Content-based Image Retrieval
— A Small Sample Learning Approach

Tao Dacheng

A Thesis Submitted in Partial Fulfillment

of the Requirements for the Degree of

Master of Philosophy

in

Information Engineering

June 2004

Abstract

With the explosive growth of the digital visual information, content based image retrieval (CBIR) became an important and active research topic. Many CBIR systems have been developed over the years based on the low-level visual features. However, the gap between the low-level visual feature and the high-level semantic content always degrades the retrieval performance. Relevance feedback (RF) is an important tool to improve the performance of CBIR. In a RF process, the user first labels a number of relevant retrieval results as positive feedbacks and some irrelevant retrieval results as negative feedbacks. Then the system refines all retrieval results based on these feedbacks. The two steps are carried out iteratively to improve the performance of image retrieval system by gradually learning the user's perception.

Many RF methods have been developed in recent years and small sample learning based methods achieved the state-of-the-art performance. In this thesis, we focus on two popular small sample learning algorithms, the Support Vector Machine (SVM) and the Biased Discriminant Analysis (BDA).

SVM classifies the relevant samples and irrelevant samples based on the support vectors, which are automatically determined by SVM learning algorithm. However, the performance of SVM-based RF is often poor when the number of labeled positive feedback samples is small. This is mainly due to three reasons: 1. SVM classifier is unstable on small size training sets; 2. SVM's optimal hyper-plane may be biased when the positive feedback samples are much less than the negative feedback samples; 3. Over-fitting due to the fact that the feature dimension is much higher than the size of the training set. In this thesis, we try to use random sampling techniques to overcome these problems. To address the first two problems, we propose an asymmetric bagging based SVM. For the third problem, we combine the random subspace method with SVM. Finally, by integrating bagging and RSM, we solve all the three problems and further improve the RF performance.

BDA is another small sample learning model in CBIR RF. In BDA model, the negative feedbacks are required to stay away from the center of positive feedbacks. Although BDA achieved satisfactory results, it also meets many problems: 1. To solve the BDA, the regularization method is used. It is well known that the method often encounters the Matrix Singular Problem (MSP) or the Small Sample Size (SSS) problem; 2. BDA assumes all positive feedbacks form a single Gaussian distribution which may not be the case for CBIR; 3. Although kernel BDA (KBDA) can overcome the single Gaussian distribution assumption to some extent, the kernel parameter tuning makes the online learning unfeasible. Motivated by the successful direct method and null-space method used in linear discriminant analysis to solve the SSS problem, we generalize them into the kernel Hilbert space to over come the SSS problem in KBDA. Because direct method and null-space method may lose some discriminant information, we propose a new full-space method to contain all discriminant information both in linear space and in kernel Hilbert Space. To avoid the parameter tuning problem and the single Gaussian distribution assumption in BDA, we construct a new nonparametric discriminant analysis (NDA) for RF in CBIR. We then generalize the regularization method, the direct method, the null-space method, and the full-space method to address the SSS problem in NDA.

At the end of the thesis, we conduct the first study on SARS radiographic image processing as an application of CBIR. In order to distinguish SARS infected regions from normal lung regions using texture features, we propose several improvements to the traditional gray-level co-occurrence texture features. We use a multi-level feature selection approach to extract texture features from a multi-resolution region based co-occurrence matrix directly for texture classification. The selected texture features can preserve most of the discriminant information in the texture image. Satisfactory results are obtained on a large set of chest radiographic images of SARS patients.

摘要

隨著因特网和數据庫上可視化信息的爆炸式增長，基于內容的圖像檢索已經成為一個非常重要的研究方向。在過去的許多年里，大量的基于底層視覺特征的圖像檢索系統已經問世。

但是，由于底層視覺特征和高層語義特征之間的鴻溝，使得圖像檢索的效果并不理想。而相關反饋正是架构在這個鴻溝上的橋梁，大大的提高了圖像檢索的效果。在一次反饋過程中，用戶首相標記一定數量的和查詢圖像相關以及不相關的圖像分別作為正反饋和負反饋。然后檢索系統根据反饋的信息來重新調整圖像數据的排序。這兩個步驟被重复的執行，直到用戶得到了一個滿意的結果。

近來出現了大量的相關反饋的方法，其中基于小樣本學習的方法取得了令人滿意的效果。在本論文中，我們主要研究兩個最為普遍的基于小樣本學習的反饋方法，他們分別是基于支撐向量机的反饋和基于有偏鑒別矢量分析的反饋。

支撐向量机通過自動學習得到的支撐向量來區分相關樣本与不相關樣本。但是，如果正反饋的樣本很少的時候基于支撐向量机的反饋方法的效果會不理想，這主要是因為下面三個原因：1. 支撐向量机對于小樣本的訓練集合是不穩定的；2. 如果正反饋和負反饋的數目相差比較多，那么支撐向量机的最优化分類平面是有偏的；3. 當底層特征的維數遠遠高于訓練樣本的數量的時候，過适應總會發生。在這篇論文里，我們使用隨机采樣的方法去克服這些問題。為了解決前兩個問題，我們提出了基于不對稱 Bagging 的支撐向量机。為了解決最后一個問題，我們提出了基于特征子空間隨机采樣的支撐向量机。最后，通過合并這兩個方法，我們提出了基于不對稱 Bagging 特征子空間隨机采樣的支撐向量机解決了所有問題。我們在論文中給出了數學上的解釋。大量的試驗証明所提出的方法是行之有效的。

基于有偏鑒別矢量分析的反饋是另外一种基于小樣本學習的反饋方法。在基于有偏鑒別矢量分析的反饋模型中，負反饋樣本被要求遠离正反饋樣本的中心。盡管基于有偏鑒別矢量分析的反饋取得了非常好的效果，但是它依然存在很多問題：1. 為了得到有偏鑒別矢量分析的最优解總會遇到矩陣奇异值

問題或者說是小樣本問題；2. 有偏鑒別矢量分析假定正反饋是服從單高斯分布的，但是這樣的假定對于圖像檢索來說并不合适；3. 盡管基于核空間的有偏鑒別矢量分析能夠避免單高斯分布的假定，但是這個方法又需要進行在線核空間參數的調整。近年來，直接方法和零空間方法成功的解決了線性鑒別矢量分析中矩陣奇异值問題。我們首先把這兩個方法推廣到了核空間的有偏鑒別矢量分析。因為直接方法和零空間方法都會丟失一些鑒別信息，我們提出了一种新的全空間分析方法。新的方法可以保留全部的鑒別信息。為了避免核空間有偏鑒別矢量分析的在線參數調整的問題，我們提出了非參數有偏鑒別矢量分析。最后我們也分別采用了直接方法，零空間方法，和全空間方法去解決非參數有偏鑒別矢量分析中的矩陣奇异值問題。試驗証明，非參數全空間有偏鑒別矢量分析能夠非常好的解決原來面臨的問題，它有效的提高了相關反饋的效果。

在論文的最后，我們研究了 SARS 醫學圖像的分類問題，作為一個圖像檢索的應用，同時它也是SARS計算机輔助治療的初步研究。為了鑒別 SARS 感染的肺部區域和正常的肺部區域，我們提出了兩种方法去改進原有的灰度級共發矩陣特征。我們采用了多層特征選擇方法去提取多分辨率的基于區域的灰度級共發矩陣進行紋理分類。大量的試驗証明，所選擇的特征能夠极大的保留具有分類能力的特征。

Table of Contents

Acknowledgments

Here I would like to acknowledge all the people who had assisted me during the past two years of my graduate studies at the Chinese University of Hong Kong. First of all, I would like to thank my supervisor, Prof. Dr. Xiaoou Tang. All gave his best in providing me with a stimulating and relax environment. Many ideas in this thesis are according to discussions I had with Dr. Tang. All the research work in this thesis is completed under his professional and careful direction. I have learnt so much from him in the past two years, not only on the research, but also on the attitudes of life. I am very fortunate to be able to complete my postgraduate study under his direction.

I would like to thank Dr. Jianzhuang Liu. He gave some beneficial suggestions to my work. All work in this thesis has been done at Multimedia Lab. I deeply appreciate the support of all current and former members of this group for creating a pleasant atmosphere, including Feng Lin, Qingshan Liu, Ying Tan, Lifeng sha, Feng Zhao, Zhifeng Li, Xiaogang Wang, Hao Liu, Bo Luo, Hua Shen, Tong Wang, Liangliang Cao, and Tianqiang Yuan. Thanks to you all for bearing me all the time.

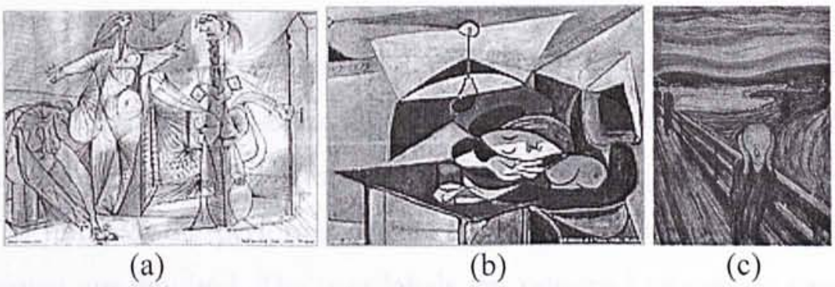
Owe my sincere thanks to all my family members, for their never fading love, care, understanding and encouragement.

Chapter 1

Introduction

1.1 Content-based Image Retrieval

With the explosive growth of image databases in terms of both size and variety, effective indexing and searching images from a large-scale database or the Internet are becoming more and more important in recent years [1][7][50][59][62][66][70].



(a) (b) (c)

Figure 1-1. (a) Picasso's "Bathers with Crab", (b) Picasso's "Girl Asleep at a Table", and (c) Munch's "The Scream"

Conventional approach relies on the key words or text description of an image to retrieve and index image data [40][60][62][70][88]. However to give all images text annotation is very difficult, because automatic annotation of an image cannot be done by the current image processing and pattern recognition techniques yet [3][6]. Moreover, an image says more than a thousand words and many images even cannot be described by text information, such as the Picasso's "Bathers with Crab", the Picasso's "Girl Asleep at a Table", and the Munch's "The Scream", which are shown in Figure 1-1. There, using the visual information of the image data to retrieve images is a reasonable approach for the nonce [35][36][45][65].

Content-based Image Retrieval (CBIR) is the techniques that retrieve semantically relevant images from an image database through the automatically extracted image features based on the color [3][8][28][39][40][60][74][103], texture [4][9][30][46][49][77][93], or shape [2][48][71][72][78] information of the images. In the past twenty years, a great deal of low-level visual features have be used for CBIR, such as the color histogram[60], color coherence vector[28], wavelet texture[4], Gabor texture[9], edge direction histogram[2], etc.

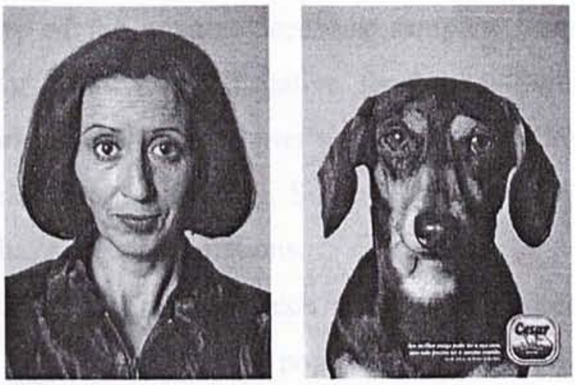


Figure 1-2. The gap between the low-level visual feature and the high-level semantic. The two objects are different but have similar low-level features.

However, the gap between the low-level visual feature and the high-level semantic of an image always leads to the poor performance of CBIR [99][100]. This point can be seen clearly from the Figure 1-2 and Figure 1-3. To bridge the gap and to improve the performance, the interactions between the user and the search engine are required. The user labels the retrieved images as semantically relevant or irrelevant, and then the system refines the retrieval results. This technique is generally named as relevance feedback (RF), which was initially developed in document retrieval [29]. RF is selected as an important modus to scale the performance of CBIR systems during the early and mid 1990's [81][82][90][91][92][95][98][101] and has been shown to provide dramatic performance boost.

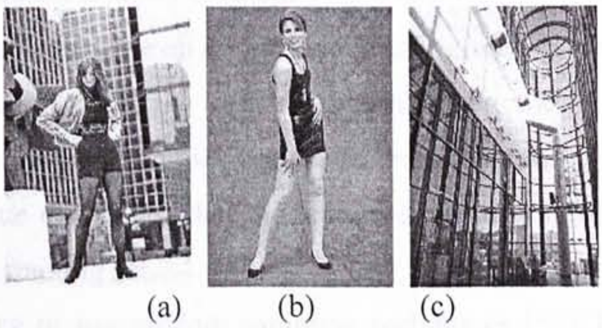


Figure 1-3. The ideal query assumption is not suitable. If (a), which includes the woman and the building, is the query image, (b) and (c) will be retrieved. Clearly, when the user focuses on the woman, (b) is a desirable image, otherwise; (c) is the right one.

Many RF methods have been developed in recent years [15][16][17][18][19] [20] [21][22][23][26][55][58][67][68]. Some approaches [97][99][100][101] adjust the weights of various features to adapt to the user's perception. Some [95][101] estimate the density of the positive feedback samples. Some [55][11] give a binary feedback for positive and negative feedbacks. Some [6][35] use the Bayesian framework to estimate the user's requirements. Some [103] use both labeled and unlabeled data for training. Some [18][68] use multi-class methods. All these methods have certain limitations.

Recently, the classification method, such as Artificial Neural Networks [42], Bayesian Analysis [6], etc., has become popularly in RF algorithms. However, the traditional classification and RF are definitely different because the user would not like to provide a large number of marked samples. To overcome this problem, small sample learning methods [27][37][52][54][55][84][85][91] are proposed in CBIR RF. Support vector machine (SVM) [26][58][67][95] and discriminant analysis (DA) [90][91] are two small sample learning methods used in CBIR RF in the recent years and obtaining the-state-of-the-art performance.

1.2 SVM based RF in CBIR

SVM [84][85] is an approximate implementation of the structure risk minimization in statistical learning theory [84][85]. It was successfully used in CBIR in the last two years. SVM classifies the relevant samples and irrelevant samples based on the support vectors, which are automatically determined by the SVM learning algorithm.

However, the performance of SVM based RF is often poor when the number of labeled positive feedback samples is small. This is mainly due to three reasons:

- SVM classifier is unstable on a small size training set;
- SVM's optimal hyper-plane may be biased when the positive feedback samples are much less than the negative feedback samples;
- Over-fitting due to the fact that the feature dimension is much higher than the size of the training set.

In this thesis, we try to use random sampling techniques [56][79] to overcome these problems [21]. To address the first two problems, we propose an asymmetric bagging based SVM. For the third problem, we combine the random subspace method (RSM) and SVM for RF [22]. Finally, by integrating bagging and RSM, we solve all the three problems, further improving the RF performance.

1.3 DA based RF in CBIR

DA [27][52] is another way to model CBIR RF. In the last two years, Fisher linear discriminant analysis (LDA) has been successfully used in face recognition [14][51][53]. It also can be used as a RF algorithm [91] for CBIR with a similar way to face recognition. LDA extracts the discriminant subspace in the low-level feature space to distinct the relevant and irrelevant samples. Then the remaining images in the database are projected into the subspace. Finally, the CBIR system uses some similarity measures to sort these images.

However, LDA based RF considers the positive and negative feedback examples equivalently. This is a lethal drawback because all positive examples are alike and each negative example is negative in its own way. With the observation, biased discriminant analysis (BDA) was developed to scale the performance of CBIR and obtained satisfactory results. In the BDA model, the negative feedbacks are required to keep away from the center of positive feedbacks. Although BDA achieves the state-of-the-art performance, it also meets many problems:

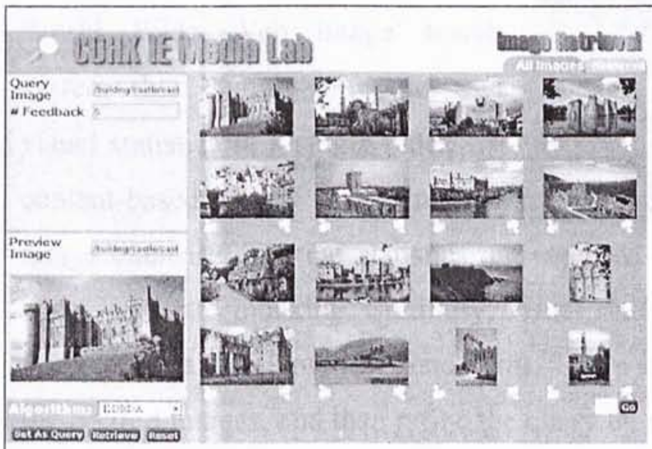
- For the BDA, the regularization method is used. It is well known that this method often encounters the Matrix Singular Problem (MSP) or the Small Sample Size (SSS) problem.
- BDA assumes all positive feedbacks from a single Gaussian distribution which may not be the case for CBIR.
- Although kernel BDA (KBDA) can circumvent the single Gaussian distribution assumption to some extent, the kernel parameter tuning makes the online learning unfeasible.

Motivated by the successful direct method [31][43] and null-space method [57] used in LDA to solve the SSS problem, we generalize them into the kernel Hilbert space to over come the SSS problem in KBDA [15][16]. Because direct LDA method and null-space method may lose some discriminant information, we propose a new full-space method [16] to contain all discriminant information both in linear space and in kernel Hilbert Space.

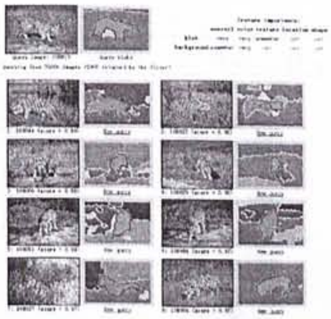
To avoid the parameter tuning problem and the single Gaussian distribution assumption in BDA, we construct a new nonparametric discriminant analysis (NDA) [19] for RF in CBIR. We then generalize the regularization method, the direct method, the null-space method, and the full-space method to address the SSS problem in NDA.

1.4 Existing CBIR Engines

CBIR for a general purpose image database is a challenging issue because the size of the database may be very large, understanding image contents is tough by a computer, and performance evaluation is difficult. Recently, a number of search engines were developed for general purpose image retrieval, such as IBM QBIC [59], VIRAGE [1], NEC AMORA [75], Bell Laboratory [5], PhotoBook [6] [24][82], Image Beagle [23], PicToSEEK [78], NETRA [86][87], WBIIS [49], etc. Here we show some image retrieval systems:



QueryGo [15][19][21] image retrieval system, developed in the department of information engineering at the Chinese University of Hong Kong, supports color, texture, and shape features. It can use different relevance feedback algotithms. Moreover, new feedback algorithms can be easily embedded into the system.



The BlobWorld [10] system supports color, shape, spatial, and texture features. It can segment each image into regions automatically, which correspond approximately to objects or parts of objects in an image. BlobWorld allows users to view the results of the segmentation of both the query image and returned results with highlight showing how the segmented features have influenced the retrieval results. The system allows querying at object level rather than on the whole image.

alltheweb
Web
News
Pictures
Video
Audio
FTP files
SEARCH
Currently searching 3,151,743,117 web pages
Copyright © 2003 Overture Services, Inc

AllTheWeb can retrieval images, audios, and Videos by text information. The system can be found at: www.alltheweb.com.

The C-BIRD system [103], developed at the Vision & Media Laboratory, Simon Fraser University, supports color, shape, and texture matching features.

Welcome to C-BIRD

C-BIRD stands for Content-Based Image Retrieval from Digital libraries

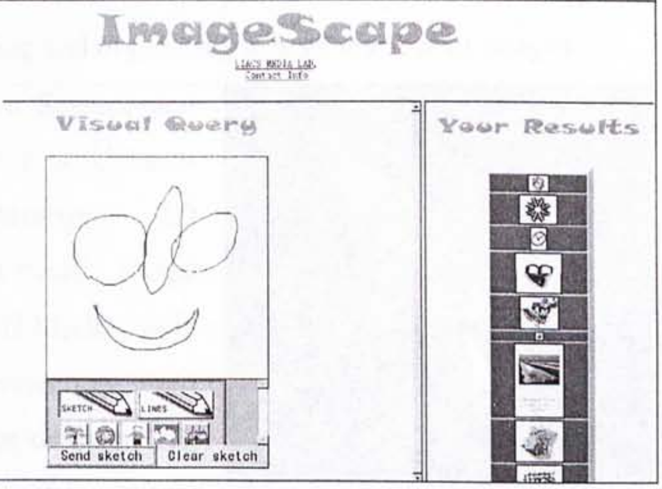
Enter the C-BIRD site using Java

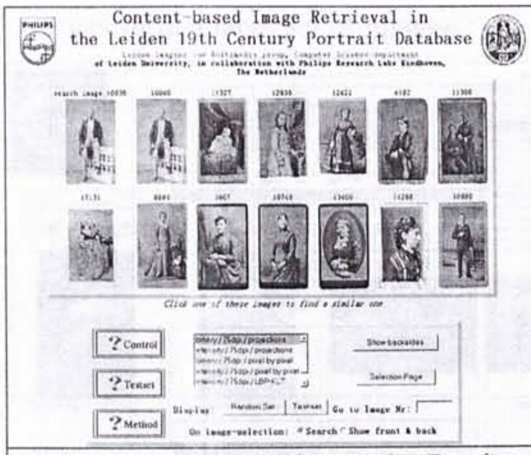
The ImageRover system [23], developed at Boston University, is a World Wide Web image search system that combines textual and visual statistics in a single index for content-based search of an Internet image database. Textual statistics are captured in a vector using a technique called latent semantic indexing. Similarly, visual statistics are captured in a feature vector using color and orientation histograms. Users initially specify keywords to describe the desired images, and then refine the query by relevance feedback.

ImageRover On-Line Demo

To select a relevant image, mark the check-box found next to the relevant image(s)...

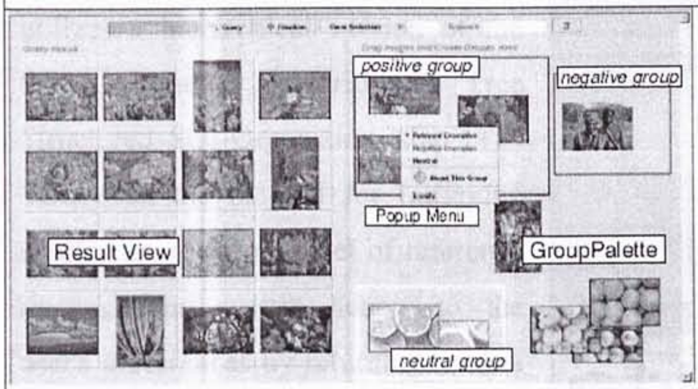
The ImageScape system [44] is a World Wide Web sketch image retrieval system.




Content-based Image Retrieval in
the Leiden 19th Century Portrait Database
of Leiden University, in collaboration with Philips Research Labs Eindhoven,
The Netherlands
Click one of these images to find a similar one
?Control
?Testset
?Method
Show backsides
Selection Page
Display:
Random Set
Testset
Go to Image Nr:
On image-selection: Search Show front & back

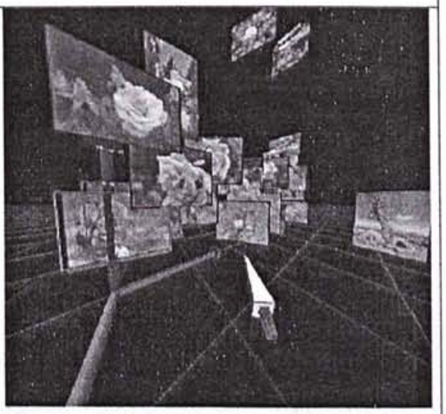
The Leiden 19th Century Portrait Database (LCPD) [33], developed at the Computer Science Department, Leiden university, supports shape and texture matching features for the retrieval of greyscale images.

Multimedia Analysis and Retrieval System (MARS) [100], developed at the Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, supports combinations of color, shape, spatial layout, and texture matching features.

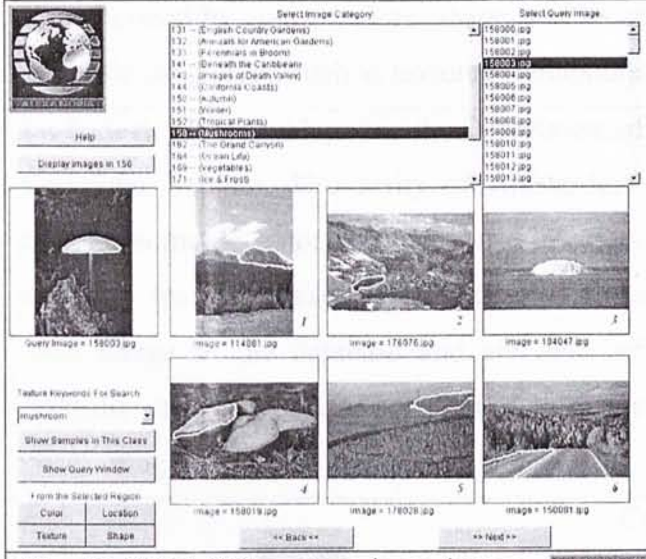


In ImageGrouper system [63] was developed at the Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign. Users can compare different combinations of query examples by dragging and grouping images on the workspace (Query-by-Group) interactively. Because the query results are displayed on another pane, the user can quickly review the results. Combining different queries is also easy. Furthermore, the concept of "image groups" is applied to annotating and organizing a large number of images.

The Beckman Institute for Advanced Science and Technology in University of Illinois at Urbana-Champaign proposed an interactive 3D visualization system for Content-based image retrieval named 3D MARS [64]. In 3D MARS, only relevant images are displayed on projection-based immersive Virtual Reality system or desktop VR. Based on the users' feedback, the system reorganizes its visualization scheme. 3D MARS eases tedious task of searching

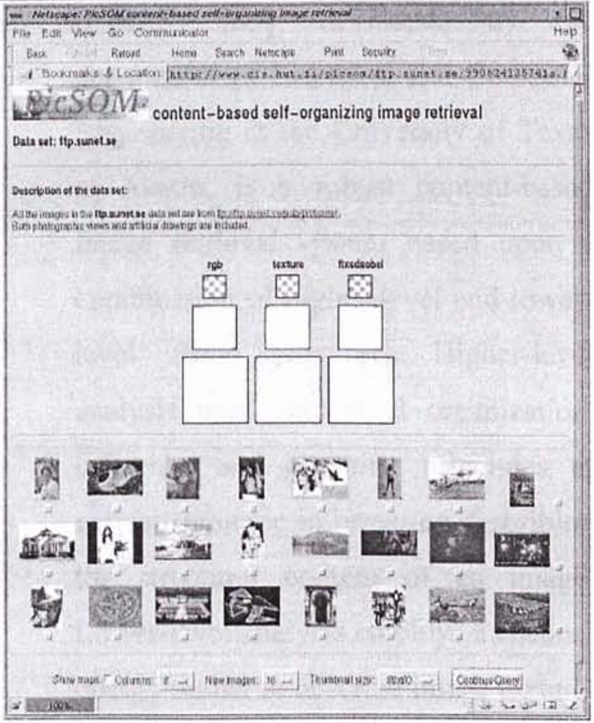


images from a large set of images.

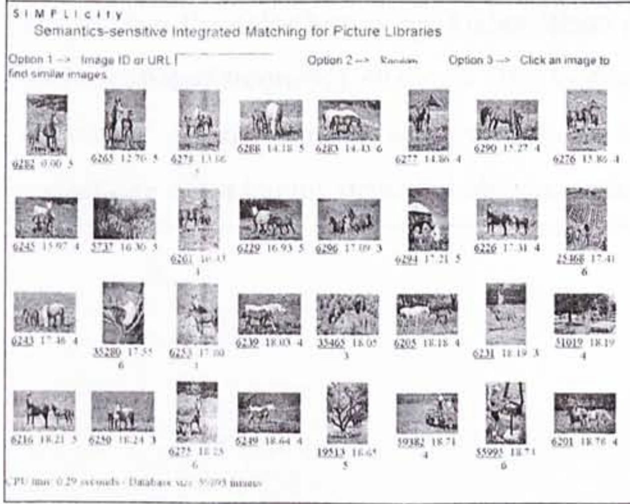


The NeTra system [87] supports colour, shape, spatial layout and texture matching features in segmented image regions to search and retrieve similar regions from an image database.

The PicSOM [42] system is an image browsing system based on the Self-Organizing Map (SOM). The system utilizes a hierarchical version of the SOM neural algorithm, Tree Structured Self-Organizing Map (TS-SOM), as the method for retrieving similar images from a set of reference images. The system adapts to the user's preferences by returning images from those SOMs where their responses have been most densely mapped.

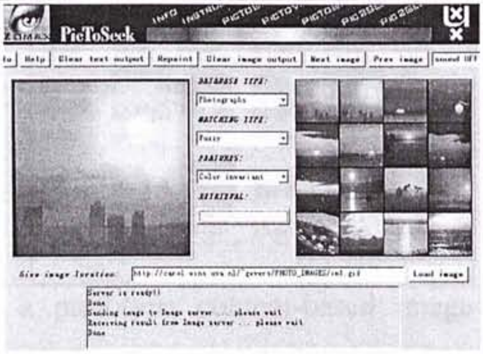

Netscape: PicSOM content-based self-organizing image retrieval
File Edit View Go Communicator
Help
Back Forward Reload Home Search Netscape Print Security Stop
Bookmarks Location: http://www.cis.hut.fi/picsom/ftp.sunet.se/990824135741s.l
PicSOM
content-based self-organizing image retrieval
Data set: ftp.sunet.se
Description of the data set:
rgb
texture
fixedsobel
Show maps Columns: 8 New images: 16 Thumbnail size: 80x80 Continue Query
100%

All the images in the ftp.sunet.se data set are from ftp://ftp.sunet.se/pub/pictures/.
Both photographic views and artificial drawings are included.


SIMPLIcity
Semantics-sensitive Integrated Matching for Picture Libraries
Option 1 --> Image ID or URL
Option 2 --> Random
Option 3 --> Click an image to find similar images
6282 0.00 5
6265 12.76 5
6278 13.16
6288 14.18 5
6283 14.43 6
6277 14.86 4
6290 15.27 4
6276 15.86 4
CPU time: 0.29 seconds - Database size: 59,895 images

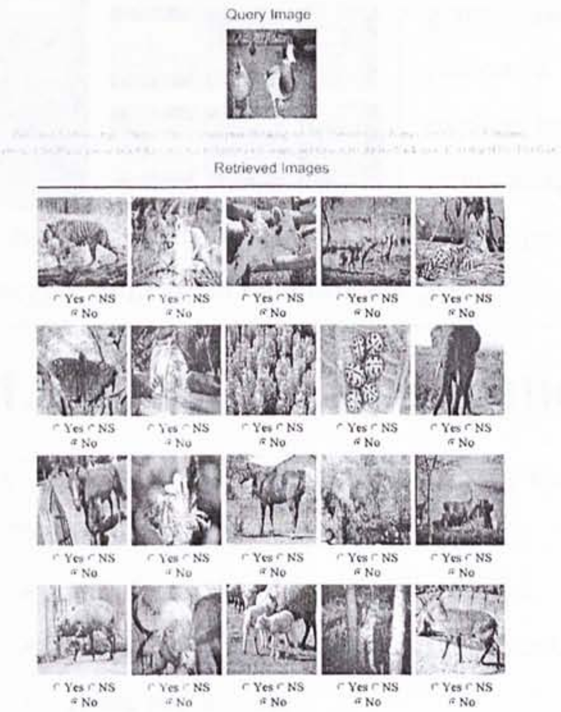
Stanford University presented the Semantics-sensitive Integrated Matching for Picture LIbraries (SIMPLIcity) [50], an image retrieval system, which uses semantics classification methods, a wavelet-based approach for feature extraction, and integrated region matching based upon image segmentation. The image is represented by a set of regions, roughly corresponding to objects, which are characterized by color, texture, shape, and location. The system classifies images into semantic categories, such as textured-nontextured, graph-photograph.

PicToSeek [78], developed in the Department of Computer Science, University of Amsterdam, uses photometric color and geometric invariant indices. Invariant features are extracted from each image in the database and are matched with the invariant feature set derived from the query image.



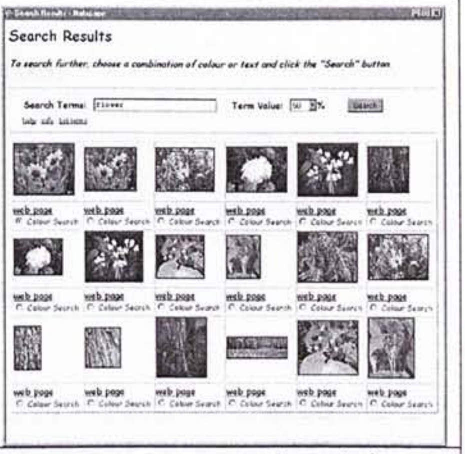
http://carol.wins.uva.nl/~gevers/PHOTO_IMAGES/...gif

```
Server is ready!!
Done
Sending image to Image server ... please wait
Receiving result from Image server ... please wait
Done
```



Content Based Image REtrieval System (CIRES) [69], developed by the department of Electrical and Computer Engineering at the University of Texas at Austin, is a robust content-based image retrieval system based upon a combination of higher-level and lower-level vision principles. Higher-level analysis uses perceptual organization, inference and grouping principles to extract semantic information describing the structural content of an image. Lower-level analysis employs a channel energy model to describe image texture, and utilizes the color histogram. Gabor filters are used to extract fractional energies in various spatial-frequency channels. The system is able to accept queries ranging from scenes of purely natural objects such as vegetation, trees, sky, etc. to images containing conspicuous structural objects such as buildings, towers, bridges, etc.

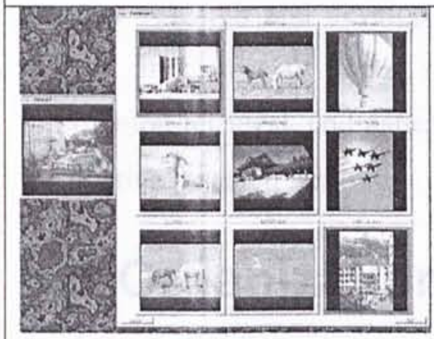
Effective WWW image retrieval systems are required to locate relevant images as more and more images used in HTML documents. Lu Guojun [32] described an approach integrating text based and content based techniques, to take advantage of their complementing strengths.



Search Results

To search further, choose a combination of colour or text and click the "Search" button

Search Terms: flower Term Value: 50 %



PicHunter [35], a prototype content-based image retrieval system, represents a simple instance of a general Bayesian framework we describe for using relevance feedback to direct a search. With an explicit model of what users would do, given what target image they want, it uses Bayes's rule to predict what the target is. This is done via a probability distribution over possible image targets, rather than by refining a query.

1.5 Practical Applications of CBIR

A wide range of applications for CBIR technology has been identified [38][62][70][80]:

- architectural and engineering design,
- art galleries and museum management,
- crime prevention,
- cultural heritage,
- education and training,
- fabric and fashion design,
- geographical information systems,
- home entertainment,
- intellectual property,
- interior design,
- journalism and advertising,
- law enforcement and criminal investigation,
- medical image classification and diagnosis,

- picture archiving and communication systems,
- remote sensing and management of earth resources,
- retailing,
- scientific database management,
- the military,
- trademark and copyright database management,
- weather forecasting, and
- web searching.

Because research and develop most issues in CBIR spread on many different aspects and most of them share with image processing, computer vision, information retrieval, and patter recognition, the progress in CBIR can inspirit all the relative research fields.

1.6 Organization of this thesis

The rest of the thesis is organized as following. In Chapter 2, we review the statistical learning theory and its approximate implementation with SVM. In Chapter 3, we review the Principle Component Analysis (PCA), Kernel PCA (KPCA), LDA, BDA, and KBDA. We also prove that KPCA combined with BDA is KBDA. In Chapter 4, we develop the random sampling based method for SVM based RF. In Chapter 5, we propose the direct method, null-space method, and full-space method for KBDA to overcome the SSS problem. Then the NDA is developed in Chapter 6. After that, a medical image classification application is described in Chapter 7. Finally, the Chapter 8 draws the conclusions of the thesis.

Chapter 2

Statistical Learning Theory and Support Vector Machine

This chapter provides an introduction on the fundamental knowledge of the statistical learning theory [84] and the Support Vector Machine (SVM) [84], which are the main theoretical background in this thesis, and have been successfully applied in the pattern recognition and multimedia information retrieval in the last years. Theses introductions is also chapter dwells entirely on the object recognition, image segmentation, information retrieval, time-series prediction, text categorization, and all their extended related fields.

2.1 The Recognition Problem

We first consider the basic problem in pattern recognition [54]. Suppose we are given a set of observations generated from an unknown probability distribution $P(\mathbf{x}, y)$

$$\mathbf{X} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)\} \text{ with } \mathbf{x}_i \in R^n,\ y_i \in \{-1, +1\}. \tag{2-1}$$

and a class of functions

$$F = \{f \mid f : R^n \mapsto \{-1, +1\}\} \tag{2-2}$$

then the basic problem is to find a function $f \in F$ that minimize a risk function

$$R[f] = \int l(y - f(\mathbf{x}), \mathbf{x})\, dP(\mathbf{x}, y) \tag{2-3}$$

where l denotes a suitable loss function, such as $l(f(\mathbf{x}), y) = (f(\mathbf{x}) - y)^2$, which indicates how differences between y and $f(\mathbf{x})$ should be penalized.

As $P(\mathbf{x}, y)$ is always unknown, therefore we cannot evaluate $R[f]$ directly. One possible solution would be to estimate the density function $P(\mathbf{x}, y)$ from the samples $\mathbf{X}$ and many theoretical and practical techniques want exactly this by some way. It is well known that density estimation is difficult and depends greatly on the previous assumptions. If the size of $\mathbf{X}$ is small, it is always impossible to estimate $P(\mathbf{x}, y)$ well. One particular simple way is to minimize the empirical risk only

$$R_{emp} = \frac{1}{m}\sum_{i=1}^{m} l\left(y_i - f(\mathbf{x}_i), \mathbf{x}_i\right). \tag{2-4}$$

When the number of the training samples is asymptotical to $+\infty$, the empirical risk will converge to the real risk. However, in pattern recognition the size of the training set is limited. Consequently, to minimize the empirical risk will always lead to the over-fitting problem. A network or function f that is too complex may fit the noise, not just the signal, leading to **over-fitting**.

Over-fitting is especially dangerous because it can easily lead to predictions that are far beyond the range of the training data with many of the common types of pattern recognition methods. Over-fitting can also produce wild predictions in many pattern recognition methods even with noise-free data. The over-fitting problem is caused by the over complex function f, which can represent the training set $\mathbf{X}$ well but cannot generalize to unseen examples. The converse leads to the under-fitting problem. A network or function f that is not sufficiently complex can fail to detect fully the signal in a complicated data set, leading to **under-fitting**.

Apparently, we need to control the complexity of the function set F to avoid the over-fitting problem and under-fitting problem. There are two methods to control the complexity of F, regularization and the structure risk minimization principle.

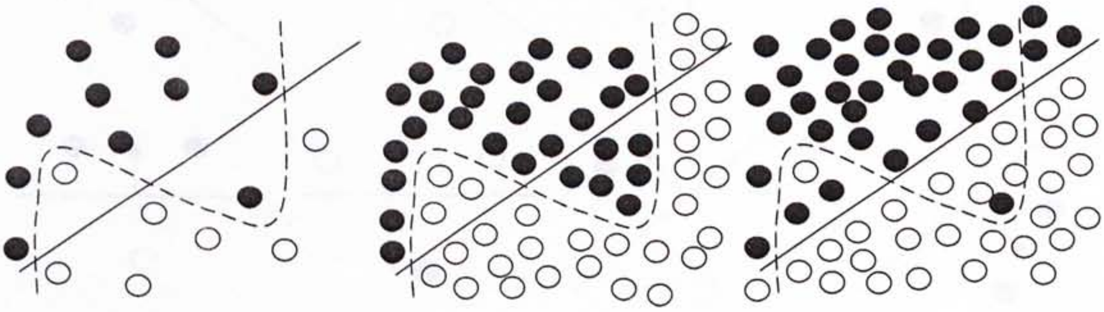


Figure 2-1. Illustration of the over-fitting dilemma: Given only a small sample (left) either, the solid or the dashed hypothesis might be true, the dashed one being more complex, but also having a smaller empirical risk. Only with a large sample we are able to see which decision reflects the true distribution more closely. If the dashed hypothesis is correct the solid would under-fit (middle); if the solid were correct the dashed hypothesis would over-fit (right).

2.2 Regularization

The method wants to minimize the empirical risk plus some penalty item, which is called the regularized risk:

$$R_{reg} = R_{emp} + \lambda\Omega(f)\,, \tag{2-5}$$

where $\Omega: F \to R^{+}$ is a regularization operator which measures the properties of the function f. The constant λ is used to control the trade-off between the empirical risk and the regularization.

2.3 The VC Dimension

Another way of controlling the complexity of F is given by the Vapnik-Chervonenkis (VC) theory [37]. The VC dimension is a property of a set of functions F. If a given set of m points can be labeled by $\{-1,+1\}$ in all possible 2^{m} ways, and for each labeling, a member of the set functions f can be found to correctly assign these labels, we say that set of points is *shattered* by F. VC dimension h of F is defined as the maximum number of training points that can be shattered by F.

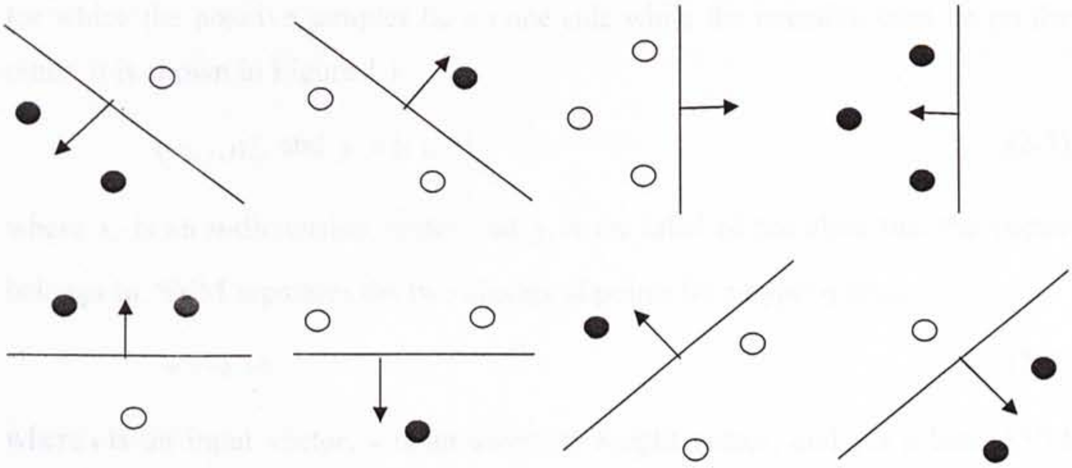


Figure 2-2. Three points in R^2, shattered by oriented lines.

Suppose the data belongs to R^2, and F consists of oriented straight lines, that is for a given line, all points on one side are assigned by 1, and the other size are assigned by -1. The maximum number of points can be shattered is 3. Consequently, the VC dimension of F (the set of oriented lines in R^2) is 3.

2.4 Structure Risk Minimization

With the definition of the VC dimension, we give out the following theorem, which is the infrastructure of the statistical learning theory.

Theorem (2-1). Let h denotes the VC-dimension of the function set F and R_{emp} is the empirical risk. For all $\delta > 0$ and $f \in F$ the following inequality bounding the risk

$$R(f) \le R_{emp}(f, \mathbf{X}) + \sqrt{\frac{h\left(\ln\frac{2m}{h}+1\right)-\ln\left(\delta/4\right)}{m}} \tag{2-6}$$

holds with probability of at least $1-\delta$ for $m > h$ over the random draw of the training samples $\mathbf{X}$.

2.5 Support Vector Machine

SVM is a learning algorithm used for various function estimation problems based on the structural risk minimization principle. The SVM creates a classifier with minimized Vapnik-Chervonenkis dimension and an upper bound on the generalization error rate. Consider a linearly separable binary classification problem (The training data is linearly separable if there exists a hyper-plane $(\mathbf{w}, b)$ for which the positive samples lie on one side while the negative ones lie on the other. It is shown in Figure 1.):

$$\{(x_i, y_i)\}_{i=1}^{N} \text{ and } y_i = \{+1, -1\} \tag{2-7}$$

where x_i is an n-dimension vector and y_i is the label of the class that the vector belongs to. SVM separates the two classes of points by a hyper-plane,

$$\mathbf{w}^T x + b = 0\,, \tag{2-8}$$

where $\mathbf{x}$ is an input vector, $\mathbf{w}$ is an adaptive weight vector, and b is a bias. SVM finds the parameters $\mathbf{w}$ and b for the optimal hyper-plane to maximize the geometric margin $2/\|\mathbf{w}\|$ subject to $y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \ge +1$, which will minimize a bound on the generalization error and will generalize best, regardless the dimension of the input space. That is we need to solve the following constrained minimization problem:

$$\begin{cases} \min\limits_{w,b} & \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ s.t. & y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1 \end{cases} \tag{2-9}$$

The solution can be found through a Wolfe dual problem with Lagrangian multiplie α_i :

$$Q(\alpha) = \sum_{i=1}^{m}\alpha_i - \sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j(\mathbf{x}_i \cdot \mathbf{x}_j)\Big/2 , \tag{2-10}$$

subject to $\alpha_i \geq 0$ and $\sum_{i=1}^{m}\alpha_i y_i = 0$.

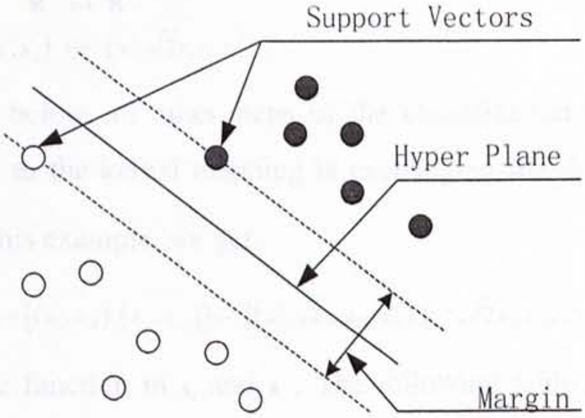


Figure 2-3. SVM for the linearly separable binary classes problem.

In the dual format, the data points only appear in the inner product. To get a potentially better representation of the data, the data points are mapped into the Hilbert Inner Product space through a replacement:

$$\mathbf{x}_i \cdot \mathbf{x}_j \rightarrow \phi(\mathbf{x}_i)\cdot\phi(\mathbf{x}_j) = K(\mathbf{x}_i, \mathbf{x}_j) , \tag{2-11}$$

where $K(.)$ is a kernel function. We then get the kernel version of the Wolfe dual problem:

$$Q(\alpha) = \sum_{i=1}^{m}\alpha_i - \sum_{i,j=1}^{m}\alpha_i\alpha_j d_i d_j K(\mathbf{x}_i \cdot \mathbf{x}_j)\Big/2 . \tag{2-12}$$

Thus for a given kernel function, the SVM classifier is given by

$$F(\mathbf{x}) = \operatorname{sgn}\left(f(\mathbf{x})\right) , \tag{2-13}$$

where $f(\mathbf{x}) = \sum_{i=1}^{l}\alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b$ is the output hyper-plane decision function of the SVM. In general, when $|f(\mathbf{x})|$ for a given pattern is high, the corresponding prediction confidence will be high. On the contrary, a low $|f(\mathbf{x})|$ of a given pattern means the pattern is close to the decision boundary and its corresponding prediction confidence will be low. Consequently, the output of SVM, $f(\mathbf{x})$ has been used to measure the dissimilarity between a given pattern and the query image, in traditional SVM based CBIR RF.

2.6 Kernel Space

Kernel method is to first process the data by some non-linear mapping Φ and then to apply the same linear algorithm in the kernel feature space.
For example:

$$\begin{aligned} \Phi: \quad & \mathbf{R}^2 \mapsto \mathbf{R}^3 \\ (x_1, x_2) & \mapsto \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right) \end{aligned} \tag{2-14}$$

Φ is carried out before all other steps of the classification methods. The only modification due to the kernel mapping is exchanging the dot product $(\mathbf{x}_i, \mathbf{x}_j)$ by $(\Phi(\mathbf{x}_i).\Phi(\mathbf{x}_j))$. In this example, we get:

$$(\mathbf{x}_i, \mathbf{x}_j) = \left((x_{i1}, x_{i2}), (x_{j1}, x_{j2})\right) \mapsto \left(\left(x_{i1}^2, \sqrt{2}x_{i1}x_{i2}, x_{i2}^2\right), \left(x_{j1}^2, \sqrt{2}x_{j1}x_{j2}, x_{j2}^2\right)\right) = K(\mathbf{x}_i, \mathbf{x}_j). \tag{2-15}$$

K is a symmetric function in $\mathbf{x}_i$ and $\mathbf{x}_j$. The following table shows some useful kernels:

Kernel Function Type	Kernel Function
Polynomial of Order p	$\left(1+(\mathbf{x}_i, \mathbf{x}_j)\right)^p$
Gaussian Radial Basis	$\exp\left(-\frac{\lVert \mathbf{x}_i - \mathbf{x}_j \rVert_2^2}{2\sigma^2}\right)$
Sigmoid Function	$\tanh\left(\kappa(\mathbf{x}_i, \mathbf{x}_j) - \theta\right)$

Chapter 3

Discriminant Analysis

In this Chapter, we first give the definitions of Principle Component Analysis (PCA) [27] and Kernel PCA (KPCA) [54], which are the base for discriminant analysis. Then we briefly discuss LDA [52], BDA [91], and KBDA [91]. Finally, we prove the KPCA combined with BDA is actually KBDA.

3.1 PCA

Given a set of N observations, $\mathbf{x}_k$, $k=1,\ldots N$, $\mathbf{x}_k \in \mathbf{R}^M$, PCA diagonalizes the covariance matrix in the input space:

$$\mathbf{C}=\frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i-\mathbf{m})(\mathbf{x}_i-\mathbf{m})^T. \tag{3-1}$$

where $\mathbf{m}=\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$.

To do this, one has to solve the Eigenvalue equation

$$\lambda\mathbf{v}=\mathbf{C}\mathbf{v} \tag{3-2}$$

for Eigenvalues $\lambda \geq 0$ and $\mathbf{v}\in\mathbf{R}^N\setminus\{0\}$.

3.2 KPCA

The section is devoted to a straightforward translation to a nonlinear scenario, in order to prepare the ground for the method. We shall now describe this computation in another dot product space $\mathbf{F}$, which is related to the input space by a possibly nonlinear map:

$$\begin{cases}\phi:\mathbf{R}^M\mapsto\mathbf{F}\\ \quad\mathbf{x}\mapsto\phi(\mathbf{x}).\end{cases} \tag{3-3}$$

Note that $\mathbf{F}$, which is referred to as the *kernel space* or the *feature space*, could have an arbitrarily large, possibly infinite, dimensionality.

Given a set of N observations, $\phi(\mathbf{x}_k)$, $k=1,\ldots N$, $\phi(\mathbf{x}_k)\in\mathbf{F}$, KPCA diagonalizes the covariance matrix in the feature space:

$$\begin{aligned}\mathbf{C} \mapsto \mathbf{C}^{\phi} &= \frac{1}{N}\sum_{i=1}^{N}\left(\phi(\mathbf{x}_i)-\mathbf{m}^{\phi}\right)\left(\phi(\mathbf{x}_i)-\mathbf{m}^{\phi}\right)^{T} \\ &= \frac{1}{N}\sum_{i=1}^{N}\left(\phi_i-\frac{1}{N}\sum_{j=1}^{N}\phi_j\right)\left(\phi_i-\frac{1}{N}\sum_{j=1}^{N}\phi_j\right)^{T}\end{aligned} \tag{3-4}$$

where $\mathbf{x}_i \mapsto \phi(\mathbf{x}_i)=\phi_i$, $\mathbf{m} \mapsto \mathbf{m}^{\phi}=\frac{1}{N}\sum_{i=1}^{N}\phi(\mathbf{x}_i)=\frac{1}{N}\sum_{i=1}^{N}\phi_i=\frac{1}{N}\mathbf{\Phi}\mathbf{1}_N^T$,

and $\mathbf{x}_i^T \cdot \mathbf{x}_j \mapsto \phi(\mathbf{x}_i)^T \cdot \phi(\mathbf{x}_j)=k(\mathbf{x}_i,\mathbf{x}_j)=k_{ij}$.

To do this, one has to solve the Eigenvalue equation

$$\lambda \mathbf{v}^{\phi}=\mathbf{C}^{\phi}\mathbf{v}^{\phi} \tag{3-5}$$

for Eigenvalues $\lambda \geq 0$ and $\mathbf{v}^{\phi} \in \mathbf{F}\setminus\{0\}$.

From the solutions of the Eigenvalue problem in the input space, the solutions $\mathbf{v}^{\phi}$ lie in the span of $\phi_1,\phi_2,\dots\phi_N$, which means $\mathbf{v}^{\phi}=\sum_{i=1}^{N}\alpha_i\phi_i=\mathbf{\Phi}\boldsymbol{\alpha}$,

where

$$\boldsymbol{\alpha}=\begin{bmatrix}\alpha_1\\ \dots\\ \alpha_N\end{bmatrix},\ \Phi=\begin{bmatrix}\phi_1 & \cdots & \phi_N\end{bmatrix}. \tag{3-6}$$

$$\begin{aligned}\left(\mathbf{v}^{\phi}\right)^{T}\mathbf{C}^{\phi}\mathbf{v}^{\phi} &= \boldsymbol{\alpha}^{T}\mathbf{\Phi}^{T}\mathbf{C}^{\phi}\mathbf{\Phi}\boldsymbol{\alpha} \\ &= \boldsymbol{\alpha}^{T}\mathbf{\Phi}^{T}\begin{bmatrix}\phi_1-\mathbf{m}^{\phi} & \dots & \phi_N-\mathbf{m}^{\phi}\end{bmatrix}\begin{bmatrix}\phi_1-\mathbf{m}^{\phi}\\ \dots\\ \phi_N-\mathbf{m}^{\phi}\end{bmatrix}\mathbf{\Phi}\boldsymbol{\alpha} \\ &= \boldsymbol{\alpha}^{T}\mathbf{\Phi}^{T}\left(\mathbf{\Phi}-\mathbf{m}^{\phi}\mathbf{1}_N^T\right)\left(\mathbf{\Phi}-\mathbf{m}^{\phi}\mathbf{1}_N^T\right)^{T}\mathbf{\Phi}\boldsymbol{\alpha} \\ &= \left(\boldsymbol{\alpha}^{T}\mathbf{\Phi}^{T}\left(\mathbf{\Phi}-\mathbf{m}^{\phi}\mathbf{1}_N^T\right)\right)\left(\boldsymbol{\alpha}^{T}\mathbf{\Phi}^{T}\left(\mathbf{\Phi}-\mathbf{m}^{\phi}\mathbf{1}_N^T\right)\right)^{T} \\ &= \left(\boldsymbol{\alpha}^{T}\mathbf{\Phi}^{T}\mathbf{\Phi}-\boldsymbol{\alpha}^{T}\mathbf{\Phi}^{T}\mathbf{m}^{\phi}\mathbf{1}_N^T\right)\left(\boldsymbol{\alpha}^{T}\mathbf{\Phi}^{T}\mathbf{\Phi}-\boldsymbol{\alpha}^{T}\mathbf{\Phi}^{T}\mathbf{m}^{\phi}\mathbf{1}_N^T\right)^{T} \\ &= \left(\boldsymbol{\alpha}^{T}\mathbf{K}-\boldsymbol{\alpha}^{T}\mathbf{\Phi}^{T}\frac{1}{N}\mathbf{\Phi}\mathbf{1}_N\mathbf{1}_N^T\right)\left(\boldsymbol{\alpha}^{T}\mathbf{K}-\boldsymbol{\alpha}^{T}\mathbf{\Phi}^{T}\frac{1}{N}\mathbf{\Phi}\mathbf{1}_N\mathbf{1}_N^T\right)^{T} \\ &= \left(\boldsymbol{\alpha}^{T}\mathbf{K}-\frac{1}{N}\boldsymbol{\alpha}^{T}\mathbf{K}\mathbf{1}_{NN}\right)\left(\boldsymbol{\alpha}^{T}\mathbf{K}-\frac{1}{N}\boldsymbol{\alpha}^{T}\mathbf{K}\mathbf{1}_{NN}\right)^{T} \\ &= \boldsymbol{\alpha}^{T}\left(\mathbf{K}-\frac{1}{N}\mathbf{K}\mathbf{1}_{NN}\right)\left(\mathbf{K}-\frac{1}{N}\mathbf{K}\mathbf{1}_{NN}\right)^{T}\boldsymbol{\alpha} \\ &= \boldsymbol{\alpha}^{T}\left(\mathbf{K}\mathbf{K}^{T}-\frac{1}{N}\mathbf{K}\mathbf{1}_{NN}\mathbf{K}^{T}\right)\boldsymbol{\alpha}\end{aligned} \tag{3-7}$$

Then we need to solve the following Eigenvalue problem:

$$\left(\mathbf{v}^{\phi}\right)^{T}\mathbf{C}^{\phi}\mathbf{v}^{\phi}=\left[\lambda_i\right]_{i=1,\dots,N}, \tag{3-8}$$

i.e.

$$\boldsymbol{\alpha}^{T}\left(\mathbf{K}\mathbf{K}^{T}-\frac{1}{N}\mathbf{K}\mathbf{1}_{NN}\mathbf{K}^{T}\right)\boldsymbol{\alpha}=\mathbf{\Lambda}. \tag{3-9}$$

The projection is:

$$y = \mathbf{v}^T \phi(\mathbf{x}) = \left(\sum_{i=1}^{N} \alpha_i \phi(\mathbf{x}_i)\right)^T \phi(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i \phi^T(\mathbf{x}_i)\phi(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i k(\mathbf{x}_i, \mathbf{x}) . \tag{3-10}$$

3.3 LDA

LDA tries to find the subspace best discriminating different classes. It is spanned by a set of vectors $\mathbf{W}$ maximizing the ratio between the within-class scatter matrix $\mathbf{S}_w$ and the between-class scatter matrix $\mathbf{S}_b$,

$$\mathbf{W} = \arg\max_{\mathbf{W}} \frac{\left\|\mathbf{W}^T \mathbf{S}_b \mathbf{W}\right\|}{\left\|\mathbf{W}^T \mathbf{S}_w \mathbf{W}\right\|} . \tag{3-11}$$

Let the training set contain c individual classes and each class C_i has N_i samples. Then $\mathbf{S}_w$ and $\mathbf{S}_b$ are defined as,

$$\begin{cases} \mathbf{S}_b = \frac{1}{N}\sum_{i=1}^{c} N_i (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T \\ \mathbf{S}_w = \frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N_i} (\mathbf{x}_j^i - \mathbf{m}_i)(\mathbf{x}_j^i - \mathbf{m}_i)^T , \mathbf{x}_j^i \in C^i \end{cases} \tag{3-12}$$

where $\mathbf{m} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i$ is the mean vector of the total training set, $\mathbf{m}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} \mathbf{x}_j^i$ is the mean vector for the individual class C_i, and $\mathbf{x}_j^i$ is the sample belonging to class C_i. $\mathbf{W}$ can be computed from the eigenvectors of $\mathbf{S}_w^{-1}\mathbf{S}_b$. If c is 2, LDA changed into Fisher discriminant analysis, otherwise, multiple discriminant analysis.

3.4 BDA

Based on "all positive examples are alike, each negative example is negative in its own way", Zhou developed Biased discriminant analysis (BDA). BDA defines the *(1+x)-class classification problem*, which means there is an unknown number of classes but the user only concerns one class.

BDA tries to find the subspace to discriminate the positive (the only class concerned by the user) and negative samples (unknown number of classes). It is spanned by a set of vectors $\mathbf{W}$ maximizing the ratio between the positive covariance matrix $\mathbf{S}_x$ and the biased matrix $\mathbf{S}_y$,

$$\mathbf{W} = \arg\max_{\mathbf{W}} \frac{\left\|\mathbf{W}^T \mathbf{S}_y \mathbf{W}\right\|}{\left\|\mathbf{W}^T \mathbf{S}_x \mathbf{W}\right\|} . \tag{3-13}$$

Let the training set contains N_x positive and N_y negative samples. Then $\mathbf{S}_x$ and $\mathbf{S}_y$ are defined as,

$$\begin{cases} \mathbf{S}_x = \sum_{i=1}^{N_x} (\mathbf{x}_i - \mathbf{m}_x)(\mathbf{x}_i - \mathbf{m}_x)^T \\ \mathbf{S}_y = \sum_{i=1}^{N_y} (\mathbf{y}_i - \mathbf{m}_x)(\mathbf{y}_i - \mathbf{m}_x)^T, \end{cases} \tag{3-14}$$

where $\mathbf{x}_i$ denote the positive samples, $\mathbf{y}_i$ denote the negative samples, $\mathbf{m}_x = \frac{1}{N_x}\sum_{i=1}^{N_x}\mathbf{x}_i$ is the mean vector of the positive samples, and $\mathbf{w}$ can be computed from the eigenvectors of $\mathbf{S}_x^{-1}\mathbf{S}_y$. Firstly, BDA minimize the variance of the positive samples. Then BDA maximize the distance between the center of the positive feedbacks and all negative feedbacks. BDA maximize the distance between the center of the positive feedbacks and all negative feedbacks.

3.5 KBDA

According to the non-linearity of the data and the successfully kernel method used in non-linear analysis, BDA was also generalized to its kernel version, named as kernel biased discriminant analysis (KBDA). To obtain the non-linear generalization, the nonlinear mapping:

$$\begin{aligned} \Phi &: \mathbf{R}^N \to \mathbf{F} \\ &\mathbf{x} \mapsto \Phi(\mathbf{x}) \end{aligned} \tag{3-15}$$

from the linear input space to nonlinear kernel feature space is used. Where the data $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in R^N$ is mapped into a potentially much higher dimensional feature space ***F***. For a given learning problem one now considers the same algorithm in ***F*** instead of $\boldsymbol{R}^N$. The idea behind KBDA is to perform the BDA in the feature space instead of the input space.

Let $\mathbf{S}_x^\phi$ and $\mathbf{S}_y^\phi$ be the "the positive with-in class scatter" and "the negative scatter with respect to positive centroid" matrices in the feature space ***F***. They can be respectively expressed as follows:

$$\begin{cases} \mathbf{S}_x^\phi = \sum_{i=1}^{Nx} \left(\varphi(\mathbf{x}_i) - \overline{\varphi}(\mathbf{x})\right)\left(\varphi(\mathbf{x}_i) - \overline{\varphi}(\mathbf{x})\right)^T = \Phi_x \Phi_x^T \\ \mathbf{S}_y^\phi = \sum_{i=1}^{Nx} \left(\varphi(\mathbf{y}_i) - \overline{\varphi}(\mathbf{x})\right)\left(\varphi(\mathbf{y}_i) - \overline{\varphi}(\mathbf{x})\right)^T = \Phi_y \Phi_y^T \end{cases} \tag{3-16}$$

$$\begin{cases} \Phi_x = \left[\left(\varphi(\mathbf{x}_1) - \overline{\varphi}(\mathbf{x})\right) \ \cdots \ \left(\varphi(\mathbf{x}_i) - \overline{\varphi}(\mathbf{x})\right) \ \cdots \ \left(\varphi(\mathbf{x}_{Nx}) - \overline{\varphi}(\mathbf{x})\right)\right] \\ \Phi_y = \left[\left(\varphi(\mathbf{y}_1) - \overline{\varphi}(\mathbf{x})\right) \ \cdots \ \left(\varphi(\mathbf{y}_i) - \overline{\varphi}(\mathbf{x})\right) \ \cdots \ \left(\varphi(\mathbf{y}_{Ny}) - \overline{\varphi}(\mathbf{x})\right)\right] \end{cases} \tag{3-17}$$

where $\overline{\varphi}(\mathbf{x}) = \frac{1}{N_x}\sum_{i=1}^{N_x}\varphi(\mathbf{x}_i)$ is the centroid of the positive samples, N_x is the number of positive samples, and N_y is the number of negative samples. KBDA determines a set of optimal discriminant basis vectors $\mathbf{W} = \{w_k\}_{k=1}^{m}$, which can be obtained to solve the following eigenvalue problem:

$$\mathbf{W} = \underset{\mathbf{W}}{\arg\max} \frac{\left\| \mathbf{W}^{\mathrm{T}} \mathbf{S}_y^{\phi} \mathbf{W} \right\|}{\left\| \mathbf{W}^{\mathrm{T}} \mathbf{S}_x^{\phi} \mathbf{W} \right\|}. \tag{3-18}$$

according to eigenvectors of $\mathbf{S}_x^{\phi-1} \mathbf{S}_y^{\phi}$.

The dimension of the feature space $\boldsymbol{F}$ is arbitrarily large, and possibly infinite. But we need not to use the exact $\Phi(x)$ to calculate $\boldsymbol{W}$, because the kernel method can be utilized to avoid to map the feature point from the linear input space to nonlinear kernel feature space based on replace the dot product with a kernel function in the input space $\boldsymbol{R}^{\mathrm{N}}$.

Theorem (3-1). Kernel PCA combined with BDA is Kernel BDA.

To prove KPCA combined with BDA equals to KBDA, we first use KPCA to project all samples in the training set to the empirical feature space, that is we will use the KPCA projection matrix to map the training set samples.

$$\varphi = \mathbf{v}\phi(\mathbf{z}) = \left(\sum_{i=1}^{N} \alpha_i \phi(\mathbf{z}_i)\right)\phi(\mathbf{z}) = \sum_{i=1}^{N} \alpha_i \phi(\mathbf{z}_i)\phi(\mathbf{z}) = \sum_{i=1}^{N} \alpha_i k(\mathbf{z}_i, \mathbf{z}) \tag{3-19}$$

$$\mathbf{S}_y = \sum_{i=Nx+1}^{N} \left(\varphi_i - \frac{1}{N_x}\sum_{l=1}^{Nx}\varphi_l\right)\left(\varphi_i - \frac{1}{N_x}\sum_{l=1}^{Nx}\varphi_l\right)^T \tag{3-20}$$

$$\mathbf{S}_x = \sum_{i=1}^{Nx} \left(\varphi_i - \frac{1}{N_x}\sum_{l=1}^{Nx}\varphi_l\right)\left(\varphi_i - \frac{1}{N_x}\sum_{l=1}^{Nx}\varphi_l\right)^T \tag{3-21}$$

$$\begin{aligned}
&\mathbf{S}_y \\
&= \sum_{i=Nx+1}^{N} \left(\sum_{j=1}^{N} \alpha_j k(\mathbf{z}_j, \mathbf{z}_i) - \frac{1}{N_x}\sum_{l=1}^{Nx}\sum_{j=1}^{N} \alpha_j k(\mathbf{z}_j, \mathbf{z}_l)\right)\left(\sum_{j=1}^{N} \alpha_j k(\mathbf{z}_j, \mathbf{z}_i) - \frac{1}{N_x}\sum_{l=1}^{Nx}\sum_{j=1}^{N} \alpha_j k(\mathbf{z}_j, \mathbf{z}_l)\right)^T \\
&= \sum_{i=Nx+1}^{N} \left(\sum_{j=1}^{N} \alpha_j \phi_j^T \phi_i - \frac{1}{N_x}\sum_{l=1}^{Nx}\sum_{j=1}^{N} \alpha_j \phi_j^T \phi_l\right)\left(\sum_{j=1}^{N} \alpha_j \phi_j^T \phi_i - \frac{1}{N_x}\sum_{l=1}^{Nx}\sum_{j=1}^{N} \alpha_j \phi_j^T \phi_l\right)^T \\
&= \sum_{i=Nx+1}^{N} \left((\mathbf{\Phi\alpha})^T \phi_i - \frac{1}{N_x}\sum_{l=1}^{Nx}(\mathbf{\Phi\alpha})^T \phi_l\right)\left((\mathbf{\Phi\alpha})^T \phi_i - \frac{1}{N_x}\sum_{l=1}^{Nx}(\mathbf{\Phi\alpha})^T \phi_l\right)^T \\
&= \sum_{i=Nx+1}^{N} \left((\mathbf{\Phi\alpha})^T \phi_i - \frac{1}{N_x}(\mathbf{\Phi\alpha})^T \mathbf{\Phi}_x \mathbf{1}_{Nx}^T\right)\left((\mathbf{\Phi\alpha})^T \phi_i - \frac{1}{N_x}(\mathbf{\Phi\alpha})^T \mathbf{\Phi}_x \mathbf{1}_{Nx}^T\right)^T \\
&= (\mathbf{\Phi\alpha})^T \left(\sum_{i=Nx+1}^{N} \left(\phi_i - \frac{1}{N_x}\mathbf{\Phi}_x \mathbf{1}_{Nx}^T\right)\left(\phi_i - \frac{1}{N_x}\mathbf{\Phi}_x \mathbf{1}_{Nx}^T\right)^T\right)(\mathbf{\Phi\alpha}) \\
&= (\mathbf{\Phi\alpha})^T \left(\mathbf{\Phi}_y - \frac{1}{N_x}\mathbf{\Phi}_x \mathbf{1}_{Nx}^T \mathbf{1}_{Ny}\right)\left(\mathbf{\Phi}_y - \frac{1}{N_x}\mathbf{\Phi}_x \mathbf{1}_{Nx}^T \mathbf{1}_{Ny}\right)^T (\mathbf{\Phi\alpha}) \\
&= \mathbf{\alpha}^T \left(\mathbf{\Phi}^T \mathbf{\Phi}_y - \frac{1}{N_x}\mathbf{\Phi}^T \mathbf{\Phi}_x \mathbf{1}_{Nx,Ny}\right)\left(\mathbf{\Phi}^T \mathbf{\Phi}_y - \frac{1}{N_x}\mathbf{\Phi}^T \mathbf{\Phi}_x \mathbf{1}_{Nx,Ny}\right)^T \mathbf{\alpha}
\end{aligned} \tag{3-22}$$

$$
\begin{aligned}
&\mathbf{S}_x \\
&=\sum_{i=1}^{Nx}\left(\sum_{j=1}^{N}\alpha_j k(\mathbf{z}_j,\mathbf{z}_i)-\frac{1}{N_x}\sum_{l=1}^{Nx}\sum_{j=1}^{N}\alpha_j k(\mathbf{z}_j,\mathbf{z}_l)\right)\left(\sum_{j=1}^{N}\alpha_j k(\mathbf{z}_j,\mathbf{z}_i)-\frac{1}{N_x}\sum_{l=1}^{Nx}\sum_{j=1}^{N}\alpha_j k(\mathbf{z}_j,\mathbf{z}_l)\right)^T \\
&=\sum_{i=1}^{Nx}\left(\sum_{j=1}^{N}\alpha_j\phi_j^T\phi_i-\frac{1}{N_x}\sum_{l=1}^{Nx}\sum_{j=1}^{N}\alpha_j\phi_j^T\phi_l\right)\left(\sum_{j=1}^{N}\alpha_j\phi_j^T\phi_i-\frac{1}{N_x}\sum_{l=1}^{Nx}\sum_{j=1}^{N}\alpha_j\phi_j^T\phi_l\right)^T \\
&=\sum_{i=1}^{Nx}\left((\mathbf{\Phi\alpha})^T\phi_i-\frac{1}{N_x}\sum_{l=1}^{Nx}(\mathbf{\Phi\alpha})^T\phi_l\right)\left((\mathbf{\Phi\alpha})^T\phi_i-\frac{1}{N_x}\sum_{l=1}^{Nx}(\mathbf{\Phi\alpha})^T\phi_l\right)^T \\
&=\sum_{i=1}^{Nx}\left((\mathbf{\Phi\alpha})^T\phi_i-\frac{1}{N_x}(\mathbf{\Phi\alpha})^T\mathbf{\Phi}_x\mathbf{1}_{Nx}^T\right)\left((\mathbf{\Phi\alpha})^T\phi_i-\frac{1}{N_x}(\mathbf{\Phi\alpha})^T\mathbf{\Phi}_x\mathbf{1}_{Nx}^T\right)^T \qquad (3\text{-}23) \\
&=(\mathbf{\Phi\alpha})^T\left(\sum_{i=Nx+1}^{N}\left(\phi_i-\frac{1}{N_x}\mathbf{\Phi}_x\mathbf{1}_{Nx}^T\right)\left(\phi_i-\frac{1}{N_x}\mathbf{\Phi}_x\mathbf{1}_{Nx}^T\right)^T\right)(\mathbf{\Phi\alpha}) \\
&=(\mathbf{\Phi\alpha})^T\left(\mathbf{\Phi}_x-\frac{1}{N_x}\mathbf{\Phi}_x\mathbf{1}_{Nx}^T\mathbf{1}_{Ny}\right)\left(\mathbf{\Phi}_x-\frac{1}{N_x}\mathbf{\Phi}_x\mathbf{1}_{Nx}^T\mathbf{1}_{Ny}\right)^T(\mathbf{\Phi\alpha}) \\
&=\mathbf{\alpha}^T\left(\mathbf{\Phi}^T\mathbf{\Phi}_x-\frac{1}{N_x}\mathbf{\Phi}^T\mathbf{\Phi}_x\mathbf{1}_{Nx,Ny}\right)\left(\mathbf{\Phi}^T\mathbf{\Phi}_x-\frac{1}{N_x}\mathbf{\Phi}^T\mathbf{\Phi}_x\mathbf{1}_{Nx,Ny}\right)^T\mathbf{\alpha}
\end{aligned}
$$

That is,

$$
\begin{aligned}
&\max\frac{\left\|\mathbf{W}^T\mathbf{S}_y^\phi\mathbf{W}\right\|}{\left\|\mathbf{W}^T\mathbf{S}_x^\phi\mathbf{W}\right\|}=\max\frac{\left\|\mathbf{W}^T\mathbf{S}_y\mathbf{W}\right\|}{\left\|\mathbf{W}^T\mathbf{S}_x\mathbf{W}\right\|} \qquad (3\text{-}24) \\
&\mathbf{\beta}=\alpha\mathbf{W}
\end{aligned}
$$

From these deductions, we can see that KPCA + BDA equals to KBDA, but we should reserve all the eigen-vectors of the KPCA procedure, otherwise, we will lose some discriminant information.

Chapter 4

Random Sampling Based SVM

Recently, classification-based RF has become a popular technique in CBIR and SVM based RF (SVMRF) has shown promising results owing to its good generalization ability [26][58][67][95]. However, when the number of positive feedbacks is small, the performance of SVMRF becomes poor. This is mainly due to the following reasons.

First, SVM classifier is unstable for small size training set, i.e. the optimal hyper-plane of SVM is sensitive to the training samples when the size of the training set is small. In SVM RF, the optimal hyper-plane is determined by the feedbacks. However, more often than not the users would only label a few images and cannot label each feedback accurately all the time. Hence the performance of the system may be poor with the inexactly labeled samples.

Second, in the RF process there are usually much more negative feedback samples than positive ones. Because of the imbalance of the training samples for the two classes, SVM's optimal hyper-plane will be biased toward the negative feedback samples. Consequently, SVMRF may mistake many query irrelevant images as relevant.

Finally, in RF, the size of the training set is much smaller than the dimension of the feature vector, thus may cause the over fitting problem. Because of the existence of noise, some features can only discriminant the positive and negative feedbacks but cannot discriminant the relevant or irrelevant images in the database. So the learned SVM classifier cannot work well for remnant images in database.

In order to overcome these problems, we design several new algorithms to improve the SVM based RF for CBIR. The key idea comes from the Classifier Committee Learning (CCL) [12][41][56][79][83]. Since each classifier has its own unique ability to classify relevant and irrelevant samples, the CCL can pool a number of weak classifiers to improve the recognition performance. We use bagging and random subspace method to improve the SVM since they are especially effective when the original classifier is not very stable.

4.1 Asymmetric Bagging SVM

Bagging [56] strategy incorporates the benefits of bootstrapping and aggregation. Multiple classifiers can be generated by training on multiple sets of samples that are produced by bootstrapping, i.e. random sampling with replacement on the training samples. Aggregation of the generated classifiers can then be implemented by majority voting rule (MVR) [41].

Experimental and theoretical results have shown that bagging can improve a good but unstable classifier significantly. This is exactly the case of the first problem of SVM based RF. However, directly using Bagging in SVM RF is not appropriate since we have only a very small number of positive feedback samples. To overcome this problem we develop a novel asymmetric Bagging strategy. The bootstrapping is executed only on the negative feedbacks, since there are far more negative feedbacks than the positive feedbacks. This way each generated classifier will be trained on a balanced number of positive and negative samples, thus solving the second problem as well. The Asymmetric Bagging SVM (ABSVM) algorithm is described in Table 1.

Table 4-1: Algorithm of Asymmetric Bagging SVM.

```
Input: positive training set $S^+$, negative training set $S^-$, weak classifier $I$ (SVM), integer $T$ (number of generated classifiers), $x$ is the test sample.
For $i = 1$ to $T$ {
    $S_i^- =$ bootstrap sample from $S^-$, with $|S_i^-| = |S^+|$.
    $C_i = I(S_i^-, S^+)$
}
$C^*(x) = aggregation\{C_i(x, S_i^-, S^+), 1 \le i \le T\}$.
Output: classifier $C^*$.
```

In ABSVM, the aggregation is implemented by Majority Voting Rule (MVR). The asymmetric Bagging strategy solves the classifier unstable problem and the training set unbalance problem. However, it cannot solve the small sample size problem. We will solve it by the Random Subspace Method (RSM) in the next section.

4.2 Random Subspace Method SVM

Similar to Bagging, RSM [79] also benefits from the bootstrapping and aggregation. However, unlike Bagging that bootstrapping training samples, RSM performs the bootstrapping in the feature space.

For SVM based RF, over fitting happens when the training set is relatively small compared to the high dimensionality of the feature vector. In order to avoid over fitting, we sample a small subset of features to reduce the discrepancy between the training data size and the feature vector length. Using such a random sampling method, we construct a multiple number of SVMs free of over fitting problem. We then combine these SVMs to construct a more powerful classifier. Thus the over fitting problem is solved. The RSM based SVM (RSVM) algorithm is described in Table 2.

Table 4-2: Algorithm of RSM SVM.

```
Input: feature set $\mathbf{F}$, weak classifier $I$ (SVM), integer $T$ (number of generated classifiers), $\mathbf{x}$ is the test sample.
1. For $i=1$ to $T$ {
2.     $\mathbf{F}_i$ = bootstrap feature from $\mathbf{F}$.
3.     $C_i = I(\mathbf{F}_i)$
4. }
5. $C^*(\mathbf{x}) = aggregation\{C_i(\mathbf{x},\mathbf{F}_i), 1 \le i \le T\}$.
Output: classifier $C^*$.
```

4.3 Asymmetric Bagging RSM SVM

Since the asymmetric Bagging method can overcome the first two problems of SVMRF and the RSM can overcome the third problem of the SVMRF, we should be able to integrate the two methods to solve all the three problems together. So we propose an Asymmetric Bagging RSM SVM (ABRSVM) to combine the two. The algorithm is described in Table 3.

Table 4-3: Algorithm of Asymmetric Bagging RSM SVM.

Input: positive training set $\mathbf{S}^{+}$, negative training set $\mathbf{S}^{-}$, feature set $\mathbf{F}$, weak classifier I (SVM), integer T_s (number of Bagging classifiers), integer T_f (number of RSM classifiers), $\mathbf{x}$ is the test sample.

```
1. For $j = 1$ to $T_s$ {
2.     $\mathbf{S}_j^{-}$ = bootstrap sample from $\mathbf{S}^{-}$.
3.     for $i = 1$ to $T_f$ {
4.         $\mathbf{F}_i$ = bootstrap sample from $\mathbf{F}$.
5.         $C_{i,j} = I(\mathbf{F}_i, \mathbf{S}_j^{-}, \mathbf{S}^{+})$.
6.     }
7. }
8. $C^{*}(\mathbf{x}) = aggregation\left\{\begin{matrix} C_{ij}(\mathbf{x}, \mathbf{F}_i, \mathbf{S}_j^{-}, \mathbf{S}^{+}) \\ 1 \le i \le T_f, 1 \le j \le T_s \end{matrix}\right\}$
```

Output: classifier C^{*}.

In order to explain why Bagging RSM strategy works, we derive the proof following a similar discussion on Bagging in [56].

Let $(y, \mathbf{x})$ be a data sample in the training set L with feature vector F, where y is the class label of the sample x. L is drawn from the probability distribution P. Suppose $\varphi(\mathbf{x}, L, F)$ is the simple predictor (classifier) constructed by the Bagging RSM strategy, and the aggregated predictor is,

$$\varphi_A(\mathbf{x}, P) = E_F E_L \varphi(\mathbf{x}, L, F). \tag{4-1}$$

Let random variables $(Y, \mathbf{X})$ be drawn from the distribution P independent of the training set L. The average predictor error, estimated by $\varphi(\mathbf{x}, L, F)$, is $e_a = E_F E_L E_{Y,\mathbf{X}}(Y - \varphi(\mathbf{X}, L, F))^2$. The corresponding error estimated by the aggregated predictor is

$$e_A = E_{Y,\mathbf{X}}(Y - \varphi_A(\mathbf{X}, P))^2. \tag{4-2}$$

Using the inequality $\frac{1}{M}\sum_{j=1}^{M}\frac{1}{N}\sum_{i=1}^{N}(z_{ij})^2 \ge \left(\frac{1}{M}\sum_{j=1}^{M}\frac{1}{N}\sum_{i=1}^{N} z_{ij}\right)^2$, we have:

$$E_F E_L \varphi^2(\mathbf{X}, L, F) \ge (E_F E_L \varphi(\mathbf{X}, L, F))^2 \tag{4-3}$$

$$E_{Y,X}E_F E_L \varphi^2(\mathbf{X},L,F) \geq E_{Y,X}\varphi_A^2(\mathbf{X},P) \tag{4-4}$$

Thus,

$$\begin{aligned} e_a &= E_{Y,X}Y^2 - 2E_{Y,X}Y\varphi_A + E_{Y,X}E_F E_L \varphi^2(\mathbf{X},L,F) \\ &\geq E_{Y,X}(Y-\varphi_A)^2 = e_A \end{aligned} . \tag{4-5}$$

Therefore, the predicted error of the aggregated method is reduced. From the inequality, we can see that the more diverse is the $\varphi(\mathbf{x},L,F)$, the more accurate is the aggregated predictor. In CBIR RF, the SVM classifier is unstable both for the training features and the training samples. Consequently, the Bagging RSM strategy can improve the performance.

Here we made an assumption that the average performance of all the individual classifier $\varphi(\mathbf{x},L,F)$, trained on a subset of feature and training set replica is similar to a classifier, which use the full feature set and the whole subset training set. This can be true when the size of feature and training data subset is adequate to approximate the full set distribution. Even when this is not true, the drop of accuracy for each simple classifier may be well compensated in the aggregation process.

From the inequality, we can see that the more diverse of the $\varphi(\mathbf{x},L,F)$, the more accurate of the aggregated predictor. Practically, the aggregated predictor is not $\varphi_A(\mathbf{x},P)$, but $\varphi_A(\mathbf{x},P')$, because the Bagging RSM strategy is used on the training set. P' and P are consistent in the probability space. If the classifier φ is stable, $\varphi_A(\mathbf{x},P')$ (it approximates to $\varphi(\mathbf{x},L,F)$) given by the Bagging RSM strategy is not as accurate as $\varphi_A(\mathbf{x},P)$. Therefore, the strategy may not work. However, if φ is unstable (weak classifiers are diverse), $\varphi_A(\mathbf{x},P')$can improve the performance. In CBIR RF, the SVM classifier is unstable both for the training features and the training samples. Consequently, the Bagging RSM strategy can improve the performance.

There are many different ways to do the aggregation. Two typical methods are hierocratic and parallel structures. The hierocratic structure of the aggregation is shown in Figure 1.

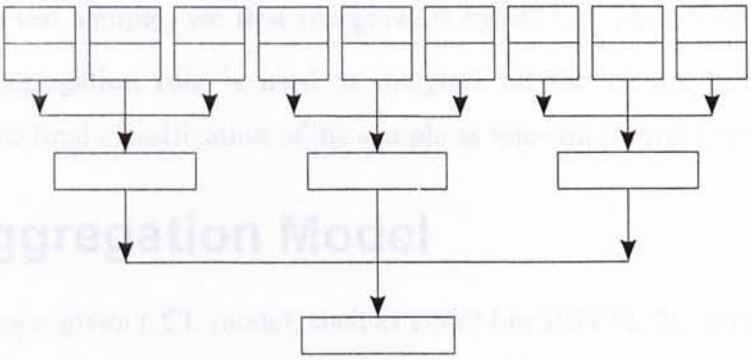


Figure 4-1. Hierocratic structure of aggregation.

For a given pattern, we first recognize it by a series of weak SVMs, which are constructed by the bootstrapping training set and features and denoted as $\{C_{ij}=C(F_i,S_j)|1\le i\le T_f,1\le j\le T_s\}$. Then we recognize it on a subset of weak classifiers $\{C_i=C(F_i,S_j)|1\le i\le T_f\}$, which are constructed on the same training examples but with different training features. At last, we use these outputs and the aggregation rule to construct the destination classifier. For example, if the aggregation rule is majority voting, we can represent it as:

$$C^*(x)=\arg\max_{y\in Y}\sum_{j|C_j(x,S_j)=\arg\max_{y\in Y}\left\{\sum_{i|C_{ij}(x,F_i,S_j)=y}1\right\}}1 \tag{4-6}$$

The parallel structure of the aggregation is shown in Figure 2.

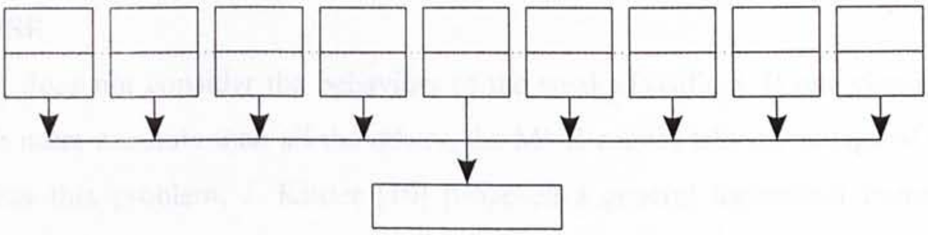


Figure 4-2. Parallel structure of aggregation.

For a given pattern, we recognize it by all weak SVMs $\{C_{ij}=C(F_i,S_j)|1\le i\le T_f,1\le j\le T_s\}$. Then, an aggregation rule is utilized to classify it as a query relevant or irrelevant. For example, if the aggregation rule is majority voting, we can represent it as:

$$C^*(x)=\arg\max_{y\in Y}\sum_{i,j|C_{ij}(x,F_i,S_j)=y}1 \tag{4-7}$$

For a given test sample, we first recognize it by all $T_f \cdot T_s$ weak SVM classifiers. Then, an aggregation rule is used to integrate all the results from the weak classifiers for final classification of the sample as relevant or irrelevant.

4.4 Aggregation Model

After training a given CCL model, such as BSVM or RSVM, the aggregation rule should be given to combine the weak classifiers. Many aggregation models have been developed, such as majority voting rule (MVR), Bayes sum rule (BSR), Bayes product rule, LSE-based weighting rule, double-layer combination, Dempster-Shafer model, and some nonlinear methods. In this paper, we only focus on the MVR and the BSR, due to their good performance in pattern classification.

1. MVR

MVR is the simplest method to combine multiple classifiers. Given a series of weak classifiers

$\{C_i(\mathbf{x}), 1 \le i \le N\}$, the MVR can be represented as:

$$C^*(\mathbf{x}) = \arg\max_{y \in \Gamma} \sum_{i: C_i(\mathbf{x}) = y} 1 . \tag{4-8}$$

MVR does not consider any individual behavior of each weak classifier. It only counts the largest number of classifiers that agree with each other.

2. BSR

MVR does not consider the behaviors of the weak classifiers. If one classifier is much more accurate than all the others, the MVR cannot take advantage of it. To address this problem, J. Kittler [10] proposed a general theoretical framework based on the Bayeian decision rule. We select BSR in the paper to aggregate multiple classifiers, because BSR outperform most of the other rules.

BSR, denotes the measurement vector used by the i^{th} classifier z_i. In the measurement space each class y_k is modeled by the probability density function $p(z_i \mid y_k)$ and its priori-probability is $p(y_k)$. Then BSR can be computed as,

$$C^*(\mathbf{x}) = \arg\max_k \left[(1\text{-R}) P(y_k) + \sum_{i=1}^{R} P(y_k \mid z_i) \right] . \tag{4-9}$$

To use the BSR in our schemes (BSVM, RSVM, and BRSVM), the probability model is required. As shown in [17], the sigmoid function combined with the output of SVM can be used to estimate the class-conditional probability for a given instance $\mathbf{x}$ by,

$$P(y_k \mid z_i) = 1 \Big/ \left\{1 + \exp\left(-\left|f_i(\mathbf{x})\right|\right)\right\}. \tag{4-10}$$

We do not need to consider $p(y_k)$ here, because the probability for an unknown sample to be query relevant or irrelevant are equal. Then BSR is simplified as,

$$C^*(\mathbf{x}) = \arg\max_k \left[\sum_{i=1}^{R} P(y_k \mid z_i)\right]. \tag{4-11}$$

4.5 Dissimilarity Measure

1. Using MVR to Combine the SVMs (MVRSVM)

For a given sample, we first use the MVR to recognize it as query relevant or irrelevant. Then we measure the dissimilarity between the sample and the query as the output of the individual SVM classifier, which gives the same label as the MVR and produces the highest confidence value (the absolute value of the decision function of the SVM classifier).

2. Using BSR to Combine the SVMs (BSRSVM)

For a given sample, we first use the BSR to recognize it as query relevant or irrelevant. Then we measure the dissimilarity between the sample and the query using the individual SVM classifier, which gives the same label as the BSR and has the highest confidence value.

3. BSR

From the definition of BSR, the output of the BSR $\sum_{i=1}^{R} P(y_k \mid x_i)$ can also be used as a dissimilarity measure between a given sample and the query.

In this chapter, we will compare all the three rules for BRSVM based RF.

4.6 Computational Complexity Analysis

From [12], we know that the computational complexity for training a SVM is $O(SVM) = O\left(n_s^3 + n_s^2 L + n_s n_f L\right)$, where n_s is the number of support vectors, n_f is feature dimension, and L is the size of the training set. From the formula of the output of SVM, the number of the support vectors n_s determines the computational complexity in the testing stage. We denote the computational complexity for a multiplication and addition of two real values as $\otimes$ and $\oplus$, respectively. Then the computational complexities of SVM, BSVM, RSVM, and BRSVM are:

Table 4-4: Algorithms' computational complexity.

	Training	Testing
SVM	$O(SVM)$	$N_s^{SVM} \cdot N_f^{SVM} \cdot (\otimes + \oplus)$
ABSVM	$T_s \cdot O(SVM)$	$T_s \cdot N_s^{ABSVM} \cdot N_f^{ABSVM} \cdot (\otimes + \oplus)$
RSMSVM	$T_f \cdot O(SVM)$	$T_f \cdot N_s^{RSMSVM} \cdot N_f^{RSMSVM} \cdot (\otimes + \oplus)$
ABRSVM	$T_s \cdot T_f \cdot O(SVM)$	$T_s \cdot T_f \cdot N_s^{ABRSVM} \cdot N_f^{ABRSVM} \cdot (\otimes + \oplus)$

4.7 QueryGo Image Retrieval System

In CBIR, we assume that the user is greedy, who expects the best possible retrieval results after each RF iterations, i.e. the search engine is required to feedback the most semantically relevant images under the previous feedback samples. Meanwhile, the user is impatient, who will never label a great deal of images in each RF iteration and only does a few numbers of iteration. To solve this type CBIR problem, the following CBIR framework QueryGo is proposed. With the proposed system, we can embed any RF algorithm easily.

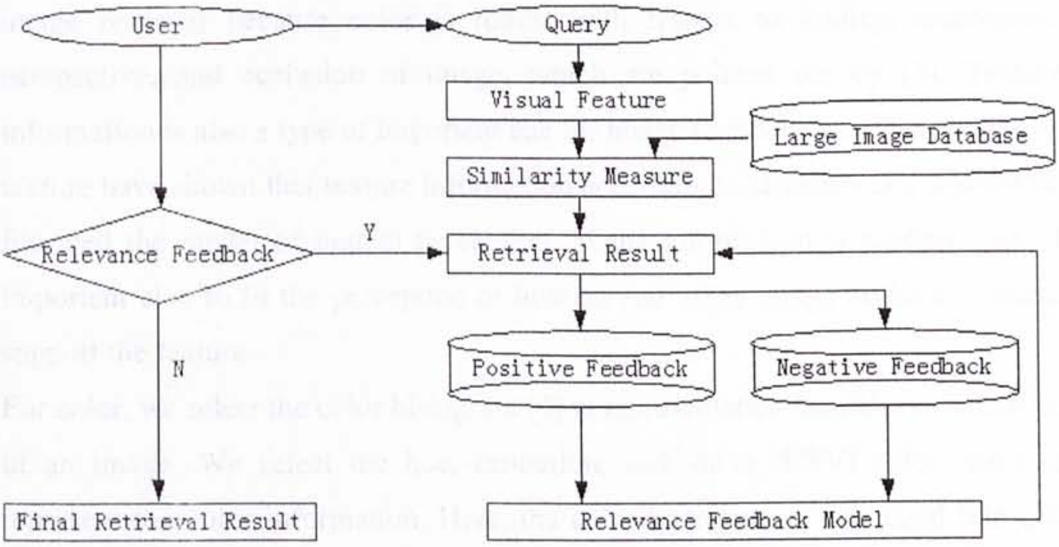


Figure 4-3. QueryGo system flow chart.

From Figure 3, when a query (image) is inputted, the low-level visual features are extracted. Then, all images in the database are sorted based on some similarity metric. If the user is satisfactory with the result, there will be no relevance feedback. However, most time, the RF is needed because of the poor performance. The user labels some top images as positive feedbacks and negative feedbacks. The user labels some top images as positive feedbacks and negative feedbacks. Using these feedback images, a RF model is trained based on certain machine learning algorithms. Then the similarity metric is updated based on the RF model. All the images are sorted again based on the renovated similarity metric. If the user is not content the result, the RF is done circularly, otherwise, the user get the final retrieval result.

The image retrieval system has been implemented with a real-world image database including 17,800 Corel images a subset of Corel Photo Gallery [39]. Corel Photo Gallery uses semantic concepts to group the photos each with 100 images. But we cannot directly use the concept information as the ground truth, because many images with similar concept bit given different label information. Meanwhile, some content absolutely dissimilar images given same label information. Because of these reasons, we re-labeled the 17, 800 images into 90 concepts.

In QueryGo, we represent images by three main features: color [3]-[5], texture [6]-[13], and shape [13]-[17]. Color information is the most important features for image retrieval because color is robust with respect to scaling, orientation, perspective, and occlusion of image, which are pointed out by [3]. Texture information is also a type of important cue for image retrieval. Previous studies on texture have shown that texture information according to structure and orientation fits well the model of human perception. Shape information is another type of important clue to fit the perception of human, and many image retrieval systems support the feature.

For color, we select the color histogram [3] to representation the color information of an image. We select the hue, saturation, and value (HSV) color space to represent the color information. Here, the color histogram is quantized into 256 levels. Because hue is the most important for human's perception, we quantized hue into 8 bins. Saturation and value are quantized into 4 bins respectively.

For texture, Wavelet texture is extracted from Y component in YCrCb space. We select the pyramid wavelet transform (PWT) [12] for image texture information representation. Image is decomposed by the traditional pyramid-type wavelet transform with Haar wavelet. In the system, the mean and standard deviation are calculated in terms of the sub-bands at each decomposed level. The decomposition procedure can be seen from the figure. PWT results in a feature vector of $2\times4\times3$ values.

For shape, the edge histogram [13] captures the spatial distribution of edges in an image. The distribution of edges is a good shape signature that is useful for image-to-image matching even when the underlying texture is not homogeneous. The edge histogram is calculated on Y component in YCrCb color space. Edges are grouped into five categories, which are horizontal, 45 diagonal, vertical, 135 diagonal, and isotropic. From the description of edge histogram, we can get a five-dimension shape feature for image retrieval.

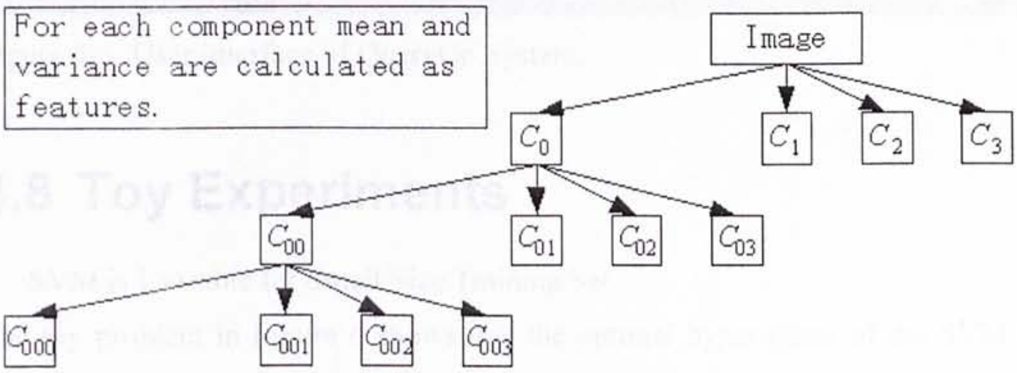
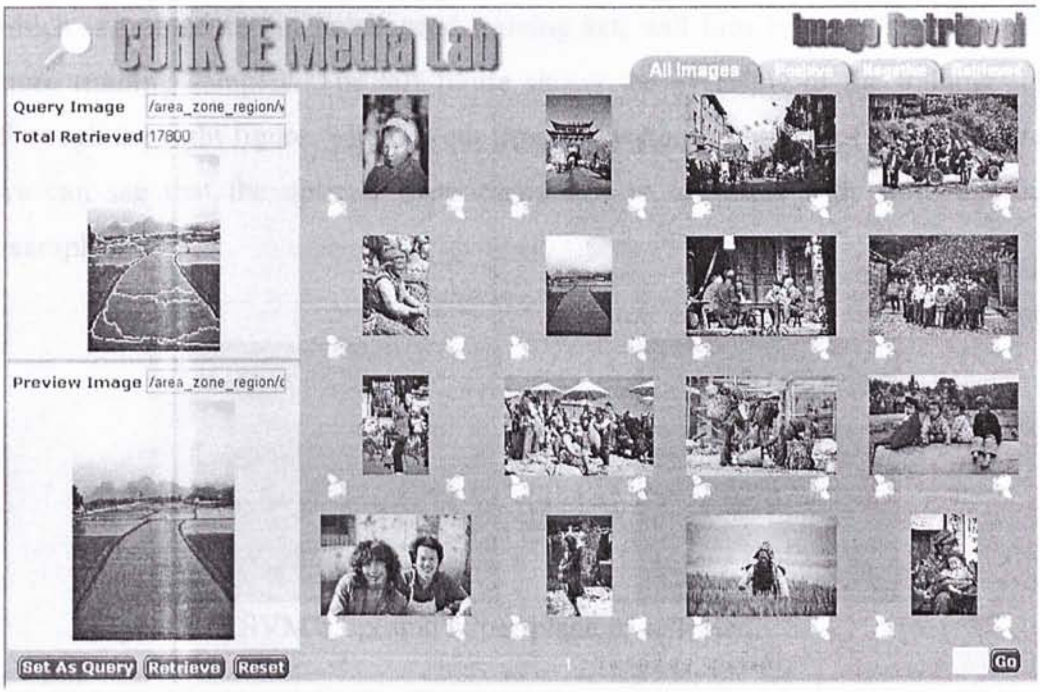


Figure 4-4. Wavelet texture feature structure.

Each feature has its own power to characterize a type of property of the content of an image. We combine the color, texture, and shape features into a feature vector, and then we normalize it into a normal distribution.

Figure 5 shows the user interface of QueryGo. In the paper, query by example is used. To scale the performance, the RF algorithms are focused here. First, user selects a query image from the thumbnail gallery and pushdown the "Set As Query" button. Second, user pushdown the "Retrieval" button, and then the images in the gallery are resorted. Third, user provides the feedback by clicking on the "thumb up" or "thumb down" button in terms of his judgment of the relevance of the retrieved image. At last, user pushdown the "Retrieval" button to resort the images in the gallery. The last two steps can be done iteratively to obtain a satisfactory performance.



Query Image /area_zone_region/\
Total Retrieved 17800
Preview Image /area_zone_region/(

Figure 4-5. User interface of QueryGo System.

4.8 Toy Experiments

1. SVM is Unstable for Small Size Training Set

The toy problem in Figure 6 shows that the optimal hyper-plane of the SVM is sensitive to the small changes of the training set. The left figure shows an optimal hyper-plane, which is trained by the original training set. The right figure shows a much different optimal hyper-plane, which is trained by the original training set with only one incremental pattern.

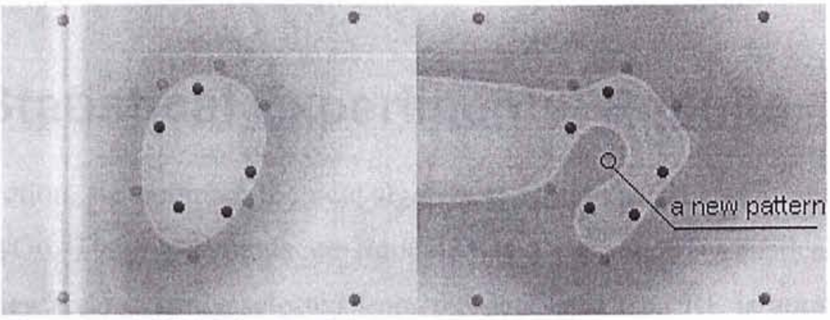


Figure 4-6. SVM is unstable.

2. SVM is Biased with Unbalanced Training Set

The toy problem in Figure 7 shows that the optimal hyper-plane of the SVM, which is trained by an unbalanced training set, will bias toward the class with more training samples. The left figure shows the overview of the training set. Through the right figure, which is cut from the bottom-right part of the left figure, we can see that the optimal hyper-plane bias to the class with more training examples.

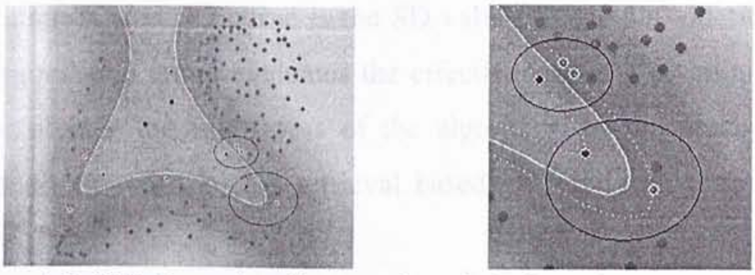


Figure 4-7. SVM's optimal hyper-plane is deflected.

3. The Visual Feature is Diverse for CBIR

This toy problem is constructed from the real data in RF. There are four positive and seven negative feedbacks. We randomly select two features to construct the SVM optimal hyper-plane for three times. They are visualized in Figure 8. We can see that the individual SVM classifiers are diverse with different features.

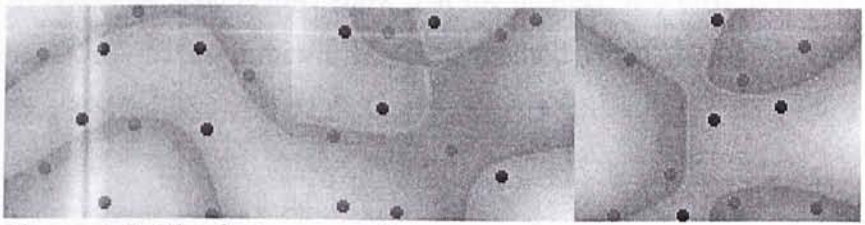


Figure 4-8. The features are diverse.

4.9 Statistical Experimental Results

In this section, we compare the new algorithms with existing algorithms through the QueryGo. The experiments are simulated by a computer automatically. First, 300 queries are randomly selected from the data, and then RF is automatically done by the computer: all query relevant images (i.e. images of the same concept as the query) are marked as positive feedbacks in the top 40 images and all the other images are marked as negative feedbacks. In general, we have about 5 images as positive feedbacks. The procedure is close to the real circumstances, because the user typically would not like to click on the negative feedbacks. Thus requiring the user to mark only the positive feedbacks in top 40 images is reasonable.

In this Chapter, precision and standard deviation (SD) are used to evaluate the performance of a RF algorithm. Precision is the percentage of relevant images in the top N retrieved images. The precision curve is the averaged precision values of the 300 queries, and SD curve is the SD values of the 300 queries' precision values. The precision curve evaluates the effectiveness of a given algorithm and SD curve evaluates the robustness of the algorithm. In the precision and SD curves 0 feedback refers to the retrieval based on Euclidean distance measure without RF.

We compare all the proposed algorithms with the original SVM based RF [5] and the constrained similarity measure SVM (CSVM) based RF [7]. We chose the Gaussian kernel $K(\mathbf{x},\mathbf{y}) = e^{-\rho\|\mathbf{x}-\mathbf{y}\|^2}$ with $\rho = 1$ (the default value in the OSU-SVM [15] MatLabTM toolbox) for all the algorithms. The performances of all the SVM algorithms are stable over a range of ρ.

1. Performance of Asymmetric Bagging SVM

Figure 9 shows the precision and SD values when using different number of SVMs in ABSVM. The results show that the number of SVMs will not affect the performance of the asymmetric Bagging method when the number of SVM is large enough.

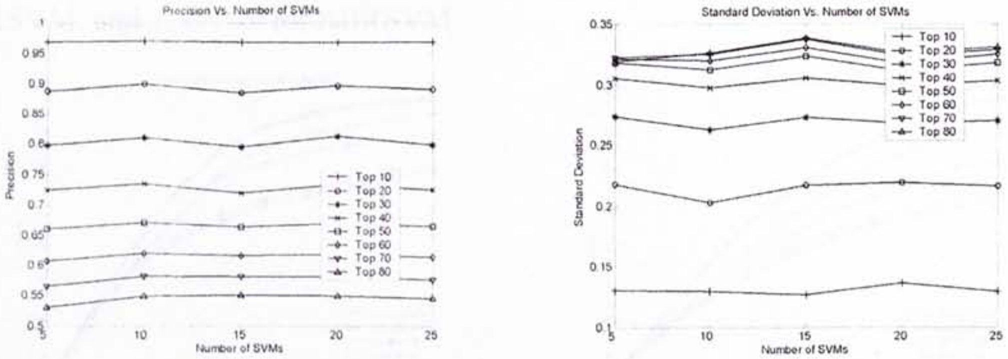


Figure 4-9. ABSVM in RF. The number of SVMs will not affect the performance of the asymmetric Bagging method when the number of SVM is enough.

Figure 12 evaluates the performance of the proposed ABRSVM based RF. In this experiment, we chose $T_s = 5$ for ABSVM. The results in Figure 12 show that the ABSVM gives a much better performance than SVM and CSVM.

2. Performance of RSM SVM

Figure 10 shows the precision and SD values when using different number of SVMs of RSVM. The results show that the number of SVMs does not affect the performance of RSVM when the number of SVMs is large enough.

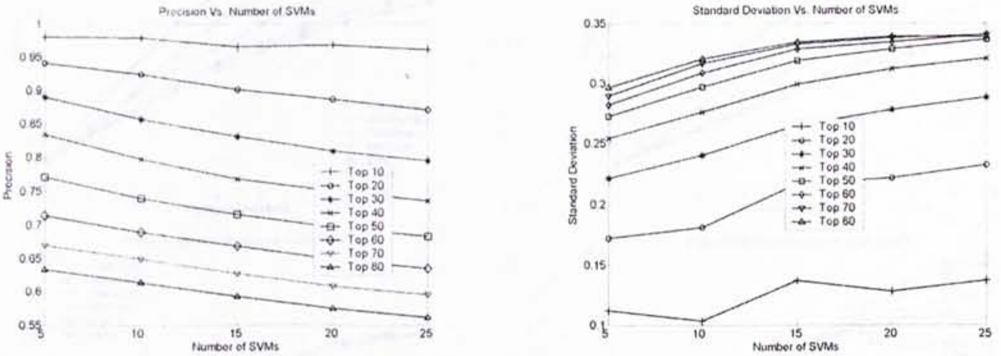
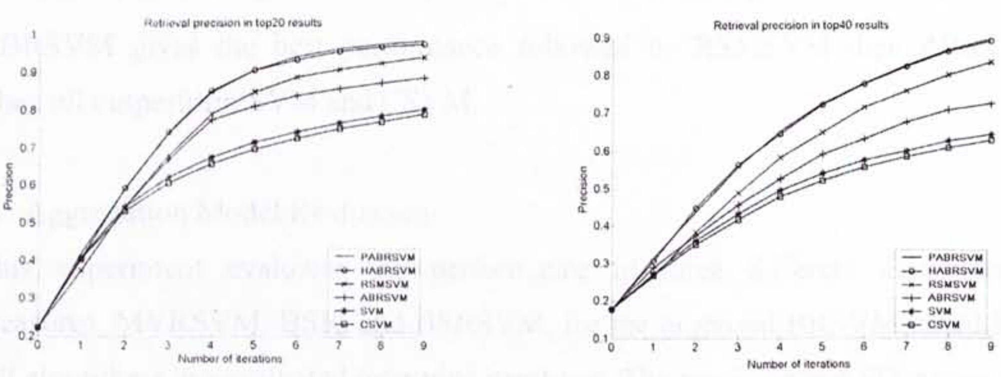


Figure 4-11. RSMSVM in RF. The number of SVMs does not affect the performance of RSVM when the number of SVMs is large enough.

Figure 12 evaluates the performance of the proposed RSMSVM based RF. In this experiment, we chose $T_f = 5$ for RSMSVM. The results in Figure 12 show that the RSMSVM gives a much better performance than SVM and CSVM.

3. Performance of Asymmetric Bagging RSM SVM

This experiment evaluates the performance of the proposed ABRSVM, ABSVM, and RSVM based RF. In this experiment, we chose $T_s = 5$ for ABSVM, $T_f = 5$ for RSVM, and $T_s = T_f = 5$ for ABRSVM.



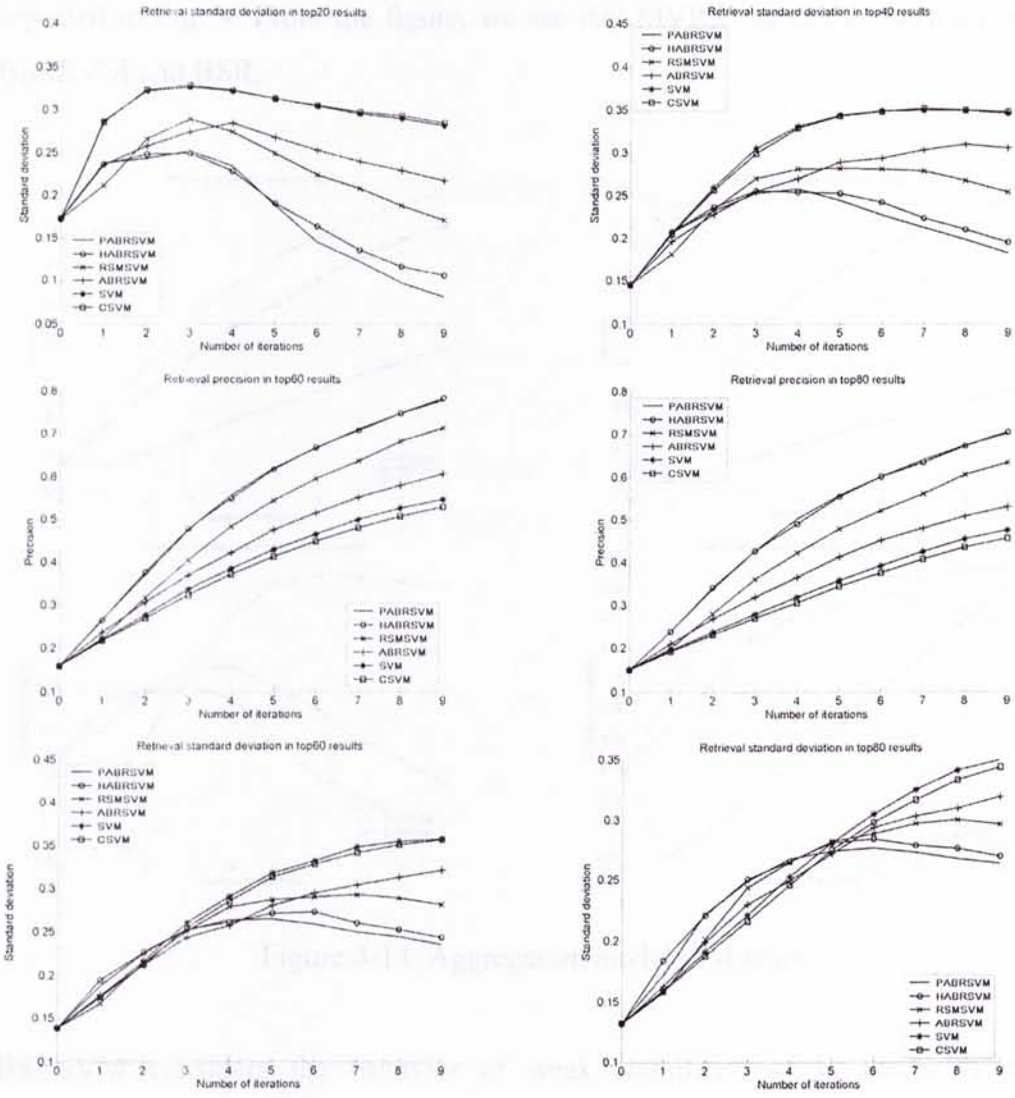


Figure 4-12. Performance of all proposed algorithms compared to existing algorithms. The algorithms are evaluated over 9 iterations.

This experiment evaluates the performance of the proposed ABRSVM, ABSVM, and RSMSVM based RF. In this experiment, we chose $T_s = 5$ for ABSVM, $T_f = 5$ for RSVM, and $T_s = T_f = 5$ for ABRSVM. The results in Figure 12 show that the ABRSVM gives the best performance followed by RSMSVM then ABSVM. They all outperform SVM and CSVM.

4. Aggregation Model Evaluation

This experiment evaluates the performance of three different dissimilarity measures, MVRSVM, BSR, and BSRSVM, for the proposed BRSVM based RF. All algorithms are evaluated over nine iterations. The precision and SD curves are reported in Fig. 9. From the figure, we see that MVRSVM can outperform both BSRSVM and BSR.

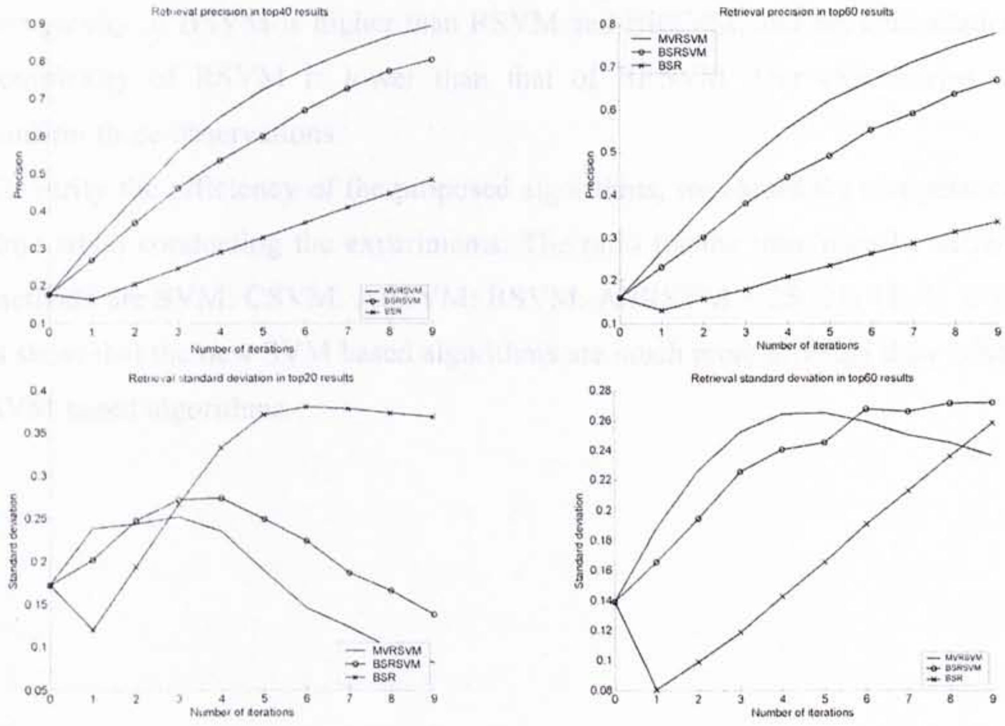


Figure 4-14. Aggregation model evaluation.

BSRSVM considers the behavior of weak classifiers, so in theory it may outperform the MVRSVM. However, according to the experimental results, its performance is worse than MVRSVM, because we cannot estimate the behavior exactly for unstable weak classifiers. From Figure 9, we find that BSR is much worse than MVRSVM and BSRSVM, because MVRSVM and BSRSVM choose the best individual SVM to measure the dissimilarity between a given image and the user's sentiment. BSR uses the averaged probabilities, which cannot be estimated exactly. Therefore, MVRSVM is the best choice.

5. Computational Complexity

Because the size of the training set is small, the overall computational complexity is mostly determined by the testing stage. The N_s for SVM RF is much bigger than that of BSVM and BRSVM, and the N_f of SVM RF is much bigger than that of RSVM and BRSVM. Thus the computational complexity of SVM RF is much higher than that of BSVM, RSVM, and BRSVM. In general, for each of the four algorithms, the inequality $N_f > N_s$ holds, because the number of feedbacks is much smaller than the dimension of the feature. Consequently, the computational complexity of BSVM is higher than RSVM and BRSVM, and the computational complexity of RSVM is lower than that of BRSVM. Our experiments will confirm these observations.

To verify the efficiency of the proposed algorithms, we record the computational time when conducting the experiments. The ratio for the time used by different methods are SVM: CSVM: ABSVM: RSVM: ABRSVM = 25: 25: 11: 3: 5. This is show that the new SVM based algorithms are much more efficient than existing SVM based algorithms.

Chapter 5
SSS Problems in KBDA RF

KBDA [91] has been used for RF mainly because it handles the positive and negative feedbacks separately. However, KBDA often suffers from the SSS problem. To overcome it, the regularization method adds small quantities to the diagonal of the scatter matrices. This apparently is not an optimal solution and sometimes it may lead to an ill-posed problem, which limits the performance of RF.

Recently, direct LDA (DLDA) [30] was proposed to solve the SSS problem in face recognition. DLDA discards the null space of between-class scatter matrix. Then the discriminant vectors are the within-class scatter matrix's eigenvectors with smallest eigenvalues. The successes of the kernel-machine based pattern classification algorithms have motivated us to generalize the idea of DLDA to BDA in the kernel feature space (DKBDA). We first project all the training samples from the input feature space to kernel feature space, and then the null-space of the negative-scatter-with-respect-to-positive-centroid matrix is removed. At last, the discriminant vectors are extracted as the positive-within-class-scatter matrix's eigenvectors with the smallest eigenvalues.

Another method to overcome the SSS problem is the null-space LDA (NLDA) [57]. NLDA extracts discriminant information from the null space of within class scatter matrix. Similar to KDBDA, we also generalize the idea of NLDA to BDA in the kernel feature space (NKBDA). We first project all the training samples from the input feature space to kernel feature space, and then the primal-space of the positive-within-class-scatter matrix is removed. At last, the discriminant vectors are extracted as the negative-scatter-with-respect-to-positive-centroid matrix's eigenvectors with the largest eigenvalues.

Clearly, both DLDA and NLDA may lose some discriminant information, because the null space of between class scatter matrix and the principle space of within-class scatter matrix, which are removed when conduct DLDA and NLDA, may contain some discriminant information. So we propose a full-space method

(FLDA) to preserve all discriminant information contained in LDA. Finally, we generalize the FLDA for KBDA as the Kernel Full-space BDA (FKBDA) [15][16].

5.1 DKBDA

5.1.1 DLDA

Yu et. al. proposed a direct LDA method, and it accepts high-dimensional data as input, and optimizes Fisher's criterion directly, without any feature extraction or dimension reduction steps, so it takes advantage of all the information within and outside of the null space of $\mathbf{S}_w$. In this approach, $\mathbf{S}_b$ is first diagonalized, and the null space of $\mathbf{S}_b$ is removed,

$$\mathbf{Y}^T\mathbf{S}_b\mathbf{Y} = \mathbf{D}_b > 0 \tag{5-1}$$

where $\mathbf{Y}$ are eigenvectors and $\mathbf{D}_b$ are the corresponding non-zero eigenvalues of $\mathbf{S}_b$. $\mathbf{S}_w$ is transformed to

$$\mathbf{K}_w = \mathbf{D}_b^{-1/2}\mathbf{Y}^T\mathbf{S}_w\mathbf{Y}\mathbf{D}_b^{-1/2}. \tag{5-2}$$

$\mathbf{K}_w$ is diagonalized by eigenanalysis,

$$\mathbf{U}^T\mathbf{K}_w\mathbf{U} = \mathbf{D}_w. \tag{5-3}$$

The LDA transformation matrix for classification is defined as,

$$\mathbf{W} = \mathbf{Y}\mathbf{D}_b^{-1/2}\mathbf{U}\mathbf{D}_w^{-1/2}. \tag{5-4}$$

In DLDA, the null space of $\mathbf{S}_b$ is first removed, and the discriminant vectors are restricted in the subspace spanned by class centers. It is assumed that the null space of $\mathbf{S}_b$ contains no discriminative information at all.

5.1.2 DKBDA

Before we deduce the kernel Direct BDA, we first introduce the kernel matrix $\mathbf{K}$:

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{xx} & \mathbf{K}_{xy} \\ \mathbf{K}_{yx} & \mathbf{K}_{yy} \end{bmatrix} \tag{5-5}$$

where

$$\mathbf{K}_{xx} = \begin{bmatrix} k(\mathbf{x}_1,\mathbf{x}_1) & \dots & k(\mathbf{x}_1,\mathbf{x}_{Nx}) \\ \dots & \dots & \dots \\ k(\mathbf{x}_{Nx},\mathbf{x}_1) & \dots & k(\mathbf{x}_{Nx},\mathbf{x}_{Nx}) \end{bmatrix},\ \mathbf{K}_{xy} = \begin{bmatrix} k(\mathbf{x}_1,\mathbf{y}_1) & \dots & k(\mathbf{x}_1,\mathbf{y}_{Ny}) \\ \dots & \dots & \dots \\ k(\mathbf{x}_{Nx},\mathbf{y}_1) & \dots & k(\mathbf{x}_{Nx},\mathbf{y}_{Ny}) \end{bmatrix},$$

$$\mathbf{K}_{yx}=\begin{bmatrix} k(\mathbf{y}_1,\mathbf{x}_1) & \dots & k(\mathbf{y}_1,\mathbf{x}_{Nx}) \\ \dots & \dots & \dots \\ k(\mathbf{y}_{Ny},\mathbf{x}_1) & \dots & k(\mathbf{y}_{Ny},\mathbf{x}_{Nx}) \end{bmatrix},\ \mathbf{K}_{yy}=\begin{bmatrix} k(\mathbf{y}_1,\mathbf{y}_1) & \dots & k(\mathbf{x}_1,\mathbf{x}_1) \\ \dots & \dots & \dots \\ k(\mathbf{y}_{Ny},\mathbf{y}_1) & \dots & k(\mathbf{y}_{Ny},\mathbf{y}_{Ny}) \end{bmatrix},$$

$\mathbf{x}_i$ is the positive feedback samples, N_x is the number of positive feedback samples, $\mathbf{y}_i$ is the negative feedback samples, N_y is the number of negative feedback samples, and $k(\dots)$ is the kernel function.

Just like Direct LDA, we begin the kernel Direct BDA from the analysis of the "the negative scatter with respect to positive centroid" matrix. Since the dimension of the $\mathbf{\Phi}_y$ could be arbitrarily infinitive, it is impossible to calculate $\mathbf{S}_y^{\phi}=\mathbf{\Phi}_y\mathbf{\Phi}_y^T$ directly and implement Eigen analysis with $\mathbf{S}_y^{\phi}$. Fortunately, this can be avoid through the following analysis:

$$\begin{aligned} &\mathbf{\Phi}_y^T\mathbf{\Phi}_y\mathbf{e}_i=\lambda_i\mathbf{e}_i\Rightarrow\mathbf{\Phi}_y\mathbf{\Phi}_y^T\left(\mathbf{\Phi}_y\mathbf{e}_i\right)=\lambda_i\left(\mathbf{\Phi}_y\mathbf{e}_i\right) \\ &\mathbf{\Phi}_y\mathbf{\Phi}_y^T\mathbf{u}_i=\lambda_i\mathbf{u}_i\Rightarrow\mathbf{u}_i=\mathbf{\Phi}_y\mathbf{e}_i \\ &\therefore\mathbf{U}=\mathbf{\Phi}_y\mathbf{E} \end{aligned} \tag{5-6}$$

The dimension of $\mathbf{\Phi}_y^T\mathbf{\Phi}_y$ is the number of negative relevance feedback samples. The following problem is to get the matrix.

$$\begin{aligned} \mathbf{\Phi}_y^T\mathbf{\Phi}_y &=\begin{bmatrix} \left(\boldsymbol{\varphi}^T(\mathbf{y}_1)-\bar{\boldsymbol{\varphi}}^T(\mathbf{x})\right) \\ \dots \\ \left(\boldsymbol{\varphi}^T(\mathbf{y}_i)-\bar{\boldsymbol{\varphi}}^T(\mathbf{x})\right) \\ \dots \\ \left(\boldsymbol{\varphi}^T(\mathbf{y}_{Nx})-\bar{\boldsymbol{\varphi}}^T(\mathbf{x})\right) \end{bmatrix}\left[\left(\boldsymbol{\varphi}(\mathbf{y}_1)-\bar{\boldsymbol{\varphi}}(\mathbf{x})\right)\ \dots\ \left(\boldsymbol{\varphi}(\mathbf{y}_j)-\bar{\boldsymbol{\varphi}}(\mathbf{x})\right)\ \dots\ \left(\boldsymbol{\varphi}(\mathbf{y}_{Nx})-\bar{\boldsymbol{\varphi}}(\mathbf{x})\right)\right] \\ &=\left[\left(\boldsymbol{\varphi}^T(\mathbf{y}_i)-\bar{\boldsymbol{\varphi}}^T(\mathbf{x})\right)\left(\boldsymbol{\varphi}(\mathbf{y}_j)-\bar{\boldsymbol{\varphi}}(\mathbf{x})\right)\right]_{\substack{i=1,2,\dots,Ny \\ j=1,2,\dots,Ny}} \\ &=\left[\boldsymbol{\varphi}^T(\mathbf{y}_i)\boldsymbol{\varphi}(\mathbf{y}_j)-\boldsymbol{\varphi}^T(\mathbf{y}_i)\bar{\boldsymbol{\varphi}}(\mathbf{x})-\bar{\boldsymbol{\varphi}}^T(\mathbf{x})\boldsymbol{\varphi}(\mathbf{y}_j)+\bar{\boldsymbol{\varphi}}^T(\mathbf{x})\bar{\boldsymbol{\varphi}}(\mathbf{x})\right]_{\substack{i=1,2,\dots,Ny \\ j=1,2,\dots,Ny}} \end{aligned} \tag{5-7}$$

Then we should calculate $\boldsymbol{\varphi}^T(\mathbf{y}_i)\boldsymbol{\varphi}(\mathbf{y}_j)$, $\boldsymbol{\varphi}^T(\mathbf{y}_i)\bar{\boldsymbol{\varphi}}(\mathbf{x})$, $\bar{\boldsymbol{\varphi}}^T(\mathbf{x})\boldsymbol{\varphi}(\mathbf{y}_j)$, and $\bar{\boldsymbol{\varphi}}^T(\mathbf{x})\bar{\boldsymbol{\varphi}}(\mathbf{x})$. The details will be seen from the following formulations

$$\begin{aligned} \bar{\boldsymbol{\varphi}}^T(\mathbf{x})\bar{\boldsymbol{\varphi}}(\mathbf{x}) &=\left(\frac{1}{N_x}\sum_{m=1}^{N_x}\boldsymbol{\varphi}(\mathbf{x}_m)\right)^T\left(\frac{1}{N_x}\sum_{i=n}^{N_x}\boldsymbol{\varphi}(\mathbf{x}_n)\right) \\ &=\frac{1}{N_x^2}\sum_{m=1}^{N_x}\sum_{n=1}^{N_x}\boldsymbol{\varphi}^T(\mathbf{x}_m)\boldsymbol{\varphi}(\mathbf{x}_n) \\ &=\frac{1}{N_x^2}\sum_{m=1}^{N_x}\sum_{n=1}^{N_x}k(\mathbf{x}_m,\mathbf{x}_n) \\ &=\frac{1}{N_x^2}\mathbf{1}_{Nx,1}^T\mathbf{K}_{xx}\mathbf{1}_{Nx,1} \end{aligned} \tag{5-8}$$

$$\begin{aligned}\bar{\varphi}^T(\mathbf{x})\varphi(\mathbf{y}_j) &= \left(\frac{1}{N_x}\sum_{m=1}^{N_x}\varphi(\mathbf{x}_m)\right)^T\varphi(\mathbf{y}_j)\\ &= \frac{1}{N_x}\sum_{m=1}^{N_x}\varphi^T(\mathbf{x}_m)\varphi(\mathbf{y}_j)\\ &= \frac{1}{N_x}\sum_{m=1}^{N_x}k(\mathbf{x}_m,\mathbf{y}_j)\end{aligned}\tag{5-9}$$

$$\begin{aligned}\varphi^T(\mathbf{y}_i)\bar{\varphi}(\mathbf{x}) &= \varphi^T(\mathbf{y}_i)\left(\frac{1}{N_x}\sum_{m=1}^{N_x}\varphi(\mathbf{x}_m)\right)\\ &= \frac{1}{N_x}\sum_{m=1}^{N_x}\varphi^T(\mathbf{y}_i)\varphi(\mathbf{x}_m)\\ &= \frac{1}{N_x}\sum_{m=1}^{N_x}k(\mathbf{y}_i,\mathbf{x}_m)\end{aligned}\tag{5-10}$$

where $\mathbf{1}_{Nx,1}$ is a column vector with all terms equal to one.

So we can obtain the following formulation according to the kernel matrix (13).

$$\begin{aligned}&\mathbf{\Phi}_y^T\mathbf{\Phi}_y\\ &=\left[\varphi^T(\mathbf{y}_i)\varphi(\mathbf{y}_j)-\varphi^T(\mathbf{y}_i)\bar{\varphi}(\mathbf{x})-\bar{\varphi}^T(\mathbf{x})\varphi(\mathbf{y}_j)+\bar{\varphi}^T(\mathbf{x})\bar{\varphi}(\mathbf{x})\right]_{\substack{i=1,2,\ldots,Nx\\ j=1,2,\ldots,Nx}}\\ &=\left[k(\mathbf{y}_i,\mathbf{y}_j)-\frac{1}{N_x}\sum_{m=1}^{N_x}k(\mathbf{y}_i,\mathbf{x}_m)-\frac{1}{N_x}\sum_{m=1}^{N_x}k(\mathbf{x}_m,\mathbf{y}_j)+\frac{1}{N_x^2}\mathbf{1}_{Nx,1}^T\mathbf{K}_{xx}\mathbf{1}_{Nx,1}\right]_{\substack{i=1,2,\ldots,Ny\\ j=1,2,\ldots,Ny}}\\ &=\mathbf{K}_{yy}-\frac{1}{N_x}\mathbf{K}_{yx}\mathbf{1}_{Nx,1}\mathbf{1}_{Ny,1}^T-\frac{1}{N_x}\mathbf{1}_{Ny,1}\mathbf{1}_{Nx,1}^T\mathbf{K}_{xy}+\frac{1}{N_x^2}\mathbf{1}_{Nx,1}^T\mathbf{K}_{xx}\mathbf{1}_{Nx,1}\mathbf{1}_{Ny,Ny}\\ &=\mathbf{K}_{yy}-\frac{1}{N_x}\mathbf{K}_{yx}\mathbf{1}_{Nx,Ny}-\frac{1}{N_x}\mathbf{1}_{Ny,Nx}\mathbf{K}_{xy}+\frac{\alpha}{N_x^2}\mathbf{1}_{Ny,Ny}\end{aligned}\tag{5-11}$$

where $\alpha=\mathbf{1}_{Nx,1}^T\mathbf{K}_{xx}\mathbf{1}_{Nx,1}$, $\mathbf{1}_{Nx,Nx}$ is a matrix with all terms equal to one.

We do Eigen analysis with (11), and obtain the none-zero space $\mathbf{E}$ of $\mathbf{\Phi}_y^T\mathbf{\Phi}_y$, i.e. $\mathbf{E}^T\mathbf{\Phi}_y^T\mathbf{\Phi}_y\mathbf{E}=\mathbf{D}_y\neq 0$. According to (6), we can easily obtain the diagonalized none-zero subspace $\mathbf{W}=\mathbf{\Phi}_y\mathbf{D}_y^{-1/2}\mathbf{E}$ of $\mathbf{\Phi}_y\mathbf{\Phi}_y^T$, i.e. $\mathbf{W}^T\mathbf{S}_y^\varphi\mathbf{W}\neq 0$. Here, we need not to calculate $\mathbf{W}=\mathbf{\Phi}_y\mathbf{D}_y^{-1/2}\mathbf{E}$. To reckon the largest Eigen values and corresponding Eigen vectors, we can in depth the analysis. It is similar with the Direct LDA, and then the "positive with-in class scatter" matrix is projected into the none-zero space:

$$\mathbf{W}^T\mathbf{S}_x^\varphi\mathbf{W}=\mathbf{D}_y^{-1/2}\mathbf{E}^T\mathbf{\Phi}_y^T\mathbf{S}_x^\varphi\mathbf{\Phi}_y\mathbf{E}\mathbf{D}_y^{-1/2}\tag{5-12}$$

From (12), we can see that to calculate $\mathbf{W}=\mathbf{\Phi}_y\mathbf{D}_y^{-1/2}\mathbf{E}$ can be avoided. The new coming problem is to reckon $\mathbf{\Phi}_y^T\mathbf{S}_x^\varphi\mathbf{\Phi}_y$. With the following deduction (13), (14), (15), (16), and (17), we can draw the conclusion that $\mathbf{\Phi}_y^T\mathbf{S}_x^\varphi\mathbf{\Phi}_y$ only relates to the kernel matrix (5), just like $\mathbf{\Phi}_y^T\mathbf{\Phi}_y$.

$$\mathbf{\Phi}_y^T\mathbf{S}_x^\varphi\mathbf{\Phi}_y=\mathbf{\Phi}_y^T\mathbf{\Phi}_x\mathbf{\Phi}_x^T\mathbf{\Phi}_y=\left(\mathbf{\Phi}_x^T\mathbf{\Phi}_y\right)^T\left(\mathbf{\Phi}_x^T\mathbf{\Phi}_y\right)\tag{5-13}$$

To reckon $\mathbf{\Phi}_y^T\mathbf{S}_x^\varphi\mathbf{\Phi}_y$, we only need to calculate $\mathbf{\Phi}_x^T\mathbf{\Phi}_y$.

$$\begin{aligned}
&\mathbf{\Phi}_x^T\mathbf{\Phi}_y \\
&=\begin{bmatrix} \left(\varphi^T(\mathbf{x}_1)-\bar{\varphi}^T(\mathbf{x})\right) \\ \cdots \\ \left(\varphi^T(\mathbf{x}_i)-\bar{\varphi}^T(\mathbf{x})\right) \\ \cdots \\ \left(\varphi^T(\mathbf{x}_{Nx})-\bar{\varphi}^T(\mathbf{x})\right) \end{bmatrix} \cdot \left[\varphi(\mathbf{y}_1)-\bar{\varphi}(\mathbf{x}) \ \ \cdots \ \ \varphi(\mathbf{y}_j)-\bar{\varphi}(\mathbf{x}) \ \ \cdots \ \ \varphi(\mathbf{y}_{Ny})-\bar{\varphi}(\mathbf{x})\right] \\
&=\left[\left(\varphi^T(\mathbf{x}_i)-\bar{\varphi}^T(\mathbf{x})\right)\left(\varphi(\mathbf{y}_j)-\bar{\varphi}(\mathbf{x})\right)\right]_{\substack{i=1,2,\ldots,Nx\\ j=1,2,\ldots,Ny}} \\
&=\left[\varphi^T(\mathbf{x}_i)\varphi(\mathbf{y}_j)-\varphi^T(\mathbf{x}_i)\bar{\varphi}(\mathbf{x})-\bar{\varphi}^T(\mathbf{x})\varphi(\mathbf{y}_j)+\bar{\varphi}^T(\mathbf{x})\bar{\varphi}(\mathbf{x})\right]_{\substack{i=1,2,\ldots,Nx\\ j=1,2,\ldots,Ny}}
\end{aligned} \tag{5-14}$$

To calculate $\mathbf{\Phi}_x^T\mathbf{\Phi}_y$, we first should to calculate $\varphi^T(\mathbf{x}_i)\bar{\varphi}(\mathbf{x})$, $\bar{\varphi}^T(\mathbf{x})\varphi(\mathbf{y}_j)$, and $\bar{\varphi}^T(\mathbf{x})\bar{\varphi}(\mathbf{x})$. Here, $\bar{\varphi}^T(\mathbf{x})\varphi(\mathbf{y}_j)$ and $\bar{\varphi}^T(\mathbf{x})\bar{\varphi}(\mathbf{x})$ are respectively calculated in (9) and (8). In the formulation (15), we reckon $\varphi^T(\mathbf{x}_i)\bar{\varphi}(\mathbf{x})$.

$$\begin{aligned}
\varphi^T(\mathbf{x}_i)\bar{\varphi}(\mathbf{x}) &= \varphi^T(\mathbf{x}_i)\left(\frac{1}{N_x}\sum_{m=1}^{N_x}\varphi(\mathbf{x}_m)\right) \\
&= \frac{1}{N_x}\sum_{m=1}^{N_x}\varphi^T(\mathbf{x}_i)\varphi(\mathbf{x}_m) \\
&= \frac{1}{N_x}\sum_{m=1}^{N_x}k(\mathbf{x}_i,\mathbf{x}_m)
\end{aligned} \tag{5-15}$$

Then we can obtain the $\mathbf{\Phi}_x^T\mathbf{\Phi}_y$ by formulation (16).

$$\begin{aligned}
&\mathbf{\Phi}_x^T\mathbf{\Phi}_y \\
&=\left[\varphi^T(\mathbf{x}_i)\varphi(\mathbf{y}_j)-\varphi^T(\mathbf{x}_i)\bar{\varphi}(\mathbf{x})-\bar{\varphi}^T(\mathbf{x})\varphi(\mathbf{y}_j)+\bar{\varphi}^T(\mathbf{x})\bar{\varphi}(\mathbf{x})\right]_{\substack{i=1,2,\ldots,Nx\\ j=1,2,\ldots,Ny}} \\
&=\left[k(\mathbf{x}_i,\mathbf{y}_j)-\frac{1}{N_x}\sum_{m=1}^{N_x}k(\mathbf{x}_i,\mathbf{x}_m)-\frac{1}{N_x}\sum_{m=1}^{N_x}k(\mathbf{x}_m,\mathbf{y}_j)+\frac{1}{N_x^2}\mathbf{1}_{Nx,1}^T\mathbf{K}_{xx}\mathbf{1}_{Nx,1}\right]_{\substack{i=1,2,\ldots,Nx\\ j=1,2,\ldots,Ny}} \\
&=\mathbf{K}_{xy}-\frac{1}{N_x}\mathbf{K}_{xx}\mathbf{1}_{Nx,Ny}-\frac{1}{N_x}\mathbf{1}_{Nx,Nx}\mathbf{K}_{xy}+\frac{\alpha}{N_x^2}\mathbf{1}_{Nx,Ny}
\end{aligned} \tag{5-16}$$

Then we can obtain $\mathbf{\Phi}_y^T\mathbf{S}_x^{\varphi}\mathbf{\Phi}_y$ according to the following deduction:

$$
\begin{aligned}
&\mathbf{\Phi}_y^T \mathbf{S}_x^\phi \mathbf{\Phi}_y \\
&= \left(\mathbf{K}_{xy} - \frac{1}{N_x}\mathbf{K}_{xx}\mathbf{1}_{Nx,Ny} - \frac{1}{N_x}\mathbf{1}_{Nx,Nx}\mathbf{K}_{xy} + \frac{\alpha}{N_x^2}\mathbf{1}_{Nx,Ny} \right)^T \\
&\cdot \left(\mathbf{K}_{xy} - \frac{1}{N_x}\mathbf{K}_{xx}\mathbf{1}_{Nx,Ny} - \frac{1}{N_x}\mathbf{1}_{Nx,Nx}\mathbf{K}_{xy} + \frac{\alpha}{N_x^2}\mathbf{1}_{Nx,Ny} \right) \\
&= \mathbf{K}_{yx}\mathbf{K}_{xy} - \frac{1}{N_x}\mathbf{K}_{yx}\mathbf{K}_{xx}\mathbf{1}_{Nx,Ny} - \frac{1}{N_x}\mathbf{K}_{yx}\mathbf{1}_{Nx,Nx}\mathbf{K}_{xy} + \frac{\alpha}{N_x^2}\mathbf{K}_{yx}\mathbf{1}_{Nx,Ny} \\
&- \frac{1}{N_x}\mathbf{1}_{Ny,Nx}\mathbf{K}_{xx}\mathbf{K}_{xy} + \frac{1}{N_x^2}\mathbf{1}_{Ny,Nx}\mathbf{K}_{xx}\mathbf{K}_{xx}\mathbf{1}_{Nx,Ny} + \frac{1}{N_x^2}\mathbf{1}_{Ny,Nx}\mathbf{K}_{xx}\mathbf{1}_{Nx,Nx}\mathbf{K}_{xy} \\
&- \frac{\alpha}{N_x^3}\mathbf{1}_{Ny,Nx}\mathbf{K}_{xx}\mathbf{1}_{Nx,Ny} - \frac{1}{N_x}\mathbf{K}_{yx}\mathbf{1}_{Nx,Nx}\mathbf{K}_{xy} + \frac{1}{N_x^2}\mathbf{K}_{yx}\mathbf{1}_{Nx,Nx}\mathbf{K}_{xx}\mathbf{1}_{Nx,Ny} \\
&+ \frac{1}{N_x^2}\mathbf{K}_{yx}\mathbf{1}_{Nx,Nx}\mathbf{1}_{Nx,Nx}\mathbf{K}_{xy} - \frac{\alpha}{N_x^3}\mathbf{K}_{yx}\mathbf{1}_{Nx,Nx}\mathbf{1}_{Nx,Ny} + \frac{\alpha}{N_x^2}\mathbf{1}_{Ny,Nx}\mathbf{K}_{xy} \\
&- \frac{\alpha}{N_x^3}\mathbf{1}_{Ny,Nx}\mathbf{K}_{xx}\mathbf{1}_{Nx,Ny} - \frac{\alpha}{N_x^3}\mathbf{1}_{Ny,Nx}\mathbf{1}_{Nx,Nx}\mathbf{K}_{xy} + \frac{\alpha^2}{N_x^4}\mathbf{1}_{Ny,Nx}\mathbf{1}_{Nx,Ny} \\
&= \mathbf{A} - \frac{1}{N_x}\mathbf{B} + \frac{1}{N_x^2}\mathbf{C} - \frac{\alpha}{N_x^3}\mathbf{D}
\end{aligned}
\tag{5-17}
$$

where $\mathbf{A} = \mathbf{K}_{yx}\mathbf{K}_{xy} + \frac{\alpha}{N_x^2}\left(\mathbf{K}_{yx}\mathbf{1}_{Nx,Ny} + \mathbf{1}_{Ny,Nx}\mathbf{K}_{xy}\right) + \frac{\alpha^2}{N_x^3}\mathbf{1}_{Ny,Ny}$,

$\mathbf{B} = \mathbf{K}_{yx}\mathbf{K}_{xx}\mathbf{1}_{Nx,Nx} + \mathbf{K}_{yx}\mathbf{1}_{Nx,Nx}\mathbf{K}_{xy} + \mathbf{1}_{Ny,Nx}\mathbf{K}_{xx}\mathbf{K}_{xy} + \mathbf{K}_{yx}\mathbf{1}_{Nx,Nx}\mathbf{K}_{xy}$,

$\mathbf{C} = \mathbf{1}_{Ny,Nx}\mathbf{K}_{xx}\mathbf{K}_{xx}\mathbf{1}_{Nx,Ny} + \mathbf{1}_{Ny,Nx}\mathbf{K}_{xx}\mathbf{1}_{Nx,Nx}\mathbf{K}_{xy} + \mathbf{K}_{yx}\mathbf{1}_{Nx,Nx}\mathbf{K}_{xx}\mathbf{1}_{Nx,Ny} + N_x\mathbf{K}_{yx}\mathbf{1}_{Nx,Nx}\mathbf{K}_{xy}$, and

$\mathbf{D} = \mathbf{1}_{Ny,Nx}\mathbf{K}_{xx}\mathbf{1}_{Nx,Ny} + N_x\mathbf{K}_{yx}\mathbf{1}_{Nx,Ny} + \mathbf{1}_{Ny,Nx}\mathbf{K}_{xx}\mathbf{1}_{Nx,Ny} + N_x\mathbf{1}_{Ny,Nx}\mathbf{K}_{xy}$.

Under the idea of Direct LDA, we do the Eigen analysis of $\hat{\mathbf{S}}_x^\phi = \mathbf{W}^T\mathbf{S}_x^\phi\mathbf{W} = \mathbf{D}_y^{-1/2}\mathbf{E}^T\mathbf{\Phi}_y^T\mathbf{S}_x^\phi\mathbf{\Phi}_y\mathbf{E}\mathbf{D}_y^{-1/2}$, and we select the eigenvectors $\mathbf{V}$ of $\hat{\mathbf{S}}_x^\phi$ with the smallest eigenvalues $\mathbf{D}_x$, i.e.

$$\mathbf{V}^T\hat{\mathbf{S}}_x^\phi\mathbf{V} = \mathbf{D}_x \tag{5-18}$$

At last we established the overall projection matrix $\mathbf{U} = \mathbf{E}\mathbf{D}_y^{-1/2}\mathbf{V}\mathbf{D}_x^{-1/2}$.

Obviously, it is possible that some diagonal values in the matrix $\mathbf{D}_x$ is zero, which means that $\mathbf{D}_x^{-1/2}$ does not exist. However, we can avoid the zero eigenvalue problem based on a changed KBDA criterion, according to [K. Liu].

The modified KBDA criterion is:

$$\mathbf{W} = \underset{\mathbf{W}}{\arg\max}\frac{\left\|\mathbf{W}^{\mathrm{T}}\mathbf{S}_y^\phi\mathbf{W}\right\|}{\left\|\mathbf{W}^{\mathrm{T}}\left(\mathbf{S}_x^\phi + \mathbf{S}_y^\phi\right)\mathbf{W}\right\|}. \tag{5-19}$$

It is clearly that the modified criterion equals to the original KBDA criterion based on the proof in [K. Liu]. Based on the modified KBDA criterion, we can avoid the singular value problem, because $\left\|\mathbf{W}^{\mathrm{T}}\mathbf{S}_y^\phi\mathbf{W}\right\| = \mathbf{I}$.

Here, we sum up the DKBDA algorithm:

1. Calculate the kernel matrix $\mathbf{K}$ based on (5).
2. Calculate $\mathbf{\Phi}_y^T\mathbf{\Phi}_y$ according to (11).
3. Do Eigen analysis on $\mathbf{\Phi}_y^T\mathbf{\Phi}_y$, and obtain the none-zero subspace $\mathbf{E}$ of $\mathbf{\Phi}_y^T\mathbf{\Phi}_y$.
4. Calculate $\mathbf{\Phi}_y^T\mathbf{S}_x^\phi\mathbf{\Phi}_y$ according to (17).
5. Do Eigen analysis on $\widehat{\mathbf{S}}_x^\phi = \mathbf{W}^T\mathbf{S}_x^\phi\mathbf{W} + \mathbf{W}^T\mathbf{S}_y^\phi\mathbf{W} = \mathbf{D}_y^{-1/2}\mathbf{E}^T\mathbf{\Phi}_y^T\mathbf{S}_x^\phi\mathbf{\Phi}_y\mathbf{E}\mathbf{D}_y^{-1/2} + \mathbf{I}$, and select the Eigen vectors $\mathbf{V}$ of $\widehat{\mathbf{S}}_x^\phi$ with the smallest Eigen values $\mathbf{D}_x$.
6. Established the overall projection matrix $\mathbf{U} = \mathbf{E}\mathbf{D}_y^{-1/2}\mathbf{V}\mathbf{D}_x^{-1/2}$.

5.2 NKBDA

5.2.1 NLDA

NLDA [8] optimizes LDA in the null space of $\mathbf{S}_w$. In NLDA, the null space of $\mathbf{S}_w$ is first calculated as:

$$\mathbf{Y}^T\mathbf{S}_w\mathbf{Y} = 0 \tag{5-20}$$

where $\mathbf{Y}$ are eigenvectors with zero eigenvalues and $\mathbf{Y}^T\mathbf{Y} = \mathbf{I}$. The between class scatter matrix $\mathbf{S}_b$ is projected onto the null space of $\mathbf{S}_w$,

$$\hat{\mathbf{S}}_b = \mathbf{Y}^T\mathbf{S}_b\mathbf{Y} \tag{5-21}$$

The eigenvectors $\mathbf{U}$ of $\hat{\mathbf{S}}_b$ with largest eigenvalues are selected to form the transformation matrix,

$$\mathbf{W} = \mathbf{Y}\mathbf{U}. \tag{5-22}$$

5.2.2 NKBDA

For null-space method in KBDA (NKBDA), we first extract the null-space of $\mathbf{S}_x^\phi$ according to,

$$\begin{aligned}\mathbf{W}^T\mathbf{S}_x^\phi\mathbf{W} &= \boldsymbol{\alpha}_{mx}^T\mathbf{\Phi}^T\mathbf{S}_x^\phi\mathbf{\Phi}\boldsymbol{\alpha}_{mx} \\ &= \boldsymbol{\alpha}_{mx}^T\mathbf{K}_x\left(\mathbf{I} - \mathbf{1}_{Nx}^x\right)\mathbf{K}_x^T\boldsymbol{\alpha}_{mx} = \mathbf{D}_x = diag\left(d_{m+1},\ldots,d_{Nx}\right)\end{aligned} \tag{5-23}$$

where $\varepsilon \geq d_{m+1} \geq d_{m+2} \geq \ldots \geq d_{Nx} = 0$. Then project $\mathbf{S}_y^\phi$ onto the null-space of S_x^ϕ by,

$$\begin{aligned}\bar{\mathbf{S}}_y^{\phi} &= \boldsymbol{\alpha}_{ny}^T \Phi^T \mathbf{S}_y^{\phi} \Phi \boldsymbol{\alpha}_{ny} \\ &= \boldsymbol{\alpha}_{ny}^T \left(\mathbf{K}_y - \mathbf{K}_y \mathbf{I}_{Ny}^{y}\right)\left(\mathbf{K}_y - \mathbf{K}_y \mathbf{I}_{Ny}^{y}\right)^T \boldsymbol{\alpha}_{ny}\end{aligned} \tag{5-24}$$

Finally, the eigenvectors of $\bar{\mathbf{S}}_y^{\phi}$ corresponding to the largest eigenvalues selected as feature basis,

$$\boldsymbol{\alpha}_{py}^T \bar{\mathbf{S}}_y^{\phi} \boldsymbol{\alpha}_{py} = \mathbf{D}_y . \tag{5-25}$$

The kernel projection matrix is: $\boldsymbol{\alpha}_{py}^T \mathbf{D}_y^{-1/2} \boldsymbol{\alpha}_{ny}^T \Phi^T$.

5.3 FKBDA

5.3.1 FLDA

Both DLDA and NLDA may lose some discriminant information. DLDA loses the information in the null space of the between class matrix. NLDA loses information in the principle space of the within class scatter matrix. In order to overcome the MSP and still preserve the discriminant information, we propose a full space LDA. For $\mathbf{S}_w$, we compute,

$$\mathbf{Y}^T \mathbf{S}_w \mathbf{Y} = \mathbf{D}_w \tag{5-26}$$

where $\mathbf{D}_w = diag(\lambda_1, \ldots, \lambda_i, \ldots, \lambda_m, \lambda_{m+1}, \ldots, 0)$, $\lambda_{m+1} \le \varepsilon\lambda_1 \le \lambda_m$, and ε is a user selected threshold value (such as 0.01).

For a given ε , the eigenvalue matrix $\mathbf{D}_w$ is then converted to $\hat{\mathbf{D}}_w = diag(\lambda_1, \ldots, \lambda_i, \ldots, \lambda_m, \lambda_{m+1}, \ldots, \lambda_{m+1})$. All values, which is smaller than λ_{m+1} , are substituted by λ_{m+1}. After the substitution, the between class scatter matrix $\mathbf{S}_b$ is projected onto the space by,

$$\hat{\mathbf{S}}_b = \hat{\mathbf{D}}_w^{-1/2} \mathbf{Y}^T \mathbf{S}_b \mathbf{Y} \hat{\mathbf{D}}_w^{-1/2} . \tag{5-27}$$

Finally, the eigenvectors $\mathbf{U}$ of $\hat{\mathbf{S}}_b$ with largest eigenvalues are selected to form the transformation matrix,

$$\mathbf{W} = \mathbf{Y} \hat{\mathbf{D}}_w^{-1/2} \mathbf{U} \mathbf{D}_b^{-1/2} . \tag{5-28}$$

5.3.2 FKBDA

For full-space method in KBDA (FKBDA), we first extract the principle and null spaces of $\mathbf{S}_y^{\phi}$, and then combine them by a weight as the full-space,

$$\begin{aligned}\mathbf{W}^T\mathbf{S}_y^\phi\mathbf{W} &= \boldsymbol{\alpha}_{fx}^T\boldsymbol{\Phi}^T\mathbf{S}_y^\phi\boldsymbol{\Phi}\boldsymbol{\alpha}_{fx} = \boldsymbol{\alpha}_{fx}^T\mathbf{K}_y\left(\mathbf{I}-\mathbf{I}_{Nx}^x\right)\mathbf{K}_y^T\boldsymbol{\alpha}_{fx} \\ &= \boldsymbol{\Lambda}_x \sim \widehat{\boldsymbol{\Lambda}}_x = diag\left(\lambda_1,\dots,\lambda_m,\lambda_{m+1},\lambda_{m+1},\dots,\lambda_{m+1}\right)\end{aligned} \tag{5-29}$$

where $\lambda_1 \geq \dots \geq \lambda_m \geq \varepsilon \geq \lambda_{m+1} \geq 0$. Then project $\mathbf{S}_y^\phi$ onto the full-space of S_x^ϕ by,

$$\begin{aligned}\widehat{\mathbf{S}}_y^\phi &= \boldsymbol{\alpha}_{fx}^T\widehat{\boldsymbol{\Lambda}}_x^{-1/2}\boldsymbol{\Phi}^T\mathbf{S}_y^\phi\boldsymbol{\Phi}\widehat{\boldsymbol{\Lambda}}_x^{-1/2}\boldsymbol{\alpha}_{fx} \\ &= \boldsymbol{\alpha}_{fx}^T\widehat{\boldsymbol{\Lambda}}_x^{-1/2}\left(\mathbf{K}_y-\mathbf{K}_y\mathbf{I}_{Ny}^y\right)\left(\mathbf{K}_y-\mathbf{K}_x\mathbf{I}_{Nx}^y\right)^T\widehat{\boldsymbol{\Lambda}}_x^{-1/2}\boldsymbol{\alpha}_{fx}\end{aligned} \tag{5-30}$$

Finally, we extract the eigenvectors of the projected $\widehat{\mathbf{S}}_y^\phi$ with largest eigenvalues as,

$$\boldsymbol{\alpha}_{py}^T\widehat{\mathbf{S}}_y^\phi\boldsymbol{\alpha}_{py} = \mathbf{D}_y \tag{5-31}$$

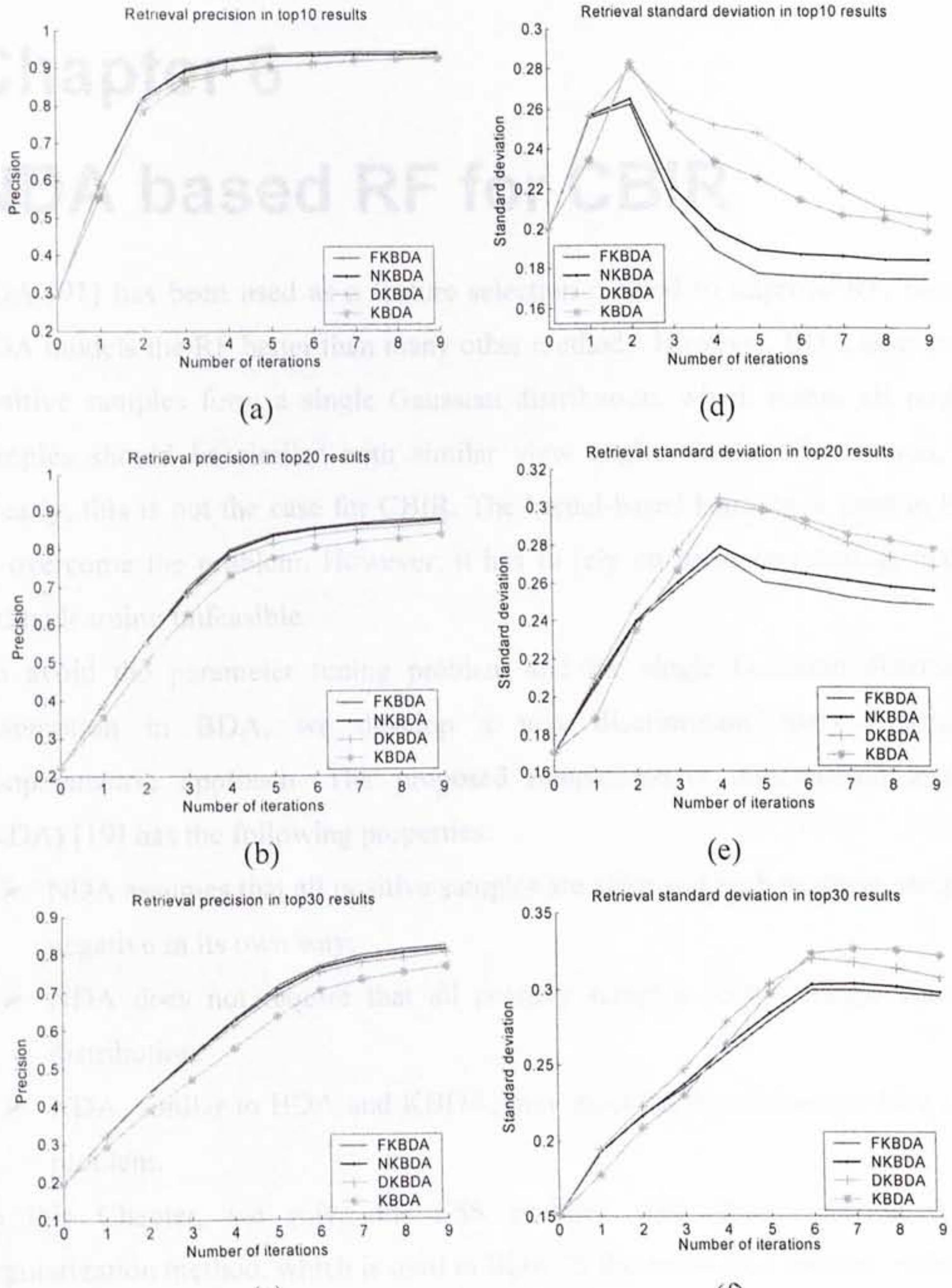
The kernel projection matrix is: $\boldsymbol{\alpha}_{py}^T\mathbf{D}_y^{-1/2}\boldsymbol{\alpha}_{fx}^T\widehat{\boldsymbol{\Lambda}}_x^{-1/2}\boldsymbol{\Phi}^T$.

5.4 Experimental Results

We evaluate the performance of the proposed algorithm using the precision and its standard deviation. Precision is the ratio of the number of retrieved relevant images to the top N retrieved images. Precision examines the effectiveness of an algorithm and the corresponding standard deviation evaluates the robustness of the algorithm. We conduct the experiment on QueryGo.

In the statistical experiment, we compare all proposed algorithms (DKBDA, NKBDA, and FKBDA) with the existing KBDA. The computer automatically did the feedback experiments with 200 queries. For each iteration, the system marked the first 5 incorrect and correct retrieved images from the top 48 matches as irrelevant and relevant examples, respectively. In the kernel based algorithms, we chose the Gaussian kernel $K(\mathbf{x},\mathbf{y}) = e^{-\rho\|\mathbf{x}-\mathbf{y}\|^2}$ with $\rho = 1/10$ because the parameter shows the best performance for FKBDA, NKBDA, DKBDA, and KBDA from a series of values.

Figure 1 shows the performance of FKBDA, NKBDA, DKBDA, and KBDA. The results show that our algorithms outperform the existing KBDA consistently both on effectiveness and robustness, meanwhile, FKBDA works best in all algorithms. In addition, the computation costs of the three methods are similar in our experiments.



(a) (d)

(b) (e)

(c) (f)

Figure 5-1. Evaluation experimental results based on the Corel database with 17, 800 images with 200 queries. (a), (b), and (c) display the retrieval precision in top 10, 20, and 30 retrieved images respectively. (d), (e), and (f) display the corresponding standard deviation of the precision curve in top 10, 20, and 30 retrieved images respectively.

Chapter 6

NDA based RF for CBIR

BDA [91] has been used as a feature selection method to improve RF, because BDA models the RF better than many other methods. However, BDA assumes all positive samples form a single Gaussian distribution, which means all positive samples should be similar with similar view angle, similar illumination, etc. Clearly, this is not the case for CBIR. The kernel-based learning is used in BDA to overcome the problem. However, it has to rely on parameter tuning, making online learning unfeasible.

To avoid the parameter tuning problem and the single Gaussian distribution assumption in BDA, we develop a new discriminant analysis using a nonparametric approach. The proposed nonparametric discriminant analysis (NDA) [19] has the following properties:

- ➢ NDA assumes that all positive samples are alike and each negative sample is negative in its own way;
- ➢ NDA does not require that all positive samples form a single Gaussian distribution.
- ➢ NDA, similar to BDA and KBDA, may meet the Small-Sample-Size (SSS) problem.

In this Chapter, we solve the SSS problem with three methods: 1. the regularization method, which is used in BDA; 2. the null-space method, which is a popular method to solve the SSS problem in linear discriminant analysis for face recognition; 3. the full-space method, which is proposed to preserve all discriminant information of NDA.

6.1 NDA

Similar to BDA, NDA is also biased toward to the positive examples. The objective function of NDA is:

$$\mathbf{W}_{opt} = \arg\max_{\mathbf{W}} \frac{\left\| \mathbf{W}^T \hat{\mathbf{S}}_y \mathbf{W} \right\|}{\left\| \mathbf{W}^T \hat{\mathbf{S}}_x \mathbf{W} \right\|}. \quad (6\text{-}1)$$

Let the training set contains N_x positive and N_y negative samples. Then $\hat{\mathbf{S}}_x$ and $\hat{\mathbf{S}}_y$ are defined as,

$$\begin{cases} \hat{\mathbf{S}}_x = \sum_{i=1}^{N_x} \left(\mathbf{x}_i - \mathbf{m}_{xi}^{kx}\right)\left(\mathbf{x}_i - \mathbf{m}_{xi}^{kx}\right)^T \\ \hat{\mathbf{S}}_y = \sum_{i=1}^{N_y} \left(\mathbf{y}_i - \mathbf{m}_{yi}^{kx}\right)\left(\mathbf{y}_i - \mathbf{m}_{yi}^{kx}\right)^T + \sum_{i=1}^{N_x} \left(\mathbf{x}_i - \mathbf{m}_{xi}^{ky}\right)\left(\mathbf{x}_i - \mathbf{m}_{xi}^{ky}\right)^T, \end{cases} \tag{6-2}$$

where $\mathbf{x}_i$ are positive samples, $\mathbf{y}_i$ are negative samples, $\mathbf{m}_{xi}^{kx} = \frac{1}{k}\sum_{l=1}^{k}\mathbf{x}_l$ is the mean vector of the k positive nearest neighbors of the i^{th} positive feedback sample $\mathbf{x}_i$, $\mathbf{m}_{yi}^{kx} = \frac{1}{k}\sum_{l=1}^{k}\mathbf{y}_l$ is the mean vector of the k negative nearest neighbors of the i^{th} positive feedback sample $\mathbf{x}_i$, $\mathbf{m}_{xi}^{ky} = \frac{1}{k}\sum_{l=1}^{k}\mathbf{x}_l$ is the mean vector of the k positive nearest neighbors of the i^{th} negative feedback sample $\mathbf{y}_i$, and $\mathbf{W}_{opt}$ can be computed from the eigenvectors of $\hat{\mathbf{S}}_x^{-1}\hat{\mathbf{S}}_y$. NDA finds the optimal feature set to maximize the margin between all positive feedbacks and all negative feedbacks in the input feature space. Because the original feature dimension is much larger than the number of the feedback samples, we can always find the subset feature to discriminant the positive and negative samples.

6.2 SSS Problem in NDA

In RF, the size of the training set is much smaller than the dimension of the feature vector, thus it may cause the SSS problem. In this Section, we will address the SSS problem using three methods, the regularization method, the null-space method, and the new full-space method.

6.2.1 Regularization method

Regularization method, which is proposed by Friedman to deal with the singularity issue, is implemented by adding small quantities to the diagonal of the scatter matrices $\hat{\mathbf{S}}_x$ and $\hat{\mathbf{S}}_y$. The regularized version of $\hat{\mathbf{S}}_x$ and $\hat{\mathbf{S}}_y$, with the dimension of the original feature space n and the identity matrix $\mathbf{I}$, are:

$$\hat{\mathbf{S}}_x^r = (1-\mu)\hat{\mathbf{S}}_x + \frac{\mu}{n} tr\left[\hat{\mathbf{S}}_x\right]\mathbf{I} \tag{6-3}$$

$$\hat{\mathbf{S}}_y^r = (1-\gamma)\hat{\mathbf{S}}_y + \frac{\gamma}{n} tr\left[\hat{\mathbf{S}}_y\right]\mathbf{I} \tag{6-4}$$

where μ and γ control the shrinkage toward a multiple of the identity matrix. $tr[.]$ is the trace operation.

It is well known that regularization method may meet the ill-posed problem. Hence, we select the null-space to overcome the ill-posed issue.

6.2.2 Null-space method

Null-space linear discriminant analysis (LDA) [8] accepts high-dimensional data as the input, and optimizes LDA in the null space of within class scatter matrix. Here, we generalize the null-space idea for NDA. The null space of $\hat{\mathbf{S}}_v$ is first calculated as:

$$\mathbf{Y}^T \hat{\mathbf{S}}_v \mathbf{Y} = 0 \tag{6-5}$$

where $\mathbf{Y}$ are eigenvectors with zero eigenvalues and $\mathbf{Y}^T \mathbf{Y} = \mathbf{I}$. $\hat{\mathbf{S}}_v$ is projected onto the null space of $\hat{\mathbf{S}}_v$:

$$\hat{\mathbf{S}}_v'' = \mathbf{Y}^T \hat{\mathbf{S}}_v \mathbf{Y} . \tag{6-6}$$

The eigenvectors $\mathbf{U}$ of $\hat{\mathbf{S}}_v''$ with largest eigenvalues are selected to form the transformation matrix as:

$$\mathbf{W} = \mathbf{YU} . \tag{6-7}$$

6.2.3 Full-space method

Null-space method loses the information in the principle space of the within class scatter matrix. In order to preserve all discriminant information, we compute features from both the null space and the principle space of $\hat{\mathbf{S}}_v$, and then integrate the two parts with a suitable weighting. A rational choice of the weighting is to select a small eigenvalue of $\mathbf{S}_w$. The algorithm first computes the eigenvalues of $\hat{\mathbf{S}}_v$ as,

$$\mathbf{Y}^T \hat{\mathbf{S}}_v \mathbf{Y} = \mathbf{D}_v , \tag{6-8}$$

where $\mathbf{D}_v = diag(\lambda_1, \ldots, \lambda_i, \ldots, \lambda_m . \lambda_{m+1}, \ldots, 0)$, $\lambda_{m+1} = \varepsilon\lambda_1$, and ε is a user selected threshold value (such as 0.01).

For a given ε, the eigenvalue matrix $\mathbf{D}_v$ is replaced by $\hat{\mathbf{D}}_v = diag(\lambda_1 \ldots . \lambda_i \ldots . \lambda_m , \lambda_{m+1} \ldots . \lambda_{m+1})$. All values, which are smaller than λ_{m+1}, are substituted by λ_{m+1}. After the substitution, $\hat{\mathbf{S}}_v$ is projected onto the space by:

$$\hat{\hat{\mathbf{S}}}_y = \hat{\mathbf{D}}_x^{-1/2}\mathbf{Y}^T\hat{\mathbf{S}}_y\mathbf{Y}\hat{\mathbf{D}}_y^{-1/2}. \tag{6-9}$$

Finally, the eigenvectors $\mathbf{U}$ of $\hat{\hat{\mathbf{S}}}_y$ with largest eigenvalues are selected to form the transformation matrix,

$$\mathbf{W} = \mathbf{Y}\bar{\mathbf{D}}_x^{-1/2}\mathbf{U}\mathbf{D}_y^{-1/2}. \tag{6-10}$$

6.3 Experimental results

In this part, a large number of statistical experiments are performed based on QueryGo. The experiments are simulated by the computer automatically. First, 300 queries are randomly selected from the data, and then RF is done by computer as: top 5 query relevant and irrelevant images are marked as positive and negative feedbacks in the top 48 images, respectively.

In this Section, precision and standard deviation (SD) are used to evaluate the performance of a RF algorithm. Precision is the ratio of the number of relevant images retrieved to the top N retrieved images. Precision curve is the averaged precision values of the 300 queries, and SD curve is the SD values of 300 queries' precision. The precision curve evaluates the effectiveness of a given algorithm and SD curve evaluates the robustness of the algorithm. In precision and SD curves, the total feedback times are 9, with 0 feedback referring to the retrieval based on Euclidean distance measure without RF.

6.3.1 K nearest neighbor evaluation for NDA

The experiment shows NDA is insensitive to the *k* value of the k-nearest-neighbor. Figure 1, 2, and 3 show the top 30 retrieved results with 3, 6, and 9 feedback iterations by the regularization method, null-space method, and full-space method, respectively. Because all curves are flat, we can draw the conclusion that NDA is insensitive to the *k* value in k nearest neighbor.

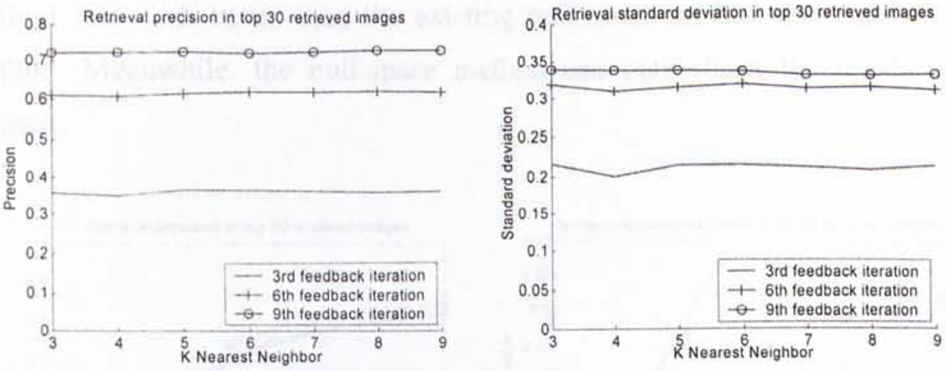


Figure 6-1. Regularization method.

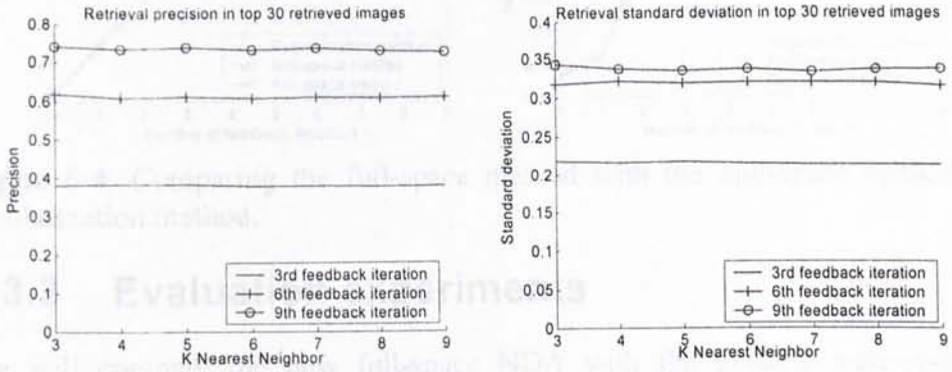


Figure 6-2. Null-space method.

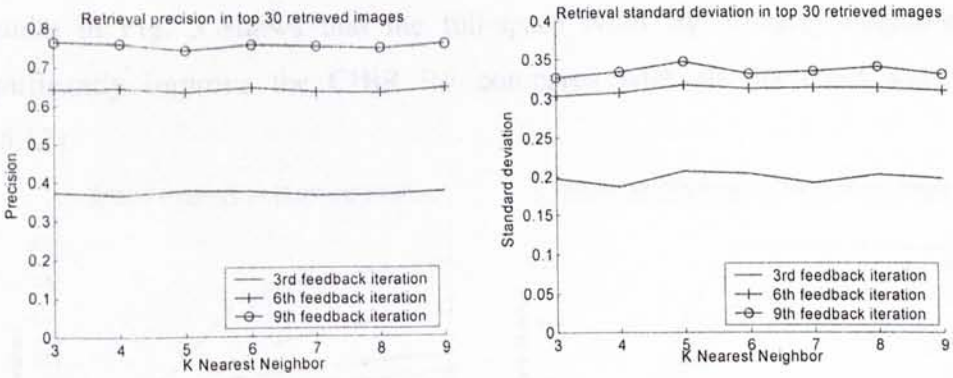


Figure 6-3. Full-space method.

6.3.2 SSS problem

Fig. 4 shows the performance of the full-space method, the null-space method, and the regularization method in NDA to solve the SSS problem. From the left subfigure in Fig. 4, we can see the precision curve of full-space method is higher than that of null-space method and regularization method, meanwhile the SD curve of full-space method is lower than that of null-space method and regularization method. Hence we can draw the conclusion the new full-space method can work better than the existing null-space method and regularization method. Meanwhile, the null-space method can outperform the regularization method.

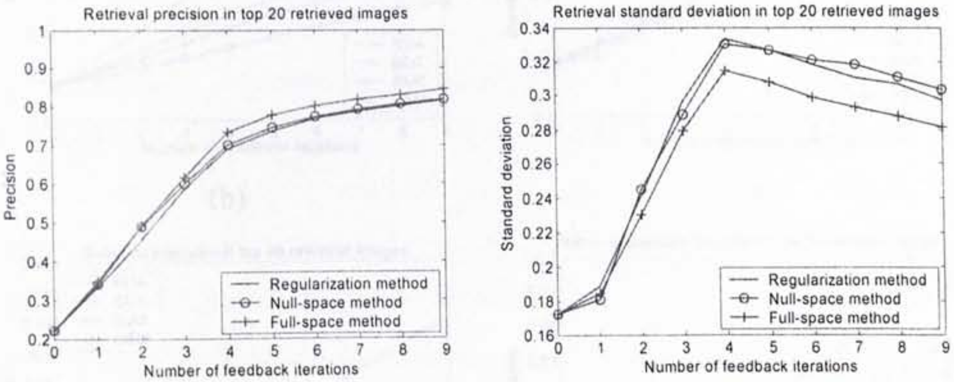
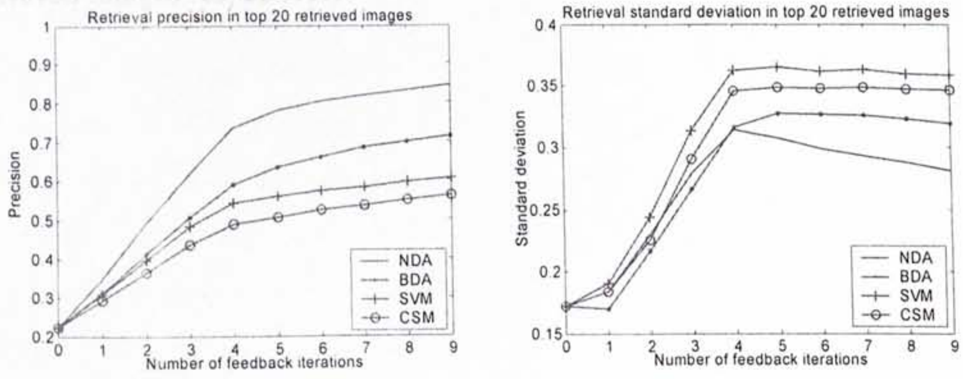


Figure 6-4. Comparing the full-space method with the null-space method and regularization method.

6.3.3 Evaluation experiments

We will compare the new full-space NDA with the existing state-of-the-art algorithms, which are BDA [5], SVM [4], and constrained SVM (CSM) [11]. Results in Fig. 5 shows that the full-space NDA by 3-nearest-neighbor can significantly improve the CBIR RF compared with all the other algorithms [4,5,11].



(a)

(d)

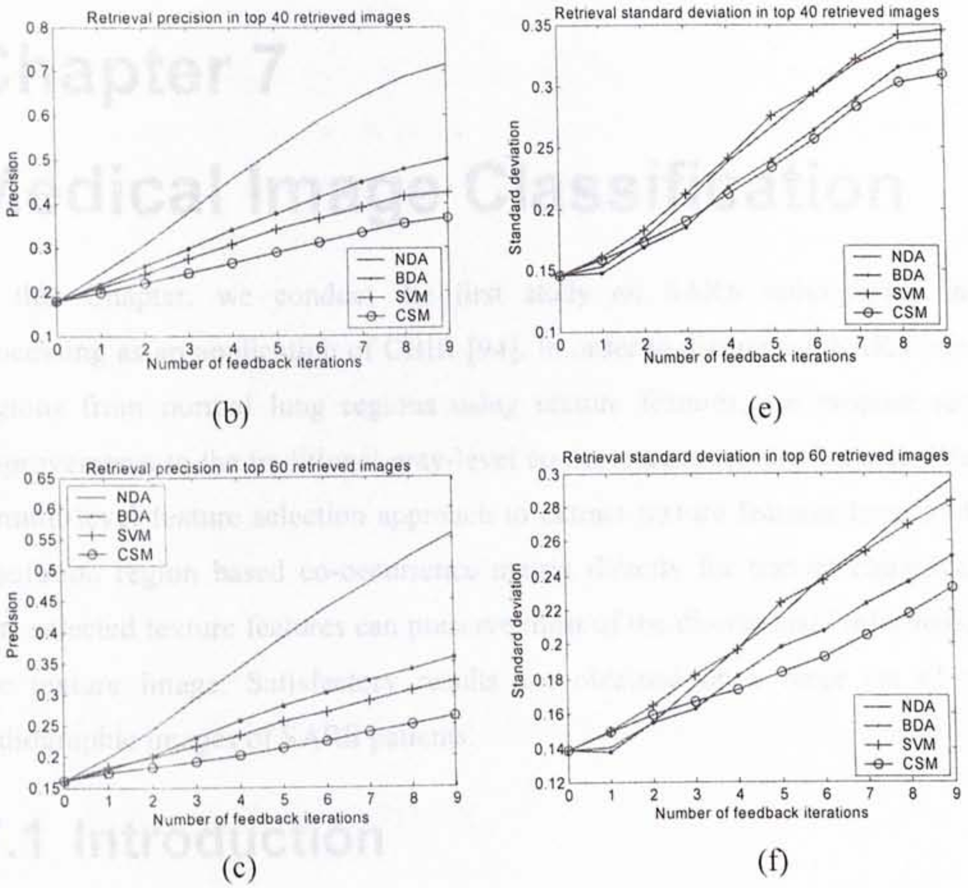


Figure 6-5. Evaluation experimental results based on the Corel database with 17, 800 images with 300 queries. (a), (b), and (c) display the retrieval precision in top 20, 40, and 60 retrieved images respectively. (d), (e), and (f) display the corresponding standard deviation of the precision curve in top 20, 40, and 60 retrieved images respectively.

Chapter 7
Medical Image Classification

In this Chapter, we conduct the first study on SARS radiographic image processing as an application of CBIR [94]. In order to distinguish SARS infected regions from normal lung regions using texture features, we propose several improvements to the traditional gray-level co-occurrence texture features. We use a multi-level feature selection approach to extract texture features from a multi-resolution region based co-occurrence matrix directly for texture classification. The selected texture features can preserve most of the discriminant information in the texture image. Satisfactory results are obtained on a large set of chest radiographic images of SARS patients.

7.1 Introduction

Severe Acute Respiratory Syndrome (SARS) outbreak in Hong Kong started in March 2003 and quickly spread to many regions around the world. By the end of the epidemic, there were 1,755 patients infected and 299 deaths in Hong Kong [56]. The main symptoms of SARS are high fever and dry cough, shortness of breath or breathing difficulties. SARS may also be associated with other symptoms including a headache. Because of the highly contagious nature of the disease and its very fast progress that often threatens the life of the patient, it is critically important to identify the disease at an early stage. However, since most of the symptoms are similar to regular pneumonia and fever, it is very difficult to give an accurate diagnosis of the disease. All currently available methods depend on laboratory testing of the virus samples from the patient, which is both costly and time consuming.

In this Chapter, we study the chest radiographs of the SARS patients to investigate a possible computer-aided approach to distinguish the SARS infected area from the normal lung area. This can be an important first step toward image based computer-aided diagnosis. Of course, it is unrealistic to expect accurate diagnosis only based on automatic computer processing of radiographic images. However, we do expect our study to be able to assist doctors with their diagnosis in the future. In addition, since for confirmed patients the chest radiographic images are taken everyday, we can also compare the progress of the images with previous patients in the database to monitor the effect of the treatment.

Because SARS regions are irregular, we cannot use shape to distinguish it from normal areas. So we focus on using texture classification to classify the SARS region. In this paper, we propose several improvements to the classic texture model, gray-level co-occurrence matrix [73], to distinguish the subtle SARS texture. We use a multi-level feature selection approach to extract texture features from a multi-resolution region based co-occurrence matrix directly for texture classification. Encouraging results are obtained on a set of chest radiographic images.

7.2 Region-based Co-occurrence Matrix Texture Feature

Co-occurrence texture features were proposed by Haralick et al. [73]. For an image with N by N pixels and G gray levels, the co-occurrence matrix for a displacement *d* in a direction *q* is defined to be a G by G matrix whose entry M(*i*, *j*) is the number of occurrences of transitions from gray level *i* to gray level *j*, given the inter-sample distance *d* and the direction *q*. The matrix gives a measure of the joint probability density of the pairs of gray levels that occur at pairs of points separated by distance *d* in the direction *q*. For a coarse texture, *d* is relatively small compared to the sizes of the texture elements; the pairs of points at separation *d* have similar intensity values. This means the matrix M has large values near its main diagonal. Conversely, for a fine texture the values in M are quite uniformly spaced. Thus, a measure of the degree of value spread around the main diagonal of M should provide a good sense of the texture coarseness. Similarly, one can extract other features to measure the directional information, contrast, correlation, etc. Haralick et al. [73] proposed 28 second-order statistic features that can be measured from this co-occurrence matrix.

Generally, the co-occurrence matrix is computed from a rectangular region or image. In our application, however, the regions are not rectangles. In order to compute the texture features, we develop a region based co-occurrence matrix:

1. Extract the marked SARS infected regions.
2. Find the maximum bounding box of each region.
3. Quantize the region with a given bin number and fill the blank part of the bounding box with –1.
4. Calculate the co-occurrence matrix $\mathbf{P}$ of the filled bounding box, and extract a sub-matrix $\mathbf{P}_s$ from $\mathbf{P}$, where $\mathbf{P}_s$ is obtained by deleting the first row and the first column of $\mathbf{P}$. The region based co-occurrence matrix is:

$$\mathbf{P}_s = \mathbf{P}_{2n,2n} \tag{7-1}$$

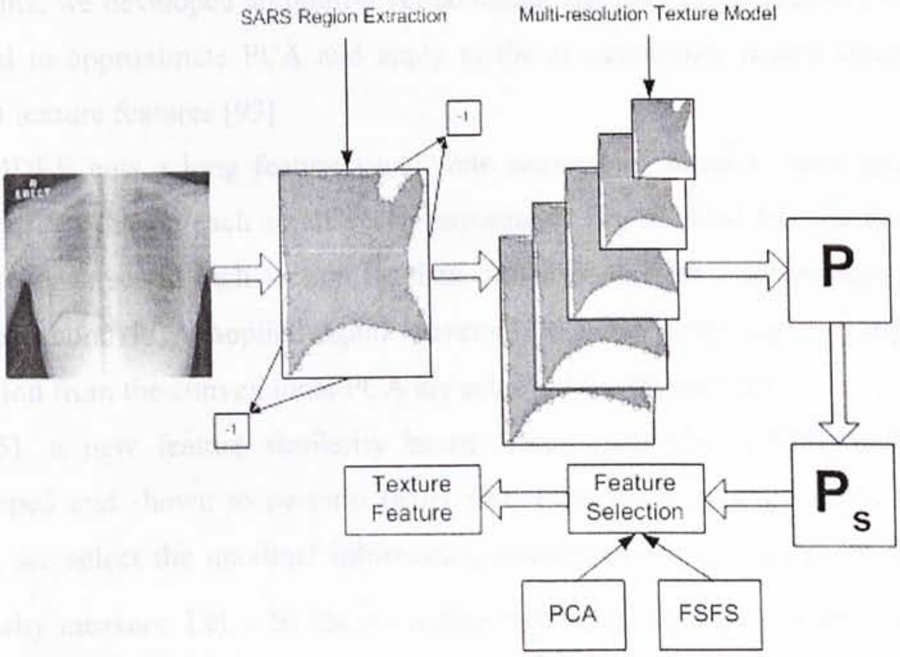


Figure 7-1. Flowchart of the multi-resolution non-rectangle region's co-occurrence matrix and texture feature extraction.

The size of the texture image is crucial for classification. To preserve more information of the texture image, we select a series scales to zoom the original image and the scale vector is: $S = [0.1\ 0.2\ 0.4\ 0.6\ 0.8\ 1.0]$, where $S = 0.6$ means the ratio between the size of the zoomed image and the original image is 0.6. For each scale, the region-based co-occurrence matrix and the corresponding statistical features are calculated.

7.3 Multi-level Feature Selection

The original texture features computed from the co-occurrence matrix are mostly based on intuitive observation of the shape and statistics of the matrix [73]. There are two drawbacks with this approach. First, there is no theoretical proof that, given a certain number of features, maximum texture information can be extracted from the co-occurrence matrix. Second, many of these features are highly correlated with each other. A better approach is to use the co-occurrence matrix as the texture feature vector directly to preserve all the information in the matrix instead of developing new functions to extract texture information. However, this again introduces two problems: the large dimensionality of the feature vector and the high-degree correlation of the neighborhood features. To alleviate these problems, we developed an multi-level dominant eigenvector estimation (MDEE) method to approximate PCA and apply to the co-occurrence matrix directly to extract texture features [93].

The MDEE cuts a long feature vector into sections of small vectors, and then performs a PCA on each small vector separately. The selected top features with large eigenvalues in each section are then combined to form a new feature vector with a second PCA applied again. Several orders of computation complexity reduction from the conventional PCA are achieved by this method.

In [65], a new feature similarity based feature selection (FSFS) method is developed and shown to perform better than PCA for feature selection. In this paper, we select the maximal information compression index (λ_2) as the feature similarity measure. Let Σ be the covariance matrix of random variables x and y. Define maximal information compression index as $\lambda_2(x,y)$ = smallest eigenvalue of Σ, i.e.,

$$\begin{aligned} 2\lambda_2 &= \left(\mathrm{var}(x)+\mathrm{var}(y)\right) \\ &\quad -\sqrt{\left(\mathrm{var}(x)+\mathrm{var}(y)\right)^2-4\,\mathrm{var}(x)\mathrm{var}(y)\left(1-\rho(x,y)^2\right)}. \end{aligned} \tag{7-2}$$

The larger the value of λ_2, the less of the dependency of the two variables. The value of λ_2 is zero means the features are linearly dependent. For feature selection, FSFS first partitions the original feature set into a number of homogeneous subsets and select a representative feature from each subset based on the similarity measure.

However, the FSFS method encounters the same problem as PCA. The computational complexity of FSFS is $O(D^2l)$, where D is the feature dimension and l is the size of the data-set. In our study, the feature dimension is 1024×6. The computational cost is too high for FSFS. In order to overcome this problem, we propose a similar multi-level approach as the MDEE method. We first apply the FSFS to feature vector for each image scale, then combine the selected features and use the FSFS again on the combined feature vector.

The flowchart of our feature selection algorithm is shown in Figure 2. For each level, we calculate the region-based co-occurrence matrix, and then FSFS or PCA is applied to the matrix directly to select first level features. We then combine all the selected features into a new feature vector and the feature selection method is used again to select the final features.

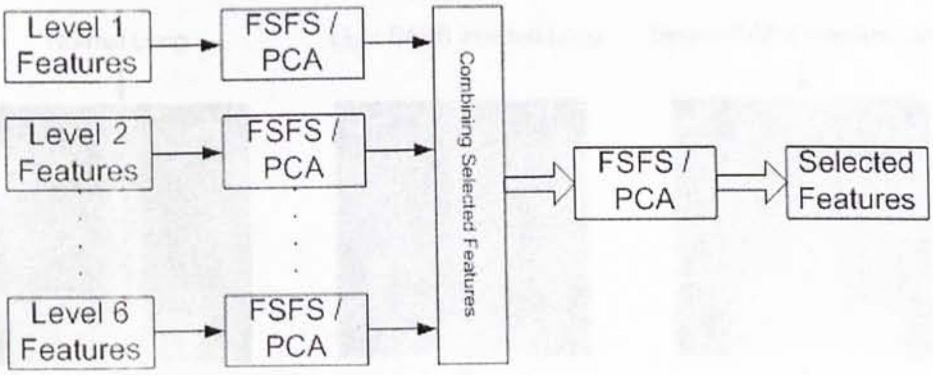


Figure 7-2. Flowchart of the multi-level feature selection method.

7.4 Experimental Results

In this Section, we use the new algorithm to classify the SARS infected region from the normal lung region in chest radiographic images. We also compare the new features with traditional co-occurrence features. We use the SVM with Gaussian Kernel as the classifier since SVM is a very effective binary classifier. All the parameters are default values in OSUSVM [34].

7.4.1 Data Set

We use the posteroanterior chest radiographs taken by the department of Diagnostic Radiology & Organ Imaging of the Prince of Wales Hospital. The digital images were obtained by digitizing the chest radiographs in the SIEMENS medical computer system. The original image has a pixel size of about 0.175mm, a matrix size of about 2000×2400, and a gray level range of 16 bits. The SARS infected regions and normal regions of all lung radiographs were labeled by doctors in the hospital as ground truth. Table 1 shows the details of the database and Figure 3 shows some sample images and SARS infected regions in the database.

Table 7-1. Image Database

	SARS	Normal
Training	37	37
Testing	126	38
Total	163	75

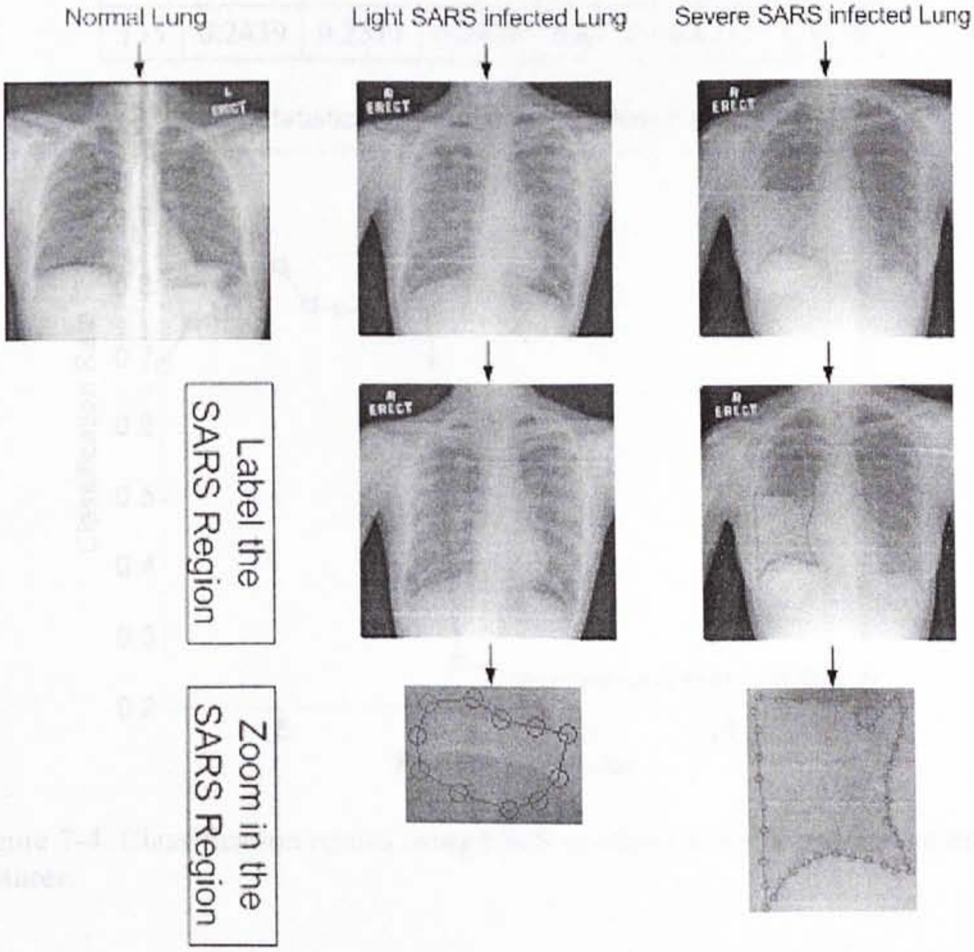


Figure 7-3. Sample images and SARS infected regions in the database.

7.4.2 Classification Using Traditional Features

We first use the traditional texture features defined in [73] to classify the images. Classification results are summarized in Table 2. The results show that the traditional feature of each co-occurrence direction in each scale cannot discriminant the SARS and normal lung regions well. Figure 4 shows the results of using FSFS to select features from all the traditional texture features (the original feature dimension is $13\times6\times4=312$). The classification result is still less than satisfactory.

Table 7-2. Traditional feature based classification. The first row is the scale value of the image and the first column is the direction of the co-occurrence matrix.

	0.1	0.2	0.4	0.6	0.8	1.0
0	0.7927	0.8110	0.8110	0.2561	0.8110	0.8171
45	0.7683	0.2500	0.2500	0.8171	0.7927	0.8171
90	0.7683	0.2561	0.8354	0.2561	0.8232	0.2561
135	0.2439	0.2317	0.2439	0.8171	0.8232	0.8476

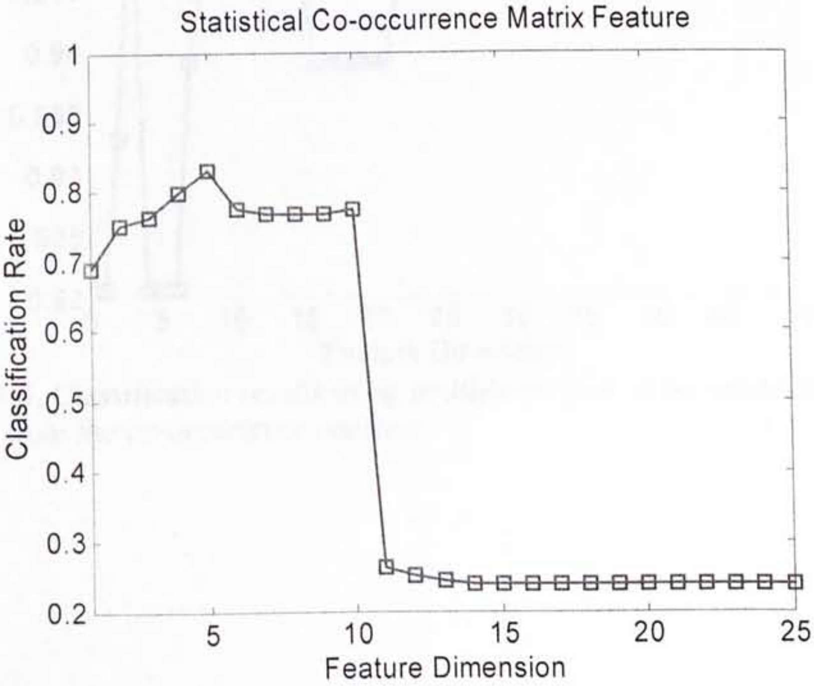


Figure 7-4. Classification results using FSFS to combine traditional co-occurrence features.

7.4.3 Classification Using the New Features

Classification results using the multilevel PCA to extract texture features directly from the co-occurrence matrices are shown in Figure 5. The recognition rate is significantly improved over the traditional features. This shows that the method can effectively preserve the discriminant texture information for SARS and normal lung region classification.

Next, we use the new multi-level FSFS method to extract the texture features directly from the original co-occurrence matrix. The recognition rate is further improved as shown in Figure 6. The highest classification rate of the method is around 97%.

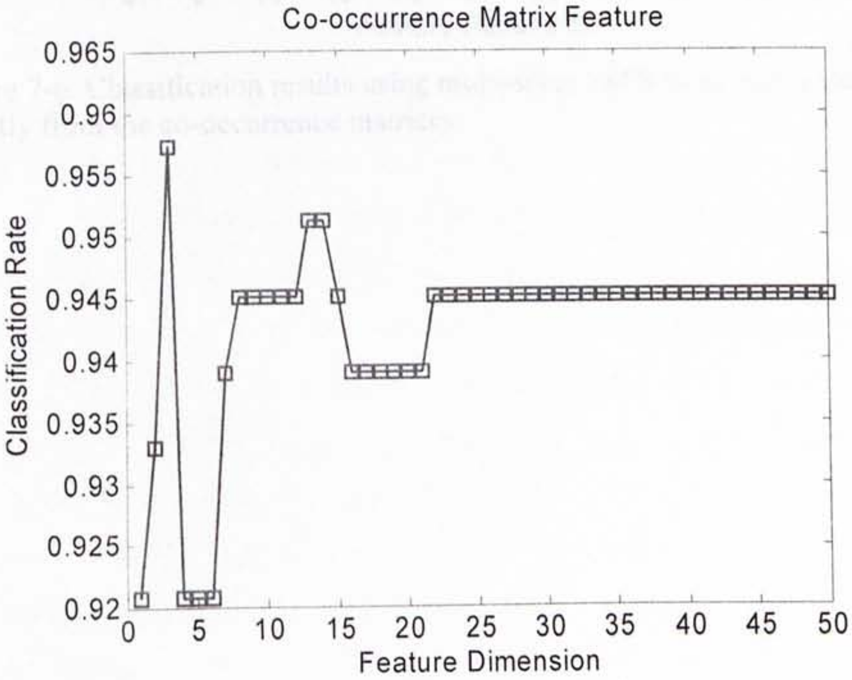


Figure 7-5. Classification results using multi-level PCA to extract texture features directly from the co-occurrence matrices.

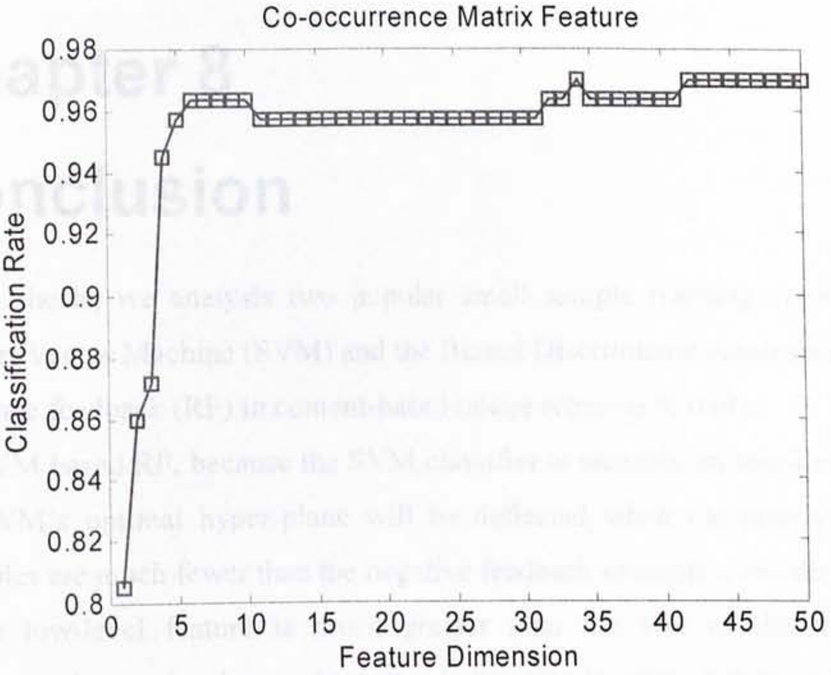


Figure 7-6. Classification results using multi-level FSFS to extract texture features directly from the co-occurrence matrices.

Chapter 8

Conclusion

In this thesis, we analysis two popular small sample learning algorithms, the Support Vector Machine (SVM) and the Biased Discriminant Analysis (BDA), for relevance feedback (RF) in content-based image retrieval (CBIR).

For SVM based RF, because the SVM classifier is unstable on small size training set, SVM's optimal hyper-plane will be deflected when the positive feedback examples are much fewer than the negative feedback examples, and the dimension of the low-level feature is much greater than the size of the training set. Consequently, we develop an Asymmetric Bagging Random Subspace Method for SVM based RF. With the new learning scheme, all the three problems in SVM based RF can be overcome to some extent. Extensive experiments on a Corel Photo database with 17, 800 images show that the new algorithm can improve the performance (both the accuracy and the efficiency) of RF significantly.

For BDA based RF, we first generalize the ideas of the Direct Linear Discriminant Analysis (DLDA) and the Null-space Linear Discriminant Analysis (NLDA) for BDA in the Hilbert space to solve the Small Sample Size (SSS) problem. Because DLDA and NLDA may lose some discriminant information, we then propose a full-space method. Finally, we implement the full-space method for kernel BDA. According to a large number of evaluation experiments in the Corel Photo Gallery, we can draw the conclusion that the proposed Direct Kernel BDA, Null-space Kernel BDA, and Full-space Kernel BDA outperform Kernel BDA consistently.

Moreover, BDA based RF assumes that all positive feedbacks form a single Gaussian distribution. This may not be the case in CBIR. Although kernel BDA can overcome the drawback to some extent, the kernel parameter tuning makes the online learning unfeasible. To avoid the parameter tuning problem and the single Gaussian distribution assumption in BDA, we construct a new nonparametric discriminant analysis (NDA). To address the small sample size problem in NDA, we introduce the regularization method and the null-space method. Because the regularization method may meet the ill-posed problem and the null-space method may lose some discriminant information, we propose here a full-space method. The proposed full-space NDA is demonstrated to outperform BDA based RF significantly with a large number of experiments in the Corel database.

Finally, as an application of CBIR and toward assisting doctors to diagnose Severe Acute Respiratory Syndrome (SARS) patients, we conduct a preliminary study on texture classification of SARS infected regions in chest radiographic images. In order to distinguish SARS infected regions from normal lung regions, we propose several improvements to the traditional gray-level co-occurrence texture features. We use a multi-level feature selection approach to extract texture features from a multi-resolution region based co-occurrence matrix directly for texture classification. The multi-level Feature Similarity-based Feature Selection algorithm is shown to be very effective in preserving most of the discriminant information in the texture image. Experiments on a large set of chest radiographic images of SARS patients give encouraging results. This is a first promising step toward computer-aided diagnosis of the disease.

Bibliography

[1] A. Gupta and R. Jain, "Visual Information Retrieval," Comm. ACM, vol. 40, no. 5, pp. 71-79, 1997.

[2] A. K. Jain and A. Vailaya, "Shape-Based Retrieval: A Case Study With Trademark Image Databases," Pattern Recognition, vol. 31 no. 9 pp. 1369-13990, 1998.

[3] A. K. Jain and A. Vailaya. "Image Retrieval Using Color and Shape," Pattern Recognition, vol. 29, no.8 pp.1233-1244, Aug. 1996.

[4] A. Laine and J. Fan, "Texture Classification by Wavelet Packet Signature," IEEE Trans. PAMI, vol. 15, no. 11, pp. 1,186-1,191, Nov. 1993.

[5] A. Natsev, R. Rastogi, and K. Shim, "WALRUS: A Similarity Retrieval Algorithm for Image Databases," SIGMOD Record, vol. 28, no. 2, pp. 395-406, 1999.

[6] A. Pentland, R.W. Picard, and S. Sclaroff, "Photobook: Tools for Content-Based Manipulation of Image Databases," In Proc. SPIE, vol. 2185, pp. 34-47, Feb. 1994.

[7] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-Based Image Retrieval at the End of the Early Years". IEEE Trans. on PAMI, vol. 22, no. 12, pp. 1349-1380, Dec. 2000.

[8] B. S Manjunath, J. Ohm, V. Vasudevan, and A. Yamada, "Color and texture descriptors," IEEE Trans. on CSVT, Vol. 11, 2001.

[9] B. S. Manjunath and W. Y. Ma. "Texture Features for Browsing and Retrieval of Image Data," IEEE Trans. on PAMI, vol.18 no. 8 pp. 837-42, Aug. 1996.

[10] C. Carson, M. Thomas, S. Belongie, J. M. Hellerstein, and J. Malik, "Blobworld: A System for Region-Based Image Indexing and Retrieval," In Proc. Visual Information Systems, pp. 509-516, June 1999.

[11] C. Nastar, M. Mitschke, and C. Meilhac, "Efficient Query Refinement for Image Retrieval," In IEEE Proc. CVPR. 1998.

[12] D. Bahler and L. Navarro, "Methods for Combining Heterogeneous Sets of Classifiers", 17th Natl. Conf. on AAAI 2000.

[13] D. Kim and C. Kim, "Forecasting Time Series with Genetic Fuzzy Predictor Ensemble". IEEE Trans. On FS. vol. 5, no. 4, pp 523-535, 1997.

[14] D. L. Swets, and J. Weng, "Discriminant analysis and eigenspace partition tree for face and object recognition from views," in Proc. of the IEEE ICAFGR, pp.192-197, Killington, Vermont, U.S.A., October 14-16, 1996.

[15] D. Tao and X. Tang, "A direct method to solve the biased discriminant analysis in kernel feature space for content based image retrieval," In Proc. IEEE ICASSP, 2004.

[16] D. Tao and X. Tang, "Kernel full-space biased discriminant analysis," In Proc. IEEE ICME, 2004.

[17] D. Tao and X. Tang, "Learning User's Perception Using Region-based SVM for Content-based Image Retrieval," In Proc. CISST, 2004.

[18] D. Tao and X. Tang, "Multi-Class Discriminant Learning for Image Retrieval," In Proc. CISST, 2004.

[19] D. Tao and X. Tang, "Nonparametric Discriminant Analysis in Relevance Feedback for Content-based Image Retrieval," In Proc. IEEE ICPR, 2004.

[20]D. Tao and X. Tang, "Orthogonal Complement Component Analysis for Positive Samples in SVM Based Relevance Feedback Image Retrieval," In Proc. IEEE CVPR, 2004.

[21]D. Tao and X. Tang, "Random Sampling Based SVM for Relevance Feedback Image Retrieval," In Proc. IEEE CVPR, 2004.

[22]D. Tao and X. Tang, "SVM-based Relevance Feedback Using Random Subspace Method Kernel full-space biased discriminant analysis," In Proc. IEEE ICME, 2004.

[23]E. Chang and B. Li, "MEGA: The Maximizing Expected Generalization Algorithm for Learning Complex Query Concepts," ACM Trans. on Information Systems, vol. 21, no. 4, pp. 347-382, Oct. 2003.

[24]F. Liu and R.W. Picard. Periodicity, "Directionality and Randomness: Wold Features for Image Modeling and Retrieval," IEEE Trans. on PAMI, 18(17) pp.722--733, July, 1996.

[25]G. Bucci, S. Cagnoni, and R. De Dominicis, "Integrating Content-Based Retrieval in a Medical Image Reference Database," Computerized Medical Imaging and Graphics, vol. 20, no. 4, pp. 231-241, 1996.

[26]G. Guo, A. K. Jain, W. Ma, and H. Zhang, "Learning Similarity Measure for Natural Image Retrieval with Relevance Feedback," IEEE Trans. on NN, vol. 12, no. 4, pp.811-820, July 2002.

[27]G. J. McLachlan, "Discriminant Analysis and Statistical Pattern Recognition," John Wiley and Sons, Inc., New York. 526 pp. 1992.

[28]G. Pass, R. Zabih, and J. Miller, "Comparing images using color coherence vectors," In Proc. ACM Multimedia 96, pp 65-73, Boston MA USA, 1996.

[29]G. Salton, "Automatic text processing," Reading, Mass., Addison-Wesley. 1989.

[30]H. Tamura, S. Mori, and T. Yamawaki, "Texture Features Corresponding to Visual Perception," IEEE Trans. on SMC, 8(6), June 1978.

[31]H. Yu and J. Yang, "A Direct LDA Algorithm for High-dimensional Data with Application to Face Recognition," Int. J. on PR, vol. 34, pp. 2067-2070, 2001.

[32]http://ausweb.scu.edu.au/aw99/papers/lu/paper.html

[33]http://nies.liacs.nl:1860/

[34]http://www.eleceng.ohio-state.edu/~maj/osu_svm/

[35]I. J. Cox, L. Miller, P. Minka, V. Papthomas, and P. Yianilos, "The Bayesian Image Retrieval System, PicHunter: Theory, Implementation and Psychophysical Experiments," IEEE Trans. on IP, vol 9, no.1, 20-37, 2000.

[36]J. Assfalg, A. del Bimbo, and P. Pala, "Using Multiple Examples for Content Based Retrieval," In Proc. ICME 2000.

[37]J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition". Int. J. on. DMKD 2, pp.121-167, 1998.

[38]J. F. Cullen, J.J. Hull, and P.E. Hart, "Document Image Database Retrieval and Browsing Using Texture Analysis," In Proc. ICDAR, pp. 718-721, 1997.

[39]J. Huang, R. Kumar, and M. Mitra, "Combining Supervised learning with color correlogram for content based image retrieval," In Proc. ACM Multimedia, pp. 325-334, Seattle, Washington, 1997.

[40]J. Huang, S. R. Kumar, M. Mitra, W. J. Zhu, and R. Zabih, "Spatial Color Indexing and Applications," Int. J. on Computer Vision, vol. 35, no. 3, pp. 245-268, 1999.

[41]J. Kittler, M. Hatef, P.W. Duin, and J. Matas, "On Combining Classifiers".

IEEE Trans. On PAMI. Vol. 20, no. 3, pp. 226-239, Mar. 1998.

[42]J. Laaksonen, M. Koskela, S. L. and E. Oja, "Self-Organizing Maps as a Relevance Feedback Technique in Content-Based Image Retrieval," Pattern Analysis & Applications, 4(2+3):140-152, June 2001.

[43]J. Lu, K. Plataniotis, and A. Venetsanopoulos, "Face Recognition Using Kernel Direct Discriminant Analysis Algorithms," IEEE Trans. on NN, vol. 14, pp. 117-126, 2003.

[44]J. M. Buijs and M. Lew, "Visual learning of simple semantics in imagescape," In Huijsmans and Smeulders, pp. 131-138. 99.

[45]J. M. Corridoni, A. del Bimbo, and P. Pala, "Image Retrieval by Color Semantics," Multimedia Systems, vol. 7, pp. 175-183, 1999.

[46]J. Mao and A. K. Jain, "Texture Classification and Segmentation Using Multiresolution Simultaneous Autoregressive Models," Pattern Recognition, 25(2), pp.173-188, 1992.

[47]J. Platt. "Probabilistic outputs for SVMs and comparisons to regularized likelihood methods". In advances in Large Margin Classifiers. MIT Press, 1999.

[48]J. P. Eakins, J.M. Boardman, and M.E. Graham, "Similarity Retrieval of Trademark Images," IEEE Multimedia, vol. 5, no. 2, pp. 53-63, Apr.-June 1998.

[49]J. Z. Wang, G. Wiederhold, O. Firschein, and X. W. Sha, "Content-Based Image Indexing and Searching Using Daubechies' Wavelets," Int'l J. Digital Libraries, vol. 1, no. 4, pp. 311-328, 1998.

[50]J. Z. Wang, J. Li, G. Wiederhold, "SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture Libraries," IEEE Trans. on PAMI, vol. 23, no. 9, pp. 947-963, Sept. 2001.

[51]K. Etemad, and R. Chellappa, "Discriminant Analysis for Recognition of Human Face Images," in Lecture Notes in Computer Science: Audioand Video- based Biometric Person Authentication, vol. 1206, pp. 127-142, 1997.

[52]K. Fukunaga, "Introduction to Statistical Pattern Recognition (2nd)," Academic Press, Boston 1990.

[53]K. Liu, Y. Cheng, J, Yang, and X. Liu, "An Efficient Algorithm for Foley-Sammon Optimal Set of Discriminant Vectors by Algebraic Method," IJPR,vol.6, pp817-829, 1992.

[54]K. R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf, "An Introduction to Kernel-based Learning Algorithms," IEEE Trans. on NN, vol 12, no. 2, Mar. 2001.

[55]K. Tieu and P. Viola, "Boosting image retrieval," In IEEE Proc. CVPR, South Carolina. 2001.

[56]K. T. Wong, G. E. Antonio, D. S. Hui et. al., "Severe acute respiratory syndrome: radiographic appearances and pattern of progression in 138 patients," Radiology, 228(2): 401-406, Aug. 2003.L. Breiman, "Bagging Predictors". Tech report 421. Sept. 1994.

[57]L. F. Chen, H.Y. Liao, M. T. Ko, J. C. Lin, and G. J. Yu, "A New LDA-based Face Recognition System Which Can Solve the Small Sample Size Problem," Int. J. on PR, vol 33, pp. 1713-1726, 2000.

[58]L. Zhang, F. Lin, and B. Zhang, "Support vector machine learning for image retrieval". In Proc. IEEE ICIP., 2001.

[59]M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by

Image and Video Content: The QBIC System," IEEE Computer, 1995.

[60]M. J. Swain and D.H. Ballard, "Color Indexing," International Journal of CV, 7(1), pp.11-32, 1991.

[61]M. Jordan and R. Jacobs,"Hierarchical Mixtures of Experts and the EM Algorithm". Int. J. on Neural Comput. vol. 6, no. 5, pp. 181-214, 1994.

[62]M. K. Mandal, F. Idris, and S. Panchanathan, "Image and Video Indexing in the Compressed Domain: A Critical Review," Image and Vision Computing, 2000.

[63]M. Nakazato and T.S. Huang, "An Interactive 3D Visualization for Content-Based Image Retrieval," In PRoc. IEEE ICME. 2001.

[64]M. Nakazato, L. Manola, and T. S. Huang, "ImageGrouper: a group-oriented user interface for content-based image retrieval and digital image arrangement," In Journal of Visual Languages and Computing, 14/4 pp.363-386, Aug., 2003.

[65]M. Pabitra, C.A. Murthy, and K. Sankar, "Unsupervised Feature Selection Using Feature Similarity," IEEE Trans. on Pattern Analysis and Machine Intelligence, 24, pp. 1-13, 2002.N. R. Howe and D.P. Huttenlocher, "Integrating Color, Texture, and Geometry for Image Retrieval," In Proc. CVPR, pp.239-247, 2000.

[66]P. Aigrain, H. Zhang, and D. Petkovic, "Content-based Representation and Retrieval of Visual Media: A State of the Art Review," Multimedia Tools and Applications, vol. 3, pp. 179-202, 1996.

[67]P. Hong, Q. Tian, and T. S. Huang. "Incorporate Support Vector Machines to Content-based Image Retrieval with Relevant Feedback," In Proc. IEEE ICIP, 2000.

[68]P. Jing, "Multi-class Relevance Feedback Content-based Image Retrieval," Computer Vision and Image Understanding, pp. 42-67 2003.

[69]Q. Iqbal and J. K. Aggarwal, "CIRES: A System for Content-based Retrieval in Digital Image Libraries," In Proc. ICRACV. Singapore, pp. 205-210, Dec. 2002.

[70]R. Brunelli, O. Mich, and C.M. Modena, "A Survey on the Automatic Indexing of Video Data," J. Visual Comm. and Image Representation, vol. 10, pp. 78-112, 1999.

[71]R. Jain, S. N. J. Murthy, P. L. J. Chen, and S. Chatterjee, "Similarity Measures for Image Databases," In Proc. SPIE, vol. 2420, pp. 58-65, Feb. 1995.

[72]R. Mehrotra and J. E. Gary, "Similar-Shape Retrieval in Shape Data Management," Computer, vol. 28, no. 9, pp. 57-62, Sept. 1995.

[73]R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," IEEE Trans. on Systems man and Cybernetics, 3, pp 610-621, 1973.S. Aksoy and R. Haralick, "Graph-Theoretic Clustering for Image Grouping and Retrieval," In Proc. IEEE CVPR, pp. 63-68, 1999.

[74]S. Berretti, A. del Bimbo, and E. Vicario, "Modeling Spatial Relationships between Color Sets," In Proc. Workshop Content-Based Access of Image and Video Libraries, 1998.

[75]S. Mukherjea, K. Hirata, and Y. Hara, "AMORE: A World Wide Web Image Retrieval Wngine," In Proc. WWW, vol. 2, no. 3, pp. 115-132, 1999.

[76]S. Stevens, M. Christel, and H. Wactlar, "Informedia: Improving Access to Digital Video," Interactions, vol. 1, no. 4, pp. 67-71, 1994.

[77]T. Chang and C. J. Kuo, "Texture analysis and classification with tree-

structured wavelet transform," IEEE Trans. on IP., Vol. 2, No. 4, Oct, pp. 429-441, 1993.

[78]T. Gevers and A. Smeulders, "Pictoseek: Combining color and shape invariant features for image retrieval," In IEEE Trans. on IP, 9(1):102-119, Jan. 2000.

[79]T. K. Ho, "The Random Subspace Method for Constructing Decision Forests". IEEE Trans. On PAMI. vol. 20, no. 8, pp. 832-844, Aug. 1998.

[80]T. K. Lau and I. King, "Montage: An Image Database for the Fashion, Textile, and Clothing Industry in Hong Kong," In Proc. ACCV, pp. 575-582, 1998.

[81]T. Kurita and T. Kato, "Learning of personal visual impression for image database systems," In Proc. ICDAR. 1993.

[82]T. P. Minka and R. W. Picard, "Interactive Learning Using a Society of Models," Pattern Recognition, vol. 30, no. 3, p. 565, 1997.

[83]V. Tresp and M. Taniguchi, "Combining Estimators Using Non-Constant Weighting Functions". Advances in NIPS. MIT Press 1995.

[84]V. Vapnik, "Statistical Learning Theory," J. Wiley, 1995.

[85]V. Vapnik, "The Nature of Statistical Learning Theory," Springer-Verlag, New York (1995).

[86]W. Y. Ma and B. S. Manjunath, "Edge Flow: A Framework of Boundary Detection and Image Segmentation," In IEEE Proc. CVPR, pp. 744-749, 1997.

[87]W. Y. Ma and B. Manjunath, "NaTra: A Toolbox for Navigating Large Image Databases," In Proc. ICIP, pp. 568-571, 1997.

[88]W. Y. Ma and H. J. Zhang, "Content-based image indexing and retrieval," in Handbook of Multimedia Computing, Borko Furht, ed. CRC Press, 1998.

[89]W. Zhao, R. Chellappa, and A. Krishnaswamy, "Discriminant analysis of principal components for face recognition," In Proc. IEEE ICAFGR, pages 336-341, 1998.

[90]X. S. Zhou, T. S. Huang, "Comparing Discriminanting Transformations and SVM for Learning during Multimedia Retrieval," In Proc. ACM Multimedia, 2001.

[91]X. S. Zhou, T. S. Huang, "Small Sample Learning During Multimedia Retrieval Using Biasmap," In Proc. IEEE CVPR, 2001.

[92]X. S. Zhou, T.S. Huang, "Relevance Feedback for Image Retrieval: a Comprehensive Review," ACM Multimedia Systems Journal, vol. 8, no. 6, pp. 536-544, Apr. 2003.

[93]X. Tang, "Texture Information in Run Length Matrices," IEEE Trans. on IP. vol. 7, No. 11, pp. 1602 - 1609, Nov. 1998.

[94]X. Tang, D. Tao, and G. E. Antonio, "Texture Classification of SARS Infected Region in Radiographic Image," In Proc. ICIP, 2004.

[95]Y. Chen, X. S. Zhou, and T. S. Huang, "One-class SVM for Learning in Image Retrieval," In Proc. IEEE ICIP, 2001.

[96]Y. Cheng, Y. Zhuang, and J. Yang, "Optimal Fisher Discriminant Analysis using the Rank Decomposition," Pattern Recognition, vol. 25, pp.101-111, 1992.

[97]Y. Ishikawa, R. Subramanya, and C. Faloutsos, "Mindreader: Querying Databases through Multiple Examples," 24th Int. Conf. on VLDB 1998, pp.433-438, 1998.

[98]Y. Rubner, L. J. Guibas, and C. Tomasi, "The Earth Mover's Distance, Multi-Dimensional Scaling, and Color-Based Image Retrieval," In Proc. DARPA Image Understanding Workshop, pp. 661-668, May 1997.

[99]Y. Rui, T. Huang, and S. Mehrotra, "Content-based Image Retrieval with Relevance Feedback in MARS," In Proc. IEEE ICIP, pp.815-818, Oct. 1997.

[100] Y. Rui, T. Huang, M. Ortega, and S. Mehrotra. "Relevance Feedback: A Power Tool in Interactive Content-based Image Retrieval," IEEE Tran. on CSVT, vol.8 no.5 pp. 644-655, Sept. 1998.

[101] Y. Rui and T. S. Huang, "Optimizing Learning in Image Retrieval," In IEEE Proc. CVPR 2000.

[102] Y. Rui. T. S. Huang, and S. F. Chang, "Image retrieval: Current techniques, promising directions and open issues," Journal of Visual Communication and Image Representation, 10(1):1-23, Mar. 1999.

[103] Y. Wu, Q. Tian, and T. S. Huang, "Discriminant EM Algorithm with Application to Image Retrieval," In IEEE Proc. CVPR. 2000.

[104] Z. N. Li, O. R. Zaiane, and Z. Tauber, "Illumination invariance and object model in content-based image and video retrieval," Journal of Visual Communication and Image Representation, 10(3):219-244, Sep. 1999.