This is to certify that the
dissertation entitled

SPARSE REPRESENTATIONS FOR IMAGE
CLASSIFICATION

presented by

Ke Huang

has been accepted towards fulfillment
of the requirements for the

Ph.D. degree in Electrical and Computer Engineering

Aviyente
Major Professor's Signature

05/03/07
Date

PLACE IN RETURN BOX to remove this checkout from your record.
TO AVOID FINES return on or before date due.
MAY BE RECALLED with earlier due date if requested.

DATE DUE	DATE DUE	DATE DUE
SEP 14 2009 05 26 09		OCT 19 2011 OCT 25 2011
10 20 11 APR 23 2011		

SPARSE REPRESENTATIONS FOR IMAGE CLASSIFICATION

By

Ke Huang

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Electrical and Computer Engineering

2007

ABSTRACT

SPARSE REPRESENTATIONS FOR IMAGE CLASSIFICATION

By

Ke Huang

In recent years, filter bank based approaches such as wavelet and wavelet packet transforms have been used extensively in image classification. One key issue in wavelet based classification methods is how to choose the best set of features (subbands). In the existing methods, each subband is evaluated separately or only the children subbands are compared with their parent subband for selection. In this thesis, we show that this subband selection method does not consider the dependence from different subbands in the selection process. In order to address this issue, we propose subband selection methods that take the dependence into account. We offer a theoretical and experimental analysis of dependence among features among different subbands. Based on this analysis, *mutual information based subband selection* (MISS) algorithm is proposed for subband selection based on feature dependence. The MISS algorithm is further improved by the *subband grouping and selection* (SGS) algorithm which combines the dependence between subbands and the evaluation score of each subband. All of these methods result in a compact set of features for efficient image classification.

The development of efficient subband selection methods for image classification motivates us to consider the more general problem of sparse representation of images for classification. We propose a new approach in the framework of the sparse representations by combining the reconstruction error, classification power and sparseness in a single cost function.

The formulation of the proposed sparse representation for image classification method is further improved by using the large margin method for the measure of discrimination. Based on this new and improved formulation, we can model the robust and sparse feature extraction with an optimization problem that can be solved by iterative quadratic programming. In order to reduce the computational complexity required for the iterative quadratic programming, we propose decomposing the robust and sparse feature extraction into two steps, with the first step being sparse reconstruction and the second step being sparse feature selection and dimension reduction. For the second step, we propose a new method called large margin dimension reduction (LMDR). LMDR integrates the idea of L1-norm support vector machine (SVM) and distance metric learning for obtaining feature representation in a low dimensional space.

Finally, to measure the goodness of features obtained by different methods, a mutual information based feature evaluation criterion is proposed. The proposed measure is independent of distance metric and classifier. Computation of mutual information in a high dimensional space is addressed by using the uncorrelated linear discrimination analysis (ULDA). The proposed computational model effectively reduces the computational complexity of computing mutual information for feature evaluation.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor, Prof. Selin Aviyente, for her guidance, encouragement, and support in every stage of my graduate study. Her knowledge, kindness, patience, and vision have provided me with lifetime benefits.

I am grateful to Profs. John Deller, Rong Jin and Hyder Radha in my thesis committee for their valuable comments and suggestions on the thesis drafts, as well as for the experience as a student with these three outstanding teachers. I would also like to thank many faculty members of MSU who were the instructors for the courses I took. The course works have greatly enriched my knowledge and provided the background and foundations for my thesis research.

I want to take this opportunity to thank Dr. Michelle Yan, Hong Shen, Sam Zheng and Jason Tyan for their support before, during and after my internship at Siemens Corporate Research. The amazing internship provided me very valuable experience in both research and development, and aligned my expectation for a future job. Friends made at Siemens Corporate Research are another valuable treasure I got.

My memory of graduate studies at MSU can never be complete without mentioning my fellow graduate students at MSU. The survivor story would be much tougher if without the accompanying by the students coming to MSU in the same year as me. Special thanks to Rensheng Sun couple for their countless help, to Yunhao Liu, Chen Wang and Zhenyun Zhuang for providing me encouragements and accommodation in the coldest winter.

This thesis draws a period for my more than 20 years of education in schools. In addition to the thanks to my teachers and classmates in the old days, I must mention my parents, without whose love, nurturing, and support I could never accomplish all these. I dedicate this thesis to my parents. Finally, thanks to my wife, Gangli, for her love and patience.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

Image classification is an important research problem that has applications including computer vision, medical imaging and biometrics. A standard image classification system consists of feature extraction, feature selection, and classifier design [1–8]. The research in this dissertation focuses on the feature extraction, selection and evaluation aspects of an image classification system in the framework of transform (filtering) based feature extraction methods. The major objective is to obtain a compact and robust set of features for a given image set using multi-resolution transforms such as the wavelets, wavelet packets, directional wavelets and Gabor functions.

Image classification usually relies on features extracted from transform or filtering operations [9]. The image classification discussed in this dissertation is also based on the filtering features. The general paradigm for feature extraction is to filter images with a set of filters and to compute energy values or other statistical values from the filtered images as features. The filters can be either data dependent or data independent. Data dependent filters are derived from a training data set. Principal component analysis (PCA) [10], linear discriminative analysis (LDA) [10,11] and K-SVD [12] are examples of data dependent filters. Wavelets [13], wavelet packet [13], directional wavelets [14] and Gabor filters [3] are example of data independent filters.

For images with different texture structures, multi-scale data independent filtering, such as wavelets, has been applied to image classification [1,2,7]. The multi-scale structure of the wavelet filters is a good fit for the multi-scale structure that often appears in texture images. In this sense, the filtering results with the wavelet filters can discriminatively describe the texture structure for classification. In this dissertation, feature extraction and selection with data independent filtering using wavelets and wavelet packet are studied. Some of the techniques discussed in the dissertation can also be applied to features obtained with data dependent filtering.

The first part of the proposed work focuses on the extraction of a set of features from wavelet and wavelet packet transforms. The proposed approach first quantifies the dependence between wavelet subbands and then exploits this property to select a compact set of features that are discriminative for the given image set. This method is then further improved by incorporating the individual discrimination power provided by each subband into the selection process.

The promising results obtained with the proposed feature selection methods motivate the formulation of a more general feature selection problem for image classification. This more general feature selection problem can be stated as "How can we choose the 'best' set of multi-resolution transforms for discriminating between different image classes?" The 'best' set of transforms is defined by the compactness (sparseness) of the selected feature set, the robustness of the features in noisy environments and the accuracy of the classification. Combining all of these requirements, we pose this more general problem using sparse representations. Sparse representations aim at finding a sparse and close approximation to a given signal using a large collection of functions, called a dictionary [15–19]. In the proposed work, this framework is modified to address the question of image classification by defining a cost function that incorporates the discrimination and reconstruction abilities of the elements in the dictionary, as well as the sparseness of the selected feature vector. As part of the proposed work, the different aspects of this problem will be investigated including the selection of the dictionary elements, different cost functions to quantify the discrimination power of the selected features and the tradeoff between reconstruction power and discrimination power.

In the third part, the formulation of the proposed sparse representation for image classification method is further improved by using the large margin method for the measure of discrimination. Based on this new and improved formulation, we can model the robust and sparse feature extraction with an optimization problem that can be solved by iterative quadratic programming. In order to reduce the computational complexity required for the iterative quadratic programming, we propose decomposing the robust and sparse feature extraction into two steps, with the first step being sparse reconstruction and the second step being sparse feature selection and dimension reduction. For the second step, we propose a new method called large margin dimension reduction (LMDR). LMDR integrates the idea of L1-norm support vector machine (SVM) and distance metric learning for obtaining feature representation in a low dimensional space.

In the fourth part of this dissertation, we propose a mutual information based feature evaluation criterion and a corresponding computational model, such that features obtained from different selection methods can be quantitatively evaluated. Traditional feature evaluation based on empirical study is subjected to the selection of distance metric and classifier. The proposed measure is independent of distance metric and classifier, and reflects more objectively the goodness of a feature representation. Computation of mutual information in a high dimensional space is usually involved in the model. To deal with this problem, we propose computing mutual information with the uncorrelated linear discrimination analysis (ULDA). The proposed computational model effectively reduces the computational complexity of computing mutual information for feature evaluation.

1.1 Overview of Contributions

The contributions of the thesis can be divided into four parts: subband selection with the wavelet packet analysis, a general framework of sparse representation for image classification, sparse representation for image classification with the large margin method and large margin dimension reduction and classifier independent feature evaluation with mutual information.

In wavelet packet based signal classification, sparse representation is realized through subband selection, i.e., selection of a subset of subbands for signal representation. The major contributions of this work can be summarized as follows:

1. The dependence among features extracted from different subbands is quantified.

2. The dependence between subbands is incorporated into the selection process to improve classification accuracy.

3. The selected features are less redundant than the existing methods, thus reducing the dimensionality of the feature vector. The reduced dimensionality of the feature vectors increases the robustness of image classification in noisy environments.

4. The proposed feature selection method is easily generalizable to other multiresolution and directional transforms.

In the second part of the proposed research, a more general framework of sparse representations for signal classification is introduced with the following contributions:

1. The current framework for the general sparse representations is modified by incorporating both the reconstruction error and the discrimination power into the optimization problem. This flexibility improves the robustness of classification in noise.

2. The proposed framework will allow the design of adaptive basis functions that can best represent and discriminate between the given signals.

3. The proposed framework will also have significant impact on a variety of classification problems such as texture classification and object discrimination.

In the third part of the proposed research, an improved formulation of sparse representations for signal classification is introduced and a new dimension reduction method is introduced with the following contributions:

1. The proposed framework of incorporating both the reconstruction error and the discrimination power into the feature extraction process is reformulated based on the large margin method. The new formulation makes the feature extraction problem solvable with the iterative quadratic programming method.

2. A more computationally efficient algorithm, large margin dimension reduction, is proposed for extracting discriminative features in the low dimensional space.

In the fourth part of the proposed research, a mutual information based measure is proposed for feature evaluation. A computational model is also proposed for practically computing mutual information in this case. The contributions can be summarized as:

1. Feature evaluation based on empirical study is first Analyzed. Based on this analysis, a mutual information based feature evaluation method that is independent of distance metric and classifier is proposed.

2. An uncorrelated linear discrimination analysis based method is proposed for computing mutual information in the high dimensional space. Compared with the existing mutual information computational model, the computational model retains the accuracy of computation with reasonable complexity.

CHAPTER 2

Wavelet Subband Selection for Image Classification

2.1 Background on Wavelet and Wavelet Packet Transforms

Wavelet decomposition and its extension, wavelet packet decomposition have gained popular applications in the field of signal/image processing and classification. For example, speech [20], EEG [21], and texture images [2,22] have been successfully classified by statistical models built in the transform domain obtained from the wavelet decomposition. Wavelet transforms enable the decomposition of the image into different frequency subbands, similar to the way the human visual system operates. This property makes it especially suitable for the segmentation and classification of texture images [1,2,7,22–27]. For the purpose of texture classification, appropriate features need to be extracted to obtain a discriminative representation as possible in the transform domain. A widely used wavelet feature is the energy of each wavelet subband [1,2,7,20,21,23,27,28]. This idea of extracting energy features from filtered images can be traced back to [29], where a bank of band-pass filters were used for image analysis. A more recent survey on filtering based methods for texture classification can be found in [9]. In early research, such as in [2,28], features from all wavelet subbands are used for texture classification. However, it is common knowledge in the area of pattern recognition that proper feature selection is likely to improve the classification accuracy with fewer number of features [10]. At the same time, the overcomplete structure of the wavelet packet transform motivates the selection of the wavelet features for classification. For the widely used energy feature, since one energy value is extracted from each subband, wavelet feature selection is thus equivalent to selecting a set of subbands for texture decomposition. Therefore, "wavelet feature selection" and "wavelet subband selection" are interchangeably used in this chapter. Note that different from the general subband selection method that aims at signal reconstruction, such as the entropy based subband selection method proposed in [30], wavelet feature selection does not require the selected subbands to reconstruct the texture image.

At this point, it is important to make a distinction between wavelet feature selection and the general feature selection discussed in pattern recognition [10, 31]. Both selection methods aim at obtaining a compact representation of the texture for classification. However, the two techniques are not exactly the same. First, for general feature selection methods, the explicit knowledge of the feature extraction process may not always be available. The input to the general feature selection process is usually a vector of values representing the different features without any *a priori* information about how these features were obtained. On the other hand, the wavelet feature selection methods can take advantage of the tree structure of the wavelet decomposition for the selection process. For example, in this chapter, the wavelet packet decomposition structure is analyzed and used in determining the optimal set of subbands. Second, some selection methods based on component analysis, such as the principle component analysis (PCA), independent component analysis (ICA), and linear discriminant analysis (LDA), usually require that all of the original feature components are available at both the training and testing stages for projecting features into a selected subspace. In the setting of wavelet packet decomposition, this means that a full decomposition is required at both the training and testing stages. On the other hand, for the wavelet feature selection method, an texture is only needed be decomposed into the wavelet subbands selected by the wavelet feature selection method during the training stage. In this sense, wavelet feature selection selection also reduces the computational complexity of feature extraction.

One of the most well-known subband selection algorithms is based on the entropy cost function [30], where the entropy of the coefficients at the children nodes is compared to the entropy at the parent node. In this method, the "best" wavelet packet tree is chosen to minimize the entropy of the representation coefficients. As such, the main goal of this method is signal reconstruction and not classification. For this reason, subband selection methods that particularly aim at achieving compact signal representations for classification have been proposed. For example, in [1], the features are only extracted from subbands with high energy values higher than a predetermined threshold value. The energy distribution over these subbands with high energy is then used as features for classification. In [32], each subband is evaluated based on the discrimination power of the extracted energy value, and the subbands with high discrimination power are selected for subsequent classification. A similar evaluation method is also employed in [21], where the decomposition tree is pruned by comparing the discrimination score of the parent and its children nodes. In [7], a neuro-fuzzy method is used for subband selection to obtain a compact representation.

One important and effective principle guiding the feature selection process is to exploit the dependence relation among different feature components [22, 25, 33–37]. For wavelet feature selection, this principle indicates that the dependence between different subbands should be investigated and utilized. However, in the existing research on wavelet feature selection [1, 7, 21, 30, 32], either each subband is evaluated separately, or only the parent subband and children subbands are compared based on a pre-determined criterion and the wavelet subband selection is based on the evaluation. When each subband is evaluated separately, it is implicitly assumed that the features from different subbands are independent. When selection is based on comparing parent and children subbands, only the dependence between parent and children subbands is considered. Usually the assumption of independence between subbands does not hold, thus degrading the classification accuracy.

The dependence between wavelet coefficients that have the "parent-child" or "sibling" relationship has been successfully measured by mutual information [38], or modelled by hidden Markov Model (HMM) [39, 40]. It should be noted that the dependence of energy features among subbands is different from the dependence of wavelet coefficients among subbands. The statistical properties of wavelet coefficients of an image have also been analyzed and modelled [41,42]. It has been shown that incorporating the dependence among wavelet coefficients improves the image compression efficiency [43] and image denoising performance. In this chapter, we are interested in the dependence of the extracted features, not the wavelet coefficients. Since the extracted features are a function of all the coefficients in a subband, the dependence between features is more complicated than the dependence between individual wavelet coefficients.

The dependence among subbands can also be interpreted as the amount of redundancy among subbands. In classification, one can take advantage of this redundancy in features by two different approaches. In the first approach, the structure of redundancy among the transform coefficients is modelled with a parametric model and the parameters of the model are extracted as features for classification. Examples of this first approach can be seen in [22,25,35], where the dependency among wavelet subband coefficients that have "parent-children" relation is modelled with the HMM and used for texture classification. In the second approach, first features are extracted from the transform coefficients and then the redundant features are removed, since the removal does not reduce information useful for classification. Algorithms presented in [33,34,36,37] belong to this second type of approaches. In [34], the "minimax entropy principle" is proposed to require that a newly added feature should be "very different" from the existing features and the degree of difference is measured by the changes in the entropy caused by the inclusion of a new feature. To illustrate the difference between the two types of approaches, consider the following simple example. Let $X = [X_1, X_2]$ be a 2-dimensional binary feature vector describing a data sample. Assume that as the prior knowledge, it is known that $X_2 \equiv \bar{X}_1$. In the first type of dependency analysis, this dependence feature can be extracted as "the second component is the complement of the first component". For the second method, the component X_2 is simply removed since the inclusion of X_2 does not provide any extra information for classification. In this chapter, we adopt the second type of approach, i.e., identifying the redundant features and then removing them. In this sense, the models developed in [38–40] are not suitable for the problem of wavelet feature selection for texture classification studied in this chapter.

In this chapter, the dependence between energy values extracted from different subbands are analyzed and incorporated for wavelet feature selection. Different from the HMM model that only models dependence between "parent-children" subbands [22, 25, 35], in this chapter the dependence between features from all subbands are taken into account in the selection process. The analysis and the proposed algorithms can be easily extended for other features, such as entropy, kurtosis, etc. Applying the same analysis to other filter bank based methods, such as Gabor decomposition, is also straightforward. The dependence between energy features extracted from different subbands is theoretically analyzed and verified by the simulation results. Based on this analysis, Mutual Information based Subband Selection (MISS) algorithm is proposed for subband selection based on feature dependence. Experimental results show that dependence is an effective criterion for selecting subbands to obtain a compact feature representation. In order to further combine the dependence between the subbands and the evaluation score of individual subbands, Subband Grouping and Selection (SGS) algorithm is proposed to incorporate both factors into the subband selection process.

The contributions of this chapter include demonstrating the dependence among extracted features; demonstrating that dependence can be used for effective subband selection; and combining dependence between features from different subbands and the evaluation score of each individual feature for better subband selection. Although classification of texture images are used to show the effectiveness of the proposed methods, the focus of this chapter is not to propose a new feature extraction method for texture classification, but to identify the dependence among different components of wavelet features and propose new subband selection methods for texture classification. This analysis and the proposed methods can be similarly applied to the classification of other signals (like EEG, speech) using the wavelet packet transform.

The organization for the rest of this chapter is as follows. Section 2.2 briefly reviews the standard 2D wavelet packet decomposition and feature extraction. In Section 2.3, the dependence between energy features from different subbands is analyzed. Due to the lack of an accurate statistical model for natural images, simulation results for covariance between energy values from different subbands are presented. Based on the analysis provided in Section 2.3, Section 2.4 proposes an algorithm for subband selection based on dependence. Section 2.5 proposes a second algorithm that combines the dependence between subbands and the evaluation score of each individual subband. Experimental results and related discussions are given in Section 5.4. Section 5.5 concludes the chapter with suggestions for possible future research.

2.2 Wavelet Packet - Feature Extraction

As an extension of the standard wavelets, wavelet packets represent a generalization of the multi-resolution analysis and use the entire family of subband decompositions to generate an over-complete representation of signals. In 2D discrete wavelet packet transform (2D-DWPT), an image is decomposed into one approximation and three detail images. The approximation and the detail images are then decomposed into a second-level approximation and detail images, and the process is repeated. The standard 2D-DWPT can be implemented with a low-pass filter h and a high-pass filter g [13]. The 2D-DWPT of an $N \times M$ discrete image A up to level *P+1* ($P \leq \min(\log_2 N, \log_2 M)$), is recursively defined in terms of the coefficients at level p as follows:

$$C^{p+1}_{4k,(i,j)} = \sum_m \sum_n h(m)h(n)C^p_{k,(m+2i,n+2j)}, \tag{2.1}$$

$$C^{p+1}_{4k+1,(i,j)} = \sum_m \sum_n h(m)g(n)C^p_{k,(m+2i,n+2j)}, \tag{2.2}$$

$$C^{p+1}_{4k+2,(i,j)} = \sum_m \sum_n g(m)h(n)C^p_{k,(m+2i,n+2j)}, \tag{2.3}$$

$$C^{p+1}_{4k+3,(i,j)} = \sum_m \sum_n g(m)g(n)C^p_{k,(m+2i,n+2j)}, \tag{2.4}$$

where C_0^0 is the image A. At each step, the image C_k^p (with the value of $C_{k,(i,j)}^p$ at the position (i,j)) is decomposed into four quarter-size images C_{4k}^{p+1}, C_{4k+1}^{p+1}, C_{4k+2}^{p+1}, C_{4k+3}^{p+1}. A two-level wavelet packet decomposition is illustrated in Figure 2.1.

2D-DWPT decomposition allows us to analyze an image simultaneously at different resolution levels and orientations. Several energy functions can be used to extract feature from each subband for classification. Commonly used energy functions include magnitude $|\bullet|$, magnitude square $|\bullet|^2$ and the rectified sigmoid $|\tanh(\alpha)\bullet|$ [9]. Both magnitude and magnitude square are parameter-free while the rectified sigmoid function can be adjusted by the parameter α. The mean and the standard deviation of subband coefficients can also be extracted as features [24, 44]. All these definitions of energy are highly correlated. In this chapter, the definition of energy based on squaring is used. The energy in different subbands is computed from the subband coefficient matrix as:

$$\sigma_p^2(k) = \sum_i \sum_j [C_{k,(i,j)}^p]^2 \tag{2.5}$$

where $\sigma_p^2(k)$ is the energy of the image projected onto the subspace at node (p,k). The energy of each subband provides a measure of the image characteristics in that subband. The energy distribution has important discriminatory properties for images and as such can be used as a feature for texture classification.

2.3 Dependence Analysis

Wavelet packet decomposition is an orthogonal transform. However, it is observed that considerable dependence exists between coefficients in different subbands, especially for coefficients from subbands having "parent-child" or "sibling" relations in the decomposition tree. The dependence between coefficients has been modelled by Hidden Markov Model (HMM) [39, 40] and utilized for better image compression efficiency [43]. However, no analysis has been done on the dependence among features extracted from different subbands. In this section, covariance is used to analyze the dependence between energy values from any two subbands. The analysis is first

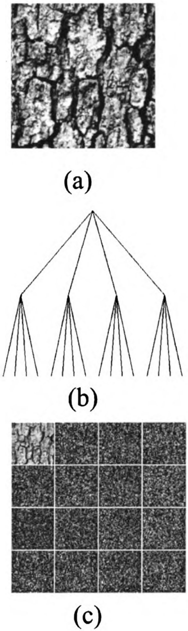


Figure 2.1. A wavelet packet decomposition tree: (a) The original image, (b) The 2-level decomposition tree, (c) Projection of the original image onto each leaf node of the tree.

conducted theoretically and then a simulation is used for a given image model.

2.3.1 Theoretical Analysis

Based on equations (2.1) to (2.4), the wavelet coefficients at level $(p+1)$ is obtained by convolving the wavelet coefficients at level p with the wavelet filters. The highest level in the wavelet packet decomposition tree is the original image A. Therefore, the coefficients in any subband can be written as a linear combination of pixel values in the original image A. The weighting coefficients in the linear combination are determined by the properties of the wavelet basis, i.e. the lengths and the values for filters g and h. Considering the definition of subband energy in equation 2.5, the energy of wavelet coefficients in a subband is a second order polynomial in terms of the pixel values of the image A. Coefficients in the polynomials are determined by the filters g and h and the decomposition level. Denote the coefficient for the term $A(i,j)A(i',j')$ as $f_k^p(i,j,i',j')$ in the representation of an energy function $\sigma_p^2(k)$, as follows:

$$\sigma_p^2(k) = \sum_i \sum_j \sum_{i'} \sum_{j'} f_k^p(i,j,i',j') A(i,j) A(i',j') \tag{2.6}$$

Due to the downsampling, some of the coefficients $f_k^p(i,j,i',j')$ are zero. Hence, the covariance between two energy values is defined as

$$\text{Cov}(\sigma_{p_1}^2(k_1), \sigma_{p_2}^2(k_2)) = E[\sigma_{p_1}^2(k_1)\sigma_{p_2}^2(k_2)] - E[\sigma_{p_1}^2(k_1)]E[\sigma_{p_2}^2(k_2)]. \tag{2.7}$$

$\text{Cov}(\sigma_{p_1}^2(k_1), \sigma_{p_2}^2(k_2))$ will be a fourth-ordered polynomial in terms of the pixel values in the image A. Using definitions (2.1) to (2.4), the correlation between energy values of two children nodes, $\sigma_1^2(0)$ and $\sigma_1^2(1)$, is:

$$\begin{aligned} &E\sigma_1^2(0)\sigma_1^2(1)] \\ &= E[\sum_i\sum_j\left(\sum_m\sum_n h(m)h(n)x(m+2i,n+2j)\right)^2 \\ &\quad \sum_p\sum_q\left(\sum_l\sum_r h(l)h(r)x(l+2p,r+2q)\right)^2\,] \end{aligned} \tag{2.8}$$

Given that g and h are FIR filters, the random variables in equation (2.8) are the different pixel values. Therefore, the covariance is a function of the 4th-order statistics of the original image A. If we have a statistical model for the pixels, the correlation value can be easily calculated.

To illustrate this, suppose that the image A is of size 4×4. The image model can be described using AR-1 Gaussian model, which is simple but often used in digital image processing [45]. A useful property of AR-1 Gaussian model is that the marginal distribution of each pixel is also Gaussian [46]. The Gaussian AR-1 process with mean μ is usually written in terms of a series of white noise innovation processes $\{E_n\}$:

$$X_n - \mu = a(X_{n-1} - \mu) + E_n \tag{2.9}$$

where $E_n \sim N(0, \sigma^2)$ are i.i.d. and $|a| < 1$. The marginal distribution is also normal:

$$X_n \sim N(\mu, \frac{\sigma^2}{1-a^2}) \tag{2.10}$$

Given the marginal distribution, equation (2.8) can be expanded as follows:

$$\begin{aligned} E[\sigma_1^2(0)\sigma_1^2(1)] &= v_1(h,g)E[X_1^4] + v_2(h,g)E[X_1^3X_2] \\ &+ v_3(h,g)E[X_1^2X_2^2] + v_4(h,g)E[X_1^2X_2X_3] + v_5(h,g)E[X_1X_2X_3X_4] \end{aligned} \tag{2.11}$$

where $\{X_1, X_2, X_3, X_4\}$ are i.i.d. Gaussian random variables corresponding to different pixels distributed as given by equation (2.10), and $\{v_1(h,g), v_2(h,g), v_3(h,g), v_4(h,g)\}$ are functions of wavelet filters.

Suppose that the size of the original image X is $N \times M$, and define $K =$

$\min(N, M)$. $v_1(h,g)E[X_1^4]$ represents those terms on the right hand side of equation (2.8) where the indices of four pixel values are exactly the same. Thus, $v_1(h,g)E[X_1^4]$ contains exactly K terms. Similarly, $v_2(h,g)E[X_1^3X_2]$ contains 4K(K-1), $v_3(h,g)E[X_1^2X_2^2]$ contains 3K(K-1), $v_4(h,g)E[X_1^2X_2X_3]$ contains 6K(K-1)(K-2) and $v_5(h,g)E[X_1X_2X_3X_4]$ contains K(K-1)(K-2)(K-3) terms from the right hand side of equation (2.8). Based on the marginal distribution, the four expected values are given as follows:

$$\begin{aligned}
&E[X_1^4] = \mu^4 + 6\mu^2\tfrac{\sigma^2}{1-a^2} + 3\left(\tfrac{\sigma^2}{1-a^2}\right)^2\\
&E[X_1^3X_2] = \mu^4 + 3\mu^2\tfrac{\sigma^2}{1-a^2}\\
&E[X_1^2X_2^2] = \left(\mu^2 + \tfrac{\sigma^2}{1-a^2}\right)^2\\
&E[X_1^2X_2X_3] = \mu^4 + \mu^2\tfrac{\sigma^2}{1-a^2}\\
&E[X_1X_2X_3X_4] = \mu^4
\end{aligned} \tag{2.12}$$

Therefore, we can obtain the covariance between the energy features from two subbands analytically. In the next section, we compute the derived covariance values for a specific set of wavelet filters.

2.3.2 Simulation

Suppose that the Haar wavelet basis is used, i.e., g={0.7071, 0.7071} and h = {-0.7071, 0.7071}. The parameters in equation (2.10) are chosen as follows: $\mu = 128$, $\sigma = 12$ and $a = 0.8$, which are standard values for a 256-valued grey image. The normalized covariance:

$$N\_Cov(X,Y) = \frac{Cov(X,Y)}{\sqrt{Cov(X,X)}\sqrt{Cov(Y,Y)}} \tag{2.13}$$

is used for measuring the dependence between energy of different subbands. With these assumptions, the covariance matrix for energy values from any two subbands in the wavelet packet decomposition tree can be easily calculated. The size of images at the first decomposition level is 2 × 2. The covariance matrix between the four

Table 2.1. Normalized energy covariance with haar basis

	$h\_\sigma_1^2(0)$	$h\_\sigma_1^2(1)$	$h\_\sigma_1^2(2)$	$h\_\sigma_1^2(3)$
$h\_\sigma_1^2(0)$	1.0000	0.5656	0.2765	-0.0669
$h\_\sigma_1^2(1)$	0.5656	1.0000	0.2782	-0.0610
$h\_\sigma_1^2(2)$	0.2765	0.2782	1.0000	-0.2705
$h\_\sigma_1^2(3)$	-0.0669	-0.0610	-0.2705	1.0000

Table 2.2. Normalized energy covariance with db4 basis

	$d\_\sigma_1^2(0)$	$d\_\sigma_1^2(1)$	$d\_\sigma_1^2(2)$	$d\_\sigma_1^2(3)$
$d\_\sigma_1^2(0)$	1.0000	0.5560	0.5635	0.0191
$d\_\sigma_1^2(1)$	0.5560	1.0000	-0.1294	0.2210
$d\_\sigma_1^2(2)$	0.5635	-0.1294	1.0000	-0.0029
$d\_\sigma_1^2(3)$	0.0191	0.2210	-0.0029	1.0000

energy values from the first decomposition level, $\sigma_1^2(0)$, $\sigma_1^2(1)$, $\sigma_1^2(2)$, $\sigma_1^2(3)$, is shown in Table 2.1. For the Daubechies' 4-tap (2 vanishing moments) filters, where g = {0.4830, 0.8365, 0.2241, -0.1294} and h = {0.1294, 0.2241, -0.8365, 0.4830}, the same 4 × 4 covariance matrix is shown in Table 2.2. In Table 2.3, the covariance between 4 energy values computed from the subbands in the first decomposition level of Haar basis and 4 energy values for the corresponding subbands in Daubechies' 4-tap basis is given.

These results show that various degrees of dependence exist between the energy features extracted from different subbands within a wavelet basis and from different wavelet bases. The wavelet basis, i.e., the high and low pass filter pair h and g, affects the degree of dependence. To analyze the dependence between energy values more

Table 2.3. Normalized energy covariance between db4 basis and haar basis

	$h\_\sigma_1^2(0)$	$h\_\sigma_1^2(1)$	$h\_\sigma_1^2(2)$	$h\_\sigma_1^2(3)$
$d\_\sigma_1^2(0)$	-0.0516	-0.0214	-0.0707	0.9870
$d\_\sigma_1^2(1)$	-0.1279	-0.0295	0.3520	0.4322
$d\_\sigma_1^2(2)$	0.0665	0.3608	0.0189	0.6090
$d\_\sigma_1^2(3)$	0.2174	0.1972	0.0339	0.0666

accurately, more complicated statistical image models are needed.

The analysis in this section illustrates that given the statistical model for image pixels, we can calculate the covariance between energy values for any two subbands. The results indicate that the wavelet basis has a great impact on the degree of dependence between energy values from different subbands. The energy values from subbands with different wavelet bases may have different degrees of dependence. Therefore, the assumption used in many image processing algorithms that different wavelet subbands yield independent energy values is incorrect. Thus, the independent subbands can be combined to form a sparse representation for classification.

2.4 Subband Selection with Dependence

In Section 2.3, an analytical procedure is given for quantifying the dependence between energy values from different subbands using different wavelet bases, given that the underlying image model is known. Although in most applications, an accurate statistical image model is not available [47], the simulation in Section 2.3 still strongly supports the hypothesis that the energy values from different subbands and even different wavelet bases may be dependent. This observation motivates us to select subbands based on dependence to generate a compact representation of the wavelet features for texture classification.

In real applications, due to the lack of suitable statistical image models, it is not likely that dependence between features can be easily calculated analytically. However, the dependence relationship between energy values from subbands can still be estimated with nonparametric methods based on a training set. More specifically, given training samples, the dependence between the energy values from different subbands can be estimated empirically from the training data. Based on this estimation, independent subbands can be chosen to achieve a compact representation for the subsequent classification. In this chapter, Mutual Information based Subband Selection (MISS) is proposed to implement this subband selection process. Given a training texture set T_n, a testing texture set T_c for classification, and a set of subbands $U = \{S_1, S_2, ..., S_N\}$ from the wavelet packet decomposition, the MISS algorithm can be formulated as follows.

1. For each texture $A_i \in T_n$, extract the energy values for all subbands. Notice that each time the texture is decomposed, the subband size is decreased by a factor of 4. When the subband size is small, the energy of the subband will not be a robust measure. In the implementation, the minimum subband size is set to 16×16.

2. Divide the subbands into 2 sets: SU and SR. Initialize SU to contain the subband that has the highest energy and $SR = U \backslash SU$.

3. Move a subband S_i from SR to SU based on the following criterion: S_i is chosen such that the feature corresponding to S_i has the smallest mutual information with the features extracted from the subbands in the set SU, i.e.

   $$i = \arg\min_j \{I(S_j; SU), S_j \in SR\} \tag{2.14}$$

   where S_i is the *i-th* wavelet subband. Here we use mutual information I to measure the dependence. This process stops when the number of subbands in SU reaches a predefined value, which can be determined by cross validation. In the experiments presented in this chapter, this value is set as a parameter and the relation between the value of this parameter to the classification accuracy is studied. Finally, the subbands in the set SU are used to construct the compact representation for the given images. Note that since the goal is to classify the images, it is not required that images are reconstructed from SU, i.e. the compact representation does not necessarily form a basis.

4. Extract energy values corresponding to the sparse representation SU for all images in the testing set T_c.

This algorithm tries to select subbands such that the energy features from these subbands are as independent as possible. It does not require that the chosen sub-

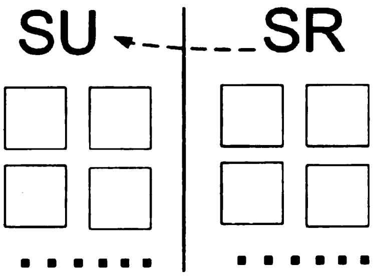


Figure 2.2. Illustration of MISS algorithm: At each step, a subband in SR with lowest mutual information with all subbands in SU is moved from SR to SU. In this figure, each square represents a single wavelet subband.

bands construct a valid wavelet packet decomposition tree, as in [2,27,30]. A simple illustration of MISS algorithm is given in Figure 2.2.

In step 3), the mutual information is used to quantify the dependence between a set of subbands in U and a single subband. Traditional measures of dependence, such as correlation and covariance are not defined between a set of variables and a single variable, while mutual information is well-defined to measure such dependencies. The computation of mutual information will be discussed in more detail in the following two subsections.

2.4.1 Mutual Information

This subsection discusses the computation of mutual information, which is used in the step 3) of the MISS algorithm. Consider two random variables $X \in \tilde{\mathrm{X}}$ and $Y \in \tilde{\mathrm{Y}}$ with a joint distribution *p(x,y)*. The mutual information between X and Y is defined as:

$$I(X;Y) = \sum_{x\in\tilde{X}} \sum_{y\in\tilde{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} = E_{XY}\left[\log\frac{p(x,y)}{p(x)p(y)}\right] = D(p(x,y)||p(x)p(y)), \tag{2.15}$$

where $D(\bullet||\bullet)$ is the Kullback-Leibler distance between two probability mass functions *p(x,y)* and *p(x)p(y)*. When the logarithm function in the definition uses a base of 2, the unit for the $I(X;Y)$ is bits. The mutual information is symmetric in X and Y, nonnegative, and is equal to zero if and only if X and Y are independent. The mutual information $I(X;Y)$ indicates how much information Y conveys about X. Given Y, the extra information required to fully describe X is given by conditional entropy $H(X|Y)$ [48]. Thus, the following equation holds:

$$I(X;Y) = H(X) - H(X|Y), \tag{2.16}$$

where $H(X)$ is the entropy of random variable X. Similar to the definition of conditional entropy, the conditional mutual information between random variables X and Y given Z is defined as:

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = E_{p(x,y,z)}[\log \frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)}]. \tag{2.17}$$

The conditional information has an interpretation similar to that of mutual information. Given the above definitions, the mutual information between a set of random variables $\{X_1, X_2, ..., X_n\}$ and a single random variable Y can be defined as:

$$I(X_1, X_2, ..., X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, X_{i-2}, ..., X_1). \tag{2.18}$$

This definition of mutual information can be used for evaluating the independence between the features extracted from a set of subbands SU and a single subband S_i. The higher the value of the mutual information, the easier it is to estimate the distribution of S_i given SU. The mutual information thus provides a criterion for the subband selection process.

2.4.2 Computation of Mutual Information

Considering the right hand side of equation (5.4), one practical difficulty in computation is the estimation of high-dimensional joint pdfs (or equivalently, conditional pdf), due to the possible large size of the subband set U. For example, if the texture

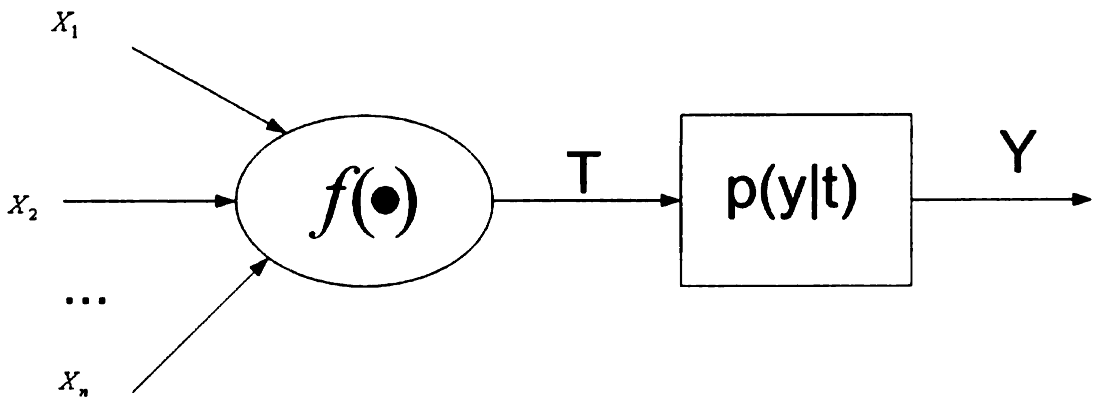


Figure 2.3. Dimension reduction in mutual information through a many-to-one mapping function

is decomposed up to level 3 using wavelet packets, then there are 85 subbands. It is impractical and unreliable to estimate such high-dimensional distributions. To avoid the problem of "curse of dimensionality", we assume that the set of random variables $\{X_1, X_2, ..., X_n\}$ provides information to Y only through a many-to-one scalar mapping $T = f(X_1, X_2, ..., X_n)$, in this sense, the mutual information $I(X_1, X_2, ..., X_n; Y)$ can be calculated as $I(T; Y)$, as illustrated in Figure 2.3.

In this model, T is obtained by processing X. Thus, applying the data-processing theorem [48] yields the following inequality:

$$I(X_1, X_2, ..., X_n; Y) \geq I(T; Y). \tag{2.19}$$

The equality is achieved if and only if $T = f(X_1, X_2, ..., X_n)$ is the sufficient statistics for $\{X_1, X_2, ..., X_n\}$ [48]. Since this is rarely true in practice, finding a proper form for the function $f(\bullet)$ is important to minimize the error caused by the inaccurate assumption. Since the true mutual information $I(X_1, X_2, ..., X_n; Y)$ is the upper bound for the estimated mutual information $I(T; Y)$, the function $f(\bullet)$ should maximize the estimation $I(T; Y)$. For the convenience of computation, $f(\bullet)$ is usually assumed to be a linear function. With this assumption, the optimization is still nontrivial. In [49–51], more approximations are proposed for computing mutual information. The approximations include replacing the definition of mutual information in equation (2.15) with Renyi's definition, and using Parzen's window method for modelling pdfs in high dimensional spaces. Even with all these assumptions and approximations, the complex iterative optimization steps can only achieve locally optimal solution that highly depends on the initialization of the numerical procedure.

A similar mutual information computation problem arises in the statistical modelling of wavelet coefficients and is solved in [38,42]. In [38], a simple linear model was adopted, as follows:

$$T = f(X_1, X_2, ..., X_n) = \sum_{i=1}^{n} W_i X_i. \tag{2.20}$$

Optimizing this model, i.e., the weighting coefficients $\{W_i\}$, incurs similar difficulties discussed in [49–51]. To avoid the difficulties, a simple equal weight model, i.e., $W_i = 1/n$, is used in [38] and the results from the computational model matches the statistical property of actual images well. An intuitive improvement on the equal weight model by exploiting the "parent-children" and "sibling" relation among wavelet coefficients is also proposed in [38], but shows similar results to that of the equal weight model. In this chapter, equal weights are used for mutual information computation, i.e., $W_i = 1/n$. In this case, T is the unbiased estimate for the mean of $\{X_i\}$.

At this point, the accuracy of this simplified model for mutual information computation and its effect on the MISS algorithm deserve further discussions. In [38], different weighting models, such as equal weight and optimal linear weights have been proposed. The experimental results show that the mutual information estimate does not vary much based on the weighting model. By using the equal weight model, the energy values in the set SU are averaged. In this process, the subbands with higher energy values play a more important role in the mutual information computation that determines the next selected subband. This approximation of mutual information computation may cause some bias in the selection of the subbands. The next selected subband is the one least dependent on those subbands in SU with high energy values. This is consistent with the current knowledge that the wavelet subbands with higher energy values are more informative for texture classification. As shown in the experimental results in the later sections, the simplified model is able to achieve the goal of wavelet feature selection in the sense that a small number of subbands selected by the MISS algorithm are able to achieve a relatively low classification error rate.

In the actual implementation, the value of mutual information between two discrete vectors X and T needs to be computed in the discrete domain. Usually, the ranges of X and T are partitioned into N_X and N_T intervals, respectively [52, 53]. The continuous pdf *p(x,t)* can be approximated by the 2D histogram $\{P_{X,T}(i,j), 1 \leq i \leq N_X, 1 \leq j \leq N_Y\}$. Similarly, the marginal distributions $p_X(x)$ and $p_T(t)$ can be estimated by $P_X(i)$ and $P_T(j)$, respectively. Based on the the estimated discrete probability distributions, the mutual information can be calculated as:

$$\hat{I}(X;T) = \sum_i \sum_j P_{X,T}(i,j) \log \frac{P_{X,T}(i,j)}{P_X(i) P_T(j)}. \tag{2.21}$$

If the random processes X, T are stationary and ergodic, then the histogram reliably approaches the pdf and $\hat{I}(X;T)$ converges to *I(X;T)*.

2.5 Subband Selection Combining Dependence and Energy

As mentioned in Section 5.1, the evaluation of individual subbands provides useful information for subband selection. One widely used evaluation criterion is the magnitude of energy computed for each subband. It has been shown that images can be accurately identified by subbands with high energy values [1, 7]. It is now clear that two factors, evaluation of individual subbands and dependence among subbands, affect the subband selection. In Section 2.4, dependence between different subbands is considered for subband selection. However, the subbands selected based on the independence criterion are not necessarily the most significant subbands for classification. In order to further improve subband selection for classification, Subband Grouping and Selection (SGS) algorithm is proposed to combine the dependence between subbands and the evaluation of individual subbands. SGS achieves this objective through two steps. First, the algorithm discovers the structure of statistical dependency between subbands, then it selects the subbands based on this structure using evaluation scores of individual subbands. In order to discover the structure of statistical dependency between subbands, SGS employs a grouping algorithm that partitions the subbands into different sets based on their dependency on each other, i.e. ideally the subbands from the same set are dependent, while subbands from different sets are independent. The proper partitioning of all the subbands reveals the structure of statistical dependency between subbands. In the second step, the subband with the highest energy from each subset is selected for classifying images. With these two steps, SGS successfully incorporates the previous research results in [1], and exploits the statistical dependency among subbands for a concise representation of the images.

2.5.1 Statistical Dependency Between Subbands

The general problem of discovering the structure of statistical dependency can be described as follows: Given a set of subbands $U = \{S_1, S_2, ..., S_N\}$, which is the full wavelet packet decomposition in our problem, find a partition $\{U_1, U_2, ..., U_M\}$ of set U such that

$$\bigcup_{i=1}^{M} U_i = U \quad \text{and} \quad U_i \cap U_j = \emptyset, i \neq j \tag{2.22}$$

where M is the number of subsets. It is desired that that the subbands from the same subset U_i are dependent on each other, while the subbands from different subsets are independent. If this property holds, then we can choose only one feature from each subset to construct a concise representation of the texture.

Discovering the structure of statistical dependency has been addressed in literature. One of the recent research results in this area was introduced in [54], where a method based on hypothesis testing is proposed. Different hypotheses on subset partitions were compared based on log-likelihood. However, the number of all possible partitions of a set with M random variables is in the order of $O(M^M)$, which is practically intractable. The hypothesis testing also requires evaluating the joint pdf of all the random variables in the same set. The estimation of the pdf in high-dimensional spaces is usually unstable and inaccurate, due to the sparsity of the training data [10]. To avoid this problem of estimating high-dimensional joint pdfs, we propose an approximate method for statistically partitioning the subbands based on pairwise dependency. More specifically, we can estimate the marginal pdf of features extracted from each subband. Using these estimated marginal pdfs, the pairwise dependency between two subbands can be quantified. A data grouping algorithm is subsequently applied to features from all subbands, generating the partition of the features from each subband. This grouping algorithm only requires pairwise dependencies, thus avoiding computations in high-dimensional space. With this method, the dependent subbands tend to be grouped into the same subset. Note that this method is an approximation to the discovery of the structure of statistical dependency. For example, given three random variables X, Y, Z, it is possible that $P(X|Y) = P(X)$, $P(X|Z) = P(X)$, while $P(X|Y, Z) \neq P(X)$ [55]. If only pairwise dependence is measured, X will not be grouped into the same subband with Y and Z, though the dependence is evident. Another limitation lies in the data grouping algorithm itself. Research on data grouping, though conducted for several decades, is still far from perfect [10]. This indicates that an accurate partitioning of the subbands such that the ones in the same subset are dependent, while the ones from different subsets are independent can only be approximated. Despite these limitations, the grouping algorithm, measuring pairwise dependency, is still able to find out the dependence structure in the data.

In this chapter, K-Means grouping algorithm is used to generate the partition of subbands. Given N subbands, M partition subsets ($N > M$), and the mean of each subset as $\{\mu_1, \mu_2, ..., \mu_M\}$, the K-Means algorithm can be described as follows [10].

1. Randomly select M subbands from N subbands and set them as initial class means $\{\mu_1, \mu_2, ..., \mu_M\}$.

2. Classify the N subbands. Each subband is assigned to the mean μ_i with which it has the highest mutual information.

3. Recompute the means $\{\mu_1, \mu_2, ..., \mu_M\}$ based on the partition of subbands from step (2).

4. Repeat steps 2) and 3) until the means do not change.

It is clear that the K-Means is an iterative algorithm. The initial selection of means ("seeds") can affect the final partition results. To avoid this randomness in final performance evaluation, we run the K-Means algorithm multiple times and the final evaluation is based on the average value from all runs.

2.5.2 Subband Grouping and Selection

The problem of discovering the structure of statistical dependency among subbands has been addressed in Section 2.5.1. Incorporating this partitioning algorithm into our method, the subband with highest energy value in each subset is chosen for texture classification. In summary, given a training texture set T_n and a test texture set T_c for classification, and a subband set $U = \{S_1, S_2, ..., S_N\}$ from wavelet packet decomposition, the SGS algorithm can be described as follows.

1. For each texture $A_i \in T_n$, extract the energy values of all subbands from wavelet packet decomposition. Note that each time the texture is decomposed, the subband size is decreased by a factor of 4. When the size is too small, the energy of the subband will not be a robust measure. In the implementation, the minimum subband size is set to 16×16.

2. For each subband S_i, estimate the marginal pdf of its energy value using Parzen method with Gaussian kernel [10] based on the training set.

3. For any two subbands S_i, S_j, use equation (2.21) to compute the mutual information between the extracted features using the pdfs from step 2.

4. Apply K-Means grouping algorithm with the pairwise mutual information as the metric, generate the partition of the subband set U: $\{U_1, U_2, ..., U_M\}$.

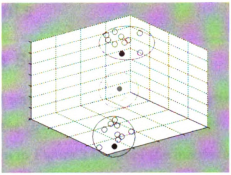


Figure 2.4. A simplified illustration of SGS: subbands are first partitioned into different groups (3 ellipses) and then the subband with highest energy (filled dot) is selected from each group.

5. Select the subband with the highest energy values from each subset. From the subset U_i, a subband SS_i is selected, such that:

$$i = \arg\max_{j}\{E(S_j), S_j \in U_i\} \tag{2.23}$$

where $E(S_j)$ is the energy feature extracted from the subband S_j and is defined by equation (2.5). The set of subbands used for representing the images is thus $SU = \{SS_1, SS_2, ..., SS_M\}$.

6. Extract subband energies of all images in the testing set T_c. Use only the energy values corresponding to the subbands in the set SU for classification.

The SGS algorithm attempts to select independent subbands with high energy values for classification. This point is reflected in steps 4 and 5 of the algorithm, where subbands are grouped according to the dependence relation and only subbands with high energy values are selected. Note that SGS does not require the chosen subbands to construct a valid wavelet packet decomposition tree, as in [2]. A simple illustration of SGS for a three dimensional space is given in Figure 2.4. Note that this figure is simplified for illustration. In the actual application, the extracted features are in a much higher dimensional space and the different subbands may not be well separated.

2.6 Experiments

2.6.1 Experiment Setting

To evaluate the proposed wavelet subband selection algorithms, classification experiments are conducted on the Brodatz texture database [56]. All of the images used in the experiments are shown in Figure 2.5. All of the images are of gray-value and have a size of 512 × 512. Each texture is divided into 16 non-overlapping subimages with a size of 128 × 128. In this way, the data set for experiments contains 54 different texture classes, with 16 images in each class, resulting in 864 images for experiments. This experimental setting of using training and testing images from the same large texture image is used commonly in evaluation of texture classification methods, such as in [1,5,23,25,26]. It is known that the size of subimages affects the classification results. Since the focus of this chapter is to demonstrate the effect of subband selection methods rather than finding a set of optimal parameters for texture classification, we fix the texture size to 128 × 128. For images at different sizes, the subband selection methods can be applied without any significant change.

The gray-scale images are first normalized to a given mean and variance. Denote $A(i,j)$ as the gray-level value of pixel (i,j) and μ and σ as the mean and variance of A, respectively. The normalized image A' is given by:

$$A'(i,j) = \begin{cases} \mu_0 + \sqrt{\frac{\sigma_0^2(A(i,j)-\mu)^2}{\sigma}}, & \text{if} \quad A(i,j) > \mu \\ \mu_0 + \sqrt{\frac{\sigma_0^2(A(i,j)-\mu)^2}{\sigma}}, & \text{if} \quad A(i,j) \leq \mu \end{cases} \tag{2.24}$$

where μ_0 and σ_0 are the predefined mean and variance of the adjusted image A', separately. In the experiment, we set $\mu_0 = 128$ and $\sigma_0 = 30$.

The experiment includes two stages: the training stage and the testing stage. Half of the data set ($27 \times 16 = 432$ images) is used for training and the other half is used for testing. In the training stage, each texture in the training set is decomposed with the wavelet packet transform up to 3 levels. Thus, for each texture, the number of subbands from a single wavelet basis is $1+4+16+64 = 85$. A pre-determined number

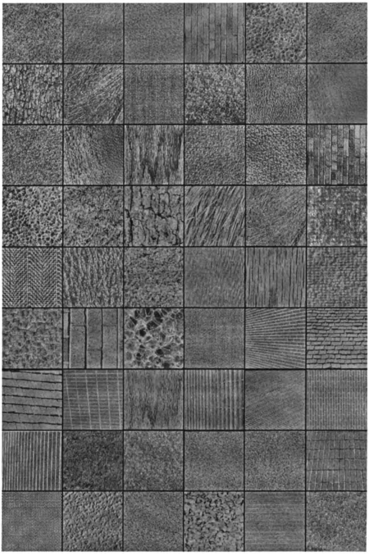


Figure 2.5. Brodatz texture images used in the experiments.

of subbands are selected by applying the subband selection algorithms on the training set T_n. In the training stage, due to the randomness of selecting the initial seeds for the K-Means grouping algorithm, SGS was run 10 times for the same training and testing data set and the results were averaged. For this experimental setting, it is found empirically that averaging on 10 runs results in a relatively stable result.

Classification experiments are conducted on the testing set, T_c, with the images decomposed only at the subbands selected in the training stage. In the testing stage, the classification error rates for different number of selected subbands are computed for performance evaluation. The classification error rate is computed using the KNN classifier with cross-validation (leave-one-out) [10]. All of the images in the testing set T_c are classified. In each round, one texture from T_c is taken out and the normalized Euclidian distance defined in equation (5.19) between this texture and all the other images from T_c are computed based on the energy features extracted from the selected subbands. The distance on energy features is defined by the normalized Euclidian distance, which was shown to be effective for measuring texture similarity [57]:

$$d(F_1, F_2) = \sum_{i=1}^{n} \left| \frac{F_1(i) - F_2(i)}{\sigma_i} \right|^2 \tag{2.25}$$

where F_1, F_2 are two different wavelet energy feature vectors, and σ_i is the standard deviation of the feature extracted from subband i, estimated from the training set T_n. The "K" most similar images (i.e., the images with the smallest distances) determine the class label of the texture to be classified by a majority vote. After each texture in the testing set T_c is classified in this way, the ratio of the number of misclassified images to the number of total images in T_c is computed as the error rate. The value "K" in "KNN" is chosen to be 16. Considering that the number of images in each class is 16, we expect that most of the 16 most similar images come from the same class as the texture to be classified.

For comparison, existing algorithms for wavelet subband selection are implemented and tested. The different methods that are implemented for comparison include: (1) the subband selection algorithm based only on the magnitude of the energy in each subband [1], referred to as "Energy" method in the following discussions; (2) the subband selection algorithm that evaluates the Fisher discrimination power of each subband [21, 32], referred to as "Fisher" in the following discussion; (3) the best wavelet packet decomposition tree based on entropy [30], referred to as "Best Tree" method in the following discussions. In the "Best Tree" method, if the entropy value of a subband is less than the sum of entropy values of its children subbands, the subband is decomposed into children subbands. This criterion is iteratively applied to the leaf nodes of the current wavelet decomposition tree. With this criterion, we can not obtain any arbitrary number of subbands, since whether a subband is decomposed or not depends on a fixed criterion. The only parameter that we can modify in this process is the decomposition level, which relates to the maximum number of subbands; (4) The subband selection algorithm that evaluates each subband with its entropy value. This method selects subbands with high entropy values for classification. This method is a combination of the ideas presented in [30], [32], and [21]. With this method, any number of subbands can be selected. This method is referred to as "Entropy" in the following discussions.

Note that the criterion used in the "Best Tree" method [30] aims at reducing the image reconstruction error, and not improving the classification accuracy. This method is included for comparison, since it is widely used for selecting the optimal wavelet packet tree structure. The four selection methods ("Energy", "Fisher", "Best Tree" and "Entropy") are tested using the same experimental settings as the two algorithms proposed in this chapter ("MISS" and "SGS"). The error rates for different number of subbands selected by different methods are reported and compared.

2.6.2 Performance with Different Wavelet Bases

Different wavelet bases define the transform filters used in the tree structure decomposition introduced in Section 2.2. Therefore, using different wavelet bases will generate different energy distributions over subbands. The relationship between wavelet bases and classification performance was empirically discussed in [5]. In this chapter, three wavelet bases are used for experiments: Haar, 'db10' (Daubechies' basis with 10 vanishing moments) and 'bio28' (biorthogonal spline wavelets having 2 vanishing moments in the decomposition filter and 8 vanishing moments in the reconstruction filter.) [13]. Since the total number of subbands for each wavelet basis is 85, the number of selected subbands is set to change from 5 to 85, with increments of 5. Figs. 2.6, 2.7 and 2.8 show the classification performance with the 'Haar', 'db10' and 'bio28' wavelet bases, respectively.

Several observations can be made from these figures. First of all, the dependence between features extracted from different subbands is useful for subband selection. This is verified by the performance of the MISS algorithm. For all of the three wavelet bases, MISS algorithm can achieve a classification accuracy comparable to the accuracy with the full decomposition using only 20 to 30 subbands. Second, the evaluation of individual subbands is also effective in subband selection. The "Energy", "Entropy" and "Fisher" methods select subbands based on the evaluation of single subbands with different evaluation functions. For the "Haar" and "db10" bases, MISS algorithm performs better than the three algorithms when smaller number (less than 30) of subbands are selected. However for the "bio28" basis, "Fisher" method performs better for the same range. This shows that both factors (dependence between different subbands and the evaluation of individual subbands) affect the subband selection for classification. Third, the most prominent result is that SGS consistently outperforms the other algorithms. This is due to the fact that both dependence between subbands and the evaluation of individual subbands are incorporated into the selection process. The advantage is more obvious when smaller number of subbands are selected. The number of selected subbands is the number of clusters in SGS algorithm. When the number of clusters is relatively small compared to the total number of subbands, the distribution of clusters reflect more reliably the underlying dependence structure, since more data points can be used to construct each cluster. Finally, the "Best Tree" algorithm does not achieve the best classification accuracy. The "Best Tree" is optimal in the sense of image reconstruction, not in the sense of classification. A small reconstruction error does not necessarily mean

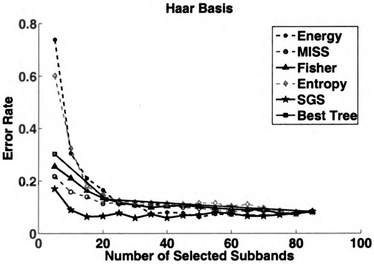


Figure 2.6. The error rates with six subband selection algorithms with the 'Haar' basis.

a small classification error. For the wavelet packet decomposition of the Brodatz texture image, we find that the optimality criterion used in the "Best Tree" method always chooses to fully decompose an image, resulting in 5, 21 and 85 subbands for level 1, 2 and 3 decompositions, respectively. This is why the "Best Tree" curves in the figures have only 3 data points.

2.6.3 Performance Using Subbands Selected from Two Wavelet Bases

Traditional subband selection algorithm confines to selecting subbands that are generated from a single wavelet basis. However, for classification, the pool of candidate subbands can be expanded to subbands generated by multiple wavelet bases. For example, we can combine subbands generated by the "Haar" basis and the "db10" basis. As long as the selected subbands form a set of features that are not correlated to each other and are discriminant in classification, it is not required that the selected subbands can be used to perfectly reconstruct the original image. Extension to selecting subbands from multiple wavelet bases does not require the modification of

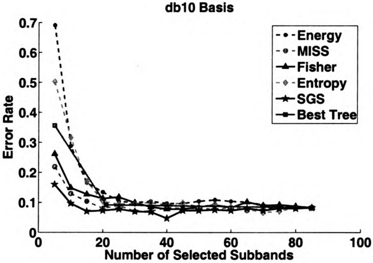


Figure 2.7. The error rates with six subband selection algorithms with the 'db10' basis.

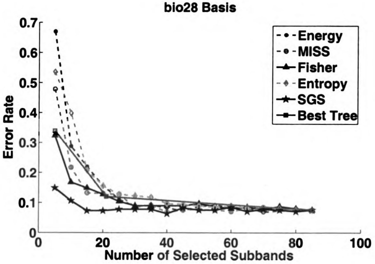


Figure 2.8. The error rates with six subband selection algorithms with the 'bio28' basis.

the algorithms described in Section 2.6.1 except that the pool of subbands to choose from has been enlarged. For the "Best Tree" algorithm, the features from subbands selected by the algorithm on different decomposition trees with different bases are simply concatenated.

In this subsection, experiments on selecting subbands from two wavelet bases are conducted for all of the subband selection methods mentioned in Section 2.6.1. Figs. 2.9, 2.10, and 2.11 show the error rates obtained by the algorithms selecting subbands from "bio28" and "db10","Haar" and "bio28", and "db10" and "Haar", respectively. Each decomposition generates 85 subbands, so the total number of subbands for selection is 170. In order to make the results directly comparable with subband selection from a single wavelet basis, the number of selected subbands is still set to change from 5 to 85, with an increment of 5.

In this modified experiment setting, the SGS still consistently outperforms other algorithms and the observations in Section 2.6.2 still hold here. It should be noted that the error rate obtained by selecting 85 subbands out of a possible 170 subbands from the two wavelet bases, regardless of the algorithm and the wavelet bases pairs used, is lower than the error rate using all 85 subbands from a single wavelet basis. This shows the advantage of selecting subbands from multiple wavelet bases or redundant dictionaries in general. More subbands can provide more information for classification and the proper selection algorithm can select the most informative subbands for classification.

The minimum error rates achieved with different wavelet bases and combination of bases by applying different subband selection methods are summarized in Table 2.4 for comparison. Note that the different minimum error rates are not necessarily obtained with the same number of subbands. Based on the results in this table, the best classification performance is achieved by applying the SGS algorithm on the combination of "bio28" and "haar" bases. An empirical study on the relation between wavelet basis and the classification results on texture classification was discussed in [5]. Based on the conclusion from [5], the amount of shift variance in the decomposition filters in the wavelet basis is much more important than the regularity of the filters.

Table 2.4. Minimum Error Rates Achieved With Different Wavelet Bases and Different Selection Methods

	Best Tree	Energy	Entropy	Fisher	MISS	SGS
bio28	0.0766	0.0688	0.0766	0.0779	0.0675	0.0649
db10	0.0831	0.0831	0.0688	0.0792	0.0714	0.0481
haar	0.0809	0.0603	0.0809	0.0765	0.0750	0.0603
bio28 and db10	0.1116	0.0785	0.0872	0.0802	0.0767	0.0488
bio28 and haar	0.1228	0.0632	0.0842	0.0895	0.0596	0.0368
haar and db10	0.1029	0.0663	0.0837	0.0767	0.0593	0.0558

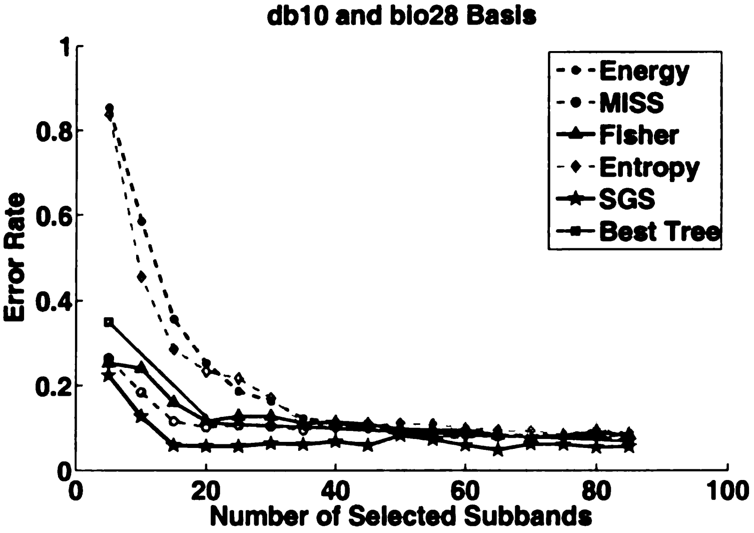


Figure 2.9. The error rates with six subband selection algorithms by selecting subbands from the 'bio28' basis and the 'db10' basis.

This was used to explain why "haar" basis performed better than the Daubechies wavelets for texture classification in [5]. Another empirical conclusion drawn from experimental results in [5] is that even-length biorthogonal filters perform better than odd-length ones. These conclusions also support our finding that the combination of "bio28" and "haar" bases achieves the best performance. Another support comes from the difference of the spaces spanned by the wavelet functions of "haar" and "bio28". As shown in Figure 2.12, the two wavelet functions have quite different shape and therefore the combination of the two is able to capture a variety of structures in texture images such as smooth curves and sharp discontinuities.

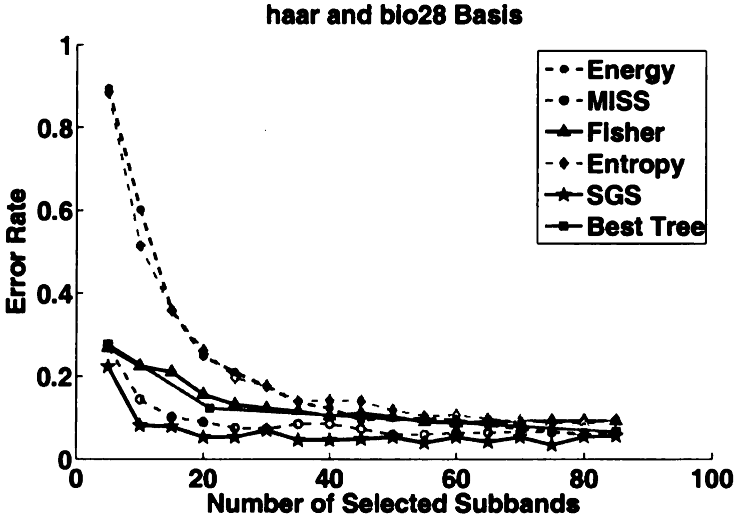


Figure 2.10. The error rates with six subband selection algorithms by selecting subbands from the 'Haar' basis and the 'bio28' basis.

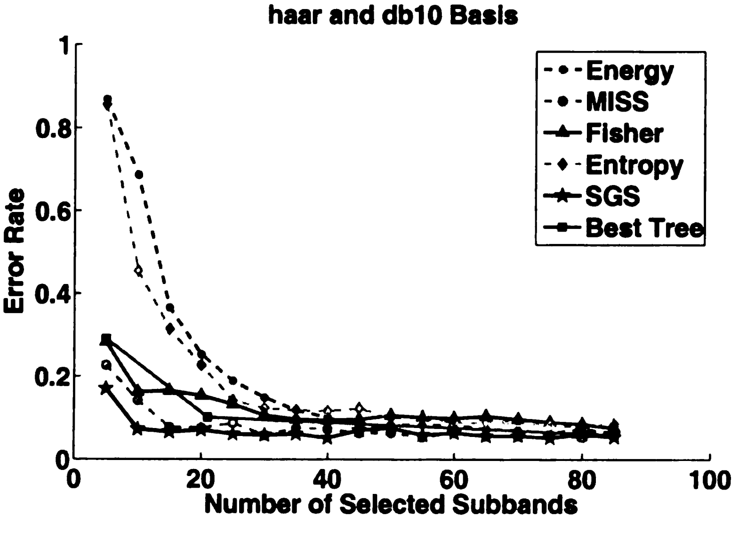


Figure 2.11. The error rates with six subband selection algorithms by selecting subbands from the 'Haar' basis and the 'db10' basis.

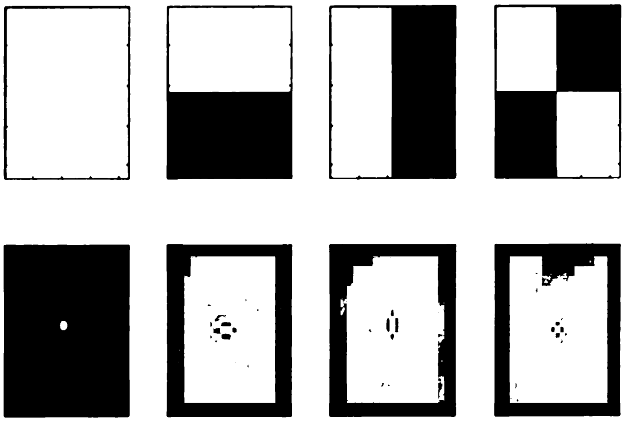


Figure 2.12. The wavelet basis functions at the first decomposition level for "haar" and "bio28", in the first row and second row, respectively. In this figure, white regions correspond to high values and dark regions correspond to low values.

2.6.4 The Clustering of Subbands in the SGS Algorithm

In order to understand why SGS performs the best compared to other methods, it is important to explore the distribution of subbands among clusters. The distribution of subbands in different clusters determined by the SGS algorithm reflects the similarity of these subbands in identifying different texture structures. As shown by the discussions and simulation results in Section 2.3, the choice of the wavelet basis and the statistical distribution of pixels in texture images jointly determine the dependence relations between energy values from different subbands. Therefore, it should be pointed out that the subband grouping results are dependent on the particular wavelet basis and the texture images. However, if a certain pattern frequently appears in the subband grouping, such a pattern may reveal an inherent link among different subbands. To make this analysis easier for visualization, we analyze the grouping results with 2-level wavelet packet decomposition, i.e., there are 21 subbands in total. The grouping algorithm is run 1000 times. Each time half of the

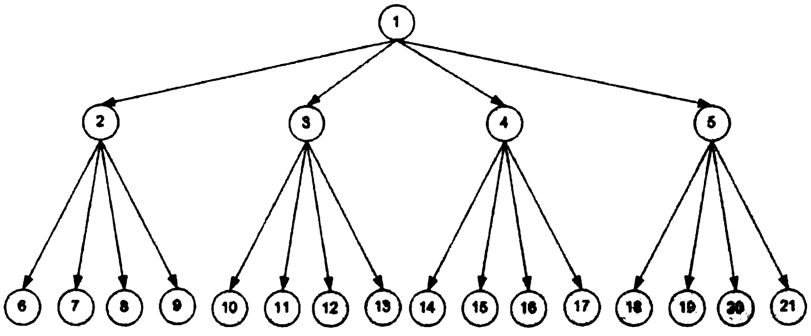


Figure 2.13. The subband labelling scheme for a 2-level wavelet packet decomposition.

864 texture images are randomly selected for decomposition and the energy features are used for grouping. When the number of clusters is set to 2, 998 out of the 1000 runs give the same grouping results: All approximation subbands (subband $1, 2$ and 6 in Figure 2.13) are in one cluster and the rest of the subbands are in another cluster. This result indicates that subbands corresponding to the same spatial frequency range usually have higher degree of dependence and tend to be grouped into the same cluster.

We further change the number of clusters in the K-Means algorithm to 4 and conduct the grouping algorithm. The number of clusters is set to 4 because there are 4 types of subbands in wavelet packet decomposition: LL, LH, HL, HH. The resulting subband clusters are used to build a co-occurrence C_M with a size of 21×21, where the value at the entry (i, j) is the number of times the i-th and j-th subbands are in the same cluster, with the subband labelling scheme given in Figure 2.13. Since the grouping algorithm is run 1000 times, the maximum value in the matrix is 1000 and the minimum value is 0. The phenomenon observed in the 2-cluster case appears again, i.e., the approximation subbands are always grouped in the same cluster with $C_M(1,2), C_M(1,6)$ and $C_M(2,6)$ larger than 990 and other co-occurrence values in rows corresponding to subbands $1, 2, 6$ are small (less than 20). Some subbands corresponding to similar spatial frequency are grouped into the same cluster most of the time. For example, $C_M(7,10), C_M(8,14), C_M(9,18), C_M(12,15), C_M(13,19), C_M(17,20)$ all have value larger than 990, which means that more than 99% of the time, these subbands are grouped into the same cluster. The common property of these subband pairs is that the paths from the root to the two subbands are swapped. For example, the path from the root to subband 7 is "LL+LH" and the path to subband 10 is "LH+LL". The empirical analysis of the grouping results is consistent with the filtering structure of the wavelet packet analysis and verifies the validity of the proposed SGS algorithm.

2.6.5 Performance on Unseen Data

The experimental results discussed so far are based on the setting that the training and testing subimages are cropped from different regions in the same large texture image. Although this experimental setting is commonly used for conducting experiments on texture classification, such as in [1,5,23,25,26], having training and testing texture images from the same large image may cause bias in the classification result. Therefore, experiments are conducted for studying the performance of the proposed algorithms on unseen data to see if the proposed algorithms are still effective. For this purpose, 9 large texture images are selected and divided into 3 classes, as shown in Figure 2.14, with one class in each row. Each of these 512 × 512 images is divided into 16 non-overlapping 128 × 128 subimages. Subimages from the first larger image in each row are used for training and the subimages from the other two images are used for testing. In this case, the training set includes 48 samples and the testing set includes 96 samples. The wavelet basis "bio28" is selected for decomposing the images. Other experimental settings are the same as in the previous sections. Given these conditions, the classification results are shown in Figure 2.15. For comparison, experiments are also conducted by using subimages from each large texture image in the training set. In this case, 5 subimages are selected from the 16 subimages of each larger texture into the training set and the remaining 11 subimages are used for testing. The classification results of this setting are shown in Figure 2.16. From Figure 2.15 and Figure 2.16, similar conclusions can be drawn as in the previous experiment setting, i.e., selection based on either the evaluation of individual subbands or the dependence between subbands is effective, and subband selection based on the two factors results in the best performance. The classification on the unseen data introduces some fluctuations in the performance curve, but does not change the conclusion on the comparison of different selection algorithms. The minimum error rates obtained by the two experimental settings are comparable. The major difference in the results is that a small error rate is achieved with fewer subbands in the second experimental setting (Figure 2.16) compared to the modified experiments with unseen data (Figure 2.15). The reason for the observed fluctuation in the error rates and the slight increase in the number of subbands selected is because of the fact that the tested images have more variability in the unseen data case. Experiments are also conducted with other two wavelet bases, i.e., "haar" and "db10", and with the three combinations of two wavelet bases. The results from these experiments are similar to those that use training and testing subimages from the same larger image except that the performance curve has more fluctuations. This consistence shows the applicability and validity of the proposed algorithms to a wider setting of texture classification problems.

2.6.6 Discussion

In most applications of wavelet transforms such as denoising and compression, the goal is to reconstruct the original signal/image. The perfect reconstruction goal constraints how the subbands are chosen. However, for classification, it is not required that the selected subbands can be used to reconstruct the original image. The major requirement for classification applications is that the features extracted from the selected subbands are uncorrelated and discriminative. With this criterion, the features from the selected subbands will form a compact and informative representation of the images for the purpose of classification.

The work presented in this chapter is motivated by the observation that most existing subband selection methods implicitly assume that features from different subbands are independent. MISS algorithm shows that the dependence can be used

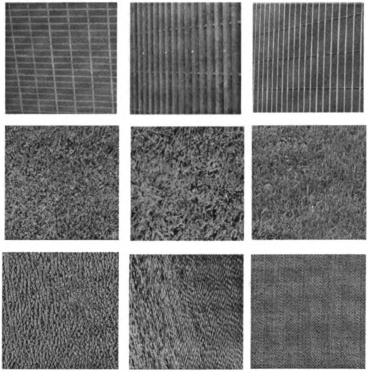


Figure 2.14. Texture images used for classification on unseen data. First row: bars; Second row: grass; Third row: knit pattern.

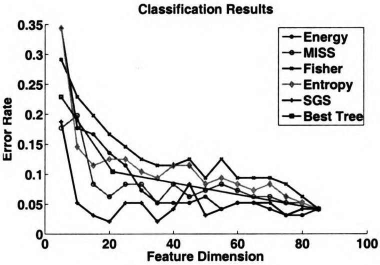


Figure 2.15. The error rates with six subband selection algorithms with the 'bio28' basis on unseen data.

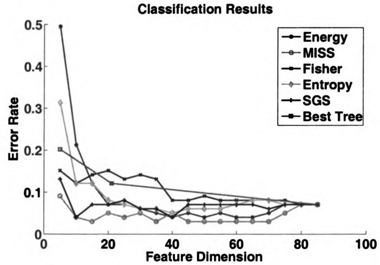


Figure 2.16. The error rates with six subband selection algorithms with the 'bio28' basis and training data from all big texture images.

for selecting a compact set of subbands for classification. However, independent features do not necessarily imply that the features are informative for classification. To select subbands that are both compact and informative, SGS algorithm is proposed to combine dependence and the evaluation of individual subbands for better subband selection. Experimental results show that SGS outperforms the algorithm based on dependence only (MISS) and the algorithms based on the evaluation of individual subbands only ("Energy", "Fisher" and "Entropy"). It has also been shown that the subband selection algorithms can be easily extended to select subbands from multiple wavelet bases.

The idea of incrementally selecting an informative subset of wavelet subbands used in the MISS algorithm bears some similarities to "feature pursuit" with "minimax entropy principle" [34]. In [34], the feature pursuit process selects a new feature that maximizes the decrease in entropy between the existing feature set and the new feature set obtained by adding the new feature to the existing feature set. The feature pursuit is proposed for texture synthesis with a parametric probability model, while the objective of the MISS algorithm is texture classification. This difference directly results in the difference in the criterion used in the feature pursuit and the MISS algorithm. For the feature pursuit, the change in entropy values of different feature sets is used for selecting new features, whereas in the MISS algorithm, the new subband is selected based on change in the mutual information.

It is also important to note that different experimental settings for the texture classification can affect the classification accuracy. For example, the distance measure, changes in the texture size and extracting features other than energy from wavelet subbands may all affect the classification accuracy. However, the focus of this chapter is not to propose a new method that can outperform the state-of-the-art texture classification methods, but rather to investigate one factor (dependence) that affects subband selection and incorporate it into the selection process. The proposed subband selection method is applicable to filterbank based classification of other signal and texture.

2.7 Conclusion

In this chapter, the dependence between features extracted from wavelet packet decomposition is investigated and incorporated into subband selection for classification. The traditional methods implicitly assume independence among features extracted from different subbands or only consider the dependence between the parent subband and the corresponding children subbands. We first demonstrate that the dependence between features extracted from different subbands exist by theoretical analysis and simulation. An algorithm exploiting the dependence, MISS, is proposed for subband selection. Experimental results show that dependence among features from different subbands is effective for subband selection. By exploiting the dependence among features from different subbands and incorporating subband selection based on individual evaluation of subbands, SGS has been shown to be more effective in subband selection. Experimental results indicate that SGS can effectively select smaller number of subbands to achieve lower classification error rates than existing subband selection methods.

CHAPTER 3

A Sparse Representation Framework for Signal Classification

3.1 Background

Sparse representations of signals have received a great deal of attentions in recent years [15–18, 35, 58, 59]. Sparse representations address the problem of finding the most compact representation of a signal in terms of a linear combination of atoms in an overcomplete dictionary. Recent developments in multi-scale and multi-orientation representation of signals, such as wavelet, ridgelet, curvelet and contourlet transforms are an important incentive for the research on sparse representations. Compared to methods based on orthonormal transforms or direct time domain processing, sparse representations usually offer better performance with their capacity for efficient signal modelling. Current research has focused on three aspects of sparse representations: pursuit methods for solving the optimization problem, such as FOCUSS [60], matching pursuit [15], orthogonal matching pursuit [16], basis pursuit [17], LARS/homotopy methods [61]; design of the dictionary, such as the K-SVD method [12]; the applications of sparse representations for different tasks, such as signal separation, denoising, coding, image inpainting [18, 19, 35, 58, 59, 62]. For instance, in [59], sparse representations are used for image separation. The overcomplete dictionary is generated by combining multiple standard transforms, including the curvelet transform, ridgelet transform and discrete cosine transform. In [35, 58], application of sparse representations to blind source separation is discussed and experimental results on EEG data analysis are demonstrated. In [62], a sparse image coding method with the wavelet transform is presented. In [18], sparse representation with an adaptive dictionary is shown to have state-of-the-art performance in image denoising. The widely used shrinkage method for image denoising is shown to be the first iteration of basis pursuit that solves the sparse representation problem [19].

In the standard framework of sparse representations, the objective is to reduce the signal reconstruction error with as few number of atoms as possible. On the other hand, discriminative analysis methods, such as LDA, are more suitable for the tasks of classification. However, discriminative methods are usually sensitive to corruption in signals due to a lack of crucial properties for signal reconstruction. In this thesis, we propose a method for sparse representations for signal classification (SRSC), which modifies the standard sparse representation framework for signal classification. We first show that replacing the reconstruction error with discrimination power in the objective function of the sparse representation is more suitable for the tasks of classification. When the signal is corrupted, the discriminative methods may fail since the information contained in the discriminative analysis may be easily distorted by noise, missing data and outliers. To address this robustness problem, the proposed SRSC approach combines discrimination power, signal reconstruction and sparsity in the objective function for classification. Our ultimate goal is to achieve a sparse and robust representation of signals in noise for effective classification.

3.1.1 Problem Formulation

The problem of finding the sparse representation of a signal in a given overcomplete dictionary can be formulated as follows. Given a $N \times M$ matrix $\mathbf{A}$ containing the elements of an overcomplete dictionary in its columns, with $M > N$ and usually $M >> N$, and a signal $\mathbf{y} \in R^N$, the problem of sparse representation is to find an $M \times 1$ coefficient vector $\mathbf{x}$, such that $\mathbf{y} = \mathbf{A}\mathbf{x}$ and $\|\mathbf{x}\|_0$ is minimized, i.e.,

$$\mathbf{x} = \min_{\mathbf{x}'} \|\mathbf{x}'\|_0 \quad \text{s.t.} \quad \mathbf{y} = \mathbf{A}\mathbf{x}. \tag{3.1}$$

where $\|\mathbf{x}\|_0$ is the ℓ_0 norm and is equivalent to the number of non-zero components in the vector $\mathbf{x}$. Finding the solution to equation (3.1) is NP hard due to its nature of combinatorial optimization. Suboptimal solutions to this problem can be found by iterative methods like the matching pursuit and orthogonal matching pursuit. An approximate solution is obtained by replacing the ℓ_0 norm in equation (3.1) with the ℓ_1 norm, as follows:

$$\mathbf{x} = \min_{\mathbf{x}'} \|\mathbf{x}'\|_1 \quad \text{s.t.} \quad \mathbf{y} = \mathbf{Ax}. \tag{3.2}$$

where $\|\mathbf{x}\|_1$ is the ℓ_1 norm. In [63], it is proved that if certain conditions on the sparsity is satisfied, i.e., the solution is sparse enough, the solution of equation (3.1) is equivalent to the solution of equation (3.2), which can be efficiently solved by basis pursuit using linear programming. A generalized version of equation (3.2), which allows noise, is to find $\mathbf{x}$ such that the following objective function is minimized:

$$J_1(\mathbf{x}; \lambda) = \|\mathbf{y} - \mathbf{Ax}\|_2^2 + \lambda \|\mathbf{x}\|_1 \tag{3.3}$$

where the parameter $\lambda > 0$ is a scalar regularization parameter that balances the tradeoff between reconstruction error and sparsity. In [64], a Bayesian approach is proposed for learning the optimal value for λ. The problem formulated in equation (3.3) has an equivalent interpretation in the framework of Bayesian decision as follows [65]. The signal $\mathbf{y}$ is assumed to be generated by the following model:

$$\mathbf{y} = \mathbf{Ax} + \varepsilon \tag{3.4}$$

where ε is white Gaussian noise. Moreover, the prior distribution of $\mathbf{x}$ is assumed to be super-Gaussian:

$$p(\mathbf{x}) \sim \exp\left(-\lambda \sum_{i=1}^{M} |x_i|^p\right) \tag{3.5}$$

where $p \in [0, 1]$. This prior has been shown to encourage sparsity in many situations, due to its heavy tails and sharp peak. Given this prior, maximum *a posteriori* (MAP) estimation of $\mathbf{x}$ is formulated as

$$\mathbf{x}_{MAP} = \arg\max_{\mathbf{x}} p(\mathbf{x}|\mathbf{y}) = \arg\min_{\mathbf{x}} -\log p(\mathbf{y}|\mathbf{x}) - \log p(\mathbf{x}) = \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_p \tag{3.6}$$

when $p = 0$, equation (3.6) is equivalent to the generalized form of equation (3.1); when $p = 1$, equation (3.6) is equivalent to equation (3.2).

3.1.2 Reconstruction and Discrimination

Sparse representations work well in applications where the original signal $\mathbf{y}$ needs to be reconstructed as accurately as possible, such as denoising, image inpainting and coding. However, for applications like signal classification, it is more important that the representation is discriminative for the given signal classes than a small reconstruction error.

The difference between reconstruction and discrimination has been widely investigated in literature. It is known that typical *reconstructive* methods, such as principal component analysis (PCA) and independent component analysis (ICA), aim at obtaining a representation that enables sufficient reconstruction, thus are able to deal with signal corruption, i.e., noise, missing data and outliers. On the other hand, *discriminative* methods, such as LDA [10], generate a signal representation that maximizes the separation of distributions of signals from different classes. While both methods have broad applications in classification, the discriminative methods have often outperformed the reconstructive methods for the classification task [11,66]. However, this comparison between the two types of methods assumes that the signals being classified are ideal, i.e., noiseless, complete(without missing data) and without outliers. When this assumption does not hold, the classification may suffer from the nonrobust nature of the discriminative methods that contains insufficient information to successfully deal with signal corruptions. Specifically, the representation provided by the discriminative methods for optimal classification does not necessarily contain sufficient information for signal reconstruction, which is necessary for removing noise, recovering missing data and detecting outliers. This performance degradation of discriminative methods on corrupted signals is evident in the examples shown in [67]. On the other hand, reconstructive methods have shown successful performance in addressing these problems. In [18], the sparse representation is shown to achieve state-of-the-art performance in image denoising. In [68], missing pixels in images are successfully recovered by inpainting method based on sparse representation. In [67, 69], PCA method with subsampling effectively detects and excludes outliers for the following LDA analysis.

All of these examples motivate the design of a new signal representation that combines the advantages of both reconstructive and discriminative methods to address the problem of *robust classification* when the signals are corrupted. The proposed method should generate a representation that contains discriminative information for classification, representative information for signal reconstruction and should be sparse. In current research [18, 68], sparse representations have proven to be effective for signal reconstruction. Existing pursuit methods provide an efficient way to solve the sparse representation problem. Therefore, we choose the sparse representation as the basic framework for the SRSC and incorporate a measure of discrimination power into the objective function. With this new method, the sparse representation obtained by the proposed method contains both crucial information for reconstruction and discriminative information for classification, which enable a reasonable classification performance in the case of corrupted signals. The three objectives: sparsity, reconstruction and discrimination may not always be consistent. Therefore, weighting factors are introduced to adjust the tradeoff among these objectives, as the weighting factor λ in equation (3.3). It should be noted that the aim of SRSC is not to improve the standard discriminative methods like LDA in the case of ideal signals, but to achieve comparable classification results with a few number of features when the signals are corrupted. A recent work [67] that aims at robust classification shares some common ideas with the proposed SRSC. In [67], PCA with subsampling proposed in [69] is applied to detect and exclude outliers in images and the rest of the pixels are used for computing LDA vectors.

3.2 Sparse Representation for Signal Classification

In this section, the SRSC problem is formulated mathematically and a pursuit method is proposed to optimize the objective function. We first replace the term measuring reconstruction error with a term measuring discrimination power to show the different effects of reconstruction and discrimination. Further, we incorporate measure of discrimination power in the framework of standard sparse representations to effectively address the problem of classifying corrupted signals. The Fisher's discrimination criterion [10] used in the LDA is applied to quantify the discrimination power. Other well-known discrimination criteria can easily be substituted.

3.2.1 Problem Formulation

Given an $N \times 1$ signal $\mathbf{y}$ and its $M \times 1$ projection on an $N \times M(M > N)$ overcomplete dictionary $\mathbf{A}$, i.e., $\mathbf{y} = \mathbf{A}\mathbf{x}$, we view $\mathbf{x}$ as the feature extracted from signal $\mathbf{y}$ for classification. The extracted feature should be as discriminative as possible between the different signal classes. Suppose that we have a set of K signals $\{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_K\}$, with representations in the overcomplete dictionary as $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_K\}$, of which K_i samples are in the class C_i, for $1 \leq i \leq C$. Mean $\mathbf{m_i}$ and variance s_i^2 for class C_i are computed in the feature space as follows:

$$\mathbf{m_i} = \frac{1}{K_i} \sum_{j \in C_i} \mathbf{x_j} \ , \quad s_i^2 = \frac{1}{K_i} \sum_{j \in C_i} \|\mathbf{x_j} - \mathbf{m_i}\|_2^2 \tag{3.7}$$

The mean of all samples are defined as: $\mathbf{m} = \frac{1}{K} \sum_{i=1}^{K} \mathbf{x}_i$. Finally, the Fisher's discrimination power can be defined as:

$$F(\mathbf{X}) = \frac{S_B}{S_W} = \frac{\left\| \sum_{i=1}^{C} K_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T \right\|_2^2}{\sum_{i=1}^{C} s_i^2}. \tag{3.8}$$

The difference between the sample means $S_B = \left\| \sum_{i=1}^{C} K_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T \right\|_2^2$ can be interpreted as the 'inter-class distance' and the sum of variance $S_W = \sum_{i=1}^{C} s_i^2$ can be similarly interpreted as the 'inner-class scatter'. Fisher's criterion is motivated by the intuitive idea that the discrimination power is maximized when the spatial distribution of different classes are as far away as possible from each other and the spatial distribution of samples from the same class are as close as possible to each other.

Let $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_K]$ be the signal matrix and $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_K]$ be the feature matrix. Replacing the reconstruction error with the discrimination power, the objective function that focuses only on classification can be written as:

$$J_2(\mathbf{X}, \lambda) = F(\mathbf{X}) - \lambda \sum_{i=1}^{K} \|\mathbf{x}_i\|_0 \tag{3.9}$$

where λ is a positive scalar weighting factor chosen to adjust the tradeoff between discrimination power and sparsity. Maximizing $J_2(\mathbf{X}, \lambda)$ generates a sparse representation that has a good discrimination power. When the discrimination power, reconstruction error and sparsity are combined, the objective function can be written as:

$$J_3(\mathbf{X}, \lambda_1, \lambda_2) = F(\mathbf{X}) - \lambda_1 \sum_{i=1}^{K} \|\mathbf{x}_i\|_0 - \lambda_2 \sum_{i=1}^{K} \|\mathbf{y}_i - \mathbf{A}\mathbf{x}_i\|_2^2 \tag{3.10}$$

where λ_1 and λ_2 are positive scalar weighting factors chosen to adjust the tradeoff between the discrimination power, sparsity and the reconstruction error. Maximizing $J_3(\mathbf{X}, \lambda_1, \lambda_2)$ ensures a representation with high discrimination power, low reconstruction error and sparsity for robust classification of corrupted signals. In the case that the signals are corrupted, the two terms $\sum_{i=1}^{K} \|\mathbf{x}_i\|_0$ and $\sum_{i=1}^{K} \|\mathbf{y}_i - \mathbf{A}\mathbf{x}_i\|_2^2$ robustly recover the signal structure, as in [18, 68]. On the other hand, the inclusion of the term $F(\mathbf{X})$ requires that the obtained representation contains discriminative information for classification. In the following discussions, we refer to the solution of the objective function $J_3(\mathbf{X}, \lambda_1, \lambda_2)$ as the features for the proposed SRSC.

3.2.2 Problem Solution

Both the objective function $J_2(\mathbf{X}, \lambda)$ defined in equation (3.9) and the objective function $J_3(\mathbf{X}, \lambda_1, \lambda_2)$ defined in equation (3.10) have similar forms to the objective function defined in the standard sparse representation, $J_1(\mathbf{x}; \lambda)$ in equation (3.3). However, the key difference is that the evaluation of $F(\mathbf{X})$ in $J_2(\mathbf{X}, \lambda)$ and $J_3(\mathbf{X}, \lambda_1, \lambda_2)$ involves not only a single sample, as in $J_1(\mathbf{x}; \lambda)$, but also all other samples. Therefore, not all the pursuit methods, such as basis pursuit and LARS/Homotopy methods, that are applicable to the standard sparse representation problem can be directly applied to optimize $J_2(\mathbf{X}, \lambda)$ and $J_3(\mathbf{X}, \lambda_1, \lambda_2)$. However, the iterative optimization methods employed in the matching pursuit and the orthogonal matching pursuit can be used in the optimization of $J_2(\mathbf{X}, \lambda)$ and $J_3(\mathbf{X}, \lambda_1, \lambda_2)$. In this thesis, we propose an algorithm similar to the orthogonal matching pursuit inspired by the simultaneous sparse approximation algorithm described in [70,71]. The pursuit algorithm for optimizing $J_3(\mathbf{X}, \lambda_1, \lambda_2)$ can be summarized as follows:

1. Initialize the residue matrix $\mathbf{R}_0 = \mathbf{Y}$ and the iteration counter $t = 0$.

2. Choose the atom from the dictionary, $\mathbf{A}$, that maximizes the objective function:

$$\mathbf{g_t} = \text{argmax}_{\mathbf{g} \in \mathbf{A}} J_3(\mathbf{g}^T \mathbf{R_t}, \lambda_1, \lambda_2) \tag{3.11}$$

3. Denote the collection of selected basis as $\mathbf{G} = [\mathbf{g_1}, \mathbf{g_2}, ..., \mathbf{g_t}]$, apply the Gram-Schmidt orthonormalization on $\mathbf{G}$ to get $\mathbf{G_1}$.

4. Determine the orthogonal projection matrix $\mathbf{P}_t$ onto the span of the chosen atoms, i.e., $\mathbf{P}_t = \mathbf{G_1}\mathbf{G_1}^T$, and compute the new residue.

$$\mathbf{R}_t = \mathbf{Y} - \mathbf{P}_t\mathbf{Y} \tag{3.12}$$

5. Increment t and return to Step 2 until t is less than or equal to a pre-determined number.

The algorithm is recursive, so the sparsity term $\lambda_1 \sum_{i=1}^{N} \|\mathbf{x}_i\|_0$ is implicitly considered in this process. The pursuit algorithm for optimizing $J_2(\mathbf{X}, \lambda)$ also follows the same steps. Note that as other greedy pursuit methods, the optimality of this pursuit algorithm is not guaranteed. Detailed analysis of this pursuit algorithm can be found in [70, 71].

3.3 Experimental Results

In order to understand the behavior of the different cost functions, two sets of experiments are conducted. In Section 3.3.1, synthesized signals are analyzed to show the difference between the features extracted by $J_1(\mathbf{X}, \lambda)$ and $J_2(\mathbf{X}, \lambda)$, which reflects the properties of reconstruction and discrimination. In Section 3.3.2, classification on real data is shown. Random noise and occlusion are added to the original signals to test the robustness of SRSC.

3.3.1 Synthetic Example

Two simple signal classes, $f_1(t)$ and $f_2(t)$, are constructed with the Fourier basis. The signals are constructed to show the difference between the reconstructive methods and discriminative methods.

$$f_1(t) = g_1 \cos t + h_1 \sin t \tag{3.13}$$

$$f_2(t) = g_2 \cos t + h_2 \sin t \tag{3.14}$$

The random variable g_1 is uniformly distributed in the interval $[0, 5]$, and g_2 is uniformly distributed in the interval $[5, 10]$. The coefficients h_1 and h_2 are uniformly distributed in the interval $[10, 20]$. Therefore, most of the energy of the signal

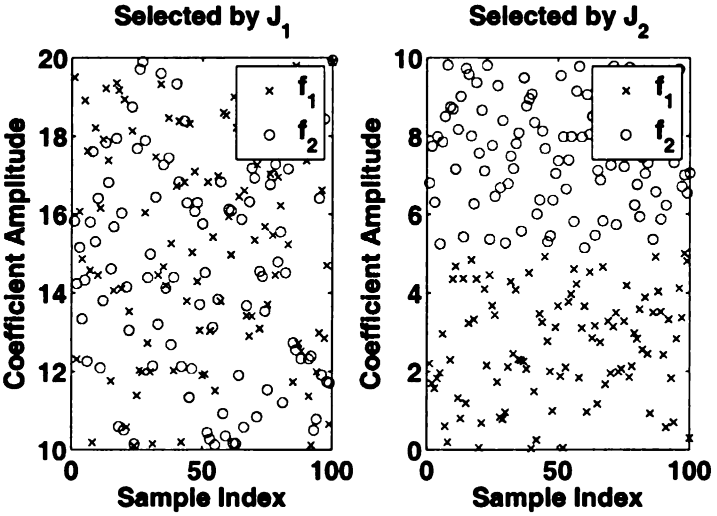


Figure 3.1. Distributions of projection of signals from two classes with the first atom selected by: $J_1(\mathbf{X}, \lambda)$ (the left figure) and $J_2(\mathbf{X}, \lambda)$ (the right figure).

can be described by the sine function and most of the discrimination power is in the cosine function. The signal component with most energy is not necessarily the component with the most discrimination power. We consider the dictionary as $\{\sin t, \cos t\}$. Optimizing the objective function $J_1(\mathbf{X}, \lambda)$ with the pursuit method described in Section 3.2 selects $\sin t$ as the first atom. On the other hand, optimizing the objective function $J_2(\mathbf{X}, \lambda)$ selects $\cos t$ as the first atom. In the simulation, 100 samples are generated for each class and the pursuit algorithm stops after the first run. The projection of the signals from both classes to the first atom selected by $J_1(\mathbf{X}, \lambda)$ and $J_2(\mathbf{X}, \lambda)$ are shown in Figure 3.1. The difference in the distribution of the two classes shown in the figures has a direct impact on the classification. Linear classification on these distributions generates a classification error rate around 0.5 for the first case and 0.0 for the second case.

3.3.2 Sparse Classification with Hand Written Digit Images

Classification with J_1, J_2 and J_3(SRSC) is also conducted on the database of USPS handwritten digits. The database contains 8-bit grayscale images of "0" through "9" with a size of 16 × 16 and there are 1100 examples of each digit. Some samples of the

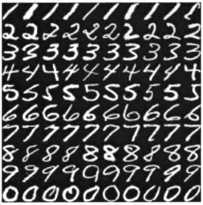


Figure 3.2. Samples of the USPS hand written digit images.

digits are shown in Figure 3.3. Following the conclusion of [72], 10-fold stratified cross validation is adopted. Classification is conducted with the decomposition coefficients (' X' in equation (3.10)) as the selected feature and support vector machine (SVM) as the classifier. The evaluation of Fisher's discrimination score implicitly assumes Gaussian distribution with equal variance for each class. Therefore, simple Euclidian distance is used to measure the distance between different samples. In this implementation, the overcomplete dictionary is a combination of Haar wavelet basis and Gabor basis. Haar basis is good at modelling discontinuities in the image, whereas the Gabor basis is good at modelling continuous image components.

In this experiment, noise and occlusion are added to the images to test the robustness of SRSC. First, white Gaussian noise with increasing levels of energy, thus decreasing levels of signal-to-noise ratio (SNR), is added to each image. Table 3.3 summarizes the classification errors obtained with different SNRs. Second, different sizes of black squares are overlayed on each image at a random location to generate occlusion (missing data). For the image size of 16×16, black squares with size of 3×3, 5×5, 7×7, 9×9 and 11×11 are overlayed as occlusion. Table 3.4 summarizes

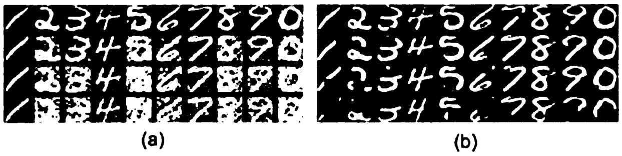


Figure 3.3. Corrupted data for the experiments: (a) with different degrees of white Gaussian noise; (b) with different sizes of occlusion.

Table 3.1. Classification errors with different levels of white Gaussian noise

	Noiseless	$20dB$	$15dB$	$10dB$	$5dB$
J_1	0.0855	0.0975	0.1375	0.1895	0.2310
J_2	**0.0605**	0.0816	0.1475	0.2065	0.2785
J_3	0.0727	**0.0803**	**0.1025**	**0.1490**	**0.2060**

the classification errors obtained with occlusion.

In Table 3.3 and Table 3.4, the minimum error rate at each noise level is highlighted with bold font. Results in Table 3.3 and Table 3.4 show that in the case that the signals are ideal (without missing data and noiseless) or nearly ideal, $J_2(\mathbf{X}, \lambda)$ is the best criterion for classification. This is consistent with the known conclusion that discriminative methods outperform reconstructive methods in classification. However, when the noise is increased or more data is missing (with larger area of occlusion), the accuracy based on $J_2(\mathbf{X}, \lambda)$ degrades faster than the accuracy based on $J_1(\mathbf{X}, \lambda)$. This indicates that the signal structures recovered by the standard sparse representation are more robust to noise and occlusion, thus yield less performance degradation. On the other hand, the SRSC demonstrates lower error rates by combing the reconstruc-

Table 3.2. Classification errors with different sizes of occlusion

	no occlusion	3 × 3	5 × 5	7 × 7	9 × 9	11 × 11
J_1	0.0855	0.0930	0.1270	0.1605	0.2020	0.2990
J_2	**0.0605**	**0.0720**	**0.1095**	0.1805	0.2405	0.3305
J_3	0.0727	0.0775	0.1135	**0.1465**	**0.1815**	**0.2590**

tion property and the discrimination power in the case that signals are corrupted.

3.3.3 Experiments with the Object Recognition

Experimental Settings

Experiments are also conducted on the classification of images from 2 different objects in the COIL20 data set [73]. The focus of this experiment is to investigate the effect of different weighting values on the cost function $J_3(\mathbf{X}, \lambda_1, \lambda_2)$. In the training stage, 12 images of each object with a size of 16 × 16 are used, and the remaining 60 images from each object are used for testing. Classification is conducted using the decomposition coefficients ('$\mathbf{X}$' in equation (3.10)) as the features and the K nearest neighborhood (KNN) classifier for the "leave-one-out" method. The Euclidian distance is used to measure the distance between features from different samples. In this implementation, the overcomplete dictionary is formed by combining Haar wavelet basis, Gabor basis and contourlet basis. Haar basis is good at modelling discontinuities in the signal. Gabor basis, on the other hand, is good at modelling continuous signal components and texture. The contourlet is a multi-resolution and multi-directional wavelet transform that can efficiently model object contours. In summary, the over-complete dictionary has 200 Gabor atoms, 441 Haar atoms and 454 contourlet atoms.

Random Gaussian noise and occlusions are added to the original images to test the performance of SRSC with different parameter settings. First, white Gaussian noise with different levels of signal-to-noise ratio (SNR), are added to each image. Second, different sizes of black squares (zero values) are overlayed on each image at the image center to generate occlusions. For the image with size 16×16, black squares with size 3×3, 5×5 and 7×7 are overlayed as occlusions. Several example images with different levels of Gaussian noise and occlusion are shown in Figure 3.4.

Since the three components of $J_3(\mathbf{X}; \lambda_1, \lambda_2, \lambda_3)$ may not be on the same scale, we first normalize the three factors before setting the weighting factors. Given a vector $\mathbf{Z} = [z_1, z_2, ..., z_M]$, where $\mathbf{Z}$ can be either one of the three items. The normal-

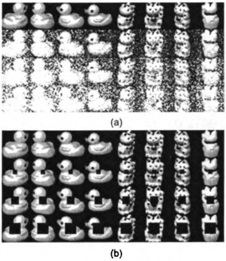


Figure 3.4. (a) Object images with different levels of white Gaussian noise; (b) Object images with different sizes of occlusion. The images in the first line of (a) and (b) are noiseless.

ization is implemented by the following equation:

$$z_i = \frac{z_i - \mu_{\mathbf{Z}}}{\sigma_{\mathbf{Z}}} \tag{3.15}$$

where $\mu_{\mathbf{Z}}$ and $\sigma_{\mathbf{Z}}$ are the mean and standard variance of **Z**, respectively.

Study on the Effect of the Different Factors

Table 3.3 summarizes the classification error rates for the different weighting factors for different SNR values. Table 3.4 summarizes the classification error rates with different sizes of occlusions. In both tables, the minimum error rate of each column is underscored. In this experiment, the pursuit algorithm stops when 547 atoms are chosen, i.e., half of the total number of atoms.

From the experimental results, we can observe several phenomena. First, in a noiseless environment, the Fisher's discrimination ratio by itself results in the lowest error rate. This result is consistent with the existing research that the discriminative

Table 3.3. Error rates with white Gaussian noise

$[\lambda_1, \lambda_2, \lambda_3]$	Noiseless	$20dB$	$15dB$	$10dB$
[0,0,1]	0.1000	0.2583	0.3083	0.3417
[1,0,0]	0.0750	0.2417	0.3417	0.3833
[1,1,0]	0.0833	0.2167	0.3000	0.3250
[0,1,1]	0.0917	0.2583	0.2750	0.3000
[1,0,1]	0.0833	0.2083	0.2833	0.3667
[1,1,1]	0.1000	0.2500	0.2667	0.3333

Table 3.4. Error rates with occlusions

$[\lambda_1, \lambda_2, \lambda_3]$	no occlusion	3×3	5×5	7×7
[0,0,1]	0.1000	0.1167	0.1500	0.2000
[1,0,0]	0.0750	0.1667	0.1750	0.2167
[1,1,0]	0.0833	0.1250	0.2417	0.3250
[0,1,1]	0.0917	0.1333	0.1583	0.2667
[1,0,1]	0.0833	0.1417	0.2250	0.3000
[1,1,1]	0.1000	0.1083	0.1333	0.2833

methods are better than the reconstructive methods for classification [11]. Second, as the signal is corrupted by Gaussian noise, or as more pixels are lost, the performance of all weighting schemes deteriorates. The different weighting schemes have different responses to noise and occlusions. In a noisy environment, the weighting vectors that combines discrimination power and reconstruction power usually achieve the best performance, as shown in the third and fourth columns of both tables. When the images are highly corrupted, such as shown in the last column of both tables, the inclusion of discriminative power actually results in a higher classification error rate. As shown in the last column of both tables, the minimum error rate is achieved with $\lambda_1 = 0$, which means that the discriminative power is not considered. These results indicate that when the signal is corrupted beyond a certain amount, the discriminative power estimated from the signal is not reliable any more, and the reconstruction error provided by the framework of sparse representations and harmonic analysis can reduce the effect of noise and missing data to improve the classification accuracy.

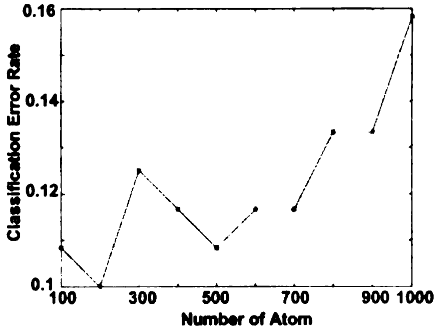


Figure 3.5. The relation between error rate and sparsity.

Study on Sparsity

The sparsity of the representation is incorporated in the objective function with the term $\lambda_2 \sum_{i=1}^{K} \|\mathbf{x}_i\|_0$. The sparsity is preferred in SRSC because it provides a compact representation and filters out noisy signal components, simultaneously. This idea is similar to that of feature selection in pattern recognition. In SRSC, the effect of sparsity can be controlled in two ways. First, we can increase the value of λ_2 in $J_3(\mathbf{X}; \lambda_1, \lambda_2, \lambda_3)$ to emphasize sparsity. Second, we can reduce the number of iterations in the pursuit algorithm, i.e., the number of selected atoms.

For the first case, we can compare the error rates for $\lambda_2 = 1$ and $\lambda_2 = 0$ in Tables 3.3 and 3.4. In most cases especially when the images are corrupted, emphasizing the sparsity of the representation by increasing λ_2 results in a more discriminative feature improving the classification accuracy. For the second case, we study the relation between sparsity and the classification error of SRSC when the images are occluded with a 3×3 square and the weighting factors are fixed: $\lambda_1 = \lambda_2 = \lambda_3 = 1$. The experimental results are shown in Figure 3.5. In this experiment, we change the number of selected atoms from 100 to 1000, with increments of 100. In this experiment, the minimum error rate is achieved with only 200 atoms from a total of 1091 atoms, i.e, with less than 20% of atoms. We also note that after 500 atoms, the error rate monotonically increases with the number of atoms. The results indicate that not all atoms provide useful information for classification. Decomposition of the signals with certain atoms actually add noise rather than improving the classification. The results also show the effectiveness of the pursuit algorithm for selecting a small number of atoms that achieve low error rate.

3.4 Conclusions

SRSC is motivated by the ongoing research in the area of sparse representations in signal processing. SRSC incorporates reconstruction error, discrimination power and sparsity for robust classification. In the current implementation of SRSC, the weight factors are empirically set to optimize the performance. Approaches to determine optimal values for the weighting factors are being investigated, following the methods similar to that introduced in [64].

The experimental results clearly show that the different parameters have significant effect on the atom selection process and correspondingly, the following feature extraction and classification performance. For example,when $\lambda_1 = 1, \lambda_2 = 0$ and $\lambda_3 = 0$, SRSC becomes a sequential version of the traditional LDA method. On the other hand, when $\lambda_1 = 0, \lambda_2 = 1$ and $\lambda_3 = 1$, SRSC becomes the traditional sparse representation. As reflected by the experimental results, a given parameter setting can not achieve the best performance for different levels of signal corruption. It is therefore interesting to investigate the optimal value for the parameters for different situations. Investigating the optimal number of iterations for the pursuit algorithm is another interesting problem, since it also has considerable effect on the classification accuracy.

CHAPTER 4

Sparse Representation for Signal Classification: An Optimization Approach

4.1 Introduction

The SRSC framework proposed in Chapter 3 presented a promising method for robust and sparse feature extraction from images for classification. However, using Fisher's ratio for measuring discrimination makes the objective function non-convex. This limitation makes it hard to find the optimal set of features. For this reason, an algorithm similar to matching pursuit was applied by selecting one atom at a time. This approach provides good classification accuracy, however, is still suboptimal. In this chapter, the SRSC method is improved by replacing the Fisher's ratio with the class margin criterion used in the support vector machine to quantify the discrimination power. Using this new measure of discrimination, the SRSC problem can be formulated as a constrained optimization problem that can be iteratively optimized with two quadratic programming problems in each iteration. In order to further reduce the computational complexity, the SRSC problem is decomposed into two steps, with the first step being sparse reconstruction and the second step being sparse classification. For the sparse classification step, a new algorithm called Large Margin Dimension Reduction (LMDR) is proposed for obtaining sparse features. The formulation of LMDR incorporates the advantages of the L1-norm SVM [74] and distance metric learning [75] into one framework by using the idea of distance metric learning to search for an optimal linear transform on the original features and using the idea of L1-norm SVM to determine significant feature components. In this way, LMDR can effectively identify the underlying features that compose the sparse feature representation for classification. LMDR is applicable to dimension reduction problems in general purpose classification, and performs better than L1-norm SVM and linear discriminative analysis (LDA) in certain circumstances.

The rest of this chapter is organized as follows. Section 4.2 briefly reviews the data representation model and defines the general notations used in this chapter. Section 4.3 proposes a new formulation of SRSC with the large margin method and its optimization with iterative quadratic programming. Dividing SRSC into two steps and employing LMDR in the second step are discussed in Section 4.4. Experiments and discussions are given in Section 4.5.

4.2 Data Model

A dictionary $\mathbf{A} \in R^{k \times d}$ is a matrix where each column corresponds to an atom. Usually the number of atoms is much larger than the dimension of each atom, i.e., $k << d$ in the traditional sparse representation problems. The set of observed signals $\mathbf{Y} = [\mathbf{y_1}, \mathbf{y_2}, ..., \mathbf{y_n}]$ is a set of k-dimensional ($\mathbf{y_i} \in R^k$) training data with class labels $\mathbf{C} = \{c_1, c_2, ..., c_n\}$ $(c_i \in Z^1)$. Based on this training data set, we want to learn the representation of $\mathbf{Y}$ using selected atoms in $\mathbf{A}$ such that $\mathbf{Y}$ is representable by the dictionary $\mathbf{A}$ and the representation also has some other desired properties. In this formulation, $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}]$ and $\mathbf{x_i} \in R^d$ is the decomposition of $\mathbf{y_i}$ on the dictionary $\mathbf{A}$. We can then use the decomposition coefficients $\mathbf{X}$ as features for classification. Mathematically, these goals can be formulated into a single objective function $G(\mathbf{X})$ that needs to be optimized. For example, in the traditional sparse representation, $G(\mathbf{X})$ includes both the signal reconstruction error and the sparsity measure. In the SRSC problem formulated in Chapter 3, $G(\mathbf{X})$ includes the signal reconstruction error, sparsity measure and discrimination measure.

4.3 SRSC with Large Margin: An Optimization Model

The large margin principle that is used in the design of support vector machines (SVM) aims at maximizing the margin between the distributions of different classes. Given linearly separable labelled features from two classes $S_l = \{(\mathbf{x_1}, c_1), (\mathbf{x_2}, c_2), ..., (\mathbf{x_n}, c_n)\}$ with $c_i \in \{-1, 1\}$, the hyperplane $(\mathbf{w}, b)$ realizes the maximum margin classifier by solving the following optimization problem:

$$\begin{aligned} &\min \quad (w.r.t. \;\; \mathbf{w}, b) \quad < \mathbf{w}, \mathbf{w} > \\ &s.t. \quad c_i(< \mathbf{w} \;, \mathbf{x_i} > +b) \geq 1 \end{aligned} \tag{4.1}$$

where the operator $<,>$ is the inner product, i.e., $< \mathbf{w}, \mathbf{x} >= \mathbf{w}^T\mathbf{x}$. An intuitive explanation of this formulation lies in the connection between the margin γ between the data distribution of two classes $\mathbf{w}$: $\gamma = \frac{1}{\|\mathbf{w}\|^2}$. An illustration of the large margin principle is shown in Figure 4.1. In the SVM method, this optimization problem can be solved by solving the dual problem of equation (4.1) with quadratic programming.

4.3.1 SRSC with Large Margin: Formulation

Similar to Fisher's ratio, the margin between two classes provides a measure for feature discrimination. One advantage of the large margin classification is that it usually has a good generalization capacity, i.e., a better performance than other classifiers on unseen data, as shown by the statistical learning theory [76, 77]. This motivates the application of the large margin principle as the discrimination measure in SRSC, forming a new objective function for SRSC:

$$\begin{aligned} &\min \quad (w.r.t. \;\; \mathbf{w}, b, \mathbf{x_i}) \quad \lambda_1 < \mathbf{w}, \mathbf{w} > + \lambda_2 \sum_{i=1}^{n} \|\mathbf{x_i}\|_1 + \lambda_3 \sum_{i=1}^{n} \|\mathbf{y_i} - \mathbf{A}\mathbf{x_i}\|_2^2 \\ &s.t. \quad c_i(< \mathbf{w} \;, \mathbf{x_i} > +b) \geq 1 \end{aligned} \tag{4.2}$$

where λ_1, λ_2 and λ_3 are non-negative weighting factors. The advantage of this opti-

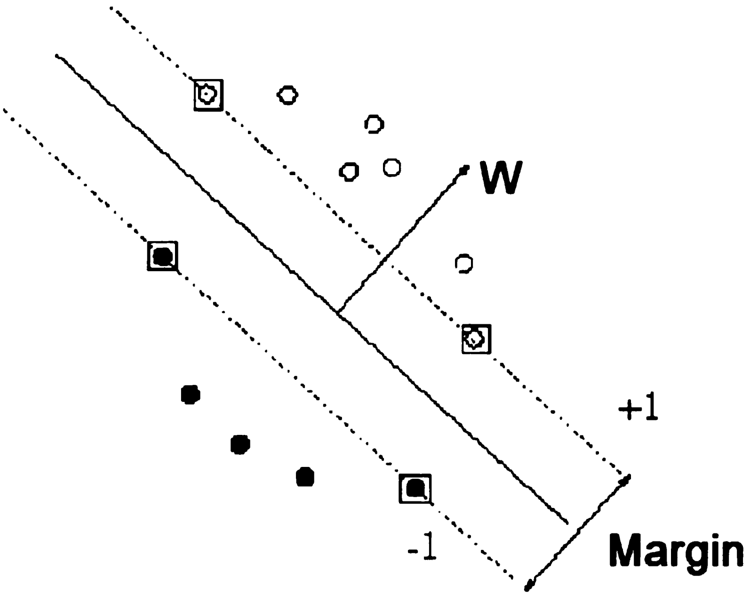


Figure 4.1. Illustration of the max-margin linear classifier.

mization formulation is that it integrates feature extraction and classifier design into a single step. Once solved, the feature is sparse, robust to noise (by incorporating the reconstruction error) and discriminative (large margin).

Since the ℓ_1 norm is not differentiable, a common trick to deal with this problem is to introduce two positive variables $\mathbf{u_i}$ and $\mathbf{v_i}$: $\mathbf{u_i}$ accounts for the positive components of $\mathbf{x_i}$; $\mathbf{v_i}$ accounts for the absolute values of the negative components of $\mathbf{x_i}$. With the two variables, $\mathbf{x_i} = \mathbf{u_i} - \mathbf{v_i}$, $\|\mathbf{x_i}\|_1 = \mathbf{1}^T(\mathbf{u_i}+\mathbf{v_i})$, $\mathbf{u_i} \geq 0$ and $\mathbf{v_i} \geq 0$. This conversion of the minimum ℓ_1 optimization to linear representation has been known as the "Least Absolute Derivation" problem. Plugging this representation into equation (4.2), the problem can be formulated as:

$$\begin{aligned} \min \quad & (w.r.t. \;\; \mathbf{w}, b, \mathbf{u_i}, \mathbf{v_i}) \quad \lambda_1 \mathbf{w}^T\mathbf{w} \; + \lambda_2 \sum_{i=1}^{n} [\mathbf{1}^T(\mathbf{u_i} + \mathbf{v_i})] + \\ & \lambda_3 \sum_{i=1}^{n} [\mathbf{y_i} - \mathbf{A}(\mathbf{u_i} - \mathbf{v_i})]^T [\mathbf{y_i} - \mathbf{A}(\mathbf{u_i} - \mathbf{v_i})] \\ s.t. \quad & c_i[\mathbf{w}^T(\mathbf{u_i} - \mathbf{v_i}) \; + b] \geq 1 \\ & \mathbf{u_i} \geq 0; \quad \mathbf{v_i} \geq 0 \end{aligned} \tag{4.3}$$

Considering that the samples are not exactly linearly separable, slack variables ξ_i are introduced. The effect of ξ_i is illustrated in Figure 4.2. For the distribution shown in Figure 4.2, the data samples from the two classes are not linearly separable and the constraints $c_i[\mathbf{w}^T(\mathbf{u_i} - \mathbf{v_i}) + b] \geq 1$ are not satisfiable. Therefore, there is no solution to the optimization problem. Slack variables, ξ_i, introduces the endurance of misclassification and extends the applicability of SVM to data that are not linearly separable. With the slack variables, the formulation can be rewritten as follows:

$$\begin{aligned} \min \quad & (w.r.t. \;\; \mathbf{w}, b, \mathbf{u_i}, \mathbf{v_i}, \xi_i) \quad \lambda_1 \mathbf{w}^T\mathbf{w} + \lambda_2 \sum_{i=1}^{n} [\mathbf{1}^T(\mathbf{u_i} + \mathbf{v_i})] + \\ & \lambda_3 \sum_{i=1}^{n} [\mathbf{y_i} - \mathbf{A}(\mathbf{u_i} - \mathbf{v_i})]^T [\mathbf{y_i} - \mathbf{A}(\mathbf{u_i} - \mathbf{v_i})] + \lambda_4 \sum_{i=1}^{n} \xi_i \\ s.t. \quad & c_i[\mathbf{w}^T(\mathbf{u_i} - \mathbf{v_i}) + b] \geq 1 - \xi_i \\ & \mathbf{u_i} \geq 0; \quad \mathbf{v_i} \geq 0; \quad \xi_i \geq 0 \end{aligned} \tag{4.4}$$

Both the objective function and the constraint of this optimization problem are in the square form. If there is only one constraint, the problem can be efficiently solved

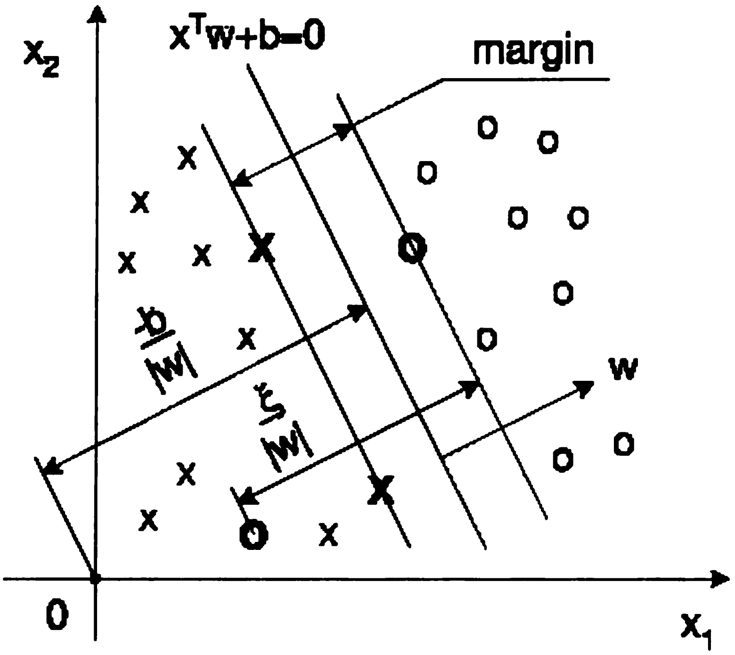


Figure 4.2. Illustration of the max-margin linear classifier with samples that are not linearly separable.

by the S-procedure (see Appendix B of [78]). However, the number of constraints is equal to the number of training samples, which is larger than 1 in all cases. This is not even a convex optimization problem, since it can be easily verified that the constraint $c_i[\mathbf{w}^T(\mathbf{u_i} - \mathbf{v_i}) + b] \geq 1$ with variables $\mathbf{w}, \mathbf{u_i}, \mathbf{v_i}$ and b is not convex. To solve this problem, we adopt the following alternating optimization approach.

4.3.2 SRSC with Large Margin: Iterative Optimization

The major source of difficulty in the optimization model in equation (4.2) comes from the objective function that integrates feature extraction (deriving $\mathbf{x_i}$ from $\mathbf{y_i}$) and classifier design ($\mathbf{w}$ and b are unknown) into a single step. Even though this integration has valid theoretical motivations, it introduces non-convexity in the optimization problem. To make the optimization tractable, an iterative optimization strategy is adopted. Mathematically, fixing the value of either $\mathbf{w}$ and $\mathbf{x_i}$ makes the optimization problem become a quadratic programming problem, which is a convex optimization problem that is tractable. The implementation of this iterative alternating optimization strategy can be formulated as:

1. Initialization: Set $\mathbf{w} = \mathbf{w_0}$ and $b = b_0$, where $\mathbf{w_0}$ and b_0 are random initial values.

2. Feature extraction step:

$$\begin{aligned}(\mathbf{u_i}, \mathbf{v_i}) = \arg\min \quad & (w.r.t. \;\; \mathbf{u_i}, \mathbf{v_i}) \;\; \{\lambda_1 \mathbf{w}^T\mathbf{w} + \lambda_2 \sum_{i=1}^{n} \left[\mathbf{1}^T(\mathbf{u_i} + \mathbf{v_i})\right] + \\ & \lambda_3 \sum_{i=1}^{n} [\mathbf{y_i} - \mathbf{A}(\mathbf{u_i} + \mathbf{v_i})]^T [\mathbf{y_i} - \mathbf{A}(\mathbf{u_i} + \mathbf{v_i})] + \lambda_4 \sum_{i=1}^{n} \xi_i\} \\ s.t. \;\; & c_i[\mathbf{w}^T(\mathbf{u_i} - \mathbf{v_i}) + b] \geq 1 - \xi_i \\ & \mathbf{u_i} \geq 0; \quad \mathbf{v_i} \geq 0; \end{aligned} \tag{4.5}$$

3. Update the parameters of the large margin classifier:

$$\begin{aligned}
(\mathbf{w}, b, \xi_i) = \arg\min \quad (w.r.t.\ \ \mathbf{w}, b, \xi_i) \quad & \{\lambda_1 \mathbf{w}^T\mathbf{w} + \lambda_2 \sum_{i=1}^{n} \left[\mathbf{1}^T(\mathbf{u_i} + \mathbf{v_i})\right] + \\
& \lambda_3 \sum_{i=1}^{n} \left[\mathbf{y_i} - \mathbf{A}(\mathbf{u_i} + \mathbf{v_i})\right]^T \left[\mathbf{y_i} - \mathbf{A}(\mathbf{u_i} + \mathbf{v_i})\right] + \lambda_4 \sum_{i=1}^{n} \xi_i\} \\
s.t. \quad c_i[\mathbf{w}^T(\mathbf{u_i} - \mathbf{v_i}) + b] \geq 1 - \xi_i & \\
\xi_i \geq 0; &
\end{aligned} \tag{4.6}$$

4. Repeat steps 2 and 3 until a given alternating step number is reached or the changes in the parameters $\mathbf{w}, b, \xi_i, \mathbf{u_i}, \mathbf{v_i}$ are less than a predefined value.

One practical problem with this approach is the computational complexity of the quadratic programming with the high-dimensional vectors $\mathbf{u_i}$ and $\mathbf{v_i}$, which have a dimension of d, i.e., the number of atoms. Setting appropriate values for the 4 weighting parameters λ_1 to λ_4 is also a nontrivial problem. Even in the traditional sparse representation where there is only one weighting parameter λ, setting the value is not trivial [64]. The classification result with inappropriate parameter setting is likely to deviate from the optimal case shown in the theoretical analysis. Preliminary numerical simulations verify our concerns on the computational complexity of the proposed method. The simulation, conducted with the YALMIP optimization toolbox [79] running in the Matlab environment on a Pentium IV PC, needs more than 10 hours to optimize a data set with 80 samples with a dictionary of 32 atoms. These practical problems in the implementation motivates a more computationally efficient solution to SRSC problem, as presented in the next section.

4.4 SRSC with Large Margin: A Two-Step Approximation

The optimization model and solution proposed in Section 4.3 for solving the SRSC problem has a solid theoretical foundation, but suffers from several difficulties in the implementation stage. The difficulties arise from the ambitious goal of integrating both signal reconstruction (decomposition of a signal over the dictionary) and dimension reduction (large margin and sparsity) into a single step, as shown in equation (4.4). To make the problem more tractable, in this section the two goals of signal reconstruction and dimension reduction are separately implemented in two steps. In the first step, sparse representation with a dictionary is used for denoising, such that noise and missing data do not affect the classification performance. In the second step, the sparsity regularization with large margin method is applied for feature selection and dimension reduction. Research on sparse reconstruction has been extensively conducted in the area of traditional sparse representation, such as [18,19,58,59,62]. In this section, the focus is on the second step, i.e., sparse feature selection and dimension reduction. For this purpose, the large margin dimension reduction (LMDR) method is proposed for the implementation of the second step. LMDR is based on combining the idea of the L1-norm SVM and distance metric learning. In the rest of this chapter, Section 4.4.1 introduces the formulation and application of the L1-norm SVM; Section 4.4.2 reviews the idea of the distance metric learning; Section 4.4.3 presents the proposed LMDR method and finally Section 4.5 conducts experimental studies.

4.4.1 L1-norm Support Vector Machine

Suppose that the features and class labels are already available for a 2-class supervised learning task. $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}\} \in R^{d \times n}$ is the collection of features and $\mathbf{C} = \{c_1, c_2, ..., c_n\} \in R^{1 \times d}$, where $c_i \in \{-1, +1\}$, is the collection of labels. In SRSC, sparsity is enforced on the feature set $\mathbf{X}$ for deriving sparse and discriminative features. One key observation is that enforcing sparsity on a feature vector $\mathbf{x_i}$ is equivalent to enforcing sparsity on the coefficient vector $\mathbf{w}$ in a linear classifier, as in the SVM. In the context of large margin linear classifiers, this is exactly what the L1-norm SVM [74] does. Formally, the L1-norm SVM can be formulated with the following optimization problem:

$$\begin{aligned} &\min \quad (w.r.t. \;\; \mathbf{w}, \xi_i) \quad \|\mathbf{w}\|_1 + \lambda \sum_{i=1}^{n} \xi_i \\ &s.t. \quad c_i(\mathbf{w}^T \mathbf{x_i} + b) \geq 1 - \xi_i \\ &\quad \xi_i \geq 0 \end{aligned} \tag{4.7}$$

Compared to the normal SVM, L1-norm SVM replaces the L2-norm, $\|\mathbf{w}\|_2$, with the L1-norm, $\|\mathbf{w}\|_1$, which enforces sparsity [17,74]. In this setting, enforcing sparsity is equivalent to feature selection. If a component of $\mathbf{w}$, w_j, has a value close to 0 that indicates that the jth component of the feature vector x_{ij} is uninformative for classification and can be removed for classification. This paradigm of feature selection has been applied to various applications, such as medical structure classification, content based image classification [80–82]. The modification from L2-norm to L1-norm also makes the optimization problem solvable by linear programming, whereas quadratic programming is required to solve the L2-norm SVM problem. Note that with this feature selection strategy, kernel method can not be used, since there is no one-to-one correspondence between the feature component and the weighting components in the kernel method [80–82]. To see this clearly, note that for the kernel method the decision function is given by $\mathbf{w}^T\Phi(\mathbf{x_i}) + b$, where $\Phi(\bullet)$ is the kernel function. Therefore, the components of $\mathbf{w}$ do not directly correspond to the components of the feature vector $\mathbf{x_i}$.

4.4.2 Distance Metric Learning

The dimension reduction with L1-norm SVM is achieved through feature selection with classifier coefficients and not through any transforms on the original feature space. However, feature selection in a transformed feature space may yield better results in terms of dimension reduction. To see this point, consider the simple example illustrated in Figure 4.3, where the distribution of two dimensional features from two classes are plotted. Figure 4.3 (a) shows the original distribution of the two classes and Figure 4.3 (b) is obtained by rotating the feature values by 45 degrees in the $2D$

space. In the distribution shown in Figure 4.3 (a), L1-norm SVM will select both of the feature components, which are almost equally important for classification. However, with the rotation the L1-norm SVM will only select one feature component in the new feature representation, since the feature in the direction of $x1$ is much more important than $x2$.

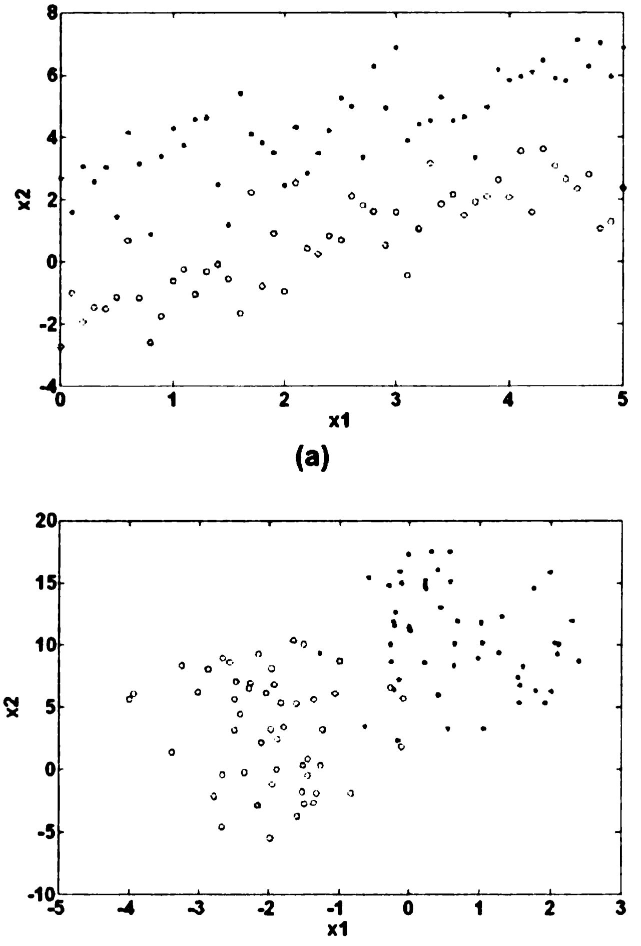
The idea of applying a transform on the original feature before selection and classification is widely adopted in the field of machine learning and pattern recognition. Typical examples of feature transform include PCA, LDA and more recently the distance metric learning [75, 83–85]. The basic idea of distance metric learning is to search for a linear transform of the original features such that in the transformed feature space the samples from the same class are close to each other and samples from different classes are far away from each other. In this sense, distance metric learning shares the same motivation as LDA. The difference is that distance metric learning adopts a different mathematical formulation to achieve this goal [75]:

$$\begin{aligned} \min_{\mathbf{A}} \quad & \textstyle\sum_{(\mathbf{x_i},\mathbf{x_j})\in S} (\mathbf{x_i}-\mathbf{x_j})^T\mathbf{A}(\mathbf{x_i}-\mathbf{x_j}) \\ s.t. \quad & \textstyle\sum_{(\mathbf{x_i},\mathbf{x_j})\in D} (\mathbf{x_i}-\mathbf{x_j})^T\mathbf{A}(\mathbf{x_i}-\mathbf{x_j}) \geq 1 \\ & \mathbf{A} \succ= 0 \end{aligned} \tag{4.8}$$

The requirement that $\mathbf{A} \succ= 0$, i.e., $\mathbf{A}$ is semi-definite positive, is necessary since negative distances between feature vectors is not allowed. In this formulation, S is the set of feature pairs $(\mathbf{x_i}, \mathbf{x_j})$ from the same class and D is the set of feature pairs $(\mathbf{x_i}, \mathbf{x_j})$ from different classes. $\mathbf{A}$ is a semi-definite positive matrix to be optimized such that in the transformed space, samples from the same class are close to each other subject to the constraint that the distances between samples from different classes are larger than a fixed value. Denoting $\mathbf{A} = \mathbf{B}^T\mathbf{B}$, where $\mathbf{B} = \mathbf{A}^{\frac{1}{2}}$, the distance between two feature vectors can be represented as:

$$(\mathbf{x_i}-\mathbf{x_j})^T\mathbf{A}(\mathbf{x_i}-\mathbf{x_j}) = (\mathbf{B}(\mathbf{x_i}-\mathbf{x_j}))^T(\mathbf{B}(\mathbf{x_i}-\mathbf{x_j})) \tag{4.9}$$

Therefore, the distance metric learning problem is equivalent to finding a linear



(a)

(b)

Figure 4.3. Illustration of the effect of transform for feature selection: (a) the original data distribution from two classes; (b) the distribution of the data after rotation.

transform on the original features with the matrix $\mathbf{B}$ and then using the Euclidian distance on the transformed features. Application of the linear transform $\mathbf{B}$ can be viewed as extracting new features from the original feature space. The successful application of the distance metric learning on some general classification and clustering tasks, such as classifying the data sets in the UCI data set [86], indicates that transforming the original features helps to derive more discriminative features. This property is useful for dimension reduction.

4.4.3 Large Margin Dimension Reduction

As discussed in Sections 4.4.1 and 4.4.2, the L1-norm SVM and distance metric learning have different properties in feature selection and transform. L1-norm SVM can identify useful feature components but only operates on the original feature space. Distance metric learning searches for a linear transform that optimizes the feature distribution in the transform space with $\mathbf{B} \in R^{d \times d}$ and does not reduce the feature dimension. An intuitive generalization is to incorporate the dimension reduction into the distance metric learning by assuming $\mathbf{B} \in R^{d \times d'}$, where $d' < d$. However, determining an appropriate value for d' is data dependent and nontrivial.

In this section, large margin dimension reduction algorithm (LMDR) is proposed to combine the advantages of both L1-norm SVM and distance metric learning. The basic idea is to search for a linear transform on the original feature space and then apply the L1-norm SVM to the transformed feature space for dimension reduction. Formally, this problem can be formulated as follows:

$$\begin{aligned}
&\min \quad (w.r.t. \;\; \mathbf{w}, b, \mathbf{B}, \xi_i) \;\; \|\mathbf{w}\|_1 + \lambda \sum_{i=1}^{n} \xi_i \\
&s.t. \;\; c_i(\mathbf{w}^T \mathbf{z_i} + b) \geq 1 - \xi_i \\
&\quad \mathbf{z_i} = \mathbf{B}\mathbf{x_i} \\
&\quad \xi_i \geq 0
\end{aligned} \tag{4.10}$$

where $\mathbf{B} \in R^{d \times d}$ and $\mathbf{z_i} = \mathbf{B}\mathbf{x_i}$ is the feature in the transformed space. Selection on the components of $\mathbf{z_i}$ is based on the value of $\mathbf{w}$, i.e., components corresponding to $w_j \sim 0$

are removed. Note that this formulation is ill-posed. Given a solution $\mathbf{w}'$ and $\mathbf{B}'$ to the optimization problem, a better solution can always be found by letting $\mathbf{w}'' = \mathbf{w}'/a$ and $\mathbf{B}'' = a\mathbf{B}'$, where $a > 1$. The solution $(\mathbf{w}'', \mathbf{B}'')$ satisfies the constraints but has a smaller ℓ_1 norm. One direct solution to this problem is to enforce regularization on $\mathbf{B}$. Since the ambiguity is caused by a scaling factor, enforcement on the norm or trace of $\mathbf{B}$ can be applied. If the trace of $\mathbf{B}$ is normalized, the problem can be formulated as follows:

$$\begin{aligned} &\min \quad (w.r.t. \;\; \mathbf{w}, b, \mathbf{B}, \xi_i) \quad \|\mathbf{w}\|_1 + \lambda \sum_{i=1}^{n} \xi_i \\ &s.t. \;\; c_i(\mathbf{w}^T \mathbf{z_i} + b) \geq 1 - \xi_i \\ &\quad \mathbf{z_i} = \mathbf{B}\mathbf{x_i} \\ &\quad trace(\mathbf{B}) = 1 \\ &\quad \xi_i \geq 0 \end{aligned} \tag{4.11}$$

Compared with L1-norm SVM, LMDR is able to generate lower dimensional features, i.e., more sparse features, with the help of the optimized linear transform $\mathbf{B}$. Compared with the distance metric learning, LMDR incorporates the L1-norm SVM to determine the dimension of the low dimensional feature. Another advantage of LMDR over the distance metric learning lies in the measure of separation in the transform feature space. For the distance metric learning, the measure is the total distance between all sample pairs from different classes and the same classes. The distance metric learning does not include a scheme to detect outliers that may exist in the data distribution. On the other hand, by using the margin between the classes and slack variables ξ_i to deal with outliers, LMDR has more generalization capacity than the distance metric learning approach. Samples with large magnitude of ξ_i can be detected as outliers and excluded from training.

It is also interesting to compare LMDR with general LDA, which derives low dimensional feature representations by maximizing the ratio of inter-class scatter to inner-class scatter. In this sense, LDA also provides a method for simultaneous feature extraction and selection. In LDA, feature extraction is through a matrix containing the eigenvectors of $\mathbf{S_W}^{-1}\mathbf{S_B}$, where $\mathbf{S_W}$ is the inner-class scatter matrix and $\mathbf{S_B}$ is the inter-class scatter matrix [10]. Feature selection is based on the relative magnitudes of the eigenvalues of $\mathbf{S_W}^{-1}\mathbf{S_B}$. By using the second order statistics of the data samples for dimension reduction, LDA assumes a Gaussian distribution for the original feature space. When the Gaussianity assumption is violated, which often happens in real applications, the performance of LDA suffers. On the other hand, LMDR does not enforce any assumptions on the distribution of the underlying data, which extends the applicability of LMDR. Another difference between LDA and LMDR lies in the discrimination measure. The margin used in LMDR for the discrimination measure has been shown to have better generalization capacity, as shown in general SVM [76, 77].

The problem with the optimization problem of LMDR is that the constraints $c_i(\mathbf{w}^T\mathbf{z_i} + b) \geq 1 - \xi_i$ are not convex, since the optimization variables $\mathbf{w}$ and $\mathbf{B}$ are entangled. Here we adopt a simple alternating optimization algorithm that breaks the entanglement between $\mathbf{w}$ and $\mathbf{B}$ as follows.

1. Initialize $\mathbf{B} = \mathbf{B_0}$, where the columns of $\mathbf{B_0}$ is the generalized eigenvectors from the equation $\mathbf{S_B x} = \lambda \mathbf{S_W x}$. This means that the transform matrix is initialized by the result of the generalized LDA problem.

2. Search for the large margin classifier given $\mathbf{B} = \mathbf{B_0}$:

$$\begin{aligned}
&(\mathbf{w_0}, b, \xi_i) = \arg\min \quad (w.r.t. \quad \mathbf{w}, b, \xi_i) \quad \|\mathbf{w}\|_1 + \lambda \sum_{i=1}^{n} \xi_i \\
&s.t. \quad c_i(\mathbf{w}^T\mathbf{z_i} + b) \geq 1 - \xi_i \\
&\qquad \mathbf{z_i} = \mathbf{B_0 x_i} \\
&\qquad \xi_i \geq 0
\end{aligned} \tag{4.12}$$

3. Search for the transform matrix given $w = w_0$:

$$\begin{aligned}
&(\mathbf{B_1}, b, \xi_i) = \arg\min \quad (w.r.t.\ \ \mathbf{B}, b, \xi_i) \quad \|\mathbf{w_0}\|_1 + \lambda \sum_{i=1}^{n} \xi_i \\
&s.t. \ \ c_i(\mathbf{w_0}^T \mathbf{z_i} + b) \geq 1 - \xi_i \\
&\quad \mathbf{z_i} = \mathbf{B}\mathbf{x_i} \\
&\quad trace(\mathbf{B}) = 1 \\
&\quad \xi_i \geq 0
\end{aligned} \tag{4.13}$$

4. Set $\mathbf{B_0} = \mathbf{B_1}$ and repeat steps 2 and 3 until the change in $\mathbf{w}$ and $\mathbf{B}$ is less than a threshold value or a given number of iterations has been reached.

The optimization formulated in equation (4.13) is clearly a linear programming problem. Actually the problem formulated in equation (4.12) is also a linear programming problem. To see this, let $\mathbf{w} = \mathbf{u} - \mathbf{v}$, where $\mathbf{u} \in R^{dx1}$, $\mathbf{v} \in R^{dx1}$ and $\mathbf{u} \geq 0, \mathbf{v} \geq 0$, then $\|\mathbf{w}\|_1 = \mathbf{1}^T(\mathbf{u}+\mathbf{v})$. Each solution of $\mathbf{u}$ and $\mathbf{v}$ corresponds to a solution of $\mathbf{w}$. The same trick was previously adopted in the basis pursuit algorithm [17] and in solving the L1-norm SVM with linear programming [74,80].

4.5 Experiments

4.5.1 Experiments with Synthetic Data

We start with a simple synthetic numerical example. Two classes of two dimensional data are generated, as illustrated in Figure 4.4 (a). Denote each 2D data sample as $[x_1, x_2]$ and given that $0 \leq x_1 \leq 5$, samples are generated as $x_2 = x_1 + a + \omega$ for the first class and $x_2 = x_1 - a + \omega$ for the second class, where $a = 2.1$ is a constant shift and ω is white Gaussian noise with mean 0 and standard deviation 1. The data distribution is designed in a way such that both feature components x_1 and x_2 play equivalently important roles in classification. Running L1-norm SVM on the data set results in $\mathbf{w} = [-0.5585, 0.6619]^T$, which indicates that we can not reduce the dimension of this feature set, both dimensions have similar weights. On the other hand, running the iterative large margin dimension reduction as in equations (4.12) and (4.13) with 2 iterations results in the following solution:

$$\mathbf{B} = \begin{bmatrix} -2.0463 & 2.1336 \\ 2.1336 & 3.0463 \end{bmatrix}, \qquad \mathbf{w} = [\ 1.5338 \quad 0\ ]^T \tag{4.14}$$

Applying **B** to the original features yields the feature distribution shown in Figure 4.4 (b). As a comparison, **B0** obtained by the generalized LDA and the corresponding **w** from the L1-norm SVM are as follows:

$$\mathbf{B0} = \begin{bmatrix} -0.1060 & 0.0673 \\ 0.1059 & 0 \end{bmatrix}, \qquad \mathbf{w} = [\ 4.4456 \quad 0\ ]^T \tag{4.15}$$

Feature distribution obtained by applying **B0** to the original features results in the new distribution shown in Figure 4.4. Applying the transform **B** on the synthetic data set results in the data distributions shown in Figure 4.4 (b). From this figure, it is clear that in the transform feature space, the first feature component x_1 is much more important for classification than the second component x_2. Simultaneously, the value of **w** correctly reflects this difference between the two feature components. In this case, the feature dimension can be effectively reduced to 1 based on the value of **w**.

From these numerical results and the distributions shown in Figure 4.4, it is clear that both LDA and the LMDR are effective for dimension reduction. From the coefficients of L1-norm SVM, as shown in equation (4.14) and (4.15), it is seen that both methods can reduce the feature dimension from 2 to 1. In order to compare the performance of classification in the reduced dimensional space, features are projected onto $1D$ space and the minimum Bayesian classification errors are determined. As shown in Figure 4.4 and indicated by equations (4.14) and (4.15), $x1$ feature component is retained for the LMDR and generalized LDA based method and the $x2$ component is discarded. In this setting, 2 samples (1.96%) are misclassified for the LMDR method and 5 samples (2.90%) are misclassified for the generalized LDA method. This comparison shows that the optimization steps in the LMDR improves the distribution of the features in the low dimensional space.

4.5.2 Experiments with the UCI Data

Experiments are also conducted on two data sets: "iris" and "glass" from the UCI data repository [86]. The properties of the two data sets are summarized in Table 4.1. For the purposes of comparison, the linear discriminative analysis (LDA) is implemented for dimension reduction. Since LDA's dimension depends directly on the number of classes, it can only derive up to $C-1$ directions for feature projection, where C is the number of different classes in the training set. This means that LDA can generate up to 2 dimensional features for "iris" and 5 dimensional features for "glass".

The experiment setting is as follows. The large-margin dimension reduction is based on a scenario of pairwise classification, as indicated by the constraints in equations 4.12 and 4.13. The generalization of LMDR to the case of multiple classes follows the same method as the general SVM, where pairwise classification ("1 vs 1") or setting one class as positive label and others as negative label ("1 vs rest") are commonly used. In our method, LMDR based dimension reduction and classification are conducted pair-wise, i.e., based on data samples from each pair of classes c_1 and c_2, a large dimension reduction scheme is trained and feature selection and classification is based on the classifier built on this pair-wise data. For a data set with C different classes, a total of $C(C-1)/2$ dimension reduction schemes and classifiers are built. Each classifier assigns either C_1 or C_2 to each sample in the testing set. The final class label of a testing sample is determined by a simple majority vote on each class label. For the large-margin dimension reduction, the feature is first subjected to a linear transform $\mathbf{B}$ and the feature components are selected by the magnitude of the components of $\mathbf{w}$. In the training process, the parameter λ in equations 4.12 and 4.13 is set by cross validation. The iterative optimization algorithm for the large-margin dimension reduction is run for 6 iterations each time. Half of the data set is randomly selected for training and the other half is used for testing. The KNN classifier with

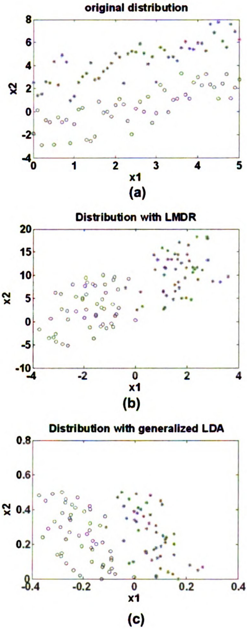


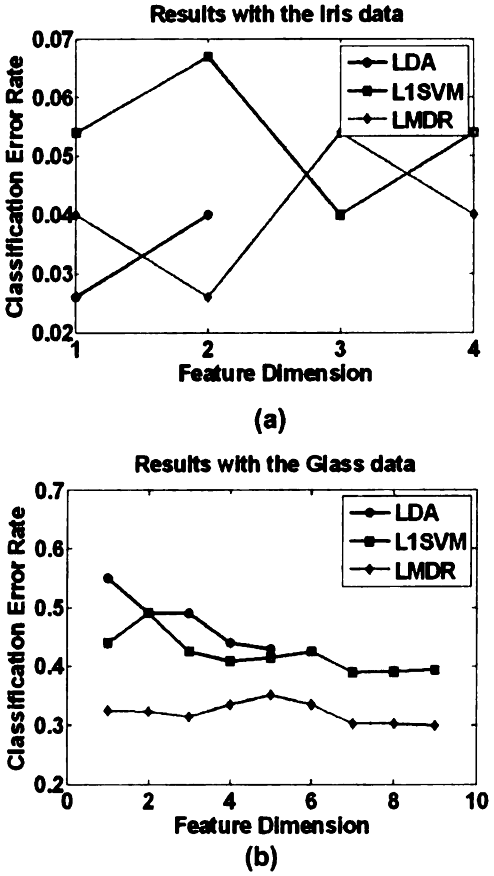
Figure 4.4. Feature distributions: (a) Original features; (b) Features obtained with LMDR; (c) Features obtained with the generalized LDA

Table 4.1. The properties of the data sets "iris" and "glass"

Data set	Number of class	Samples per class	Feature dimension
Iris	3	50	4
Glass	6	from 9 to 76	9

Euclidian distance is used for classification. The number of nearest neighbors, "K" in "KNN" is set to a quarter of the number of training samples in each pairwise classifier.

Given the above experimental settings, the classification error rates at different dimensions selected by LDA and the large-margin dimension reduction are plotted in Figure 4.5, where (a) is the result on the "iris" data set and (b) is the result on the "glass" data set. The large-margin dimension reduction method performs differently for the two data sets. For the "iris" data set, the performance of LDA is comparable with that of large-margin dimension reduction. For the "glass" data set, LMDR performs better than LDA. The performance difference on the two data sets can be explained by looking at the data sample distributions in the two data sets, as illustrated in Figure 4.6, which plots the distribution of "iris" and "glass" in the 2D space obtained by LDA. From this figure, it is clear that the data samples in "iris" are clearly separated. Another important observation is that the distributions of each class in the 2D space can be roughly approximated by a Gaussian distribution. Therefore, the distribution of "iris" satisfies LDA's assumption on the data distribution quite well and this explains why LDA performs well on the "iris" data. On the other hand, Figure 4.6 (b) shows that the distribution of "glass" data is more complex than the "iris" data, in the sense that different classes are not well separated and the distribution of each class can not be accurately approximated as a Gaussian distribution. In this case, the assumption of a Gaussian distribution required by the LDA does not hold and therefore classification based on LDA yields high error rates. Since LMDR does not assume any prior information on the data distribution and since it uses class margin as the measure of separation, it yields much better classification performance on the "glass" data set.



(a)

(b)

Figure 4.5. The relation between classification error rate and feature dimension: Comparison between LDA, L1-norm SVM and large-margin dimension reduction on (a) "iris" and (b) "glass" data sets.

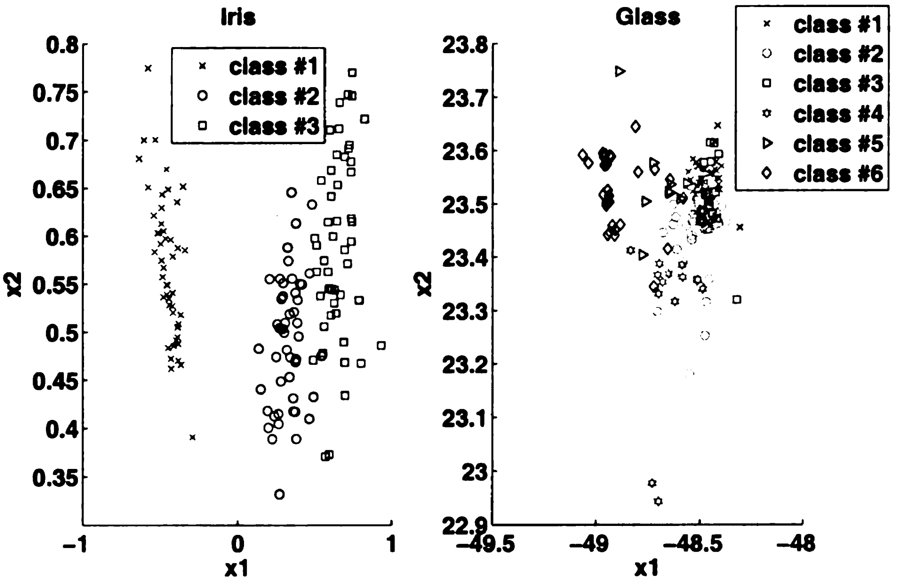


Figure 4.6. The projection of "iris" and "glass" data sets with LDA to 2D space.

4.5.3 Experiments with the Texture Data

Experiments are also conducted on texture classification with the texture images from the Brodatz data set [56]. Images from 10 texture classes, with 16 images in each class, are used for classification. Half of the images are used for training and another half is used for testing. are conducted again by decomposing the images into 3 wavelet level. In this case, each image has a feature feature of 85 dimensions. Based on the experiments conducted in Chapter 2, the features from most of the subbands are noise, in the sense that only a small portion of the features can achieve a relatively low classification error rate. The indication of this phenomenon is two sided. First, the involvement of large number of noisy feature components greatly interfere the optimization process of the LMDR method. Second, the computational complexity of the optimization process is increased. To tackle these problems, we propose the combination of L1-norm SVM and LMDR for high dimensional feature selection and transformation. For the high dimensional feature, the L1-norm SVM is first applied to identify those important feature components and discard the noisy feature components. The LMDR is then applied to the low dimensional feature this data set, and results in a very high classification error rate (higher than 70%), due to the interference of noisy features.

In Figure 4.7, the curve "LMDR" is obtained by first applying the L1-norm SVM to select 35 features from the 85 features, and then applying the LMDR on the selected 35 features. Features of different dimensions are then selected by the value of $\mathbf{w}$ in equation (4.11). The curve "L1SVM+LMDR" is obtained by first applying the L1-norm SVM to select n features from the 85 features, and then applying the LMDR on the selected n features, generating n-dimensional features in the transformed space. In this case, the value of n changes from 5 to 35, with an increment of 5.

From this figure, the performance of LMDR is worse than that of L1-norm SVM. As discussed in Chapter 2, the 35 features selected in by the L1-norm SVM contains considerable amount of noise that misleads the LMDR process. On the other hand, "L1SVM+LMDR" always performs better than the "LMDR" and better than the "L1SVM" in the low dimensional space. This illustrates that searching for an optimal linear transform for the feature can substantially improve the accuracy of classification. The phenomenon that "L1SVM+LMDR" does not performs better than "L1SVM" in the high dimensional space illustrates that noisy features can interfere the optimization of LMDR and deteriorate its performance. In this experimental setting, the minimum classification error rate is achieved by "L1SVM+LMDR" at the dimension of 15. Therefore, the combination of L1-norm SVM and LMDR is an effective strategy for dimension reduction on high dimensional feature. The advantage is two-sided: Removing the noisy features and reducing the computational complexity.

4.6 Summary

In this chapter, the SRSC problem proposed in the previous chapter is reformulated as a constrained optimization problem by using the large margin as the measure of discrimination. The formulation has solid theoretic foundation but faces practical numerical problems since multiple steps of iterative quadratic programming are involved. To deal with this problem, the SRSC is divided into two steps, with the first step

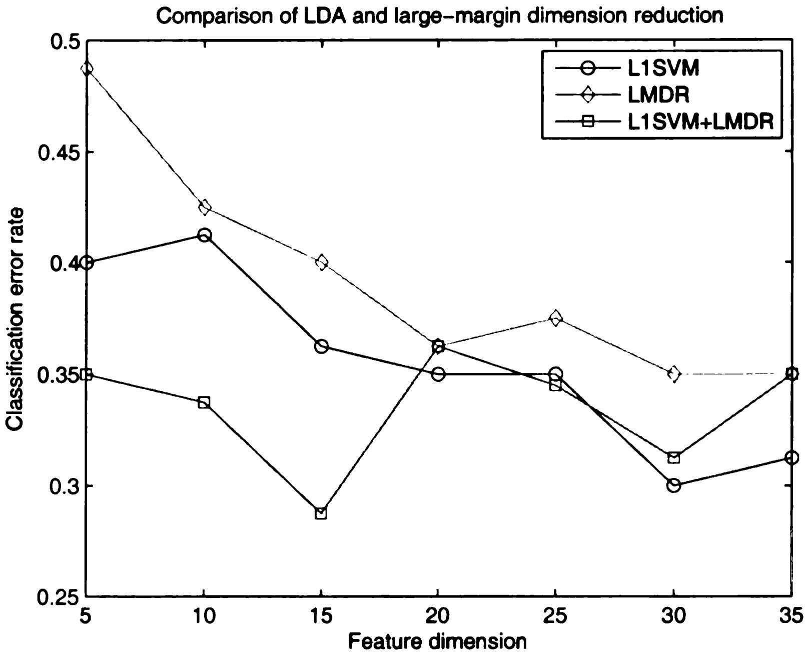


Figure 4.7. The relation between classification error rate and feature dimension: Comparison between Ll-norm SVM, LMDR and the combination of Ll-norm SVM and LMDR on texture data.

being sparse reconstruction and the second step being sparse classification. For the sparse classification, the large margin dimension reduction (LMDR) that incorporates the L1-norm SVM and distance metric learning is proposed. An efficient optimization algorithm based on iterative linear programming is proposed to solve the constrained optimization formulation. The proposed algorithm inherits the advantages of distance metric learning for finding the optimal linear feature transform and L1-norm SVM for automatically determining the dimensionality of the low-dimensional space. Experimental results on the UCI data sets show that compared with the LDA method, the proposed LMDR method achieves comparable results on dimension reduction on data samples following Gaussian distribution and better results on data samples that do not follow the Gaussian distribution.

CHAPTER 5

Feature Evaluation with Information Theoretic Measures

5.1 Introduction

All of the previous chapters in this dissertation deal with obtaining a low dimensional (sparse) feature set that is discriminative. In order to show that the features derived from the proposed methods are discriminative, empirical experiments are conducted for evaluating the "goodness" of different feature representations. This chapter discusses the problem of feature evaluation and proposes a new method for feature evaluation based on information theoretic measures that can be computed efficiently.

Traditional pattern recognition problems usually include two consecutive steps: feature extraction and classification. Feature extraction aims at generating a compact and discriminative description of the data in terms of a feature vector. Multiple features describing different properties of the sample data can be extracted. For example, color and texture features can be extracted from an image for content based image retrieval. Therefore, there is a need for evaluating the different features. For example, in the face recognition application, features based on different linear transforms, such as PCA and LDA, have been widely investigated [11,66]. At this point, it is important to note that although the problem of feature evaluation shares some common characteristics with feature selection [31], the two methods differ in their basic goal. Feature selection evaluates and selects a subset of the feature elements, while feature evaluation quantifies the performance of the different feature vectors. Taking the color and texture features for an image as an example, feature selection evaluates the different components of the color feature such as red, green and blue, while feature evaluation focuses on the performance of the color feature versus the performance of the texture feature.

Feature evaluation is usually conducted by an empirical test, i.e., the classification accuracy on given testing data sets. Feature evaluation with empirical tests can be seen in various applications, such as text categorization [87, 88] and image retrieval [89]. However, this paradigm of feature evaluation does not only depend on the feature itself. The results of empirical tests are often subject to the distance metric that determines how to measure the similarity between the two features and the choice of the classifier that determines how to assign class labels. The impact of the distance metric on the classification results is so substantial that distance metric learning, which learns a suitable distance metric for a specific classification task, has recently attracted a lot of attention [90–92]. It is also obvious that the choice of the classifier and the parameter setting in a given classifier have strong impact on classification results. Various examples of how the classifier affects classification results can be found in literature, such as the experimental results on the benchmark UCI data [86] in a recent paper [93]. Feature evaluation based on empirical tests is complicated due to the fact that tuning the distance metric and the classifier may benefit certain features and degrade the performance of other features. However, overly tuning of either the distance metric or the classifier parameters can result in distorted evaluation for different features. One interesting discussion on empirical tests can be found in [94], which shows an extreme case that given any algorithm, a data set can be designed such that the algorithm can outperform other algorithms on this data set in the sense of classification accuracy. This can also be interpreted as that even a "bad" feature in the general sense can result in a high classification accuracy with certain classification algorithms.

The constraints of the empirical test for feature evaluation motivates the need for an alternative feature evaluation method that is independent of the distance metric and the classifier. Fisher's ratio used in the linear discriminative analysis (LDA) is one such choice [10]. The Fisher's ratio measures the ratio between intra-class scatter and inter-class scatter. For the samples drawn from Gaussian distributions, a large value of Fisher's ratio indicates a large distance between samples from different classes and a small distance among samples from the same class. In [95], mutual information (MI) between class label and features is computed for feature selection. A higher MI value indicates that the certain feature components provide more information for class label. This selection measure can also provide a criterion for feature evaluation. However, the computational model for estimating MI value proposed in [95] was only applied to 1D and 2D features and may be difficult to adapt to high dimensional features.

In this chapter, mutual information between a feature and the class label is computed as a quantitative criterion for feature evaluation. The applicability of MI for feature evaluation is theoretically supported by Fano's inequality [48], which reveals the direct connection between the MI value and the probability of misclassification. An uncorrelated linear discriminative analysis (ULDA) [96] based MI computational model is proposed such that the evaluation criterion is applicable for high dimensional features. The proposed method avoids the empirical test and thus is independent of the distance metric and classifier. Therefore, the empirical test, when combined with the proposed MI based measure, can provide a more complete and accurate criterion for feature evaluation.

The rest of this chapter is organized as follows. Section 5.2 reviews the MI and Fano's inequality that motivates the usage of MI as the criterion for feature evaluation. Section 5.3 analyzes the existing computational models for MI computation and proposes a new MI computation model based on uncorrelated linear discriminative analysis. Section 5.4 conducts extensive experiments to test the validity of the proposed MI based feature evaluation method. Finally, Section 5.5 concludes this chapter and discusses the findings.

5.2 Quantitative Evaluation of Features with Mutual Information

In the following analysis, the feature is modelled as a random vector: $\mathbf{X} \in \tilde{\mathrm{X}} \subset R^d$ and the class label is modelled as a discrete random variable: $C \in \tilde{\mathrm{C}} \subset Z^1$. The joint distribution of X and C is described by *p(x,c)*. The MI between $\mathbf{X}$ and C is defined as:

$$\begin{aligned} I(\mathbf{X};C) &= \sum_{\mathbf{x}\in\tilde{X}} \sum_{c\in\tilde{C}} p(\mathbf{x},c) \log \tfrac{p(\mathbf{x},c)}{p(\mathbf{x})p(c)} = E_{\mathbf{X}C}\left[\log \tfrac{p(\mathbf{x},c)}{p(\mathbf{x})p(c)}\right] \\ &= D(p(\mathbf{x},c)||p(\mathbf{x})p(c)), \end{aligned} \tag{5.1}$$

where $D(\bullet||\bullet)$ is the relative entropy or Kullback-Leibler distance. When the logarithm function in the definition uses a base of 2, the unit for the $I(X;C)$ is bits. MI is symmetric in $\mathbf{X}$ and C, nonnegative, and is equal to zero if and only if $\mathbf{X}$ and C are independent. $I(\mathbf{X};C)$ indicates how much information $\mathbf{X}$ conveys about C. Given $\mathbf{X}$, the extra information required to fully describe C is given by the conditional entropy $H(C|\mathbf{X})$ [48].

$$I(\mathbf{X};C) = H(C) - H(C|\mathbf{X}), \tag{5.2}$$

where $H(C)$ is the entropy of the random variable C. Given this equation, mutual information computation is equivalent to the computation of two entropy values. Similar to the definition of conditional entropy, the conditional MI between random variables $\mathbf{X}$ and C given a new random variable $\mathbf{Z}$ is defined as:

$$\begin{aligned} I(\mathbf{X};C|\mathbf{Z}) &= H(\mathbf{X}|\mathbf{Z}) - H(\mathbf{X}|C,\mathbf{Z}) \\ &= E_{p(\mathbf{x},c,z)}[\log \tfrac{p(\mathbf{X},C|\mathbf{Z})}{p(\mathbf{X}|\mathbf{Z})p(C|\mathbf{Z})}]. \end{aligned} \tag{5.3}$$

The conditional MI has an interpretation similar to that of MI. Given the above definitions, the mutual information between a random vector $\mathbf{X} = [X_1, X_2, ..., X_d]$ and a random variable C can be defined as:

$$\begin{aligned} I(\mathbf{X};C) &= I(X_1, X_2, ..., X_d; C) \\ &= \sum_{i=1}^{d} I(X_i; C|X_{i-1}, X_{i-2}, ..., X_1). \end{aligned} \tag{5.4}$$

5.2.1 Fano's Inequality

Intuitively, $I(\mathbf{X};C)$ indicates how much information the feature $\mathbf{X}$ contains about the class label C. The larger $I(\mathbf{X};C)$ is, the more accurate is the estimation of the class label C from the feature $\mathbf{X}$. Mathematically, the relation between the MI and the probability of misclassification is given by Fano's inequality [48], as follows:

$$P(C \neq C') \geq \frac{H(C|\mathbf{X}) - 1}{\log(N)} = \frac{H(C) - I(\mathbf{X};C) - 1}{\log(N_c)} \tag{5.5}$$

where N_c is the number of possible values of C (i.e., the number of classes), and C' is the estimation of C based on the feature $\mathbf{X}$ predicted by a classifier. For a given classification task, $H(C)$ and $\log(N_c)$ are fixed. Thus, a large value of $I(\mathbf{X};C)$ directly results in a small value of $P(C \neq C')$. Note that Fano's inequality is not confined to any specific type of classifier or distance metric and is a suitable measure for feature evaluation.

Due to the connection between $P(C \neq C')$ and $I(\mathbf{X};C)$ revealed by Fano's inequality, MI has been widely applied to feature extraction [50, 51, 97] and feature selection [98, 99]. At this point, it should be noted that MI based feature extraction differs from the MI based feature evaluation. In the former case, the feature is unknown and chosen based on MI. In this chapter, the feature is already extracted and MI is used for the evaluation purposes. For MI based feature extraction, a distance metric and the classifier are usually known a priori to evaluate the effect of feature extraction, thus providing feedback for parameter adjustment for the extraction process. For MI based feature evaluation discussed in this chapter, no distance metric and classifier are involved.

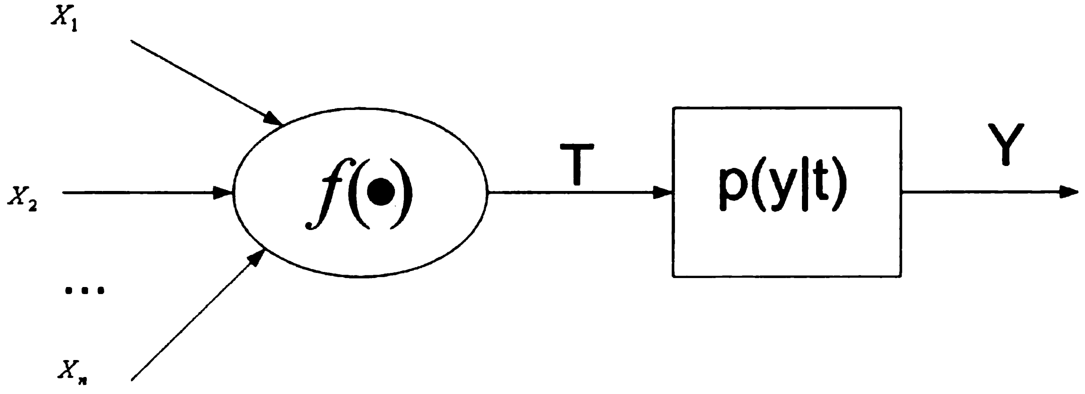


Figure 5.1. MI computation with dimension reduction. The original feature $\mathbf{X}$ is in a high-dimensional space and the transformed feature $\mathbf{Y}$ is in a low-dimensional space.

5.3 Mutual Information Computation

The difficulty of MI computation lies in estimating the joint probability density function (pdf), or equivalently, the conditional pdf between C and $\mathbf{X}$, especially in a high dimensional space with limited number of data samples. A variety of methods aiming at reducing the complexity by dimension reduction have been proposed. For example, in [100, 101], the relationship between the different components of $\mathbf{X}$ is modelled by the first order Markov chain that simplifies the MI computation, as follows:

$$I(X_i; C|X_{i-1}, X_{i-2}, ..., X_1) = I(X_i; C|X_{i-1}); \tag{5.6}$$

A more general approach for dimension reduction is searching for a transform $T(\bullet)$, such that $\mathbf{Y} = T(\mathbf{X}) \in \tilde{Y} \subset R^{d'}$ that satisfies $d' < d$. Rather than computing $I(\mathbf{X}; C)$, $I(\mathbf{Y}; C)$ is computed as an approximation for $I(\mathbf{X}; C)$. This approach has been adopted in a variety of applications, such as in [38, 49–51, 102]. This transform process can be illustrated in Figure 5.1. With this transform, a Markov chain is formed as: $\mathbf{X} \rightarrow \mathbf{Y} \rightarrow C$. According to the data processing theorem [48], $I(\mathbf{X}; C) \geq I(\mathbf{Y}; C)$. The equality holds if and only if $\mathbf{Y}$ is the sufficient statistics of $\mathbf{X}$. Based on this ineqaulity, the transform $T(\bullet)$ that maximizes $I(\mathbf{Y}; C)$ should be chosen so that the error in the approximation is minimized.

A common method for solving this dimension reduction problem is to first assume a parametric form of $T(\bullet)$ and then optimize the parameters for $T(\bullet)$ such that $I(Y;C)$ is maximized. For example, the linear transform $\mathbf{Y} = \mathbf{W}\mathbf{X}$, where $\mathbf{W} \in R^{d' \times d}$ is the transformation matrix, is often used for its simplicity. In order to estimate the optimal value for $\mathbf{W}$, $I(\mathbf{Y};C)$ is explicitly expressed as a function of $\mathbf{W}$ and the standard optimization techniques, such as gradient descent, are then used to get an optimal value for $\mathbf{W}$. This paradigm for searching for $T(\bullet)$ is adopted in [49–51] for feature extraction, i.e., deriving an informative low dimensional feature $\mathbf{Y}$ from a high dimensional feature $\mathbf{X}$. In [49–51], the pdf $p(\mathbf{Y},C)$ is modelled with the Parzen window method with a Gaussian kernel. Thus, $p(\mathbf{Y},C)$ is written as a mixture of multiple Gaussian functions, which is non-convex. The combination of non-convexity and the gradient descent method used in [49–51] can only guarantee a locally optimal solution for $\mathbf{W}$. Due to the complexity in optimizing $\mathbf{W}$, simplified transforms are used in some applications. For example, the equal weight model where $\mathbf{W} = \frac{1}{d}\mathbf{1}_{d' \times d}$, is used for computing MI between wavelet coefficients in [38].

Note that the goal of the transform $T(\bullet)$ is dimension reduction. A small value of d' reduces the complexity of MI computation. However, a larger value of $d - d'$ usually indicates a larger difference between $I(\mathbf{X};C)$ and $I(\mathbf{Y};C)$, since $\mathbf{Y}$ is not a sufficient statistics of $\mathbf{X}$ in the general case. For feature extraction, the classification is conducted with this low dimensional feature $\mathbf{Y}$. Therefore, the large difference between $I(\mathbf{X};C)$ and $I(\mathbf{Y};C)$ is not a severe problem, as long as $I(\mathbf{Y};C)$ can be maximized. However, this estimation error in MI is a problem for feature evaluation. Given two features, $\mathbf{X_1} \in R^d$ and $\mathbf{X_2} \in R^d$ and their low dimensional representations $\mathbf{Y_1} \in R^{d'}$ and $\mathbf{Y_2} \in R^{d'}$, the reliability of concluding that $I(\mathbf{X_1};C) > I(\mathbf{X_2};C)$ based on $I(\mathbf{Y_1};C) > I(\mathbf{Y_2};C)$ is influenced by the error introduced by the inaccuracy of the computational model.

The tradeoff between the computational complexity and the error in MI estimation caused by the value of d' are due to the statistical dependence among different components of $\mathbf{Y}$, i.e., $Y_1, Y_2, ...Y_{d'}$. If these components were statistically independent, the computation of $I(\mathbf{Y};C)$ would become much easier:

$$I(\mathbf{Y};C) = \sum_{i=1}^{d'} I(Y_i;C) \tag{5.7}$$

If equation (5.7) holds, the complexity of computing MI does not substantially increase with an increase in d'. Therefore, a large value of d' can be chosen to reduce the error between $I(\mathbf{X};C)$ and $I(\mathbf{Y};C)$ and the computation of the MI is still tractable. Generally, it is difficult to search for a transform that results in statistical independence among the different components of $\mathbf{Y}$. An approximate solution is to search for a linear transform that results in uncorrelated projections, i.e, $cov(Y_i, Y_j) = 0$ for $1 \leq i \neq j \leq d'$. Since statistical uncorrelation is equivalent to statistical independence only when $\mathbf{Y}$ is distributed as a Gaussian, which is not satisfied in the general case, equation (5.7) does not strictly hold and should be rewritten as:

$$I(\mathbf{Y};C) \approx \sum_{i=1}^{d'} I(Y_i;C) \tag{5.8}$$

The approximation error depends on how the distribution of $\mathbf{Y}$ deviates from a Gaussian distribution. As an approximation, this can be obtained by measuring the Gaussianity of the individual components of $\mathbf{Y}$, i.e, Y_i for $1 \leq i \leq d'$. The measure of the Gaussianity of a random variable can be conducted by computing the kurtosis or negentropy of the variable, as in the method of independent component analysis (ICA) [103].

Motivated by the above discussion, the uncorrelated linear discriminative analysis (ULDA) is chosen to obtain the transform matrix $\mathbf{W}$ in this dissertation. ULDA was proposed in [104] and improved in [96] as a generalized linear discriminative analysis method that projects a feature into statistically uncorrelated feature components. The improved ULDA is also applicable to the undersampled problem, where the dimension of the data is much larger than the number of samples. This property makes it especially suitable for processing high dimensional features. ULDA has been successfully applied to a variety of classification tasks, such as text categorization [96], face recognition [105] and gene data classification [106]. ULDA derives up to $N_c - 1$ optimal discriminative vectors that maximizes the Fisher's ratio, with an extra constraint that the different eigenvectors are $\mathbf{S_t}$-orthogonal [96], where $\mathbf{S_t}$ is the total scatter matrix of the features. Since ULDA also maximizes the Fisher's ratio, the distribution of samples on individual feature components, Y_i, achieves the goal that samples from different classes are far away from each other and samples from the same class are close to each other. In this sense, the value of $I(Y_i; C)$ is optimized compared to other linear projections. Since $I(\mathbf{Y}; C) \leq I(\mathbf{X}; C)$ according to the data processing theorem, optimizing $I(Y_i; C)$ helps reduce the estimation error between $I(\mathbf{Y}; C)$ and $I(\mathbf{X}; C)$. When using ULDA for feature evaluation, the transform matrix $\mathbf{W}$ is given by the discriminative vectors obtained by ULDA in each row.

5.3.1 Overview of Uncorrelated Linear Discriminative Analysis

Uncorrelated linear discriminative analysis (ULDA) was proposed to extract statistically uncorrelated discriminative features [96,104]. Assume that there are k classes in the training data set. Denote the mean vector, covariance matrix and *a priori* probability of the *ith* class as $\mathbf{m}_i$, $\mathbf{S}_i$ and p_i respectively. Then the between-class scatter $\mathbf{S}_b$, within-class scatter $\mathbf{S}_w$ and the total scatter matrix $\mathbf{S}_t$ are defined as follows:

$$\begin{aligned} \mathbf{S}_w &= \sum_{i=1}^{k} p_i \mathbf{S}_i \\ \mathbf{S}_b &= \sum_{i=1}^{k} p_i (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T \\ \mathbf{S}_t &= \mathbf{S}_b + \mathbf{S}_w \end{aligned} \tag{5.9}$$

In the traditional linear discriminative analysis (LDA), the objective function is to minimize the ratio of within-class scatter to between-class scatter. This goal can be achieved by solving an eigenvalue problem on $\mathbf{S}_w^{-1}\mathbf{S}_b$, provided that the within-class scatter matrix $\mathbf{S}_w$ is nonsingular. In this case, there are at most $k-1$ discriminative vectors available, since the rank of $\mathbf{S}_b$ is bounded above by $k-1$. Although $\mathbf{S}_w$ and $\mathbf{S}_b$ are symmetric matrices, generally $\mathbf{S}_w^{-1}\mathbf{S}_b$ is not symmetric. Therefore, the discriminative vectors that are the eigenvectors of $\mathbf{S}_w^{-1}\mathbf{S}_b$ are not orthogonal and the extracted features are not uncorrelated.

ULDA aims at finding the optimal discriminative vectors that are $\mathbf{S}_t$-orthogonal. Two vectors $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are $\mathbf{S}_t$-orthogonal if $\mathbf{Y}_1^T\mathbf{S}_t\mathbf{Y}_2 = 0$. Specifically, after r vectors $\phi_1, \phi_2, .., \phi_r$ have been extracted, the $(r+1)$th vector ϕ_{r+1} is obtained by solving the following optimization problem:

$$\begin{aligned} &\max_{\phi} f(\phi) = \tfrac{\phi^T \mathbf{S_b} \phi}{\phi^T \mathbf{S_w} \phi} \\ &s.t. \quad \phi_{r+1}^T \mathbf{S_t} \phi_i = 0, \quad i = 1, 2, ..., r \end{aligned} \tag{5.10}$$

One solution to this optimization problem was proposed in [104] by iteratively solving a generalized eigenvalue problem. A more computationally efficient solution called ULDA/QR algorithm was proposed in [96] based on QR decomposition and singular vector decomposition (SVD). More details on solving this optimization problem can be found in [96, 104]. In our implementation, the ULDA/QR algorithm is used for generating the discriminative vectors. It was shown that feature vectors transformed by ULDA are mutually uncorrelated [96, 104]. The proof of this property of uncorrelation is straightforward, as shown below.

Given that

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_m \end{bmatrix} = \begin{bmatrix} \varphi_1^T \\ \varphi_2^T \\ . \\ . \\ \varphi_m^T \end{bmatrix} \mathbf{X} \tag{5.11}$$

The covariance between any two feature components, y_i and y_j is computed as follows:

$$E[(y_i - E(y_i))(y_j - E(y_j))] = \varphi_j^T \mathbf{S_t} \varphi_i \tag{5.12}$$

Based on the property given in equation (5.10), $\varphi_j^T \mathbf{S_t} \varphi_i = 0$. Therefore, the feature components in the transformed space obtained with the ULDA are uncorrelated.

Table 5.1. Classification Error Rates (in percentage) on “wine”, “iris” and “glass”

Data set	Wine	Iris	Glass
Error Rate	1% to 5%	4% to 9%	25% to 30%

5.4 Experiments

5.4.1 Experiments with the UCI data

For the first part of the experiment, the goal is to show that “good” features have in general high MI between the feature vector and the class label. For this purpose, the data sets of “wine”, “iris” and “glass” from the UCI machine learning repository [86] are used for the experiment. Based on the previous experimental results with various distance metrics and classifiers on these data sets presented in literature, such as [93], we know that the classification error rates on the three data sets are in different ranges as shown in Table 5.1. Based on the results, it can be concluded that the features for “wine” is better than the features for “iris”, which is better than that of “glass”.

We apply both the equal weight model where $W = \frac{1}{d}\mathbf{1}_{d' \times d}$ and the ULDA based model for estimating MI values for the three data sets. Note that for the equal weight model, $\mathbf{Y}$ is in the one dimensional space. The results are plotted in Figure 5.2. For the ULDA based method, the feature $\mathbf{X}$ can be projected to a space with dimension up to $N_c - 1$. The data sets “wine” and “iris” have 3 classes and the “glass” data set has 6 classes. Therefore, the projected feature $\mathbf{Y} \in R^2$ for “wine” and “iris” and $\mathbf{Y} \in R^5$ for “glass”. Figure 5.2 shows that the MI estimated from the equal weight model is inconsistent with the empirical results. The inconsistency in this example indicates that the model of the equal weights is not suitable for feature evaluation. Figure 5.2 also shows that if only the first component, i.e., Y_1, obtained by the ULDA is retained, the evaluation given by the MI in the descending order is as “iris, wine, glass”, which is still inconsistent with the empirical results. However, it should be noted that the MI values with Y_1 from the ULDA on the 3 data sets are larger than the corresponding MI values obtained with the equal weight method. This indicates that the ULDA optimizes the MI estimation better than the equal weight model.

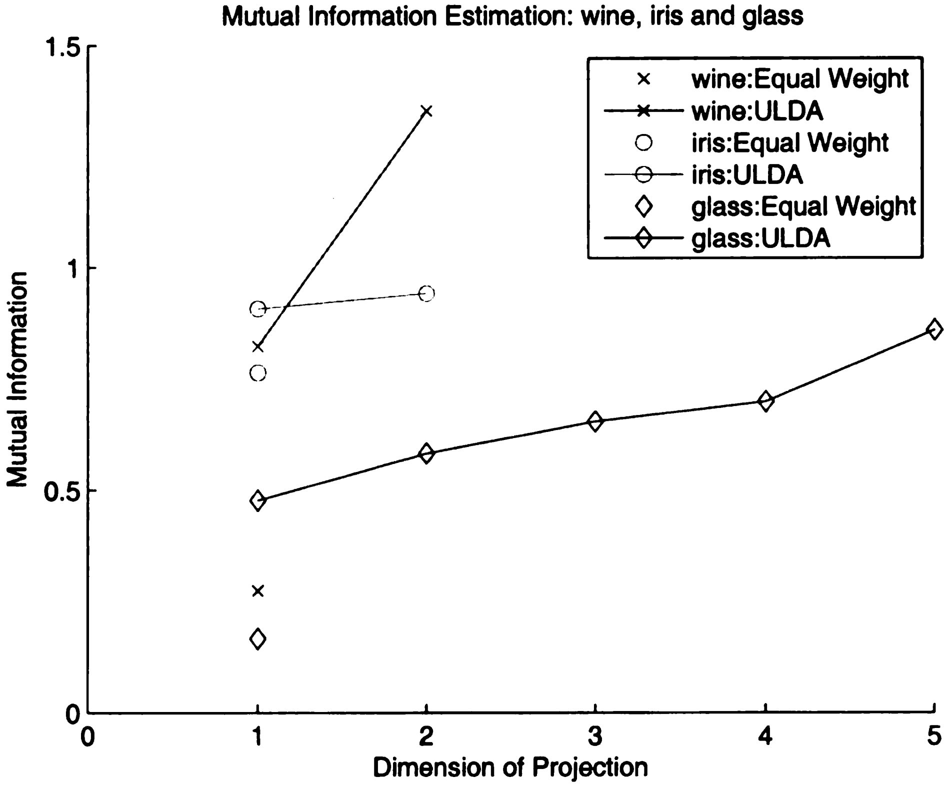


Figure 5.2. Estimation of mutual information on 3 different data sets using equal weight and LDA based method.

Finally, Figure 5.2 shows that the ULDA based method with additive approximation model $I(\mathbf{Y}; C) \approx \sum_{i=1}^{d'} I(Y_i; C)$ provides a feature evaluation measure that is consistent with the empirical classification results: "wine" is better than "iris", which is better than "glass".

5.4.2 Experiments with Energy Features for Wavelet Based Texture Classification

In the previous experimental section, the 3 sets of features are from 3 different types of objects. In this section, feature evaluation using the proposed method is conducted on 5 different wavelet features extracted from the same set of texture images for classification. Wavelet feature extraction is a two-step process. In the first step, an image is decomposed into multiple wavelet subbands. In the second step, one feature component is extracted from each subband and a feature vector is formed by concatenating all of the feature components [9, 23, 107]. Given a subband coefficient matrix $\mathbf{R}$ with size $M \times N$, 5 different definitions of energy function are used to extract features from each subband: E1 (Standard Deviation), E2 (Residual), E3 ($L1$ Energy), E4 ($L2$ Energy) and E5 (Entropy). The feature vectors based on the following 5 energy functions are extracted and evaluated in the experiment.

$$E_1 = \sqrt{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}|\mathbf{R}(i,j) - \mu|^2} \tag{5.13}$$

$$E_2 = \sum_{i=1}^{M}\sum_{j=1}^{N}|\mathbf{R}(i,j) - \mu| \tag{5.14}$$

$$E_3 = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}|\mathbf{R}(i,j)| \tag{5.15}$$

$$E_4 = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}|\mathbf{R}(i,j)|^2 \tag{5.16}$$

$$E_5 = -\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\frac{|\mathbf{R}(i,j)|}{\mu MN}\log(\frac{|R(i,j)|}{\mu MN}) \tag{5.17}$$

where μ is the mean value defined as:

$$\mu = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\mathbf{R}(i,j) \tag{5.18}$$

By definition, E1 and E2 are similar to each other and E3 and E4 are similar to each other. Based on this observation, we may expect that the classification performance is close for similar features. In the experiment, 54 texture images from the Brodatz texture database [56] are used. Each texture image with a size of 512×512 is divided evenly into 16 non-overlapping images. Therefore, there are 54 texture classes, with 16 samples in each class in this classification task. Each sample image is decomposed into 85 wavelet subbands with the Daubechies 16-tap wavelet basis. Hence, the feature dimension d is 85 for each of the 5 features. ULDA applied to the extracted features generates a new feature up to $54 - 1 = 53$ dimensions.

Based on the equation $I(\mathbf{Y}; C) \approx \sum_{i=1}^{d'} I(Y_i; C)$, we set the value of d' from 1 to 53 and plot the corresponding estimation of $I(\mathbf{Y}; C)$ versus d' for each feature type in Figure 5.3. In Figure 5.3, we note that the shapes of all 5 curves are similar. When d' is small (from 1 to somewhere between 10 and 20), the curves have a sharp slope, and the slope becomes flat after that. This indicates that the first several discriminative feature components provide more information for the classification than the other components. This is consistent with the implementation of the ULDA, which arranges eigenvectors corresponding to the large eigenvalues first. We also note that the curves of E1 (standard deviation) and E2 (residual) are close to each other, and the curves of E3 (L1 norm) and E4 (L2 norm) are close to each other, while the curve of E5 (entropy) is not close to any of the other curves. This similarity between the MI estimates is rooted in the similarity of the definition of the corresponding energy functions.

For the purpose of feature evaluation, it can be inferred from Figure 5.3 that E1 and E2 are the best, and the E5 is the worst, while E3 and E4 are in between. Note that this conclusion comes without conducting any classification experiments and is thus independent of any distance metric and classifier. It is therefore interesting to compare how this conclusion matches with empirical testing results. For this purpose, two simple classification experiments are conducted with the 5 features. For the first experiment, K nearest neighbor (KNN) classifier with "leave-one-out" method is used. The value of "K" in KNN is set to 16 and the standard Euclidian distance is used as the distance metric. Given two feature vectors $\mathbf{X_1}$ and $\mathbf{X_2}$, the Euclidian distance is the ℓ_2 norm of their difference, as in equation 5.19.

$$d_1(\mathbf{X_1}, \mathbf{X_2}) = \|\mathbf{X_1} - \mathbf{X_2}\|_2^2 \tag{5.19}$$

In the second experiment, all conditions are the same as the first one except

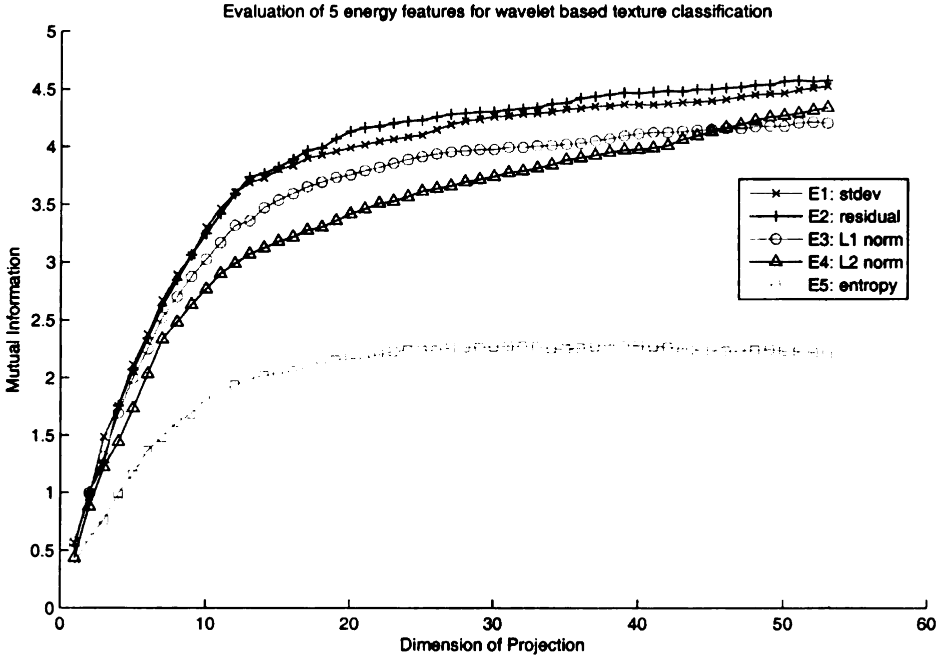


Figure 5.3. Estimation of mutual information on 5 different energy features for wavelet based texture classification.

that the distance metric is replaced by the normalized Euclidian distance, defined in equation (5.20).

$$d_2(\mathbf{X_1}, \mathbf{X_2}) = \sum_{i=1}^{d'} \left(\frac{\mathbf{X_1}(i) - \mathbf{X_2}(i)}{\sigma_i} \right)^2 \tag{5.20}$$

where σ_i is the standard deviation of all the feature components extracted from subband i. The classification error rates of the empirical tests are summarized in Table 5.2.

When comparing the error rates in the two experiments, we find that the normalization has the most obvious effect on the performance of E3 and E4. This is because the additive operation used in the definition of Euclidian distance has the effect of deemphasizing the small distance values. Therefore, the normalization greatly reduces the classification error rates for E3 and E4. Note that the normalization does not benefit classification accuracy for all features, such as for E5 in this case. Therefore, manipulation of the distance metric may benefit the evaluation of some features in the empirical tests while degrading the performance of other features. The experimental results with normalization matches the previous results of MI based feature evaluation, i.e., El and E2 are the best, and E5 is the worst, while E3 and E4 are in between. This conclusion is not completely consistent with the experimental results without normalization, where E4 performs the worst. However, it is generally believed that normalization is needed for dealing with filtering features in texture classification [57]. Therefore, the experiment with normalization reflects the performance of different features more accurately. The comparison of experimental results in Table 5.2 also illustrates that the variance introduced by the distance measure could be so large that the feature evaluation based on empirical tests could not reflect the truth. Therefore, MI based measure proposed in this chapter, combined with the empirical tests, is able to provide a more complete picture for feature evaluation.

Table 5.2. Classification Error Rates with Different Energy Functions (in percentage)

	E1	E2	E3	E4	E5
Euclidean distance	12.15	13.54	28.59	72.11	39.12
Normalized Euclidean distance	10.53	8.91	11.34	14.12	41.90

5.5 Conclusion

In this chapter, a mutual information based feature evaluation method with a corresponding computational model is proposed and studied. The empirical test for feature evaluation is influenced by the choice of the distance metric and the classifier. The proposed method evaluates features by computing the MI between the feature and the class label. The validity of this method is supported by Fano's inequality and the method is not dependent on any distance metric and classifier. To avoid the problem of estimating pdf in a high dimensional space, an ULDA based computational model is proposed for MI computation. Finally, experiments on the UCI data and wavelet features for texture classification demonstrate the promising results for using the proposed method for quantitative feature evaluation.

CHAPTER 6

Summary and Conclusions

This dissertation discusses the problem of feature extraction and selection in transform based image classification. The major objective is to obtain a compact and robust set of features for a given image set using multi-resolution transforms such as the wavelets, wavelet packets, directional wavelets, Gabor functions and other filter banks.

6.1 Summary

The first part of the proposed work focuses on the extraction of a set of features from wavelet and wavelet packet transforms. The proposed approach first quantifies the dependence between wavelet subbands and then exploits this property to select a compact set of features that are discriminative for the given image set. This method is then further improved by incorporating the individual discrimination power provided by each subband into the selection process.

The promising results obtained with the proposed feature selection methods motivate the formulation of a more general feature selection problem for image classification. This more general feature selection problem can be posed as "How can we choose the 'best' set of multi-resolution transforms for discriminating between different image classes?" The 'best' set of transforms is defined by the compactness (sparseness) of the selected feature set, the robustness of the features in noisy environments and the accuracy of the classification. Combining all of these requirements, we pose this more general problem using sparse representations. Sparse representations aim at finding a sparse and close approximation to a given signal using a large collection of functions, called a dictionary. In the proposed work, this framework is modified to address the question of image classification by defining a cost function that incorporates the discrimination and reconstruction abilities of the elements in the dictionary, as well as the sparseness of the selected feature vector. As part of the proposed work, the different aspects of this problem will be investigated including the selection of the dictionary elements, different cost functions to quantify the discrimination power of the selected features and the tradeoff between reconstruction power and discrimination power.

In the third part, the formulation of the proposed sparse representation for image classification method is further improved by using the large margin method for the measure of discrimination. Based on this new and improved formulation, we can model the robust and sparse feature extraction with an optimization problem that can be solved by iterative quadratic programming. In order to reduce the computational complexity required for the iterative quadratic programming, we propose decomposing the robust and sparse feature extraction into two steps, with the first step being sparse reconstruction and the second step being sparse feature selection and dimension reduction. For the second step, we propose a new method called large margin dimension reduction (LMDR). LMDR integrates the idea of L1-norm support vector machine (SVM) and distance metric learning for obtaining feature representation in a low dimensional space.

In the fourth part, we propose a mutual information based feature evaluation criterion and its computational model, such that the features obtained from different selection methods can be quantitatively evaluated. Traditional feature evaluation based on the empirical study is subjected to the selection of the distance metric and the classifier. The proposed measure is independent of distance metric and classifier, and is supposed to more objectively reflect the discrimination power of a feature representation. Computation of mutual information in a high dimensional space is usually involved in the model. To deal with this problem, we propose computing mutual information with the uncorrelated linear discrimination analysis (ULDA). The proposed computational model effectively reduces the computational complexity of computing mutual information for feature evaluation.

6.2 Future Work

6.2.1 Basis Design for Sparse Image Classification

The idea of basis design is motivated by the sparse representations for image classification discussed in Chapter 3 and the existing data adaptive iterative basis design methods [12, 108, 109], which aim to derive an efficient data adaptive dictionary for sparse representations. The goal is to design a data adaptive dictionary that is efficient for sparse and accurate image classification. The basis design method can follow the same process as discussed in [12, 108, 109], which interleaves solving the sparse representation with a dictionary and optimizing each dictionary entry with the fixed coefficients obtained from the solution of the sparse representation.

If this approach is successful, it would provide a different objective function and a different rule for updating the dictionary specific to image classification, thus improving the classification accuracy. Formally, the iteration algorithm may be described as follows.

1. Randomly initialize the dictionary $\mathbf{A} \in R^{N \times M}$.

2. *Sparse representation stage*: Use the pursuit algorithm introduced in Section 3.2.2 to optimize the objective function for sparse representations for signal classification, i.e., equation(3.10),

$$J_3(\mathbf{X}, \lambda_1, \lambda_2) = F(\mathbf{X}) - \lambda_1 \sum_{i=1}^{K} \|\mathbf{x}_i\|_0 - \lambda_2 \sum_{i=1}^{K} \|\mathbf{y}_i - \mathbf{A}\mathbf{x}_i\|_2^2 \tag{6.1}$$

3. *Dictionary update stage*: For each $m = 1, 2, ..., M$, determine the set of signal samples that use the dictionary entry D_m: $\omega_m = \{Y_i | 1 \le i \le N, X_i(m) \neq 0\}$. If ω_m has samples from more than one class, apply the LDA on the data set ω_m and replace D_m with the eigenvector corresponding to the largest eigenvalue of the LDA. If all samples in ω_m are from the same class, keep D_m unchanged.

4. Iterate steps 2 and 3 for a given number of iterations.

The motivation for the dictionary update stage is to increase the value of $F(\mathbf{X})$ in equation (6.1), and therefore increase the value of the objective function. The biggest challenge of this method is to prove the convergence of the iterative algorithm. Another major problem is that after the dictionary entry is updated in step 3, the entry may no longer be used by the same group of samples. This is also a common problem for the current iterative basis design methods, as no discussions of convergence appears in [12, 108, 109].

6.2.2 Combination of Data-Driven and Model Based Methods for Sparse Signal Representation and Classification

The existing basis design methods can be divided into two categories: data-driven and model based. The two types of dictionary are designed based on different strategies. The data-driven method derives its dictionary from a set of training signals and expects the dictionary to adapt to the signals. Therefore, the dictionary is usually dependent on the training samples. Some examples of data driven basis design include principal component analysis (PCA) [10], linear discriminative analysis (LDA) [10], independent component analysis (ICA) [110], non-negative matrix factorization (NMF) [111, 112] and K-SVD method [12]. The dictionaries derived from the data driven model are usually more efficient for signal representation than the model based method, given that the signal to be represented has the same statistical properties as the training samples. For example, the dictionary derived from the PCA method achieves the optimal linear transform for representing a high dimensional signal with a low dimensional representation in the sense of minimum mean square error (MSE). NMF has been shown to be able to decompose a signal into non-negative parts. The disadvantage of the data driven methods is that the dictionary needs to be updated whenever new samples are added to the training set. This property is also undesirable in some applications like compression, where the dictionary is also needed to be encoded. This data driven dictionary design has been successfully adopted in various applications, such as PCA/LDA/NMF for face recognition [66, 112, 113], K-SVD for image denoising [18], ICA for blind source separation [114] and NMF for spectra recovery in brain signals [115].

On the other hand, the model based method is independent of any training data. The general approach for this basis design method is to capture the desired properties of the dictionary, such as the signal singularities or continuities, and then design the dictionary that achieves the proposed properties through methods like filter design. In this sense, the model based method tries to summarize the common characteristics of signals and develops a set of templates in representing the signals efficiently. A typical example of this kind of method is the traditional Fourier transform, which uses sinusoid functions as the templates to model the signal components. Recent research in the wavelet basis design have generated new dictionaries that can be used to efficiently model signal components. Examples include the ridgelet [116, 117], curvelet [117], contourlet [118], bandlet [14], beamlet [119] and directionlets [120]. Most of these dictionaries inherit the properties of the traditional wavelets, i.e., the multi-resolution analysis (MRA) for analyzing signal structure at multiple scales and multiple shifts, and the self-similarity over different scales. Aside from these inherited advantages, these dictionaries also provide direction as another degree of freedom for signal analysis. The advantage of the model based method is that the dictionary is data independent, which is favorable in applications like compression. The disadvantage is that the basis in the dictionary may not be very efficient for signal representation, when compared with the dictionaries obtained with the data driven methods. Some typical applications of the wavelet dictionaries include image denoising [117], inpainting [68], coding [43] and recognition as discussed in Chapter 2.

Since the data driven method and the model based method are based on different strategies and have different advantages, it is natural to combine their benefits. In essence, the model based method designs a series of templates (bases) that can efficiently capture the common characteristics of signals. Given that the signal space is usually much larger than the space that can be efficiently (sparsely) spanned by the dictionary, it is desired that the variability of the "common characteristics" is reduced such that the dictionary from the model based method has a better chance of representing the signals efficiently. A direct method to reduce signal variability is to reduce the signal size, since the variance is exponentially proportional to the size. For example, for a length N sequence with an amplitude range of m bits, the number of possible signals is $(2^m)^N$. This strategy is used for the K-SVD basis design for sparse image representation in [12], where images are divided into 8×8 non-overlapping blocks for analysis and the size of each basis is adapted to these blocks.

In the new method, we may use a different strategy to reduce the signal variance. We notice that some data driven methods are good at extracting the "common characteristics" of signals to be analyzed. For example, the "eigenface" obtained by the PCA [66], "fisherface" obtained by the LDA [66] and "face parts" obtained by the NMF [112] are the effective "common characteristics" for face image representation. Therefore, we propose a two-stage approach where a data driven method is used to reduce theh variance in the given signal space and to capture the information with a few components in the first stage and a sparse representation with an overcomplete dictionary is applied in the second stage.

6.2.3 Saliency Analysis with Sparse Representation

Recent applications of sparse representations to computer vision problems such as object recognition [121] and visual tracking [122] motivates us to apply the sparse representation method for discriminative signal analysis, especially for saliency analysis. Generally speaking, saliency analysis aims at finding information-rich object parts to discriminate different classes of objects. For example, car parts, like wheels and side windows in various backgrounds, are searched by moving windows in the spatial domain and extracted as salient templates for car detection in [123]. In [124,125], the salient regions, such as human face, bicycle boundary, are detected by the discrete cosine transform (DCT) for discriminating different classes of objects in images. In [126], a PCA subspace is first generated with samples of all classes, and then a subset of PCA bases that can best represent each class is searched for. With the framework of the sparse representation and the various new wavelet bases with powerful signal modelling capacity, we feel that the sparse representations may be able to provide a better solution to the problem of saliency analysis.

In the proposed method, the images from each class can be decomposed through the standard sparse representation with a combination of multiple dictionaries. The combined dictionary should include basis functions that can efficiently model different structures, such as DCT for local structure and curvelets for recurrent texture structure. A larger dictionary is likely to have a higher probability of efficiently representing structures specific to different images. However, a large dictionary also means a higher computational complexity for solving the sparse representation problem. Therefore, there is a tradeoff between the versatility of the dictionary and the computation cost. We expect that salient regions from different classes can be discriminated based on their respective sparse representations. This approach for saliency analysis with sparse representations is similar to finding a matched filter for images from an overcomplete dictionary. If the dictionary is more versatile, it is more likely that one or more "matched filters" can be found. Designing a "matched filter" for each class may not be practical, considering that the common features of samples in a class may not be easily described. The method of sparse representations is promising thanks to the recent developments in the wavelet basis design that have generated dictionaries with powerful image modelling capacity.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] T. Chang and C. Kuo, "Texture analysis and classification with tree-structured wavelet transform," *IEEE Transactions on Image Processing*, vol. 2, no. 4, pp. 429–441, 1993.

[2] A. Laine and J. Fan, "Texture classification by wavelet signature," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 11, pp. 1186–1191, 1993.

[3] A.K. Jain, S. Prabhakar, and H. Lin, "A multichannel approach to fingerprint classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 4, pp. 348–359, 1999.

[4] A. Jain, S. Prabhakar, L.Hong, and S. Pankanti, "Filterbank-based fingerprint matching," *IEEE Transactions on Image Processing*, vol. 9, no. 5, pp. 846–859, 2000.

[5] A. Mojsilovic, M. Popovic, and D.Rackov, "On the selection of an optimal wavelet basis for texture characterization," *IEEE Transactions on Image Processing*, vol. 9, no. 12, pp. 2043–2050, 2000.

[6] H. Choi and R.G. Baraniuk, "Multiscale image segmentation using wavelet-domain hidden markov models," *IEEE Transactions on Signal Processing*, vol. 10, no. 9, pp. 1309–1321, 2001.

[7] M. Acharyya, R. De, and M. Kundu, "Extraction of feature using m-band wavelet packet frame and their neuro-fuzzy evaluation for multitexture segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1639–1644, 2003.

[8] R. Datta, D. Joshi, J. Li, and J. Wang, "Image retrieval: Ideas, influences, and trends of the new age," Tech. Rep., Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA 16802, USA., 2006.

[9] T. Randen and J.H. Husoy, "Filtering for texture classification: A comparative study," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 4, pp. 291–310, 1999.

[10] R. Duda, P. Hart, and D. Stork, *Pattern classification (2nd ed.)*, Wiley-Interscience, 2000.

[11] A. Martinez and A. Kak, "PCA versus LDA," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228–233, 2001.

[12] M. Aharon, M. Elad, and A. Bruckstein, "The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[13] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, 1999.

[14] E.Le Pennec and S. Mallat, "Sparse geometric image representations with bandelets," *IEEE Transactions on Image Processing*, vol. 14, no. 4, pp. 423–438.

[15] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, pp. 3397–3415, 1993.

[16] Y. Pati, R. Rezaiifar, and P. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *27th Annual Asilomar Conference on Signals, Systems, and Computers*, 1993.

[17] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1999.

[18] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3736–3745, 2006.

[19] M. Elad, B. Matalon, and M. Zibulevsky, "Image denoising with shrinkage and redundant representation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[20] G. Choueiter and J. Glass, "A wavelet and filter bank framework for phonetic classification," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2005, vol. 1, pp. 933–936.

[21] N. Ince, A. Tewfik, and S.Arica, "Classification of movement eeg with local discriminant bases," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2005, vol. 5, pp. 413–416.

[22] G. Fan and X. Xia, "Wavelet-based texture analysis and synthesis using hidden Markov models," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 50, no. 1, pp. 106–120, 1 2003.

[23] G. Wouwer, P. Scheunders, and D.Dyck, "Statistical texture characterization from discrete wavelet representations," *IEEE Transactions on Image Processing*, vol. 8, no. 4, pp. 592–598, 1999.

[24] C. Garcia, G.Zikos, and G. Tziritas, "Wavelet packet analysis for face recognition," *Image and Vision Computing*, vol. 18, no. 4, pp. 289–297, 2000.

[25] M. Do and M. Vetterli, "Rotation invariant texture characterization and retrieval using steerable wavelet-domain hidden markov models," *IEEE Transactions on Multimedia*, vol. 4, no. 4, pp. 517–527, 12 2002.

[26] M. Do and M. Vetterli, "Wavelet-based texture retrieval using generalized gaussian density and Kullback-Leibler distance," *IEEE Transactions on Image Processing*, vol. 11, no. 2, pp. 146–158, 2 2002.

[27] S. Arivazhagan and L. Ganesan, "Texture classification using wavelet transform," *Pattern Recognition Letters*, vol. 24, no. 10, pp. 1513–1521, 2003.

[28] M. Unser, "Texture classification and segmentation using wavelet frames," *IEEE Transactions on Image Processing*, vol. 4, no. 11, pp. 1549–1560, 1995.

[29] K. Laws, "Rapid texture identification," in *SPIE Vol. 238 Image Processing for Missile Guidance*, 1980, pp. 376–380.

[30] R. Coifman and M. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 713–718, 1992.

[31] A. Jain and D. Zongker, "Feature selection: Evaluation, application, and small sample performance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 2, pp. 153–158, 1997.

[32] N. Rajpoot, "Local discriminant wavelet packet basis for texture classification," in *Proceedings SPIE Wavelets X*, 2003, vol. 5207, pp. 774–783.

[33] D. Koller and M. Sahami, "Toward optimal feature selection," in *Proceedings of International Conference on Machine Learning*, 1996, pp. 284–292.

[34] S. Zhu, N. Wu, and D. Mumford, "Minimax entropy principle and its application to texture modeling," *Neural Computation*, vol. 9, no. 8, 1997.

[35] Y. Li, A. Cichocki, S. Amari, et al., "Sparse representation and its applications in blind source separation," in *Proceedings of Neural Information Processing Systems*, 2003.

[36] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *Journal of Machine Learning Research*, vol. 5, no. 10, pp. 1205–1224, 10 2004.

[37] H. Chen and P. Varshney, "Feature subset selection with applications to hyperspectral data," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2005, vol. 2, pp. 249–252.

[38] J. Liu and P. Moulin, "Information-theoretic analysis of interscale and intrascale dependencies between image wavelet coefficients," *IEEE Transactions on Image Processing*, vol. 10, no. 11, pp. 1647–1658, 2001.

[39] M. Crouse, R. Nowak, and R. Baraniuk, "Wavelet-based statistical signal processing using hidden markov models," *IEEE Transactions on Signal Processing*, vol. 46, no. 4, pp. 886–902, 1998.

[40] G. Fan and X. Xia, "Improved hidden markov models in the wavelet-domain," *IEEE Transactions on Signal Processing*, vol. 49, no. 1, pp. 115–120, 2001.

[41] J. Portilla and E. P.Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," in *International Journal of Computer Vision*, 2000, number 1, pp. 49–71.

[42] E. Simoncelli, "Modeling the joint statistics of images in the wavelet domain," in *Proceedings of SPIE, 44th Annual Meeting*, 1999.

[43] J. Shapiro, "Embedded imagecoding using zerotrees of wavelet coeficients," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3445–3462, 1993.

[44] L. Ma, T. Tan, Y. Wang, and D. Zhang, "Personal identification based on iris texture analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1519–1533, 2003.

[45] A. Jain, *Fundamentals of Digital Image Processing*, Englewood Cliffs, NJ:Prentice-Hall, 1989.

[46] G.K. Grunwald, R.J. Hyndman, L. Tedesco, and R. Tweedie, "A unified view of linear AR(1) models," Tech. Rep., Department of Preventive Medicine and Biometrics, University of Colorado Health Sciences Center, Denver, CO, USA., 1996.

[47] A. Srivastava, A. Lee, E. Simoncelli, and S. Zhu, "On advances in statistical modeling of natural images," *Journal of Mathematical Imaging and Vision*, vol. 18, no. 1, pp. 17–33, 2003.

[48] T.M. Cover and J.A. Thomas, *Elementsof Information Theory*, Wiley, 1991.

[49] J. Principe, D. Xu, and J. Fisher, *Information Theoretic Learning*, pp. 265–319, Wiley, 1999.

[50] K. Torkkola, "Feature extraction by non-parametric mutual information maximization," *Journal of Machine Learning Research*, vol. 3, pp. 1415–1438, 2003.

[51] K. Hild, D. Erdogmus, K. Torkkola, and J. Principe, "Feature extraction using information-theoretic learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, pp. 1385–1392, 2006.

[52] R. Moddemeijer, "On estimation of entropy and mutual information of continuous distributions," *Signal Processing*, vol. 16, no. 3, pp. 233–246, 1989.

[53] L. Paninski, "Estimation of entropy and mutual information," *Neural Computation*, vol. 15, pp. 1191–1253, 2003.

[54] A. Ihler, J. Fisher, and A. Willsky, "Nonparametric hypothesis tests for statistical dependency," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2234–2249, 2004.

[55] S.M. Ross, *Introduction to Probability Models (7th edition)*, Academic Press, 2000.

[56] P. Brodatz, *Texture: A Photographic Album for Artists and Designers*, Dover, New York, USA, 1966.

[57] W. Ma and B. Manjunath, "Texture features and learning similarity," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1996, pp. 425–430.

[58] Y. Li, A. Cichocki, and S. Amari, "Analysis of sparse representation and blind source separation," *Neural Computation*, vol. 16, no. 6, pp. 1193–1234, 2004.

[59] J. Starck, M. Elad, and D. Donoho, "Image decomposition via the combination of sparse representation and a variational approach," *IEEE Transactions on Image Processing*, vol. 14, no. 10, pp. 1570–1582, 2005.

[60] I. Gorodnitsky and B. Rao, "Sparse signal reconstruction from limited data using focuss: a re-weighted minimum norm algorithm," *IEEE Transactions on Signal Processing*, vol. 45, no. 3, pp. 600–616, 1997.

[61] I. Drori and D. Donoho, "Solution of L1 minimization problems by LARS/Homotopy methods," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2006, vol. 3, pp. 636–639.

[62] B. Olshausen, P. Sallee, and M. Lewicki, "Learning sparse image codes using a wavelet pyramid architecture," in *Proceedings of Neural Information Processing Systems*, 2001, pp. 887–893.

[63] D. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Transactions on Information Theory*, vol. 47, no. 7, pp. 2845–2862, 2001.

[64] Y. Lin and D. Lee, "Bayesian L1-Norm sparse learning," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2006, vol. 5, pp. 605–608.

[65] D. Wipf and B. Rao, "Sparse bayesian learning for basis selection," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2153–2164, 2004.

[66] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[67] S. Fidler, D. Skocaj, and A. Leonardis, "Combining reconstructive and discriminative subspace methods for robust classification and regression by subsampling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 3, pp. 337–350, 2006.

[68] M. Elad, J. Starck, P. Querre, and D. Donoho, "Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA)," *Journal on Applied and Computational Harmonic Analysis*, vol. 19, pp. 340–358, 2005.

[69] A. Leonardis and H. Bischof, "Robust recognition using eigenimages," *Computer Vision and Image Understanding*, vol. 78, pp. 99–118, 2000.

[70] J. Tropp, A. Gilbert, and M. Strauss, "Algorithms for simultaneous sparse approximation. part I: Greedy pursuit," *Signal Processing, special issue on Sparse approximations in signal and image processing*, vol. 86, no. 4, pp. 572–588, 2006.

[71] J. Tropp, A. Gilbert, and M. Strauss, "Algorithms for simultaneous sparse approximation. part II: Convex relaxation," *Signal Processing, special issue on Sparse approximations in signal and image processing*, vol. 86, no. 4, pp. 589–602, 2006.

[72] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *International Joint Conferences on Artificial Intelligence*, 1995, pp. 1137–1145.

[73] S. Nene, S. Nayar, and H. Murase, "Columbia object image library (coil-20)," Tech. Rep., Department of Computer Science, Columbia University, New York City, NY, USA., 1996.

[74] J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani, "1-norm support vector machines," in *Proceedings of Neural Information Processing Systems*, 2004.

[75] E. Xing, A. Ng, M. Jordan, and S. Russell, "Distance metric learning, with application to clustering with side-information," in *Proceedings of Neural Information Processing Systems*, 2002.

[76] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, 1995.

[77] C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[78] S. Boyd and L. Vandenberghe, *Covex Optimization*, Cambridge University Press, 2004.

[79] YALMIP Toolbox, "http://control.ee.ethz.ch/ joloef/yalmip.php," .

[80] J. Bi, K. Bennett, M. Embrechts, and C. Breneman, "Dimensionality reduction via sparse support vector machine," *Journal of Machine Learning Research*, vol. 3, pp. 1229–1243, 2003.

[81] J. Bi, Y. Chen, and J. Wang, "A sparse support vector machine approach to region-based image categorization," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2005, vol. 1, pp. 1121- 1128.

[82] Y. Chen, J. Bi, and J. Wang, "Miles: Multiple-instance learning via embedded instance selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 1931- 1947, 2006.

[83] K. Weinberger, J. Blitzer, and L. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Proceedings of Neural Information Processing Systems*, 2005.

[84] R. Rosales and G. Fung, "Learning sparse metrics via linear programming," in *Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 367 - 373.

[85] G. Fung, R. Rosales, and R. Rao, "Feature selection and kernel design via linear programming," in *Proceedings of Internatioanl Joint Conference on Artificial Intelligence*, 2007, pp. 786 - 791.

[86] D. Newman, S. Hettich, C. Blake, and C. Merz, "UCI repository of machine learning databases," 1998.

[87] Y. Yang and J. Pedersen, "A comparative study on feature selection in text categorization," in *Proceedings of International Conference on Machine Learning*, 1997, pp. 412–420.

[88] T. Liu, S. Liu, Y. Chen, and W. Ma, "An evaluation on feature selection for text clustering," in *Proceedings of International Conference on Machine Learning*, 2003.

[89] P. Howarth and S. Ruger, "Evaluation of texture features for content-based image retrieval," in *International Conference on Image and Video Retrieval*, 2004, pp. 326–334.

[90] K. Weinberger, J. Blitzer, and L. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Proceedings of Neural Information Processing Systems*, 2005, pp. 1473–1480.

[91] G. Lebanon, "Metric learning for text documents," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 497–508, 2006.

[92] G. Lanckriet, N. Cristianini, P. Bartlett, L. Ghaoui, and M. Jordan, "Learning the kernel matrix with semidefinite programming," *Journal of Machine Learning Research*, vol. 5, pp. 27–72, 2004.

[93] L. Li, "Multiclass boosting with repartitioning," in *Proceedings of International Conference on Machine Learning*, 2006, pp. 569 – 576.

[94] D. LaLoudouana and M. B. Tarare, "Data set selection," in *Proceedings of Neural Information Processing Systems*, 2002.

[95] D. Xu and J. Principe, "Feature evaluation using quadratic mutual information," in *Proceedings of International Joint Conference on Neural Networks*, 2001, vol. 1, pp. 459–463.

[96] J. Ye, R. Janardan, Q. Li, and H. Park, "Feature extraction via generalized uncorrelated linear discriminant analysis," in *Proceedings of International Conference on Machine Learning*, 2004, p. 895902.

[97] E.G. Maes and P. Beauseroy, "Mutual informatio-based feature extraction on the time-frequency plane," *IEEE Transactions on Signal Processing*, vol. 50, no. 4, pp. 779–790, 2002.

[98] N. Kwak and C.H. Choi, "Input feature selection by mutual information based on parzen window," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1667–1671, 2002.

[99] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226 - 1238, 2005.

[100] J. Denzler and C.M. Brown, "Information theoretic sensor data selection for active object recognition and state estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 145–157, 2001.

[101] Y. Zhang and Q. Ji, "Sensor selectiion for active information fusion," in *Proc. of the National Conference on Artificial Intelligence (AAAI)*, 2005, pp. 1229–1234.

[102] S. Kaski and J. Peltonen, "Informative discriminant analysis," in *Proceedings of International Conference on Machine Learning*, 2003, pp. 329–336.

[103] A. Hyvarinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[104] Z. Jin, J. Yang, Z. Hu, and Z. Lou, "A theorem on the uncorrelated optimal discriminant vectors," *Pattern Recognition*, vol. 34, no. 10, pp. 2041 2047, 2001.

[105] Z. Jin, J. Yang, Z. Hu, and Z. Lou, "Face recognition based on the uncorrelated discriminant transformation," *Pattern Recognition*, vol. 34, no. 10, pp. 1405 1416, 2001.

[106] J. Ye, T. Li, T. Xiong, and R. Janardan, "Using uncorrelated discriminant analysis for tissue classification with gene expression data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 1, no. 4, pp. 181–190, 2004.

[107] C. Pun and M. Lee, "Log-polar wavelet energy signature for rotation and scale invariant texture classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 590–603, 2003.

[108] B. Olshausen and D. Field, "Sparse coding with an overcomplete basis set: A strategy employed by v," *Vision Research*, vol. 37, no. 23, pp. 3311–3325, 1997.

[109] K. Kreutz-Delgado, J. Murray, B. Rao, K. Engan, T. Lee, , and T. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural Computation*, vol. 15, pp. 349–396, 2003.

[110] A. Hyvärinen, "Survey on independent component analysis," *Neural Computing Surveys*, vol. 2, pp. 94–128, 1999.

[111] D. Lee and H. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[112] M. Spratling, "Learning image components for object recognition," *Journal of Machine Learning Research*, vol. 7, pp. 793–815, 2006.

[113] S. Li, X. Hou, and H. Zhang, "Learning spatially localized, parts-based representation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2001.

[114] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing*, John Wiley and Sons, 2002.

[115] P. Sajda, S. Du, T. Brown, R.Stoyanova, D. Shungu, X. Mao, and L. Parra, "Nonnegative matrix factorization for rapid recovery ofconstituent spectra in magnetic resonance chemical shift imaging of the brain," *IEEE Transactions on Medical Imaging*, vol. 23, no. 12, pp. 1453- 1465, 2004.

[116] A. Flesia, H. Hel-Or, A. Averbuch, E. Candes, R. Coifman, and D. Donoho, "Digital implementation of ridgelet packets," in *Beyond Wavelets.* 2002, Academic Press.

[117] J. Starck, E.J. Candes, and D.L.Donoho, "The curvelet transform for image denoising," *IEEE Transactions on Signal Processing*, vol. 11, no. 6, pp. 670–684, 2002.

[118] M.N. Do and M. Vetterli, "Contourlets," in *Beyond Wavelets.* 2003, Academic Press.

[119] D. Donoho and X. Huo, "Beamlets andmultiscale image analysis," *Multiscale and Multiresolution Methods, Springer Lecture Notes in Computational Science and Engineering*, vol. 20, pp. 149–196, 2002.

[120] V. Velisavljevic, B. Beferull-Lozano, M. Vetterli, and P. Dragotti, "Directionlets: Anisotropic multidirectional representation with separable filtering," *IEEE Transactions on Image Processing*, vol. 15, no. 7, pp. 1916–1933, 2006.

[121] T. Pham and A. Smeulders, "Sparse representation for coarse and fine object recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 555–567, 2006.

[122] O. Williams, A. Blake, and R. Cipolla, "Sparse bayesian learning for efficient visual tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1292–1304, 2005.

[123] S. Agarwal and D. Roth, "Learning a sparse representation for object detection," in *Proceedings of the 7th European Conference on Computer Vision*, 2002, vol. 4, pp. 113–130.

[124] D. Gao and N. Vasconcelos, "Discriminant saliency for visual recognition from cluttered scenes," in *Proceedings of Neural Information Processing Systems*, 2004.

[125] D. Gao and N. Vasconcelos, "Integrated learning of saliency, complex features, and object detectors from cluttered scenes," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2005, vol. 2, pp. 282 - 287.

[126] N. Vaswani and R. Chellappa, "Principal components null space analysis for image and vedio classification," *IEEE Transactions on Image Processing*, vol. 15, no. 7, pp. 1816–1830, 2006.