation energy levels at different spatial frequencies, they are called orientation maps which are shown in Figure 2.15.

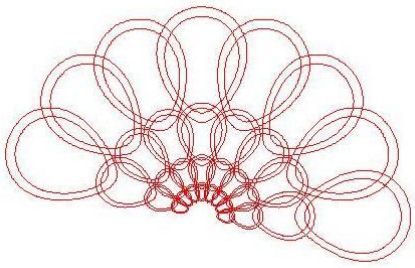


Figure 2.14. *A jet of Gabor filters at 4 scales with 8 orientations for each scale (N_s=4, N_o=8).*

③ The resulting image features cannot be directly used as the GIST features. The reasons are twofold: a) the orientation maps can usually tabulate up to millions of dimensions, depending on the resolution of the original image. Such dimensionality exceeds the maximum capacity of most machine learning algorithms for recognition and demands a substantial amount of system resources. In retrieval applications, high dimensional features severely slow down the retrieving process and thereby cause performance degradation; b) although the orientation maps capture scene features precisely, the structural characteristics do not match pixel to pixel, even when orientation maps are from scenes sharing semantic membership. The reality is, scenes from the same class often present similar structural properties and these properties can be present in different parts of the scenes. Due to these reasons, the original orientation maps are often down-sampled into tiny thumbnail images to accommodate spatial variations of scene structures. The simplest way of down-sampling is to divide each orientation map into a few rectangular blocks and use the average energy in each block to represent its feature, as shown in Figure 2.16. Thus, the dimensions of orientation maps can be reduced to a few hundreds, providing efficiency and flexibility.

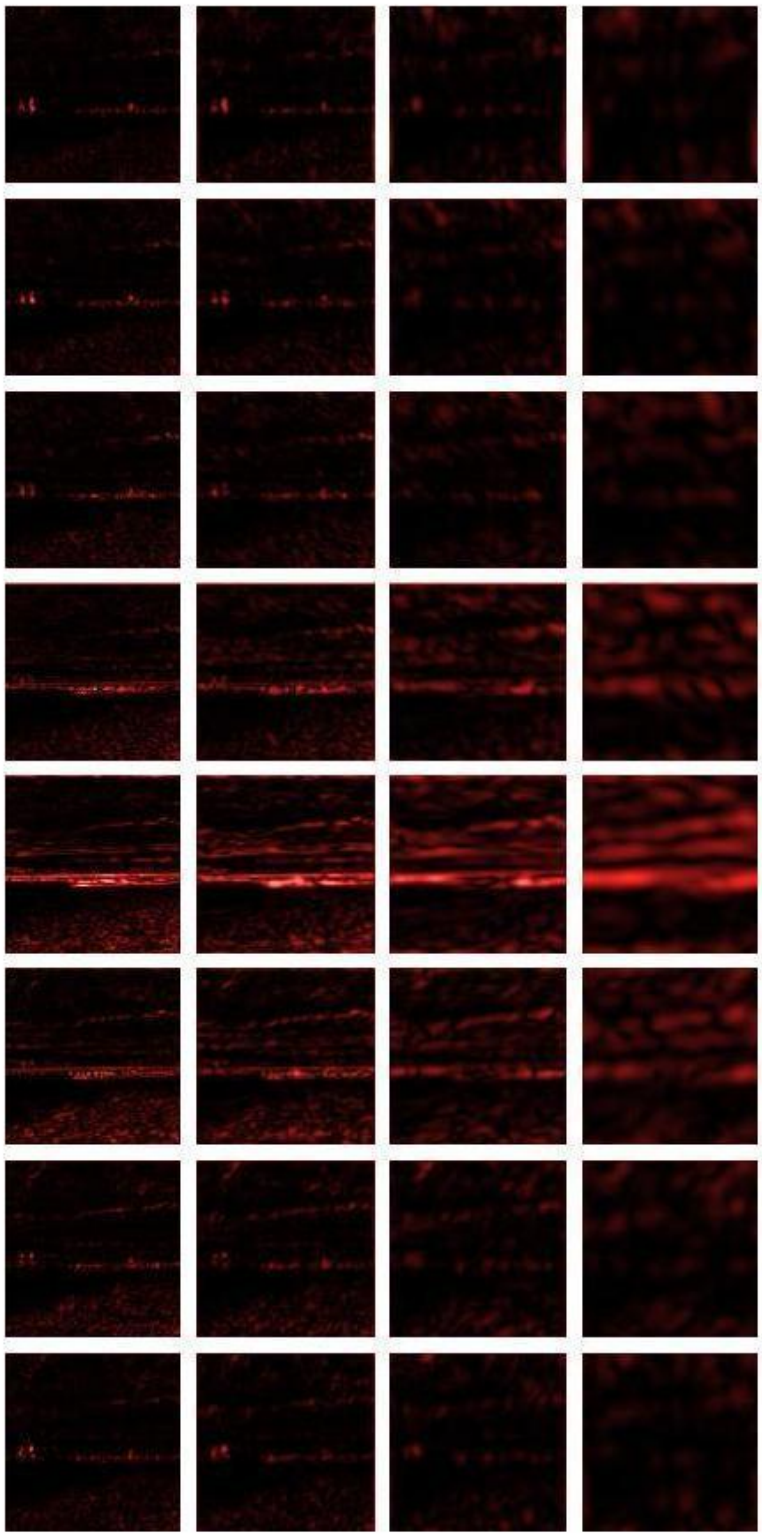


Figure 2.15. *Orientaion maps of a coast scene. The horizontal axis refers to (4) scales and the vertical one denotes (8) orientation.*

It should be noted that the sample orientation maps shown in Figure 2.15 are manipulated to highlight the filter response. Although higher energy regions are marked in red, the Gabor filter outputs are not 3-D arrays.

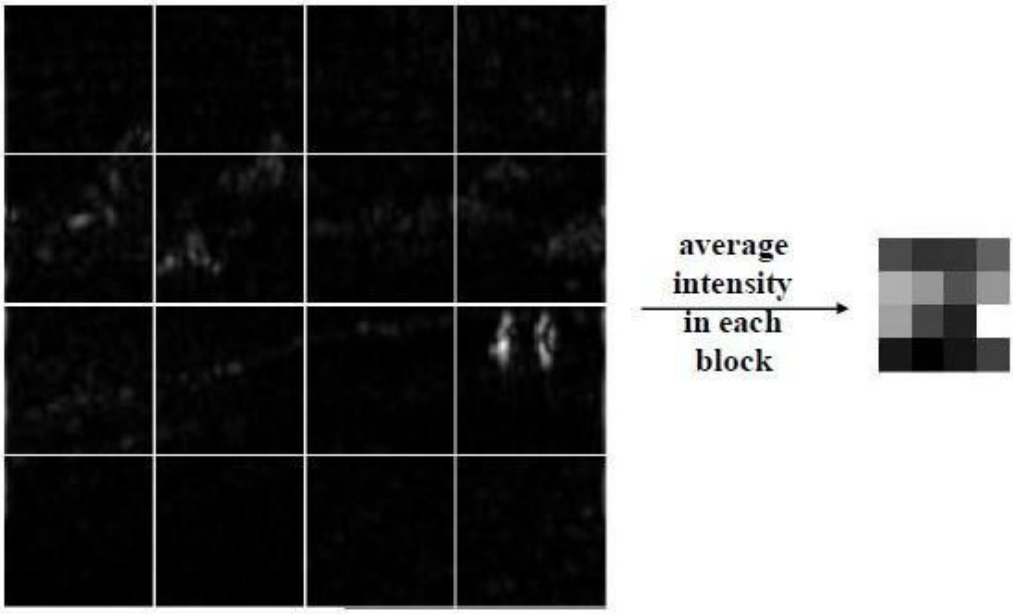


Figure 2.16. *Downsample a sample orientation map.*

Figure 2.16 shows an instance of down-sampling process, the original orientation map is divide into a 4×4 grid, and then the average intensity in each block is used as the final feature. These down-sampled orientation maps are shown in Figure 2.17.

The orientation thumbnails are the GIST features of a scene. They are concatenated to form the final feature vector. These feature vectors are used as a form of scene representation for scene retrieval or classification.

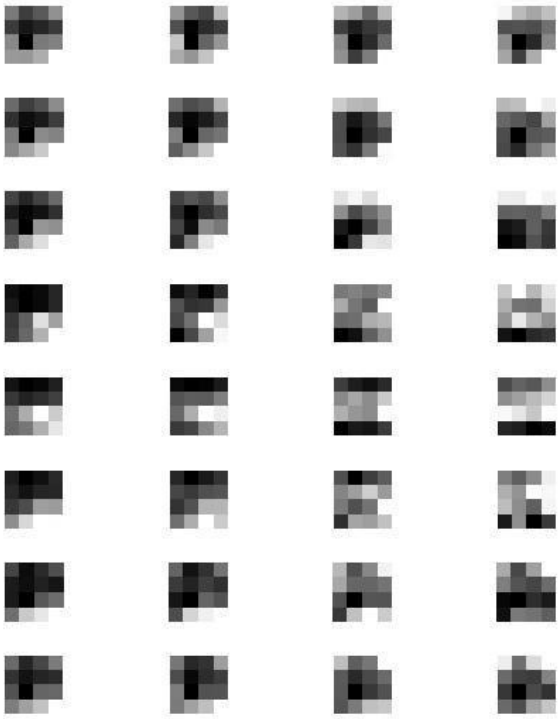


Figure 2.17. *Down-sampled orientation maps of a coast scene. These images reprent the GIST features of a scene.*

2.3. Local Binary Pattern

Texture analysis as a standalone research subject has been given considerable attention in the fields of computer vision and image processing as it is an important part of many tasks involving classification, detection or image segmentation. Notably, it is highly useful in detecting defects in industrial production as well as medical image analysis.

A major theme in this area of research is finding a suitable measure to account for illumination, scale and rotation variations. These variations stem from the unpredictability of camera settings, such as focal length, position of the camera, time of exposure and etc. In order to ensure even or uniform representation of textures from the same class under different imaging conditions, powerful texture descriptors should be robust to illumination, scale and rotation variations. Additionally, computational complexity is also a major concern for texture analysis as it has an impact on resource consumption and performance. It is shown that numerous texture extraction approaches have failed to meet that requirement [107].

There have been studies exploring and seeking the possibility of accommodating all three variations [1], [24], [149]. But a more popular theme is to deal with one at a time. In handling rotation invariance, studies have been focusing either on the development of new, rotation invariant texture features or on the modification of existing texture descriptors to improve robustness to rotation variations. In the former case, specially devised features include generalized cooccurrence matrices [30], polarograms [29] and texture anisotropy [23]. In the case of feature modification, robustness to rotation invariance is usually facilitated by converting well-known texture features [45], [49], [52], [72], [74], [84], [85], [104], [150], such as the Gabor filters.

Similar to Local Binary Pattern (LBP), several studies have proposed to realize texture matching with both illumination and rotation invariant features [22], [148]. However, in these studies, only one type of illumination shift is considered, which is modelled as linear transformation. This assumption has cast serious doubt on the validity and applicability of the proposed method. Particularly, the realization of illumination invariance is achieved by normalizing images with global histogram equalization, which does not account for local variations.

The concept of LBP was originally proposed in [96]. It is designed such that most monotonic transformation of intensity shift can be reasonably filtered out. The computation of the original LBP is surprisingly simple, and yet the descriptor is discriminative, operating primarily on a local neighbourhood of every image pixel. Pietikäinen et al. further extended the original LBP in [103] to take into consideration local rotation variations. In a rather simplistic transformation, LBP features are shifted such that similar local textural characteristics can yield similar features regardless of the presence of rotational variations. Furthermore, Ojala et al. [97] have identified several unique patterns called “uniform” patterns that account for over 90% of all local patterns. These uniform patterns are said to present the most significant discriminative powers over other patterns.

The most recent state-of-the-art scene classification study is mostly based on the promise of LBP. Introduced by Wu et al., the so called CENsus TRansform hISTogram (CENTRIST) [147] is dubbed as a new scene descriptor that fundamentally encodes the dominant scene structures with LBP. For the purpose of dimensionality reduction, CENTRIST incorporates PCA in its basic form which has not only reduced sharply the dimensions of image features, but also significantly improved accuracy for scene classification. In contrast to LBP, CENTRIST features are computed on multiple spatial levels of a scene in a hierarchical manner [73], which has greatly contributed to its excellent performance.

2.3.1. Circularly symmetric neighbourhood

The LBP local neighbourhood can be generalized in the following manner: suppose we have a texture image that is achromatic by default or converted to gray-scale from a colour space. In a local neighbourhood of the image, Texture *T* is defined as a spatial distribution of intensity levels from the centre pixel and the *P* neighbouring pixels:

$$T = t(g_c, g_0, ..., g_{P-1}) \tag{13}$$

where g_c denotes the intensity level of the centre pixel, and $g_p\,(p = 0,...,P-1)$ represents the gray-scale value of one of the *P* neighbouring pixels that are equally spaced around the centre pixel g_c and form a circularly symmetric neighbourhood around it with a radius *R*. In a Cartesian coordination system, assuming that the coordinates of g_c are given as (0,0), the neighbouring pixel g_p occupying the symmetric circle can be denoted by coordinates $(-R\sin(2\pi p/P), R\cos(2\pi p/P))$. The value of *P* and *R* can be taken with respect to the demand of the applications. Various compositions of *P*s and *R*s are shown in Figure 2.18. The intensity level of a neighbour that lies between image pixels can be interpolated using their gray-scale levels.

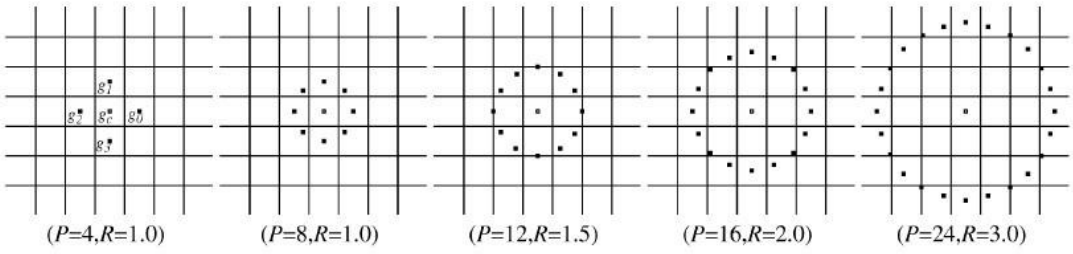


Figure 2.18. *Examples of circularly symmetric neighbours with varying values of P and R. [97]*

2.3.2. Gray-scale invariant representation

Gray-scale variations are usually unevenly distributed in different regions of an image and the types of illumination shifts are hardly identical among different images. How-

ever, the illumination transformation observed in a small region, such as the circularly symmetric neighbour mentioned in the previous section, is highly correlated to the intensity values of neighbouring pixels. The original LBP is directly based on this observation. The gray-scale variation is effectively filtered out by subtracting the intensity value of the centre pixel from those of all neighbouring pixels, the mathematical formulation of which is shown in equation (14).

$$T = t(g_c, g_0 - g_c, ..., g_{P-1} - g_c) \tag{14}$$

In such a way of representation, the illumination shift is effectively attenuated, which strengthens the gray-scale invariant property of the LBP descriptor. By decorrelating the intensity value of the centre pixel from the circularly symmetric neighbours, the texture T now can be characterized by two parts: the general illumination level of the whole texture g_c and the relative gray-scale levels of the whole neighbourhood with respect to the centre pixel. This decorrelation conception can be further simplified by the following equation:

$$T \approx t(g_c)t(g_0 - g_c, ..., g_{P-1} - g_c) \tag{15}$$

It should be noted that equation (15) is an approximation of the original texture, since it assumes the independence of intensity level between the centre pixel and its neighbours. Even though this assumption does not hold for every situation and may very well incur loss of information, Ojala et al. argue that the realization of illumination invariance outweighs a slight loss of information. In addition, the overall intensity level of the small neighbourhood $t(g_c)$ does not contribute much to the extraction of textural features and therefore, the loss of information does not have any effect on textural details. Since most textural information is encapsulated in the relative intensity levels of circular neighbours, the mathematical formulation of texture T can be further simplified and approximated by the following equation:

$$T \approx t(g_0 - g_c, ..., g_{P-1} - g_c) \tag{16}$$

Equation (16) is a highly discriminative representation of texture. Different textural characteristics exhibit distinctive types of intensity encoding. And the direction of the intensity level differences can be an indicator of the gradient of the texture. For example, a uniform distribution can be characterized by zero intensity differences between the centre pixel and its circular neighbours. An edge feature is usually identified by a sharp difference in one direction that represents the edge gradient. In the case of a spot, the textural feature is characterized by the pan-directional differences among all circular neighbours.

Since the most important property of textural features lies within the direction of gradient, the absolute difference between intensity levels do not carry as much weight as the signs of the differences. In addition, there still exists slight illumination variations shift from one region to another. Hence, using the signs of differences between neighbouring pixels can further provide robustness to illumination variations and simplification for textural feature representation. Equation (17) shows the mathematical formulation.

$$T \approx t(s(g_0 - g_c),\ldots,s(g_{P-1} - g_c)) \tag{17}$$

where $s(x)$ is a thresholding function that has the following form:

$$s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \tag{18}$$

After the completion of all previous operations, the small circularly symmetric neighbourhood is represented by a string of 0s and 1s, which constitutes the binary pattern, (and hence the name.) Every unique textural characteristic can be encoded with a distinctive LBP sequence. To simplify notation, each string of binary numbers is encoded into an integer by multiplying each binary position with a binomial factor 2^p, as shown in equation (19). Thus, the textural pattern of any circular neighbourhood can be represented as an integer and this integer is used to replace the centre pixel. This will result in an encoded image that heightens the textural structures of the original image. The encoded images from several scene categories are shown in Figure 2.19.

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p \tag{19}$$

It should be noted that other thresholding functions or orders of binary strings are also acceptable, since the textural pattern extracted is independent to encoding schemes. As long as the same scheme is used consistently across the image, the textural structures will be encoded uniquely and coherently.

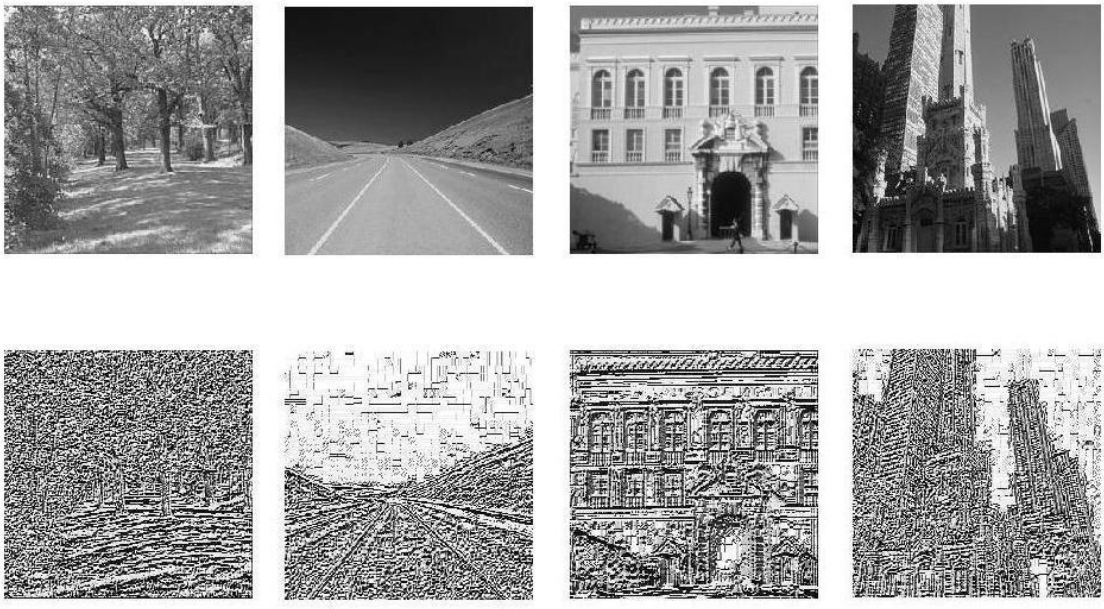


Figure 2.19. *Sample images (top) and their LBP encoded images (bottom). Image categories from left to right are: forest, highway, inside city, tall building.*

The LBP feature presented above is slightly different from its most basic form proposed in [96]. In the most primitive formulation, LBP only considers the N_8 neighbours

of the centre pixel, which is similar to $LBP_{8,1}$ without taking into account circular symmetry. The $LBP_{P,R}$ proposed in this section can be seen as a generalization of LBP. Due to circular symmetry, the property of rotation invariance can be more easily achieved.

2.3.3. Rotational invariance

The number of different values resulted from LBP encoding with P neighbouring pixels residing in a circular line amounts to 2^P. But these 2^P values do not translate into 2^P distinctive textural patterns. Suppose a sequence of intensity difference values is already computed and so is the value of $LBP_{P,R}$. But if the circular neighbours are rotated counter-clockwise one spacing altogether, the most significant bit will become the least significant bit, resulting in a shift of $LBP_{P,R}$ value, unless the sequence takes the value of all 1s or 0s. However, it does not matter if or how a circular neighbourhood is rotated, the textural pattern remains the same. In other words, similar but rotated patterns should be transformed numerically to have the same $LBP_{P,R}$ value. In light of the rotational property of circular symmetry, Ojala et al. attempt to realize rotational invariance by applying the following operation on the sequence of binary values:

$$LBP_{P,R}^{ri} = \min\{ROR(LBP_{P,R}, i) \mid i = 0,1,..., P-1\} \tag{20}$$

where $LBP_{P,R}^{ri}$ denotes the LBP that accounts for rotation variations and $ROR(LBP_{P,R}, i)$ refers to the operation of circularly right shifting the sequence of binary values of $LBP_{P,R}$ in total of i times. In much simpler words, the rotational invariant property is achieved by shifting clockwise the sequence of binary values a number of times so that the most 0 values can occupy the most significant bits.

This rotation invariance operation is quite similar to the one proposed in [103], in which only the N_8 neighbours of the centre pixel are accounted for. The only difference is, the N_8 neighbours are much more straight-forward and do not necessitate pixel interpolation. There are 36 distinctive rotation invariant LBP patterns when P is taken the value 8 and these $LBP_{8,R}^{ri}$ patterns are shown in Figure 2.20.

2.3.4. Uniform patterns

Pietikäinen et al. have demonstrated in [103] that the 36 unique binary patterns of $LBP_{8,R}^{ri}$ do not have the same probability of occurrence in real world textural images. Some $LBP_{8,R}^{ri}$ patterns emerge far more often than others and these frequently occurring patterns may account up to 90% instances of all patterns. In addition, the 8 circular neighbours are spaced $45°$ apart, which is a coarse sampling scheme and may not provide sufficient information on the local neighbourhood.

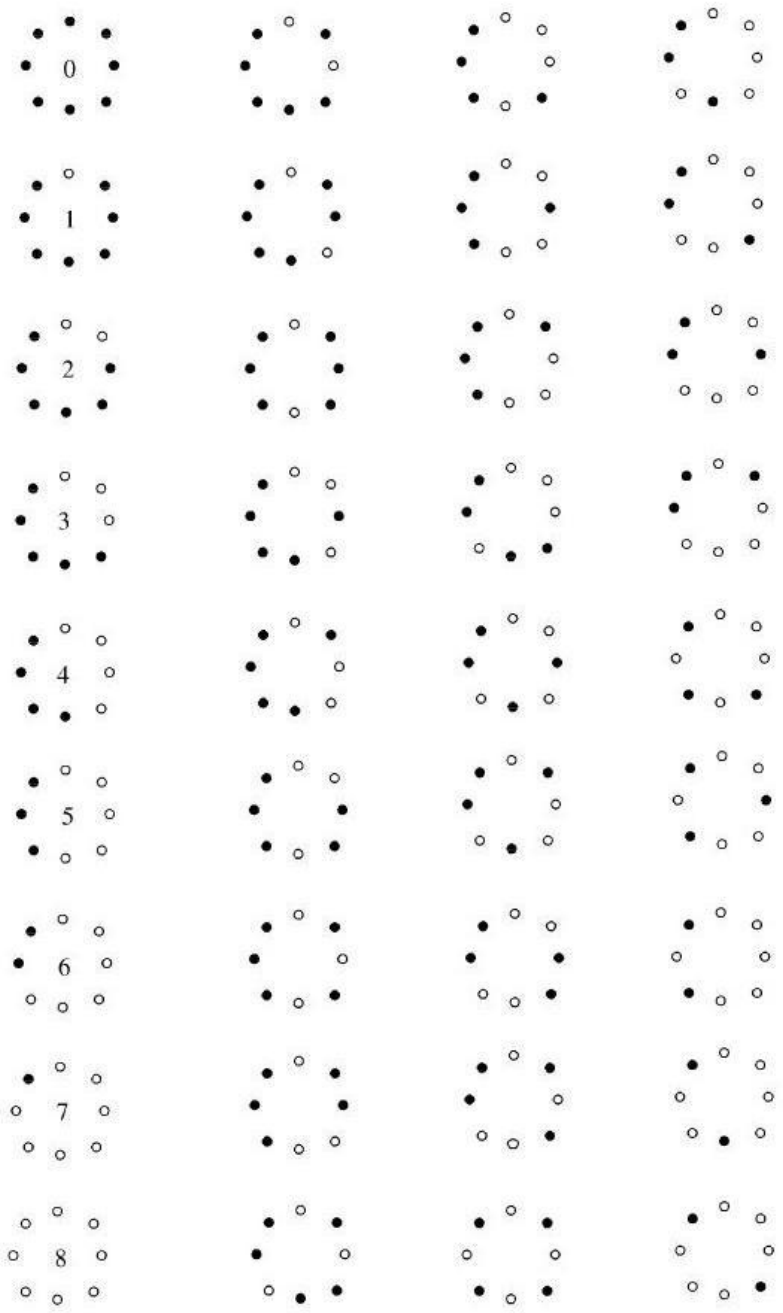


Figure 2.20. *The set of 36 distinctive rotation invariant LBP patterns that sampled around a circularly symmetric neighbourhood of 8 pixels. The black and white dots represent binary values of 0 and 1 respectively. The first column shows the nine uniform patterns. [97]*

Ojala et al. have identified the textual patterns that occur more frequently than others and they name these patterns “uniform” patterns. The first column of Figure 2.20 shows the nine uniform rotation invariant patterns that are extracted in a circularly symmetric neighbourhood of 8 pixels. These nine patterns represent the most basic textural structures, such as bright dots, dark dots and edges that present different types of gradient directions.

Upon further inspection, Ojala et al. have discovered that these uniform patterns share a unique attribute—the sequence of binary values is marked by limited 0/1(0 to 1 or 1 to 0) transitions. As shown in the first column of Figure 2.20, pattern 0 and pattern 8 do not present any 0/1 transitions at all; whereas other seven patterns exhibit 0/1 transitions 2 times. In counting the number of 0/1 transitions, one technicality should be cleared. The tabulation should start from the least significant bit g_0 and end with it as well, in a circular fashion. For example, in the case of a 11111110_2 binary sequence, the counting starts from the least significant bit 0, goes towards the most significant bit 1 and comes back to 0, which would result in two 0/1 transitions.

In an attempt to mathematically define the property of uniform patterns, Ojala et al. have devised a uniformity measure U to formally represent the number of 0/1 transitions. Experimental evidence [97] has suggested that binary patterns with a U that do not exceeds the value 2 cover the majority of all patterns and thus are defined as uniform patterns, denoted by $LBP_{P,R}^{riu2}$. And this definition can be formulated mathematically with the following equation:

$$LBP_{P,R}^{riu2} = \begin{cases} \sum_{p=0}^{P-1} s(g_p - g_c), & \text{if } U(LBP_{P,R}) \le 2 \\ P+1, & \text{otherwise} \end{cases} \tag{21}$$

where $U(LBP_{P,R})$ can be denoted as:

$$\begin{aligned} U(LBP_{P,R}) = & | s(g_{P-1} - g_c) - s(g_0 - g_c) | \\ & + \sum_{p=1}^{P-1} | s(g_p - g_c) - s(g_{p-1} - g_c) | \end{aligned} \tag{22}$$

Equation (21) suggests that if the uniformity measure tabulates to or below the value 2, the binary encoding operation $LBP_{P,R}^{riu2}$ is defined as the integer number that the sequence formulates. However, once the 0/1 transitions of a binary sequence exceeds the number of 2, the binary pattern will be ignored and all the non-uniform patterns are assigned trivial value P+1. According to equation (21), there will be P+1 possible uniform patterns in total if a circular neighbourhood is sampled at P pixels around the centre position.

After $LBP_{P,R}^{riu2}$ feature extraction operations, the encoded image is represented in the form of a histogram which is the finalized texture feature. The use of rotational invariance operation and uniformity measures has greatly reduced the dimensions of the histogram feature, from 2^P dimensions to a mere P+1 feature length without any type of quantization. In a sense, the reason why uniform patterns present better discriminative capacity is that the uniform operation has eliminated histogram bins that otherwise would not be as frequently occupied as other bins. Traditionally, these trivial histogram bins would be grouped together with important ones that signify uniform patterns and thereby dilute the descriptive power of uniform patterns. By clustering these bins to-

gether to form the non-uniform histogram bin, quantization becomes less of a necessity and the discriminative ability of uniform patterns can be fully accentuated.

Compared to the original LBP feature, the $LBP_{P,R}^{riu2}$ descriptor allows different configurations of extraction parameter setup. As shown in Figure 2.18, different combinations of P and R values can result in different sampling schemes. Such flexibility in textural feature extraction accommodates different types of textures and allows the transition between different spatial scales. The greater the value of P is, the better represented the circular neighbourhood is, thanks to finer sampling. However, the parameter P is highly correlated to the value of R. For instance, when one pixel spacing is taken as the radius R, it is unwise to sample more than 8 (P=8) pixels along the circle, otherwise oversampling will occur. Oversampling will undoubtedly result in numerical redundancy as well as computational complexity. Therefore, it is always a consideration to achieve a balance between the number of neighbouring pixels and the value of the specified radius R.

2.3.5. The CENTRIST descriptor

The CENTRIST (short for CENsus TRansform hISTogram) descriptor [147] is a newly proposed image feature that is devised primarily to extract structural details of real-world scenes. It is based on the Census Transform (CT) [151], a strikingly similar concept to LBP, and can be considered as a remarkably promising application of the LBP features in scene recognition tasks. The most attractive characteristic of the CENTRIST feature is that the descriptor operates globally on scene images and thereby ensures simplicity and efficiency in computation. In addition, because of its global nature and subsequent dimensionality reduction measures, the CENTRIST descriptor extract scene features in relatively small dimensions. Wu et al. have demonstrated in their experiment that CENTRIST descriptor can produce state-of-the-art performance on scene recognition applications.

The first step of CENTRIST feature extraction operations is the Census Transform. Similar to the LBP encoding scheme, the CT operator also compares neighbouring pixel values to the intensity level of the centre pixel. However, in the CT transform, the comparison only concerns the N_8 neighbours—the immediate neighbours of the centre pixel, which is a particular case of the general $LBP_{P,R}$ operation without taking into consideration neighbourhood circular symmetry. Additionally, in contrast to the thresholding function shown in equation (18), the thresholding function used in the CENTRIST descriptor has the reverse effect and is shown in the following equation:

$$s(x) = \begin{cases} 1, & x \le 0 \\ 0, & x > 0 \end{cases} \tag{23}$$

Equation (23) specifies that if the intensity level of a neighbouring pixel is larger than that of the centre pixel, the pixel is encoded with 0; otherwise, the value of the neighbouring pixel is replaced with 1. A visualization of such encoding scheme is

shown in Figure 2.21. Since the CENTRIST descriptor does not operate on a circularly symmetric local neighbourhood, the ordering of the binary values is also different from $LBP_{P,R}$. As opposed to the counter clockwise formulation of binary sequence in $LBP_{P,R}$, the CENTRIST descriptor scans the binary values horizontally from left to right and then order them from top to bottom to produce the binary pattern. The final CT value is an integer value computed from the binary sequence. As the CENTRIST descriptor only involves the centre pixel's 8 immediate neighbours, the possible range of the CT values is between 0 and 255. An illustration of complete CT transform is shown in Figure 2.21. Similar to the $LBP_{P,R}$ feature, the CT transform is also invariant to local illumination variations.

$$\begin{array}{|c|c|c|} \hline 32 & 64 & 96 \\ \hline 32 & \mathbf{64} & 96 \\ \hline 32 & 32 & 96 \\ \hline \end{array} \Rightarrow \begin{array}{ccc} 1 & 1 & 0 \\ 1 & & 0 \\ 1 & 1 & 0 \end{array} \Rightarrow (11010110)_2 \Rightarrow \mathrm{CT} = 214$$

Figure 2.21. *An example of CT transform. [147]*

The LBP descriptor is originally designed to analyze the characteristics of textures. And texture images are usually tiny image patches that measure up to 64×64 resolutions at most. This is the primary reason why LBP only considers the histogram of the whole image. In the case of scene images, however, a global histogram is incapable of representing the relationship between different structures of a scene which provides essential information to scene recognition. In addition, scene images are usually of 256×256 in dimension or more, a mere global histogram is too coarse an analysis to capture the textural details of the whole scene. Due to such practicality consideration, Wu et al. propose to represent scene images in a multilevel spatial representation approach, similar to the concept employed in [73]. Also known as *spatial pyramid*, this spatial representation scheme divides each scene into a number of blocks on a few spatial scales. Different levels of spatial scales are generated through sequential Gaussian smoothing of the image and down-sampling operations. As a scene image is processed to a coarser scale, the number of blocks also decreases. Figure 2.22 shows the multilevel spatial representation approach used in the CENTRIST descriptor. From left to right, the divided images refer to the representation at level 2, 1 and 0 respectively. The number of blocks extracted is defined by $2^l \times 2^l$, where l stands for the level number. This means that at level 2, the division will result in sixteen blocks, level 1 four blocks and level 0 just one block, as shown in Figure 2.22. Additionally, in order to counter artifacts caused by block division, the grid is shifted to the centre of the image at each level to create overlapping blocks (shown in dashed lines). This three level representation will eventually lead to 31 blocks (16+9+4+1+1) in total. With each block being de-

scribed by a 256-bin CT histogram, the total dimensions of the features tabulate to 7936 (31×256).

Since the resulting feature dimensions consume too much system resources, Wu et al. propose to use dimensionality reduction measures to scale down the feature vector. Principal Component Analysis (PCA) is performed on the CENTRIST features so that redundancy among different dimensions of the features can be effectively mitigated. By projecting the original features onto orthogonal eigenvectors, the original CENTRIST features are effectively decorrelated. Wu et al. have selected the first 40 components of the CT histogram which keep most of the data variance. With the help of PCA, the dimensions of the final CENTRIST features are reduced to 1240 (31×40), when the multilevel scene representation scheme (Figure 2.22) is used.

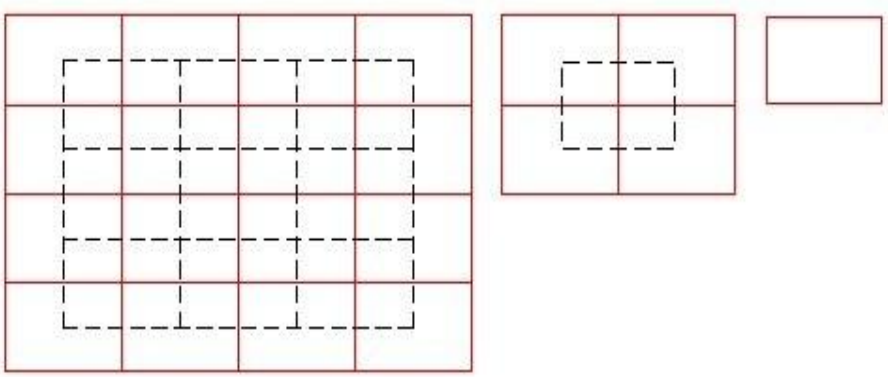


Figure 2.22. *Multilevel spatial representation. From left to right, the images represent divisions at level 2, 1 and 0 respectively. [147]*

In one experiment, Wu et al. demonstrate the discriminative capacity of the CT transform and histogram. It is shown that the CT encoding mechanism captures the textural and structural details of image content. In this image reconstruction experiment, several small image patches that depict different numerical numbers and alphabets are used as the inputs. These patches are shown on the left side of the three jointed images in Figure 2.23. The CT histogram for each input image patch is extracted for comparison. During the initial stage, the pixels in each image patch are randomly swapped, two pixels at a time. After a certain times of the swapping operations, the image patches are usually beyond recognition and are shown in the middle of each image group. Then in the reconstruction stage, a pair of pixels in each image patch is again randomly selected and swapped and at the same time, the CT histogram is computed and compared to that of the original image patch. The swapping operation finally terminates when the CT histograms of the swapped image patch matches that of the original input image patch. The final output is shown on the right side in each image group. It is quite conspicuous that by using the CT histogram as a signature of image content, the scrabbled image patches are recovered to their initial state once their CT histograms match those of the original ones. This experiment unequivocally demonstrates that the CT histogram is a

powerful indicator of the image visual appearance and provides a uniquely encoded representation of the original image.

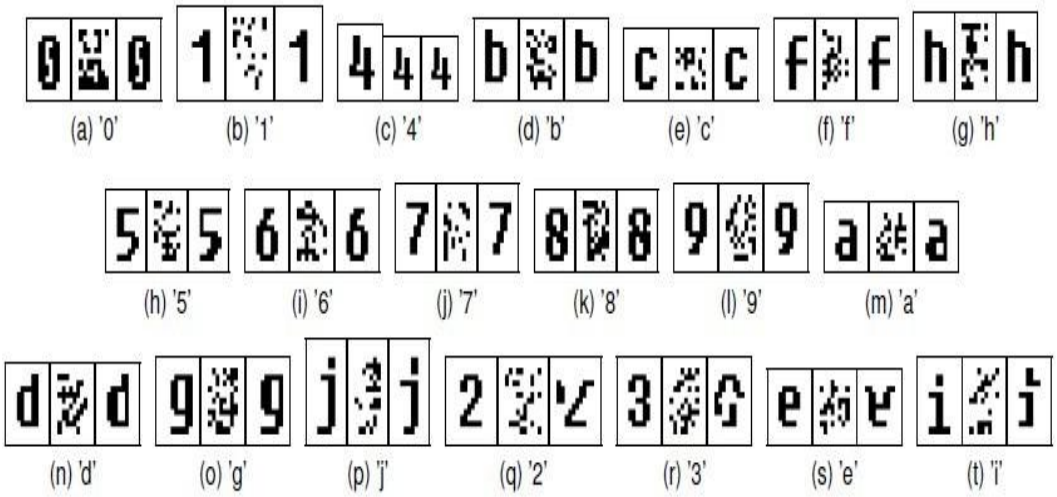


Figure 2.23. *Image reconstruction experiment. In each image group, from left to right, image patches of the input, the initial scrabbled patch and the output patch are shown.[147]*

2.4. Local features

Local image features have gained tremendous traction recently in the computer vision community. Since the introduction of the Scale-Invariant Feature Transform (SIFT) [82] in 1999, studies on local features have been a major theme in the field and contributed significantly to solutions to numerous computer vision problems, such as local image matching [80], object recognition [8], [26], [37], [38], [39], [41], [42], [67], [81], [91], [122], [152], scene understanding [32], [40], [73], [75], human detection [13], [27], [28], [124], [146], image auto-stitching [14], [15] and many others. Compared to global features, local descriptors concentrate on all or a selected subset of interest regions and effectively represent the local properties or spatial and orientation distribution of each region.

Even though local descriptors have been producing state-of-the-art recognition results, they remain to be designated tools for object recognition. There have been scene recognition studies [32], [40], [73], [75] that are modelled on the utilization of local features representing local distribution of appearance characteristics. Despite the promising results [73] garnered by some studies, local features are always plagued by computational complexity and high dimensional features, and the concept of describing local appearance of scenes does not consistently agree with behavioural findings on human perception. We argue that the computational burden of local features is not well warranted, so we briefly present the essential information about local features for the sake of completion.

2.4.1. Definition

The difference between global and local features is that global descriptors often represent images as a whole or in a few relatively large blocks; whereas local features tend to concentrate on a small image patch that often contains around 100 to 200 pixels. However, such local operations should not be confused with local transformation in global descriptors. After all, before the formulation of feature vectors, almost all descriptors, global or local, operate locally on individual image pixels or a small neighbourhood for preprocessing. For instance, the LBP operates on a small neighbourhood of often 3×3 pixels and the output of Gabor filters are orientation maps that have the same resolution as the original image. And yet, it is the mark of global features to represent the whole image with only a fraction of the features generated, e.g., the GIST feature only represents the whole image in a few blocks. In the case of local features, an image should be described by the combination of all extracted features, which usually results in high dimensionality in feature space.

Due to such high dimensional nature, local features are often not computed on every region of the image, but on regions of interest that are stable and robust to occlusion, illumination, rotation and scale variations and affine transformations. Since one major application of local features is object recognition, the same region should be detectable under different viewing conditions, which is a property called repeatability. There are various types of interest region detectors that vary in invariance properties, repeatability and complexity. And they are presented in the following section.

2.4.2. Interest region detectors

As not all regions in an image carry equally important information, using specific detectors to find distinguished regions of interest is a common practice for local feature extraction. In fact, the pruning effect offered by these detectors can be seen as a signature for local descriptors, since the proposals of new interest region detectors and local descriptors often go hand in hand [80]. Generally speaking, interest point detectors can be divided into three categories [112]: corner-based detectors, region-based detectors and other detection methods.

Corner-based detectors tend to have a preference to regions that are marked by sharp transition and strong intensity change in multiple directions, the very characteristics that suggest the presence of edges. These detectors have an excellent capacity for edge-like region detections, but often fail to distinguish uniform regions. On the other hand, region-based detectors favour salient regions that are marked by high contrast with surrounding regions, e.g., a region of a bright blob. Other methods may use local information (entropy) as relevant consideration for region detection or select interest regions based on the human perception system. The most notable interest region detectors are summarized as follows:

- Harris/Hessian point detectors, Harris-Laplace, Hessian-Laplace [36], [53], [92]
- Difference of Gaussian region detector [80]

- Harris or Hessian affine region detectors [90]
- Maximally Stable Extremal Regions (MSER) [89]
- Entropy Based Salient Region detector (EBSR) [61], [62], [63], [64]
- Intensity Based Regions/Edge Based Regions (IBR/EBR) [138], [139], [140]

2.4.3. Interest region descriptors

Upon obtaining a number of interest regions, they should be properly described so that the local properties can be represented and heightened. Same as the case in global features, it is theoretically plausible to use the concatenation of pixel intensity levels of the interest region for description. But the curse of high dimensionality also applies to local features and a pixel-wise matching scheme does not provide invariance to illumination, rotation or other shifts. To this end, several interest region descriptors have been proposed to tackle such predicament. And these descriptors can be generally divided, according to [91], [112], into three categories, namely distribution-based descriptors, filter-based descriptors and other approaches.

The most widely used local descriptors perhaps all belong to the distribution-based category. These descriptors evaluate the intricate properties by way of a local histogram. The histogram usually concerns the localization of interest points in the region of interest as well as the evaluation of gradient orientations. The SIFT descriptor [82] falls into this category and is the most prominent feature in this realm. Figure 2.24 illustrates how different oriented gradients are binned into a local histogram based on the locations of the gradients and their orientations. The second category includes descriptors that concern the use of filters to extract local properties. Notable approaches are differential invariant descriptor [116], steerable filter [46] and complex filter [7]. The local descriptors that are cast into the third category often rely on the statistics of local pixel intensities. One approach is as simple as comparing the intensity values of local regions between two images, using a measure called cross-correlation which can be perceived as a similarity measure. Other applications tend to summarize the local details or colour distribution with moment invariants [144].

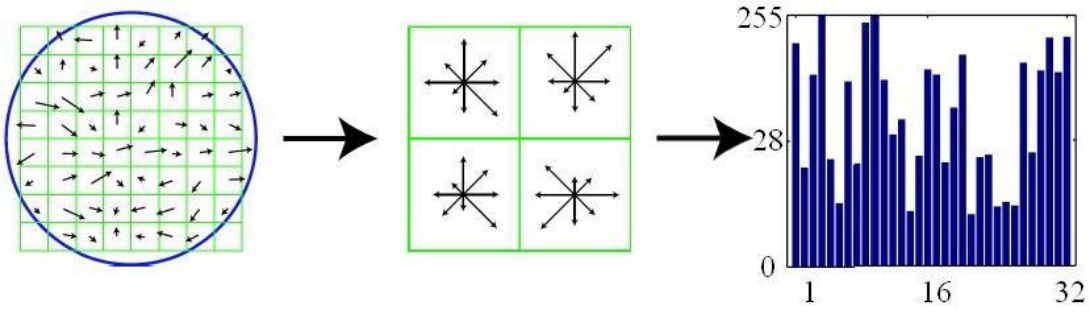


Figure 2.24. *Illustration of the formulation of the SIFT descriptor. The image on the left depicts the gradient magnitude and orientation at each sample point. These magnitudes and orientations are localized on one of four grids depending on the location of the sample point (middle). Finally, they are concatenated into a histogram. [82], [112]*

2.5. The bag of words (BoW) representation

In this section, we introduce an image representation approach that breaks the barrier of feature localization. In traditional image matching methods, dissimilarity measures are computed between two full-sized images or their corresponding blocks, which enforces strong spatial correspondence. In reality, however, the corresponding image regions between two images can vary greatly in terms of spatial location, imposing significant computational complexity if cross-location region matching is considered. But in a bag of words (BoW) representation [123], each image is perceived as an orderless composite of visual words (Figure 2.25). Without any spatial information, image matching is proceeded as a process of establishing the correspondence between visual words.

The BoW model is highly contingent on the utilization of local features. In fact, the inception of BoW seems to be an enthused response to the increasing popularity of local features. The “word” in the BoW stands for a local description of an image patch. Since these “words” are in fact local features, BoW is also referred to as “bag of features” by some scholars. Due to this dependency on local descriptors and sophistication of the algorithm, BoW often imposes undesirable computational cost and resource consumption on the application system.

2.5.1. Introduction

The term bag of words is borrowed from the realm of text/document classification, in which a document is represented by not every word in the document, but the most important ones. This notion stems from the observation that the document label is highly correlated to the most significant words in the document. For instance, if an email is filled by words such as “win”, “money”, “cash”, “lottery” and etc., there is an exceptionally high probability that it is a spam email which should be automatically placed into the junk folder.

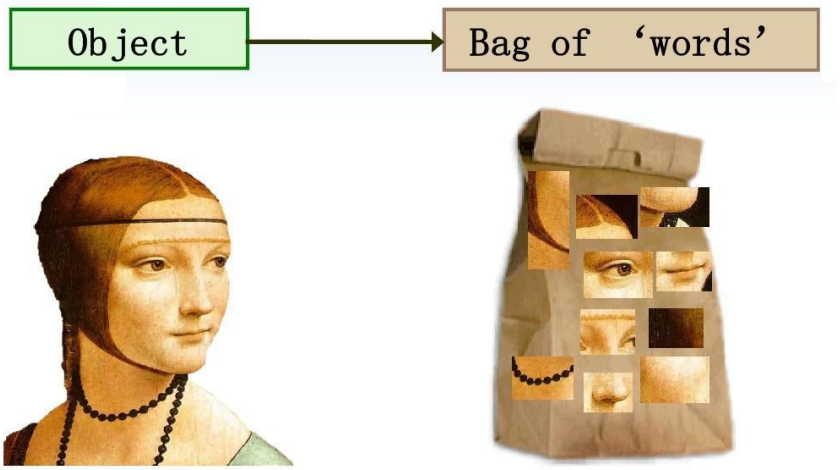


Figure 2.25. *An illustration of the bag of words (BoW) representation. An image is defined by a "bag" of local features of the image. Courtesy of Fei-Fei Li from the Vision Lab at Stanford.*

Similarly, in the field of computer vision, BoW specifies the concept of perceiving each image as a bag of visual words and each visual word refers to the local description of an image patch. As shown in Figure 2.25, the direct relation between the image on the left and the BoW on the right is a unique representation, which lends a great deal of discriminative capacity to the approach. Similar to the BoW in document classification, the BoW representation of an image is expressed in terms of a histogram. The histogram reflects the number of occurrences of a few visual words and interpretation of the occurrence histogram leads to its classification.

2.5.2. Implementation procedure

The implementation of the BoW approach is mostly based on the BoW operations in document classification. First, the visual words of an image are extracted to represent the content of the image; second, a vocabulary or code-book has to be built based on the visual words extracted from a considerable amount of images that are carefully selected from all image categories so that the vocabulary can be a comprehensive representation of frequently occurring visual appearances; finally, the visual words from each image can be assigned to an index in the vocabulary, which will result in a word occurrence histogram as the final image feature.

- Local feature extraction: In the original BoW proposals, the visual words are described around a number of key points detected using one of many interest region detectors. This is the common practice of feature extraction for object recognition applications. In scene recognition applications, however, the semantics of a scene depends on the distribution of both edge-like patterns and uniform patterns. Current region detectors can hardly fully satisfy both measures. Therefore, in the case of scene recognition, local image features are extracted on a densely sampled grid with overlapping image regions that measure 16×16 in resolution to fully represent the visual appearance of a scene [40], [73]. Figure 2.26 shows an example of a densely sample grid employed in [73]. Although theoretically, any local features can be used for visual word extraction, most scholars in the field prefer the SIFT features.

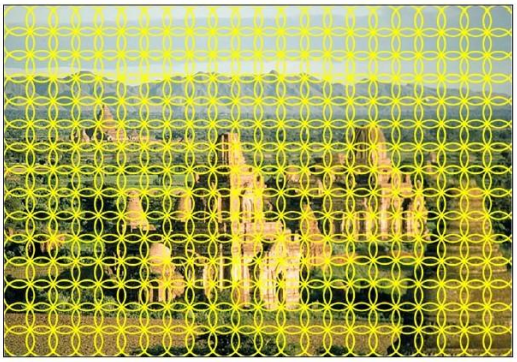


Figure 2.26. *An example of densely sampled grid. Each image patch is of 16×16 in resolution and spaced 8 pixels apart. Courtesy of Svetlana Lazebnik.*

- Formulation of the vocabulary: A representative visual vocabulary should cover almost all the variations of visual appearances in the application. To this end, a large quantity of visual words should be used to establish the vocabulary. Specifically, the visual words of images that fully represent the visual diversity of every image category are pooled together for vector quantization (VQ), as shown in the last row of Figure 2.27. The most widely used VQ technique is K-means clustering, although more sophisticated VQ algorithms are also applicable. Special attention should be paid to the size of the vocabulary: a under-quantized vocabulary (with too many visual words) can lead to misassigning similar visual words into different vocabulary positions; whereas a over-quantized vocabulary (with insufficient visual words) do not possess adequate discriminative ability to differentiate between distinctive visual appearances, resulting in undesirable mismatch.

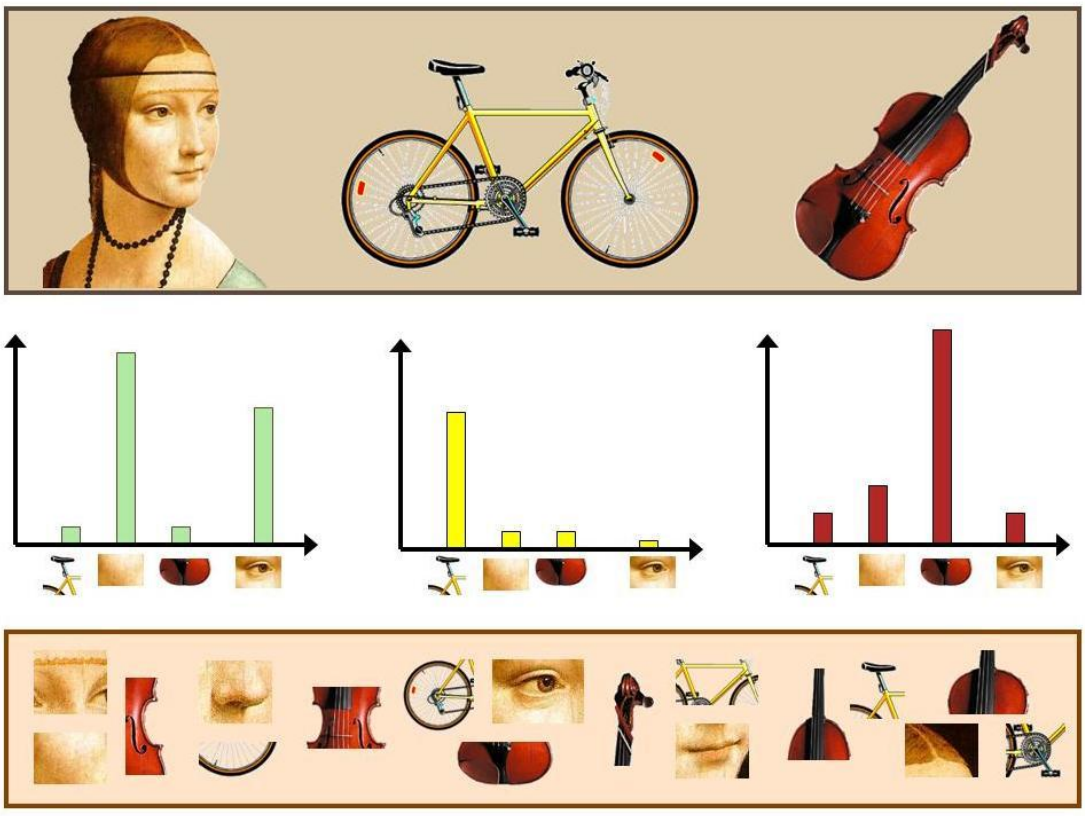


Figure 2.27. *An example to show the implementation procedure for the BoW representation. The first row shows an image from each category. The last row signifies the process of visual word pooling and quantization. The final product of vector quantization is a vocabulary which is shown below the horizontal axes of the histograms in the middle row. The histograms are generated from assigning the visual words from the image to positions in the vocabulary. Courtesy of Fei-Fei Li from the Vision Lab at Stanford.*

- Generation of histograms: Once a vocabulary is properly formulated, the local features of every image patch in each image from the dataset can now be assigned to one of the positions in the vocabulary according to a predetermined dissimilarity measure. The histogram of an image is the final BoW representation and it specifies the makeup of visual words for the image. The significant correlation between the histogram and the image it represents ensures the discriminative capacity of the BoW approach. An example of these histograms is shown in the middle row of Figure 2.27. The positions on the horizontal axis stand for words in the vocabulary.

3. LEARNING AND CLASSIFICATION MODELS

For scene classification tasks, a proper form of machine learning algorithm has to be used for automatic scene category inference. In this chapter, the basic notion of machine learning is briefly introduced. This overview covers the formulation of the problem, types of algorithms and popular approaches.

In particular, the two classes of models in machine learning, namely the generative model and the discriminative model, will be introduced and further explained in terms of general conception. Additionally, the strengths and weaknesses of both models will be summarized. One particular approach in the class of discriminative model—support vector machines (SVMs)—will be discussed independently because several scene classification studies have demonstrated its superiority over other approaches in this line of application [73], [99], [147]. The SVMs will be used extensively in the following two chapters to show the effectiveness of the proposed methods.

3.1. Definition

Machine learning is a major component in the research of Artificial Intelligence (AI). It refers to the formulation of algorithms or systems that enable a machine to improve performance based on given data and self-evolve in AI tasks, such as recognition, classification, prediction, planning etc.

In his widely praised book *Machine Learning*, Tom M. Mitchell has given a more accurate and formal definition of machine learning in technical terms: “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.” [93] Despite the clarification of this definition provides, it hardly offers any insight into the necessity of machine learning. Preference for incorporating machine learning over a static algorithm into systems is given to one or some of these cases:

- When the only known aspects of a task are the identities of inputs and outputs, one cannot design a system to define the exact mechanism of the task. In such cases, where the exact relationship between inputs and outputs is beyond comprehension, the only viable solution is to approximate the intricacy of the task by using available examples. Such approximation can be modelled by machine learning algorithms.
- For some well-defined tasks, certain aspects of the working environment cannot be fully understood or modelled. The use of machine learning algorithms for

such cases can provide flexibility for the system so that it can adapt to the real working environment by itself.

- Some working environments are not static. Rather some characteristics change over time. In such occasions, the system can resort to machine learning algorithms for adjusting the structure of the system and render proper or evolved outputs with respect to the new environment.
- For certain tasks, new information regarding the mechanism emerges constantly. It is for designers' best interest to devise a system with machine learning algorithms such that the system itself can evolve over time with newly acquired knowledge about the task.
- The complexity of some tasks may be too overwhelming to be fully implemented into the system by human designers. In this case, a better approach is to leverage the evolutionary property of machine learning algorithms to allow the system gradually discover the sophisticated definition of the tasks and automatically adapt to it.
- There might be hidden variables or relationship between the observed and target variables. Proper machine learning algorithms can often help to model the hidden process and extract the hidden variables or relationship.

For scene classification, the objective of the task is to render a proper label for each image given its scene features. The only information is the knowledge of scene features for all images and the correlation between some scenes and their features. Clearly, the task is only defined by a few examples and the exact structure of the system is yet to be known. In this case, proper machine learning algorithms can approximately establish the relationship between scene labels and features.

3.2. Types of learning

Since the advent of machine learning algorithms, there have been studies that focus on different types of learning algorithms. But some research themes have been given more attention and precedence than others. These approaches can be categorized according to the availability of data labels—the target variables of the function that maps the inputs (observed variables) to the system outputs (target variables). If there are a large amount of labelled data for inferring the mapping function, with which the system can take new data and predict their labels, this type of learning is called *supervised learning*. If the correspondence between the observed variables and the target variables is unknown (no labelled data), then this algorithm is called *unsupervised learning*. In this case, the primary objective of the algorithm is to make proper partitions among the data points according to some implicit rules regarding the properties of the data. The learning algorithm that comes in between is *semi-supervised learning* which deals with the situation where only a limited number of labelled data are available. This type of learning algorithm can be useful to refine the previously learned mapping function with the help of unlabelled data.

In the case of scene recognition, most studies employ supervised learning to render label inferences. Hence in the following two chapters, only supervised learning algorithms are considered for testing the effectiveness of proposed scene features.

For scene classification tasks, the performance of any supervised learning algorithm relies directly on the representation of scene images. The better scene images are represented, the more effectively the algorithm will detect the patterns. Thus, under the same learning algorithm with the same experimental setup, the performance of the learning algorithm directly reflects the discriminative capacity of scene features.

Another major concern for using machine learning algorithms is the dimensionality of observed variables. In a general sense, the higher the dimensions of scene features are, the better scene images are represented. Higher dimensions of features contain more information for the machine learning algorithm to render a proper prediction. However, with the increase of dimensionality in the feature space, the volume of the space also grows. To avoid sparsity in the feature space, the amount of training data should be increased exponentially so that these features can be used to generate viable statistical predictions. This phenomenon is called the curse of dimensionality and should be avoided in any machine learning application.

3.3. Generative model

In machine learning, a generative model refers to a model that is capable of directly generating observable variables. The primary objective of building a generative model is to estimate model parameters by using the observed data points. It takes the form of a joint probability distribution function between the observed variables and target variables through the link of some hidden variables. Since the final objective of machine learning is to classify the observed variables into one or a number of categories, which often requires the estimation of the posterior conditional probability distribution, the generative model can be used to predict via the combination of the generative joint and prior probability distributions according to Bayes' rule.

Figure 3.1. *A graphical illustration of the generative model.*

Figure 3.1 shows an illustration of a graphical model of the generative approach. In the figure, *F* stands for the image features and *C* refers to the category the image belongs to. The direction of the array indicates that the generative model is established

through the conditional probability distribution of $P(F \mid C)$—the probability distribution of feature *F* given the category label *C* of the image—and the marginal distribution of *C* according to Bayes' rule which is specified in the following equation:

$$P(F,C) = P(F \mid C)P(C) \tag{24}$$

Since this is not a direct modelling of the classification decision probability distribution $P(C \mid F)$, the generative model often requires the introduction of hidden variables to establish the relationship between the image feature and its category label.

In the case of scene recognition, local features are often implicitly required in order to build a generative model for classification purposes, since the variations of visual appearance throughout a scene should be fully accounted for. These local features are used to generate the intermediate variables "themes" (e.g., rocks, water, sky, etc.) which establish the link between scene features and category label. The utilization of these intermediate variables not only necessitates detailed description of local structures, but also mediates the model-building process in a hierarchical manner. In [40], Fei-Fei et al. propose to generate such a model based on a hierarchical Bayesian text approach using latent Dirichlet allocation (LDA) [11]. The general concept of the modelling process is shown in Figure 3.2. The category label *c* is linked to the codewords *w* of image patches in a scene through the intermediate variables of mixture of themes π and the patch themes *z*. In other words, the generative model of a scene category is represented by the probability of a mixture of patch themes (e.g., rock, sky, road etc.) with respect to the themes of image patches from the same category.

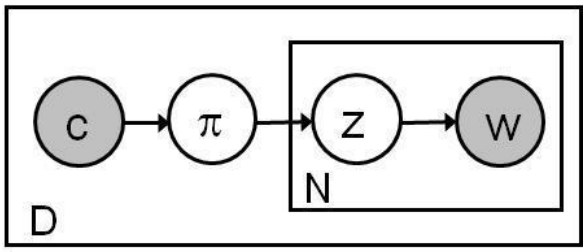


Figure 3.2. *A graphical illustration of latent Dirichlet allocation (LDA).Courtesy of Fei-Fei Li in Vision Lab at Stanford.*

One major advantage of the generative model in scene classification is that by introducing intermediate variables into the modelling process, the system is highly intuitive and related to document classification. The generative model is a fully statistical model that can be easily translated into classification. Since the probability distributions of both observed and target variables are modelled, this approach is adaptive to introduction of new knowledge regarding the relationship between variables. And the generative model should be a top choice when the prior distribution is given for the classification task.

However, the implementation process that leads to generation of the model is rather tedious. The problem of modelling the theme topics from image patches may be ill-posed since some image patches may exhibit different themes. In other words, precise segmentation is not guaranteed, which could lead to ambiguous patch identification. Additionally, spatial information is not incorporated into the modelling process, nor is strong geometric information. This partly explains the redundancy of the implementation process and the vocabulary.

3.4. Discriminative model

A discriminative model is marked by the direct mapping between observed variables and target variables in the form of a conditional probability distribution $P(C|F)$, as shown in Figure 3.3, where *F* and *C* denote the feature vectors and category labels respectively, consistent with the notions used in the generative model. In a sense, the discriminative approach appears to be a black box in which the relationship between observable variables and the prior probability distribution cannot be discovered. Without going through proper statistical modelling, the discriminative approach is more straightforward.

Figure 3.3. *A graphical model of the discriminative approach.*

There are numerous machine learning algorithms or architectures that fit into this category. Some of the most widely used are logistic regression, support vector machines (SVMs), artificial neural networks etc. In scene recognition applications, the SVM approach has been given precedence over other algorithms due to its maturity and the availability of open-source programs. The popularity of the SVMs necessitates further explanation on this subject in the next section.

3.5. Support vector machines (SVMs)

Originally proposed by Boser et al. in [12], the support vector machine approach is a supervised machine learning solution specifically devised for pattern recognition. Based on justified statistical learning theories, the SVMs are capable of learning from training examples to construct a set of hyperplanes that separate the data points into two different classes with the maximum margin in the high dimensional feature space. SVM provides a direct mapping from observed variables to target variables, which is the result of

learning the parameters of a mapping function. The mapping function is usually selected and optimized according to a set of constraints, among them is the minimization of training errors. Even though other types of SVMs are proposed after their initial presentation, the original SVM is a strictly discriminative approach.

Being a general pattern recognition method, SVM has been used in a variety of applications. The modular characteristic also adds to its popularity since users need not to worry about the implementation details of the algorithm. The publishing of SVM library [21] provides users with an easy access to pattern recognition, which has tremendously benefited other studies.

3.5.1. Problem definition

In mathematical term, the primary objective of SVMs is to learn a mapping: $\chi \mapsto y$, where $x \in \chi$ represents the feature vector (observed variables) and $y \in y$ denotes the class label (target variables). In a binary classification problem, where $x \in R^n$ (n dimensional feature space), $y \in \{\pm 1\}$, the goal is to learn a classifier defined in equation (25), with function parameters α, given the training set $(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)$. Once the mapping function is learned, the algorithm can take into new unlabeled data points and render a label with a certain level of confidence.

$$y = f(x, \alpha) \tag{25}$$

Since there are many functions that can separate the training data with respect to their labels, as shown in Figure 3.4, the optimization process should be constrained with additional requirements so that the best hyperplane is selected. One requirement of SVM is that the selected hyperplane should measure the largest distance between classes; the other is the constraint on the training error which should be kept to a minimum.

According to linear separability of data points from different classes, SVMs can be linear or nonlinear. Nonlinear SVM is simply an extension to linear SVM. It first maps linearly inseparable data points in the original space to a higher dimensional feature space, in which the data points can be linearly separable again. Such mapping is realized through the use of kernel functions.

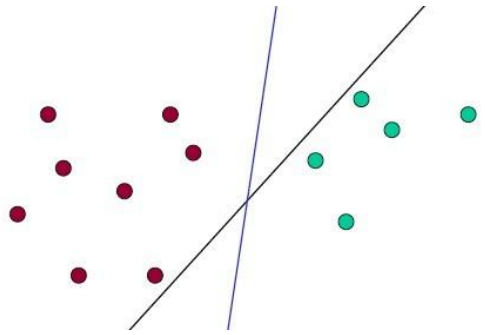


Figure 3.4. *An example of separating data points from two different classes. There exist many hyperplanes that can manage such data separation.*

3.5.2. Linear SVM

In the case of linear SVM, a hyperplane can be expressed as a linear combination of the dimensions of observed data point x_i ($w \cdot x_i - b = 0$, where w is the normal vector, b denotes the offset of the hyperplane from the origin and $\cdot$ the dot product.) In association with class labels, the hyperplane should be defined such that the following equations are satisfied:

$$\begin{cases} x_i \cdot w + b \geq +1, \text{ if } \mathrm{y}_i = +1 \\ x_i \cdot w + b \leq -1, \text{ if } \mathrm{y}_i = -1 \end{cases} \tag{26}$$

Equation (26) can be rewritten as a single form:

$$y_i(w \cdot x_i + b) \geq 1 \tag{27}$$

The data points that lie exactly on the hyperplanes should satisfy $x_i \cdot w + b = +1$ or $x_i \cdot w + b = -1$ and are called support vectors. Recall that in the definition of SVM, the hyperplanes are constrained to have the maximum margin. Since the margin between them can be defined as $\frac{2}{\|w\|}$ according to simple geometric deduction, the objective can be morphed to minimize $\|w\|$. Combining the objective and the constraint, we can arrive to the theoretical formulation of linear SVM which is to minimize $\|w\|$, subject to equation (27).

Since the norm of $w\|w\|$ involves square root operations, it is difficult to minimize. Without changing the general solution, the problem is reformulated to: $\min(\|w\|^2)$, subject to equation (27). This formulation can be expressed with Lagrange function that has the following form:

$$L_p = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m} \alpha_i [y_i (x_i \cdot w + b) - 1] \tag{28}$$

where α_i are the Lagrange multipliers. This problem can be solved with quadratic programming.

3.5.3. Nonlinear SVM—the kernel trick

Since the normal vector can be expressed as $w = \sum_{i=1}^{m} \alpha_i y_i x_i$, alternatively the linear SVM can be formulated as the following optimization problem:

$$\begin{aligned} L(\alpha) &= \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \\ &= \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k(x_i, x_j) \end{aligned} \tag{29}$$

subject to $\alpha_i \geq 0$ and $\sum_{i=0}^{m} \alpha_i y_i = 0$. In equation (29), the dot product is expressed in a linear kernel function $k(x_i, x_j)$ which can be easily extended to nonlinear cases.

It is not always guaranteed that in the original feature space, data points are linearly separable. In such cases, the training data should be first transformed into a higher dimensional feature space in which the linear separability property is satisfied so that linear SVM can be directly applied to perform pattern recognition. Suppose a feature transformation function can be denoted as $\phi(x_i)$, then the kernel function in equation (29) can be expressed as

$$k(x_i, x_j) = \phi(x_i) \cdot \phi(x_j) \tag{30}$$

Thus the nonlinear SVM problem is transformed into linear SVM in a higher dimensional feature space by replacing the linear kernel function with a nonlinear one. There are many types of nonlinear kernel functions, the most widely used in scene recognition, however, is the Radial Basis Function (RBF):

$$k(x_i, x_j) = \exp(-\gamma \left\| x_i - x_j \right\|^2), \quad \gamma > 0 \tag{31}$$

An example of using RBF as a kernel function for nonlinear class separation is shown in Figure 3.5.

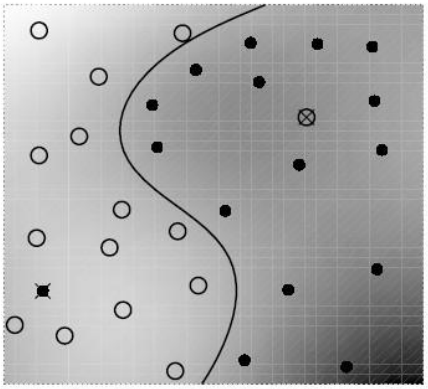


Figure 3.5. *Nonlinear separation of data points from two different classes. The nonlinearity of the boundary indicates that the observed variables are in the original feature space.*

4. ARP-GIST SCENE FEATURE

In this chapter, a novel scene descriptor is introduced. Inspired by the Angular Radial Partitioning (ARP) scheme, the ARP-GIST descriptor resolves ambiguity of spatial structures that are previously unaddressed in the original GIST features. By applying rough spatial layout estimation, the original GIST descriptor can only represent scene features at a coarse resolution. By further dividing rectangular blocks on a roughly sampled grid into a few angular bins, structural details within each block can be better represented and accentuated, accompanied by the multiscale-orientation analysis of GIST. With the help of one dimensional (1-D) discrete Fourier Transform (DFT), the stricter spatial conformity imposed by additional angular partitioning can be loosened to allow positional flexibility in a circular fashion within each rectangular block. In other words, the use of angular partitioning extracts structural details without enforcing their positions. Thus, the novel scene descriptor ARP-GIST provides a balance between detail extraction and spatial conformity.

We will show in the experimental section of this chapter that the proposed scene descriptor can better represent spatial structures of a scene than the original GIST, vindicated on two publically available scene image datasets, one of which is the testbed for validity of the original GIST, as shown in the first chapter. We also compare the performance of ARP-GIST with the popular BoW approach. Even though the classification accuracy on scene images obtained by ARP-GIST is only marginally superior to that of BoW, the performance improvement of the proposed feature is warranted in terms of computational cost and resource consumption. Such performance superiority also indicates that at a single level, spatial layout properties are essential to efficient recognition of a scene. Without including spatial information, scene representation will result in unjustifiable redundancy.

The proposal of ARP-GIST is originally introduced by the author of this thesis in [78].

4.1. Angular radial partitioning (ARP)

ARP has been successfully applied in content-based image retrieval (CBIR) [19], sketch-based image retrieval (SBIR) applications [20] and object recognition [9]. It employs both angular and radial partitioning that is similar to the polar coordinate system. One main advantage of ARP is its ability to capture intricate structures in an angular-spatial manner, as opposed to a simple spatial distribution in a rectangular partitioning scheme. Since the resulting blocks are arranged on a circle, it is much easier to achieve rotational invariance with ARP. Figure 4.1 demonstrates a typical ARP strategy.

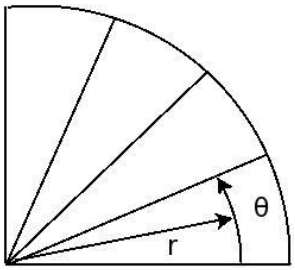


Figure 4.1. *An illustration of Angular Radial Partitioning. r corresponds to radial bins and* θ *denotes angular position that can be quantized into several bins.*

Spatial layout is an important part of a scene image as it carries essential information regarding its category. In order to preserve relative spatial layout of a scene and allow moderate intra-class variations in scenes from each class, i.e., the presence of the stove can be in the middle of the image or to its left centre, the GIST descriptor is computed on an *N*-by-*N* grid. Even though this coarse partitioning scheme has yielded significant success in terms of recognition accuracy in scene classification tasks, it fails to represent spatial structures efficiently within a block as the averaging operator often renders different structures indistinguishable, resulting in mismatch among scene categories. Figure 4.2 shows an example of such a deficiency. It is clear that even though the spatial structures are visually different for human observers, the GIST feature vectors cannot really discriminate between the two distinct images.

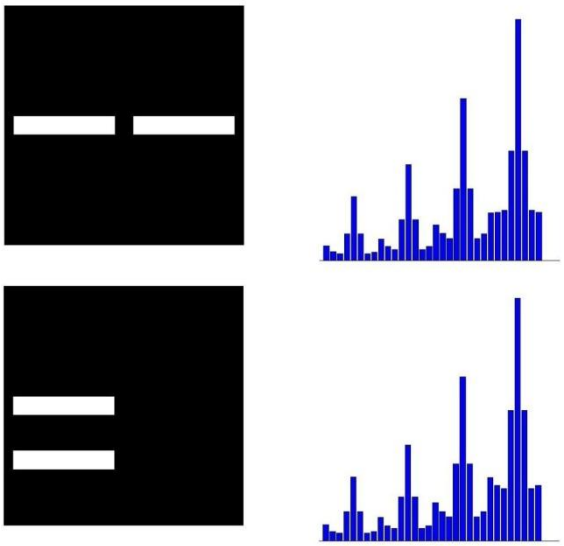
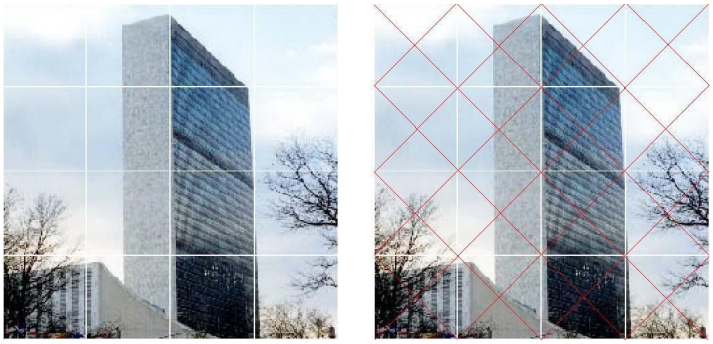


Figure 4.2. *An illustration of the limitations of the GIST descriptor. Two distinctive blocks are shown on the left and their corresponding GIST features on the right. The striking similarity between the two feature vectors suggest that they are not capable of distinguishing between the different structures in those two blocks.*

For a better representation of the spatial structures of a scene, we propose a strategy that builds on the success of the original GIST feature. In addition to the *N*-by-*N* rectan-

gular partitioning (Figure 4.3 (a)), we further divide each block using ARP into *A* angular bins, which not only extracts the coarse spatial layout but also the finer angular distribution in a scene. To avoid coincidence with further rectangular partitioning, we use the upper right diagonal as the starter of ARP in each block, as illustrated in Figure 4.3 (b).



(a) (b)

Figure 4.3. *Demonstration of rectangular partitioning and ARP: (a) image partitioned in a 4-by-4 grid, (b) ARP in addition to original rectangular partitioning (A=8).*

Figure 4.4 shows the same example in Figure 4.2 but with additional ARP. Since these two blocks are divided into 4 additional angular bins, the dissimilarity between the two resulting feature vectors becomes significant enough to distinguish the two different structures.

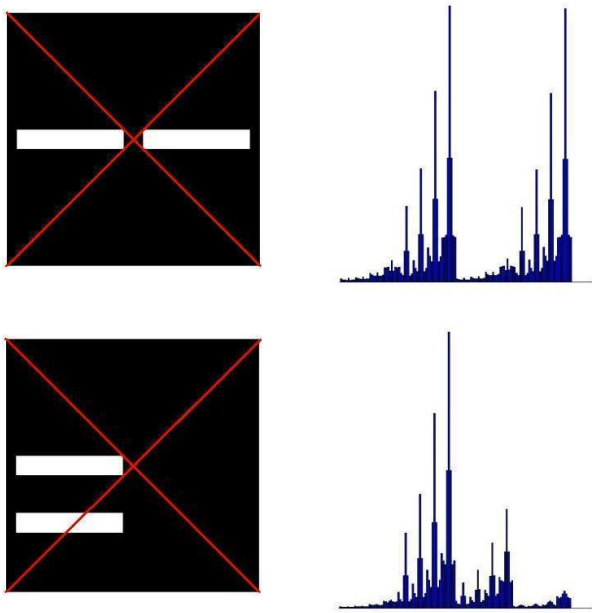


Figure 4.4. *An illustration of the dicriminative power of ARP. With additional angular partitioning, the two distinctive blocks on the left can be represented differently in the feature space, shown on the left.*

4.2. Positional invariance

Even though ARP can better delineate the spatial structure in a block, the risk is the same with other types of blocks: over-partitioning. The idea of dividing an image into blocks is to preserve some spatial layout in the process of recognition or matching. Finer partitioning means stricter layout confinement, which is not the case for different scene images in the same category. This is the reason why the original GIST descriptor is calculated on a 4-by-4 grid instead of an 8-by-8 one. Experiments (see experimental results section) have shown that over-partitioning will not improve classification accuracy, and sometimes may even induce accuracy erosion. This is also true with ARP. Further dividing the 4-by-4 grid can sometimes degrade the leeway gained by better representing the structure since the same spatial structures in different scene images within the same category often enjoy spatial freedom within an area of the image, i.e., a computer can be at different positions along the surface of the desk.

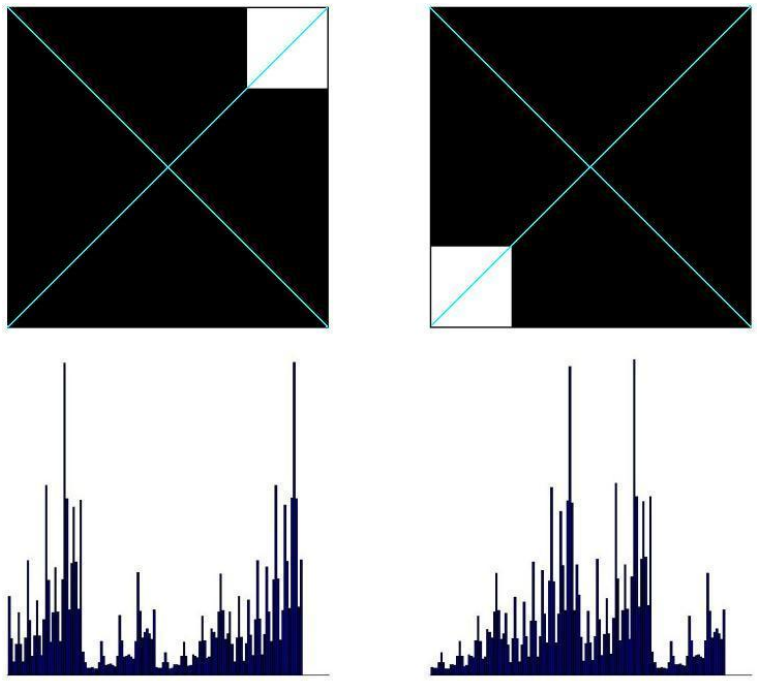


Figure 4.5. *A toy example to show the undesirable effect of spatial conformity imposed by ARP.With additional angular partitioning, the same block structure (top row) can result in different feature vectors (bottom row).*

This point can be illustrated by a toy example shown in Figure 4.5. The top row depicts two image blocks that exhibit the same structure at different locations. If ARP is used in conjunction with the GIST descriptor, the resulting feature vectors (bottom row) will not match each other due to the fact that the same structures fall into different angular bins.

In light of such dilemma, the proposed method utilizes the discrete Fourier transform (DFT) to achieve rotational or positional invariance.

Let I denote an image block and A denote the number of angular partitions. The angle in each bin can be computed as $\theta = 2\pi / A$. Then the i[th] element of the feature vector of one block can be formulated as follows:

$$f(i) = \frac{1}{S} \sum_{\theta=\frac{i2\pi}{A}}^{\frac{(i+1)2\pi}{A}} I(\theta) \tag{32}$$

for i = 0, 1, 2…A-1, where S is the total number of image pixels that fall into each angular bin.

If the block is rotated counter clockwise $\tau = l2\pi / A$ radian (l = 0, 1, 2…A-1) around the centre of the block, then the image block, denoted as, I_τ, can be represented by the following equation:

$$I_\tau(\theta) = I(\theta - \tau) \tag{33}$$

Through simple mathematical deduction, we can arrive at the relationship between $f_\tau(i)$ and $f(i)$:

$$f_\tau(i) = f(i-l) \tag{34}$$

Clearly there is a distinction between $f_\tau(i)$ and $f(i)$. But with a simple 1-D DFT, the similarity between the two features can be easily observed. After applying DFT to $f_\tau(i)$ and $f(i)$, we obtain:

$$F(u) = \frac{1}{A} \sum_{i=0}^{A-1} f(i) e^{-j2\pi ui/A} \tag{35}$$

$$F_\tau(u) = \frac{1}{A} \sum_{i=0}^{A-1} f_\tau(i) e^{-j2\pi ui/A} \tag{36}$$

$$= \frac{1}{A} \sum_{i=0}^{A-1} f(i-l) e^{-j2\pi ui/A} \tag{37}$$

$$= \frac{1}{A} \sum_{i=-l}^{A-1-l} f(i) e^{-j2\pi u(i+l)/A} \tag{38}$$

$$= e^{-j2\pi ul/A} F(u) \tag{39}$$

According to equation (39), the DFT of the rotated feature vector is merely multiplied by a certain angle to that of the original one. Note that the magnitudes of the DFT coefficients of both feature vectors are the same, that is $\|F_\tau(u)\| = \|F(u)\|$. Therefore, we use the norm of 1-D DFT coefficients to achieve rotational or positional invariance.

It should be clarified that the term rotational variation referred in this occasion does not stand for the rotation of scene structures around a centre point. Rather it denotes the positional shift around the angular partitioning of a block. Therefore, we will refer to this property as positional invariance for the rest of the section lest confusion with rotational invariance be provoked.

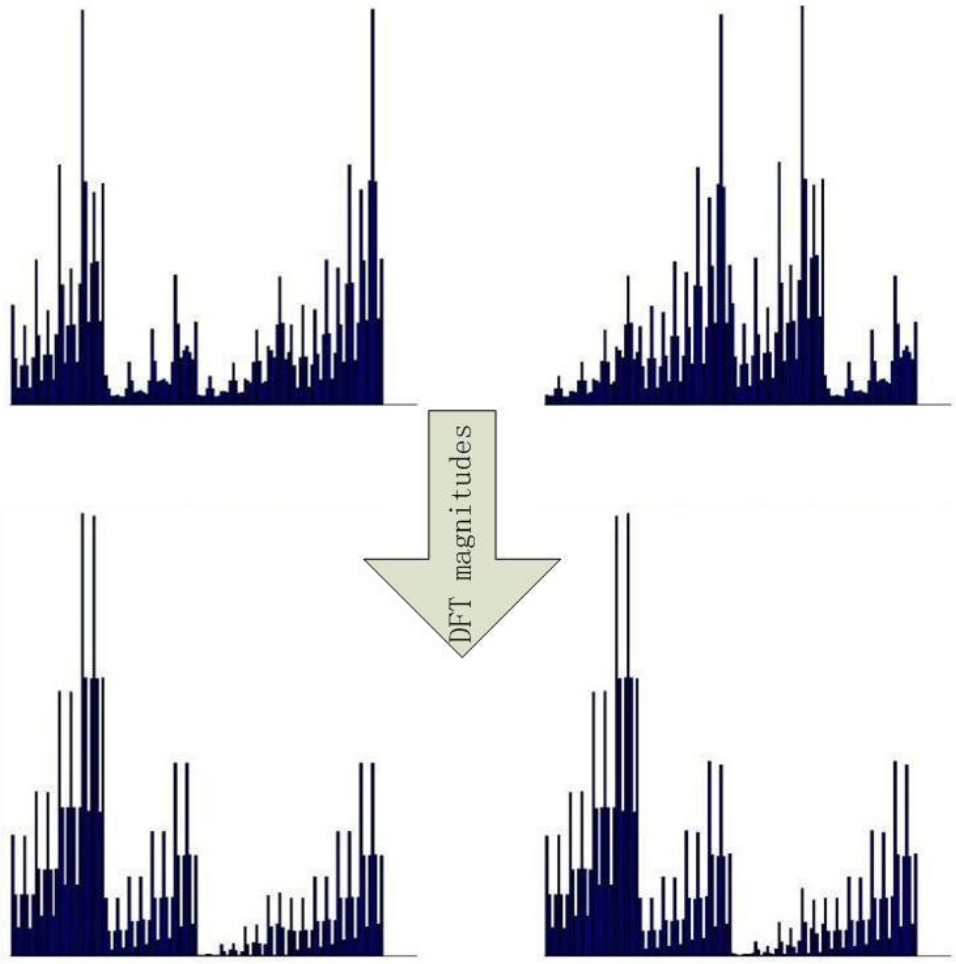


Figure 4.6. *An illustration of the effctiveness of 1-D DFT. The magnitudes of DFT ensure that the same structure will lead to the same feature vector regardless of its position.*

Figure 4.6 shows the gap-bridging effect of 1-D DFT transformation. The top row in the figure shows the original ARP-GIST feature vectors of the same structure shown in Figure 4.5 without DFT. Due to the periodical property of the DFT, the magnitude of the DFT coefficients for the two feature vectors, shown on the bottom row, can effectively discard the angular positions of the spatial structures. This shows that the proposed feature extraction method will render scene features based on visual appearance of the scene structures in the block without concentrating too much on their exact locations, which points to the positional invariant property of the proposed descriptor.

In addition, the DFT operations do not reduce the discriminative capacity of the ARP-GIST descriptor. Figure 4.7 shows the effect of DFT on different scene structures in a block. It is quite obvious that even with additional operations of DFT, the final feature vectors (composing of the magnitudes of the coefficients) for the two distinctive block structures after applying ARP on each block can still differentiate the visual dissimilarity between them. Thus, the application of DFT is capable of avoiding both false positive and false negative matching in scene recognition tasks.

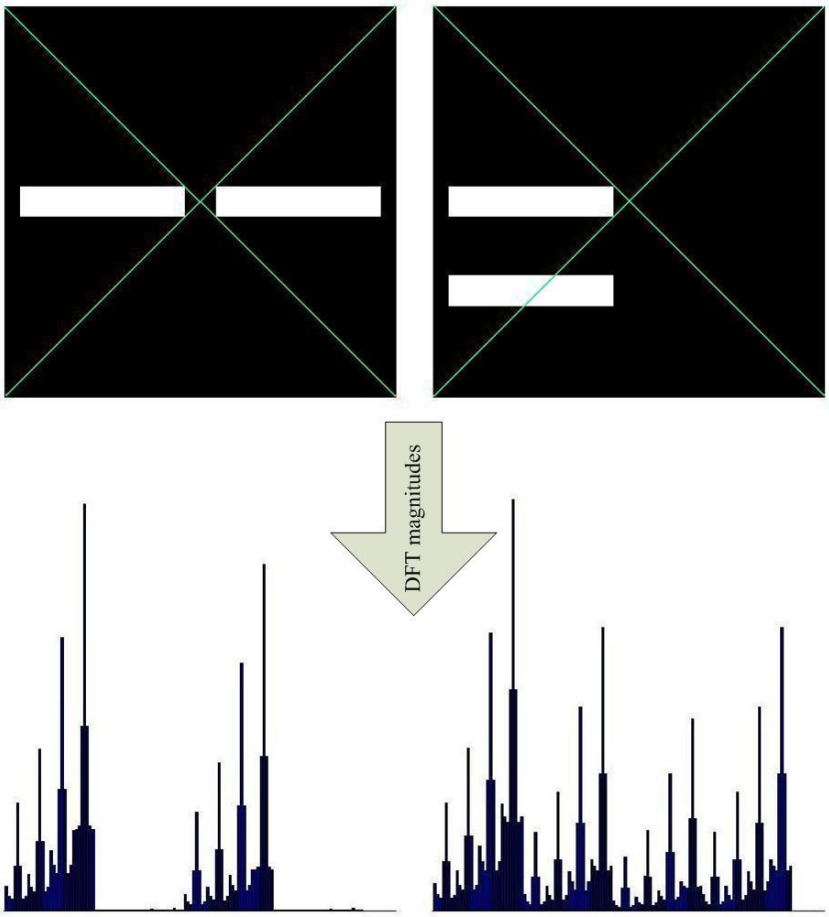


Figure 4.7. *The resulting ARP-GIST feature with DFT from two different block structures, as shown in previous figures.*

4.3. Implementation procedure

First, a gray-scale image is pre-processed by a whitening filter to preserve dominant structural details and then normalized with respect to local contrast (equation (11)). The pre-processed image is then passed through a cascade of Gabor filters (equation (12)) in *S* scales with *O* orientations at each scale. Each of these *S*×*O* images (orientation maps), representing the original image at one orientation in each scale, is then divided on an *N*-by-*N* grid. (For the original GIST feature, the average intensity is computed in each block to represent its feature. The final output is a concatenated feature vector of *S*×*O*×*N*×*N* dimensions.)

Instead of taking the average value within each block on the *N*-by-*N* grid, we further partition each block into *A* angular bins using ARP. To avoid over-partitioning, only angular partitioning is considered; in other words, the number of radial partitioning is set to 1 for all blocks. Then the average intensity level is computed in each angular bin, followed by a 1-D DFT on the angular bins in each block and then taking the magnitudes of the coefficients to achieve positional invariance. Finally, the feature vector is obtained by concatenating all the DFT transformed bins in the image across all the ori-

entations and scales, resulting in an $S \times O \times N \times N \times A$ dimensional feature vector. Figure 4.8 shows the complete block diagram for the proposed method.

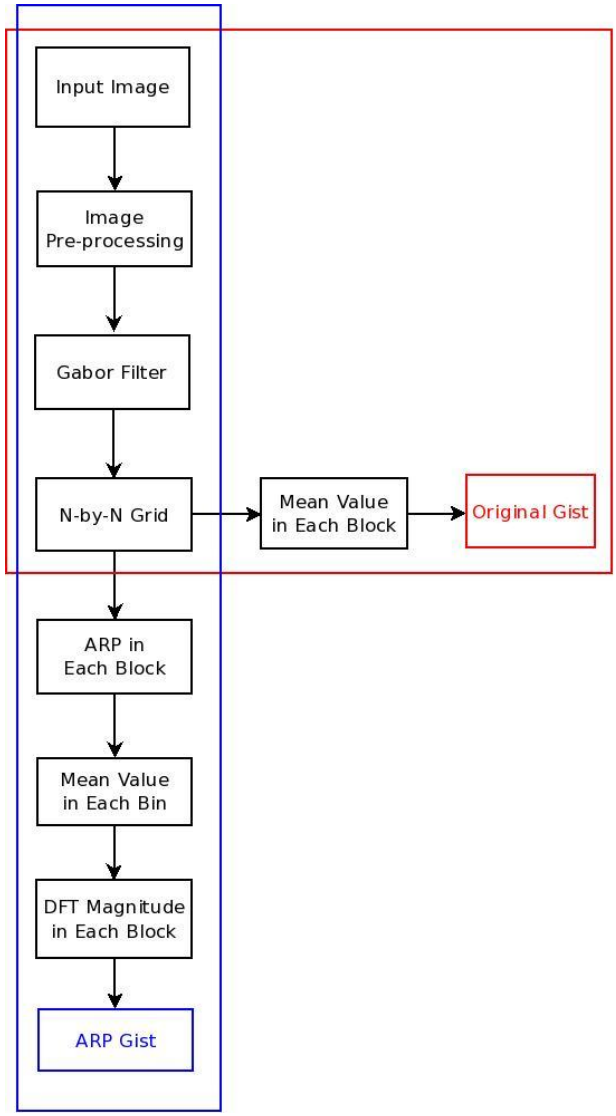


Figure 4.8. *Flowchart of the original GIST operations and the proposed ARP-GIST descriptor.*

4.4. Experimental setup and results

In this section, we present the experimental setup to evaluate the effectiveness of the proposed ARP-GIST descriptor. In order to demonstrate its superiority, we adopt the same image and parameter settings as used in the proposal of the original GIST descriptor.

4.4.1. Image normalization

Since the algorithm is based on the spatial structures within scene images, we consider only the luminance component, for which we use the mean pixel values of the R, G, B

channels. In order to ensure comparability, all images are resized to the resolution of 256×256 using bilinear interpolation; therefore the aspect ratio of each image is completely ignored. This is in line with the experimental setup used by Oliva et al. in their implementation.

It should be noted that other types of colour space can also be applied to the experiments. Notably, any colour space that allocates one component directly to luminance, such as the YUV and the CIELAB colour space (initiated by CIE, Commission Internationale de L'éclairage - International Commission on Illumination), can be considered. In such cases, there is no need to normalize but to use the luminance channel for ARP-GIST experiments. Since scene images are often loaded onto the RGB colour space, in this case, colour space transformation has to be used to convert colour components. (see appendix for colour space transformation equations between RGB and YUV, and between RGB and CIELAB.)

4.4.2. Parameter settings for feature extraction

The parameters for image pre-processing (image whitening and local contrast normalization) are kept the same as with the original GIST, and so are the parameters for Gabor filters. The images are filtered by a jet of Gabor filters at 4 scales, with 8 orientation channels at each scale. For the original GIST descriptor, each image is divided into $N \times N$ (N=4, 8) blocks and the average is taken in each block. Hence, the total dimensions of the feature vector for each image are $4 \times 8 \times N \times N = 32N^2$ for the original GIST.

ARP is applied to each block on a 4-by-4 grid. The number of angular partitioning (A) used in our experiment is 3, 4, 5 and 6 respectively to fully evaluate the performance of ARP-GIST. In each angular bin, we take the average value to represent the feature of that bin, resulting in a feature vector of size $4 \times 8 \times 4 \times 4 \times A = 512A$.

4.4.3. Classifier training

SVM training and testing are conducted 1000 times so that generality can be achieved. We randomly select 100 images in each category for training and the rest for testing. This processing is done 1000 times to ensure effective comparison between the proposed algorithm and the original GIST descriptor. Note that the comparison is based on the same 1000 sets of training and testing data.

In our experiment, we use Gaussian Radial Basis Function (RBF) as the kernel to build one-versus-all classifiers, the scaling factor in equation (31) is defined in our experiment as the following:

$$\gamma = \frac{1}{p \times f} \tag{40}$$

where p is the kernel parameter, which is set to 0.003 in all experiments, and f is the number of dimensions of the feature vector.

The confusion matrix for every training/testing set is recorded during each run. The final classification accuracy is the average value of the mean of the confusion matrix diagonal.

4.4.4. Results on the spatial envelope dataset

Sample images from each category of the spatial envelope (SE) dataset are shown in Figure 1.3. The performance comparison between the original GIST and the ARP-GIST feature is summarized in Table 4.1. The average accuracy rates are percentage numbers over all 8 categories.

Table 4.1. *Comparison of classification accuracy on the SE dataset.*

Method		Classification Accuracy
Original GIST	N=4	83.2661±0.7757
	N=8	83.2664±0.7417
ARP-GIST	A=2	84.5626±0.7358
	A=3	**84.7671±0.7141**
	A=4	84.6186±0.7097
	A=5	84.2832±0.6986
	A=6	83.6655±0.7177

As the results summarized in Table 4.1 indicate, the average classification accuracy obtained by the original GIST descriptor (N=4) is 83.2661%, with a standard deviation of 0.7757. Note that this is slightly lower than the 83.7% reported by Oliva et al. because of different training configurations: in our experiment, we have selected 1000 different training/testing sets in SVM to evaluate average performance. In contrast, the proposed ARP-GIST has shown improvement over the original, with the best configuration (A=3) yielding a classification accuracy of 84.7671%. To show the validity of the improvement, we have also tested the original GIST on an 8-by-8 (N=8) grid (resulting in a feature vector of 2048 dimensions, equivalent to ARP-GIST when A is set to 4), which results in 83.2664% accuracy. This is in parallel to the 4-by-4 grid GIST with almost intelligible accuracy improvement. As observed in Table 4.1, the proposed algorithm has the superiority in terms of classification accuracy.

4.4.5. Results on the UIUC 15-category dataset

As an extension to the SE dataset, sample images from additional categories are shown in Figure 1.4. Table 4.2 shows the performance comparison between GIST and ARP-GIST, along with classification accuracy rates achieved by single level BoW.

In this dataset, the classification accuracy achieved by the original GIST is only 72.6739% with a standard deviation of 0.7133. In contrast to the previous dataset, the over-partitioned original GIST (8-by-8 grid) has suffered a slight accuracy erosion, with

an accuracy rate of 72.4312%. On the other hand, the proposed ARP-GIST has yielded classification rates above 74%. The best result (75.2474%) is obtained when the number of angular partitioning is set to 4. Note that the feature vector dimension in this configuration is the same with 8-by-8 grid of the original GIST. To show the significance of performance improvement obtained by ARP-GIST, we have also included the classification rates achieved by BoW algorithm [73] in Table 4.2. The BoW feature is based on image patches on a densely sampled grid, without the usual process of interest point detection. The SIFT descriptor is computed on each image patch. The experiment is conducted on 200 (M=200) and 400 (M=400) vocabulary size models. It is evident that even without building the codebook, saving significant computational cost, ARP-GIST is still superior to the BoW model. (i.e., note that images used in BoW model are not normalized to 256×256 resolution. If normalized, the model will suffer significant accuracy degradation [106].)

Table 4.2. *Comparison of classification accuracy on 15 scene category dataset.*

Method		Classification Accuracy
Original GIST	N=4	72.6739±0.7133
	N=8	72.4312±0.7068
ARP-GIST	A=2	74.4612±0.6864
	A=3	75.0379±0.6811
	A=4	**75.2474±0.6717**
	A=5	74.8499±0.6713
	A=6	74.2130±0.6777
BoW [73]	M=200	72.2±0.6
	M=400	74.8±0.3

4.5. Conclusion

Built on the original GIST descriptor, the proposed ARP-GIST descriptor utilizes the effectiveness of angular partitioning to capture the finer details of scene images. With the DFT transform and magnitude of its coefficients, ARP-GIST allows positional invariance of scene structures within a rectangular block. The proposed method not only preserves rough spatial layout, but also provides flexibility in each block, achieving a balance between spatial constraints and freedom. Experiments on two datasets have shown that the proposed method is superior to the original GIST and rivals the state-of-the-art BoW model in terms of classification accuracy and computational cost.

5. GIST-LBP FEATURE AND BLOCK RANKING

In a scene image retrieval system, the dimensionality of scene features should be kept to a minimum for fast response without compromising retrieval performance. Most scene features, however, are over 200 dimensions even when dimensionality reduction methods are incorporated. (In the case of local features, the total feature dimensions are often in the magnitude of 100 thousand, so the feature extraction process is time-consuming even if the semantic concepts used as image features are of 20 dimensions.)

In this chapter, we present a novel scene image retrieval framework (as shown in Figure 5.1) that achieves state-of-the-art retrieval performance with compact and low-dimensional feature vectors. The proposed framework leverages the discriminative powers of both the GIST descriptor and LBP. The low dimensionality of feature vectors is realized through the use of PCA and subsequent block ranking, aspiring to reach a balance between retrieval accuracy and efficiency. With block ranking, feature dimensions can be further reduced to a mere 150 or less while achieving superior retrieval performance with top-ranked blocks. Furthermore, the user can explicitly specify the total dimensions of low level features to be used in the retrieving process with respect to retrieval speed and accuracy requirement, making the system more user-friendly.

Experiments show that the proposed framework manages to improve performance by more than 12% over feature fusion and approximately 30% over individual feature.

The original proposal of the work is penned by the author of this thesis and submitted to European Conference of Computer Vision 2012 [77].

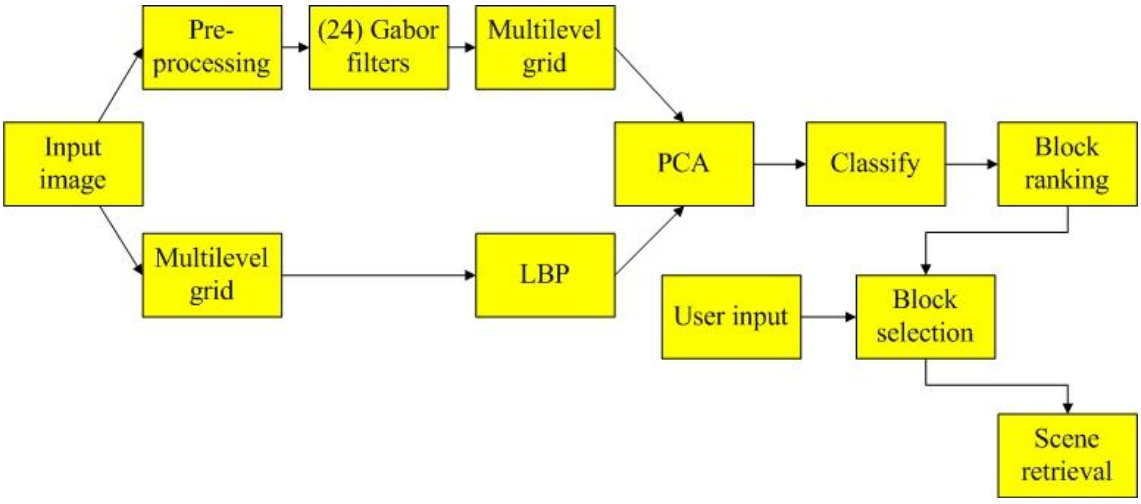


Figure 5.1. *The framework of the GIST-LBP scene retrieval system with block ranking.*

5.1. Multilevel scene representation

Compared to the performance of GIST, the fairly inferior classification result obtained by using single level sparse SIFT feature suggests that a single level representation of a scene using image features, albeit how discriminative they might be, without incorporating spatial information can produce inferior outcome. Contrary to object recognition, where spatial information is filtered through image segmentation, efficient scene categorization requires the innate presence of scene logic, i.e., in a coast scene, the sky often occupies the upper half of the image whereas the beach or ocean can be observed in the bottom half. The spatial layout is essential to scene recognition as by diving images into several regions, the variations of visual appearance become more pronounced across different categories within the same region. Such high degree of variations can therefore loosen the discriminative requirement of image features. That is probably the reason why two 1000-word vocabularies used by sparse SIFT performs as poorly as 56.6% accuracy, while a mere 400-word code-book can accurately classify more than 81% of images in a spatial pyramid scheme.

Similarly, we represent scene images in a multilevel fashion. Each image is processed at three levels, with the highest level being the original image and presenting the most structural details. Lower level images are obtained by applying a Gaussian smoothing filter and subsequently down-sampled using bilinear interpolation by a factor of 2 at each level. We employ 3 levels in total, with image size of the lowest level being only 1/16 of the original image. At the first and second levels, the images are partitioned on non-overlapping 4×4 and 2×2 grids. In addition, on the assumption that the most important area in an image is closely located inside the field of view, we also divide the central parts of the first level image into a 3×3 grid, and the second and third level images 1×1 grids to avoid over-partitioning. An illustration of the multilevel scene representation approach is shown in Figure 5.2. It should be noted that this two part dividing scheme is equivalent to overlapping partitioning and very similar to the one used in the CENTRIST descriptor [147].

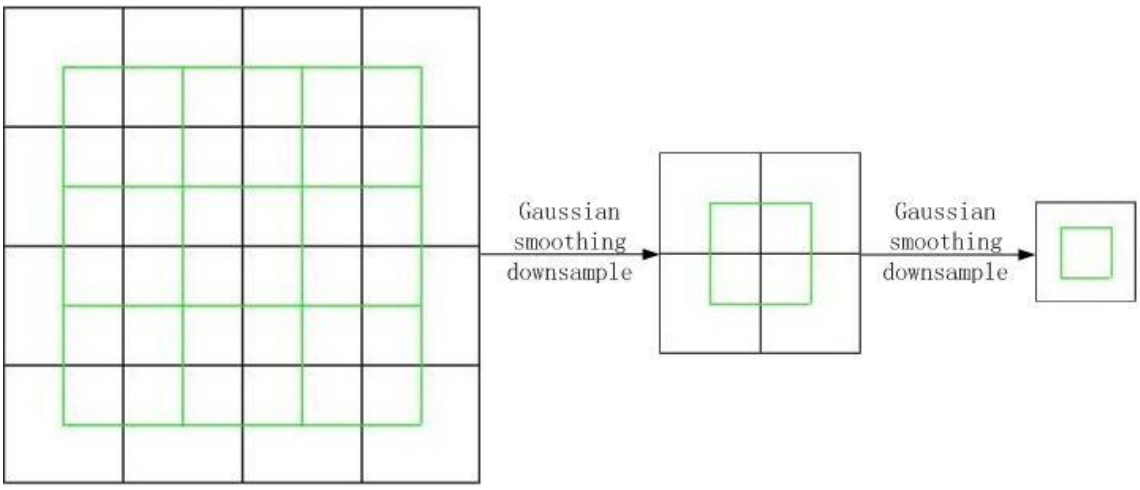


Figure 5.2. *An illustration of multilevel scene representation.*

5.2. The GIST-LBP scene descriptor

The GIST-LBP scene descriptor is a combinatory feature extraction method based on the original GIST descriptor and a variant of the LBP feature. The process involves separate feature extraction using each descriptor in a multilevel scene representation manner. Due to the high dimensional nature of the combined feature, a feature selection process is then applied to reduce its dimensions.

For feature selection, we apply PCA on features extracted in each block. From preliminary experimental results we observed the high compactness of the proposed GIST-LBP feature: a mere 10 principal components can preserve most of its discriminative capacity. Therefore, the proposed feature extraction process will lead to a low-dimensional and yet highly descriptive scene feature.

5.2.1. GIST feature extraction

We use the standard operations for GIST feature extraction: images are preprocessed (equation (11)) using a whitening filter and normalized with regard to local contrast so that the structural details are heightened and gray-scale invariance is achieved. Then the immediate output is fed through a jet of Gabor filters (equation (12)) with a total of 32 filter banks (4 scales and 8 orientations) before collecting the average energy in each block.

With multilevel scene representation, the final GIST feature amounts to $(4\times4+3\times3+2\times2+1+1+1)\times32=1024$ dimensions in total. (The dimensionality of the GIST feature used in this work is double the size of the original GIST due to the multilevel representation approach.)

5.2.2. LBP feature extraction

As explicitly explained in Chapter 2, there are a few variants of the LBP descriptor. The $LBP_{P,R}^{riu2}$ operator, for instance, is a fully integrated descriptor that accounts for rotational variations and is capable of detecting uniform patterns. In the case when P=8 (8 circular neighbours) and R=1 (1 pixel spacing), there are only 9 (P+1) uniform patterns and 1 non-uniform trivial pattern. In other words, without any form of quantization, the total number of histogram bins is 10, which is inadequate to describe the textural structures in one block with unequivocal certainty. On the other hand, without any consideration of rotational invariance and uniform patterns, the resulting histogram bins will tabulate to 256 dimensions, assuming no quantization is applied, which would impose significant computational burden.

In the proposed GIST-LBP feature, we seek to reach a balance between the two extremes and propose to use the $LBP_{8,1}^{u2}$ operator for extracting LBP features. In such a setting, we consider a circularly symmetric neighbourhood of 8 pixels besides the centre pixel as the basic unit for LBP feature extraction. The rotational invariance property of the descriptor is ignored while we concern only uniform patterns. This setting will result

in 59 possible unique patterns in total. Therefore, the final LBP feature will amount to (4×4+3×3+2×2+1+1+1)×59=1888 dimensions with multilevel representation taken into consideration.

5.2.3. Feature selection with PCA

Since high dimensional feature vector can exert too much computational cost on retrieval systems and increase responding time exponentially, feature selection methods should be considered for computational efficiency. The most widely used feature selection algorithm is Principal Component Analysis (PCA), which transforms the original, usually correlated data into uncorrelated variables under the new orthogonal bases. The transformation is devised such that the first few components carry the most variance of the whole data set, which makes it the ideal technique for feature selection in an unsupervised fashion. One of the attractive characteristics of PCA is that it minimizes the reconstruction errors from the principal components, allowing erasion of insignificant information in the original variables without losing essential details.

A toy example of such orthogonal linear transformation is shown in Figure 5.3. In the original feature space on the left, feature one and two can be considered as a projection to the original coordination system which is often non-orthogonal, unlike the axes shown in the figure. This nonorthogonality indicates correlation among variables. Through PCA transformation, an orthogonal coordination system can be extracted from the original data and these data are projected onto the new orthogonal axes, marked green on the right side of the figure. The variable projected onto the first principal component direction carries the most information of all the original variables. The principal component directions are also called eigenvectors and the projections are called scores, coefficients or eigenvalues.

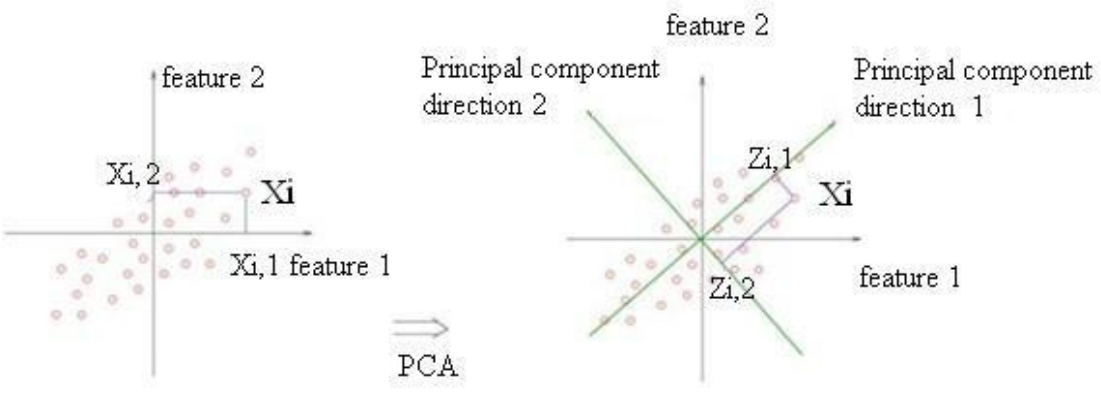


Figure 5.3. *A toy example of PCA in a two dimensional space.*

The PCA transformation can be implemented through single value decomposition (SVD) of the original variable matrix X or eigenvalue desomposition of the covariance matrix of X. The matrix X is arranged such that the rows correspond to variables and

the columns represent observations of each variable. First, the variable matrix should be centred to zero mean by subtracting observation mean of each column. Then the eigenvectors of the covariance matrix X^TX can be derived from SVD whose equation is depicted as following:

$$X=W\Sigma V^T \tag{41}$$

where W is the eigenvector matrix of XX^T, Σ is a rectangular diagonal matrix that contains the ranks of the SVD and V is the eigenvector matrix of X^TX. Finally, the original variables X can be projected onto their eigenvectors contained in V.

Instead of using two separate eigen systems for projecting the GIST and LBP features, we use the combined feature from both sources in each block to derive the eigenvectors so that the block-wise fusion of both features can complement each other and increase the discriminative capacity of block features. Preliminary results from experiments show that classification accuracy using a limited subset (less than 10 principal components) of variables significantly improves over correlated and non-orthogonal data. The reason for this result might be twofold: first, the LBP transformation process leaves vast numerical redundancy unhandled as the common circular neighbours from adjacent cells are accounted for twice in computation (shown in Figure 5.4), as reasoned by Wu et al. in [147]; second, the multilevel scene representation scheme involves overlapping image regions, which can be further decorrelated by using multi-block PCA transformation.

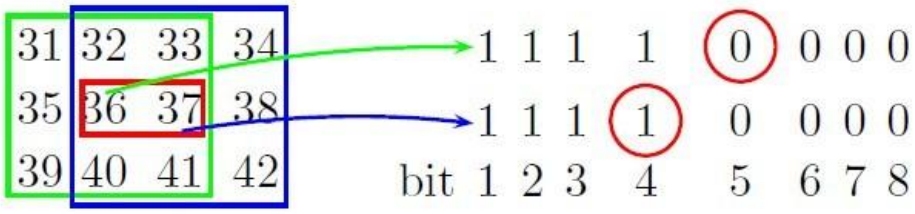


Figure 5.4. *Numerical redundancy in LBP transformation. The pixels in red that are encoded into two binary sequences are highly correlated. [147]*

Without losing generality, however, the eigenvectors for PCA transformation are derived from each block and by retaining only the 10 principal components, we arrive at a feature dimension of 32×10=320 in total.

5.3. Block ranking method

In the feature space, each block carries certain descriptive information of the whole scene and not all block features contribute equally to the semantics or visual appearance of the scene: some block features exhibit dominant scene structures (e.g., texture of trees in a forest scene) that are relevant to its semantic label while others only provide trivial information (e.g., human presence in a forest scene.)

With such rationale, we present in this chapter a block ranking approach with respect to scene image retrieval. We will show that even coarsely labelled scene images can be used to estimate the distribution of block features in the feature space. Through retrieval estimation, blocks are ranked according to compactness of feature distribution, with the scene features from top-ranked blocks representing the most discriminative information for scene retrieval. The number of blocks used for scene retrieval can be determined in correspondence to accuracy demand and resource consumption limit.

5.3.1. Scene labelling

In order to estimate the clustering effect of block feature distribution in the feature space, scene images are first labelled based on their global GIST-LBP features. This is equivalent to a scene recognition task in which a classifier takes in the feature vector of a scene and renders a statistical prediction of its categorical label.

Since SVM has been shown to work exceptionally well with scene categorization applications, we also use SVM to label scenes from image database. Although linear SVM can produce relatively good results, we choose to use RBF (equation (31)) as a kernel to SVM, which according to our experiments performs better than its linear counterpart. The kernel parameter γ is computed according to equation (40), in which the parameter p is set to 0.003 in all experiments and f refers to the number of dimensions of feature vectors, (the GIST-LBP feature tabulates to 320 dimensions). Images from each category are randomly divided into training set (100 images) and testing set for cross validation. The SVM is configured as one-versus-all.

5.3.2. Estimation of block feature distribution and block ranking

Not all blocks within a scene contribute equally to effective retrieval of visually and semantically similar images. On one hand, every image is defined according to the semantics of major scene regions. It is not uncommon to observe structures or details that are dispensable or irrelevant to the identification of a scene. For example, the presence of a cruise ship is not as essential as the texture of ocean or beach when it comes to classifying a scene as "coast". And sometimes, these irrelevant structures may even occupy a large portion of a scene and ultimately lead to misclassification. In a scene retrieval task, these blocks in a query will eventually "weigh the scene down"—scene features that correspond to irrelevant blocks will carry too much weight in the distance measure and semantically inaccurate scenes may be returned as a top-ranked retrieval. In order to achieve a better performance, blocks that fall into this category should be discarded in measuring distance between feature vectors. On the other hand, if each image feature is viewed as a combination of block features, different block features exhibit different distribution among scene images in the feature space. Hypothetically speaking, for one particular scene category, the block that can retrieve semantically relevant images may be located on the lower right corner, and for others, it may be in the centre of the scene. What is equally possible is that for a scene image, certain blocks

are closely distributed together with those of semantically different images. If these block features are included in the final feature set for computing the distance measure, unrelated images will be pulled closer to the query image as a result of equal block weighting.

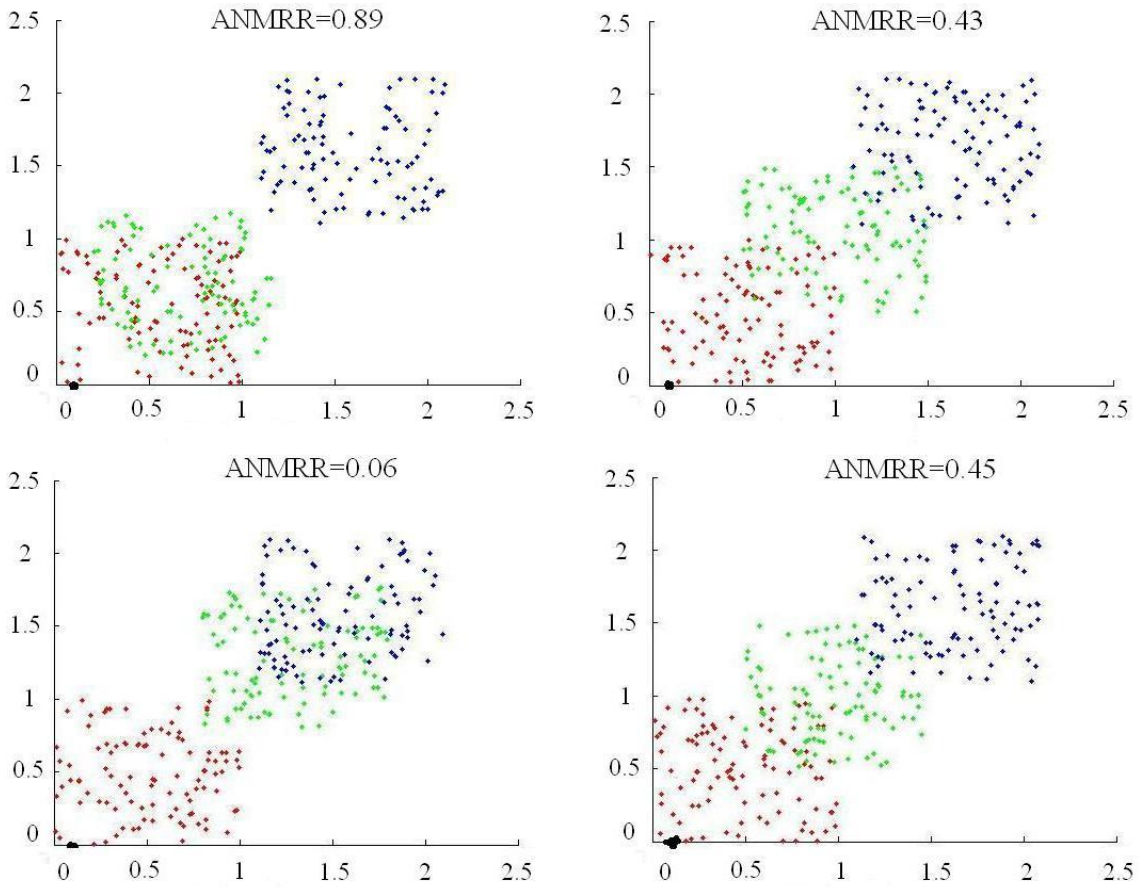


Figure 5.5. *A toy example of using ANMRR to estimate block feature distribution in the feature space.*

Upon this observation, we propose to estimate feature distribution of each block in the feature space by incorporating the labels inferred from previous classification stage. The block features are split with respect to their natural block boundaries, putting the combined GIST-LBP features of each block in the same space. The distribution of image features across different blocks in the feature space present different characteristics, which is determined by the distance between feature points. Without losing generality, euclidean distance is used as the distance measure for estimating image feature distribution for each scene block. In order to assign a ranking score for each block feature in the feature space, we use Averaged Normalized Modified Retrieval Rate (ANMRR) [86] to estimate how closely the block features from the same category are distributed. Figure 5.5 shows a toy example of using ANMRR as a ranking score to estimate clustering distribution of image features across different blocks.

ANMRR is originally proposed to evaluate the retrieval performance of image descriptors. It adopts for each query image a relevant rank threshold *K* which is defined as $\min(4\times NG(q), 2\times GTM)$, where *NG(q)* is the number of ground truth data of the query image and *GTM* is the maximum of all *NG(q)*s in question. The true ranks of retrieved images are recorded and accumulated except that the ranks of images after *K* are penalized as 1.25×*K*. The average rank of a query image *AVR(q)* is obtained by dividing the sum of ranks over *NG(q)*. After deriving *NMRR(q)* using equation (42), *ANMRR* is simply the average *NMRR* computed over all queries.

$$NMRR(q) = \frac{AVR(q) - 0.5\times(1+NG(q))}{1.25\times K - 0.5\times(1+NG(q))} \tag{42}$$

The block features ranked by ANMRR can be selected according to their distribution in the feature space. The user then can decide with respect to the number of block features to use in the scene retrieval process.

5.4. Experimental results

In this section, we report experimental results that attest to the effectiveness of the proposed GIST-LBP feature and block ranking algorithm. We compare the retrieval performance of the GIST-LBP feature with standalone GIST and LBP features in conjunction with the proposed feature using block ranking algorithm.

5.4.1. Image dataset and experimental setup

Since the UIUC 15-category dataset (shown in Figure 1.3 and Figure 1.4) include all the scene images from the SE dataset, we proceed to conduct the experiments directly on the former one. In order to simplify computation, all images are normalized to 256×256 with bilinear interpolation and we perform feature extraction on the average pixel values across R, G, B channels.

After GIST-LBP feature extraction, 100 scene features from each category are randomly selected to form the original feature matrix for deduction of the eigenvectors in PCA transformation. Subsequently, each scene feature is projected onto the eigenvectors so that the original scene features can be decorrelated. The projections on the first 10 principal component directions are selected as the feature for each scene block, resulting in the 320 dimensions GIST-LBP scene features.

5.4.2. Best 15 block features retrieval

Since higher ranked blocks have more discriminative power than the lower ranked ones due to their compact feature distribution in the feature space, we only use the image features from the top 15 blocks for similarity retrieval. With 10 principal components in each block, this set up renders the feature vector 150 dimensions in total. In this experiment, images are randomly split into training and test sets from every category, with 100 from each category as training images and the rest for testing. To evaluate the aver-

age performance of the proposed method, the splitting is done 100 times and so is the cross validation. The evaluation of retrieval performance is based solely on test images since the labels for training images are already specified. Such performance is also evaluated using ANMRR given in equation (42). We compare the performance of the proposed method with that of using only GIST, LBP and GIST-LBP combined. Figure 5.6 shows the top 3 ranked blocks marked with a rectangular square.

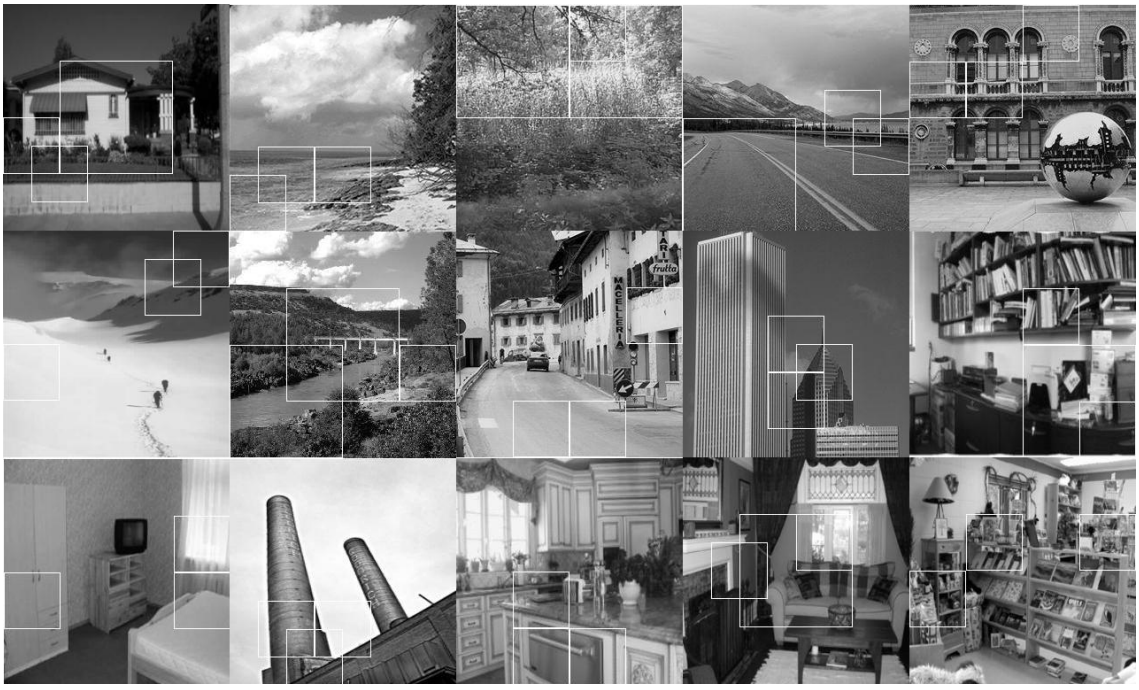


Figure 5.6. *Sample scene images from each category with the top 3 ranked blocks marked in white.*

Table 5.1 shows that the proposed algorithm garners a significant performance improvement over existing method. The proposed GIST-LBP feature increases the retrieval performance 19.18% over the original GIST descriptor and 29.07% over LBP, with only 320 dimensions in the feature space. If block ranking is used during retrieval, further performance improvement is observed. The retrieval rate of 0.3787 achieved with block ranking is a performance increase of 29.12% over using the GIST descriptor alone, 37.80% over LBP scene retrieval and 12.30% more effective than using the GIST-LBP feature.

Table 5.1*. Performance comparison between the proposed method and competing scene features.*

Method	ANMRR
GIST	0.5343
LBP	0.6088
GIST-LBP	0.4318
Block Ranking	0.3787

Another advantage of the proposed method is that it is flexible in terms of the number of feature dimensions. The user can determine the number of block features used for retrieving semantically and visually similar images. As shown in Figure 5.7, the retrieval performance of the system improves as the number of block features used increases, but performance starts to degrade when more than half of the blocks are included. The flatness in the middle of the curve suggests that increase in the number of block features does not necessarily ensure equivalent increase in retrieval performance. The user can consider the balance between retrieval accuracy and resource consumption. For instance, when computational cost is of primary concern, the user can choose to use as low as 10 blocks or so to initiate the query without sacrificing too much retrieval performance.

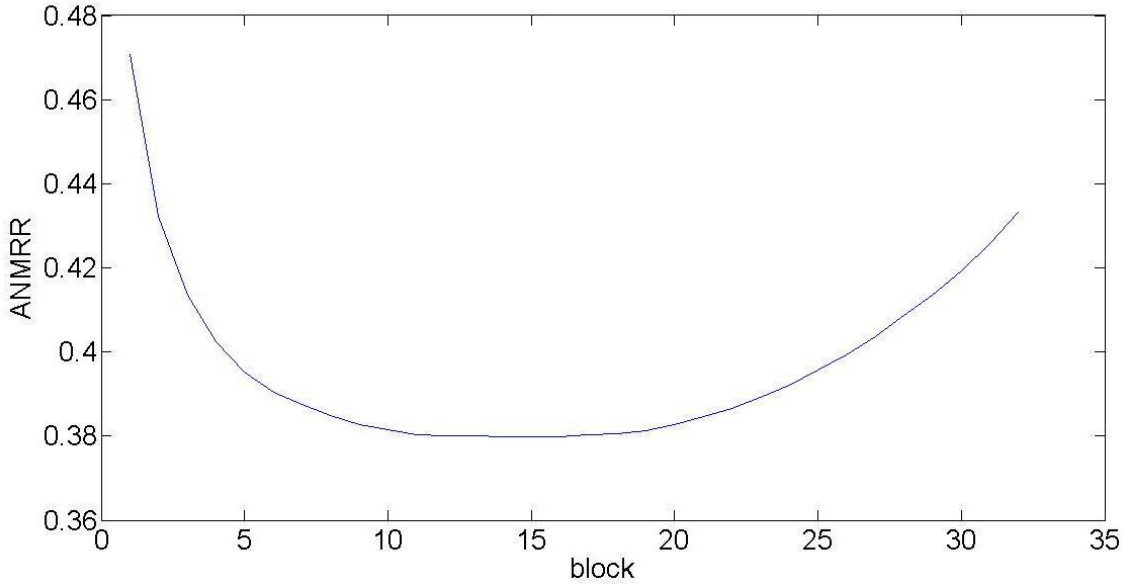


Figure 5.7. *ANMRR scores with respect to the number of block features used in the scene retrieval stage*

5.5. Conclusion

In this chapter, we have introduced a novel scene descriptor GIST-LBP. It incorporates the descriptive ability of both the original GIST feature and a variant of the LBP descriptor. Compared to other prevailing scene features, the GIST-LBP operation yields a low-dimensional feature vector. When used in conjunction with a multilevel scene representation scheme, GIST-LBP poses more discriminative capacity in terms of scene recognition accuracy and retrieval performance than using GIST or LBP feature alone, as evidenced by experimental results.

The block ranking approach introduced in this chapter can further improve scene retrieval performance by identifying the most discriminative block features of a scene. By using the top-ranked block features, those that limit the description of the visual and

semantic appearance of a scene are effectively rejected. In addition, the block feature selection process can significantly reduce the total dimensionality of a scene image, which contributes positively to retrieval efficiency. The user can further dictates the number of feature dimensions used for initiating the process in consideration to the retrieval performance and resource consumption, providing significant flexibility for scene retrieval.

6. CONCLUSION

In this thesis, we have presented several leading scene image classification and retrieval approaches in the field of computer vision. Depending on the level of recognition and specific requirements of scene categorization, different levels of abstraction with respect to scene features can be adopted for feature extraction. As the nature of Alma project suggests, semantic label inference is sufficient for this thesis work, which is the primary motivation for adopting a holistic representation on scene perception. There are a few advantages to such representation: first, by perceiving a scene as a whole, the necessity for scene region or object segmentation and recognition no longer exists; second, a holistic representation can often translates to a global description of the scene, which ensures low dimensionality of the resulting scene feature; finally, several studies have shown that human perception on scene images in rapid scene classification tasks is in agreement with this representation.

Due to this rationalization, we are in favour of global scene features over local descriptions. This preference stems not only from the requirement of this project, but also from the efficiency point of view. It is granted that local features can be quite powerful in describing intricate local structures. Nevertheless, applying local descriptors to scene images will no doubt impose high computational cost and render extremely high dimensional feature vectors. Furthermore, the adoption of local features almost ensures the application of the bag of words (BoW) representation which is characterized by a vocabulary building process, adding another layer of computational burden to the process. Even though the classification process with BoW representation of local features can be facilitated by the discriminative classification approach, the classification accuracy achieved is hardly justified.

In contrast, global scene descriptors can yield promising classification results without imposing heavy demands on computational expenses. The GIST feature is a leading scene descriptor that captures the Spatial Envelope (SE) properties of a scene, namely its degree of *naturalness, openness, roughness, expansion* and *ruggedness*. Several studies have shown that these properties carry the semantic information of a scene with roughly localized information. By extracting the spatial layout and dominant structures of a scene, the GIST descriptor encodes the SE properties in a relatively low dimensional feature vector. In the feature space, GIST features from scene images that share the same semantic category are clustered closely together. Such characteristic substantiates promising application of GIST in scene classification or retrieval tasks.

Similarly, local binary pattern (LBP) is also capable of extracting dominant textural information from scene images. The LBP descriptor proceeds to encode local textural

pattern of a symmetrically circular neighbourhood in a single integer. Such transformation conveniently encodes any textural details in a scene and effectively mitigates the artifact of illumination variations of any kind. The efficiency of LBP can be further augmented by retaining only the uniform patterns that account for the most discriminative variations. In doing so, the dimensionality of LBP histogram is significantly reduced and meanwhile trivial patterns are effectively rejected to avoid mismatch. Other study has demonstrated that the use of LBP in conjunction with multilevel scene representation can achieve state-of-the-art classification performance in a scene recognition application.

Based on the promising performance of the GIST feature, we have proposed a novel scene descriptor, ARP-GIST, which seeks to represent scene structures more precisely. By applying additional angular partitioning in each image block, more structural details can be extracted. In addition, the use of Angular Radial Partitioning (ARP) scheme allows the realization of positional invariance inside each block. With additional operations of 1-D discrete Fourier transform (DFT) on the angular bins in each block, the periodical property of DFT provides reasonable positional flexibility without tempering structural details. Experimental results illustrate the claim that the proposed descriptor outperforms the original GIST feature on two publically available datasets by reaching a balance between the level of scene structure extraction and the extent of spatial conformity.

In an attempt to incorporate the descriptive capacity of both the GIST and LBP features, we propose to combine them in a multilevel scene representation manner with the help of feature selection to reduce the dimensionality of the final feature. The resulting GIST-LBP descriptor can extract scene properties and encode them in a highly compact feature vector (of 320 dimensions). Such a low-dimensional feature can be exceptionally efficient in scene recognition and retrieval tasks since it imposes little computational cost and resource consumption onto the system. And yet, the compactness of GIST-LBP does not decrease its discriminative ability. In fact, experimental results show that GIST-LBP has yielded superior performance over either unreduced feature used alone in terms of retrieval rate. Prior to initiating scene retrieval with GIST-LBP, the feature can be further reduced by a block ranking algorithm that seeks to select the most discriminative block features in the feature space. Once the block features have been ranked according to their distribution, the user can cast the deciding vote on the number of feature dimensions for the retrieval process with respect to consideration on resource consumption. By rejecting the least discriminative block features, scene images can be retrieved more accurately, according to scene retrieval experiments.

REFERENCES

[1] Alata, O., Cariou, C., Ramananjarasoa, C., Najim, M.: Classification of Rotated and Scaled Textures Using HMHV Spectrum Estimation and the Fourier-Mellin Transform. IEEE International Conference on Image Processing, vol. 1, pp. 53-56, 1998.

[2] Amadasun, M., King, R.: Textural features corresponding to textural properties. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans - TSMCA, vol. 19, no. 5, pp. 1264-1274, 1989.

[3] Atick, J., Redlich, A.: What does the retina know about natural scenes? Neural Computation, 4:196–210, 1992.

[4] Baddeley, R.: The correlational structure of natural images and the calibration of spatial representations. Cognitive Science, vol. 21, pp. 351–372, 1997.

[5] Barnard, K., Forsyth, D.A.: Learning the semantics of words and pictures. Proceedings of the International Conference on Computer Vision, Vancouver, Canada, pp. 408–415, 2001.

[6] Barrow, H.G., Tannenbaum, J.M.: Recovering intrinsic scene characteristics from images. International Conference on Computer Vision Systems - ICVS, 1978.

[7] Baumberg, A.: Reliable feature matching across widely separated views. IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 774-781, June 2000.

[8] Bay, H., Tuytelaars, T., Van Gool, L.J.: Surf: Speeded up robust features. European Conference on Computer Vision, vol. 1, pp. 404–417, 2006.

[9] Belongie, S., Malik, J.: Shape matching and object recognition using shape contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24(4), pp. 509-522, 2002.

[10] Biederman, I.: Aspects and extension of a theory of human image understanding. In Computational Processes in Human Vision: An Interdisciplinary Perspective, 1988.

[11] Blei, D., Ng, A., Jordan, M.: Latent Dirichlet allocation. Journal of Machine Learning Research, vol. 3, pp. 993–1022, January 2003.

[12] Boser, B.E., Guyon, I.M.; Vapnik, V.N.; A training algorithm for optimal margin classifiers. 5th Annual ACM Workshop on COLT, pp. 144–152, Pittsburgh, PA, 1992.

[13] Broggi, A., Bertozzi, M., Rose, M.D., Felisa, M., Rakotomamonjy, A., Suard, F.: A Pedestrian Detector Using Histograms of Oriented Gradients and a Support Vector Machine Classificator. IEEE International Conference on Intelligent Transportation Systems, pp. 144-148, Seattle, WA, USA, 2007.

[14] Brown, M., Lowe, D.G.: Automatic Panoramic Image Stitching using Invariant Features. International Journal of Computer Vision, vol. 74(1), pp. 59-73, 2007.

[15] Brown, M., Lowe, D.G.: Recognising Panoramas. International Conference on Computer Vision, pp. 1218-1225, Nice, France, 2003.

[16] Carson, C., Belongie, S., Greenspan, H., Malik, J.: Blobworld: Image segmentation using Expectation-Maximization and its Application to Image Querying. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, pp. 1026–1038, 2002.

[17] Carson, C., Belongie, S., Greenspan, H., and Malik, J.: Region-based image querying. IEEE Workshop on Content-Based Access of Image and Video Libraries, pp. 42–49, 1997.

[18] Carson, C., Thomas, M., Belongie, S., Hellerstein, J.M., and Malik, J.: Blobworld: A system for region-based image indexing and retrieval. In Third International Conference on Visual Information Systems, Springer-Verlag, 1999.

[19] Chalechale, A., Mertins, A., Naghdy, G.: Edge image description using angular radial partitioning. IEE Proceedings - Vision, Image and Signal Processing, vol. 151(2), 2005.

[20] Chalechale, A., Naghdy, G., Mertins, A.: Sketch-based image matching using angular partitioning. IEEE Transactions on Systems, Man, and Cybernetics, vol. 35(1), pp. 28-41, 2005.

[21] Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, vol. 2, pp. 1-27, 2011.

[22] Chen, J.-L., Kundu, A.: Rotation and Gray Scale Transform Invariant Texture Identification Using Wavelet Decomposition and Hidden Markov Model. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 16, pp. 208-214, 1994.

[23] Chetverikov, D.: Experiments in the Rotation-Invariant Texture Discrimination Using Anisotropy Features. International Conference on Pattern Recognition, pp. 1071-1073, 1982.

[24] Cohen, F.S., Fan, Z., Patel, M.A.: Classification of Rotated and Scaled Texture Images Using Gaussian Markov Random Field Models. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 13, pp. 192-202, 1991.

[25] Cover, T.M., Hart, P.E.: Nearest neighbor pattern classification. IEEE Transactions on Information Theory, vol. 13 (1), pp. 21–27, 1967.

[26] Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. Workshop on Statistical Learning in Computer Vision, pp. 1–22, 2004.

[27] Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection, IEEE Conference on Computer Vision and Pattern Recognition, pp. 886-893, 2005.

[28] Dalal, N., Triggs, B., Schmid, C.: Human Detection Using Oriented Histograms of Flow and Appearance. European Conference on Computer Vision (2), pp. 428-441, 2006.

[29] Davis, L.S.: Polarograms: A New Tool for Image Texture Analysis, Pattern Recognition, vol. 13, pp. 219-223, 1981.

[30] Davis, L.S., Johns, S.A., Aggarwal, J.K.: Texture Analysis Using Generalized Cooccurrence Matrices. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 1, pp. 251-259, 1979.

[31] de Bonet, J.S., Viola, P.: Structure driven image database retrieval. Advances in Neural Information Processing, vol. 10, pp. 866–872, 1997.

[32] Deng, J., Berg, A., Li, K., Fei-Fei, L.: What does classifying more than 10,000 image categories tell us? Proceedings of the 12th European Conference of Computer Vision (ECCV), 2010.

[33] Deng, D., Zhang, J.: Combining Multiple Precision-Boosted Classifiers for Indoor-Outdoor Scene Classification. International Conference on Information Technology and Applications - ICITA, pp. 720-725, 2005.

[34] Divvala, S., Hoiem, D., Hays, J., Efros, A., Hebert, M.: An empirical study of context in object detection. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1271-1278, 2009.

[35] Douze, M., Ramisa, A., Schmid, C.: Combining attributes and fisher vectors for efficient image retrieval. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 745-752, 2011.

[36] Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. John Wiley & Sons, 2000.

[37] Duygulu, P., Barnard, K., de Freitas, J., Forsyth, D.: Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. European Conference on Computer Vision (ECCV), 2002.

[38] Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. IEEE Conference on Computer Vision and Pattern Recognition, Workshop on Generative-Model Based Vision, 2004.

[39] Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models for 101 object categories. Computer Vision and Image Understanding, 2007.

[40] Fei-Fei, L., Perona, P.: A Bayesian Hierarchical Model for Learning Natural Scene Categories. IEEE Conference on Computer Vision and Pattern Recognition, 2005.

[41] Fergus, R., Fei-Fei, L, Perona, P., Zisserman, A.: Learning Object Categories from Google's Image Search. IEEE International Conference on Computer Vision, 2005.

[42] Fergus, R., Fei-Fei, L, Perona, P., Zisserman, A.: Learning object categories from Internet image searches. IEEE Special Issue on Internet Vision, vol. 98/8, 2010.

[43] Field, D.J.: Relations between the statistics of natural images and the response properties of cortical cells. Journal of Optical Society of America, 4:2379–2394, 1987.

[44] Field, D.J.: What is the goal of sensory coding? Neural Computation, vol. 6, pp. 559–601, 1994.

[45] Fountain, S.R., Tan, T.N.: Efficient Rotation Invariant Texture Features for Content-Based Image Retrieval. Pattern Recognition, vol. 31, pp. 1725-1732, 1998.

[46] Freeman, W.T., Adelson, E.H.: The design and use of steerable filters. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 13(9), pp. 891-906, 1991.

[47] Goldberger, J., Roweis, S.T., Hinton, G.E., Salakhutdinov, R.: Neighbourhood Components Analysis. Neural Information Processing Systems - NIPS , 2004.

[48] Gorkani, M.M., Picard, R.W.: Texture orientation for sorting photos “at a glance”. International Conference on Pattern Recognition, Jerusalem, vol. I, pp. 459–464, 1994.

[49] Greenspan, H., Belongie, S., Goodman, R., Perona, P.: Rotation Invariant Texture Recognition Using a Steerable Pyramid. International Conference on Pattern Recognition, vol. 2, pp. 162-167, 1994.

[50] Guerin-Dugue, A. and Oliva, A.: Classification of scene photographs from local orientations features. Pattern Recognition Letters, vol. 21, pp. 1135–1140, 2000.

[51] Gupta, L., Pathangay, V., Patra, A., Dyana, A., Das, S.: Indoor versus Outdoor Scene Classification Using Probabilistic Neural Network. Eurasip Journal on Advances in Signal Processing, vol. 2007, pp. 1-11, 2007.

[52] Haley, G.M., Manjunath, B.S.: Rotation-Invariant Texture Classification Using a Complete Space-Frequency Model. IEEE Transactions on Image Processing, vol. 8, pp. 255-269, 1999.

[53] Harris, C., Stephens, M.: A combined corner and edge detector. Proceedings of Alvey Vision Conference, pp. 147-151, 1988.

[54] Hays, J., Efros, A.A.: im2gps: estimating geographic information from a single image. IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[55] Hays, J., Efros, A.A.: Scene completion using millions of photographs. ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH, vol. 26, 2007.

[56] Heaps, C., Handel, S.: Similarity and features of natural textures. Journal of Experimental Psychology-human Perception and Performance, vol. 25, no. 2, pp. 299-320, 1999.

[57] Henderson, J.M., Hollingworth, A.: High level scene perception. Annual Review of Psychology, vol. 50, pp. 243–271, 1999.

[58] Ikizler-Cinbis, N., Sclaroff, S.: Object, scene and actions: Combining multiple features for human action recognition. European Conference on Computer Vision, pp. 494-507, 2010.

[59] Jones, J. P. and Palmer, L.: An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. Journal of Neurophysiology, vol. 58, pp. 1233–1258, 1987.

[60] Joubert, O., Fize, D., Rousselet, G., Fabre, M.: Rapid categorization of Natural or Man-made scene contexts: different effects with amplitude and phase alterations. Journal of Vision, 2007.

[61] Kadir, T., Boukerroui, D., Brady, M.: An analysis of the scale saliency algorithm. Technical report, Robotics Research Laboratory, Department of Engineering Science, University of Oxford, 2003.

[62] Kadir, T., Brady, M.: Saliency, scale and image description. International Journal of Computer Vision, vol. 45(2), pp. 83-105, 2001.

[63] Kadir, T., Brady, M.: Scale saliency: A novel approach to salient feature and scale selection. International Conference on Visual Information Engineering, pp. 25-28, 2003.

[64] Kadir, T., Zisserman, A., Brady, M.: An affine invariant salient region detector. European Conference on Computer Vision, pp. 228-241, 2004.

[65] Kalogerakis, E., Vesselova, O., Hays, J., Efros, A.A., Hertzmann, A.: Image sequence geolocation with human travel priors. IEEE International Conference on Computer Vision, 2009.

[66] Kane, M. J., Savakis, A. E.: Bayesian Network Structure Learning and Inference in Indoor vs. Outdoor Image Classification. International Conference on Pattern Recognition - ICPR, vol. 2, pp. 479-482, 2004.

[67] Ke, Y., Sukthanka, R.: PCA-SIFT: a more distinctive representation for local image descriptors. IEEE Conference on Computer Vision and Pattern Recognition, 2004.

[68] Kim, W., Park, J., Kim, C.: A Novel Method for Efficient Indoor-Outdoor Image Classification. Journal of Signal Processing Systems - JSPS, vol. 61, no. 3, pp. 251-258, 2010.

[69] Kimchi, R.: Primacy of wholistic processing and global/local paradigm: a critical review, Psychological Bulletin, vol. 112, pp. 24–38, 1992.

[70] Kimchi, R.: Uniform connectedness and grouping in the perceptual organization of hierarchical patterns. Journal of Experimental Psychology: Human Perception and Performance, vol. 24, pp. 1105–1118, 1998.

[71] Kovac, J., Peer, P., Solina, F.: Automatic natural and man-made scene differentiation using perspective geometrical properties of the scenes. 15th International Conference on Systems, Signals and Image Processing, 2008.

[72] Lam, W.-K., Li, C.-K.: Rotated Texture Classification by Improved Iterative Morphological Decomposition. IEE Proceedings on Vision, Image, and Signal Processing, vol. 144, pp. 171- 179, 1997.

[73] Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. IEEE Conference on Computer Vision and Pattern Recognition, vol. II, pp. 2169–2178, 2006.

[74] Leung, M.M., Peterson, A.M.: Scale and Rotation Invariant Texture Classification, 26th Asilomar Conference on Signals, Systems, and Computers. vol. 1, pp. 461-465, 1992.

[75] Li, L.-J., Socher, R., Fei-Fei, L.: Towards Total Scene Understanding: Classification, Annotation and Segmentation in an Automatic Framework. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009. (Oral)

[76] Lipson, P., Grimson, E., Sinha, P.: Configuration based scene classification and image indexing. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Puerto Rico, pp. 1007–1013, 1997.

[77] Liu, W., Kiranyaz, S., Gabbouj, M.: Efficient Scene Retrieval Using GIST-LBP Feature with Block Ranking. Submitted to European Conference on Computer Vision, 2012.

[78] Liu, W., Kiranyaz, S., Gabbouj, M.: Robust Scene Classification By Gist With Angular Radial Partitioning. International Symposium on Communications, Control and Signal Processing-ISCCSP, 2012

[79] Loschky, L. C., Larson A.M.: Localized information is necessary for scene categorization, including the Natural/Man-made distinction. Journal of Vision, vol. 8, no. 1, pp. 4-4, 2008.

[80] Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, vol. 60 (2), pp. 91-110, 2004.

[81] Lowe, D.G.: Local feature view clustering for 3D object recognition. IEEE Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii, pp. 682-688, 2001.

[82] Lowe, D.G.: Object recognition from local scale-invariant features. International Conference on Computer Vision, Corfu, Greece, pp. 1150-1157, 1999.

[83] Luo, J., Savakis, A. E.: Indoor vs outdoor classification of consumer photographs using low-level and semantic features. Image Processing, IEEE International Conference - ICIP, pp. 745-748, 2001.

[84] Madiraju, S.V.R., Liu, C.C., Rotation Invariant Texture Classification Using Covariance. International Conference on Image Processing. vol. 2, pp. 655-659, 1994.

[85] Manian, V., Vasquez, R.: Scaled and Rotated Texture Classification Using a Class of Basis Functions. Pattern Recognition, vol. 31, pp. 1937-1948, 1998.

[86] Manjunath, B.S., Salembier, P., Sikora, T.: Introduction to mpeg-7. John Wiley and Sons, Ltd., San Francisco, 2002.

[87] Marr, D.: Vision. WH Freeman: San Francisco, CA, 1982.

[88] Marr, D., Hildreth, E.C.: Theory of edge detection. Proceedings of the Royal Society of London B, vol. 207, pp. 187–217, 1980.

[89] Matas, J., Chum, O., Martin, U., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. British Machine Vision Conference, vol. I, pp. 384-393, 2002.

[90] Mikolajczyk, K., Schmid, C.: An affine invariant interest point detector. European Conference on Computer Vision, vol. I, pp. 128-142, 2002.

[91] Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27(10), pp. 1615-1630, 2005.

[92] Mikolajczyk, K., Schmid, C.: Indexing based on scale invariant interest points. IEEE International Conference on Computer Vision, vol. I, pp. 525-531, 2001.

[93] Mitchell, T.: Machine Learning, McGraw Hill. ISBN 0-07-042807-7. pp. 2, 1997.

[94] Nanayakkara, N.D., Samarabandu, J., Fenster, A.: Prostate segmentation by feature enhancement using domain knowledge and adaptive region based operations. Physics in Medicine and Biology, vol. 51(7), pp. 1831-1848, 2006.

[95] Navon, D.: Forest before trees: the precedence of global features in visual perception. Cognitive Psychology, vol. 9, pp. 353–383, 1977.

[96] Ojala, T., Pietikainen, M., Harwood, D.: A Comparative Study of Texture Measures with Classification Based on Feature Distributions. Pattern Recognition, vol. 29, pp. 51-59, 1996.

[97] Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on pattern analysis and machine intelligence, vol. 24(7), pp. 971–987, 2002.

[98] Oliva, A., Schyns, P. G.: Diagnostic color blobs mediate scene recognition. Cognitive Psychology, vol. 41, pp. 176–210, 2000.

[99] Oliva, A., Torralba, A.: Building the gist of a scene: the role of global image features in recognition. Visual Perception, Progress in Brain Research, vol. 155, 2006.

[100]Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the Spatial Envelope. International Journal of Computer Vision, vol. 42, pp. 145–175, 2001.

[101]Oliva, A., Torralba, A., Guerin-Dugue, A., Herault, J.: Global semantic classification using power spectrum templates. In Proceedings of The Challenge of Image Retrieval, Electronic Workshops in Computing series, Springer-Verlag: Newcastle, 1999.

[102]Payne, A., Singh, S.: Indoor vs. outdoor scene classification in digital photographs. Pattern Recognition - PR, vol. 38, no. 10, pp. 1533-1545, 2005.

[103]Pietikainen, M., Ojala, T., Xu, Z.: Rotation-Invariant Texture Classification Using Feature Distributions. Pattern Recognition, vol. 33, pp. 43-52, 2000.

[104]Porat, M., Zeevi, Y.: Localized Texture Processing in Vision: Analysis and Synthesis in the Gaborian Space. IEEE Transactions on Biomedical Engineering, vol. 36, pp. 115-129, 1989.

[105]Potter, M.C.: Meaning in visual search. Science, 187:965–966, 1975.

[106]Quattoni, A., Torralba, A.: Recognizing Indoor Scenes. IEEE Conference on Computer Vision and Pattern Recognition, pp. 413–420, 2009.

[107]Randen T., Husoy, J.H.: Filtering for Texture Classification: A Comparative Study. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 21, pp. 291-310, 1999.

[108]Rao, A. R., Lohse, G. L.: Identifying High Level Features of Texture Perception. Graphical Models/graphical Models and Image Processing/computer Vision, Graphics, and Image Processing - CVGIP, vol. 55, no. 3, pp. 218-233, 1993.

[109]Rensink, R. A., O'Regan, J. K., Clark, J. J.: To see or not to see: the need for attention to perceive changes in scenes. Psychological Science, vol. 8, pp. 368–373, 1997.

[110]Rogowitz, B., Frese, T., Smith, J., Bouman, Kalin, E.: Perceptual image similarity experiments. Human Vision and Electronic Imaging, SPIE, vol. 3299, pp. 576–590, 1998.

[111]Rosch, E., Mervis, C.B.: Family resemblances: Studies in the internal structure of categories. Cognitive Psychology, vol. 7, pp. 573–605, 1975.

[112]Roth, P.M., Winter, M.: Survey of Appearance-Based Methods for Object Recognition. Institute for Computer Graphics and Vision, Graz University of Technology, Austria, 2008.

[113]Russell, B.C., Torralba, A., Liu, C., Fergus, R., Freeman, W.T.: Object recognition by scene alignment. Advances in Neural Information Processing Systems, 2007.

[114]Samulon, A.: Separation of man-made and natural patterns in high-altitude imagery of agricultural areas. IEEE Transactions on Circuits and Systems, vol. 22, no. 5, pp. 450-463, 1975.

[115]Sanocki, T., Reynolds, S.: Does figural goodness influence the processing and representation of spatial layout. Investigative Ophthalmology and Visual Science, vol. 41, pp. 723, 2000.

[116]Schmid, C., Mohr, R.: Local grayvalue invariants for image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, pp. 530-535, 1997.

[117]Schyns, P.G. and Oliva, A.: From blobs to boundary edges: evidence for time- and spatial-scale-dependent scene recognition. Psychol. Sci., vol. 5, pp. 195–200, 1994.

[118]Serranol, N., Savakis A. E., Luo, J.: A Computationally Efficient Approach to Indoor/Outdoor Scene Classification. International Conference on Pattern Recognition - ICPR, vol. 4, pp. 146-149 vol.4, 2002.

[119]Shashua, A., Ullman, S.: Structural saliency: the detection of globally salient structures using a locally connected network. Proceedings of the 2nd International Conference on Computer Vision, Tempa, FL, pp. 321–327, 1988.

[120]Siagian, C., Itti, L.: Rapid biologically-inspired scene classification using features shared with visual attention. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, pp. 300-312, 2007.

[121]Silberman, N., Fergus, R.: Indoor Scene Segmentation using a Structured Light Sensor. 3DRR Workshop, International Conference on Computer Vision, 2011.

[122]Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering Objects and Their Location in Images. International Conference on Computer Vision, October, 2005.

[123]Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. International Conference on Computer Vision, vol. 2, pp. 1470-1477, 2003.

[124]Suard, F., Rakotomamonjy, A., Bensrhair, A., Broggi, A.: Pedestrian Detection using Infrared images and Histograms of Oriented Gradients. IEEE Intelligent Vehicles Symposium, pp. 206-212, Tokyo, Japan, 2006.

[125]Sugase, Y., Yamane, S., Ueno, S., Kawano, K.: Global and fine information coded by single neurons in the temporal visual cortex. Nature, vol. 400, pp. 869–873, 1999.

[126]Swain, M.J., Ballard, D.H.: Color indexing. International Journal of Computer Vision - IJCV , vol. 7(1), pp. 11-32, 1991.

[127]Switkes, E., Mayer, M.J., Sloan, J.A.: Spatial frequency analysis of the visual environment: anisotropy and the carpentered environment hypothesis. Vision Research, vol. 18, pp.1393–1399, 1978.

[128]Szummer, M., Picard, R.W.: Indoor-Outdoor Image Classification. IEEE International Workshop on Content-Based Access of Image and Video Database, pp. 42 - 51, 1998.

[129]Tamura, H., Mori, S., and Yamawaki, T.: Textural features corresponding to visual perception. IEEE Transactions on Systems, Man and Cybernetics, vol. 8, pp. 460–473, 1978.

[130]Tao, Li., Kim, Y.: An efficient neural network based indoor-outdoor scene classification algorithm. International Conference on Consumer Electronics (ICCE), pp. 317 - 318, 2010.

[131]Thorpe, S., Fize, D. and Marlot, C: Speed of processing in the human visual system. Nature, vol. 381, pp. 520–522, 1996.

[132]Torralba, A., Fergus, R., Weiss, Y.: Small codes and large image databases for recognition. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-8, 2008.

[133]Torralba, A., Oliva, A.: Depth estimation from image structure. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, pp. 1226–1238, 2002.

[134]Torralba, A., Oliva, A.: Scene organization using discriminant structural templates. International Conference on Computer Vision, pp. 1253–1258, 1999.

[135]Traherne, M., Singh, S.: An Integrated Approach to Automatic Indoor Outdoor Scene Classification in Digital Images. Intelligent Data Engineering and Automated Learning - IDEAL, pp. 511-516, 2004.

[136]Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of Cognitive Neuroscience, vol. 3 (1), pp. 71–86, 1991.

[137]Turk, M., Pentland, A.: Face recognition using eigenfaces. IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–591, 1991.

[138]Tuytelaars T., Van Gool, L.J.: Content-based image retrieval based on local affinely invariant regions. International Conference on Visual Information and Information Systems, pp. 493-500, 1999.

[139]Tuytelaars T., Van Gool, L.J.: Matching widely separated views based on affine invariant regions. International Journal of Computer Vision, vol. 1(59), pp. 61-85, 2004.

[140]Tuytelaars, T., Van Gool, L.J.: Wide baseline stereo matching based on local, affinely invariant regions. British Machine Vision Conference, pp. 412-422, 2000.

[141]Tversky, B. and Hemenway, K: Categories of environmental scenes. Cognitive Psychology, vol. 15, pp. 121–149, 1983.

[142]Vailaya, A., Figueiredo, M., Jain, A., and Zhang, H.J.: Content-based hierarchical classification of vacation images. In Proceedings of the International Conference on Multimedia, Computing and Systems, 1999.

[143]Vailaya, A., Jain, A., Zhang, H. J.: On image classification: city images vs. landscapes. Pattern Recognition, vol. 31, pp. 1921–1935, 1998.

[144]Van Gool, L.J., Moons, T., Ungureanu, D.: Affine photometric invariants for planar intensity patterns. European Conference on Computer Vision, vol. 1, pp. 642-651, 1996.

[145]Van der Schaaf, A., Van Hateren, J.H.: Modeling of the power spectra of natural images: Statistics and information. Vision Research, vol. 36, pp. 2759–2770, 1996.

[146]Wang, X., Han, T.X., Yan, S.: An hog-lbp human detector with partial occlusion handling. International Conference on Computer Vision, pp. 32-39, 2009.

[147]Wu, J., Rehg, J.M.: CENTRIST: A visual descriptor for scene categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33(8), pp. 1489-1501, 2011.

[148]Wu, W.-R., Wei, S.-C.: Rotation and Gray-Scale Transform-Invariant Texture Classification Using Spiral Resampling, Subband Decomposition and Hidden Markov Model. IEEE Transactions on Image Processing, vol. 5, pp. 1423-1434, 1996.

[149]Wu, Y., Yoshida, Y.: An Efficient Method for Rotation and Scaling Invariant Texture Classification. IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 4, pp. 2519-2522, 1995.

[150]You, J., Cohen, H.A.: Classification and Segmentation of Rotated and Scaled Textured Images Using Texture `Tuned' Masks. Pattern Recognition, vol. 26, pp. 245-258, 1993.

[151]Zabih, R., Woodfill, J.: Non-parametric local transforms for computing visual correspondence. European Conference on Computer Vision, vol. 2, pp. 151–158, 1994.

[152]Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: A comprehensive study. International Journal of Computer Vision, vol. 73 (2), pp. 213–238, 2007.

APPENDIX

- **From RGB to YUV:**

Y = 0.299R + 0.587G + 0.114B
U = -0.147R - 0.289G + 0.436B
V = 0.615R - 0.515G - 0.100B

- **From YUV to RGB:**

R = Y + 1.140V
G = Y - 0.395U - 0.581V
B = Y + 2.032U

- **From RGB to XYZ:**

X = 0.412453R + 0.357580G + 0.180423B
Y = 0.212671R + 0.715160G + 0.072169B
Z = 0.019334R + 0.119193G + 0.950227B

- **From XYZ to RGB:**

R = 3.240479X - 1.537150Y - 0.498535Z
G =-0.969256X + 1.875992Y + 0.041556Z
B = 0.055648X - 0.204043Y + 1.057311Z

- **From XYZ to CIELAB:**

$$L = \begin{cases} 116(Y/Yn)^{1/3} - 16, & \text{when } Y/Yn > 0.008856 \\ 903.3(Y/Yn)^{1/3}, & \text{when } Y/Yn \le 0.008856 \end{cases}$$

a = 500[f (X/Xn) - f (Y/Yn)]
b = 200[f (Y/Yn) - f (Z/Zn)]

$$f(t) = \begin{cases} t^{1/3} - 16, & \textit{when } t > 0.008856 \\ 7.787t + 16/116, & \text{when } t \le 0.008856 \end{cases}$$

Yn = 1.0, Xn = 0.950455 and Zn = 1.088753 denote the reference white

- **From CIELAB to XYZ:**

$Y = Yn * P^3$
$X = Xn * (P + a/500)^3$
$Z = Zn * (P - b/200)^3$

$P = (L +16)/116$