Content-based image retrieval with statistical machine learning.

Zhang, Lining.

2013

Zhang, L. (2013). Content-based image retrieval with statistical machine learning. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/54889>

<https://doi.org/10.32657/10356/54889>

CONTENT-BASED IMAGE RETRIEVAL WITH STATISTICAL MACHINE LEARNING

Zhang Lining

School of Electrical and Electronic Engineering

A Thesis Submitted to the Nanyang Technological University
in Partial Fulfilment of the Requirements for the Degree of
Doctor of Philosophy

January, 2013

Acknowledgments

Professor Lipo Wang is an excellent advisor and very nice friend! During the past three years for my PhD study in Singapore, I owed him a lot, because I have obtained persistent guidance from him when I was confused; and I have obtained everlasting support from him when I was depressed. I am anxious to have his comments and suggestions, because they are precious stairs for my research. I would like to express my deepest thanks and gratitude to Professor Lipo Wang.

I would like to thank my second supervisor and friend Professor Weisi Lin for his suggestions, encouragements, and funding support. All his efforts towards my study have significantly improved many aspects of my development.

I would like to thank my fantastic collaborator and friend Professor Dacheng Tao at the University of Technology Sydney for his valuable inspirations and suggestions during my PhD study.

I would like to thank Professor Shuicheng Yan at the National University of Singapore for his kindly providing a position in the Learning and Vision Research Group during my transition period in Singapore. His great enthusiasm and profound knowledge on computer vision has impressed me immensely.

I need to thank the financial support from the Institute for Media Innovation, the College of Engineering, and the School of Electrical and Electronic Engineering. Without them, I have no chance to finish my PhD study in Singapore.

I would also like to thank Ms. Leow-How Seok Lai at the Media Technology Laboratory for her help on computing resources. Without her help, I would not have obtained my experimental results in time.

I would like to thank my good friends, Dr. Chenwei Deng, Dr. Anmin Liu, Mr. Yuming Fang, Mr. Zhouye Gu, Ms. Lu Dong, Ms. Huan Yang, Dr. Jiang Qian, Dr.

Haijian Zhang, Dr. Junyang Pan, Ms. Ye Luo, Ms. Lu Wang, Ms. Zhen Zuo, Mr. Li Wang, Mr. Yangtao Wang, Mr. Lifan Zhao, Mr. Renliang Weng, Mr. Guang Hua, Mr. Yuan Wang, Mr. Jun Deng, Mr. Kwang Sing Yew, Mr. Yinjun Zou, Dr. Tao Gan, Mr. Qi Dong, Mr. Yushuang Tian, Mr. Gang Yu, Mr. Hongxing Wang, Mr. Zhou Ren, Mr. Sixuan Zhao, etc. I would also like to thank all good friends at the Media Technology Laboratory and the Institute for Media Innovation for their friendly support and help.

Last but not least, I was so lucky to meet a fantastic and pretty girl, who understood me, encouraged me, and supported me so much all the time. Moreover, I owe my family a lot for many things. I thank my kindly parents and all my family members for their generous understanding and infinite support of my PhD study.

Contents

Acknowledgments	i
Summary	vii
List of Figures	viii
List of Tables	xiv
List of Abbreviations	xvii
List of Symbols	xix
1 Introduction	1
1.1 Content-Based Image Retrieval	1
1.1.1 Overview	1
1.1.2 Open Issues	4
1.2 Statistical Machine Learning	6
1.2.1 Unsupervised Learning	6
1.2.2 Supervised Learning	7
1.2.3 Weakly Supervised Learning	7
1.2.4 Semi-supervised Learning	7
1.2.5 Active Learning	8
1.3 Thesis Organization and Contributions	8
2 Background Review	15
2.1 Content-Based Image Retrieval	15
2.1.1 Conventional RF Methods	15
2.1.2 Collaborative Image Retrieval	17
2.2 Statistical Machine Learning	18
2.2.1 Unsupervised Learning	19

2.2.2	Supervised Learning	21
2.2.3	Weakly Supervised Learning	24
2.2.4	Semi-supervised Learning	25
2.2.5	Active Learning	26
3	Generalized Biased Discriminant Analysis	29
3.1	Motivation	29
3.2	Related Work	32
3.2.1	Biased Discriminant Analysis	32
3.2.2	Differential Scatter Discriminant Criterion	33
3.3	Generalized Biased Discriminant Analysis	34
3.4	Experimental Results	38
3.4.1	Parameter Sensitivity	40
3.4.2	Statistical Experimental Results	43
3.4.3	Discussions	47
3.5	Summary	49
4	Semi-supervised Biased Maximum Margin Analysis	51
4.1	Motivation	51
4.2	Related Work	54
4.2.1	Traditional SVM-based RF for CBIR	54
4.2.2	The Graph Embedding Framework	56
4.3	BMMA and SemiBMMA for SVM-based RF	58
4.4	Content-Based Image Retrieval	66
4.5	Experimental Results	69
4.5.1	Intrinsic Problems in Traditional SVM-based RF	69
4.5.2	Feature Extraction Based on Different Methods	70
4.5.3	Statistical Experimental Results	71
4.6	Summary	83

5	Geometric Optimum Experimental Design	85
5.1	Motivation	85
5.2	Related Work	89
5.2.1	Selecting Most Informative Samples	89
5.2.2	Optimum Experimental Design	90
5.3	Geometric Optimum Experimental Design	93
5.3.1	Problem Definition	93
5.3.2	Geometric Optimum Experimental Design	94
5.4	Collaborative Image Retrieval	102
5.5	Experimental Results	104
5.5.1	Experiments on Synthetic Datasets	104
5.5.2	Experiments on a Real-World Image Database	107
5.5.3	Parameter Sensitivity	115
5.5.4	Discussions	116
5.6	Summary	117
6	Conjunctive Patches Subspace Learning with Side Information	119
6.1	Motivation	119
6.2	Related Work	122
6.2.1	Exploiting RF Log Data	122
6.2.2	Subspace Learning and Distance Metric Learning	123
6.3	Conjunctive Patches Subspace Learning with Side Information	124
6.3.1	Problem Definition	124
6.3.2	Conjunctive Patches Subspace Learning with Side Information . .	127
6.4	Collaborative Image Retrieval	136
6.5	Experimental Results	138
6.5.1	Experiments with Synthetic Datasets	140
6.5.2	Experiments on the CIR System with RF Log Data	141
6.5.3	Parameter Sensitivity	148
6.6	Summary	151

7 Thesis Conclusions and Future Work	153
7.1 Thesis Conclusions	153
7.2 Future Work	157
Publications	159
References	161

Summary

Content-based image retrieval (CBIR) has attracted intensive attention in the computer vision community during the last decades. Relevance feedback (RF) is a powerful tool to bridge the gap between low-level visual features and high-level semantic concepts in CBIR. Although many algorithms have obtained promising performance in various practical applications, CBIR is still an open research topic mainly due to the difficulties in bridging the semantic gap. In this thesis, we mainly focus on applying statistical machine learning techniques to maximize the potential of conventional RF methods to significantly improve the performance of CBIR.

To alleviate the small-sized training data problem in conventional discriminant analysis based RF, i.e., biased discriminant analysis (BDA), a generalized BDA (GBDA) method is developed based on the differential scatter discriminant criterion (DSDC). By redesigning the between-class scatter matrix and integrating the locality preserving principle, GBDA can also avoid the Gaussian distribution assumption for the positive feedback samples and the overfitting problem in BDA. The new method can outperform BDA and its extensions significantly, as shown by a large number of empirical studies.

To incorporate the asymmetric property of training data with conventional classification based RF, i.e., support vector machine (SVM)-based RF, a biased maximum margin analysis (BMMA) method is designed based on the graph embedding framework to separate the positive and negative feedback samples by a maximum margin in the reduced subspace. By introducing a Laplacian regularizer to BMMA, semi-supervised BMMA (SemiBMMA) is also proposed to utilize the information of unlabeled samples for SVM-based RF. Experiments on a real-world image database have demonstrated that the proposed scheme combined with SVM-based RF can better model the RF procedure and reduce the performance degradation caused by the asymmetric property of training data.

To select the most informative samples for the user to label, a geometric optimum experimental design (GOED) method is proposed to select multiple representative samples in the database as the most informative ones. GOED can alleviate the small-sized training data problem by leveraging the geometric structure of unlabeled samples in the reproducing kernel Hilbert space (RKHS), and thus can further enhance the performance of image retrieval. By minimizing the expected average prediction variance on the test data, GOED has a clear geometric interpretation to select a set of the most representative samples in the database iteratively with the global optimum. Moreover, the new method is label-independent and can effectively avoid various potential problems caused by insufficient and inexactlly labeled samples in RF. Extensive experiments on both synthetic datasets and a real-world image database have confirmed the advantages of GOED.

To exploit the RF log data, conjunctive patches subspace learning (CPSL) with side information is developed. CPSL can directly learn a semantic concept subspace from the RF log data with a set of similar and dissimilar pairwise constraints without using any explicit class label information, and this is more practical and useful in many real-world applications. CPSL can be formulated as a constraint optimization problem, and an efficient algorithm is presented to solve this task with closed-form solutions. Moreover, the new method can also learn a distance metric but performs more effectively and efficiently when dealing with high-dimensional data. The effectiveness of CPSL in exploiting the RF log data to improve the performance of CBIR has been demonstrated by a large number of empirical studies.

List of Figures

1.1	The organization of the thesis.	9
3.1	APs of GBDA with different k values in top N results based on 400 independent experiments. (a) top 10, (b) top 30, and (c) top 50.	41
3.2	APs and ARs of GBDA with different β_2 values in top 30 results based on 500 independent experiments. (a) APs and (b) ARs.	42
3.3	APs of GBDA compared with the RF methods, i.e., NBDA, MBA, DBDA, BDA, SVM, and CSVM, based on 500 independent experiments. All the results are evaluated over 9 rounds of RF. 0 round of RF refers to the retrieved results based on the Euclidean distance metric without any RF information.	44
3.4	SDs of GBDA compared with the RF methods, i.e., NBDA, MBA, DBDA, BDA, SVM, and CSVM, based on 500 independent experiments. All the results are evaluated over 9 rounds of RF. 0 round of RF refers to the retrieved results based on the Euclidean distance metric without any RF information.	46
3.5	APs and ARs of the compared algorithms, i.e., NBDA, MBA, DBDA, BDA, SVM, and CSVM, after the ninth round of RF. (a) APs and (b) ARs.	48
4.1	A typical set of positive and negative feedback samples at each round of RF.	52
4.2	(a) red dots are the positive feedback samples and blue dots are the negative feedback samples in the original space, (b) the BMMA method, and (c) the positive and negative feedback samples in the maximum margin subspace.	63

4.3	The illustration of the optimal hyperplane comparison between BMMA SVM and SemiBMMA SVM for two classes of feedback samples.	65
4.4	The flowchart of our CBIR system.	68
4.5	The optimal hyperplane of SVM is diverse for different combinations of the features.	70
4.6	The optimal hyperplane of SVM is unstable and complex when dealing with a small number of training feedback samples.	71
4.7	Performance comparisons of four different feature extraction methods, i.e., LDA, BDA, MFA, and BMMA, for two classes of feedback samples, i.e., positive feedback samples and negative feedback samples, in RF. In experiments, red “o” and blue “Δ” denote the positive and negative feedback samples, respectively. Black dotted lines, blue dot dash lines, green dot dash lines, and red full lines indicate LDA, BDA, MFA, and BMMA, respectively. (a)-(f) show the experimental results of the four feature extraction methods when handling the feedback samples with various different distributions, respectively.	72
4.8	Example categories used in the small-scale image database.	74
4.9	APs with 30 categories in top 20 results of the compared algorithms, i.e., SVM, OCCA SVM, and BMMA SVM, after the second round of RF. . .	75
4.10	APs of the compared algorithms, i.e., SVM, OCCA SVM, and BMMA SVM, after the first and second rounds of RF. (a) the first round of RF and (b) the second round of RF.	76
4.11	APs of the compared algorithms, i.e., SemiBMMA SVM, BMMA SVM, OCCA SVM, BDA SVM, SVM, and One SVM, in top 10-60 results with fivefold cross validation. All the methods are evaluated over 9 rounds of RF. 0 round of RF refers to the retrieved results based on the Euclidean distance metric without any RF information.	78
4.12	SDs of the compared algorithms, i.e., SemiBMMA SVM, BMMA SVM, OCCA SVM, BDA SVM, SVM, and One SVM, in top 10-60 results with fivefold cross validation. All the methods are evaluated over 9 rounds of RF. 0 round of RF refers to the retrieved results based on the Euclidean distance metric without any RF information.	80

4.13	The performance of the proposed scheme for four different query images, i.e., bobsled, cloud, cat, and car, after the fourth round of RF.	82
5.1	A toy example to show the mechanism of SVMactive in selecting the most informative unlabeled samples for the user to label. The red solid dot, green cross point, and open circles indicate the positive feedback sample, negative feedback sample, and unlabeled samples, respectively. SVMactive can select the ambiguous unlabeled samples “A” and “B” for their low prediction confidence but cannot select the more representative unlabeled sample “B”.	87
5.2	The framework of our CIR system. Different from conventional CBIR systems, our system can select the most informative images in the database for the user to label and return the most semantically relevant images as the final results.	102
5.3	Some example images in the Corel image database.	104
5.4	The illustration of the unstable problem of the optimal hyperplane of SVM when dealing with small-sized training data. The open circles, triangles, and crosses denote the positive feedback samples, negative feedback samples, and unlabeled samples, respectively. The white solid line indicates the optimal hyperplane of SVM, which separates the positive and negative feedback samples. The unlabeled samples closest to the optimal hyperplane of SVM will be identified as the most informative ones. . . .	105
5.5	The performance comparison between TED and GOED in selecting the most representative samples. The black dots and blue crosses indicate the relevant and irrelevant unlabeled samples in the database, respectively. The red open circles are the most representative samples selected by TED and GOED in the database.	106
5.6	APs with 30 categories in top 20 results of the compared algorithms, i.e., GOED, LSRR, and Baseline, after the second round of RF.	108
5.7	APs of the compared algorithms, i.e., GOED, LSRR, and Baseline, after the first and second rounds of RF. (a) the first round of RF and (b) the second round of RF.	109

5.8	APs of GOED compared with the algorithms, i.e., TED, SVMactive, SVM, and LSRR. All the algorithms are evaluated over 9 rounds of RF. 0 round of RF refers to the retrieved results based on the Euclidean distance metric without any RF information.	111
5.9	SDs of GOED compared with the algorithms, i.e., TED, SVMactive, SVM, and LSRR. All the algorithms are evaluated over 9 rounds of RF. 0 round of RF refers to the retrieved results based on the Euclidean distance metric without any RF information.	112
5.10	The performance sensitivity of GOED with regard to the data-dependent kernel deformation parameter γ . (a) APs and (b) ARs.	115
6.1	Different similar relations between pairs of images based on different semantic concept subspaces in a multi-dimensional low-level visual feature space. (a) four images with low-level visual features, (b) similar in the shape subspace, (c) similar in the size subspace, and (d) similar in the texture subspace.	120
6.2	The illustration of the local discriminative patch for an image and its associated nearest similar and dissimilar images. Circular solid dots and square solid dots denote labeled similar and dissimilar images, respectively. For each given image, minimizing the objective function of the local discriminative patch will pull the nearest similar images towards this image while pushing the nearest dissimilar images away from this image in the reduced subspace.	129
6.3	The illustration of the local geometric patch for an image and its associated nearest similar images. Circular solid dots denote labeled similar images. For each given image, the local geometric patch aims to preserve the local geometric information of labeled similar images before and after projection. Minimizing the objective function of the local geometric patch will reconstruct the given image from its associated nearest similar images with a minimal error in the reduced subspace.	130

6.4	The illustration of the local weakly similar patch for an image and its associated nearest labeled and unlabeled images. Solid dots and hollow dots denote labeled and unlabeled images, respectively. For each given image, the local weakly similar patch attempts to impose local consistency constraints on this image and its associated nearest images. Minimizing the objective function of the local weakly similar patch will incorporate the weakly similar information of unlabeled images in the reduced subspace.	132
6.5	The framework of our CIR system. Different from conventional CBIR systems, our system can integrate conventional RF schemes with an offline RF log data exploiting scheme.	136
6.6	Some example images in the log database.	137
6.7	The performance comparisons of three different subspace learning methods, i.e., LDA, BDA, and CPSL, for two sets of samples, i.e., similar samples and dissimilar samples, in RF. In experiments, red “+” and blue “o” denote similar and dissimilar samples, respectively. Black dotted lines, green dot dash lines, and red full lines indicate LDA, BDA, and CPSL, respectively. (a)-(f) show the experimental results of three subspace learning methods when handling samples with various different distributions, respectively.	139
6.8	APs and ARs of the compared algorithms, i.e., CPSL, Euclidean distance metric, Mahalanobis distance metric, RCA, DCA, and Xing, for the 200 log data. (a) APs and (b) ARs.	142
6.9	APs and ARs of the compared algorithms, i.e., CPSL, Euclidean distance metric, Mahalanobis distance metric, RCA, DCA, and Xing, for the 600 log data. (a) APs and (b) ARs.	145
6.10	APs in top 50 results and top 100 results of the three different RF methods, i.e., SVM-based RF, CPSL-I SVM-based RF, and CPSL-II SVM-based RF, for image retrieval.	146
6.11	The AP surface of CPSL subject to different k_1 and k_2 for the two log databases, i.e., (a) 200 log data and (b) 600 log data.	149
6.12	Performance of CPSL with different β_1 for the two log databases, i.e., (a) 200 log data and (b) 600 log data.	150

6.13 Performance of CPSL with features projected onto the subspaces with different dimensions for the two log databases, i.e., (a) 200 log data and (b) 600 log data.	151
---	-----

List of Tables

3.1	Generalized Biased Discriminant Analysis	39
3.2	APs in top N results of the compared algorithms, i.e., NBDA, MBA, DBDA, BDA, SVM, and CSVM, after the ninth round of RF	47
3.3	ARs in top N results of the compared algorithms, i.e., NBDA, MBA, DBDA, BDA, SVM, and CSVM, after the ninth round of RF	47
4.1	Semi-supervised Biased Maximum Margin Analysis	67
4.2	APs and SDs in top N results of the compared algorithms, i.e., SemiBMMA SVM, BMMA SVM, OCCA SVM, BDA SVM, SVM, and OneSVM, after the ninth round of RF (APs $\pm$ SDs)	81
5.1	Geometric Optimum Experimental Design	101
5.2	APs and SDs in top N results of the compared algorithms, i.e., GOED, TED, LSRR, SVMactive, and SVM, after the ninth round of RF (APs $\pm$ SDs).	114
6.1	APs in top N results of the compared algorithms, i.e., Euclidean distance metric, Mahalanobis distance metric, RCA, DCA, Xing, and CPSL, for the log database with 200 log data	142
6.2	ARs in top N results of the compared algorithms, i.e., Euclidean distance metric, Mahalanobis distance metric, RCA, DCA, Xing, and CPSL, for the log database with 200 log data	142
6.3	APs in top N results of the compared algorithms, i.e., Euclidean distance metric, Mahalanobis distance metric, RCA, DCA, Xing, and CPSL, for the log database with 600 log data	145

6.4	ARs in top N results of the compared algorithms, i.e., Euclidean distance metric, Mahalanobis distance metric, RCA, DCA, Xing, and CPSL, for the log database with 600 log data	145
-----	---	-----

List of Abbreviations

AP	Average Precision
AR	Average Recall
BDA	Biased Discriminant Analysis
BMMA	Biased Maximum Margin Analysis
CBIR	Content-Based Image Retrieval
CIR	Collaborative Image Retrieval
CPSL	Conjunctive Patches Subspace Learning
CSVM	Constrained Support Vector Machine
DBDA	Direct Biased Discriminant Analysis
DCA	Discriminative Component Analysis
DKBDA	Direct Kernel Biased Discriminant Analysis
DSDC	Differential Scatter Discriminant Criterion
FDC	Fisher Discriminant Criterion
GBDA	Generalized Biased Discriminant Analysis
GOED	Geometric Optimum Experimental Design
LDA	Linear Discriminant Analysis
LE	Laplacian Eigenmap
LLE	Locally Linear Embedding
LMNN	Large Margin Nearest Neighbor
LPP	Locality Preserving Projections
LSRR	Least-Square Ridge Regression
MARS	Multimedia Analysis and Retrieval System
MBA	Marginal Biased Analysis
MCML	Maximally Collapsing Metric Learning
MDS	Multi-Dimensional Scaling
NBDA	Null Space Biased Discriminant Analysis
NCA	Neighborhood Component Analysis
MFA	Marginal Fisher Analysis
MMC	Maximum Margin Criterion
OED	Optimum Experimental Design
OCCA	Orthogonal Complement Component Analysis

OneSVM	One-class Support Vector Machine
PCA	Principle Component Analysis
RCA	Relevant Component Analysis
QBIC	Query By Image Content
RF	Relevance Feedback
RKHS	Reproducing Kernel Hilbert Space
SD	Standard Deviation
SDA	Semi-supervised Discriminant Analysis
SemiBMMA	Semi-supervised Biased Maximum Margin Analysis
SMO	Sequential Minimal Optimization
SVM	Support Vector Machine
SVMactive	Support Vector Machine Active Learning
TED	Transductive Experiment Design
WLD	Weber's Law Descriptors

List of Symbols

b	The bias
c	The number of classes
f	The objective function of a regularized learning problem
f	The evaluation functions on the labeled samples
g	The evaluation functions on the unlabeled samples
h	The dimension of the high-dimensional space
i	The index of the sample
k	The number of nearest neighborhood samples
$k(\cdot, \cdot)$	The original kernel function
$\tilde{k}(\cdot, \cdot)$	The deformed kernel function
l	The dimension of the low-dimensional space
m	The mean of the training samples
n	The number of samples
n^+	The number of positive feedback samples
n^-	The number of negative feedback samples
s	The number of support vectors
u_i	The i th eigenvector
v	The weights of the within-class scatter matrix
w	The weights of the between-class scatter matrix
x_i	The i th sample
x^+	The positive feedback sample
x^-	The negative feedback sample
$\bar{x}^+$	The mean vector of the positive feedback samples
α	The coefficients of support vector machine
β	The trade off parameter
ε	The measurements noise
ρ	The kernel parameter
ξ	The error term
$\hat{w}$	The optimal solution to the least-squares problem
λ_i	The i th eigenvalue
σ^2	The constant variance
δ^2	The variance of the heat kernel

$\phi(\cdot)$	The kernel mapping function
$\psi(\cdot)$	The reconstruction errors
$\ f\ _I$	The smooth penalty term
$\ f\ _K$	The induced norm
C	The penalty parameter
$\hat{C}$	The covariance matrix of a chunklet
C_w	The covariance matrix
D	The diagonal matrix of the graph Laplacian
D^+	The diagonal matrix of the within-class scatter weighting matrix
D^-	The diagonal matrix of the between-class scatter weighting matrix
D^u	The diagonal matrix of the intrinsic graph for unlabeled samples
D^m	The diagonal matrix of the manifold regularization term
D^I	The diagonal matrix of the intrinsic graph Laplacian
D^P	The diagonal matrix of the penalty graph Laplacian
H_K	The reproducing kernel Hilbert space
$\tilde{H}_{\tilde{K}}$	The new deformed reproducing kernel Hilbert space
I	The identity matrix
G	The graph
G^I	The intrinsic graph
G^P	The penalty graph
K	The kernel Gram matrix
$\tilde{K}$	The deformed kernel Gram matrix
L	The graph Laplacian
L^I	The intrinsic graph Laplacian
L^P	The penalty graph Laplacian
L^U	The intrinsic graph Laplacian for unlabeled samples
$L(\cdot, \cdot)$	The loss function
M	The optimal Mahalanobis distance metric
$N(\cdot)$	The set of the nearest neighbor samples
$ N^S $	The number of nearest positive samples for each positive sample
$ N^p $	The number of nearest sample pairs with different class labels
R^h	The h -dimensional space
R^l	The l -dimensional space
S_b	The between-class scatter matrix
S_w	The within-class scatter matrix
S_+	The positive within-class scatter matrix
S_-	The negative between-class scatter matrix
S_U	The intrinsic graph for the unlabeled samples
U	The projection matrix
W	The weighting matrix
W^+	The weighting matrix of the within-class scatter
W^-	The weighting matrix of the between-class scatter

W^m	The weighting matrix of the manifold regularization term
W^u	The weighting matrix of the intrinsic graph for unlabeled samples
W^I	The weighting matrix of the intrinsic graph
W^P	The weighting matrix of the penalty graph
X	The samples
X^+	The positive feedback samples
X^-	The negative feedback samples
Y	The new representations of samples
Y^+	The new representations of the positive feedback samples
Y^-	The new representations of the negative feedback samples
Z	The most informative samples

Chapter 1

Introduction

1.1 Content-Based Image Retrieval

1.1.1 Overview

With the rapid popularization of digital cameras and mobile phone cameras, we have witnessed an explosive growth in the size of digital image collections. Image retrieval, which is an area of the study concerned with searching and browsing images from a large-scale image database, has attracted much attention among researchers in the fields of image processing, computer vision, and database management both from the academia and industry during the last decades [1–3].

A popular framework of image retrieval is to first annotate the images by textual descriptions, and then use conventional text based database management techniques to perform the image retrieval task [4–6]. During recent years, various research works, such as data modeling, multi-dimensional indexing, and query evaluation, have been widely carried out regarding the text-based image retrieval study. However, there are still some severe limitations of these methods, which make them far from working in real-world applications, especially when the size of image collections is extremely large. A key step in these text-based image retrieval studies is to manually annotate numerous images with textual descriptions by the user, and then the textual descriptions are used to index the images in the database. However, it is quite a challenging task for the user and almost prohibited in real-world applications. Although there are some promising advances of image annotation techniques in recent years [7–9], to automatically annotate an image is

still far beyond the current technology in the computer vision community. Moreover, it is almost impossible to describe the rich content of an image by using limited words.

To overcome the challenges encountered by conventional text-based image retrieval techniques, content-based image retrieval (CBIR) was first introduced in the early 1990s [1–3]. That is, instead of being manually annotated by the textual descriptions, query images will be indexed by their own visual content (e.g., color [10–12], texture [12–14], and shape [15–17]), which can be automatically extracted from images. Subsequently, many techniques in this research area have been developed and many CBIR systems, both from the academia and industry, have been widely designed. Below, we review some representative systems and show their distinct characteristics.

The query by image content (QBIC) system [18], which is the first commercial CBIR system developed by IBM, can support queries based on example images, sketches, drawings, and selected color and texture patterns. The framework and techniques of the QBIC system have profound effects on later CBIR systems. The Virage company developed a content-based image search engine called Virage [19], which can support arbitrary combinations of color, texture, and structure features as the visual queries. With Virage, the user can adjust the weights associated with the visual features to their own preferences. The Photobook system [20] designed by the MIT media laboratory is a set of interactive tools for browsing and searching images, and the system consists of three subbooks from which color, texture, and shape features are extracted, respectively. The user can then query images based on the corresponding features in each of the three subbooks. The multimedia analysis and retrieval system (MARS) [21] was developed at the University of Illinois at Urbana-Champaign, which is an interdisciplinary research effort involving multiple research communities: database management, computer vision, and information retrieval. Apart from these aforementioned systems, there are still many other famous systems for CBIR, which can be found in three excellent surveys on this topic [1–3].

Although this mechanism has established a general framework for image retrieval from a new perspective, there are still many open research issues to be solved before such systems can be applied to real-world applications. The early work in the CBIR community mainly focuses on building up fully automated systems and trying to find the best visual features to describe the content of images. However, the low-level visual

features extracted from images may not accurately characterize the high-level semantic concepts, since such computer vision techniques are not there yet [2]. Consequently, the gap between low-level visual features and high-level semantic concepts usually leads to poor performance for CBIR [1–3]. However, an image retrieval task is usually different from a conventional computer vision task since the user is always an indispensable part of the CBIR system. Therefore, an effective solution to bridge the semantic gap is trying to explore the synergy of the user and the system in an image retrieval procedure [1–3, 22].

Relevance feedback (RF), initially developed in the text-based information retrieval community [23], was introduced as a powerful tool to involve the user in the loop to enhance the performance of CBIR during the early and mid 1990s. The need for involving the user in the system stems from the fact that different semantic concepts live in different subspaces and each image can live in many different semantic concept subspaces [22]. Especially, different users at different times may have different interpretations for the same image, which thus makes an offline image retrieval task unfeasible. A straightforward way of involving the user in image retrieval is to ask the user to tune the parameters of the system to adapt to the user’s preferences. However, it is usually too much a burden for a common user. In general, a more feasible scheme of the interaction between the user and the system is to first let the user label a number of semantically relevant and irrelevant images in RF, which are the positive and negative feedback samples, respectively. And then, the CBIR system can refine all retrieved results based on these feedback samples provided by the user. The two steps are carried out iteratively to improve the performance of a CBIR system by gradually learning the user’s preferences.

Despite broad interests in constructing various RF methods based on different assumptions for the positive and negative feedback samples, it is boring and tedious for the user to be asked to label a large number of samples in RF. Given the difficulties in capturing the user’s preferences, multiple rounds of RF are actually required to achieve satisfactory results for a conventional image retrieval task, which will significantly limit the capability of RF for real-world applications. To reduce the labeling efforts of the user, a large number of studies have attempted to accelerate this procedure by leveraging various auxiliary information. In this thesis, the paradigm of leveraging the auxiliary information for image retrieval can be referred to as collaborative image retrieval (CIR).

Some of the research work in this category intends to select a set of the most informative samples in the database, which should be labeled by the user in RF and used as the training data to define an effective similarity metric for image retrieval. This is mainly because conventional RF methods can only select the relevant and irrelevant images in top returned results for the user to label, and this is not appropriate since top returned results may not be the most informative ones for image retrieval. On the other hand, from a long-term perspective, the RF log data can also be used as an important resource to enhance the performance of conventional CBIR systems. Recently, a number of studies have attempted to address the challenges encountered by conventional RF methods by resorting the RF log data [24–27]. In these studies, the system can accumulate the RF information provided by a number of users, which can be regarded as the RF log data in image retrieval. Therefore, besides the low-level visual features, each pair of images can also be associated with a set of similar and dissimilar pairwise constraints judged by users. During the past a few years, much research work has been carried out to exploit the RF log data to further enhance the performance of conventional CBIR systems [24–27].

1.1.2 Open Issues

CBIR has attracted much attention recently. However, the gap between low-level visual features and high-level semantic concepts usually leads to poor performance for CBIR. Although substantial research work has been widely conducted, CBIR is still an open research topic mainly due to the difficulties in bridging this semantic gap. In the following subsection, we will illustrate several challenging issues in the CBIR community as follows:

1.1.2.1 The Small-Sized Training Data Problem

In CBIR, the user is usually impatient and would not like to label a large number of samples at each round of RF. Hence, the number of training data will be much less than the dimension of the visual feature space. For such small-sized training data, most of the statistical machine learning techniques cannot obtain stable performance, and this will lead to the undersample problem encountered by most of the conventional RF methods for an image retrieval task. And thus, it is the main obstacle impeding the performance of conventional RF methods for CBIR.

1.1.2.2 The Asymmetric Property of Training Data Problem

In CBIR, most of the conventional RF methods treat the positive and negative feedback samples equally and ignore the asymmetric property of training data. However, this cannot be appropriate since all positive feedback samples share a common concept while each negative feedback sample differs in diverse concepts. The desired output of image retrieval is not necessarily a binary decision on each image as given by a traditional classifier. As a consequence, the unreasonable treatment on the two classes of feedback samples has severely degraded the performance of conventional RF methods for CBIR.

1.1.2.3 The Selection of Most Informative Samples Problem

In CBIR, most of the conventional RF methods can only select the relevant and irrelevant images in top returned results for the user to label. However, this cannot be appropriate since top returned results may not be the most informative ones to define an effective similarity metric for image retrieval. Worse of all, all top returned results labeled by the user may be relevant images, and thus standard classification methods cannot be applied due to the lack of irrelevant ones. Therefore, the key issue then becomes how to select the most informative samples in the database for the user to label and thus to further enhance the performance of image retrieval.

1.1.2.4 The Exploitation of RF Log Data Problem

Conventional CBIR systems cannot accumulate and utilize the RF information provided by a number of users, which can be regarded as the RF log data in image retrieval. From a long-term perspective, the RF log data is an important and useful resource to enhance the performance of the system. In the log database, besides the low-level visual features, each pair of images can also be associated with a set of similar and dissimilar pairwise constraints provided by users. As a consequence, it is more appropriate to design an effective scheme to fully exploit the RF log data with pairwise constraints, and then use the acquired information to further enhance the performance of image retrieval.

The aforementioned issues play a key role in the CBIR community. This thesis mainly focuses on exploring solutions to these issues in the CBIR research. There are also some other open issues that need to be investigated in the future, such as the imbalanced class distribution of the positive and negative feedback samples.

1.2 Statistical Machine Learning

Statistical machine learning theory was first introduced in the late 1960s. However, it is almost a purely theoretical analysis of the problem of function estimation from a given collection of data. During the last decade, there has been a surge of interest in applying statistical machine learning techniques to a variety of real-world applications, particularly in computer vision tasks [28–34]. Recently, we have witnessed various statistical machine learning techniques have already contributed or are still contributing to the computer vision community [32, 35–38]. Over the past decades, a large number of statistical machine learning techniques have been extensively studied in the machine learning community [39–45]; however, there are still few comprehensive work to investigate the effectiveness of these techniques conformably to maximize the potential of conventional RF methods for CBIR.

In general, statistical machine learning studies a variety of different classes of problems. In terms of different ways of studying the methodology, we can typically categorize the studies on statistical machine learning as unsupervised learning, supervised learning, weakly supervised learning, semi-supervised learning, and active learning, as well as others. Each of them has been separately studied during the past decades. Here, in this section, we briefly introduce these techniques in each category.

1.2.1 Unsupervised Learning

Unsupervised learning considers a learning task from a collection of data without label information. The techniques in this category aim to discover the intrinsic distribution or the geometric structure of data. One of the most popular study in this area is dimension reduction, which has been widely used in the computer vision community, such as face recognition [28, 29], gait recognition [46], and multimedia data management [32]. Many algorithms, such as principle component analysis (PCA) [39], Laplacian eigenmap (LE) [40], and locally linear embedding (LLE) [41], have excellent performance in a range of applications, especially in various computer vision tasks. Despite the fact that this problem has been widely studied for decades and a variety of advanced methods have been proposed, many challenging problems in unsupervised learning are still being actively studied in the current research.

1.2.2 Supervised Learning

Supervised learning can effectively exploit a collection of data with explicit class label information. These methods are powerful tools in data visualization and pattern classification. One of the most important supervised learning techniques is the Fisher’s linear discriminant analysis (LDA) [42], which has been widely employed in many fields such as biometrics [29, 47], bioinformatics [48], and multimedia information retrieval [30–32]. In a classification task, each input data instance is always associated with a corresponding training label, which can be considered as the response from a supervisor. A broad family of statistical machine learning theories have been widely studied to achieve the risk minimization and the generalization maximization in learning tasks. These theories have guided the generation of many new types of supervised learning methods for real-world applications. Among them, the support vector machine (SVM) [49, 50], has shown state of the art performance in a wide range of practical applications, especially in pattern classification [51–56].

1.2.3 Weakly Supervised Learning

Weakly supervised learning considers the problem of learning from a collection of data with a set of similar and dissimilar pairwise constraints. Most of the weakly supervised learning methods can only learn a Mahalanobis distance metric from the training data that are presented in the form of pairwise constraints (or side information) [43, 57], in which each pairwise constraint indicates whether the corresponding two samples are similar or dissimilar for a particular task. In general, selecting a good distance metric can significantly enhance the performance of a learning task. Therefore, learning appropriate distance metrics for various practical problems has been an open issue in recent research.

1.2.4 Semi-supervised Learning

Supervised learning methods cannot show good performance when there are insufficient labeled samples for a learning task. Semi-supervised learning under such a scenario is widely designed to significantly improve the generalization ability of supervised learning by leveraging abundant unlabeled samples in the database [44, 58–63]. The common

motivation of semi-supervised learning methods [44, 62, 64–68] is trying to exploit the intrinsic geometric structure of unlabeled samples by restricting the inductive prediction to comply with this geometry. The manifold regularization principle [44, 68], one of the most representative techniques in this category, assumes that the geometry of the intrinsic data probability distribution is supported on a low-dimensional manifold. The manifold approximation term and supervised learning models are combined together under the conventional regularization framework [69], which can smooth the model output along the manifold estimated from the unlabeled samples. Semi-supervised learning has been demonstrated to be a promising approach, affording improved performance compared with conventional supervised learning when only limited labeled samples are available.

1.2.5 Active Learning

Active learning is well-known for getting the necessary information by labeling as few samples as possible. For a conventional supervised learning task, the labels of training samples are usually expensive to obtain. In the machine learning community, there has been a long tradition of research on active learning [70–74]. In general, discriminative models aim to choose the most ambiguous samples to label [34], whereas generative models tend to select the most representative samples [73]. In statistics, active learning can be referred to as experimental design. The sample is referred to as experiment, and its label is referred to as measurement. The study of optimum experimental design (OED) [45] is concerned with the selection of most informative experiments to measure given that conducting an experiment is expensive. Classical OED methods include A-OED, D-OED, and E-OED, which are to maximize the confidence in a given model by minimizing some measure of the estimated parameter covariance [45]. Active learning methods have been widely employed as powerful tools to reduce the labeling efforts of conventional supervised learning in various learning tasks.

1.3 Thesis Organization and Contributions

RF is a powerful tool to involve the user in the system to narrow down the semantic gap between low-level visual features and high-level semantic concepts in the CBIR research. However, there are four types of problems in conventional RF methods: 1) the

Figure 1.1: The organization of the thesis.

small-sized training data problem, when the number of training data is much less than the dimension of the visual feature space; 2) the asymmetry property of training data problem, when most of the conventional RF methods treat the positive and negative feedback samples equally; 3) the selection of most informative samples problem, when most of the conventional RF methods can only select the relevant and irrelevant samples in top returned results; and 4) the exploitation of RF log data problem, when conventional CBIR systems cannot accumulate and utilize the RF information provided by a number of users. To alleviate the small-sized training data problem, we develop a generalized biased discriminant analysis (GBDA) method based on a biased discriminant analysis (BDA) model for CBIR in Chapter 3, as shown in Figure 1.1. To incorporate the asymmetric property of training data, we propose a biased maximum margin analysis (BMMA) method, which can be combined with the popular classification based RF, i.e., SVM-based RF, for CBIR. Semi-supervised BMMA (SemiBMMA) is also developed to utilize the information of unlabeled samples by introducing a Laplacian regularizer to

BMMA in Chapter 4, as shown in Figure 1.1. To select the most informative samples, we propose a geometric optimum experimental design (GOED) method to find multiple representative samples in the database as the most informative ones for the user to label, and this can alleviate the labeling efforts of conventional RF methods in Chapter 5, as shown in Figure 1.1. Moreover, conventional CBIR systems cannot accumulate and utilize the RF log data. In Chapter 6, we propose conjunctive patches subspace learning (CPSL) with side information to explicitly exploit the RF log data to improve the performance of CBIR.

In Chapter 2, we review the background of the research that are closely related to our work in this thesis, namely CBIR and statistical machine learning. Especially, we first classify the conventional RF methods into four categories, i.e., feature selection based methods, discriminant analysis based methods, classification based methods, as well as others. We analyze the problems encountered by these conventional RF methods and review the CIR study by leveraging the auxiliary information. Then, we also review the popular statistical machine learning techniques widely used in the computer vision community, i.e., unsupervised learning, supervised learning, weakly supervised learning, semi-supervised learning, and active learning. More reviews specific to the topic of each major chapter will be given at the start of the chapter.

In Chapter 3, we focus on the discriminant analysis based RF, i.e., BDA, because the need for RF schemes stems from the fact that different semantic concepts live in different subspaces and each image can live in many different semantic concept subspaces, and the selection of such a semantic concept subspace cannot be done offline. Subspace learning or discriminant analysis based RF has attracted much attention to bridge the semantic gap in the CBIR community recently. However, the user is usually impatient and would not like to label a large number of samples at each round of RF. Therefore, the number of training samples will be much less than the dimension of the visual feature space. The BDA model finds a subspace, which maximizes the ratio between the trace of the projected between-class scatter matrix and the trace of the projected positive within-class scatter matrix. The solution is given by an eigenvalue decomposition of the product of the inverse of the positive within-class scatter matrix and the between-class scatter matrix. In RF, BDA cannot be directly applied in this straightforward way, because

the small-sized training data problem will lead to a deficient positive within-class scatter matrix, i.e., the positive within-class scatter matrix is singular. In this chapter, we first review two aspects of the related discriminant analysis based RF, namely BDA and the differential scatter discriminant criterion (DSDC). We then analyze the problem of BDA and review some extensions related to this singular problem. The relationship between the Fisher discriminant criterion (FDC) and DSDC is discussed in detail. After that, we develop a new discriminant analysis based RF method, i.e., generalized BDA (GBDA), for CBIR. The proposed GBDA method defines the separation of different classes as a trace difference rather than a trace ratio, and this can avoid the singular problem in the original BDA model. Moreover, to avoid the Gaussian distribution assumption for the positive feedback samples, GBDA defines the between-class scatter matrix by resorting to interclass nearest neighborhood samples, thereby extracting the most discriminative information. Finally, to alleviate the overfitting problem, GBDA integrates the locality preserving principle; therefore a smooth and locally consistent transformation can also be learned. Extensive experiments on a real-world image database have shown that the proposed GBDA can significantly improve the performance of BDA and its related extensions.

In Chapter 4, we mainly focus on the classification based RF, i.e., SVM-based RF, because regarding the positive and negative feedback samples as two different groups and aiming to find a classifier to identify these two classes from each other, RF can be considered as an online binary classification problem. Hence, classification based RF has become a popular topic in the CBIR community during the last decade. However, conventional classification based RF, e.g., SVM-based RF, can only treat the positive and negative feedback samples equally although this is not appropriate since all positive feedback samples share a common concept while each negative feedback sample differs in diverse concepts. In this chapter, we first review the popular classification based RF, i.e., SVM-based RF and the graph embedding framework. These are the fundamental materials of this chapter. With an observation that different semantic concepts live in different subspaces and each image can live in many different semantic concept subspaces, we then develop a BMMA method to separate the positive feedback samples from the negative ones by a maximum margin in the reduced subspace. The traditional SVM-based RF combined with BMMA can better model the RF procedure and reduce the

performance degradation caused by the asymmetric property of training data. Moreover, SemiBMMA can effectively integrate the information of unlabeled samples by introducing a Laplacian regularizer to BMMA, and thus alleviates the overfitting problem caused by a small number of training feedback samples. To show the effectiveness of the proposed scheme combined with SVM-based RF, we compare it with the traditional SVM RF and some other related existing RF methods on a real-world image database. Experimental results show that the proposed scheme combined with SVM-based RF can significantly improve the performance of CBIR.

In Chapter 5, a GOED method is developed to select representative samples in the database as the most informative ones for the user to label. We first review two areas of the research that are closely related to our work in this chapter, i.e., 1) selecting most informative samples and 2) OED. Then, we formulate the traditional RF in CBIR as an active learning problem, and propose a novel active learning framework by leveraging the geometric structure of unlabeled samples in the reproducing kernel Hilbert space (RKHS) to select a set of the most informative samples for the user to label. Based on this active learning framework and a transductive experiment design (TED) method in OED, we present a novel active learning method, i.e., GOED, by leveraging the geometric structure of unlabeled samples in RKHS to simultaneously select multiple representative samples in the database as the most informative ones for the user to label in image retrieval. The proposed GOED can alleviate the small-sized training data problem by leveraging the geometric structure of unlabeled samples in RKHS, and thus further enhances the performance of image retrieval. By minimizing the expected average prediction variance on the test data, GOED has a clear geometric interpretation to select the most representative samples in the database iteratively with the global optimum, which is more effective and efficient for the user to label. Extensive experiments on both synthetic datasets and a real-world image database have shown the advantages of the proposed GOED for CIR.

In Chapter 6, CPSL with side information is presented to effectively exploit the RF log data. We first review two areas of the research that are closely related to our work in this chapter, i.e., 1) exploiting RF log data and 2) subspace learning and distance metric learning. Then, we propose a regularized weakly supervised subspace learning framework when the training samples are associated with only a set of similar and dissimilar pairwise constraints. Based on this framework, we propose the CPSL method to

learn such a semantic concept subspace by exploiting the RF log data. Especially, CPSL can effectively integrate the discriminative information of labeled images, the geometric information of labeled images, and the weakly similar information of unlabeled images. This process is conducted by building different kinds of local patches for each image, and then aligning those different kinds of patches together to learn a consistent coordinate through a regularized learning framework. We formally formulate this method as a constraint optimization problem, and then present an effective algorithm to solve this task with closed-form solutions. Extensive experiments on both synthetic datasets and a real-world image database demonstrate the effectiveness of the proposed scheme in improving the performance of CBIR by exploiting the RF log data.

Chapter 7 concludes.

The main contributions of this thesis are as follows.

1) A generalized biased discriminant analysis (GBDA) method has been developed as a principled way to deal with the small-sized training data problem in the conventional discriminant analysis based RF, i.e., BDA, for CBIR. Unlike BDA, GBDA can avoid the singular problem encountered by BDA by adopting the differential scatter discriminant criterion (DSDC). Moreover, by redesigning the between-class scatter matrix and integrating the locality preserving principle, we know that the proposed method can also alleviate the Gaussian distribution assumption for the positive feedback samples and the overfitting problem in BDA. Extensive experiments have shown that the proposed GBDA can significantly outperform the conventional discriminant analysis based RF, i.e., BDA, and its enhanced versions.

2) A biased maximum margin analysis (BMMA) method has been developed to effectively incorporate the asymmetry property of training data with the conventional classification based RF, i.e., SVM-based RF, for CBIR. By introducing a Laplacian regularizer to BMMA, we also design a semi-supervised BMMA (SemiBMMA) method to utilize the information of unlabeled samples for SVM-based RF. The traditional SVM-based RF combined with BMMA can better model the RF procedure and reduce the performance degradation caused by the asymmetric property of training data. SemiBMMA can integrate the information of unlabeled samples into SVM-based RF and effectively alleviates the overfitting problem caused by the small number of training feedback samples. Extensive experiments on a real-world image database have shown that the proposed scheme

combined with the traditional SVM-based RF can significantly improve the performance of CBIR.

3) A geometric optimum experimental design (GOED) method has been developed to select multiple representative samples in the database as the most informative ones for the user to label. Especially, GOED can alleviate the small-sized training data problem by leveraging the geometric structure of unlabeled samples in the reproducing kernel Hilbert space (RKHS), and thus can further enhance the performance of image retrieval. Moreover, by minimizing the expected average prediction variance on the test data, GOED has a clear geometric interpretation to select a set of the most representative samples in the database iteratively with the global optimum. Compared with the popular SVM active learning (SVMactive), GOED is label-independent and can effectively avoid various potential problems caused by insufficient and inexactlly labeled samples in RF, which is more appropriate and useful for image retrieval. Extensive experiments on both synthetic datasets and a real-world image database have been conducted to show the advantages of the proposed GOED for CIR.

4) Conjunctive patches subspace learning (CPSL) with side information has been developed to explicitly exploit the RF log data. The significance of this method is that it can directly learn a semantic concept subspace from a set of similar and dissimilar pairwise constraints without using any class label information, which is more practical and useful in real-world applications, since explicit class label information for each image might be too expensive to obtain in many real-world applications. The proposed CPSL method can effectively learn a reliable subspace both from the labeled and unlabeled images through a regularized learning framework by exploiting the RF log data. Especially, we first formulate this problem as a constraint optimization problem, and then present an efficient algorithm to solve this task with closed-form solutions. Compared with previous distance metric learning methods, which usually involve a convex optimization procedure or a semidefinite programming procedure, our method can also learn a distance metric but performs more effectively and efficiently when dealing with high-dimensional data.

Chapter 2

Background Review

2.1 Content-Based Image Retrieval

Content-based image retrieval (CBIR) has grown tremendously as an emerging technology recently. However, the gap between low-level visual features and high-level semantic concepts usually leads to poor performance for CBIR. Relevance feedback (RF) is one of the most powerful tools to narrow down the semantic gap by letting the user label semantically relevant and irrelevant images with the query image, which are the positive and negative feedback samples, respectively. Despite the success, it is not appropriate to require the user to label a large number of samples in RF. To reduce the labeling efforts of conventional RF methods, collaborative image retrieval (CIR) has attempted to accelerate this procedure by resorting to the most informative samples in the database or the RF log data. To describe our work clearly, in this section, let us briefly review two areas of the research regarding the conventional RF methods and the CIR study.

2.1.1 Conventional RF Methods

RF is an important tool to involve the user in the system to narrow down the semantic gap in the CBIR research. In an RF procedure, the user first labels a number of relevant images as the positive feedback samples and some irrelevant images as the negative feedback samples. Then, the CBIR system refines the retrieved results based on these feedback samples. The two steps are carried out iteratively to improve the performance of CBIR by gradually learning the user's preferences. During the last decade, various RF methods have been constructed based on different assumptions for the positive and

negative feedback samples. In view of the characteristics of the studies in this category, we briefly classify these conventional RF methods into four categories.

Feature selection based methods adjust weights associated with various dimensions of the visual feature space to adapt to the user’s preferences [75–79]. These methods can enhance the importance of those dimensions that help in retrieving similar images and reduce the importance of those dimensions that hinder the performance of CBIR. Alternatively, features can also be selected by boosting techniques in which a strong classifier is obtained as a weighted sum of weak classifiers along different feature dimensions [80]. Some early approaches, which include the well-known multimedia analysis and retrieval system (MARS) [76], MindReader [77], and query re-weighting approaches [2], can be categorized into this group.

Discriminant analysis based methods either find a low-dimensional subspace of the visual feature space, such that the positive and negative feedback samples are well separated after projecting onto this subspace [81–85] or define a $(1+x)$ -class discriminant analysis problem and find a subspace within which to separate the one positive class from an unknown number of negative classes [32, 86–88]. However, the small-sized training data problem, i.e., the number of training samples is much smaller than the dimension of the visual feature space, will lead to the singular problem of the positive within-class scatter matrix and significantly impedes the performance of conventional discriminant analysis based RF, e.g., biased discriminant analysis (BDA), for CBIR. To alleviate the singular problem, a direct method [86] or a full space method [89] can be utilized to enhance the performance of the discriminant analysis based RF. Furthermore, to speed up the RF procedure, incremental learning can also be employed [86].

Classification based methods regard the positive and negative feedback samples as two different classes and aim to find a classifier to identify the two classes from each other. Many binary classification based methods were designed to treat the positive and negative feedback samples equally [35–37, 90]. One-class support vector machine (SVM) can estimate the density of the positive feedback examples [91] but ignores the information contained in the negative feedback samples. Derived from the one-class SVM, a biased SVM inherits the merits of one-class SVM but incorporates the negative feedback samples [92]. Random sampling techniques were applied to alleviate the unstable, biased,

and overfitting problems in the traditional SVM-based RF for CBIR [36]. Li et al. proposed a multitasking SVM method [90] by adapting a cotraining technique and a random sampling method to incorporate the information of unlabeled samples. In [93], Tao et al. made an assumption that positive feedback samples were included in a set, negative feedback samples split into a small number of subsets, and a series of kernel marginal convex machines had been developed between the one positive group and several negative subgroups.

Besides these aforementioned methods, there are still some **other techniques** to construct RF for CBIR, such as self-organizing maps [94], EM algorithms [95], and Gaussian mixture models [96]. Due to the limited space, we are unable to enumerate all existing approaches, and more conventional RF methods can be found in an excellent survey [22].

2.1.2 Collaborative Image Retrieval

Conventional RF methods cannot select the most informative samples in the database for the user to label. SVM active learning (SVMactive) can select ambiguous samples as the most informative ones for the user to label with the help of the optimal hyperplane of SVM, and thus alleviates the labeling efforts of the user. However, the optimal hyperplane of SVM is usually unstable and inaccurate with small-sized training data, which is always the case in image retrieval since the user would not like to label a large number of samples and cannot label each sample accurately all the time. To alleviate the insufficient training data problem, Wang et al. proposed to modify SVMactive with a transductive SVM by engaging unlabeled samples in the database [38]. In [97], Hoi et al. combined some semi-supervised learning techniques (i.e., Gaussian fields and harmonic functions [62]) with the traditional SVMactive, which can also effectively exploit the information of unlabeled samples. To find multiple informative samples for the user to label in each iteration, Dagli et al. proposed to use a diversity measure based information theory approach to select a set of diverse and informative samples simulatively [98]. Goh et al. adopted an active learning method by incorporating the angular diversity measure for an image retrieval task [99]. Moreover, in [27], a semi-supervised SVM batch mode active learning method was also developed to address the small-sized training data problem by leveraging the geometric structure of unlabeled samples in the reproducing kernel Hilbert

space (RKHS) and to identify the most informative samples via a min-max framework. Despite the vast research work during the past a few years, SVMactive methods always require an initial optimal hyperplane to identify the most informative samples for the user to label. However, this optimal hyperplane will not always be accurate with insufficient and inexactlabeled feedback samples, and this is always the case in image retrieval.

Despite broad interests in constructing RF approaches, an online learning task can be tedious and boring for the user. Recently, a number of studies have attempted to address the challenges encountered by conventional RF methods by resorting to the RF log data [24–27]. In these studies, the system can accumulate the RF information provided by a number of users, which can be regarded as the RF log data. Therefore, besides the low-level visual features, each pair of images can also be associated with a set of similar and dissimilar pairwise constraints judged by users. During the past several years, much research work has been carried out based on this new paradigm of image retrieval. In [24, 25], manifold learning algorithms were applied to learn an exquisite manifold structure from the log data, which could better reflect the semantic relation among different images. In [100], Muller et al. suggested a weighting scheme by exploiting the RF log data for CBIR. In [26], Hoi et al. proposed a log based RF technique with SVM by engaging the RF log data in a regular online RF task. In [27], the authors proposed a distance metric learning technique by exploiting the RF log data with pairwise constraints and showed the effectiveness of the proposed scheme by comparing it with some representative distance metric learning techniques for image retrieval.

2.2 Statistical Machine Learning

Statistical machine learning techniques have already contributed and are still contributing to the computer vision community [28–34, 39–41, 47]. Each of them has been separately studied in the machine learning community for a long period. In this section, we briefly review these statistical machine learning techniques, which have been widely applied to various computer vision tasks.

2.2.1 Unsupervised Learning

In many practical applications, it may be expensive to assign labels to samples. In this situation, unsupervised learning methods are often used to discover the unknown knowledge from a large collection of unlabeled samples. A well-known methodology among various unsupervised learning methods is dimension reduction. In this subsection, let us briefly review several representative methods for dimension reduction.

2.2.1.1 Principle Component Analysis (PCA)

Although PCA [39] is a reconstructive dimension reduction model, it has been successfully applied to various classification tasks in the computer vision community, such as biometrics [29, 101–103] and multimedia data management [104, 105]. This method extracts the principal eigenspace associated with a number of training samples $x_i \in R^h (1 \leq i \leq n)$. Let $S = (1/n) \sum_{i=1}^n (x_i - m)(x_i - m)^T$ be the covariance matrix or the total-class scatter matrix of all training samples $x_i (1 \leq i \leq n)$, where $m = (1/n) \sum_{i=1}^n x_i$ is the mean value of n training samples. One solves the eigenvalue equation $\lambda u_i = S u_i$ for eigenvalues $\lambda_i \geq 0$. The projection matrix U^* is spanned by the first l eigenvectors with the largest eigenvalues, i.e., $U^* = [u_i^*]_{i=1}^l$. If x is a new sample, then it is projected to $y = (U^*)^T(x - m)$. The vector y is used in place of x for data representation and classification.

2.2.1.2 Laplacian Eigenmap (LE)

LE [40] was proposed based on the spectral graph theory [106]. It is an unsupervised nonlinear dimension reduction method and aims to preserve the neighbor structure of data, which has been widely used for data clustering [40], face recognition [28], and image retrieval [24].

Given a k nearest neighbor graph G with the weight matrix W , which is defined as

$$W_{ij} = \begin{cases} 1, & \text{if } x_i \in N(x_j) \text{ or } x_j \in N(x_i), \\ 0, & \text{otherwise,} \end{cases} \quad (2.1)$$

where $N(x_i)$ denotes the set of k nearest neighbor samples of x_i .

The optimal embedding result can be obtained by solving the following minimization problem, i.e.,

$$\min_y \sum_{i,j=1}^n (y_i - y_j)^2 W_{ij} = \min_y 2\text{tr}(y^T L y), \quad (2.2)$$

where $y = [y_1, y_2, \dots, y_n]$; $L = D - W$ is the graph Laplacian and $D_{ii} = \sum_{j=1}^n W_{ij}$. The objective function with the choice of weight matrix W_{ij} incurs a heavy penalty if nearest samples x_i and x_j are embedded far apart. Therefore, minimizing it is an attempt to ensure that if x_i and x_j are close then y_i and y_j are close as well.

The optimal embedding result is given by the eigenvector corresponding to the smallest eigenvalue of the following generalized eigenvalue decomposition problem, i.e.,

$$L y = \lambda D y. \quad (2.3)$$

2.2.1.3 Locally Linear Embedding (LLE)

LLE [41] assumes the data reside on a nonlinear submanifold and each of the local neighborhood is linear. The local geometry of the local neighborhood can be characterized by the linear coefficients, which can be used to reconstruct each sample from its nearest neighbor samples.

Given a k nearest neighbor graph G with a weight matrix W , the reconstruction errors can be measured by the cost function

$$\begin{aligned} \psi(W) &= \sum_{i=1}^n \left\| x_i - \sum_{j=1}^k W_{ij} x_j \right\|^2 \\ \text{s.t. } &\sum_{j=1}^k W_{ij} = 1, \end{aligned} \quad (2.4)$$

which adds up the squared distances between all samples and their corresponding reconstruction results. It should be noted that W_{ij} will vanish for samples with long distances. Considering the problem of mapping the original samples to a line, each sample on the line can be represented as a linear combination of its nearest samples with coefficients

W_{ij} . Let $y = (y_1, y_2, \dots, y_n)^T$ be such a map. A reasonable criterion for choosing a good embedding result is to minimize the following loss function, i.e.,

$$\psi(y) = \sum_{i=1}^n \|y_i - \sum_{j=1}^k W_{ij} y_j\|^2. \quad (2.5)$$

The loss function is based on locally linear reconstruction errors, but here we fix the weights W_{ij} as the coefficients calculated in Eq.(2.4) while optimizing the embedding y_i . It can be shown that the optimal embedding result is given by the eigenvector corresponding to the smallest eigenvalue of the following eigenvalue problem, i.e.,

$$(I - W)^T(I - W)y = \lambda y, \quad (2.6)$$

where I is an $n \times n$ identity matrix.

Define the matrix $W_{LLE} = W + W^T - W^T W$, we can rewrite the eigenvalue problem as

$$(I - W_{LLE})y = \lambda y \Rightarrow W_{LLE}y = (1 - \lambda)y. \quad (2.7)$$

Thus, the optimal embedding result is given by the eigenvector corresponding to the largest eigenvalue of the eigenvalue problem as

$$W_{LLE}y = \lambda y. \quad (2.8)$$

2.2.2 Supervised Learning

Supervised learning considers the problem of estimating the responses from a set of training samples with explicit class label information. They are widely used for various classification tasks in real-world applications. In this subsection, let us briefly review two representative supervised learning methods, i.e., linear discriminant analysis (LDA) and SVM, which are widely used in the computer vision community.

2.2.2.1 LDA

LDA is a typical discriminative dimension reduction model, which has been widely used in real-world applications, such as biometrics [29, 47], bioinformatics [48], and multimedia information retrieval [30–32].

LDA finds in the high-dimensional feature space a low-dimensional subspace, in which different classes of samples are well separated after projecting onto this subspace. The subspace is spanned by a set of vectors, which are denoted as $U = [u_1, \dots, u_l]$. It is assumed that a set of training samples are available. The training set are divided into c classes. The i th class contains n_i samples $x_{i,j}$ ($1 \leq j \leq n_i$) and has an expected mean of $m_i = (1/n_i) \sum_{j=1}^{n_i} x_{i,j}$. The between-class scatter matrix S_b and the within-class scatter matrix S_w are defined by

$$\begin{cases} S_b = \frac{1}{n} \sum_{i=1}^c n_i (m_i - m)(m_i - m)^T, \\ S_w = \frac{1}{n} \sum_{i=1}^c \sum_{j=1}^{n_i} (x_{i,j} - m_i)(x_{i,j} - m_i)^T, \end{cases} \quad (2.9)$$

where $n = \sum_{i=1}^c n_i$ is the size of training samples; and $m = (1/n) \sum_{i=1}^c \sum_{j=1}^{n_i} x_{i,j}$ is the expected mean of all training samples. The projection matrix U^* of LDA is defined by

$$U^* = \arg \max_U \text{tr}((U^T S_w U)^{-1} U^T S_b U). \quad (2.10)$$

The projection matrix U^* is computed from the largest l eigenvectors of $S_w^{-1} S_b$, under the assumption that S_w is invertible. If c is equal to 2, LDA reduces to the Fisher's LDA [42]; otherwise, LDA is known as the Rao's multiple discriminant analysis [107].

LDA cannot fully utilize the discriminative information contained in the covariances of different classes; it models each class with a single Gaussian density; and it fairly considers all Kullback-Leibler (KL) divergences between different classes. Therefore, LDA has the corresponding three problems, i.e., 1) heteroscedastic problem, 2) multimodal problem, and 3) class separation problem, which have been identified in previous research [108]. Moreover, in practical applications, e.g., biometric research and multimedia management, LDA encounters the small-sized training data problem or the singular problem, because the number of training data is much less than the dimension of the feature space [109].

2.2.2.2 SVM

SVM enjoys solid theoretical foundations [49, 50], which has been widely applied to and demonstrated outstanding performance in many real-world applications, such as handwriting recognition [33, 110], objection recognition [51–53], face detection [111–113], and text categorization [54, 55].

In theory, SVM can be interpreted from the statistical regularization framework [69]. Especially, SVM can be formulated as a regularized learning problem in the reproducing kernel Hilbert space (RKHS), i.e.,

$$f = \arg \min_{f \in H_K} \frac{1}{n} \sum_{i=1}^n (1 - y_i f(x_i))_+ + \gamma_1 \|f\|_K^2, \quad (2.11)$$

where n is the number of training samples; $(\cdot)_+$ is the hinge loss in which $(a)_+ = a$ if a is positive and zero otherwise; y_i is the label of the training sample x_i ; $\gamma_1 > 0$; and $\|f\|_K$ is an induced norm in an appropriately chosen RKHS H_K defined over a kernel Gram matrix K [49].

In practice, SVM is usually formulated as another formula as

$$\begin{aligned} \min_{w, \xi, b} & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i \\ \text{s.t.} & y_i (w^T \phi(x_i) - b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, i = 1, 2, \dots, n, \end{aligned} \quad (2.12)$$

where C is a penalty parameter of the error term ξ_i ; $\phi(\cdot)$ is a kernel mapping function; and the labels y_i are either +1 or -1 for a conventional binary classification problem.

The solution to this optimization problem can be found by introducing a Lagrangian. It then can be reformulated as a dual form as a quadratic programming problem, i.e.,

$$\begin{aligned} \max_a & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j (\phi(x_i) \cdot \phi(x_j)) \\ \text{s.t.} & \sum_{i=1}^n \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C, i = 1, 2, \dots, n. \end{aligned} \quad (2.13)$$

This optimization problem is a typical quadratic programming problem, which can be solved effectively by a standard quadratic programming technique or by some other

available advanced methods, e.g., sequential minimal optimization (SMO) techniques [114].

2.2.3 Weakly Supervised Learning

Weakly supervised learning considers the problem of learning from a collection of data with a set of similar and dissimilar pairwise constraints. In [115], Wagstaff et al. suggested a k-means clustering algorithm by introducing the pairwise constraints. In [43], Xing et al. studied the problem of learning an optimal Mahalanobis distance metric from the contextual information in terms of pairwise constraints (or side information). Discriminative component analysis (DCA) was proposed to incorporate the dissimilar pairwise constraints [116], which can show slightly better performance compared with RCA for some datasets. Lately, an information-theoretic metric learning approach was proposed to express the weakly supervised learning problem as a Bregman optimization problem [117]. To effectively exploit the unlabeled samples, Hoi et al. proposed a Laplacian regularized metric learning approach [118]. Moreover, Wu et al. proposed to learn a Bergman distance function with pairwise constraints and showed that the approach can learn nonlinear distance functions for a semi-supervised clustering task [119]. Bar-Hillel et al. proposed a relevant component analysis (RCA) [57] method to exploit only similar pairwise constraints for a distance metric learning task, which is simple to calculate but enjoys comparable performance with that of [43]. Due to the popularity of the weakly supervised distance metric learning methods, let us briefly review these two techniques in this subsection.

The problem of distance metric learning is to find an optimal Mahalanobis distance metric that is used to measure the distance between two data instances as $d_M(x_i, x_j) = \sqrt{(x_i - x_j)^T M (x_i - x_j)}$, where M must be positive semi-definite to satisfy the properties of a metric, i.e., non-negativity and triangle inequality. The goal of weakly supervised distance metric learning is to find an optimal Mahalanobis distance metric M by using the contextual information in terms of pairwise constraints. Given the pairwise constraints S and D as

$$\begin{aligned} S &= \{(i, j) \mid x_i \text{ and } x_j \text{ are judged to be similar}\}, \\ D &= \{(i, j) \mid x_i \text{ and } x_j \text{ are judged to be dissimilar}\}, \end{aligned} \tag{2.14}$$

Xing et al. formulated the problem of weakly supervised distance metric learning into the following convex programming problem [43] as

$$\begin{aligned} \min_M \quad & \sum_{(x_i, x_j) \in S} \|x_i - x_j\|_M^2 \\ \text{s.t.} \quad & \sum_{(x_i, x_j) \in D} \|x_i - x_j\|_M > 1. \end{aligned} \quad (2.15)$$

In this equation, the optimal distance metric M is found by minimizing the sum of squared distances between pairs of samples with similar constraints S , and meanwhile satisfying the constraint that the sum of distances between pairs of samples with dissimilar constraints D is large than 1. In other words, this method tries to minimize the distances between similar samples and maximize the distances between dissimilar samples simultaneously.

RCA is trying to identify and scale down global unwanted variabilities within the data [57]. RCA changes the feature used for data representations via a global linear transformation, in which relevant dimensions are assigned with large weights. The relevant dimensions are estimated by chunklets, each of which is defined as a group of data instances linked together with similar pairwise constraints. More specifically, given a data set $X = \{x_i\}_{i=1}^n$ and c chunklets $\hat{C}_j = \{x_{i,j}\}_{i=1}^{n_j}$, RCA calculates the following matrix, i.e.,

$$\hat{C} = \frac{1}{n} \sum_{j=1}^c \sum_{i=1}^{n_j} (x_{i,j} - m_j)(x_{i,j} - m_j)^T, \quad (2.16)$$

where m_j denotes the mean of the i th chunklet; $x_{i,j}$ denotes the i th data instance in the j th chunklet; and n is the number of data instances. The optimal Mahalanobis distance metric is equal to the inverse of the matrix $\hat{C}$, i.e., $M = \hat{C}^{-1}$. Compared with [43], RCA enjoys the merits of simple implementation and good computational efficiency.

2.2.4 Semi-supervised Learning

Semi-supervised learning considers the problem of learning from a small number of labeled samples and a large number of unlabeled samples. Recently, a number of semi-supervised learning methods have been shown to be superior to classical supervised learning methods. In general, we briefly classify these methods into two categories. The first group of

methods assumes that the samples in the same cluster will share similar labels. Based on this assumption, Nigam et al. applied the EM algorithm on a mixture of multinomials for text classification, and showed that the accuracy of learned text classifiers can be improved by augmenting a small number of labeled samples with a large number of unlabeled samples [120]. The transductive SVM was proposed to maximize the margin in the presence of unlabeled samples and learn a decision boundary that traverses through low data-density regions [60]. Self-training [59] and co-training [58] iteratively label some unlabeled samples according to the predictions of the current classifier, and retrain a new classifier with the additional labeled samples. The second group of methods is graph-based [61–63]. The manifold regularization principle [44], one of the most representative techniques, assumes that the geometry of the intrinsic data probability distribution is supported on a low-dimensional manifold. There are many different ways to approximate the geometric structure of unlabeled samples [40, 121–123]. A popular way is to compute the graph Laplacian in an unsupervised manner from the unlabeled samples by using LE in the feature space [40, 123]. The manifold approximation term and supervised learning models are combined together under the conventional regularization framework [69], which can smoothen the model output along the manifold estimated from the unlabeled samples [44, 68]. In [68], Geng et al. proposed an ensemble manifold regularization framework for automatically and implicitly estimating the hyperparameters of manifold regularization. By providing some initial guesses of manifolds, the ensemble manifold regularization can learn to combine them for a conditionally optimal estimation of the intrinsic manifold. More semi-supervised learning methods can be found in a comprehensive survey [124].

2.2.5 Active Learning

Active learning aims to select the most informative unlabeled samples for labeling. Given a set of unlabeled samples $X = \{x_1, \dots, x_n\}$ in R^h , these methods in this category will find a subset $Z = \{z_1, \dots, z_k\} \subseteq X$ which contains the most informative samples. That is, if the samples $z_i (i = 1, \dots, k)$ are labeled and used as the training data, we can predict the labels of unlabeled samples more precisely. Most of these active learning methods are conducted iteratively to select the samples with highest classification uncertainty

for manually labeling. One of the most popular active learning methods is SVMactive developed by Tong and Koller [34], which measures the classification uncertainty of an unlabeled sample according to how far the example is away from the optimal hyperplane of SVM, which is widely used in text categorization [56] and image retrieval [34].

In statistics, active learning can be referred to as optimum experimental design (OED) [45]. In general, we consider a linear regression model, i.e.,

$$y = w^T x + \varepsilon, \quad (2.17)$$

where $w \in R^h$ is the parameter vector; y is the real-valued output; and ε is the measurement noise with zero mean and constant variance σ^2 . OED attempts to select the most informative samples to learn a prediction function $f(x) = w^T x$ so that the expected prediction variance can be minimized. Given a set of measured samples $(z_1, y_1), \dots, (z_k, y_k)$, the most popular method is the least squares method, in which we minimize the sum of squared errors, i.e.,

$$J(w) = \sum_{i=1}^k (y_i - f(z_i))^2. \quad (2.18)$$

Let $Z = [z_1, \dots, z_k]^T$ and $y = [y_1, \dots, y_k]^T$. The optimal solution to Eq.(2.18) is calculated as

$$\hat{w} = (ZZ^T)^{-1}ZY. \quad (2.19)$$

It can be proved that $\hat{w}$ is an unbiased estimation of w [125] and its covariance can be expressed as

$$C_w = \sigma^2(ZZ^T)^{-1}. \quad (2.20)$$

Conventional OED methods can select the most informative samples, i.e., $z_1, \dots, z_k$, by minimizing the size of the estimation covariance, i.e., Eq.(2.20). Three widely used criteria are trace of C_w (i.e., A-OED), determinant of C_w (i.e., D-OED), and maximum eigenvalue of C_w (i.e., E-OED).

Chapter 3

Generalized Biased Discriminant Analysis

3.1 Motivation

Relevance feedback (RF) [22] is a powerful tool to bridge the gap between low-level visual features and high-level semantic concepts [1–3] in content-based image retrieval (CBIR) research. In general, RF focuses on the interactions between the user and the system by letting the user label some relevant and irrelevant images with the query image, which are the positive and negative feedback samples, respectively. However, RF is different from conventional machine learning problems because the user would not like to label a large number of feedback samples [1–3].

During the last decade, various RF methods have been developed based on different assumptions for the positive and negative feedback samples. One-class support vector machine (SVM) estimates the density of positive feedback samples but ignores the negative ones [91]. Two-class SVM can identify the positive and negative feedback samples from each other but treats the two classes equally [35]. By precisely parameterizing the positive feedback samples, negative feedback samples, and unlabeled samples, a manifold learning method was proposed to find the intrinsic coordinate of the image low-level visual features [88]. In [93], a series of kernel marginal convex machines were developed between the one positive class and several negative subclasses. The results indicated that clustering the negative feedback samples into several subclasses can indeed improve the overall performance of CBIR.

Despite broad interests in constructing various RF methods [35, 88, 91, 93], the need for RF actually stems from the fact that different semantic concepts live in different subspaces and each image can live in many different semantic concept subspaces, and the selection of such a semantic concept subspace cannot be done offline [32]. Discriminant analysis is one of the most popular small sample learning methods for subspace selection, and has been widely used in recent years and achieved state of the art performance in various real-world applications [29, 79, 126–131]. During the last decade, Fisher’s linear discriminant analysis (LDA) has been successfully used in the computer vision community [29, 79, 127–129]. LDA was first used to select the most discriminative feature subset in CBIR. And then, the remaining images in the database were projected onto this subspace and some similarity measure was used to sort the images. However, with LDA, all negative feedback samples are deemed equivalent, and this is a severe limitation of this method because all positive feedback samples are alike and each negative feedback sample is negative in its own way. With this observation, biased discriminant analysis (BDA) was developed to deal with the RF scheme in CBIR and obtained a more satisfactory result [32]. In the BDA model, all negative feedback samples are required to stay away from the centroid of the positive feedback samples and all positive feedback samples are clustered together. The BDA model provides a good solution to the RF scheme, because there is an unknown number of classes in CBIR but the user is only interested in one class [32, 86]. BDA finds a subspace, which maximizes the ratio between the trace of the projected between-class scatter matrix and the trace of the projected positive within-class scatter matrix. The solution is given by an eigenvalue decomposition of the product of the inverse of the positive within-class scatter matrix and the between-class scatter matrix.

Just like LDA, BDA always suffers from the small-sized training data problem or the singular problem [126], because the rank of the positive within-class scatter matrix is deficient, i.e., the inverse of the positive within-class scatter matrix does not exist. And this is always the case in RF for CBIR, since it is not reasonable to require the user to label a large number of feedback samples. Traditionally, the singular problem is solved by the regularization method [32, 132], although this is not a good choice as pointed out by previous work on face recognition [79, 131, 133]. Additionally, the BDA model makes a strong assumption that all positive feedback samples form a single Gaussian distribution

[32], and this is always not true in real-world applications. These are the main obstacles impeding the performance of BDA for CBIR. Various research efforts have shown that the high-dimensional samples possible reside on or close to a nonlinear manifold of the ambient space [41, 134, 135].

Yu et al. proposed a dimension reduction method based on the hybrid analysis of principal component analysis and LDA, which can better integrate the descriptive and discriminative information for a specific data distribution [130]. However, using this method, it is necessary to find the best parameter pair setting for a special data distribution. To alleviate the singular problem and the Gaussian distribution assumption for the positive feedback samples, direct BDA (DBDA) and direct kernel BDA (DKBDA) [86] were proposed to enhance the performance of BDA by utilizing the direct idea [79] and the kernel trick. However, this approach still discards the null space of the between-class scatter matrix, which still contains the important discriminative information, as pointed out in literatures [129, 131]. Additionally, the kernel parameter tuning makes the online learning infeasible. As a variant of marginal Fisher analysis [136], margin biased analysis (MBA) was introduced to construct the RF scheme and had shown better performance than BDA [87]; however, it still suffers from the intrinsic singular problem as in BDA.

In this chapter, we propose a generalized BDA (GBDA) method for CBIR. To avoid the singular problem in BDA, GBDA is designed based on the differential scatter discriminant criterion (DSDC) [46, 126, 137–140], which defines the interclass separability as a trace difference for the between-class scatter matrix and the within-class scatter matrix rather than a trace ratio. Furthermore, to avoid the Gaussian distribution assumption for the positive feedback samples, the between-class scatter is specially designed by resorting to nearest neighbor samples in different classes. Additionally, to reduce the overfitting problem, the locality preserving principle emerging from the manifold learning community [41, 134, 135], which measures the local smoothness of the feature transformation, is integrated to regularize the between-class separability term. Therefore, a locally smooth and consistent transformation can also be learned for CBIR.

The rest of this chapter is organized as follows. In Section 3.2, BDA and DSDC are briefly reviewed. In Section 3.3, the GBDA method is introduced for CBIR. A large number of experiments on a real-world image database that validate the effectiveness of GBDA are given in Section 3.4. Section 3.5 gives the summary of this chapter.

3.2 Related Work

3.2.1 Biased Discriminant Analysis

In [32], Zhou and Huang proposed the BDA model as a principled way to deal with the RF scheme for CBIR, which is actually a $(1+x)$ -class discriminant analysis problem. This assumption for classes in CBIR is reasonable, because there is an unknown number of classes but the user is only concerned with the only one class semantically related to the query image.

The BDA model aims to find a subspace to discriminate the positive feedback samples from the negative ones, and this is a variant of two-class LDA based on the Fisher discriminant criterion (FDC) [126]. It is spanned by a set of vectors U , which can be obtained by maximizing the ratio between the trace of the projected between-class scatter matrix S_- and the trace of the projected positive within-class scatter matrix S_+ , i.e.,

$$U^* = \arg \max_U \frac{\text{tr}(U^T S_- U)}{\text{tr}(U^T S_+ U)}. \quad (3.1)$$

In experiments, there are n^+ positive feedback samples and n^- negative feedback samples in the training set. Then, S_- and S_+ can be defined as

$$\begin{cases} S_- = \sum_{i=1}^{n^-} (x_i^- - \bar{x}^+)(x_i^- - \bar{x}^+)^T, \\ S_+ = \sum_{i=1}^{n^+} (x_i^+ - \bar{x}^+)(x_i^+ - \bar{x}^+)^T, \end{cases} \quad (3.2)$$

where x_i^+ denotes the positive feedback sample; x_i^- denotes the negative feedback sample; and $\bar{x}^+ = \frac{1}{n^+} \sum_{i=1}^{n^+} x_i^+$ is the mean vector of the positive feedback samples. In general, the projection matrix U^* can be computed from the eigenvectors of $S_+^{-1} S_-$, corresponding to the largest eigenvalues. Because the number of training feedback samples is usually much less than the dimension of the visual feature space, this will lead to a degenerated S_+ , i.e., it is the so-called small-sized training data problem or the singular problem of the positive within-class scatter matrix. In the past decade, a lot of techniques have been proposed to alleviate the singular problem in LDA [29, 79, 128, 129], which have shown good performance for face recognition.

3.2.2 Differential Scatter Discriminant Criterion

Basically, in order to describe the class separability, we should convert the separability measure to a number, which should increase when the size of the between-class scatter matrix increases or the size of the within-class scatter matrix is smaller [126]. Different from the FDC, DSDC [46, 126] defines the separability measure as a trace difference for the projected between-class scatter matrix and the projected within-class scatter matrix rather than a trace ratio, i.e.,

$$U^* = \arg \max_{U^T U = I} \text{tr}(U^T S_b U) - \beta \text{tr}(U^T S_w U). \quad (3.3)$$

In Eq.(3.3), β is a nonnegative tuning parameter; and $U = \{u_1, \dots, u_l\} \in R^{h \times l}$, $l \ll h$, $U^T U = I$ is the projection matrix. The imposed orthogonality relationship between the projection directions, i.e., $U^T U = I$, is more effective for preserving the intrinsic geometric structure of data as shown in recent research [141, 142]. The matrix S_b is the between-class scatter matrix, which describes the interclass dispersion, while S_w is the within-class scatter matrix, which describes the intraclass compactness. Both scatter matrices of the DSDC have similar meanings with those of the FDC, and describe the interclass dispersion and the intraclass compactness, respectively. It is easy to verify that the solution of Eq.(3.3) is equivalent to solving the maximum of the Lagrange function

$$L(U, \lambda) = \sum_{k=1}^l u_k^T (S_b - \beta S_w) u_k - \lambda_k (u_k^T u_k - 1), \quad (3.4)$$

with multipliers λ_k 's. Letting $(\partial L(u_k, \lambda_k) / \partial u_k) = 0$, with $k = 1, \dots, l$, we can have

$$(S_b - \beta S_w) u_k = \lambda_k u_k, k = 1, \dots, l. \quad (3.5)$$

Thus, the problem is translated into finding the leading eigenvectors of $(S_b - \beta S_w)$, and hence, we need not calculate the inverse of S_w , which allows us to avoid the singular problem of the within-class scatter matrix easily.

Strictly speaking, the solution of Eq. (3.3) is equivalent to the FDC, only if the parameter β in DSDC is calculated as $\text{tr}(U_{opt}^T S_b U_{opt}) / \text{tr}(U_{opt}^T S_w U_{opt})$ [46, 126]. Therefore, the optimal β can only be obtained by the alternating projection method [46]. However,

for CBIR, because the distribution of the testing set diverges from that of the training set, a manually chosen value of β always achieves better prediction results than the calculated value. As demonstrated in [137, 139], when the within-class scatter matrix S_w is singular, the discriminant vectors of DSDC are approaching the discriminant vectors of the null space method of FDC at $\beta \rightarrow \infty$ [128]. When β is set properly, DSDC can show much better performance than the existing methods, which deal with the singular problem of the within-class scatter matrix in FDC.

3.3 Generalized Biased Discriminant Analysis

Let us denote the high-dimensional space as R^h and the low-dimensional intrinsic space as R^l . For convenience, we define $X^+ = \{x_i^+\}_{i=1}^{n^+} \in R^{h \times n^+}$ as the positive feedback samples, $X^- = \{x_i^-\}_{i=1}^{n^-} \in R^{h \times n^-}$ as the negative feedback samples in R^h . Then, we use $X = \{x_i\}_{i=1}^n = [X^+, X^-] \in R^{h \times n}$, $n = n^+ + n^-$ to denote all the feedback samples. We denote the embedding transformation for all the feedback samples by $f : X^+ \rightarrow Y^+$ and $X^- \rightarrow Y^-$. Therefore, after the embedding transformation, in R^l , the feedback samples matrix can be represented as $Y = \{y_i\}_{i=1}^n = \{Y^+, Y^-\} \in R^{l \times n}$. For simplicity, we restrict the embedding transformation to be linear, which can be defined by a projection matrix $U \in R^{h \times l}$ ($l \ll h$). Then, the low-dimensional representation of the sample x_i can be given as $y_i = U^T x_i \in R^l$.

The separability part of GBDA is based on the DSDC, i.e.,

$$J_s(U) = \arg \max_{U^T U = I} J_1(U) - \beta_1 J_2(U), \quad (3.6)$$

where $J_1(U)$ is the between-class scatter matrix and describes the interclass dispersion; $J_2(U)$ is the within-class scatter matrix and describes the intraclass compactness; and β_1 is a tuning parameter, which reflects the trade-off between the two goals.

In BDA [32], the between-class scatter matrix is defined as the negative scatter with respect to the positive centroid in the feature space. However, this is not a reasonable way to describe the separability of the two classes except that all positive feedback samples are drawn from a single Gaussian distribution, and this is always not the case in

CBIR. Recently, to take the nonlinearity of the sample distribution into account, non-parametric models have been developed for discriminant analysis methods and achieved improvement [126]. By reformulating the within-class scatter matrix and the between-class scatter matrix defined in LDA, Sugiyama et al. proposed the local LDA [143]. In [136], Yan et al. introduced a marginal Fisher analysis (MFA), which characterizes the interclass dispersion and the intraclass compactness by the sum of the distances between k nearest interclass neighbor samples and k nearest intraclass neighbor samples, respectively. Different from the conventional nonparametric discriminant analysis (NDA), MFA adopts a penalty graph and an intrinsic graph to characterize the interclass dispersion and the intraclass compactness, respectively, and it is largely inspired by the recent manifold learning community [41, 134, 135]. Motivated by these nonparametric techniques [136, 143], we implement the interclass dispersion by only selecting the samples pairs near the boundary, i.e.,

$$\begin{aligned}
 J_1(U) &= \arg \max_U \frac{1}{n_1} \sum_{i=1}^{n^+} \sum_{j \in N(i)} \|y_i^+ - y_j^-\|^2 + \frac{1}{n_1} \sum_{i=n^++1}^n \sum_{j \in N(i)} \|y_i^- - y_j^+\|^2 \\
 &= \arg \max_U \sum_{i=1}^n \sum_{j=1}^n w_{ij} \|y_i - y_j\|^2 \\
 &= \arg \max_U 2tr(Y(D^- - W^-)Y^T) \\
 &= \arg \max_U tr(U^T X(D^- - W^-)X^T U),
 \end{aligned} \tag{3.7}$$

where n_1 is the total number of k nearest interclass sample pairs between the positive class and the negative class; $N(i)$ is a set of indices of samples that are k nearest sample pairs among different classes for each sample x_i ; and w_{ij} is a weighting coefficient, which is defined as

$$w_{ij} = \begin{cases} \frac{1}{n_1}, & \text{if } 1 \leq i \leq n^+ \text{ and } n^+ + 1 \leq j \leq n, j \in N(i) \text{ or } i \in N(j), \\ \frac{1}{n_1}, & \text{if } n^+ + 1 \leq i \leq n \text{ and } 1 \leq j \leq n^+, j \in N(i) \text{ or } i \in N(j), \\ 0, & \text{else,} \end{cases} \tag{3.8}$$

where $W^- = \{w_{ij}\} \in R^{n \times n}$ is a symmetric matrix, and its entry is the weighting coefficient w_{ij} ; $D^- \in R^{n \times n}$ is a diagonal matrix, and its i th entry is $\sum_{j=1}^n w_{ij}$. Based on the definition of w_{ij} , we can see that the weighting coefficients encode both the class label information and the neighborhood relationship in the high-dimensional space. All of

these selected interclass sample pairs are used to capture the discriminative information between different classes.

To implement the intraclass compactness, similar to the original BDA, the definition of the intraclass compactness should also only bias towards the positive feedback samples in the positive class. Therefore, we only preserve the positive class compactness for the intraclass compactness, i.e.,

$$\begin{aligned}
 J_2(U) &= \arg \min_U \frac{1}{n_2} \sum_{i=1}^{n^+} \sum_{j=1}^{n^+} \|y_i^+ - y_j^+\|^2 \\
 &= \arg \min_U \sum_{i=1}^{n^+} \sum_{j=1}^{n^+} v_{ij} \|y_i^+ - y_j^+\|^2 \\
 &= \arg \min_U 2tr \left(Y^+ (D^+ - W^+) Y^{+T} \right) \\
 &= \arg \min_U tr \left(U^T X^+ (D^+ - W^+) X^{+T} U \right),
 \end{aligned} \tag{3.9}$$

where n_2 is the total number of sample pairs in the positive class; and v_{ij} is the weighting coefficient, which is defined as

$$v_{ij} = \begin{cases} \frac{1}{n_2}, & \text{if } 1 \leq i, j \leq n^+ \text{ and } i \neq j, \\ 0, & \text{if } 1 \leq i, j \leq n^+ \text{ and } i = j, \end{cases} \tag{3.10}$$

where $W^+ = \{v_{ij}\} \in R^{n^+ \times n^+}$ is a symmetric matrix; and D^+ is a diagonal matrix and its i th entry is $\sum_{j=1}^{n^+} v_{ij}$.

To reduce the risk of overfitting, we introduce the notion of local consistence into Eq.(3.6) to regularize the objective of separability, which is emerging from the manifold learning community [41, 134, 135]. Recently, a large number of nonlinear or linear methods have been proposed to discover the intrinsic manifold structure of the samples in the high-dimensional space. For example, Laplacian eigenmaps (LE) [40] preserves the similarities among nearest neighbor samples. These approaches yield impressive results on both benchmark artificial data sets and real-world data sets. Therefore, it is reasonable to expect that integrating the essential manifold structure of the positive feedback samples will further improve the performance of CBIR.

We choose the locality preserving projections (LPP) [144], which is the direct linearization of LE [40], to regularize the separability between different classes. LPP implements the local consistency principle by preserving the similarity among the nearest

neighbor samples and is widely used for face recognition [28] and image retrieval [84]. The local consistency for the positive feedback samples can be defined as

$$\begin{aligned}
 J_3(U) &= \arg \min_U \frac{1}{n_3} \sum_{i=1}^{n^+} \sum_{j \in S(i)} \|y_i^+ - y_j^+\|^2 \omega_{ij} \\
 &= \arg \min_U \sum_{i=1}^{n^+} \sum_{j \in S(i)} \|y_i^+ - y_j^+\|^2 m_{ij} \\
 &= \arg \min_U 2\text{tr} \left(Y^+ (D^m - W^m) Y^{+T} \right) \\
 &= \arg \min_U \text{tr} \left(Y^+ (D^m - W^m) Y^{+T} \right) \\
 &= \arg \min_U \text{tr} \left(U^T X^+ (D^m - W^m) X^{+T} U \right),
 \end{aligned} \tag{3.11}$$

where $\omega_{ij} = \exp(-\|x_i^+ - x_j^+\|^2 / \delta^2)$ is the heat kernel according to LE [40] and LPP [144], which reflects the affinity of the sample pairs; $S(i)$ is the set of indices of the nearest neighbor samples in the positive class for the positive feedback sample x_i ; n_3 is the total number of k nearest positive feedback sample pairs for all positive feedback samples; and the weighting coefficient m_{ij} can be defined as

$$m_{ij} = \begin{cases} \frac{\omega_{ij}}{n_3}, & \text{if } 1 \leq i, j \leq n^+, j \in S(i) \text{ or } i \in S(j), \\ 0, & \text{else,} \end{cases} \tag{3.12}$$

where $W^m = \{m_{ij}\} \in R^{n^+ \times n^+}$ is a symmetric matrix, which reflects the local geometry of the positive feedback samples in the high-dimensional space; and D^m is a diagonal matrix, and its i th entry is $\sum_{j=1}^{n^+} m_{ij}$.

According to He and Niyogi [40], a definition in Eq.(3.11) corresponds to the approximation of $\int_M \|\nabla f(x)\|^2$, the manifold on which the positive feedback samples reside. Minimizing the objective function can encourage the consistent output for the positive feedback samples in the high-dimensional space, and this will result in transforming with high local smoothness and best local geometry preservation. Hence, a smooth transformation that is expected to be less likely to overfit the training samples can be learned by this manifold regularization.

To sum up, the GBDA method can be formulated by combining the above two terms, i.e., $J_s(U)$ and $J_3(U)$, together, as given in Eq.(3.13).

$$\begin{aligned}
 J(U) &= J_s(U) - \beta_2 J_3(U) = J_1(U) - \beta_1 J_2(U) - \beta_2 J_3(U) \\
 &= \arg \max_U \sum_{i=1}^n \sum_{j=1}^n w_{ij} \|y_i - y_j\|^2 - \beta_1 \sum_{i=1}^{n^+} \sum_{j=1}^{n^+} v_{ij} \|y_i^+ - y_j^+\|^2 - \beta_2 \sum_{i=1}^{n^+} \sum_{j \in S(i)} m_{ij} \|y_i^+ - y_j^+\|^2 \\
 &= \arg \max_U \text{tr} \left(U^T X (D^- - W^-) X^T U \right) - \beta_1 \text{tr} \left(U^T X^+ (D^+ - W^+) X^{+T} U \right) \\
 &\quad - \beta_2 \text{tr} \left(U^T X^+ (D^m - W^m) X^{+T} U \right) \\
 &= \arg \max_U \text{tr} \left(U^T \begin{pmatrix} X(D^- - W^-)X^T - \beta_1 X^+(D^+ - W^+)X^{+T} \\ -\beta_2 X^+(D^m - W^m)X^{+T} \end{pmatrix} U \right), \tag{3.13}
 \end{aligned}$$

where β_2 is the regularization coefficient controlling the trade-off between the two objectives, i.e., the separability and the local consistency. By imposing the constraint $U^T U = I$ on Eq.(3.13), the optimal solution can be calculated by an eigenvalue decomposition, and the low-dimensional space is spanned by the l eigenvectors U associated with the l largest eigenvalues.

We empirically set the value of β_1 in RF for CBIR based on experiments, since the optimal β_1 may not be the best for classification. For the elements v_{ij} of the within-class scatter matrix, we have also normalized them by setting $v_{ij} = 1/n_2$. Therefore, from the view point of normalization, the two terms in $J_s(U)$ are actually balanced, i.e., we can set $\beta = 1$ in RF for simplicity. The value of β_2 is used to trade off the separability and the local consistency. However, how to tune the regularization coefficient and balance the two objectives is still an open question. Intuitively, a larger value of β_2 will result in a solution that can enlarge $J_s(U)$ and diminish $J_3(U)$, and therefore, it will lead to enhance the separability and encourage the local consistency. In the following experiments, we present the sensitivity of GBDA in relation to the parameter β_2 , and then select the value that shows the best performance.

The proposed GBDA method for CBIR is summarized in Table 3.1.

3.4 Experimental Results

We have implemented a CBIR system based on the Corel image database, which includes 10,763 images with 80 different semantic concepts [145]. To represent the images, we choose three groups of low-level visual features. For color, we utilize the color histogram

Table 3.1: Generalized Biased Discriminant Analysis

Input: $X = \{x_i\}_{i=1}^{n^++n^-}$, $X^+ = \{x_i\}_{i=1}^{n^+}$, and $X^- = \{x_i\}_{i=n^++1}^{n^++n^-}$ indicate all feedback samples, positive feedback samples, and negative feedback samples, respectively.
Step 1: Calculate the between-class scatter matrix $S_- = X(D^- - W^-)X^T$.
Step 2: Calculate the positive within-class scatter matrix $S_+ = X^+(D^+ - W^+)X^{+T}$.
Step 3: Calculate the manifold regularization term according to $X^+(D^m - W^m)X^{+T}$.
Step 4: Combine the three terms together by the tuning parameters, i.e., β_1 and β_2 , $\Phi = X(D^- - W^-)X^T - \beta_1 X^+(D^+ - W^+)X^{+T} - \beta_2 X^+(D^m - W^m)X^{+T}$.
Step 5: Do the eigenvalue decomposition with Φ , and select l eigenvectors U of Φ with the first l largest eigenvalues.
Step 6: For a given testing sample x , the GBDA transformation is $f(x) = U^T x$.
Output: $f(z)$ indicates the embedded testing sample.

[10] to represent the color information, and quantize hue and saturation into eight bins and value into four bins. We use the Weber’s law descriptors (WLD) [146] to represent the local features of the images, which will result in a feature vector of 240 values. For shape, the edge directional histogram from the Y component is adopted to capture the spatial distribution of the edges [12]. Five categories, including horizontal, 45° diagonal vertical, 135° diagonal, and isotropic directions, are calculated to form the shape features. All of these features are combined into a feature vector, which results in a vector with 510 values. Then, all feature components are normalized to normal distributions with zero mean and one standard deviation to represent the images.

In experiments, 500 query images are randomly selected from the Corel image database, and then, RF is automatically conducted by the CBIR system. In some initial experiments, we note that the number of relevant images (i.e., images with the same concept as the query image) at each round of RF may range from zero to the number of images displayed to the user and the number of irrelevant images (i.e., images with different concepts as the query image) may range from the number of images displayed to the user to zero. Therefore, we design the following RF procedure: at each round of RF, top 20 results resulting from the resorted results are serially examined from the top; the first 5 semantically relevant images are marked as the positive feedback samples and the first 5 semantically irrelevant images are marked as the negative feedback samples unless fewer

such images are found in top 20 results, in which case the fewer number of samples found are used as the feedback samples. Note that the images which have been selected at previous rounds of RF are excluded from later selections. All the labeled images in the RF iterations are used to train an RF model.

We use average precision (AP), average recall (AR), and standard deviation (SD) to evaluate the performance of the RF methods. AP refers to the percentage of relevant images in top returned results presented to the user, and is calculated as the averaged values of all the query images. AR shows the fraction of the related images that are successfully retrieved, and is defined as the percentage of retrieved images among all relevant images in the database. SD indicates the stability of different RF methods, and is calculated for all APs to describe the robustness of the methods.

3.4.1 Parameter Sensitivity

In order to select a proper quantity of k nearest interclass sample pairs to describe the discriminative information, we first show the performance comparison of GBDA on different quantities of k nearest interclass sample pairs. Figure 3.1 shows the top 10, 30, and 50 results after the third, fifth, seventh, and ninth rounds of RF with different k values from 3 to 13 based on 400 independent experiments.

As we can see from Figure 3.1, the AP curves change slightly for different k values. Even a small number of interclass sample pairs can capture the discriminative information well and obtain good performance with regard to APs. Therefore, in the following subsection, we simply select $k = 4$ for all the following experiments.

Then, we show the sensitivity of GBDA with regard to different values of the parameter β_2 , and empirically set the parameter as a value in a sequence, i.e., $\{2^i, i = -10, -9, \dots, 9, 10\}$. Figure 3.2 (a) and (b) give AP and AR curves in top 30 results after the fifth and ninth rounds of RF with different β_2 values based on 500 independent experiments, respectively. In experiments, we find that the parameter β_2 significantly affects the results. As shown in Figure 3.2 (a), we can see that, when β_2 is small enough, i.e., the local consistency contributes little to the formulation, the performance degrades significantly. When β_2 becomes larger, the GBDA method shows much better performance with regard to APs and ARs. However, if β_2 is too large, the performance may

(a)

(b)

(c)

Figure 3.1: APs of GBDA with different k values in top N results based on 400 independent experiments. (a) top 10, (b) top 30, and (c) top 50.

Figure 3.2: APs and ARs of GBDA with different β_2 values in top 30 results based on 500 independent experiments. (a) APs and (b) ARs.

degenerate. This is mainly due to oversmoothing. From the results, we can see that for this problem, the algorithm achieves best performance when β_2 is set as $\beta_2 = 2^6$. Therefore, in the following experiments, we empirically set the trade-off parameter $\beta_2 = 2^6$. It is convinced that the parameter β_2 can be further tuned to achieve better performance. This analysis above also indicates the important role of the local consistency for improving the generalization ability.

3.4.2 Statistical Experimental Results

In this subsection, we focus on the comparison of the proposed GBDA with the original BDA [32] and some of its variants, namely, the enhanced DBDA [86], null space BDA (NBDA) [128], and MBA [87], all of which are linear embedding methods and obtain much better performance compared with the original BDA. Simultaneously, SVM-based RF methods including SVM [35] and constrained SVM (CSVM) [37] are also compared to evaluate the performance of GBDA. BDA, MBA, DBDA, and NBDA are all based on the FDC [126].

For MBA, we empirically set the within-class compactness parameter k_1 according to LPP and the between-class separability parameter $k_2 = 4$. Due to the high-dimensional features, both the original BDA and MBA will encounter the small-sized training data problem or the singular problem, and hence, in the experiments that follow, a regularization method is used to solve the singular problem. The enhanced version of the original BDA, DBDA [86], is solved by the direct method, which first removes the null space of the negative scatter with respect to the positive centroid, and then, the eigenvectors of the positive within-class matrix corresponding to the smallest eigenvalues are extracted as the most discriminative directions. We choose the Gaussian kernel function for SVM and CSVM because it achieves the best performance for all kernel based algorithms with different parameters. For all SVM based methods, we use the Ohio-State-University SVM to implement the classification. All of the parameters are set identically as the descriptions in the corresponding papers [32, 86, 87].

As can be seen in Figure 3.3, all of these RF methods can show significantly better performance than the initial retrieval results without any RF information, and this indicates that the RF information provided by the user are very helpful in improving the

Figure 3.3: APs of GBDA compared with the RF methods, i.e., NBDA, MBA, DBDA, BDA, SVM, and CSVM, based on 500 independent experiments. All the results are evaluated over 9 rounds of RF. 0 round of RF refers to the retrieved results based on the Euclidean distance metric without any RF information.

performance of the system. The proposed GBDA method consistently outperforms BDA, MBA, DBDA, NBDA, SVM, and CSVM for RF. The subfigures in Figure 3.3 show that APs of the 500 experiments in top 20, 40, 60, 80, 100, and 120 results. We can see that GBDA can achieve better performance compared with the original BDA. The unreasonable Gaussian distribution assumption for the positive feedback samples in the original BDA and the regularization method are the main reasons, which usually result in poor performance of BDA for CBIR. DBDA solves the singular problem by the direct method, and can achieve better performance than the original BDA. However, much discriminative information contained in the null space of S_b is discarded. MBA effectively extracts the discriminative information from the marginal samples; however, it still suffers from the singular problem, which causes serious stability problems for MBA.

The proposed GBDA can extract the most discriminative information from the k nearest neighborhood interclass samples, but never encounters the singular problem. Basically, GBDA is an effective approach that can work in the whole input space rather than only in the principal space of the between-class scatter matrix [79] or in the null space of the within-class scatter matrix [128]. Therefore, GBDA can keep more discriminative information. By introducing the manifold regularization, a locally smooth and consistent transformation can be learned, which is expected to be less vulnerable to overfit the training samples as shown in Subsection 3.4.1.

According to Figure 3.3, we note that APs of GBDA clearly show superior performance. In the case of top 40 results, we note that after the sixth round of RF, the AP can achieve 60 percent, while the original BDA and its variants usually require 8 rounds of RF. Simultaneously, all the discriminant analysis based methods are much better than the SVM based methods, and the proposed GBDA is much better than BDA and its variants. When more top results are considered, GBDA can significantly outperform the other methods. The SDs corresponding to the six methods in top 20, 40, 60, 80, 100, and 120 results for the 500 experiments are given in Figure 3.4. For top 20 results, the GBDA method is much more stable and effective than other methods. Then, for other top results, the SDs of GBDA are similar to the other discriminant analysis based methods. Although the SDs of CSVM are much smaller than those of other methods, the APs of CSVM are much lower than the discriminant analysis based methods. This shows that the proposed GBDA is more effective and stable in overall performance.

Figure 3.4: SDs of GBDA compared with the RF methods, i.e., NBDA, MBA, DBDA, BDA, SVM, and CSVM, based on 500 independent experiments. All the results are evaluated over 9 rounds of RF. 0 round of RF refers to the retrieved results based on the Euclidean distance metric without any RF information.

Table 3.2: APs in top N results of the compared algorithms, i.e., NBDA, MBA, DBDA, BDA, SVM, and CSVM, after the ninth round of RF

Algorithms	Top 20	Top40	Top60	Top80	Top100	Top120	Top140
GBDA	83.35	64.91	53.18	45.20	39.09	34.34	30.73
NBDA	80.98	61.09	49.65	42.22	36.70	32.61	29.39
MBA	82.23	59.88	48.50	41.13	35.46	31.19	27.83
DBDA	81.86	59.73	47.84	40.27	34.87	30.08	27.64
BDA	78.33	58.26	45.87	39.33	33.79	29.77	26.74
SVM	71.50	53.48	43.79	36.98	32.01	28.19	25.17
CSVM	72.08	50.53	39.33	32.51	27.77	24.54	22.03

Table 3.3: ARs in top N results of the compared algorithms, i.e., NBDA, MBA, DBDA, BDA, SVM, and CSVM, after the ninth round of RF

Algorithms	Top 20	Top40	Top60	Top80	Top100	Top120	Top140
GBDA	14.18	21.88	26.77	30.21	32.47	33.93	35.27
NBDA	13.74	20.46	24.73	27.84	30.02	31.81	33.17
MBA	14.19	20.69	25.23	28.25	30.44	31.86	33.07
DBDA	13.48	19.44	23.14	25.78	27.68	29.18	30.43
BDA	13.38	18.54	22.78	26.25	28.11	29.60	30.81
SVM	12.42	18.54	22.78	25.44	27.32	28.69	29.79
CSVM	12.64	17.43	20.16	22.19	23.50	24.73	25.79

To better illustrate the performance of the compared algorithms, we also give the AP scope curves and the AR scope curves after the ninth round of RF in Figure 3.5. The horizontal axis is the number of top ranked results used in evaluation and the vertical axis is the AP or the AR measured in top results. As we can see from Figure 3.5, it is evident that GBDA substantially outperforms the original BDA method and other methods. The detailed results of APs and ARs in top results for different algorithms after the ninth round of RF are shown in Table 3.2 and Table 3.3, respectively. According to the two tables, we can observe that GBDA is more tolerant and achieves more promising results compared with other methods when the retrieval process is stopped. Specifically, GBDA enjoys almost all the best performance except the AR in top 20 results. Based on the aforementioned observations, we can empirically conclude that GBDA is more effective than the other compared algorithms in experiments.

3.4.3 Discussions

In general, for CBIR, the BDA method perform much better than the conventional discriminant analysis based methods, e.g., LDA. This is mainly because, when the user provides a query image, he/she would like to get more semantically relevant images

(a)

(b)

Figure 3.5: APs and ARs of the compared algorithms, i.e., NBDA, MBA, DBDA, BDA, SVM, and CSVM, after the ninth round of RF. (a) APs and (b) ARs.

which all share a common concept, but never cares about the irrelevant images, all of which differ in various concepts. FDC is an effective technique for feature extraction and pattern classification. However, for a task with high-dimensional data such as images, Fisher's LDA encounters a fundamental difficulty caused by the singular problem of the within-class scatter matrix, which is known to cause the serious stability problem in LDA. DSDC is an effective, stable and efficient criterion, which can also represent the class separability well, but does not have the singular problem.

Basically, devising a reasonable similarity metric, e.g., Mahalanobis distance metric, plays an important role for an image retrieval task. Different from previous work [32, 88, 93], in this study, we have shown that, by alleviating the unstable numerical computation problem (i.e., the singular problem) and the unreasonable model (i.e., the Gaussian distribution) in a metric learning algorithm (i.e., BDA), the performance of the system can be significantly improved for an image retrieval task.

Several aspects can be improved regarding the RF method and image representations. For example, the kernel machines can be incorporated to enhance the performance of GBDA as in the previous work [32, 86, 87]; newly proposed features may outperform the traditional ones, e.g., a sparse coding representation [147].

3.5 Summary

Generalized biased discriminant analysis (GBDA) is proposed to enhance the conventional biased discriminant analysis (BDA) for content-based image retrieval (CBIR). BDA is one of the most important discriminant analysis based RF methods for CBIR; however, the small-sized training data problem will lead to the singular problem of the positive within-class scatter matrix, and the Gaussian distribution assumption for the positive feedback samples is also not very reasonable. To avoid the singular problem in BDA, GBDA is designed based on the differential scatter discriminant criterion (DSDC), which defines the interclass separability as a trace difference for the between-class scatter matrix and the within-class scatter matrix rather than a trace ratio. Furthermore, to avoid the Gaussian distribution assumption for the positive feedback samples, the between-class scatter matrix is specially designed by resorting to nearest neighbor samples in different

classes. Additionally, to reduce the overfitting problem, the locality preserving principle emerging from the manifold learning community, which measures the local smoothness of the feature transformation, is integrated to regularize the between-class separability term. Therefore, a locally smooth and consistent transformation can also be learned for CBIR. Extensive experiments on a real-world image database have shown that the proposed GBDA significantly outperforms the original BDA and its enhance versions (namely, DBDA, NBDA, and MBA), as well as SVM and CSVM.

Chapter 4

Semi-supervised Biased Maximum Margin Analysis

4.1 Motivation

Relevance feedback (RF) [22] is a powerful tool to improve the performance of content-based image retrieval (CBIR) [1–3]. In this procedure, the user first labels a number of semantically relevant and irrelevant images with the query image, which are the positive and negative feedback samples, respectively. And then, the CBIR system can refine the retrieved results based on these two groups of feedback samples provided by the user. The two steps are carried out iteratively to enhance the performance of the CBIR system by gradually learning the user’s preferences.

Many RF methods have been developed based on different assumptions for the positive and negative feedback samples in recent years. They either adjust the weights of visual features to adapt to the user’s preferences [76] or estimate the density of the positive feedback samples [91]. Moreover, discriminant analysis based methods have also been widely used to construct the RF schemes for CBIR [32, 86, 87, 130]. However, most of these methods can only work well with certain limitation. The method used in [76] to construct the RF scheme is heuristic. The density estimation method in [91] totally ignores the information contained in the negative feedback samples. The discriminant analysis based methods often suffer from the small-sized training data problem or the singular problem of the positive within-class scatter matrix [32, 86, 87, 130].

Regarding the positive and negative feedback samples as two different classes and aiming to find a classifier to identify these two groups from each other, RF in CBIR can

Figure 4.1: A typical set of positive and negative feedback samples at each round of RF.

be considered as an online classification problem. Consequently, classification based RF methods have become a popular research topic in CBIR. Recently, many classification based methods were developed to deal with the positive and negative feedback samples in RF. One-class support vector machine (OneSVM) can estimate the density of positive feedback samples [91]. Zhou and Huang proposed a $(1+x)$ -class classification method in which there is an unknown number of classes but the user is only interested in the one related to the query image [32]. Moreover, a multiclass based RF method was also designed to scale the performance of the CBIR system [148].

Among these classifiers, the two-class support vector machine (SVM) [49, 50] is one of the most popular classification methods widely used in recent years and has obtained state of the art performance in classification for its good generalization ability [33, 51–53, 110–113]. The SVM can achieve a minimal structural risk by minimizing the Vapnik-Chervonenkis dimensions [49, 50]. Guo et al. developed a constrained similarity measure for image retrieval [37], which learns a boundary that divides the images in the database into two groups, and the images inside the boundary are ranked by their Euclidean distances to the query image. SVM active learning can select samples close to the optimal hyperplane of SVM as the most informative ones for the user to label [34]. Random sampling techniques were applied to alleviate the unstable, biased, and overfitting problems

in SVM-based RF for CBIR [36]. Li et al. proposed a multitasking SVM-based RF method by adapting a cotraining technique and a random sampling method to alleviate the small number of training feedback samples problem [90]. Nevertheless, most of these SVM-based RF methods ignore the basic difference between the two distinct groups of feedback samples, i.e., all positive feedback samples share a common concept with the query image while each negative feedback sample differs in diverse concepts. For instance, a typical set of feedback samples in RF is shown in Figure 4.1. All samples labeled as the positive feedback samples share a common concept (i.e., elephant), while each sample labeled as the negative feedback sample differs in diverse concepts (i.e., flower, horse, banquet, hill, etc.). Traditional SVM-based RF methods treat the positive and negative feedback samples equally [35–37, 90], although this is not appropriate since all positive feedback samples share a common concept with the query image while each negative feedback sample differs in diverse concepts. Directly using the SVM as an RF scheme is potentially damaging the performance of the CBIR system. One problem stems from the fact that different semantic concepts live in different subspaces and each image can live in many different semantic concept subspaces, and it is the goal of the RF scheme to figure out “which one” [32]. However, it will be a burden for traditional SVM-based RF methods to tune the internal parameters to adapt to the changes of such semantic concept subspaces. Such difficulties have severely degraded the effectiveness of the traditional SVM-based RF for CBIR. Additionally, it is problematic to incorporate the information of unlabeled samples into the traditional SVM-based RF for CBIR, although the unlabeled samples are very helpful in constructing the optimal classifier, alleviating the noise, and enhancing the performance of the system.

To explore solutions to these two aforementioned problems in the current technology, we propose a biased maximum margin analysis (BMMA) and a semi-supervised BMMA (SemiBMMA) for the traditional SVM-based RF based on the graph embedding framework [136]. The proposed scheme is mainly based on the following: 1) the effectiveness of treating the positive and negative feedback samples unequally [32, 86, 87]; 2) the significance of the optimal subspace or the feature subset in interactive image retrieval; 3) the success of the graph embedding framework in characterizing the intrinsic geometric structure of data in a high-dimensional space [41, 134, 136]; and 4) the convenience of the

graph embedding framework in constructing semi-supervised learning techniques. With the incorporation of BMMA, labeled positive feedback samples are mapped as close as possible, whereas labeled negative feedback samples are separated from the labeled positive feedback samples by a maximum margin in the reduced subspace. The traditional SVM-based RF combined with BMMA can better model the RF procedure and reduce the performance degradation caused by the asymmetric property of training data. The SemiBMMA can incorporate the information of unlabeled samples into the RF procedure and effectively alleviates the overfitting problem caused by a small number of training feedback samples. To show the effectiveness of the proposed scheme combined with SVM-based RF, we will compare it with the traditional SVM-based RF and some other related existing methods for RF based on a real-world image database. Extensive experiments demonstrate that the proposed scheme can significantly improve the performance of the traditional SVM-based RF for image retrieval.

This chapter is organized as follows. In Section 4.2, the related previous work, i.e., the traditional SVM-based RF for CBIR and the graph embedding framework, are briefly reviewed, respectively. In Section 4.3, BMMA and SemiBMMA for SVM-based RF are proposed. A CBIR system and the image representations are given in Section 4.4. In Section 4.5, a large number of experiments, which validate the effectiveness of the proposed scheme are conducted. Finally, the summary of this chapter is given in Section 4.6.

4.2 Related Work

4.2.1 Traditional SVM-based RF for CBIR

In this subsection, we briefly introduce the traditional SVM-based RF for CBIR. The SVM implements the structure risk minimization by minimizing the Vapnik-Chervonenkis dimensions [49]. Consider a linearly separable binary classification problem as

$$\{(x_1, y_1), \dots, (x_n, y_n)\} \text{ and } y_{i=1, \dots, n} = \{+1, -1\}, \quad (4.1)$$

where x_i denotes an h -dimensional sample; n is the number of training samples; and y_i is the label of the class that the sample x_i belongs to. The objective function of SVM aims to find an optimal hyperplane to separate the two classes, i.e.,

$$w^T x + b = 0, \quad (4.2)$$

where x is an input sample; w is a weight; and b is a bias. The SVM attempts to find the two parameters w and b for the optimal hyperplane by maximizing the geometric margin $2/\|w\|$, subject to

$$y_i(w^T x_i + b) \geq 1, \quad (4.3)$$

The solution of the objective function can be found through a Wolfe dual problem with the Lagrangian multiplied by α_i , i.e.,

$$Q(\alpha) = \sum_{i=1}^n \alpha_i - \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) / 2, \quad (4.4)$$

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

In general, in the dual problem, the samples appear only in the inner product, which can often be replaced with a positive definite kernel function for better performance.

$$x_i \cdot x_j \rightarrow \Phi(x_i) \cdot \Phi(x_j) = K(x_i, x_j), \quad (4.5)$$

where $K(\cdot)$ is the kernel Gram matrix. The kernel version of the Wolfe dual problem is

$$Q(\alpha) = \sum_{i=1}^n \alpha_i - \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j K(x_i \cdot x_j) / 2. \quad (4.6)$$

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

$$F(x) = \text{sgn}(f(x)), \quad (4.7)$$

where $f(x) = \sum_{i=1}^s \alpha_i y_i K(x_i, x) + b$ is the output optimal hyperplane of SVM; and s is the number of support vectors.

Generally, the output of SVM (i.e., $f(x)$) is usually used to measure the similarity between a given image and the query image for the traditional SVM-based RF in CBIR. The generalization performance of SVM depends mainly on the number of support vectors. Orthogonal complement component analysis (OCCA) decreases the number of support

vectors by finding a subspace, in which all positive feedback samples are merged [149]. However, it still totally ignores the information contained in negative feedback samples, and this is very helpful in finding a homogeneous subspace. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training samples, since in general, the larger the margin, the lower the generalization error of the classifier.

4.2.2 The Graph Embedding Framework

To describe our scheme clearly, let us first review the graph embedding framework introduced in [136]. Generally, for a classification problem, the sample set can be represented by a matrix $X = [x_1, x_2, \dots, x_n] \in R^{h \times n}$, where n indicates the total number of samples and h is the feature dimension. Let $G = \{X, W\}$ be an undirected similarity graph, which is called the intrinsic graph, with a vertex set X and a similarity matrix $W \in R^{n \times n}$. The similarity matrix W is real and symmetric, and measures the similarity between pairs of vertices; W can be constructed using various similarity criteria. The corresponding diagonal matrix D and the Laplacian matrix L^I of the graph G can be defined as

$$L^I = D - W, D_{ii} = \sum_{j \neq i} W_{ij}, \forall i = 1, \dots, n. \quad (4.8)$$

Graph embedding of the graph G is defined as an method to determine the low-dimensional representations $Y = [y_1, y_2, \dots, y_n] \in R^{l \times n}$ of the vertex set X , where l is lower than h for the dimension. The column vector y_i is the embedding result for the sample x_i , which preserves the similarity between pairs of vertices in the original high-dimensional space. Then, in order to characterize the difference between pairs of vertices in the original high-dimensional space, a penalty graph $G^p = \{X, W^p\}$ is also defined, where the vertex set X is the same as that of the intrinsic graph G , but the edge weight matrix W^p corresponds to the similarity characteristics that are to be suppressed in the low-dimensional space. For a dimension reduction problem, direct graph embedding requires an intrinsic graph G , while a penalty graph G^p is not a necessary input. Then, the similarity between pairs of vertices can be maintained according to the graph preserving criterion as

$$y^* = \arg \min_{tr(YL^PY^T)=c} \sum_{i \neq j} \|y_i - y_j\|^2 W_{ij} = \arg \min_{tr(YL^PY^T)=c} tr(YL^IY^T), \quad (4.9)$$

where $tr(\cdot)$ is the trace of a square matrix; c is a constant; and L^P is the constraint matrix. L^P may typically be a diagonal matrix for scale normalization or express more general constraints among vertices in the penalty graph G^p , and it describes the similarity between vertices that should be avoided. L^P is the Laplacian matrix of G^p , similar to Eq.(4.8), which can also be defined as

$$L^P = D^p - W^p, D_{ii}^p = \sum_{j \neq i} W_{ij}^p, \forall i = 1, \dots, n, \quad (4.10)$$

where W^p is the similarity matrix of the penalty graph G^p to measure the difference between pairs of vertices in G^p .

The graph embedding framework preserves the intrinsic property of samples in two ways: for larger similarity between samples x_i and x_j , the distance between y_i and y_j should be smaller to minimize the objective function. Conversely, smaller similarity between x_i and x_j should lead to a larger distance between y_i and y_j . Hence, through the intrinsic graph G and the penalty graph G^p , the similarity and the difference among pairs of vertices in a graph can be approximately preserved in the embedding.

In [30], based on the graph embedding framework, Eq.(4.9) can be resolved by converting it into the following trace ratio formulation, i.e.,

$$Y^* = \arg \max_Y \frac{tr(YL^PY^T)}{tr(YL^IY^T)}. \quad (4.11)$$

Generally, if the constraint matrix represents only scale normalization, this formulation can be directly solved by a standard eigenvalue decomposition. However, for a more general constraint matrix, it can be approximately solved with a generalized eigenvalue decomposition by transforming the objective function into a more tractable approximate form $\arg \max_Y tr((YL^IY^T)^{-1}(YL^PY^T))$. Recent research has shown that Eq.(4.11) can be solved efficiently by gradient descent and the new method has shown promising performance compared with these conventional solutions [108].

The low-dimensional representations of vertices can be approximately obtained from a linear embedding, i.e., $y_i = U^T x_i$, where U is the embedding matrix, the objective function Eq.(4.11) can be changed to

$$U^* = \arg \max_U \frac{\text{tr}(U^T X L^P X^T U)}{\text{tr}(U^T X L^I X^T U)}. \quad (4.12)$$

During the past decade, a number of manifold learning based methods have been proposed to capture the intrinsic geometry property [40, 41, 134, 144, 150, 151], In [136], Yan et al. claimed that all of the aforementioned manifold learning methods can be mathematically unified within the graph embedding framework described in this subsection. They also proposed a marginal Fisher analysis (MFA), which takes both the manifold geometry and the class label information into consideration. However, MFA still suffers from the singular problem when the number of training samples is much less than the dimension of the feature space.

4.3 BMMA and SemiBMMA for SVM-based RF

With the observation that all positive feedback examples are alike and each negative feedback example is negative in its own way, the two groups of feedback samples have distinct properties in CBIR [32]. However, the traditional SVM-based RF treats the positive and negative feedback samples equally.

To alleviate the performance degradation when directly using the SVM as an RF scheme for CBIR, we explore solutions based on the argument that different semantic concepts live in different subspaces and each image can live in many different semantic concept subspaces [32]. We formally formulate this problem as a general subspace learning problem, and propose a BMMA method for the traditional SVM-based RF. In the reduced subspace, the negative feedback samples, which differ in diverse concepts with the query image, are separated by a maximum margin from the positive feedback samples, which share a similar concept with the query image. Therefore, we can easily map the positive and negative feedback samples onto a semantic concept subspace in accordance with human perception of the image content.

To utilize the information of unlabeled samples in the database, we introduced a Laplacian regularizer to BMMA, which will lead to SemiBMMA for the traditional SVM-based RF. The resulting Laplacian regularizer is largely based on the notion of local consistency [44, 152], which is inspired by the recently emerging manifold learning community [41, 134, 135], and can deal with the scenario when the number of labeled samples is generally small and vast amounts of unlabeled samples are readily available in the database. This is because the resulting Laplacian regularizer can effectively describe the weakly similar relationship between pairs of unlabeled samples.

Then, the remaining images in the database are projected onto this resulting semantic concept subspace and a similarity metric is applied to sort the images based on the new representations. For SVM-based RF, the distance to the hyperplane of the classifier is the criterion to distinguish the query relevant samples from the query irrelevant ones. After the projection step, all positive feedback samples are clustered together, while negative feedback samples are well separated from the positive feedback samples by a maximum margin. Therefore, the resulting optimal hyperplane of SVM in this subspace will be much simpler and better than in the original high-dimensional space.

Different from classical subspace learning methods, e.g., principle component analysis and linear discriminant analysis (LDA), which can only see the linear global Euclidean structure of samples, BMMA aims to learn a projection matrix such that in the projected space, the positive feedback samples have high local within-class compactness, but the samples with different labels have high local between-class separability. To describe our method clearly, first we introduce some notations of this method.

At each round of RF, there are n samples $X = \{x_1, x_2, \dots, x_n\} \in R^h$. For simplicity, we assume the first n^+ samples $X^+ = \{x_i^+\}_{i=1}^{n^+} \in R^{h \times n^+}$ as positive feedback samples, the next n^- samples $X^- = \{x_i^-\}_{i=1}^{n^-} \in R^{h \times n^-}$ as negative feedback samples, and all others are unlabeled samples $x_i (n^+ + n^- + 1 \leq i \leq n)$. Let $l(x_i)$ be the class label of a sample x_i , we denote $l(x_i) = 1$ for the positive feedback samples, $l(x_i) = -1$ for the negative feedback samples, and $l(x_i) = 0$ for the unlabeled samples. To better show the relationship between the proposed scheme and the graph embedding framework, we use similar notations and equations in the original graph embedding framework, which provides us a general platform to develop various new methods for dimension reduction.

First, two different graphs are formed: 1) the intrinsic graph G , which characterizes the local similarity of the positive feedback samples; and 2) the penalty graph G^p , which characterizes the local difference of the feedback samples.

For all positive feedback samples, we first compute the pairwise distances between each pair of positive feedback samples. Then, for each sample, we find its k_1 nearest positive feedback samples, which can be represented as a sample set N_i^s , and put edges between x_i and its k_1 nearest positive feedback samples. The intrinsic graph is characterized as

$$\begin{aligned} S_I &= \sum_i \sum_{j: j \in N_i^s \text{ or } i \in N_j^s} \|U^T x_i - U^T x_j\|^2 W_{ij} \\ &= 2tr(U^T X(D - W)X^T U), \end{aligned} \quad (4.13)$$

$$W_{ij} = \begin{cases} 1/|N^s|, & \text{if } l(i) = 1 \text{ and } l(j) = 1, i \in N_j^s \text{ or } j \in N_i^s \\ 0, & \text{else,} \end{cases} \quad (4.14)$$

where D is a diagonal matrix, whose diagonal elements are calculated by $D_{ii} = \sum_j W_{ij}$; $|N^s|$ denotes the total number of k_1 nearest positive feedback sample pairs for each positive feedback sample. Basically, the intrinsic graph measures the total average distance of $|N^s|$ nearest sample pairs, and is used to characterize the local within-class compactness of all positive feedback samples.

For the penalty graph G^p , its similarity matrix W_{ij}^p represents geometric or statistical properties to be avoided and is used as a constraint matrix in the graph embedding framework. In BMMA, the penalty graph G^p is constructed to represent the local separability between the positive and negative classes. More strictly speaking, we expect that the total average margin between the sample pairs with different class labels should be as large as possible.

For each feedback sample, we find its k_2 nearest feedback samples with different class labels and put edges between corresponding pairs of feedback samples with weights W_{ij}^p .

Then, the penalty graph can be formed as

$$\begin{aligned} S_P &= \sum_i \sum_{j: j \in N_i^p \text{ or } i \in N_j^p} \|U^T x_i - U^T x_j\|^2 W_{ij}^p \\ &= 2tr(U^T X(D^p - W^p)X^T U), \end{aligned} \quad (4.15)$$

$$W_{ij}^p = \begin{cases} 1/|N^p|, & \text{if } l(i) = 1 \text{ and } l(j) = -1, i \in N_j^p \text{ or } j \in N_i^p, \\ 0, & \text{else,} \end{cases} \quad (4.16)$$

where D^p is a diagonal matrix, whose diagonal elements are calculated by $D_{ii}^p = \sum_j W_{ij}^p$; $|N^p|$ denotes the total number of k_2 nearest sample pairs with different class labels. Similarly, the penalty graph measures the total average distance of $|N^p|$ nearest sample pairs in different classes, and is used to characterize the local between-class separability of feedback samples.

In the following, we will show how to utilize the graph embedding framework to develop algorithms based on the designed intrinsic and penalty graphs. Different from the original formulation of the graph embedding framework in [136], the BMMA method optimizes the objective function in a trace difference form instead, i.e.,

$$\begin{aligned}
 U^* &= \arg \max_U 2tr(U^T X(D^p - W^p)X^T U) - 2tr(U^T X(D - W)X^T U) \\
 &= \arg \max_U tr(U^T X(D^p - W^p)X^T U) - tr(U^T X(D - W)X^T U) \\
 &= \arg \max_U tr(U^T X L^P X^T U) - tr(U^T X L^I X^T U) \\
 &= \arg \max_U tr(U^T X(L^P - L^I)X^T U).
 \end{aligned} \tag{4.17}$$

As given in Eq.(4.17), we can notice that the objective function works in two ways, and tries to maximize $tr(U^T X L^P X^T U)$ and at the same time minimize $tr(U^T X L^I X^T U)$. Intuitively, we can analyze the meaning of the objective function in Eq.(4.17) geometrically. By formulating the objective function as a trace difference form, we can regard it as the total average local margin between the positive and negative feedback samples. Therefore, Eq.(4.17) can be used as a criterion to distinguish different classes. In [138], a maximum margin criterion (MMC) was presented as an objective function with a similar formulation. The differences between BMMA and MMC are the definitions of the interclass separability and the intraclass compactness. In MMC, both of the interclass separability and the intraclass compactness are defined as the same in LDA, and MMC is developed with the assumption that the samples of each class are of a Gaussian distribution. In BMMA, the intraclass compactness is characterized by the sum of the distances between each positive feedback sample and its nearest samples in the same class, and the interclass separability is defined by resorting to interclass marginal samples. Without prior assumptions for the data distribution, BMMA can find more reliable low-dimensional representations of the original data compared with MMC in [138].

In order to remove an arbitrary scaling factor in the projection, we additionally require that U is constituted by unit vectors, i.e., $u_k^T u_k = 1, k = 1, 2, \dots, l$. This means that we need to solve the following constraint optimization problem as

$$\begin{aligned} & \max_U \text{tr} (U^T X (L^P - L^I) X^T U) \\ & = \sum_{k=1}^l u_k^T X (L^P - L^I) X^T u_k \\ & \text{s.t. } u_k^T u_k - 1 = 0, k = 1, 2, \dots, l. \end{aligned} \quad (4.18)$$

Note that we may also use other constraints instead. For example, we may require $\text{tr}(U^T X L^I X^T U) = 1$ and then maximize $\text{tr}(U^T X L^P X^T U)$. It is easy to check that the above approach with such a constraint in fact results in the traditional MFA [136]. The only difference is that Eq.(4.18) involves a constraint optimization problem, whereas the traditional MFA solves an unconstrained optimization problem. The motivation for using the constraint $u_k^T u_k = 1, k = 1, 2, \dots, l$ is to avoid calculating the inverse of $X L^I X^T$, which leads to the potential singular problem. In order to solve the above constraint optimization problem, we introduce a Lagrangian

$$L(u_k, \lambda_k) = \sum_{k=1}^l u_k^T X (L^P - L^I) X^T u_k - \lambda_k (u_k^T u_k - 1), \quad (4.19)$$

with the multipliers λ_k 's. The Lagrangian $L(u_k, \lambda_k)$ should be maximized with respect to both λ_k and u_k . The condition is that, at the stationary point, the derivatives of $L(u_k, \lambda_k)$ respect to u_k must vanish, i.e.,

$$\frac{\partial L(u_k, \lambda_k)}{\partial u_k} = (X (L^P - L^I) X^T - \lambda_k I) u_k = 0, k = 1, 2, \dots, l, \quad (4.20)$$

and therefore,

$$X (L^P - L^I) X^T u_k = \lambda_k u_k, k = 1, 2, \dots, l, \quad (4.21)$$

which means that the λ_k 's are the eigenvalues of $X (L^P - L^I) X^T$, and u_k 's are the corresponding eigenvectors. Thus, we have

$$J(U) = \sum_{k=1}^l u_k^T X (L^P - L^I) X^T u_k = \sum_{k=1}^l \lambda_k u_k^T u_k = \sum_{k=1}^l \lambda_k. \quad (4.22)$$

Figure 4.2: (a) red dots are the positive feedback samples and blue dots are the negative feedback samples in the original space, (b) the BMMA method, and (c) the positive and negative feedback samples in the maximum margin subspace.

Therefore, the objective function is maximized when U is composed of the largest eigenvectors of $X(L^P - L^I)X^T$. By imposing constraint $u_k^T u_k = 1, k = 1, 2, \dots, l$, we need not calculate the inverse of $XL^I X^T$, and this allows us to avoid the singular problem easily.

The BMMA method can be illustrated in Figure 4.2.

In the previous subsection, we have formulated the BMMA method and shown that the optimal projection matrix can be obtained by a standard eigenvalue decomposition on a matrix. Then, the problem is how to determine the optimal dimension for RF, i.e., the projected subspace. To achieve such a goal, we give the details of determining the optimal dimension.

In general,

$$\max_U \text{tr}(U^T X(L^P - L^I)X^T U) = \sum_{i=1}^l \lambda_i, \quad (4.23)$$

where λ_i 's are the associated eigenvalues, and we have

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{d-1} \geq 0 \geq \lambda_d \geq \dots \geq \lambda_l. \quad (4.24)$$

To maximize the margin between the positive feedback samples and the negative feedback samples, we should preserve all the eigenvectors associated with the positive

eigenvalues. However, as indicated in [149], for image retrieval the orthogonal complement components are essential to capture the same concept shared by all positive feedback samples. Based on this observation, we should also preserve the components associated with zero eigenvalues although they do not contribute to maximize the margin. This technique can effectively preserve more geometric properties of feedback samples in the original high-dimensional feature space. Therefore, the optimal dimension of the projected subspace just corresponds to the number of nonnegative eigenvalues of the matrix. Therefore, compared with the original formulation of the graph embedding framework in [136], the new formulation Eq.(4.18) can easily avoid the intrinsic singular problem and also provide us with a simple way to determine the optimal dimension for this problem.

Biased discriminant analysis (BDA) and its kernel version BiasMap were first proposed to address the asymmetry property of training data in interactive image retrieval. However, to use BDA, the singular problem and the Gaussian assumption for the positive feedback samples are two major challenges. While the kernel method BiasMap cannot exert its normal capability since the feature dimension is much higher than the number of training samples. Additionally, it is still problematic to determine the optimal dimension of BDA and BiasMap for CBIR. Different from the original BDA, our BMMA method is a local discriminant analysis method, which does not make any assumption on the distribution of the data. Biased towards the positive feedback samples, maximizing the objective function in the projected space can push the nearest negative feedback samples away from the positive feedback samples while pulling the nearest positive feedback samples towards the positive feedback samples. Therefore, the definition in Eq.(4.18) can maximize the overall average margin between the positive feedback samples and the negative feedback samples. In such a way, each sample in the original space is mapped onto a low-dimensional local maximum margin subspace in accordance with human perception of the image content.

Since the graph embedding technique is an effective way to capture the intrinsic geometric structure in the original feature space, we propose a way to incorporate the unlabeled samples based on the intrinsic graph, which is helpful in capturing the geometric structure of unlabeled samples and alleviating the overfitting problem. In the following, we design a regularization term based on intrinsic graph for the unlabeled samples in the image database.

Figure 4.3: The illustration of the optimal hyperplane comparison between BMMA SVM and SemiBMMA SVM for two classes of feedback samples.

For each unlabeled sample $x_i (n^+ + n^- + 1 \leq i \leq n)$, we expect that the nearest unlabeled samples are likely to have the similar low-dimensional representations. Specifically, for each unlabeled sample, we find its k_1 nearest neighborhood unlabeled samples, which can be represented as a sample set N_i^u , and put edges between each unlabeled sample and its k_1 nearest unlabeled samples. Then, the intrinsic graph for the unlabeled samples is characterized as

$$\begin{aligned} S_U &= \frac{1}{2} \sum_i \sum_{j: j \in N_i^s \text{ or } i \in N_j^s} \|U^T x_i - U^T x_j\|^2 W_{ij}^u \\ &= \text{tr} (U^T X (D^u - W^u) X^T U) \\ &= \text{tr} (U^T X L^U X^T U), \end{aligned} \quad (4.25)$$

$$W_{ij}^u = \begin{cases} \frac{1}{|N|^n} \exp(-\|x_i - x_j\|^2 / \delta^2), & \text{if } l(i) = l(j) = 0, i \in N_j^u \text{ or } j \in N_i^u, \\ 0, & \text{else,} \end{cases} \quad (4.26)$$

which reflects the affinity of sample pairs; D^u is a diagonal matrix whose diagonal elements are calculated by $D_{ii}^u = \sum_j W_{ij}^u$; $|D^u|$ denotes the total number of nearest unlabeled

sample pairs for each unlabeled sample; and $L^U = D^u - W^u$ can be known as a Laplacian matrix. Hence, we call this term as a Laplacian regularizer.

The motivation for introducing this term is largely inspired by the regularization principle, which is the key to enhance the generalization and the robust performance of machine learning approaches in practical applications. There are a lot of possible ways to choose a regularizer for the proposed BMMA. In this work we chose the Laplacian regularizer, which is largely inspired by the recently emerging manifold learning community. Actually, this scheme can preserve weakly (probably correct) similar relationship between all unlabeled sample pairs, and thus effectively integrate the similarity information of unlabeled samples into BMMA. By integrating the Laplacian regularizer into the supervised BMMA, we can easily obtain SemiBMMA for SVM-based RF, i.e.,

$$U^* = \arg \max_U \text{tr} (U^T X (L^P - L^I - \beta L^U) X^T U), \quad (4.27)$$

where β is used to trade off the contributions of labeled samples and unlabeled samples. Similarly, the solution of Eq.(4.27) is obtained by conducting a standard eigenvalue decomposition and is calculated as a set of eigenvectors.

The difference between BMMA SVM and SemiBMMA SVM for two classes of feedback samples can be shown in Figure 4.3. The SemiBMMA method is illustrated in Table 4.1.

Then, all unlabeled samples in the database are projected onto this maximum margin subspace. After the projection, the traditional SVM-based RF is conducted on the new representations. Finally, similar to the traditional SVM-based RF, we can measure the degree of similarity through the output of SVM, i.e., $|f(x)|$.

4.4 Content-Based Image Retrieval

In experiments, we use a subset of the Corel Photo Gallery as the test data to evaluate the performance of the proposed scheme. The original Corel Photo Gallery includes plenty of semantic categories, each of which contains 100 or more images. However, some of the categories are not suitable for image retrieval, since some images with different concepts are in the same category and many images with the same concept are in different

Table 4.1: Semi-supervised Biased Maximum Margin Analysis

Input: $X = \{x_1, x_2, \dots, x_n\} \in R^h$ indicate all feedback samples and unlabeled samples, which include the positive feedback samples set X^+ , the negative feedback samples set X^- , and the unlabeled samples.
Step 1: Construct the supervised intrinsic graph G according to Eq.(4.14), and calculate the matrix value XL^IX^T .
Step 2: Construct the supervised penalty graph G^p according to Eq.(4.16), and calculate the matrix value XL^PX^T .
Step 3: Construct the Laplacian regularizer according to Eq.(4.26), and calculate the matrix value XL^UX^T .
Step 4: Calculate the projection matrix U^* according to an eigenvalue decomposition on the matrix $X(L^P - L^I - \beta L^U)X^T$.
Step 5: Calculate the new representations: project all positive feedback samples, negative feedback samples, and remaining samples in the database onto the reduced subspace, respectively, i.e., $Y^+ = U^{*T}X^+$ and $Y^- = U^{*T}X^-$.
Output: The new representations of the positive and negative feedback samples, i.e., Y^+ and Y^- , in the reduced subspace.

categories. Therefore, the existing categories in the original Corel Photo Gallery are ignored and reorganized into 80 conceptual classes based on the ground truth, e.g., lion, castle, aviation, train, dog, autumn, cloud, tiger, etc. Finally, the test database comprises totally 10,763 real-world images.

Given a query image by the user, the CBIR system is expected to return more semantically relevant images after each round of RF [3]. However, at each round of RF, the number of relevant images is usually very small because of the semantic gap. At the same time, the user would not like to label a large number of samples in RF. The user also expects to obtain more relevant images within only a few rounds of RF. Keeping the size of labeled relevant images small and the rounds of RF few are two key issues in designing an image retrieval system. Therefore, we devise the following CBIR framework accordingly to evaluate the RF methods.

From the flowchart in Figure 4.4, we can notice that when a query image is provided by the user, the system first extracts the low-level visual features. Then, all images in the database are sorted based on a similarity metric, i.e., Euclidean distance metric. If

Figure 4.4: The flowchart of our CBIR system.

the user is satisfied with the returned results, the image retrieval process is ended, and the results are presented to the user. However, because of the semantic gap, most of the time, the user is not satisfied with the first round of returned results. Then, she/he will label the most semantically relevant images as the positive feedback samples in top returned results. All of the remaining images in top returned results are automatically labeled by the system as the negative feedback samples. Based on the small-sized positive and negative feedback samples, an RF model associated with a new similarity metric can be trained based on various existing techniques. All images in the database are resorted based on the new similarity metric. After each round of RF, the user will check whether the returned results are satisfied. If the user is satisfied with the results, then the image retrieval process is ended; otherwise, the RF procedure repeats until the user is satisfied with the returned results.

In general, the image representation is a key problem in CBIR. The images are usually represented by low-level visual features, e.g., color [10–12], texture [12–14], and shape [15–17], each of which can capture the content of an image to some extent.

For color, we first extract three moments: color mean, color variance, and color skewness in each color channel (L, U, and V), respectively. Thus, a 9-dimensional color moment is employed in our experiments. Then, a 256-dimensional HSV color histogram is calculated. Both the hue and the saturation are quantized into eight bins and the values

are quantized into four bins. These two kinds of color features are used to represent the color information.

Compared with the classical global texture descriptors (e.g., Gabor features and wavelet features), the local dense features show good performance in describing the content of an image. The Weber's law descriptors (WLD) [146] are adopted as the texture descriptors, which are mainly based on the mechanism of human perception on a pattern. The WLD result in a feature vector of 240 values.

We employ the edge directional histogram [15] from the Y component in the YCrCb space to capture the spatial distribution of the edges. The edge direction histogram is quantized into five categories including horizontal, 45° diagonal, vertical, 135° diagonal, and isotropic directions to represent the edge information.

Generally, these features are combined into a feature vector, which will result in a vector with 510 values. Then, all feature components are normalized to normal distributions with zero mean and one standard deviation to represent the images.

4.5 Experimental Results

4.5.1 Intrinsic Problems in Traditional SVM-based RF

An image is usually represented by a high-dimensional feature vector in CBIR. However, one key issue in RF is that which subset of the features can reflect the basic property of the feedback samples in different classes and thus benefits the construction of the optimal classifier. This problem can be illustrated from some real-world data in RF. There are five positive feedback samples and five negative feedback samples. We randomly select two features to construct the optimal hyperplane of SVM for three times. As shown in Figure 4.5, we can see that the resulting optimal hyperplane of SVM is diverse with different combinations of the features.

It is essential to obtain a satisfactory classifier when the number of the available feedback samples is small, which is always the case in RF, especially at the first a few rounds of RF. Therefore, we first show a simple example to simulate the unstable problem of SVM when dealing with a small number of training feedback samples. The open circles in Figure 4.6 indicate the positive feedback samples and the plus points indicate the

negative feedback samples in RF. Figure 4.6 (a) shows the optimal hyperplane, which is trained by the original training samples. Figure 4.6 (b) and (c) show different optimal hyperplanes, which are trained by the original training samples with only one and two incremental positive feedback samples, respectively. From Figure 4.6, we can see that the optimal hyperplane of SVM changes sharply when a new incremental sample is integrated into the original training set. Additionally, we can also notice that the optimal hyperplane of SVM is very complex when the training feedback samples have a complicated distribution.

Note that similar results have been indicated in previous research [36]. However, different from [36], in this subsection, we have shown some other problems in the traditional SVM-based RF, that is, the distinct property of feedback samples in RF and the unstable and complex optimal hyperplane of the traditional SVM at the first a few rounds of RF.

4.5.2 Feature Extraction Based on Different Methods

Six experiments are conducted to compare BMMA with LDA, BDA, and MFA, in finding the most discriminative directions. We plot the directions, which correspond to the largest eigenvalue of the decomposed matrices of LDA, BDA, MFA, and BMMA, respectively. From these examples, we can clearly notice that LDA can find the best discriminative direction when the samples from each class are distributed as a Gaussian distribution with similar covariance matrices, as shown in Figure 4.7 (a) and (d), but it may confuse when the data distribution is more complicated, as given in Figure 4.7 (b), (c), (e), and (f). Biased towards the positive feedback samples, BDA can find the

Figure 4.5: The optimal hyperplane of SVM is diverse for different combinations of the features.

Figure 4.6: The optimal hyperplane of SVM is unstable and complex when dealing with a small number of training feedback samples.

direction that the positive feedback samples are well separated with the negative feedback samples when the positive feedback samples have a Gaussian distribution, but it may also confuse when the distribution of the positive feedback samples is more complicated. For instance, in Figure 4.7 (b), BDA can find the direction to distinguish the positive feedback samples from the negative ones. However, in Figure 4.7 (c), (e), and (f), when the positive feedback samples have a more complicated distribution, BDA obviously fails. MFA can also find the discriminative direction when the distribution of the negative feedback samples is simple, as shown in Figure 4.7 (a), (b), and (c). But when the negative feedback samples pose a more complicated distribution, MFA will fail as in Figure 4.7 (d), (e), and (f). Biased towards the positive feedback samples, BMMA can find the most discriminative direction in the six experiments based on local analysis, since it does not make any assumptions for the distributions of the positive and negative feedback samples. It should be noted that BMMA is a linear method, and therefore, we only give the compared results of related linear methods.

4.5.3 Statistical Experimental Results

In this subsection, we evaluate the performance of the proposed methods on a real-world image database. We use average precision (AP) and standard deviation (SD) to evaluate the performance of the compared algorithms. AP refers to the percentage of relevant images in top returned results presented to the user, and is calculated as the averaged values for all query images. SD is computed from the SD value of APs for all query images, and describes the stability of different algorithms. AP evaluates the

Figure 4.7: Performance comparisons of four different feature extraction methods, i.e., LDA, BDA, MFA, and BMMA, for two classes of feedback samples, i.e., positive feedback samples and negative feedback samples, in RF. In experiments, red “o” and blue “ Δ ” denote the positive and negative feedback samples, respectively. Black dotted lines, blue dot dash lines, green dot dash lines, and red full lines indicate LDA, BDA, MFA, and BMMA, respectively. (a)-(f) show the experimental results of the four feature extraction methods when handling the feedback samples with various different distributions, respectively.

effectiveness of a given algorithm, and the corresponding SD evaluates the robustness of the algorithm. AP is the major evaluation criterion, which evaluates the effectiveness of the compared algorithms. We empirically select the parameters $k_1 = k_2 = 4$ according to the manifold learning approaches. Considering the computable efficiency, we randomly select 300 unlabeled samples at each round of RF. For the trade-off parameter between the labeled and unlabeled samples, we simply set $\beta = 1$. For all SVM based methods, the kernel parameters can be obtained by the grid search methods. In this work, we choose the Gaussian kernel and the kernel parameters are selected optimally, i.e.,

$$K(x_i, x_j) = e^{-\rho|x_i - x_j|^2}, \rho = 0.001. \quad (4.28)$$

Note that, the kernel parameters and the kernel type can significantly affect the performance of the system. For different image databases, we should tune the kernel parameters and the kernel type carefully. In experiments, we determine the kernel parameters from a series of values according to the performance. Moreover, much better performance can be achieved by tuning the kernel parameters further for different query images.

4.5.3.1 Experiments on a Small-Scale Image Database

To show the effectiveness of the proposed BMMA combined with the traditional SVM-based RF in dealing with the asymmetry property of training data, the first experiment is conducted on a small-scale image database, which includes 3,899 images with 30 different categories. We use all 3,899 images in the 30 categories as the query images. Some example categories used in this experiment are shown in Figure 4.8. To avoid the potential problem caused by the asymmetry amount of the positive and negative feedback samples [36], we select an equal number of positive and negative feedback samples in this experiment. In practice, the first 5 query relevant images and first 5 query irrelevant images in top 20 returned results in the previous round of RF are automatically labeled as the positive and negative feedback samples, respectively.

In [149], OCCA was proposed to only analyze the positive feedback samples for the traditional SVM-based RF in image retrieval. Hence, we compare the performance of BMMA combined with the traditional SVM (BMMA SVM), OCCA combined with the

Figure 4.8: Example categories used in the small-scale image database.

traditional SVM (OCCA SVM), and the traditional SVM in this subsection. In real-world applications, it is not practical to require the user to label a large number of samples in RF. Therefore, a small number of training feedback samples will cause the severe unstable problem in the traditional SVM-based RF (as shown in Section 4.5.1). Figure 4.9 shows the APs in top 20 results after the second round of RF for all 30 categories. The baseline describes the initial retrieved results without any RF information. Especially, at the beginning, the Euclidean distance metric in the high-dimensional feature space is used to measure the similarity between the query image and the images in the database. After the user provides the RF information, the traditional SVM, BMMA SVM, and OCCA SVM methods are then applied to rank the images in the database. As shown in Figure 4.9, the performance of the three methods varies with different categories. For some easy categories, all of the three methods can perform well (e.g., Categories 2 and 4). For some hard categories, all of the three methods perform poorly (e.g., Categories 18, 20, and 24). After the second round of RF, all of the three methods are significantly better than the baseline, and this indicates that the RF information provided by the user are very helpful in improving the performance of the system.

Figure 4.10 shows the AP scope curves of the compared algorithms after the first and second rounds of RF. We can notice that both the BMMA SVM and the OCCA SVM can perform much better the traditional SVM-based RF on the entire scope, especially after the first round of RF. The main reason is that after the first round of RF, the number of training samples is especially small (e.g., usually 8-10 samples totally), and this will make the traditional SVM-based RF perform extremely poor. The BMMA SVM method and the OCCA SVM method can significantly improve the performance of the traditional SVM-based RF by treating the positive and negative feedback samples

Figure 4.9: APs with 30 categories in top 20 results of the compared algorithms, i.e., SVM, OCCA SVM, and BMMA SVM, after the second round of RF.

unequally. Therefore, we can conclude that this scheme, which asymmetrically treats the feedback samples (i.e., biased towards the positive feedback samples), can significantly improve the performance of the traditional SVM-based RF, which treats the two different classes of feedback samples equally. As shown in Figure 4.10 (b), with the number of RF iterations increasing, the performance difference between the enhanced methods and the traditional SVM-based RF gets small. Generally, by iteratively adding the user's RF information, more feedback samples will be obtained as the training data, and this will make the performance of the traditional SVM-based RF much more stable. Meanwhile, the dimensions of BMMA and OCCA decrease with the increasing of the positive feedback samples. Consequently, the performance of BMMA SVM and OCCA SVM will be degraded by the overfitting problem. Therefore, the performance difference between the enhanced methods and the traditional SVM-based RF gets small. However, the performance after the first a few rounds of RF is usually most important, since the user would not like to provide more rounds of RF. At the first a few rounds of RF, the classifier trained based on few labeled training data is not reliable, but its performance can be improved when more samples are labeled by the user at the subsequent rounds of RF. Both BMMA and OCCA can significantly improve the performance of the traditional

Figure 4.10: APs of the compared algorithms, i.e., SVM, OCCA SVM, and BMMA SVM, after the first and second rounds of RF. (a) the first round of RF and (b) the second round of RF.

SVM-based RF after the first 2 rounds of RF. Therefore, we can conclude that BMMA can effectively integrate the asymmetric property of training data into the image retrieval process and thus enhances the performance of the system.

4.5.3.2 Experiments on a Large-Scale Image Database

In this part, we first design a scheme to model the real-world image retrieval process. In a real-world image retrieval system, a query image is usually not in the database. To simulate such an environment, we use a fivefold cross validation database to evaluate the compared algorithms. More precisely, we divide the whole image database into five subsets with an equal size. Therefore, there are 20 percent images per category in each subset. At each run of cross validation, one subset is selected as the query set, and the other four subsets are used as the database for image retrieval. Then, 400 query images are randomly selected from the query set and RF is automatically conducted by the system. For each query image, the system retrieves and ranks the images in the database. Finally, 9 rounds of RF automatically carried out by the system in this experiment.

At each round of RF, top 20 returned results are selected from the image database and labeled as the positive and negative feedback samples serially. Generally, in a real-world

image retrieval system, the irrelevant images usually largely outnumber the relevant ones. To simulate such a case in the system, the first 3 query relevant images are labeled as the positive feedback samples and all other query irrelevant images in top 20 returned results are automatically labeled as the negative feedback samples. Note that, the images which have been selected at previous rounds of RF are excluded from later selections. The experimental results are shown in Figure 4.11 and Figure 4.12. APs and SDs are computed from the fivefold cross validation.

To show the effectiveness of the proposed scheme, we compare them with the traditional SVM-based RF, OCCA SVM, BDA SVM, and OneSVM. The traditional SVM treats the RF as a strict two-class classification problem with an equal treatment on both the positive and negative feedback samples. OCCA tries to find a subspace, in which all positive feedback samples are merged, and then the traditional SVM is implemented to conduct the image retrieval procedure. For BDA, we select all the eigenvectors corresponding to the eigenvalues larger than 1 percent of the maximum eigenvalue, and then the traditional SVM are used to classify the query relevant and irrelevant images, which is a common way to select the dimension of the feature space. OneSVM assumes the RF procedure as a one-class classification problem and estimates the distribution of the target images in the feature space.

Figure 4.11 and Figure 4.12 show AP and SD curves of the compared algorithms, respectively. SemiBMMA SVM outperforms all other compared algorithms on the entire scope. Both BMMA and OCCA can improve the performance of the traditional SVM-based RF, as shown in Figure 4.11 (a), (b), (c), and (d). Compared with OCCA SVM, BMMA SVM performs much better for all top results, since BMMA takes both the positive and negative feedback samples into consideration. However, both BMMA and OCCA will encounter the overfitting problem, i.e., both of them combined with SVM-based RF will degrade the performance of the traditional SVM-based RF after a few rounds of RF although they can improve the performance of the traditional SVM at the first a few rounds of RF. As we can see from Figure 4.11 (d), (e), and (f), with the increase of rounds of RF, OCCA SVM performs poorly compared with the traditional SVM-based RF. At the same time, the performance difference between BMMA SVM and SVM gets smaller. SemiBMMA combined with SVM-based RF can significantly improve

Figure 4.11: APs of the compared algorithms, i.e., SemiBMMA SVM, BMMA SVM, OCCA SVM, BDA SVM, SVM, and One SVM, in top 10-60 results with fivefold cross validation. All the methods are evaluated over 9 rounds of RF. 0 round of RF refers to the retrieved results based on the Euclidean distance metric without any RF information.

the performance of SVM-based RF, since it can effectively utilize the asymmetric property of training data and integrate the information of unlabeled samples into the construction of the optimal classifier. For top 10 results, BDA SVM can achieve a better result than SVM-based RF. However, BDA still discards much useful information contained in the orthogonal complement components of the positive feedback samples. Therefore, BDA combined with SVM-based RF performs much worse than SVM-based RF, as shown in Figure 4.11 (b), (c), (d), (e), and (f). Although OneSVM tries to estimate the distribution of the target images in the feature space, it cannot work well without the help of the negative feedback samples.

Considering the stability of the compared algorithms, we can notice that SemiBMMA SVM and BMMA SVM perform best among all the algorithms in top 10, top 20, and top 30 results. Although OneSVM shows good stability in top 40, top 50, and top 60 results, its APs are too low.

We should indicate that the performance difference of the compared algorithms between experiments in Subsection 4.5.3.1 and experiments in this subsection is mainly caused by the different experimental setting. Because the numbers of the positive and negative feedback samples are equal in Subsection 4.5.3.1, while the negative feedback samples largely outnumber the positive feedback samples in Subsection 4.5.3.2. Additionally, the performance of SemiBMMA SVM does not perform better compared with that of BMMA SVM at the first 2 rounds of RF for most of the retrieved results. This is mainly because the maximum margin between the two different classes is essentially important when the number of training feedback samples is extremely small.

The detailed results of the compared algorithms after the ninth round of RF are given in Table 4.2. As can be seen, SemiBMMA combined with SVM-based RF can integrate all the available information into the RF procedure and achieves much better performance compared with other algorithms for all top results. BMMA SVM still obtains satisfactory performance compared with the traditional SVM-based RF and OCCA SVM. Therefore, we can conclude that the proposed BMMA and SemiBMMA combined with the traditional SVM-based RF have shown much better performance than the traditional SVM-based RF (i.e., directly using the SVM as an RF scheme) for CBIR.

Figure 4.12: SDs of the compared algorithms, i.e., SemiBMMA SVM, BMMA SVM, OCCA SVM, BDA SVM, SVM, and One SVM, in top 10-60 results with fivefold cross validation. All the methods are evaluated over 9 rounds of RF. 0 round of RF refers to the retrieved results based on the Euclidean distance metric without any RF information.

Table 4.2: APs and SDs in top N results of the compared algorithms, i.e., SemiBMMA SVM, BMMA SVM, OCCA SVM, BDA SVM, SVM, and OneSVM, after the ninth round of RF (APs $\pm$ SDs)

Algorithms	SemiBMMA SVM	BMMA SVM	OCCASVM	BDASVM	SVM	OneSVM
Top 10	0.9165 $\pm$ 0.1505	0.8914 $\pm$ 0.1663	0.8658 $\pm$ 0.2162	0.8747 $\pm$ 0.1638	0.8265 $\pm$ 0.3011	0.5122 $\pm$ 0.3795
Top 20	0.8727 $\pm$ 0.2008	0.8363 $\pm$ 0.2213	0.8154 $\pm$ 0.2622	0.7612 $\pm$ 0.2520	0.7856 $\pm$ 0.3138	0.3428 $\pm$ 0.3149
Top 30	0.7840 $\pm$ 0.2319	0.7452 $\pm$ 0.2502	0.7151 $\pm$ 0.2767	0.6339 $\pm$ 0.2722	0.6996 $\pm$ 0.3134	0.2878 $\pm$ 0.2797
Top 40	0.7058 $\pm$ 0.2491	0.6645 $\pm$ 0.2642	0.6315 $\pm$ 0.2811	0.5465 $\pm$ 0.2712	0.6290 $\pm$ 0.3088	0.2546 $\pm$ 0.2588
Top 50	0.6396 $\pm$ 0.2542	0.5982 $\pm$ 0.2687	0.5638 $\pm$ 0.2761	0.4829 $\pm$ 0.2631	0.5706 $\pm$ 0.2993	0.2331 $\pm$ 0.2404
Top 60	0.5854 $\pm$ 0.2548	0.5431 $\pm$ 0.2640	0.5078 $\pm$ 0.2660	0.4319 $\pm$ 0.2497	0.5216 $\pm$ 0.2847	0.2147 $\pm$ 0.2255
Top 70	0.5416 $\pm$ 0.2523	0.4964 $\pm$ 0.2549	0.4629 $\pm$ 0.2544	0.3925 $\pm$ 0.2362	0.4806 $\pm$ 0.2783	0.2000 $\pm$ 0.2115
Top 80	0.5025 $\pm$ 0.2460	0.4564 $\pm$ 0.2430	0.4246 $\pm$ 0.2407	0.3596 $\pm$ 0.2229	0.4462 $\pm$ 0.2687	0.1886 $\pm$ 0.2000
Top 90	0.4865 $\pm$ 0.2387	0.4224 $\pm$ 0.2298	0.3941 $\pm$ 0.2287	0.3320 $\pm$ 0.2102	0.4166 $\pm$ 0.2581	0.1789 $\pm$ 0.1890

4.5.3.3 Visualization of Image Retrieval Results

In the previous subsections, we have presented some statistical quantitative results of the proposed scheme. In this subsection, we show the visualization of image retrieval results. In this experiment, we randomly select some images (e.g., bobsled, cloud, cat, and car) as the query images, and perform the RF procedure based on the ground truth. For each query image, we do 4 rounds of RF. At each round of RF, we randomly select some query relevant and irrelevant images as the positive and negative feedback samples from the first screen, which contains 20 images totally. The numbers of selected positive and negative feedback samples are about 4. We choose them according to the ground truth of the images, i.e., whether they share the same concept with the query image or not. Figure 4.13 shows the experimental results. The query images are given as the first image of each row. We show top 1 to top 10 images of initial results without any RF information and SemiBMMA SVM after 4 rounds of RF, respectively. And the incorrect results are highlighted by green boxes. From the results, we can notice that our proposed scheme can significantly improve the performance of the system. For the first, second, and fourth query images, our system produces 10 relevant images out of top 10 results. For the third query image, our system produces 9 relevant images out of top 10 results. Therefore, SemiBMMA SVM can effectively detect the homogeneous concept shared by all positive feedback samples, and hence improve the performance of the system.

Figure 4.13: The performance of the proposed scheme for four different query images, i.e., bobsled, cloud, cat, and car, after the fourth round of RF.

4.6 Summary

Regarding the positive and negative feedback samples as two different classes, RF can be considered as an online binary classification problem. However, the conventional classification based RF, i.e., SVM-based RF, can only treat the positive and negative feedback samples equally although this is not appropriate since all positive feedback samples share a common concept while each negative feedback sample differs in diverse concepts. With the observation that different semantic concepts live in different subspaces and each image can live in many different semantic concept subspaces, a biased maximum margin analysis (BMMA) method is introduced to separate the positive feedback samples from the negative ones by a maximum margin in the reduced subspace. The traditional SVM-based RF combined with BMMA can better model the RF procedure and reduce the performance degradation caused by the asymmetric property of training data. Moreover, semi-supervised BMMA (SemiBMMA) can effectively integrate the information of unlabeled samples by introducing a Laplacian regularizer to BMMA and thus effectively alleviate the overfitting problem caused by a small number of training feedback samples. Extensive experiments on both synthetic datasets and a real-world image database have been conducted to show the proposed scheme combined with the traditional SVM-based RF can significantly improve the performance of the CBIR system.

Chapter 5

Geometric Optimum Experimental Design

5.1 Motivation

A variety of conventional relevance feedback (RF) methods [22] have been designed to involve the user in the system to reduce the semantic gap between low-level visual features and high-level semantic concepts in content-based image retrieval (CBIR) [35–37, 75–79, 81–85, 90]. Despite the success, it is boring and tedious for the user to be asked to label a large number of samples in RF. To reduce the labeling efforts of the user, collaborative image retrieval (CIR) attempts to accelerate this procedure by leveraging various auxiliary information. During the past a few years, vast research work has been conducted regarding the CIR study [24–27, 34, 38, 97, 98]. Some of the research tries to select a set of most informative samples from the database, which should be labeled by the user in RF and used as the training data to define an effective similarity metric for image retrieval [34, 38, 97, 98]. Support vector machine active learning (SVMactive) is one of the most popular methods in this category [34], which has been widely used to identify the ambiguous samples as the most informative ones by leveraging the optimal hyperplane of SVM. This technique always requires the user to label the unlabeled samples, which are closest to the optimal hyperplane of SVM for their low predication confidence.

To explain the mechanism of SVMactive, a simple toy example is illustrated in Figure 5.1. There are two labeled samples (i.e., the red solid dot “●” for the positive feedback

sample while the green cross point “ $\times$ ” for the negative feedback sample) and several unlabeled ones (i.e., open circles “ $\circ$ ”). The six samples distribute along a line and the coordinates on the horizontal axis indicate their positions. By using SVM, the optimal hyperplane of the classifier f , which separates the two labeled feedback samples with a maximum margin, crosses position 0 as shown in Figure 5.1 with the dashed line. According to the most ambiguous criterion, i.e., the samples closest to f having the maximum ambiguity, we can get that “A” and “B” have the maximum and identical ambiguity because they have the same distances, i.e., 0.5 for both, to the optimal hyperplane. Therefore, “A” and “B” should be labeled by the user and used as the training data in RF. If we can choose only one sample for labeling, it is more reasonable to label “B” than “A” since more unlabeled samples are distributed around “B” and thus “B” is more effective than “A” to represent the distribution of unlabeled samples in the database. However, SVMactive can only select the ambiguous samples for the user to label although labeling representative ones may bring more useful information for achieving much better performance. Moreover, the optimal hyperplane of SVM is always unstable with small-sized training data [36, 153], i.e., this hyperplane is always sensitive when the size of the training feedback samples is small. Generally, in RF, the user would only label a small number of samples and cannot label each sample accurately all the time. Therefore, the optimal hyperplane of SVM cannot always be accurate with insufficient and inexactlly labeled feedback samples. During the last decade, a large number of studies have been carried out to further enhance the performance of SVMactive for image retrieval [27, 38, 97, 98]. Nevertheless, these SVMactive methods always require an initial classifier model to identify the most informative samples and thus cannot be directly applied when there are no training data, which is however not appropriate for CIR [27, 34, 38, 97, 98].

In the machine learning community, there has been a long tradition of research on active learning [70–74]. In general, discriminative models aim to choose the most ambiguous samples, e.g., SVMactive methods [34], and generative models tend to select the most representative samples [73]. In statistics, active learning can be typically referred to as experimental design. The sample is referred to as experiment, and its label is referred to as measurement. The study of optimum experimental design (OED) [45] is concerned

Figure 5.1: A toy example to show the mechanism of SVMactive in selecting the most informative unlabeled samples for the user to label. The red solid dot, green cross point, and open circles indicate the positive feedback sample, negative feedback sample, and unlabeled samples, respectively. SVMactive can select the ambiguous unlabeled samples “A” and “B” for their low prediction confidence but cannot select the more representative unlabeled sample “B”.

with the selection of the most informative experiments to measure given that conducting an experiment is expensive. Classical OED methods include A-OED, D-OED, and E-OED, which are to maximize the confidence in a given model by minimizing some measure of the estimated parameter covariance [45]. However, classical OED methods cannot characterize the quality of predictions on test data if the test data are given beforehand. Transductive experimental design (TED) with different solutions [154, 155] was proposed to directly assess the quality of prediction on the given test data, which thus has a clear geometric interpretation to select the most representative samples in the database and has yielded impressive results for text categorization compared with classical OED methods (i.e., A-OED, D-OED, and E-OED). Nevertheless, most of the experimental design methods can only evaluate the labeled samples while ignore abundant unlabeled samples although the unlabeled samples can also provide useful information and further enhance the performance of image retrieval. In RF, more often than not, the effort of labeling samples is generally laborious and vast amounts of unlabeled samples are readily available in the database. Various semi-supervised learning approaches under such a scenario have been widely developed to improve the generalization ability of supervised learning by leveraging the geometric structure of unlabeled samples [44, 62, 65–67, 156, 157].

In addition, conventional experimental design studies can only select one sample after another with a greedy strategy [158] or involve a semidefinite procedure with high computational burden [159] when dealing with abundant unlabeled samples, which is however not appropriate and thus will significantly limit their potential applications to CIR.

In this chapter, we propose a novel active learning method, i.e., geometric optimum experimental design (GOED), by leveraging the geometric structure of unlabeled samples in the reproducing kernel Hilbert space (RKHS) to simultaneously select multiple representative samples in the database as the most informative ones for the user to label in image retrieval. The proposed scheme is largely inspired by the recent manifold regularization principles [44, 68], which are usually the key in semi-supervised learning to significantly improve the generalization ability of supervised learning. In general, the new method can be divided into three independent stages. It first learns a data-dependent kernel function from both the labeled and unlabeled samples, which can be constructed by a conventional kernel function and then warped by a data-dependent norm to reflect the geometric structure of unlabeled samples in RKHS, and thus alleviates the small-sized training data problem in RF [66]. By minimizing the expected average prediction variance on the test data as suggested by TED [154, 155], GOED has a clear geometric interpretation to select representative samples in the database as the most informative ones. Finally, this kernel function can be used to identify a set of the most informative samples iteratively with the global optimum for the user to label, and this is more effective and efficient for CIR. Compared with the popular SVMactive [27, 34, 38, 97, 98], our method is label-independent and thus can effectively avoid various potential problems caused by insufficient and inexact labeled feedback samples in SVMactive methods for image retrieval. Extensive experiments on both synthetic datasets and a real-world image database have demonstrated the advantages of the proposed GOED for CIR.

The rest of this chapter is organized as follows. Section 5.2 briefly reviews the related work. In Section 5.3, we propose the GOED method for CIR. A CIR system is introduced in Section 5.4. In Section 5.5, we first give the experimental results on both synthetic datasets and a real-world image database, and then show the analysis of the important parameter in GOED. Section 5.6 summarizes this chapter.

5.2 Related Work

To describe our method clearly, let us first review two areas of the research that are closely related to our work in this chapter, i.e., (1) selecting most informative samples and (2) OED.

5.2.1 Selecting Most Informative Samples

During the past decade, various conventional RF methods have been widely developed based on different assumptions for the positive and negative feedback samples. However, it is more appropriate for image retrieval to achieve satisfactory results within as few rounds of RF as possible since an online learning task is usually tedious and boring for the user [22, 32, 35–37, 75, 86, 87, 90, 160, 161].

Active learning is well-known for getting the necessary information by labeling as few samples as possible. SVMactive is one of the most popular techniques in this category for CIR, which has attracted much attention during the last decade [27, 34, 38, 97, 98]. However, SVMactive always suffers from some limitations when it is directly applied to the application of image retrieval since the user would not like to label a large number of samples in RF. To alleviate the small-sized training data problem, Wang et al. proposed to modify SVMactive with a transductive SVM by engaging unlabeled samples in the database [38]. In [97], Hoi et al. combined some semi-supervised learning techniques (i.e., Gaussian fields and harmonic functions [62]) with the traditional SVMactive, which can also effectively exploit the information of unlabeled samples [97]. To find multiple informative samples for the user to label in each iteration, Dagli et al. proposed to use a diversity measure based information theory approach [98] to select a set of diverse and informative samples simulatively. Moreover, in [27], a semi-supervised SVM batch mode active learning method was also developed to address the small-sized training data problem by leveraging the geometric structure of unlabeled samples in RKHS and identify the most informative samples via a min-max framework. Despite the vast research work during the past a few years, SVMactive methods always require an initial optimal hyperplane to identify the most informative samples for the user to label. However, this optimal hyperplane will not always be accurate with insufficient and inexactlly labeled feedback samples, which is always the case in image retrieval.

Besides the aforementioned methods, some other research efforts have also been devoted to select the most informative samples for the user to label. In [162], a batch model active learning framework was proposed to employ the Fisher information matrix as an ambiguous measurement to select the most informative samples, which is fundamentally based on a probabilistic framework of the kernel logistic regression model. The author of [158] employed an experimental design criterion to identify one sample after another with a greedy strategy, which does not have a clear interpretation to the selected samples and is usually not the optimal solution to select multiple informative samples for the user to label. Moreover, in [163], a dimension reduction technique was proposed to combine a structure information based data selection strategy to target the user intension and reduce the labeling efforts of the user for web image retrieval within a unified framework. In addition, some other effective techniques [164, 165] can also be employed to enhance the performance of the system when dealing with small-sized training data for objective recognition in the computer vision community, which is similar to the scenario in conventional RF.

5.2.2 Optimum Experimental Design

Given a set of unlabeled samples $X = \{x_1, \dots, x_n\}$ in R^h , active learning aims to find a subset $Z = \{z_1, \dots, z_l\} \subseteq X$ which contains the most informative samples. That is, if the samples $z_i (i = 1, \dots, l)$ are labeled and used as the training data, we can predict the labels of unlabeled samples more precisely.

In statistics, active learning is typically referred to as experimental design and considers a linear regression model

$$y = w^T x + \varepsilon, \tag{5.1}$$

where y is the observation; $w \in R^h$ is the parameter vector; $x \in R^h$ is the variable; and ε is an unknown error with zero mean and constant variance σ^2 . We define $f(x) = w^T x$ to be the prediction on an input variable x and the parameter vector w . Given a set of labeled samples $(z_1, y_1), \dots, (z_l, y_l)$, a popular way to learn the prediction function f is to estimate w by minimizing the following objective function, i.e.,

$$J(w) = \sum_{i=1}^l (w^T z_i - y_i)^2 + \gamma_1 \|w\|^2, \quad (5.2)$$

where $\gamma_1 > 0$; and $\|\cdot\|^2$ is the vector 2-norm. Let $Z = [z_1, \dots, z_l]$ and $y = [y_1, \dots, y_l]^T$. The optimal solution to Eq.(5.2) is given by

$$\hat{w} = (ZZ^T + \gamma_1 I)^{-1} Z y. \quad (5.3)$$

Note that the introduced regularization term can improve the numerical stability of the solution since $ZZ^T + \gamma_1 I$ is of full-rank. It has been shown that $\hat{w}$ is an unbiased estimation of w and its covariance matrix is given by $\sigma^2 C_w$, where C_w is the inverted Hessian of $J(w)$, i.e.,

$$C_w = \left(\frac{\partial^2 J(w)}{\partial w \partial w^T} \right)^{-1} = (ZZ^T + \gamma_1 I)^{-1}. \quad (5.4)$$

Conventional OED methods can select the most informative samples, i.e., $z_1, \dots, z_l$, by minimizing some measure of the estimated parameter covariance, i.e., Eq.(5.4). The typical criteria are trace of C_w (i.e., A-OED), determinant of C_w (i.e., D-OED), and maximum eigenvalue of C_w (i.e., E-OED) [45].

Conventional OED methods (i.e., A-OED, D-OED, and E-OED) do not have a clear interpretation to the selected informative samples and cannot characterize the quality of predictions on test data if the test data are given beforehand. To address this problem, TED aims to select representative samples in the database as the most informative ones by directly minimizing the prediction variance on the test data [154, 155]. For simplicity, we assume that the test data are given by $X = [x_1, \dots, x_n]$, and thus the average prediction variance on the test data X is given by

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n x_i^T \text{Cov}(\hat{w}) x_i \\ &= \frac{\sigma^2}{n} \sum_{i=1}^n x_i^T (ZZ^T + \gamma_1 I)^{-1} x_i \\ &= \frac{\sigma^2}{n} \text{tr}(X^T (ZZ^T + \gamma_1 I)^{-1} X). \end{aligned} \quad (5.5)$$

In order to minimize Eq.(5.5), we attempt to find a subset $Z \subseteq X$ which can minimize $\text{tr}(X^T (ZZ^T + \gamma_1 I)^{-1} X)$. However, it can be verified that this optimization problem is

NP-hard and thus infeasible to find the global optimum. In [154], after some mathematical derivations, the problem of Eq.(5.5) can be reformulated as

$$\min_{\alpha_i \in R^l} \sum_{i=1}^n \|x_i - Z\alpha_i\|^2 + \gamma_1 \|\alpha_i\|^2. \quad (5.6)$$

Eq.(5.6) aims to find the selection coefficients α_i for the most informative samples. The first term in Eq.(5.6) shows that the samples Z selected by TED are the most representative samples in the test data X . That is, the selected samples $z_i (i = 1, \dots, l)$ can be used to reconstruct the data X most precisely [154]. In [154], Yu et al. proposed a sequential greedy method to select one sample z_i at a time. However, the greedy solution to this problem is always suboptimal.

To obtain the global optimum, a convex relaxation of Eq.(5.6) was carefully designed in [155]. By introducing auxiliary variables $\beta = (\beta_1, \dots, \beta_n)$ as a data selection coefficient to control the inclusion of samples into the training data, the optimization problem of Eq.(5.6) can be reformulated as

$$\begin{aligned} \min_{\alpha_i, \beta \in R^n} \sum_{i=1}^n \left(\|x_i - X\alpha_i\|^2 + \sum_{j=1}^n \frac{\alpha_{i,j}^2}{\beta_j} \right) + \lambda \|\beta\|_1 \\ \text{s.t. } \beta_i \geq 0, i = 1, \dots, n, \end{aligned} \quad (5.7)$$

where $\alpha_i = (\alpha_{i,1}, \dots, \alpha_{i,n})^T$; λ is the sparse regularization parameter; and $\|\cdot\|_1$ denote the l_1 norm. As indicated by the well-known LASSO method [166], the minimization of the l_1 norm $\|\beta\|_1$ will lead to a sparse β . It is easy to check that, if $\beta_j = 0$, then all $\alpha_{1,j}, \dots, \alpha_{n,j}$ must be zero; otherwise the objective function goes to infinity. Therefore, the j th sample will not be selected. It has been verified that the optimization of Eq.(5.7) is convex, and thus the global optimal solution can be obtained. For more details, please refer to [155].

Conventional experimental design methods (i.e., A-OED, D-OED, E-OED, and TED) are designed based on supervised learning models (i.e., least-square ridge regression (LSRR)), which are usually not appropriate for CIR, since the effort of labeling samples is generally laborious and vast amounts of unlabeled samples are readily available in the database.

5.3 Geometric Optimum Experimental Design

In this section, we first formulate the conventional RF in image retrieval as an active learning problem, and then propose a novel active learning method, i.e., GOED, to select multiple representative samples in the database as the most informative ones for the user to label. GOED can alleviate the small-sized training data problem by leveraging the geometric structure of unlabeled samples in RKHS. Compared with the popular SVMactive, our method is label-independent and can effectively avoid various potential problems caused by insufficient and inexact feedback samples, and is more effective and useful for CIR.

5.3.1 Problem Definition

Let $\phi : R^h \rightarrow H_K$ denote a feature mapping function from an input space R^h to RKHS H_K , and $\phi(x)$ denotes the sample x in RKHS. Various RF methods have been developed based on different assumptions on the positive and negative feedback samples during the past a few years [35, 36, 75, 83, 86, 90]. However, how to deal with these feedback samples is still an open question in the current CBIR research. In this chapter, regarding the positive and negative feedback samples as two different classes, the conventional RF can be considered as an online binary classification problem and a classifier is expected to learn for image retrieval, which can thus be used as a similarity metric to measure the similarity between a given sample x and the query image via

$$y = \text{sign}(f(x)), \quad (5.8)$$

where the classifier is assumed to be $f(x) = w^T \phi(x)$ in this chapter. Here, if $y = 1$, then x belongs to the positive class; otherwise, it belongs to the negative class. It should be noted that a bias term can be incorporated into this form by expanding the weight and sample vector as $w \leftarrow [w, b]$ and $\phi(x) \leftarrow [\phi(x), 1]$. Generally, given a sample x , when $|f(x)|$ is large, the corresponding prediction confidence will be high. Meanwhile, a small value of $|f(x)|$ means that the sample x is close to the optimal hyperplane of f , and thus its corresponding prediction confidence will be low.

Suppose we have a set of labeled feedback samples $(z_1, y_1), \dots, (z_l, y_l)$, where y_i is the label of a sample z_i , a common principle for learning a good classifier f is to minimize the following structural risk in RKHS [69]

$$\hat{f} = \arg \min_{f \in H_K} \left(J(w) = \sum_{i=1}^l L(y_i, f(z_i)) + \gamma_1 \|f\|_K^2 \right), \quad (5.9)$$

where $L(\cdot, \cdot)$ is the loss function (e.g., hinge-loss in SVM and squared-loss in LSRR); $\gamma_1 > 0$; and $\|f\|_K$ is an induced norm in an appropriately chosen RKHS H_K defined over a kernel Gram matrix K [49].

In general, the generic problem of active learning for RF is the following. Given a set of unlabeled samples $X = \{x_1, \dots, x_n\} \in R^h$ in the database, we aim to find a subset $Z = \{z_1, \dots, z_l\} \subseteq X$, which contains the most informative ones for the user to label. In other words, the samples $z_i (i = 1, \dots, l)$ can improve the performance of image retrieval the most if they are labeled and used as the training data to acquire a more effective similarity metric for image retrieval.

With the observation that the closer to the optimal hyperplane of f a sample is, the lower its prediction confidence is, SVMactive methods have been widely designed to select the unlabeled samples closest to the optimal hyperplane of f as the most informative ones so as to achieve maximal refinement on the hyperplane between the two classes [27, 34, 38, 97, 98]. However, the optimal hyperplane of f is often sensitive to the training data when the size of training feedback samples is small, and this is always the case in image retrieval since the user would not like to label a large number of samples and also cannot label each sample accurately in RF all the time. Moreover, the aforementioned SVMactive methods always require an initial classifier to identify the most informative samples and thus cannot be directly applied to CIR when there are no training data.

5.3.2 Geometric Optimum Experimental Design

Conventional active learning methods (e.g., SVMactive and OED) are developed based on supervised learning models [34, 45, 154, 155], i.e., Eq.(5.9). Semi-supervised learning is often designed to significantly improve the generalization ability of supervised learning by leveraging abundant unlabeled samples in the database. The common motivation

of semi-supervised learning methods [44, 62, 64–68] is to exploit the intrinsic geometric structure of unlabeled samples by restricting the inductive prediction to comply with this geometry. The manifold regularization principle [44, 68], one of the most representative techniques, assumes that the geometry of the intrinsic data probability distribution is supported on a low-dimensional manifold. There are many different ways to approximate the geometric structure of unlabeled samples [40, 121–123]. A popular way is to compute the graph Laplacian in an unsupervised manner from the unlabeled samples by using the Laplacian eigenmap in the feature space [40, 123]. The manifold approximation term and supervised learning models are combined together under the conventional regularization framework [69], which can smoothen the model output along the manifold estimated from the unlabeled samples [44, 68].

To alleviate the small-sized training data problem in RF, in this subsection, we propose a novel active learning framework by leveraging the geometric structure of unlabeled samples in RKHS to simultaneously select multiple informative samples for the user to label. The proposed scheme is largely inspired by the recent manifold regularization principle in the machine learning community [44, 68], which is usually the key in semi-supervised learning to significantly improve supervised learning by exploiting the geometric structure of unlabeled samples, i.e.,

$$\hat{f} = \arg \min_{f \in H_K} \left(J(w) = \sum_{i=1}^l L(y_i, f(z_i)) + \gamma_1 \|f\|_K^2 + \gamma_2 \|f\|_I^2 \right), \quad (5.10)$$

where the additional $\|f\|_I^2$ is a smooth penalty term to reflect the intrinsic geometric structure of unlabeled samples. Parameters γ_1 and γ_2 are used to balance between the loss function and two regularization terms. The manifold regularization term $\|f\|_I^2$ plays a key role in semi-supervised learning and models the classifier output smoothness along the manifold estimated from both the labeled and unlabeled samples in the database [44, 68].

The proposed active learning framework shares a similar objective function with that of the manifold regularization framework in [44]. However, our method can effectively select the most informative samples for the user to label. Different from conventional active learning methods (i.e., SVMactive and OED), the new scheme can effectively

alleviate the small-sized training data problem by exploiting the geometric structure of abundant unlabeled samples as in various semi-supervised learning methods [44, 62, 65–67, 156, 157]. In CIR, the system can effectively select the most informative samples for the user to label, which is an active learning problem. And then, if the most informative samples are labeled by the user, the system can use both the labeled and unlabeled samples to learn a classifier f and thus to measure the similarity between a given sample and the query image; this is actually a semi-supervised learning problem.

Given the above active learning framework for CIR, the key issue to attack this problem is first to design an appropriate RKHS induced by a kernel Gram matrix K , and thereafter find an effective way to identify the most informative samples for the user to label. In the following, we will study to define the kernel Gram matrix K and also to identify the most informative samples $z_i (i = 1, \dots, l)$ for the user to label.

Kernel trick is widely applied in the hope of discovering the nonlinear structure of the data by mapping the original data into a higher dimensional RKHS [167]. The popular kernel functions include Gaussian and Polynomial ones. However, the nonlinear structure captured by a data-independent kernel function may not be consistent with the intrinsic geometric structure of the data [44, 66]. To capture the geometric structure of the data, in this work, we employ a data-dependent kernel function, which can be constructed by a conventional kernel function (e.g., Gaussian or Polynomial ones) from both the labeled and unlabeled samples with a kernel deformation principle [66]. For clarity, below we briefly review the kernel deformation principle introduced in [66].

Let H_K denote the original RKHS induced by a kernel Gram matrix K , and $\tilde{H}_{\tilde{K}}$ denote the new deformed RKHS induced by a new kernel Gram matrix $\tilde{K}$. In [66], the authors assume the following relationship between the two spaces, i.e.,

$$\langle f, g \rangle_{\tilde{H}_{\tilde{K}}} = \langle f, g \rangle_{H_K} + \mathbf{f}^T M \mathbf{g}, \quad (5.11)$$

where $\mathbf{f} = (f(x_1), \dots, f(x_n))$ and $\mathbf{g} = (g(x_1), \dots, g(x_n))$ are evaluation functions on both the labeled and unlabeled samples X ; and M is used to capture the geometric structure of the data X . The deformation term, i.e., $\mathbf{f}^T M \mathbf{g}$, is introduced to assess the geometric relationship between the two functions f and g based on both the labeled and unlabeled

samples. According to [66], a new kernel function $\tilde{k}(\cdot, \cdot)$ can thus be derived associated with a new deformed RKHS $\tilde{H}_{\tilde{K}}$ by

$$\tilde{k}(x_i, x_j) = k(x_i, x_j) - \gamma \mathbf{k}_{x_i}^T (I + MK)^{-1} M \mathbf{k}_{x_j}, \quad (5.12)$$

where $k(\cdot, \cdot)$ is the original kernel function (e.g., Gaussian or Polynomial ones) defined in H_K with its corresponding kernel Gram matrix $K = [k(x_i, x_j)]_{n \times n}$; $\mathbf{k}_{x_i}$ is defined as $\mathbf{k}_{x_i} = [k(x_i, x_1), \dots, k(x_i, x_n)]^T$; γ is the kernel deformation parameter and used to balance the original kernel function and the deformation term; and I is an identity matrix. The key issue now is the choice of M , which should be designed with respect to the intrinsic geometric structure of the data X .

To capture the geometric structure of the data, we can turn to the graph Laplacian L as suggested by [44, 66]. Here, the graph Laplacian L is defined as $L = D - W$ or $L = D^{-\frac{1}{2}}(W - D)D^{-\frac{1}{2}}$ if normalized. The matrix $W \in R^n \times R^n$ is the data adjacency graph, wherein each element W_{ij} is an edge weight between two samples x_i and x_j . In the diagonal matrix $D \in R^n \times R^n$, the i th entry $D_{ii} = \sum_{j=1}^n W_{ij}$. Various extensions of the graph Laplacian have been proposed in [144]. A simple one is the following

$$W_{ij} = \begin{cases} 1, & \text{if } x_i \in N(x_j) \text{ or } x_j \in N(x_i), \\ 0, & \text{otherwise,} \end{cases} \quad (5.13)$$

where $N(x_i)$ denotes the k nearest neighbors of the sample x_i . The graph Laplacian provides the following smoothness penalty on the graph

$$\mathbf{f}^T L \mathbf{f} = \sum_{i=1}^n (f(x_i) - f(x_j))^2 W_{ij}. \quad (5.14)$$

According to He and Niyogi [144], a definition in Eq.(5.14) corresponds to the approximation of $\int_M \|\nabla f(x)\|^2$, the manifold on which the data X reside. As indicated in [66], by setting $M = L$, the modified kernel Gram matrix $\tilde{K}$ allows us to reconstruct algorithms for semi-supervised classification and reinterpret them within the standard framework of supervised learning [44, 66]. In particular, Laplacian SVM and Laplacian LSRR will become the standard SVM and LSRR by using the new kernel function (i.e., Eq.(5.12)), respectively. In this work, by adopting the squared-loss function as in experiment design methods [45], we can reformulate the active learning framework (i.e., Eq.(5.10)) as a supervised learning model in the new deformed RKHS, i.e.,

$$\hat{f} = \arg \min_{f \in \tilde{H}_{\tilde{K}}} \left\{ J(w) = \sum_{i=1}^l (w^T \tilde{\phi}(z_i) - y_i)^2 + \gamma_1 \|w\|^2 \right\}, \quad (5.15)$$

where $\tilde{\phi}(z_i)$ denotes the sample z_i in $\tilde{H}_K$, which can reflect the geometric structure of unlabeled samples in the database [66]. From the representation theorem [167], we know that w can be represented as a linear combination of $\tilde{\phi}(z_i), i = 1, \dots, l$, i.e.,

$$w = \sum_{i=1}^l v_i \tilde{\phi}(z_i) = \tilde{\phi}(Z)v, \quad (5.16)$$

where $v = [v_1, \dots, v_l]^T \in R^l$ is the expansion coefficient vector. Bringing Eq.(5.16) back into Eq.(5.15), we can derive the following objective function, i.e.,

$$\hat{f} = \arg \min_{f \in \tilde{H}_K} \left\{ J(v) = \|\tilde{K}_z v - y\|^2 + \gamma_1 v^T \tilde{K}_z v \right\}, \quad (5.17)$$

where $y = [y_1, \dots, y_l]^T$; and $\tilde{K}_z \in R^{l \times l}$ is defined only on the labeled set $\tilde{\phi}(Z) = [\tilde{\phi}(z_1), \dots, \tilde{\phi}(z_l)]$ with its entries computed as in the kernel Gram matrix $\tilde{K}$. By setting $\frac{\partial J(v)}{\partial v} = 0$, we can get the optimal solution to Eq.(5.17) as

$$\hat{v} = (\tilde{K}_z + \gamma_1 I)^{-1} y. \quad (5.18)$$

For an input sample x , we can predict its label by

$$f(x) = \sum_{i=1}^l \tilde{k}(x, z_i) \hat{v}_i, \quad (5.19)$$

where $\tilde{k}(\cdot, \cdot)$ is the new kernel function defined in Eq.(5.12). Therefore, Eq.(5.19) can be used to measure the similarity between an input sample x and the query image. Then, the next problem is how to identify a set of the most informative samples $\tilde{\phi}(z_i), i = 1, \dots, l$ for the user to label. Let us consider the active learning framework with a supervised learning model in $\tilde{H}_{\tilde{K}}$, i.e., Eq.(5.15). We define $\tilde{\phi}(X) = [\tilde{\phi}(x_1), \dots, \tilde{\phi}(x_n)]$ as the set of all samples and $\tilde{\phi}(Z) = [\tilde{\phi}(z_1), \dots, \tilde{\phi}(z_l)]$ as the selected most informative ones in $\tilde{H}_{\tilde{K}}$. Motivated by TED [154, 155], we attempt to minimize the expected average prediction variance on the test data $\tilde{\phi}(X)$ in $\tilde{H}_{\tilde{K}}$. Similar to Eq.(5.6), the optimization problem can thus be reformulated accordingly as

$$\min_{\alpha_i \in R^l} \sum_{i=1}^n \|\tilde{\phi}(x_i) - \tilde{\phi}(Z)\alpha_i\|^2 + \gamma_1 \|\alpha_i\|^2. \quad (5.20)$$

Consequently, GOED tends to select the representative samples $\tilde{\phi}(Z) = [\tilde{\phi}(z_1), \dots, \tilde{\phi}(z_l)]$ that can span a linear space to retain most of the information of $\tilde{\phi}(X)$ in $\tilde{H}_{\tilde{K}}$, which thus has a clear geometric interpretation to the selected informative samples $\tilde{\phi}(Z)$ as TED [154, 155]. Moreover, different from SVMactive methods, GOED does not depend on the labels $y_i (i = 1, \dots, l)$, but only on the training data $\tilde{\phi}(Z) = [\tilde{\phi}(z_1), \dots, \tilde{\phi}(z_l)]$, which can effectively avoid various potential problems caused by insufficient and inexactlly labeled samples in RF.

Similarly, we can introduce a convex relaxation of Eq.(5.20) to obtain the global optimum, i.e.,

$$\begin{aligned} \min_{\alpha_i, \beta \in R^n} \sum_{i=1}^n \left(\|\tilde{\phi}(x_i) - \tilde{\phi}(X)\alpha_i\|^2 + \sum_{j=1}^n \frac{\alpha_{i,j}^2}{\beta_j} \right) + \lambda \|\beta\|_1 \\ \text{s.t. } \beta_j \geq 0, j = 1, \dots, n, \end{aligned} \quad (5.21)$$

where $\alpha_i = (\alpha_{i,1}, \dots, \alpha_{i,n})^T$; λ is the sparse regularization parameter; and $\|\cdot\|_1$ denotes the l_1 norm. The minimization of the l_1 norm $\|\beta\|_1$ can lead to a sparse coefficient β . Further, when $\beta_j = 0$, all $\alpha_{1,j}, \dots, \alpha_{n,j}$ must be zero and the j th sample will not be selected as the most representative one. Therefore, β can be used as the data selection coefficient. As has been stated earlier, this optimization problem (i.e., Eq.(5.21)) is convex, and therefore the global optimum can be obtained [155]. In the following, we will discuss how to solve this problem.

Let D_β be a diagonal matrix with entries $\beta_1, \dots, \beta_n$ and we have

$$\sum_{j=1}^n \frac{\alpha_{i,j}^2}{\beta_j} = \alpha_i^T D_\beta^{-1} \alpha_i. \quad (5.22)$$

With some simple algebraic steps, we get

$$\begin{aligned}
 & \sum_{i=1}^n \left(\|\tilde{\phi}(x_i) - \tilde{\phi}(X)\alpha_i\|^2 + \sum_{j=1}^n \frac{\alpha_{i,j}^2}{\beta_j} \right) + \lambda \|\beta\|_1 \\
 &= \sum_{i=1}^n \left((\tilde{\phi}(x_i) - \tilde{\phi}(X)\alpha_i)^T (\tilde{\phi}(x_i) - \tilde{\phi}(X)\alpha_i) + \alpha_i^T D_\beta^{-1} \alpha_i \right) + \lambda \|\beta\|_1 \\
 &= \sum_{i=1}^n \left(\tilde{\phi}(x_i)^T \tilde{\phi}(x_i) - 2\alpha_i^T \tilde{\phi}(X)^T \tilde{\phi}(x_i) + \alpha_i^T \tilde{\phi}(X)^T \tilde{\phi}(X) \alpha_i + \alpha_i^T D_\beta^{-1} \alpha_i \right) + \lambda \|\beta\|_1.
 \end{aligned} \tag{5.23}$$

And then, taking the derivative of Eq.(5.23) with respect to α_i and requiring it to be zero, we have

$$-2\tilde{\phi}(X)^T \tilde{\phi}(x_i) + 2\tilde{\phi}(X)^T \tilde{\phi}(X) \alpha_i + 2D_\beta^{-1} \alpha_i = 0. \tag{5.24}$$

Thus, we can get α_i as

$$\alpha_i = \left(D_\beta^{-1} + \tilde{\phi}(X)^T \tilde{\phi}(X) \right)^{-1} \tilde{\phi}(X)^T \tilde{\phi}(x_i). \tag{5.25}$$

Let $\tilde{K}_i$ be i th column vector of $\tilde{K}$, i.e.,

$$\begin{aligned}
 \tilde{K}_i &= \left(\tilde{\phi}(x_1)^T \tilde{\phi}(x_i), \dots, \tilde{\phi}(x_n)^T \tilde{\phi}(x_i) \right)^T \\
 &= \tilde{\phi}(X)^T \tilde{\phi}(x_i).
 \end{aligned} \tag{5.26}$$

Since $\tilde{\phi}(X)^T \tilde{\phi}(X) = \tilde{K}$, Eq.(5.25) can be rewritten as

$$\alpha_i = (D_\beta^{-1} + \tilde{K})^{-1} \tilde{K}_i. \tag{5.27}$$

Once α_i is calculated, we can fix α_i and find the minimum solution for β_j . Taking the derivative of Eq.(5.21) with respect to β_j and requiring the derivative to be zero, we can have

$$\sum_{i=1}^n \left(-\frac{\alpha_{i,j}^2}{\beta_j^2} \right) + \gamma = 0. \tag{5.28}$$

Finally, we get the data selection coefficient

$$\beta_j = \sqrt{\frac{\sum_{i=1}^n \alpha_{i,j}^2}{\gamma}}. \tag{5.29}$$

Table 5.1: Geometric Optimum Experimental Design

Input: The image database X with n unlabeled samples, the number of selected most informative samples l , the number of nearest neighbors k , the regularization parameter γ_1 , the data-dependent kernel deformation parameter γ , the sparse parameter λ , and the kernel type.
Step 1: Construct a nearest neighbor graph with the weight matrix W as in Eq.(5.13) on the data X and calculate the graph Laplacian $L = D - W$.
Step 2: Compute the conventional kernel Gram matrix K with the input kernel type and let $M = L$.
Step 3: Compute the data-dependent kernel Gram matrix $\tilde{K}$ according to Eq.(5.12).
Step 4: Initialize $\alpha_{i,j} = 1$, and iteratively compute β_j and α_i according to Eq.(5.29) and Eq.(5.27), respectively, until convergence.
Step 5: Rank the samples in X according to $\beta_j (j = 1, \dots, n)$ in a descending order and return the top l samples as the most informative ones Z .
Step 6: The l selected most informative samples Z should be labeled by the user and used as the training data to obtain a classifier f according to Eq.(5.17) and Eq.(5.18).
Output: The classifier f , which can be used as a similarity metric to measure the similarity between a given sample x and the query image, i.e., Eq.(5.19).

Figure 5.2: The framework of our CIR system. Different from conventional CBIR systems, our system can select the most informative images in the database for the user to label and return the most semantically relevant images as the final results.

$\alpha_{i,j}$ and β_j can be computed iteratively according to Eq.(5.27) and Eq.(5.29), respectively. The objective function of Eq.(5.21) is convex, and thus the global optimum can be obtained after iterations.

We rank the samples in X according to the data selection coefficient β in a descending order, and the top l samples are selected as the most informative ones Z . The system requires the user to label the relevant and irrelevant images in Z as the positive and negative feedback samples in RF, respectively. Finally, these samples can be used to train a classifier f according to Eq.(5.17) and Eq.(5.18), and then the classifier can be used as a similarity metric to measure the similarity between a given sample x and the query image based on the output of f , i.e., Eq.(5.19).

The GEOD for CIR can be summarized in Table 5.1.

5.4 Collaborative Image Retrieval

In this section, we first give an overview of our CIR system, which can effectively select the most informative images in the database for the user to label and return the most semantically relevant images as the final results. The CIR system assumes that the user expects the best possible retrieval results after each round of RF. Meanwhile, the user will not label a large number of samples at each round of RF and only do a few rounds of RF. To deal with this type of scenario, the following CIR framework is proposed. With the proposed system, we can embed many kinds of RF methods easily.

As shown in Figure 5.2, when a query image is provided, the low-level visual features are first extracted. Then, all images in the database are sorted based on a predefined similarity metric. If the user is satisfied with the results, the image retrieval process is ended. However, most of the time, RF is actually required because of the poor performance of the system. The CIR system requires the user to label some top informative images as the positive and negative feedback samples by leveraging the auxiliary information of unlabeled samples in the database. Using these samples as the training data, an RF model can be obtained based on certain machine learning techniques. The similarity metric can thus be updated together with the RF model. Then, all images are resorted based on the recalculated similarity metric. If the user is satisfied with the refined results, RF is no longer required and the system gives the final results, which are most semantically relevant to the query image. Otherwise, RF will be performed iteratively.

From Figure 5.2, it can be noticed that our CIR system is different from conventional CBIR systems, which can only present relevant and irrelevant images in top returned results for the user to label and is not proper since top returned results will not always be the most informative ones. Our CIR system can effectively select representative images in the database as the most informative ones for the user to label and return the most semantically relevant images as the final results. In Figure 5.2, we can see that when a query image is submitted, a set of the most informative relevant and irrelevant images can be selected from the database, which can then be used as the training data to define a new similarity metric. Thus, the CIR system can accomplish a retrieval task with fewer iterations than conventional RF methods based CBIR systems and alleviate the labeling efforts of the user with the help of the most informative samples in the database.

We manually divide 10,763 images from the Corel photo gallery into 80 concept groups, such as tiger, lion, elephant, horse, etc., to perform empirical evaluation of the CIR system as in previous research on CBIR [86, 153, 168, 169]. Some example images in the Corel image database are shown in Figure 5.3. In our experiments, we select the same visual features, i.e., color, local descriptors [146], and edges, to represent the images as in Chapter 3 and Chapter 4. The representation of an image will result in a feature vector with 510 values.

Figure 5.3: Some example images in the Corel image database.

5.5 Experimental Results

In this section, we will evaluate the potential of GOED in selecting representative samples in the database as the most informative ones for the user to label in image retrieval. We design the experiments for performance evaluation in four aspects. First of all, we use synthetic datasets to visualize the unstable problem in SVMactive when dealing with small-sized training data. Moreover, in this subsection, we also show the effectiveness of GOED in selecting the most representative samples compared with TED based on a synthetic dataset. Second, we investigate the GOED method for CIR by comparing it with a conventional RF scheme based CBIR system. Then, we report the performance of GOED and compare it with some popular RF schemes for image retrieval based on a large real-world image database. Finally, we study the sensitivity of the important parameter in the proposed GOED method.

5.5.1 Experiments on Synthetic Datasets

5.5.1.1 SVM is Unstable with Small-Sized Training Data

SVMactive is one of the most popular methods to identify the most informative samples for CIR, which requires the user to label the unlabeled samples closest to the optimal hyperplane of SVM because of their low predication confidence. However, the optimal hyperplane of SVM is always unstable with small-sized training data, i.e., this optimal hyperplane is always sensitive to the training data when the size of training feedback samples is small, and this is always the case in image retrieval since the user would

Figure 5.4: The illustration of the unstable problem of the optimal hyperplane of SVM when dealing with small-sized training data. The open circles, triangles, and crosses denote the positive feedback samples, negative feedback samples, and unlabeled samples, respectively. The white solid line indicates the optimal hyperplane of SVM, which separates the positive and negative feedback samples. The unlabeled samples closest to the optimal hyperplane of SVM will be identified as the most informative ones.

not like to label a large number of feedback samples. And thus it is not appropriate to directly use this hyperplane to identify the most informative samples for the use to label. To visualize the unstable problem with small-sized training data, first we use a toy example to illustrate the sensitivity of the optimal hyperplane of SVM in RF. In Figure 5.4, we use the open circles, triangles, and crosses to denote the positive feedback samples, negative feedback samples, and unlabeled samples in the database, respectively. Figure 5.4 (a) shows the optimal hyperplane of SVM, which is trained by the original feedback samples and adopted to identify the most informative unlabeled samples for the user to label. The unlabeled samples closest to the optimal hyperplane of SVM will be identified as the most informative ones for the user to label. However, as shown in Figure 5.4 (b) and (c), much different optimal hyperplanes will be trained by the original training data with one and two incremental positive feedback samples, respectively. Consequently, the optimal hyperplane of SVM is usually unstable and it is not appropriate to directly use this hyperplane to identify the most informative samples when dealing with small-sized training data.

5.5.1.2 GOED can Select the Most Representative Samples in the Database

GOED aims to select representative samples in the database as the most informative ones for the user to label, which is fundamentally based on TED with a convex solution in the

Figure 5.5: The performance comparison between TED and GOED in selecting the most representative samples. The black dots and blue crosses indicate the relevant and irrelevant unlabeled samples in the database, respectively. The red open circles are the most representative samples selected by TED and GOED in the database.

new deformed RKHS [155]. The difference between the two algorithms is whether the geometric structure of unlabeled samples is fully exploited. To visualize the effectiveness of GOED in selecting the most representative samples, we give a simple toy example to compare the effectiveness of the two algorithms. In Figure 5.5, we use two sets of samples on two circles to indicate the relevant and irrelevant unlabeled samples with the query image in the database. Especially, the black dots on the small circle represent the relevant unlabeled samples and the blue crosses on the big circle denote the irrelevant unlabeled samples, respectively. The eight most informative samples in the database are required to be selected by TED and GOED for the user to label. Both of the two algorithms use the Gaussian kernel function. As we can see, all samples selected by TED are from the set of relevant samples on the small circle, while GOED can select four relevant samples on the small circle and four irrelevant samples on the big circle, respectively. Clearly, the informative samples selected by GOED can better represent the original database (i.e., two circles), which should be labeled by the user and used as the training data to acquire a more accurate RF model. This is mainly because more unlabeled samples will share the labeling information if the most representative samples can be labeled by the user and used as the training data in RF. It should be noted that we do not compare our method with SVMactive because SVMactive cannot be applied in this case due to the lack of training data.

5.5.2 Experiments on a Real-World Image Database

In this subsection, we will evaluate the effectiveness of the proposed GOED in selecting representative samples in the database as the most informative ones for the user to label in image retrieval based on two experiments: first, we investigate the GOED method for CIR by comparing it with a conventional RF scheme based CBIR system and thus show the potential of CIR in improving the performance of image retrieval; then, we show the performance of our CIR system by selecting the most informative samples for the user to label and compare it with some popular RF methods for image retrieval based on a real-world image database.

We use average precision (AP), standard deviation (SD), and average recall (AR) to evaluate the performance of the compared algorithms. AP is the major evaluation criterion, which evaluates the effectiveness of the compared algorithms. In experiments, we empirically set the regularization parameter γ_1 as $1e-3$ and the sparse regularization parameter as 100 in experiments. For all kernel based methods, we choose the Gaussian kernel function and select the best kernel parameters by the grid search methods. Considering the computable efficiency, we do not use all unlabeled images in the database but only the top 500 returned results at the previous round of RF. The number of nearest neighbors k is empirically set as 4 according to [144]. It should be noted that both the SVM and SVMactive methods adopt the hinge-loss function, and all other methods (i.e., GOED, TED, and LSRR) use the squared-loss function. In experiments, the data-dependent kernel deformation parameter γ is empirically set as 0.1, which will be analyzed carefully in the next subsection.

5.5.2.1 Performance Evaluation on a Small-Scale Image Database

In this part, we intend to examine how effective the proposed GOED is when selecting the most informative samples for the user to label in image retrieval, and thus evaluate the potential of CIR compared with the conventional RF based CBIR systems, which can only select top returned results for the user to label. In general, top returned results are not the most informative ones to enhance the performance of image retrieval, which is however not appropriate. Basically, the proposed GOED is designed based on the LSRR model

Figure 5.6: APs with 30 categories in top 20 results of the compared algorithms, i.e., GOED, LSRR, and Baseline, after the second round of RF.

in RKHS. However, LSRR can only require the user to label the top returned results. Therefore, in experiments, we compare the proposed GOED with LSRR and show the performance difference between the two algorithms. The experiments are conducted on a small-scale image database, which includes 3,139 images with 30 different categories. All 3,139 images in the 30 categories are used as the query images to evaluate the compared algorithms. To avoid the potential problems caused by an imbalanced number of the positive and negative feedback samples [36], we select an equal number of relevant and irrelevant images as the positive and negative feedback samples, respectively. For GOED, the first 3 relevant images and first 3 irrelevant images in the 20 most informative samples selected by the algorithm are labeled as the positive and negative feedback samples, respectively, whereas for LSRR, the first 3 relevant images and first 3 irrelevant images in top 20 returned results are marked as the feedback samples.

Figure 5.6 shows APs in top 20 returned results after the second round of RF for all 30 categories. The baseline describes the initial retrieved results without any RF information. Especially, at the beginning, the Euclidean distance metric in the high-dimensional feature space are used to measure the similarity between the query image and the images in the database. After the user provides the RF information, GOED

Figure 5.7: APs of the compared algorithms, i.e., GOED, LSRR, and Baseline, after the first and second rounds of RF. (a) the first round of RF and (b) the second round of RF.

and LSRR can be applied to train a classifier, and thus define a new similarity metric to resort the images in the database. As shown in Figure 5.6, the performance of the two algorithms varies with different categories. For some easy categories, both of the two algorithms can perform well (e.g., Categories 2 and 3). For some hard categories, both of the two algorithms poorly perform (e.g., Categories 16, 18, 22, and 23). After the second round of RF, both of the two algorithms can show much better performance than the baseline and therefore significantly improve the performance of image retrieval. Compared with LSRR, GOED can perform much better for most of the 30 categories (e.g., Categories 1, 4, 19, 20, and 27), which indicates that GOED is more effective for image retrieval since this method can automatically select the most informative samples for the user to label.

Figure 5.7 (a) and (b) show APs of the compared algorithms after the first and second rounds of RF, respectively. As shown in Figure 5.7, we can notice that both GOED and LSRR can show much better performance than the baseline on the entire scope, particularly after the second round of RF. This is mainly because by iteratively cumulating the RF information, more feedback samples will be used as the training data, and this will significantly enhance the performance of the system. Compared with LSRR, GOED can show much better performance on the entire scope, since GOED can effectively select the most informative samples for the user to label while LSRR can only label the top returned results.

Based on the aforementioned results, we notice that GOED is more effective than LSRR, and thus show the potential of CIR by selecting the most informative samples for the user to label in improving the performance of image retrieval. We should highlight that, at the first round of RF, there are no training data and thus conventional SVMactive cannot be applied to select the most informative samples for the user to label. Compared with SVMactive, GOED is label-independent and thus will be more appropriate for CIR.

5.5.2.2 Performance Evaluation on a Large-Scale Image Database

In this part, we first design a scheme to model the real-world image retrieval process. In a real-world image retrieval system, a query image is usually not in the database. To simulate such an environment, we use a fivefold cross validation database to evaluate the compared algorithms. More precisely, we divide the whole image database into five subsets with an equal size. Therefore, there are 20 percent per category in each subset. At each run of cross validation, one subset is selected as the query set, and the other four subsets are used as the database for image retrieval. Then, 500 query images are randomly selected from the query set, and RF is automatically conducted by the system. For each query image, the system retrieves and ranks the images in the database. Finally, 9 rounds of RF are automatically carried out by the system in this experiment.

To show the effectiveness of the proposed GOED, we compare it with TED, LSRR, SVMactive, and SVM. Out of these five algorithms, GOED, TED, and SVMactive are active learning methods, whereas LSRR and SVM are standard classification methods. At each round of RF, 20 images are picked from the database and examined serially to mark as the positive or negative feedback samples. In general, in a real-world image retrieval system, the irrelevant images (i.e., the images differ in different concepts with the query image) usually largely outnumber the relevant ones (i.e., the images share a same concept with the query image). To simulate such a case in the system, the first 3 relevant images are labeled as the positive feedback samples, and all other irrelevant images in the 20 images are automatically labeled as the negative feedback samples. The images that have been selected in previous rounds of RF are excluded from later selections. It should be noted that for active learning based RF methods (i.e., GOED,

Figure 5.8: APs of GOED compared with the algorithms, i.e., TED, SVMactive, SVM, and LSRR. All the algorithms are evaluated over 9 rounds of RF. 0 round of RF refers to the retrieved results based on the Euclidean distance metric without any RF information.

Figure 5.9: SDs of GOED compared with the algorithms, i.e., TED, SVMactive, SVM, and LSRR. All the algorithms are evaluated over 9 rounds of RF. 0 round of RF refers to the retrieved results based on the Euclidean distance metric without any RF information.

TED, and SVMactive), the 20 images are selected from the database by the algorithms themselves, whereas for conventional classification based RF methods (i.e., SVM and LSRR), the 20 images are composed of the top 20 returned results at the previous round of RF, which is the most popular way to select the feedback samples in the existing research of CBIR. In this experiment, we calculate the APs and the SDs over the 500 query images at different positions from top 10 to top 60 to obtain the APs and ARs, and all experimental results are computed from the fivefold cross validation.

Figure 5.8 and Figure 5.9 show the APs and the SDs of the compared algorithms, respectively. As shown in Figure 5.8, GOED consistently outperforms all the other compared algorithms on the entire scope. GOED and TED are mainly designed based on the LSRR model in RKHS; however, both of them can significantly improve the performance of LSRR by selecting the most informative samples for the user to label. Compared with TED, GOED performs much better for all top results, since GOED tries to find the most informative samples for the user to label by leveraging the manifold structure of the data on which the classifier is as smooth as possible. SVMactive cannot show better performance than GOED and TED, since the optimal hyperplane of SVM is usually not very stable and accurate with small-sized training data in a high dimensional space (i.e., about tens of samples in a 510 dimensional space in this chapter). Therefore, it is not appropriate to directly use the optimal hyperplane of SVM to identify the most informative samples when the number of training feedback samples is small. Different from SVMactive, GOED and TED can select representative samples in the database as the most informative ones for the user to label, which are actually label-independent and therefore more appropriate for image retrieval. This is mainly because the user would not like to label a large number of samples and cannot label each sample accurately in RF. Moreover, we should indicate that SVMactive can only be applied when there is an initial classifier. Therefore, it cannot be applied at the first round of RF. In experiments, we use the standard SVM to build an initial classifier. And thus, the performance of SVMactive and SVM are the same for all the results after the first round of RF. When considering more rounds of RF, SVMactive can get little improvement over the standard SVM. Compared with LSRR, SVM can show slightly better performance than LSRR for all top results especially after a few rounds of RF since the loss function in SVM (i.e.,

Table 5.2: APs and SDs in top N results of the compared algorithms, i.e., GOED, TED, LSRR, SVMactive, and SVM, after the ninth round of RF (APs $\pm$ SDs).

Method	GOED	TED	LSRR	SVMactive	SVM
Top 10	0.9432 $\pm$ 0.2054	0.9320 $\pm$ 0.1800	0.7532 $\pm$ 0.3656	0.8447 $\pm$ 0.2522	0.8379 $\pm$ 0.2809
Top 20	0.8404 $\pm$ 0.2491	0.8124 $\pm$ 0.2366	0.6666 $\pm$ 0.3995	0.8021 $\pm$ 0.2791	0.7563 $\pm$ 0.3068
Top 30	0.7714 $\pm$ 0.2675	0.7381 $\pm$ 0.2579	0.5547 $\pm$ 0.3661	0.7329 $\pm$ 0.2985	0.6474 $\pm$ 0.3088
Top 40	0.7156 $\pm$ 0.2737	0.6805 $\pm$ 0.2633	0.4802 $\pm$ 0.3392	0.6443 $\pm$ 0.2956	0.5633 $\pm$ 0.2978
Top 50	0.6684 $\pm$ 0.2756	0.6362 $\pm$ 0.2637	0.4271 $\pm$ 0.3169	0.5719 $\pm$ 0.2839	0.5028 $\pm$ 0.2848
Top 60	0.6255 $\pm$ 0.2761	0.5900 $\pm$ 0.2598	0.3839 $\pm$ 0.2963	0.5167 $\pm$ 0.2707	0.4552 $\pm$ 0.2716
Top 70	0.5884 $\pm$ 0.2653	0.5530 $\pm$ 0.2542	0.3497 $\pm$ 0.2784	0.4705 $\pm$ 0.2568	0.4167 $\pm$ 0.2563
Top 80	0.5522 $\pm$ 0.2568	0.5188 $\pm$ 0.2458	0.3210 $\pm$ 0.2616	0.4317 $\pm$ 0.2428	0.3848 $\pm$ 0.2414
Top 90	0.5201 $\pm$ 0.2418	0.4877 $\pm$ 0.2359	0.2966 $\pm$ 0.2461	0.3999 $\pm$ 0.2301	0.3574 $\pm$ 0.2279

hinge-loss) has much better generalization ability than the loss function in LSRR (i.e., squared-loss).

Regarding the stability of the compared algorithms, we can also notice that GOED and TED perform best among top 10, 20, 30, and 40 results as shown in Figure 5.9. Then, for other top results, the performance of GOED is similar to other compared algorithms. The detailed results of the compared algorithms after 9 rounds of RF are shown in Table 5.2. As given in Table 5.2, GOED achieves much better performance compared with other algorithms for all top results. TED still obtains satisfactory performance, as compared with SVM, LSRR, and SVMactive. Therefore, we can conclude that the proposed GOED has shown its effectiveness in selecting representative samples in the database for the user to label in image retrieval.

It should be noted that the performance difference between this work and the previous research is mainly due to different setting. As indicated in [34], to obtain a stable classifier, the user is usually required to randomly select a large number of samples to label at the first round of RF and a large number pool samples closest to optimal hyperplane of SVM (e.g., more than 20 samples in a 144-dimensional space), which can then be used to train a classifier and identify the ambiguous samples for the user to label. However, in practice, the number of positive feedback samples is usually much less than that of negative ones [36], and it is also not appropriate to require the user to label a large amount of samples in RF. In this work, we select at most 3 relevant images and all other irrelevant images in 20 returned results in a 510-dimensional space as the feedback samples, which is similar to a real-world image retrieval process. Moreover,

Figure 5.10: The performance sensitivity of GOED with regard to the data-dependent kernel deformation parameter γ . (a) APs and (b) ARs.

we use a fivefold cross validation database to evaluate the effectiveness of the compared algorithms, and this is different from the previous study [34].

5.5.3 Parameter Sensitivity

In this subsection, we study the sensitivity of GOED with regard to the kernel deformation parameter γ in constructing the data-dependent kernel function, which is the key in GOED to significantly improve the performance of image retrieval by leveraging the geometric structure of unlabeled samples in RKHS. With the observation that the difference between GOED and TED is whether the geometric structure of unlabeled samples is fully exploited, we show the performance difference between the two methods. The analysis is performed based on experiments conducted on the small-scale image database given in Subsection 5.5.2.1. In experiments, we use all samples in the database to construct the data-dependent kernel function and 500 images are randomly selected from the database and used as the query images. For each query image, the system retrieves and ranks the images in the database and 4 rounds of RF are automatically carried out by the system. At each round of RF, 20 most informative samples selected by the algorithms are examined from the top. The first 3 relevant and first 3 irrelevant images in the

20 selected informative samples are automatically marked as the positive and negative feedback samples by the system, respectively.

In experiments, we empirically set the kernel deformation parameter γ as a value in a sequence, i.e., $\{2^i, i = -10, -9, \dots, 2, 3\}$. Figure 5.10 shows how APs and ARs of GOED vary with γ in top 10, 30, and 50 results after 4 rounds of RF. The dashed lines indicate the corresponding results of TED. We can notice that the parameter γ can significantly affect the performance of GOED with regard to APs and ARs. As shown in Figure 5.10, when γ is small enough, the APs and ARs of GOED will approximate the corresponding results of TED. This is mainly because GOED will result in TED when the parameter γ is set as 0, i.e., GOED will be a standard supervised learning method and cannot exploit the geometric structure of unlabeled samples as TED. When γ becomes larger, GOED shows much better performance regarding APs and ARs. However, if γ is too large, the performance may be degenerated, and this is mainly due to the overdeformation in the new RKHS. As we can see, GOED achieves consistently good performance with γ varying from $2^{-4} \sim 2^{-2}$. The parameter can be further tuned to achieve better performance. The analysis above also indicates that the GOED can effectively alleviate the small-sized training data problem by leveraging the geometric structure of unlabeled samples in RKHS.

5.5.4 Discussions

The last two decades have witnessed the significance of RF provided by the user in improving the performance of image retrieval [22, 32, 35–37, 75, 86, 87, 90, 160, 161]. Various RF schemes have been widely developed based on different assumptions for the positive and negative feedback samples. However, conventional RF methods can only require the user to label the relevant and irrelevant images in top returned results, and this is not appropriate since top returned results may not be the most informative ones to define an effective similarity metric for image retrieval.

SVMactive is one of the most important techniques to select the most informative samples for the user to label and thus alleviate the labeling efforts of the user in image retrieval [27, 34, 38, 97, 98]. This method always requires the user to label the unlabeled samples closest to the optimal hyperplane of SVM, which are ambiguous and have low

prediction confidence. However, the optimal hyperplane of SVM is always unstable and inaccurate with small-sized training data, and this is always the case in image retrieval since the user would not like to label a large number of feedback samples and cannot label each sample accurately.

GOED can select representative samples in the database as the most informative ones for the user to label. Compared with SVMactive, GOED is label-independent and thus can effectively avoid various potential problems caused by insufficient and inexactlly labeled samples in RF. Compared with TED, GOED can constrain the classifier prediction to be smooth over the manifold estimated from the unlabeled samples in RKHS and thus can alleviate the small-sized training data problem by leveraging the geometric structure of unlabeled samples.

Active learning has been widely investigated in the machine learning community [70–74]. Discriminative models, e.g., SVMactive, aim to select the most ambiguous samples and generative models attempt to identify the most representative samples. In general, when the number of training samples is large, the methods which select the most ambiguous samples (e.g., SVMactive) may outperform the methods which select the most representative samples (e.g., GOED). This is mainly because with a large amount of training samples, the classifier will be more accurate and thus the ambiguous samples selected by the algorithm can provide the most amount of information. However, for image retrieval, it is not appropriate to require the user to label a large number of feedback samples. Consequently, when the number of the training feedback samples is small, GOED is more appropriate to select representative samples in the database as the most informative ones for the user to label compared with the popular SVMactive for CIR.

5.6 Summary

In this chapter, a geometric optimum experimental design (GOED) method is introduced to select multiple representative samples in the database as the most informative ones for the user to label in RF. Especially, GOED can alleviate the small-sized training data problem by leveraging the geometric structure of unlabeled samples in RKHS, and thus can further enhance the performance of image retrieval. By minimizing the expected

average prediction variance on the test data, GOED has a clear geometric interpretation to select the most representative samples in the database iteratively with the global optimum, and it is more effective and efficient for the user to label. Compared with the popular SVMactive, our method is label-independent and thus can avoid various potential problems caused by insufficient and inexactlly labeled feedback samples, and is more appropriate and useful for image retrieval. Extensive experiments on both synthetic datasets and a real-world image database have shown the advantages of the proposed GOED for CIR.

Chapter 6

Conjunctive Patches Subspace Learning with Side Information

6.1 Motivation

Conventional relevance feedback (RF) involves the user in the system to improve the performance of content-based image retrieval (CBIR) by gradually learning the user's preferences [35–37, 75–79, 81–85, 90]. Despite broad interests in constructing various RF methods, an online learning task can be tedious and boring for the user. Given the difficulties in capturing the user's preferences, multiple rounds of RF are actually required to achieve satisfactory results for an image retrieval task, which will limit the capability of RF for real-world applications. Active learning can be used to select the most informative samples for the user to label in RF, and this can significantly accelerate this procedure in image retrieval.

However, conventional CBIR systems cannot accumulate and utilize the RF information provided by a number of users, which can be regarded as the RF log data. This is mainly because from a long-term perspective, the RF log data can also be an important and useful resource to enhance the performance of the CBIR system. Recently, a number of studies have attempted to address the challenges encountered by conventional RF methods by resorting to the RF log data [24–26, 100, 168, 170, 171]. This new paradigm can be considered as collaborative image retrieval (CIR), which can alleviate the labeling efforts of conventional RF methods by leveraging the RF log data. In these studies, the system can accumulate the RF information provided by a number of users, and this will

Figure 6.1: Different similar relations between pairs of images based on different semantic concept subspaces in a multi-dimensional low-level visual feature space. (a) four images with low-level visual features, (b) similar in the shape subspace, (c) similar in the size subspace, and (d) similar in the texture subspace.

form a large-scale database associated with the RF log data. Therefore, besides low-level visual features, each pair of images can also be associated with a set of similar and dissimilar pairwise constraints judged by users. During the past several years, a lot of research work has been conducted regarding the new paradigm for image retrieval. In [24, 25], manifold learning algorithms were applied to learn an exquisite manifold structure from the RF log data, which can better reflect the semantic relation among images. In [100], Muller et al. suggested a weighting scheme by exploiting the RF log data for image retrieval. In [26], Hoi et al. proposed a log based RF scheme with SVM by engaging the log information in a regular online RF task. In [168], the authors proposed a distance metric learning technique by exploiting the log database with pairwise constraints and showed the effectiveness of the proposed scheme compared with some representative distance metric learning techniques for image retrieval. To sum up, we can notice that the key issue for CIR is to design an effective scheme to fully exploit the RF log data, and then use the acquired information to enhance the performance of the system.

Various approaches and schemes have been investigated for CIR; however, there is still little work on explicitly evaluating subspace learning methods in exploiting the RF log data, although the subspace learning methods play a vital role in many multimedia retrieval tasks. Let us first use a toy example to show the importance of subspace learning methods in defining the similar relation between a pair of images, which is usually the key issue in exploiting the RF log data for CIR. In image retrieval, images are usually represented by a set of low-level visual features with various semantic concepts (e.g.,

color, shape, texture, etc) in a high-dimensional space. With an assumption that different semantic concepts live in different subspaces and each image can live in many different semantic concept subspaces, Figure 6.1 (a) shows four images, each of which is associated with a number of semantic concepts (i.e., color, shape, texture, and size). However, for CIR, it is problematic for the user to determine the similar relation between a pair of images in the original multi-dimensional space (i.e., color, shape, texture, and size) due to the semantic gap. By selecting one-dimensional semantic concept subspace, defining the similar relation between a pair of images will be easy and obvious. Figure 6.1 (b), (c), and (d) show three different kinds of similar relations in three different semantic concept subspaces, respectively (i.e., Figure 6.1 (b) in the shape subspace, Figure 6.1 (c) in the size subspace, and Figure 6.1 (d) in the texture subspace).

Subspace learning methods [172] are powerful tools for various tasks in the computer vision community [86, 173, 174], e.g., face recognition [28], image retrieval [32], and gait recognition [46]. However, most of these traditional subspace learning methods (e.g., linear discriminant analysis (LDA)) normally need to acquire explicit class label information [172]. For CIR, explicit class label information for each image might be too expensive to obtain. Compared with the explicit class label information of each image, similar and dissimilar pairwise constraints between a pair of images can be acquired much easier when the RF log data is available [100]. Therefore, it is more attractive to directly learn a semantic concept subspace from the similar or dissimilar pairwise constraints without using any class label information. Recently, learning distance metrics with similar and dissimilar pairwise constraints (or side information [43]) has been actively studied [43, 116, 168, 175] in the machine learning community. Despite the active research efforts during the past a few years, most of these approaches in this group have involved a high computational burden when dealing with high-dimensional images, which will significantly limit their potential applications to CIR.

In this chapter, we propose a novel framework of subspace learning when the training images are associated with only similar and dissimilar pairwise constraints, i.e., conjunctive patches subspace learning (CPSL) with side information, to explicitly exploit the RF log data for CIR. The proposed CPSL method can effectively learn a reliable subspace both from the labeled and unlabeled images through a regularized learning framework

by exploiting the RF log data in image retrieval. Especially, we formally formulate this method as a constraint optimization problem, and then present an efficient algorithm to solve this task with closed-form solutions. Compared with previous distance metric learning methods [43, 168, 175], which usually involve a convex optimization procedure or a semidefinite programming procedure, our method can also learn a distance metric but can perform more effectively and efficiently when dealing with high-dimensional data.

The rest of this chapter is organized as follows. Section 6.2 reviews the related work; CPSL with side information is detailed in Section 6.3; a CIR system is introduced in Section 6.4; in Section 6.5, we first give the experimental results on both synthetic datasets and a real-world image database, and then show some analysis to the important parameters in CPSL; Section 6.6 summarizes this chapter.

6.2 Related Work

To describe our method clearly, let us first review two areas of the research that are closely related to our work in this chapter, i.e., (1) exploiting RF log data and (2) subspace learning and distance metric learning.

6.2.1 Exploiting RF Log Data

During the past a few years, various conventional RF methods have been developed. However, it is still a big problem to effectively bridge the semantic gap between low-level visual features and high-level semantic concepts.

Besides conventional RF methods, there are some emerging research interests in exploiting the RF log data [26, 170] for image retrieval. In [26], Hoi et al. proposed a log based RF scheme with SVM by engaging the log data for a regular online RF task. In this scheme, the user first labels some similar and dissimilar images at a few rounds of RF, and then the images in the database that are similar to the current labeled images are included in the pool of log data for training some conventional RF models, e.g., SVM-based RF. Besides the SVM based method with log data, some other efforts are also investigated in exploiting the RF log data. For instance, manifold learning techniques expect to learn an exquisite manifold structure based on the RF log data in image retrieval [24, 25]. In [100], Muller et al. proposed a feature weighting scheme by exploiting

the user’s historical relevance judgements for image retrieval. Moreover, some distance metric learning techniques have also been widely investigated to learn a good Mahalanobis distance metric by exploiting the user’s historical similar and dissimilar judgements on the feedback samples, and thus enhance the performance of the system. [168, 171].

6.2.2 Subspace Learning and Distance Metric Learning

In view of the close relation between subspace learning and distance metric learning, we briefly classify the two groups of studies into three categories within a unified framework, i.e., unsupervised learning, supervised learning with explicit class label information, and weakly supervised learning with pairwise constraints (or side information [43]).

Unsupervised learning methods do not use any class label information and usually exploit the intrinsic distribution or the geometric structure of data. Examples in this category include the well-known algorithms, such as principal component analysis (PCA) [172] and multi-dimensional scaling (MDS) [176]. Moreover, there are also some recent manifold learning approaches, which are locally linear embedding (LLE) [41], ISOMAP [134], Laplacian eigenmaps (LE) [40], locality preserving projections (LPP) [144], etc.

Supervised learning methods can effectively exploit a collection of data with explicit class label information. Well-known approaches in this category include Fisher’s LDA [172], marginal Fisher analysis (MFA) [136], and some recently proposed methods, such as neighborhood component analysis (NCA) [177], large margin nearest neighbor (LMNN) [178], and maximally collapsing metric learning (MCML) [179].

Our work is closely related to the third category of research. Let us briefly introduce several representative methods below.

Most of the weakly supervised learning methods can only learn a Mahalanobis distance metric from a collection of data, which are presented in the forms of pairwise constraints (or side information [43]), in which each pairwise constraint indicates whether the corresponding two samples are similar or dissimilar for a particular task. In [43], Xing et al. proposed a distance metric learning method (called Xing hereafter) and formulated this task as a convex optimization problem, which can be solved by an iterative projection algorithm. And then, a series of research work has been conducted with regard to this category of study. In [175], a relevant component analysis (RCA) method was proposed

to exploit similar pairwise constraints for distance metric learning. In details, given pairwise constraints, RCA first forms a set of “chunklets”, each of which is defined as a group of samples linked together by similar pairwise constraints. The optimal distance metric learned by RCA can be computed as the inverse of the average covariance matrix of the chunklets. RCA is simple to calculate, but ignores the dissimilar pairwise constraints. Discriminative component analysis (DCA) was proposed to incorporate the dissimilar pairwise constraints [116], which can show slightly better performance compared with RCA for some datasets. Lately, an information-theoretic metric learning approach was proposed to express the weakly supervised learning problem as a Bregman optimization problem [117]. To effectively exploit the unlabeled samples, Hoi et al. proposed a Laplacian regularized metric learning method and applied the generated solution to image retrieval and clustering [168]. In [119], Wu et al. proposed to learn a Bergman distance function with side information and showed that this approach can learn nonlinear distance functions for a semi-supervised clustering task.

6.3 Conjunctive Patches Subspace Learning with Side Information

In this section, we propose a novel framework of weakly supervised subspace learning, i.e., CPSL with side information, to explicitly exploit the RF log data for CIR. The proposed CPSL can directly learn a semantic concept subspace from the similar and dissimilar pairwise constraints without using any class label information, and this is more practical and useful for CIR, since explicit class label information for each image might be too expensive to obtain in many practical applications.

6.3.1 Problem Definition

To facilitate the discussion, let us first introduce some necessary notations. Assume that we are given a set of n images in the h -dimensional visual feature space $X = [x_1, x_2, \dots, x_n] \in R^h$, and two sets of similar and dissimilar pairwise constraints among the images as

$$\begin{aligned} S &= \{(i, j) \mid x_i \text{ and } x_j \text{ are judged to be similar}\}, \\ D &= \{(i, j) \mid x_i \text{ and } x_j \text{ are judged to be dissimilar}\}, \end{aligned}$$

where S is the set of similar pairwise constraints and D is the set of dissimilar pairwise constraints. Each pairwise constraint (i, j) indicates whether the two images x_i and x_j are similar or dissimilar judged by users in RF. It should be noted that it is not necessary for all images in X to be involved in S or D .

In this work, we use the low-level visual features in a high-dimensional space to represent the images. Although the low-level visual features of images are embedded in a high-dimensional space, the semantic concepts of images actually live in a low-dimensional subspace. Here, in this work, the high-dimensional space R^h is the low-level visual feature space and the low-dimensional subspace R^l is the high-level semantic concept subspace. Therefore, our objective is to find a mapping function to select an effective semantic concept subspace R^l from R^h for bridging the semantic gap. To learn such a semantic concept subspace, one can assume there is some corresponding linear mapping $U \in R^{h \times l}$ for a possible subspace, and then we can obtain the low-dimensional semantic representation as $Y = U^T X \in R^{l \times n}$, where each column of Y is $y_i = U^T x_i \in R^l$.

To measure the similarity between two images y_i and y_j in the low-dimensional semantic concept subspace, we adopt the Euclidean distance metric for its simplicity and robustness. The Euclidean distance between two images in the low-dimensional semantic concept subspace can be calculated as

$$d(y_i, y_j) = \sqrt{(U^T x_i - U^T x_j)^T (U^T x_i - U^T x_j)} = \sqrt{(x_i - x_j)^T U U^T (x_i - x_j)}. \quad (6.1)$$

Let $M = U U^T$, then

$$d(y_i, y_j) = \sqrt{(x_i - x_j)^T M (x_i - x_j)}. \quad (6.2)$$

Therefore, learning a mapping matrix U is actually equivalent to learning an efficient Mahalanobis distance metric M in the original high-dimensional space, or more concretely, learning a proper Mahalanobis distance metric M in R^h .

During the past a few years, a variety of approaches have been proposed to learn such an optimal Mahalanobis distance metric M from the training samples that are given in forms of side information [43, 116, 119, 168, 175, 180]. However, most of these methods are imperfect for CIR, since they either require solving a convex optimization problem with gradient decent and iterative projections [43, 119, 175] or involve to solve a semi-definite programming problem [168, 180], which often suffer from a large computational burden and limit their potential applications to high-dimensional images. Moreover, most of these methods, which can learn Mahalanobis distance metrics from the training samples, are unable to explicitly give the new representations of data in the new metric space. Considering this, in this work we expect to learn a mapping matrix U instead of a Mahalanobis distance metric M . From another point of view, we can also learn a Mahalanobis distance metric M by resorting to the mapping matrix U , i.e., $M = UU^T$.

In this chapter, we present a novel regularized weakly supervised subspace learning framework to explicitly exploit the RF log data for CIR, i.e.,

$$U^* = \arg \min_{U \in R^{h \times l}} f(U, X_l, S, D) + \beta_1 g(U, X_l, S) + \beta_2 r(U, X_l, X_u),$$

where U is the mapping matrix; $f(\cdot)$ is a loss function defined over the labeled images X_l with the associated constraints S and D to reflect the discriminative information; $g(\cdot)$ is a regularizer defined over the labeled images X_l with the associated similar constraints S , which models the geometric information of labeled image pairs; $r(\cdot)$ is also a regularizer, which is defined over the labeled images X_l and unlabeled images X_u on the target mapping matrix U ; β_1 and β_2 are two trade-off parameters, which are used to balance the three terms. The above regularized subspace learning framework is largely inspired by the recent regularization principle in the machine learning community, which is usually the key to enhance the generalization and robustness performance of machine learning techniques. The regularization principle has played a vital role in alleviating the overfitting problem in the machine learning community [181]. For instance, the regularization principle is the most critical aspect in ensuring the good generalization performance in SVM [49, 50, 182]. Similarly, the regularization method is also an effective technique to enhance the stable performance of Fisher's LDA when dealing with small-sized training data in a high-dimensional space [49].

Given the above weakly supervised subspace learning framework, the key issue to attack this problem is to design one appropriate loss function $f(\cdot)$, two regularizer terms $g(\cdot)$ and $r(\cdot)$, and afterward find an efficient algorithm to solve this problem. In the following subsections, we will study some principles for formulating the reasonable loss function, the regularizer terms, and also discuss the solutions to this problem.

6.3.2 Conjunctive Patches Subspace Learning with Side Information

In this work, the CIR system reduces the semantic gap by exploiting the RF log data judged by users in image retrieval and finding a semantic concept subspace to reflect the similar relation between image pairs, thereby further enhancing the performance of the system. We use a linear mapping matrix U to approximate this semantic concept subspace, and then the images in this subspace can be represented as $Y = U^T X = [y_1, y_2, \dots, y_n] \in R^{l \times n} (l < h)$ with $y_i \in R^l$ for an image $x_i \in R^h$. Therefore, in this reduced semantic concept subspace, an improved performance is expected.

In this subsection, we present the CPSL method to learn such a mapping matrix U . Especially, CPSL can effectively integrate the discriminative information of labeled images, the geometric information of labeled images, and the weakly similar information of unlabeled images. Motivated by the recent patch alignment framework [183], this process is conducted by building different kinds of local patches for each image, and then aligning those different kinds of patches together to learn a consistent coordinate through the above regularized subspace learning framework. A patch is a local area, which is formed by an image and its associated nearest images. Particularly, in CPSL, we build three different kinds of patches, which are: 1) local discriminative patches for labeled images to represent the discriminative information; 2) local geometric patches for labeled images to represent the geometric information; and 3) local weakly similar patches for labeled and unlabeled images to represent the weakly similar information of unlabeled images.

6.3.2.1 Local Discriminative Patches for Labeled Images

Given samples with side information, a popular principle for learning a distance metric M is to minimize the distances between samples with similar pairwise constraints and to

maximize the distances between samples with dissimilar pairwise constraints simultaneously, which can be referred to as a min-max principle. In [43], Xing et al. formulated the weakly supervised distance metric learning problem as a constraint convex optimization problem, i.e.,

$$\begin{aligned} \min_M \quad & \sum_{(x_i, x_j) \in S} \|x_i - x_j\|_M^2 \\ \text{s.t.} \quad & \sum_{(x_i, x_j) \in D} \|x_i - x_j\|_M \geq 1. \end{aligned} \quad (6.3)$$

Eq.(6.3) attempts to find the optimal metric M by minimizing the sum of squared distances between samples with similar pairwise constraints, and meanwhile enforcing the sum of distances between samples with dissimilar pairwise constraints larger than or equal to 1. Following this principle, [168] defined two loss functions by minimizing the sum of squared distances between all samples with similar pairwise constraints and maximizing the sum of squared distances between all samples with dissimilar pairwise constraints. Although the above distance metric learning approaches have been demonstrated to be effective for some test data sets, they are essentially linear global methods, and therefore might fail to find the nonlinear structure hidden in the high-dimensional visual feature space.

Following this min-max principle, we define a new loss function for discriminative information preservation. Particularly, for each image x_i associated with a discriminative patch $X_{d(i)} = [x_i, x_{i_1}, x_{i_2}, \dots, x_{i_{k_1}}, x_{i_{k_1+1}}, x_{i_{k_1+2}}, \dots, x_{i_{k_1+k_2}}]$, wherein $x_{i_1}, x_{i_2}, \dots, x_{i_{k_1}}$, i.e., the k_1 nearest images of x_i with similar pairwise constraints, and $x_{i_{k_1+1}}, \dots, x_{i_{k_1+k_2}}$, i.e., the k_2 nearest images with dissimilar pairwise constraints. We define the discriminative loss function as the average difference between two kinds of squared distances over this patch. That is, the discriminative loss function attempts to minimize the average squared distances between each image x_i and its associated k_1 nearest images with similar constraints; meanwhile, it tries to maximize the average squared distances between each image x_i and its associated k_2 nearest images with dissimilar constraints. The illustration of the local discriminative patch for an image is given in Figure 6.2. Especially, for the new representations of each patch, i.e., $y_i, y_{i_1}, y_{i_2}, \dots, y_{i_{k_1}}, y_{i_{k_1+1}}, y_{i_{k_1+2}}, \dots, y_{i_{k_1+k_2}}$, we expect that the loss function between k_1 nearest images with similar constraints and k_2 nearest images with dissimilar constraints will be minimized as much as possible, i.e.,

Figure 6.2: The illustration of the local discriminative patch for an image and its associated nearest similar and dissimilar images. Circular solid dots and square solid dots denote labeled similar and dissimilar images, respectively. For each given image, minimizing the objective function of the local discriminative patch will pull the nearest similar images towards this image while pushing the nearest dissimilar images away from this image in the reduced subspace.

$$f(y_i) = \min_{y_i} \sum_{j=1}^{k_1} \|y_i - y_{i_j}\|^2 \frac{1}{k_1} - \gamma \sum_{j=k_1+1}^{k_1+k_2} \|y_i - y_{i_j}\|^2 \frac{1}{k_2}. \quad (6.4)$$

Eq.(6.4) aims to find the new representation y_i for each sample x_i . To rewrite Eq.(6.4) in a more compact form, we have

$$\begin{aligned} f(y_i) &= \min_{y_i} \sum_{j=1}^{k_1} \|y_i - y_{i_j}\|^2 \frac{1}{k_1} - \gamma \sum_{j=k_1+1}^{k_1+k_2} \|y_i - y_{i_j}\|^2 \frac{1}{k_2} \\ &= \min_{y_i} \text{tr}(Y_{d(i)} \begin{bmatrix} -e_{k_1+k_2}^T \\ I_{k_1+k_2} \end{bmatrix} \text{diag}(w_i) [-e_{k_1+k_2}, I_{k_1+k_2}] Y_{d(i)}^T) \\ &= \min_{y_i} \text{tr}(Y_{d(i)} L_{d(i)} Y_{d(i)}^T), \end{aligned} \quad (6.5)$$

where $w_i = [\overbrace{1/k_1, \dots, 1/k_1}^{k_1}, \overbrace{-\gamma/k_2, \dots, -\gamma/k_2}^{k_2}]^T$; the parameter γ is used to balance the two squared distances; $I_{k_1+k_2}$ is a $(k_1 + k_2) \times (k_1 + k_2)$ identity matrix; $L_{d(i)} = \begin{bmatrix} \sum_{j=1}^{k_1+k_2} (w_i)_j & -w_i^T \\ -w_i & \text{diag}(w_i) \end{bmatrix}$; the vector $e_{k_1+k_2} = [1, \dots, 1]^T \in R^{k_1+k_2}$; and $d(i)$ encodes the discriminative information over this local discriminative patch.

Local Geometric Patch

Figure 6.3: The illustration of the local geometric patch for an image and its associated nearest similar images. Circular solid dots denote labeled similar images. For each given image, the local geometric patch aims to preserve the local geometric information of labeled similar images before and after projection. Minimizing the objective function of the local geometric patch will reconstruct the given image from its associated nearest similar images with a minimal error in the reduced subspace.

6.3.2.2 Local Geometric Patches for Labeled Images

Although the discriminative loss function for each labeled image can capture the discriminative information well, it is empirically known that the geometric information of labeled images can help to find the intrinsic semantic concept subspace. In the past a few years, various geometry based subspace learning methods were proposed to recover the geometric structure of samples in a high-dimensional space. LE [40] minimizes the average Laplacian operators over the manifold of samples and LPP [144] is a linearization version of LE. ISOMAP [134] tries to preserve the pairwise geodesic distances, which can also be used to effectively recover the intrinsic structure of samples in the high-dimensional space. LLE [41] uses the reconstruction coefficients in the high-dimensional space to reconstruct samples from their neighboring samples in the low-dimensional space with minimal errors. In this work, we utilize the LLE technique to preserve the local geometric information for semantic concept subspace learning.

For each image x_i associated with a geometric patch $X_{g(i)} = [x_i, x_{i_1}, x_{i_2}, \dots, x_{i_{k_1}}]$,

wherein $x_{i_1}, x_{i_2}, \dots, x_{i_{k_1}}$, i.e., the k_1 nearest samples of x_i with similar pairwise constraints. As we can see in Figure 6.3, this work assumes that the new representation y_i of an image x_i can be linearly reconstructed by its k_1 nearest images with similar constraints with a minimal error, i.e.,

$$g(y_i) = \min_{y_i} \left\| y_i - \sum_{j=1}^{k_1} c_{i_j} y_{i_j} \right\|^2. \quad (6.6)$$

Eq.(6.6) aims to find the new representation y_i for each sample x_i by using the linear combination coefficients vector c_i . It is used to preserve the local geometric information of labeled images with similar constraints before and after projection, and the linear combination coefficient vector c_i is required to reconstruct x_i from its k_1 nearest similar images with a minimal error, i.e.,

$$\begin{aligned} \min_{c_i} & \left\| x_i - \sum_{j=1}^{k_1} c_{i_j} x_{i_j} \right\|^2 \\ \text{s.t.} & \sum_{j=1}^{k_1} c_{i_j} = 1. \end{aligned} \quad (6.7)$$

Eq.(6.7) is used to obtain the linear combination coefficients vector c_i . To solve this problem, we have $c_{i_j} = \sum_{p=1}^{k_1} G_{jp}^{-1} / (\sum_{s=1}^{k_1} \sum_{t=1}^{k_1} G_{st}^{-1})$ with a local gram matrix $G_{jp} = (x_i - x_{i_j})^T (x_i - x_{i_p})$ as described in [41].

For simplicity, we rewrite Eq.(6.6) in a more compact form. By attaching k_2 nearest dissimilar images of x_i with the local geometric patch $X_g(i)$, we have

$$\begin{aligned} g(y_i) &= \min_{y_i} \left\| y_i - \sum_{j=1}^{k_1} c_{i_j} y_{i_j} \right\|^2 \\ &= \min_{y_i} \left\| y_i - \sum_{j=1}^{k_1} c_{i_j} y_{i_j} - \sum_{j=k_1+1}^{k_1+k_2} 0 \cdot y_{i_j} \right\|^2 \\ &= \min_{y_i} \left\| y_i - \sum_{j=1}^{k_1+k_2} \bar{c}_{i_j} y_{i_j} \right\|^2 \\ &= \min_{y_i} \text{tr} \left(Y_{d(i)} \begin{bmatrix} 1 & -\bar{c}_i^T \\ -\bar{c}_i & \bar{c}_i \bar{c}_i^T \end{bmatrix} Y_{d(i)}^T \right) \\ &= \min_{y_i} \text{tr} (Y_{d(i)} L_{g(i)} (Y_{d(i)})^T), \end{aligned} \quad (6.8)$$

where $Y_{d(i)} = [y_i, y_{i_1}, y_{i_2}, \dots, y_{i_{k_1}}, y_{i_{k_1+1}}, \dots, y_{i_{k_1+k_2}}]$; $L_{g(i)} = \begin{bmatrix} 1 & -\bar{c}_i^T \\ -\bar{c}_i & \bar{c}_i \bar{c}_i^T \end{bmatrix}$ with $\bar{c}_i =$

Figure 6.4: The illustration of the local weakly similar patch for an image and its associated nearest labeled and unlabeled images. Solid dots and hollow dots denote labeled and unlabeled images, respectively. For each given image, the local weakly similar patch attempts to impose local consistency constraints on this image and its associated nearest images. Minimizing the objective function of the local weakly similar patch will incorporate the weakly similar information of unlabeled images in the reduced subspace.

$[c_i^T, \underbrace{0, \dots, 0}_{k_2}]^T$; and $g(i)$ is used to encode the geometric information over this local geometric patch.

6.3.2.3 Local Weakly Similar Patches for Labeled and Unlabeled Images

Recent research has shown that unlabeled samples may be helpful to improve the classification performance. During the last decade, various semi-supervised approaches have attracted increasing amount of attention. In [65], semi-supervised discriminant analysis (SDA) was proposed to find a projection which respects the discriminative structure inferred from the labeled samples, as well as the intrinsic geometric structure inferred from both labeled and unlabeled samples. In [168], Hoi et al. introduced a Laplacian regularizer to a supervised metric learning approach and showed that the semi-supervised metric learning method can learn an effective distance metric by exploiting the unlabeled samples when labeled samples are limited and noisy. Inspired by the recent advance in the semi-supervised learning research, in this part, we design a new regularizer term based on the labeled and unlabeled images, and then introduce this term to our regularized subspace learning framework to find an effective semantic concept subspace.

Unlabeled images are attached to the labeled images: $X = [x_1, \dots, x_n, x_{n+1}, \dots, x_{n+n_u}]$, where the first n images are judged by users in RF, and the remaining n_u images have no label information. For each image $x_i \in X, i = 1, \dots, n+n_u$, we first find its k_3 nearest samples $x_{i_1}, \dots, x_{i_{k_3}}$ in all images including both labeled and unlabeled ones. And then the image x_i and its associated k_3 nearest images $X_{u(i)} = [x_i, x_{i_1}, \dots, x_{i_{k_3}}]$ form a local weakly similar patch. The key to semi-supervised learning is the prior assumption of the local consistency. For subspace learning methods, it can be interpreted as nearest samples will have similar low-dimensional representations. The local weakly similar patch for an image is illustrated in Figure 6.4. Particularly, for the new representations of each patch, i.e., $Y_{u(i)} = [y_i, y_{i_1}, \dots, y_{i_{k_3}}]$, we minimize the sum of the weighted squared distances between y_i and $y_{i_1}, \dots, y_{i_{k_3}}$, and we have

$$r(y_i) = \min_{y_i} \sum_{j=1}^{k_3} \|y_i - y_{i_j}\|^2 \frac{w_{i,j}}{k_3}. \quad (6.9)$$

Eq.(6.9) aims to find the new representation y_i for each sample x_i . Similarly, to rewrite the local weakly similar patch into a compact form, we can rephrase Eq.(6.9) of the patch of y_i as

$$\begin{aligned} r(y_i) &= \min_{y_i} \sum_{j=1}^{k_3} \|y_i - y_{i_j}\|^2 \frac{\omega_{i,j}}{k_3} \\ &= \min_{y_i} \text{tr} \left(Y_{u(i)} \begin{bmatrix} \sum_{j=1}^{k_3} \bar{w}_{i,j} & -\bar{w}_i^T \\ -\bar{w}_i & \text{diag}(\bar{w}_i) \end{bmatrix} Y_{u(i)}^T \right) \\ &= \min_{y_i} \text{tr}(Y_{u(i)} L_{u(i)} Y_{u(i)}^T), \end{aligned} \quad (6.10)$$

where the weight $\omega_{i,j} = \exp(-\|x_i - x_j\|^2/\delta^2)$ is the Laplacian heat kernel according to LE [40]; the patch $Y_{u(i)} = [y_i, y_{i_1}, \dots, y_{i_{k_3}}]$; $L_{u(i)} = \begin{bmatrix} \sum_{j=1}^{k_3} \bar{w}_{i,j} & -\bar{w}_i^T \\ -\bar{w}_i & \text{diag}(\bar{w}_i) \end{bmatrix}$; the vector $\bar{w}_i = \underbrace{[\frac{w_{i,1}}{k_3}, \dots, \frac{w_{i,j}}{k_3}]}_{k_3}$; and $u(i)$ encodes the weakly similar information between labeled images and unlabeled images.

6.3.2.4 CPSL with Side Information

Each of the constructed patches has its own coordinate system. To get a consistent coordinate, we can first align each of these three different kinds of patches together to

obtain a consistent coordinate according to an alignment trick [183, 184], respectively. For each image x_i , the associated patch $Y_i = [y_i, y_{i_1}, \dots, y_{i_k}]$ can be rewritten as $Y_i = Y S_i$, where $Y = [y_1, \dots, y_N]$, $N = n + n_u$ is the number of labeled and unlabeled images and $S_i = R^{N \times (k+1)}$ is the selection matrix. And S_i is defined according to [183, 184] as

$$(S_i)_{st} = \begin{cases} 1, & \text{if } s = F_i(t), \\ 0, & \text{else,} \end{cases} \quad (6.11)$$

where $F_i = [i, i_1, \dots, i_k]$ is the index vector for samples in Y_i .

And then, to get the mapping matrix U , we can integrate all the three different kinds of patches defined in Eq.(6.4), Eq.(6.6), and Eq.(6.9) together through the regularized subspace learning framework in Eq.(6.3), i.e.,

$$\begin{aligned} & \min_U f(U, X_l, S, D) + \beta_1 g(U, X_l, S) + \beta_2 r(U, X_l, X_u) \\ &= \sum_{i=1}^n \min_U \text{tr}(Y_{d(i)} L_{d(i)} Y_{d(i)}^T) + \beta_1 \sum_{i=1}^n \min_U \text{tr}(Y_{d(i)} L_{g(i)} Y_{d(i)}^T) + \beta_2 \sum_{i=1}^{n+n_u} \min_U \text{tr}(Y_{u(i)} L_{u(i)} Y_{u(i)}^T) \\ &= \min_U \text{tr} \left(\sum_{i=1}^n Y_{d(i)} L_{d(i)} Y_{d(i)}^T \right) + \beta_1 \text{tr} \left(\sum_{i=1}^n Y_{d(i)} L_{g(i)} Y_{d(i)}^T \right) + \beta_2 \text{tr} \left(\sum_{i=1}^{n+n_u} Y_{u(i)} L_{u(i)} Y_{u(i)}^T \right) \\ &= \min_U \text{tr} \left(Y \left(\sum_{i=1}^n S_{d(i)} L_{d(i)} S_{d(i)}^T \right) Y^T \right) + \beta_1 \text{tr} \left(Y \left(\sum_{i=1}^n S_{d(i)} L_{g(i)} S_{d(i)}^T \right) Y^T \right) \\ &\quad + \beta_2 \text{tr} \left(Y \left(\sum_{i=1}^{n+n_u} S_{u(i)} L_{u(i)} S_{u(i)}^T \right) Y^T \right) \\ &= \min_U \text{tr} \left(U^T X \left(\sum_{i=1}^n S_{d(i)} L_{d(i)} S_{d(i)}^T \right) X^T U \right) + \beta_1 \text{tr} \left(U^T X \left(\sum_{i=1}^n S_{d(i)} L_{g(i)} S_{d(i)}^T \right) X^T U \right) \\ &\quad + \beta_2 \text{tr} \left(U^T X \left(\sum_{i=1}^{n+n_u} S_{u(i)} L_{u(i)} S_{u(i)}^T \right) X^T U \right) \\ &= \min_U \text{tr} (U^T X (D + \beta_1 G + \beta_2 Q) X^T U), \end{aligned} \quad (6.12)$$

where D encodes the discriminative information and $D = \sum_{i=1}^n (S_{d(i)} L_{d(i)} S_{d(i)}^T)$; G encodes the geometric information and $G = \sum_{i=1}^n (S_{d(i)} L_{g(i)} S_{d(i)}^T)$; Q encodes the weakly similar information of unlabeled images and $Q = \sum_{i=1}^{n+n_u} (S_{u(i)} L_{u(i)} S_{u(i)}^T)$; and $\beta_1, \beta_2 > 0$ are tuning parameters, which are used to trade off the contributions of the three different terms.

The above regularized subspace learning framework can be further improved. Because, in the extreme case, when the two trade-off parameters $\beta_1 \rightarrow 0$ and $\beta_2 \rightarrow 0$, the above optimization problem will result in trivial solutions by shrinking the entire space,

i.e., obtaining the optimal solution of $U^* = 0$. Therefore, we should impose some constraints on the mapping matrix U on Eq.(6.12), and then the problem can be converted to a constraint optimization problem of the mapping matrix U .

Remark I: To avoid trivial solutions and find the mapping matrix U , various different constraints may be used to impose on this optimization problem, which will lead to different constraint optimization problems. A simple constraint $tr(U^T U) = 1$ can be imposed on the objective function. This problem will result in a standard eigenvalue decomposition and U is the eigenvector corresponding to the smallest non-zero eigenvalue. This method always produces rank-one solutions. In other words, the original input space will be projected onto a line by this transformation. However, in many cases it is desirable to obtain a compact low-dimensional representation of the original input feature space.

Remark II: Various distance metric learning methods with side information have been designed to learn such a distance metric M . However, some of these methods define the loss function based on the second-order statistical properties of training samples, which is similar with the discriminative loss function in CPSL, and thus involve to solve a semidefinite programming problem [168, 180]. For example, in [180], Ghodsi et al. defined a loss function, which attempts to minimize the squared induced distances between similar samples while maximizing the squared induced distances between dissimilar samples. Additionally, two constraints are also imposed on this loss function to avoid trivial solutions, i.e.,

$$\begin{aligned} \min_M \quad & \frac{1}{|S|} \sum_{(x_i, x_j) \in S} \|x_i - x_j\|_M^2 - \frac{1}{|D|} \sum_{(x_i, x_j) \in D} \|x_i - x_j\|_M^2 \\ \text{s.t.} \quad & M \succeq 0, \quad tr(M) = 1, \end{aligned} \quad (6.13)$$

where the first constraint ensures a valid metric, and the second constraint excludes trivial solutions where all distances are zeros. This loss function is then converted into a linear objective function and solved by semidefinite programming for finding a proper distance metric M . However, the computational burden of this method is too high, and this significantly limits its potential applications to high-dimensional images.

Although various different constraints can be imposed on Eq.(6.12) to avoid trivial solutions, they are actually arbitrary. Considering this, we impose $U^T U = I$ on Eq.(6.12), to avoid trivial solutions, which can be solved by conducting a standard eigenvalue decomposition and the mapping matrix U is formed by the l eigenvectors associated with

Figure 6.5: The framework of our CIR system. Different from conventional CBIR systems, our system can integrate conventional RF schemes with an offline RF log data exploiting scheme.

the first l smallest eigenvalues. This constraint optimization problem can also lead to closed-form solutions as in [168, 180] but without the runtime inefficiency. Additionally, we can easily obtain the distance metric M by resorting to the mapping matrix U .

6.4 Collaborative Image Retrieval

In this section, we firstly give an overview of our CIR system, which can systematically integrate the user's relevance judgements with a conventional RF scheme for image retrieval. The CIR system assumes that the user expects the best possible results for each query image, i.e., the system is usually required to return the most semantically relevant images based on the previous rounds of RF. Meanwhile, the user will never label a large number of images at each round of RF and only do a few rounds of RF. To deal with this type of scenario, the following CIR framework is proposed.

As shown in Figure 6.5, when a query image is provided, the low-level visual features are first extracted. Then, all images in the database are sorted based on a predefined similarity metric. If the user is satisfied with the results, the image retrieval process can be ended. However, most of the time, the RF scheme is actually required because of the

poor performance of the system. The CIR system requires the user to label some top similar and dissimilar images as the positive and negative feedback samples, respectively. Based on the RF information provided by the user, an RF model can be trained based certain machine learning technique. The similarity metric can be updated together with the RF model. Then, all images are resorted based on the recalculated similarity metric. If the user is satisfied with the refined results, RF is no longer required (i.e., we denote “No” in Figure 6.5) and the system gives the final retrieved results. On the contrary, RF will be performed iteratively (i.e., we denote “Yes” in Figure 6.5).

From Figure 6.5, it can be noticed that our CIR system is different from conventional RF based CBIR systems. The CIR system integrates conventional RF schemes with an offline RF log data exploiting scheme. In Figure 6.5, we can see that our CIR system first collects the user online RF information, which can be stored in a log database. If the RF log data is unavailable, the CIR system performs exactly like conventional CBIR systems. When the RF log data is available, the algorithm can effectively exploit the RF log data. Thus, the CIR system can accomplish the image retrieval task in less iterations than conventional RF based CBIR systems with the help of the RF log data.

Figure 6.6: Some example images in the log database.

We manually divide 10,763 images from the Corel photo gallery into 80 concept groups, such as tiger, lion, elephant, horse, etc., to perform empirical evaluation of the CIR system as in previous research on CBIR [86, 153, 168, 169]. In our experiments, we select the visual features, i.e., color, local descriptors [146], and edges, to represent the images as in previous chapters. The representation of an image will result in a feature vector with 510 values.

Collecting the RF log data is an important step for CIR. However, to our best knowledge, there is no public dataset for the application of exploiting the RF log data in image retrieval. Moreover, in RF, different users are likely to have different opinions on judging similar and dissimilar images with the query image. In our experiments, to conduct objective evaluation and effectively investigate the performance of weakly supervised learning methods, we have to provide a reliable log database to run these weakly supervised algorithms. It is not difficult to build a log database based on an existing real-world database, e.g., the Corel image database. Here, we first randomly select 10 classes according to the ground truth of images from the Corel image database and form an RF log database, which contains 1,385 real-world images. And then, different from the supervised learning task, we divide each class of the database into two groups with an equal size. Therefore, the RF log database comprises 20 groups with 10 different concepts. We randomly select 10 and 30 images uniformly from each group, and therefore we can gather two labeled log databases. The similar constraints are imposed on the images within the same group, while the dissimilar constraints are imposed on the images with different semantic concepts. Finally, we can obtain two log databases with different number of log data, i.e., 200 log data and 600 log data. Some example images in the database are shown in Figure 6.6.

6.5 Experimental Results

In this section, we evaluate the performance of the proposed method in exploiting the RF log data for CIR. We design the experiments for performance evaluation in four aspects. First of all, we use six synthetic datasets to illustrate the effectiveness of the discriminative loss function in finding the most discriminative directions in RF. Secondly, we investigate the performance of the proposed method by exploiting the log database for image retrieval without RF. Then, we report the performance of our CIR system by exploiting the user online RF log data and compare it with a conventional RF scheme (i.e., SVM-based RF) based CBIR system. Finally, we study the sensitivity of important parameters of CPSL.

Figure 6.7: The performance comparisons of three different subspace learning methods, i.e., LDA, BDA, and CPSL, for two sets of samples, i.e., similar samples and dissimilar samples, in RF. In experiments, red “+” and blue “o” denote similar and dissimilar samples, respectively. Black dotted lines, green dot dash lines, and red full lines indicate LDA, BDA, and CPSL, respectively. (a)-(f) show the experimental results of three subspace learning methods when handling samples with various different distributions, respectively.

6.5.1 Experiments with Synthetic Datasets

In order to visualize the effectiveness of the discriminative loss function (i.e., Eq.(6.4)) of CPSL in finding the most discriminative directions in RF, the first experiment is conducted on six synthetic datasets. At each round of RF, the user judges a set of similar and dissimilar images with the query image, which are the positive and negative feedback samples, respectively. The positive and negative feedback samples are generated with various strikingly different distributions since the distributions of feedback samples are usually complicated in real-world applications. Regarding the set of positive feedback samples and the set of negative feedback samples as two different classes, LDA treats the two different classes equally. Based on the assumption that all positive feedback examples are alike, and each negative feedback example is negative in its own way, BDA [32] was proposed to formulate RF as a $(1+x)$ -class discriminant analysis problem. However, it is still not very reasonable to conclude that all positive feedback samples come from a class with a Gaussian distribution. Actually, each positive feedback sample is similar with each of the remaining positive feedback samples, and each negative feedback sample is dissimilar with each positive feedback sample. Consequently, different from traditional supervised learning problems (e.g., LDA and BDA), RF is intrinsically a weakly supervised learning problem and can involve only similar and dissimilar pairwise constraints for the feedback samples. Any unreasonable assumption for the class label information of the feedback samples will result in the performance degradation.

From Figure 6.7, we clearly note that LDA can find the best discriminative directions only when the set of positive feedback samples and the set of negative feedback samples are distributed as Gaussian with similar covariance matrices, as shown in Figure 6.7 (a), but may be confused when the distribution of feedback samples is more complicated, as given in Figure 6.7 (b), (c), (d), (e), and (f). Regarding RF as a $(1+x)$ -class discriminant analysis problem, BDA can only find the direction that the positive feedback samples are well separated with the negative ones when the positive feedback samples have a Gaussian distribution, e.g., Figure 6.7 (c) and (f). However, BDA may also be confused when the distribution of positive feedback samples is more complicated, as shown in Figure 6.7(b), (d), and (e). The discriminative loss function in CPSL only involves the local similar and dissimilar pairwise constraints of feedback samples and does not impose any class

label constraint on the feedback samples, and this is more appropriate for RF in image retrieval. Consequently, the discriminative loss function in CPSL can effectively find the most discriminative subspaces compared with classical supervised subspace learning methods with explicit class label information.

6.5.2 Experiments on the CIR System with RF Log Data

In this subsection, we will evaluate the effectiveness of CPSL based on two experiments: first, we investigate the CPSL method by exploiting the log database for an image retrieval task without RF. And then, we show the performance of our CIR system by exploiting the RF log data and compare it with a conventional RF scheme (i.e., SVM-based RF) based CBIR system on a real-world image database.

6.5.2.1 Performance Evaluation by Exploiting Log Data for Image Retrieval

In this part, we intend to examine if the proposed method is comparable or better than previous representative weakly supervised metric learning methods in the distance metric learning community. We compare CPSL with two major distance metrics (i.e., Euclidean distance metric and Mahalanobis distance metric), three representative weakly supervised metric learning methods (i.e., RCA [175], DCA [116], and Xing [43]). In experiments, we do not compare the proposed method with supervised learning techniques since they require the explicit class label information, which are not suitable for CIR. Moreover, in this subsection, CPSL does not involve any unlabeled samples for fair comparison with RCA, DCA, and Xing. Parameters in each method were determined empirically to achieve its best performance in this work. The parameter sensitivity of CPSL will be carefully analyzed in the next subsection.

All of the compared algorithms are conducted on two log databases as described in Section 6.4, i.e., a log database with 200 log data and a log database with 600 log data. In experiments, 500 query images are first randomly selected from the database and then image retrieval is automatically conducted by a computer. We use the widely used average precision (AP) and average recall (AR) to evaluate the performance of the compared algorithms as in previous chapters. In experiments, we calculate APs and ARs

Figure 6.8: APs and ARs of the compared algorithms, i.e., CPSL, Euclidean distance metric, Mahalanobis distance metric, RCA, DCA, and Xing, for the 200 log data. (a) APs and (b) ARs.

Table 6.1: APs in top N results of the compared algorithms, i.e., Euclidean distance metric, Mahalanobis distance metric, RCA, DCA, Xing, and CPSL, for the log database with 200 log data

Top	10	20	30	40	50	60	70	80	90	100
Euc	55.71	49.46	45.10	41.97	39.46	37.62	36.13	34.78	33.62	32.52
Mah	26.93	21.02	18.10	16.40	15.21	14.28	13.49	12.96	12.63	12.33
RCA	57.68	53.54	50.57	48.20	46.06	44.46	43.02	41.54	40.20	38.97
DCA	59.59	55.08	51.37	48.73	46.68	44.98	43.33	41.79	40.51	39.32
Xing	56.20	52.02	48.95	46.76	44.95	43.20	41.61	40.35	39.12	37.94
CPSL	59.52	55.45	52.72	50.23	48.63	47.11	45.62	44.29	43.09	41.89

Table 6.2: ARs in top N results of the compared algorithms, i.e., Euclidean distance metric, Mahalanobis distance metric, RCA, DCA, Xing, and CPSL, for the log database with 200 log data

Top	10	20	30	40	50	60	70	80	90	100
Euc	5.84	8.63	11.01	13.16	15.09	17.03	18.87	20.54	22.21	23.71
Mah	2.87	3.75	4.55	5.30	6.02	6.68	7.30	7.96	8.64	9.29
RCA	6.06	9.34	12.30	15.05	17.52	19.95	22.23	24.29	26.26	28.10
DCA	6.18	9.50	12.32	15.00	17.56	19.98	22.15	24.18	26.16	28.02
Xing	5.91	9.09	11.93	14.61	17.13	19.41	21.54	23.64	25.58	27.39
CPSL	6.31	9.79	13.03	15.95	18.85	21.56	24.07	26.46	28.75	30.85

over the 500 query images at different positions from top 10 to top 150 to obtain the AP curves and the AR curves.

Figure 6.8 shows the experimental results of the compared algorithms on the database with 200 log data. The detailed results are given in Table 6.1 and Table 6.2. From the results, we can draw several observations. First, we notice that directly using the Euclidean distance metric in the high-dimensional visual feature space is not proper because of the semantic gap. Moreover, a simple Mahalanobis distance metric does not outperform the Euclidean distance metric. In fact, when the number of the log data (i.e., 200 log data) is much less than the dimension of the visual feature space (i.e., 510-dimensional space in this work), the covariance of log data is singular, which significantly degrades the performance of the Mahalanobis distance metric for image retrieval. To avoid the singular problem, a regularization item ($\sigma^2 I, \sigma^2 = 0.01$) is added to the covariance matrix in experiments, which is widely used to enhance the generalization ability of an algorithm. And then, all of the metric learning methods (i.e., RCA, DCA, Xing, and CPSL) can perform better than the Euclidean distance metric by exploiting the log database.

In experiments, the optimal metric learned by RCA is computed as the inverse of the average covariance matrix of the chunklets. Similar to the Mahalanobis distance metric, RCA will also encounter a singular covariance matrix when dealing with high-dimensional images. In experiments, RCA is preceded by constraints based LDA which reduces the dimension to that of CPSL as described in [175]. By doing this, we notice that RCA can show much better performance than the Euclidean distance metric by exploiting the similar pairwise constraints. DCA incorporated the dissimilar constraints into RCA and was formulated as a trace ratio. In [116], the authors proposed to attack this problem by using a direct method as in Fisher’s LDA [79]. However, much discriminative information in the null space of the dissimilar scatter matrix has been discarded in solving this problem [131]. Although DCA incorporates the dissimilar pairwise constraints into RCA, the performance of DCA has been significantly degraded due to the problem of numerical computation in handling this trace ratio. Actually, DCA cannot show better performance than RCA for some results, as shown in top 70 results to top 100 results in Table 6.2. Xing et al. formulated the weakly supervised metric learning as a convex optimization

problem, which can be solved by an iterative projection algorithm. However, this method will involve a high computational burden when dealing with high-dimensional images (i.e., 510-dimensional space in this work), which is always the case in CBIR. CPSL can learn a distance metric M by resorting to the mapping matrix U and solve the formulated constraint optimization problem with a standard eigenvalue decomposition, which is more effective and efficient when handling high-dimensional images but will not meet the problem of numerical computation.

From the results, we can notice that the proposed CPSL can significantly outperform the two major distance metrics and three compared metric learning approaches for overall evaluation. Moreover, we also conduct the same comparisons on the database with 600 log data and the results are shown in Figure 6.9, Table 6.3, and Table 6.4. Similar to the experimental results on the database with 200 log data, CPSL can also show much better performance than the compared weakly supervised metric learning methods when dealing with 600 log data. Additionally, the performance of weakly supervised learning methods on the 600 log data is much better than the corresponding results on the 200 log data since more training data are involved to train a reliable distance metric for image retrieval. It should be noted that the results of the Euclidean distance metric on 600 log data is the same as the corresponding results on 200 log data since no training procedure is involved. Compared with the results on 200 log data, the Mahalanobis distance metric cannot show better performance on 600 log data since similar and dissimilar constraints are actually not utilized to calculate the metric. Moreover, it is difficult to obtain a reliable and stable Mahalanobis distance metric when the number of log data is small and the dimension of data is high. Therefore, it is not proper to directly use the Mahalanobis distance metric for image retrieval when exploiting the RF log data.

6.5.2.2 Performance Evaluation on Our CIR System

In this part, we show the performance of our CIR system by exploiting the user's online RF log data on a large database with 10,763 Corel images and compare it with a conventional RF scheme based CBIR system. SVM-based RF is one of the most popular techniques for image retrieval, which regards RF as a strict online binary classification problem. However, it totally ignores the asymmetric property of training data, that is, all

Figure 6.9: APs and ARs of the compared algorithms, i.e., CPSL, Euclidean distance metric, Mahalanobis distance metric, RCA, DCA, and Xing, for the 600 log data. (a) APs and (b) ARs.

Table 6.3: APs in top N results of the compared algorithms, i.e., Euclidean distance metric, Mahalanobis distance metric, RCA, DCA, Xing, and CPSL, for the log database with 600 log data

Top	10	20	30	40	50	60	70	80	90	100
Euc	55.71	49.46	45.10	41.97	39.46	37.62	36.13	34.78	33.62	32.52
Mah	21.45	17.60	16.09	15.18	14.51	13.87	13.36	12.86	12.62	12.44
RCA	62.04	57.44	54.03	51.44	49.22	47.25	45.44	43.75	42.40	41.19
DCA	64.25	59.86	56.32	53.68	51.36	49.39	47.51	45.76	44.29	42.87
Xing	60.09	55.48	52.26	49.49	47.25	45.33	43.71	42.02	40.71	39.50
CPSL	71.11	67.33	64.35	61.91	59.93	58.10	56.53	54.81	53.08	51.47

Table 6.4: ARs in top N results of the compared algorithms, i.e., Euclidean distance metric, Mahalanobis distance metric, RCA, DCA, Xing, and CPSL, for the log database with 600 log data

Top	10	20	30	40	50	60	70	80	90	100
Euc	5.84	8.63	11.01	13.16	15.09	17.03	18.87	20.54	22.21	23.71
Mah	2.32	3.16	4.05	4.92	5.75	6.50	7.24	7.92	8.70	9.51
RCA	6.44	9.94	13.06	15.95	18.61	21.09	23.38	25.47	27.57	29.61
DCA	6.68	10.34	13.57	16.55	19.30	21.90	24.27	26.44	28.59	30.56
Xing	6.28	9.64	12.68	15.43	17.97	20.34	22.58	24.56	26.57	28.47
CPSL	7.63	11.99	16.01	19.76	23.38	26.79	30.07	33.03	35.72	38.28

Figure 6.10: APs in top 50 results and top 100 results of the three different RF methods, i.e., SVM-based RF, CPSL-I SVM-based RF, and CPSL-II SVM-based RF, for image retrieval.

positive feedback samples share a common concept while each negative feedback sample differs in diverse concepts. Moreover, it does not take into account the unlabeled samples although they are very helpful in constructing the optimal classifier. With an assumption that different semantic concepts live in different subspaces and each image can live in many different semantic concept subspaces, it is the goal of RF to figure out “which one”. However, it will be a burden for SVM-based RF to tune the internal parameters to adapt to the changes of such subspaces. In this subsection, we show that our CIR system can effectively address the two drawbacks by offline exploiting the user’s online RF log data.

First, 400 query images are randomly selected from the database and RF is automatically conducted by a computer. At each round of RF, the first 3 relevant images are marked as the positive feedback samples and all other irrelevant images in top 20 results are marked as the negative feedback samples. The procedure is close to real-world circumstances since the irrelevant images usually largely outnumber the relevant ones in real-world applications.

We compare the traditional SVM-based RF with two new SVM-based RF schemes, i.e., CPSL-I SVM-based RF and CPSL-II SVM-based RF. The traditional SVM implements RF in the original high-dimensional low-level visual feature space. CPSL-I SVM-based RF first exploits the user’s online RF log data by finding a semantic subspace,

in which all positive feedback samples are clustered and all negative feedback samples are separated from all positive feedback samples as much as possible. And then SVM implements RF in the reduced semantic subspace. CPSL-II SVM-based RF incorporates the information of unlabeled samples into CPSL-I SVM-based RF through a regularized learning framework. In experiments, CPSL-I and CPSL-II are implemented by setting the trade-off parameters $\beta_2 = 0$ and $\beta_2 = 1/(n^u)$, respectively. For patch building parameters, we set $k_1, k_2, k_3 = 4$ according to manifold learning approaches [40, 41, 134, 136, 144]. Considering the computable efficiency, we randomly select $n_u = 400$ unlabeled images at each round of RF. The optimal dimensions of the reduced subspace for CPSL-I and CPSL-II are empirically set in experiments to preserve more geometric information of the images. For all SVM based methods, we empirically set the kernel parameters to achieve their best performance in experiments. Figure 6.10 (a) and (b) show APs in top 50 results and top 100 results, respectively.

CPSL-I can effectively exploit the user’s online RF log data and find a semantic concept subspace, in which all positive feedback samples are clustered and all negative feedback samples are separated from the positive feedback samples as much as possible. And then, SVM-based RF is implemented in this reduced semantic concept subspace for an image retrieval task. From the results, we notice that CPSL-I SVM-based RF can outperform the traditional SVM-based RF by exploiting the user’s online RF log data. However, the performance difference between CPSL-I SVM-based RF and the traditional SVM-based RF gets smaller after a few rounds of RF because of the overfitting problem. CPSL-II SVM-based RF can effectively integrate the information of unlabeled samples through a regularized learning framework into the construction of the classifier and alleviates the overfitting problem encountered by CPSL-I SVM-based RF. As shown in Figure 6.10, when considering more rounds of RF, CPSL-II SVM-based RF is more effective than both CPSL-I SVM-based RF and the traditional SVM-based RF.

In experiments, it is difficult to calculate the actual time when conducting an image retrieval task and the mapping matrix U can be obtained by using a standard eigenvalue decomposition. The time cost to calculate U is $O((n + n_u)^3)$. Afterwards, we project all images to this semantic concept subspace and then apply the new similarity metric with respect to the query image to sort all images in the database. The time cost for

calculating the Euclidean distance in the l -dimensional semantic subspace between the query image and all images in the database is $O(nl)$, wherein n is the cardinality of the database. Therefore, for a query image, the time cost for CPSL is $O((n + n_u)^3) + O(nl)$. And the time cost for a conventional CBIR system in the h -dimensional visual feature space is $O(nh)$. In general, for a CBIR system, the cardinality of the database n is very large and $h \gg l$; therefore, the proposed method is very effective for image retrieval.

6.5.3 Parameter Sensitivity

In this subsection, we study the parameter sensitivity of the CPSL method for an image retrieval task. The analysis is performed based on experiments conducted on two log databases (i.e., 200 log data and 600 log data). In experiments, we analyze some factors: k_1 and k_2 in Eq.(6.4) for patch building, the trade-off parameter β_1 in Eq.(6.12), and the dimension of the projected subspace for CPSL. First, 500 query images are randomly selected from the database, and then image retrieval is automatically conducted by a computer. The APs in top 50 results are utilized for overall performance evaluation.

6.5.3.1 Evaluation on Number of Nearest Images

The two parameters k_1 and k_2 in Eq.(6.4) play an important role in building the local discriminative patch, and this is the most critical aspect in CPSL. Generally, for a local discriminative patch, k_1 is the number of similar images which are involved to describe the compactness of the patch, and k_2 is the number of dissimilar images which are used to characterize the dispersiveness of the patch. Both of the two parameters (i.e., k_1 and k_2) reveal the data information from different aspects. In experiments, the trade-off parameter β_1 is set as 0 for alleviating the effect of the geometric information and the reduced dimensions for the two sets of log data are empirically fixed at 11 and 17, respectively. By varying k_1 and k_2 , Figure 6.11(a) and (b) show the AP surface of CPSL subject to different k_1 and k_2 for the two log databases, respectively. From Figure 6.11, we can notice that the two parameters k_1 and k_2 can significantly affect the performance of CPSL in learning the semantic concept subspace for image retrieval. As given in Figure 6.11 (a), when k_1 and k_2 are larger than 4 and 10, respectively, the system can show more stable performance for 200 log data. Similarly, in Figure 6.11(b), when k_1 and k_2 are

Figure 6.11: The AP surface of CPSL subject to different k_1 and k_2 for the two log databases, i.e., (a) 200 log data and (b) 600 log data.

larger than 8 and 10, respectively, CPSL can achieve more satisfactory results for 600 log data. Generally, smaller values of k_1 and k_2 mean that fewer similar and dissimilar images are involved to construct the local discriminative patch, and therefore insufficient training samples lead to degenerated performance of the system.

6.5.3.2 Evaluation on Trade-Off Parameter β_1

Empirically, the geometric information is useful for finding the semantic concept subspace. In this part, we turn to investigate the influence of the trade-off parameter β_1 in Eq.(6.12) for CPSL when building the local discriminative patch and the local geometric patch for labeled log data. A small β_1 reflects the importance of separating the dissimilar samples from the similar ones, i.e., CPSL focuses on the local discriminative information but ignores the local geometric information. Figure 6.12 shows the performance of CPSL with different β_1 , from which we can have the following observations.

When β_1 is small, e.g., $\beta_1 = 0$, the performance is unsatisfactory. This is because that in this situation the local discriminative information is mainly preserved while important local geometric information within labeled images with similar pairwise constraints is less considered. The performance of CPSL increases when β_1 is growing and reaches the optimal value at $\beta_1 = 5$. And then, APs decrease when β_1 is larger than this best setup,

Figure 6.12: Performance of CPSL with different β_1 for the two log databases, i.e., (a) 200 log data and (b) 600 log data.

in which case the local geometric information dominates the local patch and the local discriminative information is ignored.

Therefore, both the discriminative information and the geometric information can reflect the important information contained in local patches from different aspects for complimentary. A suitable combination of them is essential to achieve good performance for CPSL.

6.5.3.3 Evaluation on Dimension of the Projected Subspace

Different from the weakly supervised distance metric learning methods [43, 116, 175], the proposed CPSL aims to learn a mapping matrix, which can find a low-dimensional semantic concept subspace from the original high-dimensional visual feature space. To find out an appropriate dimension of the projected semantic concept subspace, we have investigated the influence of the dimension in the following experiments. Figure 6.13 shows the performance of CPSL with features projected onto the subspaces with different dimensions. From Figure 6.13, we can notice that when the projected dimension is too low, (e.g., less than 11 and 17, respectively), the reduced subspace is insufficient to encode the semantic concepts of images, which makes the performance poor. When

Figure 6.13: Performance of CPSL with features projected onto the subspaces with different dimensions for the two log databases, i.e., (a) 200 log data and (b) 600 log data.

the dimension equals or closes to that of the original high-dimensional space (i.e., 510-dimensional space in this work), no or less benefit can be obtained from subspace learning. From the experimental results, we can notice that CPSL can achieve its best performance with the dimension of 11 and 17 for the two log databases, respectively. Moreover, lower dimensional data can lead to a less computational burden than higher dimensional data for image retrieval.

6.6 Summary

In this chapter, we have studied the problem of subspace learning with side information and presented a novel subspace learning method, termed as conjunctive patches subspace learning (CPSL) with side information to exploit the RF log data for collaborative image retrieval (CIR). The proposed method can effectively integrate the discriminative information of labeled images, the geometric information of labeled images, and the weakly similar information of unlabeled images together through a regularized learning framework. We have formally formulated this subspace learning problem as a constraint optimization task, and then present an effective algorithm to solve this problem with closed-form solutions. Extensive experimental results on both synthetic datasets and

a real-world image database have shown the effectiveness of the proposed method in exploiting the RF log data for CIR.

Chapter 7

Thesis Conclusions and Future Work

7.1 Thesis Conclusions

This thesis investigates content-based image retrieval (CBIR) [1–3]. This is because CBIR has many open problems, although it has been deemed as one of the most promising technologies in the field of information technology and it has many potential practical applications, e.g., multimedia information retrieval [19, 185, 186], biomedical information management [187–189], copy detection [190–192], digital libraries [193–195], architectural design [196], etc.

Relevance feedback (RF) [22] is a powerful tool to reduce the gap between low-level visual features and high-level semantic concepts in the CBIR system by letting the user label semantically relevant and irrelevant images as the positive and negative feedback samples, respectively. However, RF is different from conventional machine learning problems because the user would not like to label a large number of feedback samples [1–3]. Although substantial research work has been widely conducted, CBIR is still an open research topic mainly due to the difficulties in bridging the semantic gap.

The first problem to be attacked in this thesis is the small-sized training data problem. In CBIR, the user is usually impatient and would not like to label a large number of samples at each round of RF. Hence, the number of training samples will be much less than the dimension of the visual feature space. The small-sized training data problem will lead to the singular problem in conventional discriminant analysis based RF methods, e.g., biased discriminant analysis (BDA). To alleviate the singular problem in the BDA model,

we propose a generalized BDA (GBDA) method by adopting the differential scatter discriminant criterion (DSDC) [46, 126, 137–140], which defines the interclass separability as a trace difference for the between-class scatter matrix and the within-class scatter matrix rather than a trace ratio. By redesigning the between-class scatter matrix, GBDA can also avoid the Gaussian distribution assumption for the positive feedback samples in the BDA model. Moreover, to reduce the overfitting problem, the locality preserving principle emerging from the manifold learning community [41, 134, 135], which measures the local smoothness of the feature transformation, is also integrated to regularize the interclass separability. Therefore, a locally smooth and consistent transformation can also be learned. Extensive experiments on a real-world image database have shown that the proposed GBDA can significantly outperform the original BDA and its enhanced versions.

The second problem is the asymmetric property of training data problem. In CBIR, most of the conventional classification based RF methods, e.g., support vector machine (SVM)-based RF, treat the positive and negative feedback samples equally although this is not appropriate since all positive feedback samples share a common concept while each negative feedback sample differs in diverse concepts. To incorporate the asymmetric property of training data, we propose a biased maximum margin analysis (BMMA) method, which can be combined with the popular classification based RF, i.e., SVM-based RF, to better model the RF procedure and reduce the performance degradation caused by the asymmetric property of training data. Moreover, semi-supervised BMMA (SemiBMMA) is also introduced to integrate the information of unlabeled samples to BMMA and thus effectively alleviates the overfitting problem caused by a small number of training feedback samples. Extensive experiments on both synthetic datasets and a real-world image database have been conducted to show the proposed scheme combined with SVM-based RF can significantly improve the performance of the CBIR system.

The third problem is the selection of most informative samples problem. In CBIR, most of the conventional RF methods can only select relevant and irrelevant images in top returned results for the user to label although this is not a good way since top returned results may not be the most informative ones to define an effective similarity metric for image retrieval. To select the most informative samples, we propose a geometric

optimum experimental design (GOED) method to find multiple representative samples in the database as the most informative ones for the user to label and thus to alleviate the labeling efforts of conventional RF methods. The proposed GOED can alleviate the small-sized training data problem by leveraging the geometric structure of unlabeled samples in the reproducing kernel Hilbert space (RKHS), and can further enhance the performance of image retrieval. By minimizing the expected average prediction variance on the test data, GOED has a clear geometric interpretation to select the most representative samples in the database iteratively with the global optimum, which is more effective and efficient for the user to label. Extensive experiments on both synthetic datasets and a real-world image database have shown the advantages of the proposed GOED.

The fourth problem is the exploitation of RF log data problem. Conventional CBIR systems cannot accumulate and utilize the RF information provided by a number of users, which can be regarded as the RF log data in image retrieval. From a long-term perspective, the RF log data is an important and useful resource to enhance the performance of the CBIR system. To exploit the RF log data, we propose conjunctive patches subspace learning (CPSL) with side information to directly learn a semantic concept subspace from the RF log data with a set of similar and dissimilar pairwise constraints. Especially, CPSL can effectively integrate the discriminative information of labeled images, the geometric information of labeled images, and the weakly similar information of unlabeled images together. This process is conducted by building different kinds of local patches for each image, and then aligning those different kinds of patches together to learn a consistent coordinate through a regularized learning framework. We formally formulate this problem as a constraint optimization problem, and then present an effective algorithm to solve this task with closed-form solutions. Extensive experiments on both synthetic datasets and a real-world image database demonstrate the effectiveness of the proposed scheme in improving the performance of CBIR by exploiting the RF log data.

In summary, this thesis deals with non-trivial problems in CBIR, which are how to apply statistical machine learning techniques to maximize the potential of conventional RF methods to significantly improve the performance of the CBIR system. The primary contributions of the thesis are as follows:

1) We have developed a GBDA method as a principled way to alleviate the small-sized training data problem or the singular problem in the conventional discriminant analysis based RF, i.e., BDA, for CBIR. Moreover, by redesigning the between-class scatter matrix and integrating the locality preserving principle, GBDA can also alleviate the Gaussian distribution assumption for the positive feedback samples and the overfitting problem in BDA.

2) We have proposed a BMMA method to incorporate the asymmetric property of training data with the popular conventional classification based RF, i.e., SVM-based RF, for CBIR. BMMA can separate the positive feedback samples from the negative ones by a maximum margin in the reduced subspace. The SVM-based RF combined with BMMA can better model the RF procedure and reduces the performance degradation caused by the asymmetric property of training data. Moreover, by introducing a Laplacian regularizer to BMMA, we also design a SemiBMMA method to utilize the information of unlabeled samples for SVM-based RF to alleviate the overfitting problem caused by the small number of training feedback samples.

3) We have designed a GOED method to select multiple representative samples in the database as the most informative ones for the user to label. GOED can alleviate the small-sized training data problem by leveraging the geometric structure of unlabeled samples in RKHS and also has a clear geometric interpretation to select a set of representative samples in the database iteratively with the global optimum. Moreover, GOED is label-independent and can avoid various potential problems caused by insufficient and inexact labeled samples in RF, which is more appropriate and useful compared with SVM active learning (SVMactive) for image retrieval.

4) Lastly, we have developed CPSL with side information to explicitly exploit the RF log data for collaborative image retrieval (CIR). The significant of this method is that it can directly learn a semantic concept subspace from a set of similar and dissimilar pairwise constraints without using any class label information, which is more practical and useful in real-world applications. Compared with previous distance metric learning methods, which usually involve a convex optimization procedure or a semidefinite programming procedure, CPSL can also learn a distance metric but performs more effectively and efficiently when dealing with high-dimensional data.

7.2 Future Work

Although a number of promising results have been reported to address the intrinsic problems in CBIR in this thesis, it is interesting to further exploit the intrinsic connections between these statistical machine learning techniques. And there are still several open issues that need to be further investigated regarding the CBIR research in future work. An important one is the imbalanced distribution of positive and negative feedback samples, and this is one of the most critical issues in RF. It is mainly because in a real-world CBIR system, there are usually more negative feedback samples than positive ones. Because of the imbalance of training data for the two classes, traditional machine learning techniques will be biased toward the negative feedback samples. Consequently, conventional RF methods may mistake many query irrelevant images as relevant ones.

The second challenge is the computational efficiency issue. Although we have developed efficient algorithms to attack the problems in conventional RF methods for CBIR, some of these methods involve a dense-matrix eigenvalue decomposition, which can be computationally expensive both in time and memory. Therefore, an effective computational technique is required to alleviate this drawback.

Then, an important challenge in building a CBIR system is to account for the difference between human perception of similarity and machine perception of similarity. Machine perception usually considers overall similarity between the query image and images in the database, whereas human perception may be sometimes difficult to describe and involves semantic understanding of images. Recent research from neuroscience has demonstrated that biologically inspired models are attractive in visual recognition. Human perception based models aim to provide information which area is important in an image from human perception. This may help priority learning, pattern recognition, active vision, etc, for an image retrieval task.

Moreover, the localization of the query object in an image is a key part in a conventional CBIR system. Recent research has shown that by localizing actions in a crowded-scene video, the performance of an action search system can be significantly enhanced [197]. As a consequence, more research efforts are required to design effective object localization techniques in an image and they can further improve the performance of the CBIR system.

Last but not least, large-scale web data can be utilized to facilitate the current CBIR system. Recently, thousands of millions of data in various formats (e.g., texts, images, videos, etc.) are available on the Internet. When people are dealing with the task at hand, a natural and effective way is to ask for help from the Internet in terms of various data. In the computer vision community, several methods have been proposed to learn robust classifiers by leveraging the web data from online sources (e.g., Flickr, Youtube, photoSIG, etc.) [198–200]. Therefore, it is more attractive to utilize the large-scale web data to help the user retrieve images from an image database and further enhance the performance of the CBIR system.

Publications

- (i) L. Zhang, L. Wang, and W. Lin, “Generalized Biased Discriminant Analysis for Content Based Image Retrieval”, IEEE Transactions on Systems, Man, Cybernetics, Part B - Cybernetics , Vol.42, No.1, pp.282-290, Feb., 2012. [Chapter 3]
- (ii) L. Zhang, L. Wang, and W. Lin, “Semi-supervised Biased Maximum Analysis for Interactive Image Retrieval”, IEEE Transactions on Image Processing , Vol.21, No.4, pp.2294-2308, April, 2012. [Chapter 4]
- (iii) L. Zhang, L. Wang, and W. Lin, “Conjunctive Patches Subspace Learning with Side Information for Collaborative Image Retrieval”, IEEE Transactions on Image Processing, Vol.21, No.8, pp. 3707-3720, Aug., 2012. [Chapter 6]
- (iv) L. Zhang, L. Wang, and W. Lin, “Geometric Optimum Experimental Design for Collaborative Image Retrieval”, IEEE Transactions on Circuits and Systems for Video Technology (in press). [Chapter 5]
- (v) L. Zhang, L. Wang, and W. Lin, “Laplacian Regularized Subspace Learning for Interactive Image Re-ranking”, In: Proceedings of the 2012 IEEE International Joint Conference on Neural networks, Brisbane, Australia, 10-15 Jun., 2012.
- (vi) L. Zhang, L. Wang, and W. Lin, “Biased Subspace Learning for SVM Relevance Feedback in Content Based Image Retrieval”, In: Proceedings of the 2011 IEEE International Conference on Information, Communications and Signal Processing, Singapore, 13- 16 Dec., 2011.
- (vii) L. Zhang, L. Wang, and W. Lin, “A Sematic Subspace Learning Method to Exploit Relevance Feedback Log Data for Image Retrieval”, In: Proceedings of the 2013 IEEE Symposium Series on Computational Intelligence, Singapore, April, 2013.

References

- [1] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1349–1380, Dec. 2000.
- [2] 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*, vol. 10, no. 1, pp. 39–62, 1999.
- [3] R. Datta, D. Joshi, J. Li, and J. Z. Wang, “Image retrieval: ideas, influences, and trends of the new age,” *ACM Computing Surveys*, vol. 40, pp. 1–60, May 2008.
- [4] N. Chang and K. Fu, “Query-by-pictorial-example,” *IEEE Transactions on Software Engineering*, vol. 6, pp. 519–524, Nov. 1980.
- [5] N. Roussopoulos, C. Faloutsos, and T. Sellis, “An efficient pictorial database system for psql,” *IEEE Transactions on Software Engineering*, vol. 14, pp. 639–650, May 1988.
- [6] S. K. Chang, C. W. Yan, D. C. Dimitroff, and T. Arndt, “An intelligent image database system,” *IEEE Transactions on Software Engineering*, vol. 14, pp. 681–688, May 1988.
- [7] L. Cao, J. Luo, and T. S. Huang, “Annotating photo collections by label propagation according to multiple similarity cues,” in *Proceedings of the 16th ACM International Conference on Multimedia*, pp. 121–130, 2008.
- [8] S. Chang, D. Ellis, W. Jiang, K. Lee, A. Yanagawa, A. C. Loui, and J. Luo, “Large-scale multimodal semantic concept detection for consumer video,” in *Proceedings*

- of the International Workshop on Workshop on Multimedia Information Retrieval*, pp. 255–264, 2007.
- [9] J. Li and J. Z. Wang, “Real-time computerized annotation of pictures,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 985–1002, June 2008.
- [10] M. J. Swain and D. H. Ballard, “Color indexing,” *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991.
- [11] G. Pass, R. Zabih, and J. Miller, “Comparing images using color coherence vectors,” in *Proceedings of the 4th ACM International Conference on Multimedia*, pp. 65–73, 1996.
- [12] Y. Rubner, J. Puzicha, C. Tomasi, and J. M. Buhmann, “Empirical evaluation of dissimilarity measures for color and texture,” *Computer Vision and Image Understanding*, vol. 84, no. 1, pp. 25 – 43, 2001.
- [13] H. Tamura, S. Mori, and T. Yamawaki, “Textural features corresponding to visual perception,” *IEEE Transactions on Systems, Man and Cybernetics- Part B: Cybernetics*, vol. 8, no. 6, pp. 460 –473, June,1978.
- [14] J. Mao and A. K. Jain, “Texture classification and segmentation using multiresolution simultaneous autoregressive models,” *Pattern Recognition*, vol. 25, no. 2, pp. 173 – 188, 1992.
- [15] A. K. Jain and A. Vailaya, “Image retrieval using color and shape,” *Pattern Recognition*, vol. 29, no. 8, pp. 1233 – 1244, 1996.
- [16] C. W. Niblack, R. Barber, W. Equitz, M. D. Flickner, E. H. Glasman, D. Petkovic, P. Yanker, C. Faloutsos, and G. Taubin, “Qbic project: querying images by content, using color, texture, and shape,” in *SPIE’s Symposium on Electronic Imaging: Science and Technology*, pp. 173–187, 1993.
- [17] A. K. Jain and A. Vailaya, “Shape-based retrieval: A case study with trademark image databases,” *Pattern Recognition*, vol. 31, pp. 1369–1390, Sept. 1998.

- [18] 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," *Computer*, vol. 28, pp. 23–32, 1995.
- [19] J. R. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. Jain, and C. Shu, "Virage image search engine: an open framework for image management," in *Proceedings of the Society of Photo-Optical Instrumentation Engineers Conference Series*, vol. 2670, pp. 76–87, 1996.
- [20] A. Pentland, R. Picard, and S. Sclaroff, "Photobook: Content-based manipulation of image databases," *International Journal of Computer Vision*, vol. 18, no. 3, pp. 233–254, 1996.
- [21] T. S. Huang and M. R. Naphade, "Multimedia analysis and retrieval system (mars) project," in *Proceedings of the 33rd Annual Clinic on Library Application of Data Processing - Digital Image Access and Retrieval*, 1996.
- [22] X. S. Zhou and T. S. Huang, "Relevance feedback in image retrieval: A comprehensive review," *Multimedia Systems*, vol. 8, pp. 536–544, Apr. 2003.
- [23] G. Salton, ed., *Automatic text processing*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1988.
- [24] X. He, W. Ma, and H. Zhang, "Learning an image manifold for retrieval," in *Proceedings of the 12th ACM International Conference on Multimedia*, pp. 17–23, 2004.
- [25] J. He, M. Li, H. Zhang, H. Tong, and C. Zhang, "Manifold-ranking based image retrieval," in *Proceedings of the 12th ACM International Conference on Multimedia*, 2004.
- [26] C. H. Hoi, M. R. Lyu, and R. Jin, "A unified log-based relevance feedback scheme for image retrieval," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 4, pp. 509 – 524, 2006.

- [27] C. H. Hoi, R. Jin, J. Zhu, and M. R. Lyu, “Semisupervised svm batch mode active learning with applications to image retrieval,” *ACM Transactions on Information System*, vol. 27, pp. 16:1–16:29, May 2009.
- [28] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang, “Face recognition using laplacian-faces,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 328–340, Mar. 2005.
- [29] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, “Eigenfaces vs. fisherfaces: recognition using class specific linear projection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 711–720, July 1997.
- [30] D. L. Swets and J. Weng, “Hierarchical discriminant analysis for image retrieval,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, pp. 386–401, May 1999.
- [31] D. L. Swets and J. Weng, “Using discriminant eigenfeatures for image retrieval,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, pp. 831–836, Aug. 1996.
- [32] X. S. Zhou and T. S. Huang, “Small sample learning during multimedia retrieval using biasmap,” in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 11–17, 2001.
- [33] D. Gorgevik, D. Cakmakov, and V. Radevski, “Handwritten digit recognition by combining support vector machines using rule-based reasoning,” in *Proceedings of the 23rd International Conference on Information Technology Interfaces*, pp. 139–144, 2001.
- [34] S. Tong and E. Chang, “Support vector machine active learning for image retrieval,” in *Proceedings of the 9th ACM International Conference on Multimedia*, pp. 107–118, 2001.
- [35] P. Hong, Q. Tian, and T. S. Huang, “Incorporate support vector machines to content-based image retrieval with relevance feedback,” in *Proceedings of the IEEE International Conference on Image Processing*, pp. 750–753, 2000.

- [36] D. Tao, X. Tang, X. Li, and X. Wu, “Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 1088 – 1099, July 2006.
- [37] G. Guo, A. Jain, W. Ma, and H. Zhang, “Learning similarity measure for natural image retrieval with relevance feedback,” *IEEE Transactions on Neural Networks*, vol. 13, pp. 811 – 820, July 2002.
- [38] L. Wang, K. Chan, and Z. Zhang, “Bootstrapping SVM active learning by incorporating unlabelled images for image retrieval,” in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2003*.
- [39] I. T. Jolliffe and MyiLibrary, *Principal component analysis*. Wiley Online Library, 2002.
- [40] M. Belkin and P. Niyogi, “Laplacian eigenmaps and spectral techniques for embedding and clustering,” in *Advances in Neural Information Processing Systems*, 2002.
- [41] S. T. Roweis and L. K. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” *Science*, vol. 290, no. 5500, p. 2323, 2000.
- [42] R. A. Fisher, “The use of multiple measurements in taxonomic problems,” *Annals of Human Genetics*, vol. 7, no. 2, pp. 179–188, 1936.
- [43] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell, “Distance metric learning with application to clustering with side-information,” in *Advances in Neural Information Processing Systems*, pp. 521–528, 2003.
- [44] M. Belkin, N. Partha, and V. Sindhwani, “Manifold regularization: A geometric framework for learning from labeled and unlabeled examples,” *Journal of Machine Learning Research*, vol. 7, pp. 2399–2434, Dec. 2006.
- [45] A. C. Atkinson and A. N. Donev, *Optimum Experimental Designs*. Oxford,U.K., Oxford Univ. Press, 2007.

- [46] D. Tao, X. Li, X. Wu, and S. J. Maybank, “General tensor discriminant analysis and gabor features for gait recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1700–1715, 2007.
- [47] T. K. Kim and J. Kittler, “Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image,” *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 318–327, Mar. 2005.
- [48] S. Dudoit, J. Fridlyand, and T. P. Speed, “Comparison of discrimination methods for the classification of tumors using gene expression data,” *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 77–87, 2002.
- [49] V. N. Vapnik, *The nature of statistical learning theory*. Springer Verlag, 2000.
- [50] L. Wang, *Support Vector Machines: Theory and Applications*. Springer Berlin, 2005.
- [51] D. Gao, J. Zhou, and L. Xin, “Svm based detection of moving vehicles for automatic traffic monitoring,” in *Proceedings of the International Conference of Intelligent Transportation Systems*, pp. 745–749, 2001.
- [52] Q. Zhao and J. C. Principe, “Support vector machines for sar automatic target recognition,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 37, no. 2, pp. 643–654, 2001.
- [53] M. Pontil and A. Verri, “Support vector machines for 3d object recognition,” *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 637–646, June 1998.
- [54] A. K. Jain and B. Yu, “Automatic text location in images and video frames,” *Pattern Recognition*, vol. 31, no. 12, pp. 2055–2076, 1998.
- [55] T. Joachims, “Text categorization with suport vector machines: Learning with many relevant features,” in *Proceedings of the 10th European Conference on Machine Learning*, pp. 137–142, 1998.

- [56] S. Tong and D. Koller, “Support vector machine active learning with applications to text classification,” *Journal of Machine Learning Research*, vol. 2, pp. 45–66, 2002.
- [57] B. H. Aharon, T. Hertz, N. Shental, and D. Weinshall, “Learning a mahalanobis metric from equivalence constraints,” *Journal of Machine Learning Research*, vol. 6, pp. 937–965, Dec. 2005.
- [58] A. Blum and T. Mitchell, “Combining labeled and unlabeled data with co-training,” in *Proceedings of the 7th Annual Conference on Computational Learning Theory*, pp. 92–100, 1998.
- [59] C. Rosenberg, M. Hebert, and H. Schneiderman, “Semi-supervised self-training of object detection models,” in *Proceedings of the 7th IEEE Workshops on Application of Computer Vision*, pp. 29–36, 2005.
- [60] T. Joachims, “Transductive inference for text classification using support vector machines,” in *Proceedings of the 6th International Conference on Machine Learning*, pp. 200–209, 1999.
- [61] M. Belkin and P. Niyogi, “Semi-supervised learning on riemannian manifolds,” *Machine Learning*, vol. 56, pp. 209–239, June 2004.
- [62] X. Zhu, Z. Ghahramani, and J. Lafferty, “Semi-supervised learning using gaussian fields and harmonic functions,” in *Proceedings of the 20th International Conference on Machine Learning*, vol. 20, pp. 912–920, 2003.
- [63] X. Zhu and J. Lafferty, “Harmonic mixtures: combining mixture models and graph-based methods for inductive and scalable semi-supervised learning,” in *Proceedings of the 22nd International Conference on Machine Learning*, pp. 1052–1059, 2005.
- [64] M. Belkin and P. Niyogi, “Using manifold structure for partially labeled classification,” in *Advances in Neural Information Processing Systems*, vol. 15, pp. 929–936, 2002.

- [65] D. Cai, X. He, and J. Han, “Semi-supervised discriminant analysis,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2007.
- [66] V. Sindhwani, P. Niyogi, and M. Belkin, “Beyond the point cloud: from transductive to semi-supervised learning,” in *Proceedings of the 22nd International Conference on Machine Learning*, pp. 824–831, 2005.
- [67] O. Chapelle, B. Schölkopf, A. Zien, *et al.*, *Semi-supervised learning*. MIT press Cambridge, MA:, 2006.
- [68] B. Geng, D. Tao, C. Xu, L. Yang, and X. Hua, “Ensemble manifold regularization,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 1227–1233, 2012.
- [69] T. Evgeniou, M. Pontil, and T. Poggio, “Regularization networks and support vector machines,” *Advances in Computational Mathematics*, vol. 13, no. 1, pp. 1–50, 2000.
- [70] B. Settles, “Active learning literature survey,” tech. rep., University of Wisconsin Madison.
- [71] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, “Active learning with statistical models,” *Journal of Artificial Intelligence Research*, vol. 4, pp. 129–145, 1996.
- [72] R. Liere and P. Tadepalli, “Active learning with committees for text categorization,” in *Proceedings of the National Conference on Artificial Intelligence*, pp. 591–597, 1997.
- [73] A. McCallum and K. Nigam, “Employing em and pool-based active learning for text classification,” in *Proceedings of the 5th International Conference on Machine Learning*, pp. 350–358, 1998.
- [74] G. Schohn and D. Cohn, “Less is more: Active learning with support vector machines,” in *Proceedings of the 7th International Conference on Machine Learning*, pp. 839–846, 2000.

- [75] Y. Rui and T. S. Huang, “Optimizing learning in image retrieval,” in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 236 –243, 2000.
- [76] Y. Rui, T. S. Huang, and S. Mehrotra, “Content-based image retrieval with relevance feedback in mars,” in *Proceedings of the IEEE International Conference on Image Processing*, pp. 815 –818, 1997.
- [77] Y. Ishikawa, R. Subramanya, and C. Faloutsos, “Mindreader: Querying databases through multiple examples,” in *Proceedings of the 24rd International Conference on Very Large Data Bases*, pp. 218–227, 1998.
- [78] M. L. Kherfi and D. Ziou, “Relevance feedback for cbir: a new approach based on probabilistic feature weighting with positive and negative examples,” *IEEE Transactions on Image Processing*, vol. 15, no. 4, pp. 1017–1030, 2006.
- [79] H. Yu and J. Yang, “A direct lda algorithm for high-dimensional data – with application to face recognition,” *Pattern Recognition*, vol. 34, no. 10, pp. 2067 – 2070, 2001.
- [80] K. Tieu and P. Viola, “Boosting image retrieval,” *International Journal of Computer Vision*, vol. 56, pp. 17–36, Jan. 2004.
- [81] X. He, D. Cai, and J. Han, “Learning a maximum margin subspace for image retrieval,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 2, pp. 189 –201, 2008.
- [82] J. Yu and Q. Adviser-Tian, *Generalized discriminant analysis in content-based image retrieval*. The University of Texas at San Antonio, 2007.
- [83] J. Yu and Q. Tian, “Semantic subspace projection and its applications in image retrieval,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 4, pp. 544 –548, 2008.
- [84] K. Lu and X. He, “Image retrieval based on incremental subspace learning,” *Pattern Recognition*, vol. 38, no. 11, pp. 2047 – 2054, 2005.

- [85] D. Cai, X. He, and J. Han, “Regularized regression on image manifold for retrieval,” in *Proceedings of the International Workshop on Multimedia Information Retrieval*, pp. 11–20, 2007.
- [86] D. Tao, X. Tang, X. Li, and Y. Rui, “Direct kernel biased discriminant analysis: a new content-based image retrieval relevance feedback algorithm,” *IEEE Transactions on Multimedia*, vol. 8, no. 4, pp. 716–727, 2006.
- [87] D. Xu, S. Yan, D. Tao, S. Lin, and H. Zhang, “Marginal fisher analysis and its variants for human gait recognition and content-based image retrieval,” *IEEE Transactions on Image Processing*, vol. 16, no. 11, pp. 2811–2821, 2007.
- [88] W. Bian and D. Tao, “Biased discriminant euclidean embedding for content-based image retrieval,” *IEEE Transactions on Image Processing*, vol. 19, no. 2, pp. 545–554, 2010.
- [89] D. Tao and X. Tang, “Kernel full-space biased discriminant analysis,” in *Proceedings of the IEEE International Conference on Multimedia and Expo*, pp. 1287–1290, 2004.
- [90] J. Li, N. Allinson, D. Tao, and X. Li, “Multitraining support vector machine for image retrieval,” *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3597–3601, 2006.
- [91] Y. Chen, X. S. Zhou, and T. Huang, “One-class svm for learning in image retrieval,” in *Proceedings of IEEE International Conference on Image Processing*, pp. 34–37, 2001.
- [92] C. H. Hoi, C. Chan, K. Huang, M. R. Lyu, and I. King, “Biased support vector machine for relevance feedback in image retrieval,” in *Proceedings of the IEEE International Joint Conference on Neural Networks, 2004.*, vol. 4, pp. 3189–3194.
- [93] D. Tao, X. Li, and S. J. Maybank, “Negative samples analysis in relevance feedback,” *IEEE Transactions on Knowledge and Data Engineering*, pp. 568–580, 2007.

- [94] J. Laaksonen, M. Koskela, S. Laakso, and E. Oja, "Picsom - content-based image retrieval with self-organizing maps," *Pattern Recognition Letters*, vol. 21, no. 13-14, pp. 1199 – 1207, 2000.
- [95] Y. Wu, Q. Tian, and T. Huang, "Discriminant-em algorithm with application to image retrieval," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 222–227, 2000.
- [96] F. Qian, M. Li, L. Zhang, H. Zhang, and B. Zhang, "Gaussian mixture model for relevance feedback in image retrieval," in *Proceedings of the IEEE International Conference on Multimedia and Expo, 2002*, vol. 1, pp. 229–232.
- [97] C. H. Hoi and M. R. Lyu, "A semi-supervised active learning framework for image retrieval," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 302–309, 2005.
- [98] C. Dagli, S. Rajaram, and T. S. Huang, "Leveraging active learning for relevance feedback using an information theoretic diversity measure," *Image and Video Retrieval*, pp. 123–132, 2006.
- [99] K. Goh, E. Y. Chang, and W. Lai, "Multimodal concept-dependent active learning for image retrieval," in *Proceedings of the 12th ACM International Conference on Multimedia*, pp. 564–571, 2004.
- [100] H. Muller, T. Pun, and D. Squire, "Learning from user behavior in image retrieval: Application of market basket analysis," *International Journal of Computer Vision*, vol. 56, pp. 65–77, 2004.
- [101] C. Liu, "Gabor-based kernel pca with fractional power polynomial models for face recognition," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 26, pp. 572–581, May 2004.
- [102] X. Xie and K. Lam, "Gabor-based kernel pca with doubly nonlinear mapping for face recognition with a single face image," *IEEE Transaction on Image Processing*, vol. 15, pp. 2481–2492, Sept. 2006.

- [103] J. Yang, D. Zhang, J. Yang, and B. Niu, “Globally maximizing, locally minimizing: Unsupervised discriminant projection with applications to face and palm biometrics,” *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 650–664, Apr. 2007.
- [104] L. Xu and I. King, “A pca approach for fast retrieval of structural patterns in attributed graphs,” *IEEE Transaction on System Man and Cyberetics-Part B: Cybernetics*, vol. 31, pp. 812–817, Oct. 2001.
- [105] B. Celia and I. F. Rajam, “An efficient content based image retrieval framework using machine learning techniques,” in *Proceedings of the 2nd International Conference on Data Engineering and Management*, pp. 162–169, 2012.
- [106] F. R. K. Chung, *Spectral graph theory*. Amer Mathematical Society, 1997.
- [107] C. R. Rao, “The utilization of multiple measurements in problems of biological classification,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 10, no. 2, pp. 159–203, 1948.
- [108] D. Tao, X. Li, X. Wu, and S. Maybank, “Geometric mean for subspace selection,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 2, pp. 260–274, 2009.
- [109] R. Duda, P. Hart, and D. Stork, *Pattern classification*, vol. 2. Citeseer, 2001.
- [110] L. N. Teow and K. F. Loe, “Robust vision-based features and classification schemes for off-line handwritten digit recognition,” *Pattern Recognition*, vol. 35, no. 11, pp. 2355–2364, 2002.
- [111] A. Tefas, C. Kotropoulos, and I. Pitas, “Using support vector machines to enhance the performance of elastic graph matching for frontal face authentication,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 7, pp. 735–746, 2001.
- [112] E. Hjelmås and B. Low, “Face detection: A survey,” *Computer Vision and Image Understanding*, vol. 83, no. 3, pp. 236–274, 2001.

- [113] S. Romdhani, P. Torr, B. Scholkopf, and A. Blake, “Computationally efficient face detection,” in *Proceedings of the 8th IEEE International Conference on Computer Vision*, vol. 2, pp. 695–700, 2001.
- [114] B. Schölkopf, C. Burges, and A. Smola, *Advances in kernel methods: support vector learning*. MIT press, 1998.
- [115] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl, “Constrained k-means clustering with background knowledge,” in *Proceedings of the 8th International Conference on Machine Learning*, pp. 577–584, 2001.
- [116] C. H. Hoi, W. Liu, M. R. Lyu, and W. Ma, “Learning Distance Metrics with Contextual Constraints for Image Retrieval,” in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 2072–2078, 2006.
- [117] J. V. Davis and I. S. Dhillon, “Structured metric learning for high dimensional problems,” in *Proceeding of the 14th ACM International Conference on Knowledge Discovery and Data mining*, pp. 195–203, ACM, 2008.
- [118] C. Hoi, R. Jin, J. Zhu, and M. Lyu, “Semi-supervised SVM batch mode active learning for image retrieval,” in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–7, IEEE, 2008.
- [119] L. Wu, R. Jin, C. H. Hoi, J. Zhu, and N. Yu, “Learning brezman distance functions and its application for semi-supervised clustering,” in *Advances in Neural Information Processing Systems*, 2009.
- [120] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell, “Text classification from labeled and unlabeled documents using em,” *Machine learning*, vol. 39, no. 2, pp. 103–134, 2000.
- [121] A. Grigoryan, “Heat kernels on weighted manifolds and applications,” *Cont. Math*, vol. 398, pp. 93–191, 2006.
- [122] D. L. Donoho and C. Grimes, “Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data,” in *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 10, pp. 5591–5596, 2003.

- [123] M. Belkin and P. Niyogi, “Laplacian eigenmaps for dimensionality reduction and data representation,” *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, 2003.
- [124] X. Zhu, “Semi-supervised learning literature survey,” Tech. Rep. 1530, Computer Sciences, University of Wisconsin-Madison, 2005.
- [125] R. T. T. Hastie and J. Friedman, *The Elements of Statistical Learning*. Springer, 2009.
- [126] K. Fukunaga, *Introduction to statistical pattern recognition*. Academic Press Professional, Inc. San Diego, CA, USA, 1990.
- [127] C. Liu and H. Wechsler, “Enhanced fisher linear discriminant models for face recognition,” in *Proceedings of the 14th International Conference on Pattern Recognition*, pp. 1368 –1372, 1998.
- [128] L. Chen, H. M. Liao, M. Ko, J. Lin, and G. Yu, “A new lda-based face recognition system which can solve the small sample size problem,” *Pattern Recognition*, vol. 33, no. 10, pp. 1713 – 1726, 2000.
- [129] X. Wang and X. Tang, “Unified subspace analysis for face recognition,” in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 679 –686, 2003.
- [130] J. Yu, Q. Tian, T. Rui, and T. Huang, “Integrating discriminant and descriptive information for dimension reduction and classification,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 3, pp. 372 –377, 2007.
- [131] H. Gao and J. W. Davis, “Why direct lda is not equivalent to lda,” *Pattern Recognition*, vol. 39, no. 5, pp. 1002 – 1006, 2006.
- [132] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge Univ. Press, 1990.
- [133] K. Liu, Y. Cheng, J. Yang, and L. Xiao, “An efficient algorithm for foley–sammon optimal set of discriminant vectors by algebraic method,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 6, no. 05, pp. 817–829, 1992.

- [134] J. B. Tenenbaum, V. Silva, and J. C. Langford, “A global geometric framework for nonlinear dimensionality reduction,” *Science*, vol. 290, no. 5500, p. 2319, 2000.
- [135] H. S. Seung and D. D. Lee, “The manifold ways of perception,” *Science*, vol. 290, no. 5500, pp. 2268–9, 2000.
- [136] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin, “Graph embedding and extensions: A general framework for dimensionality reduction,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40–51, 2007.
- [137] F. Song, D. Zhang, D. Mei, and Z. Guo, “A multiple maximum scatter difference discriminant criterion for facial feature extraction,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 37, no. 6, pp. 1599–1606, 2007.
- [138] H. Li, T. Jiang, and K. Zhang, “Efficient and robust feature extraction by maximum margin criterion,” *IEEE Transactions on Neural Networks*, vol. 17, no. 1, pp. 157–165, 2006.
- [139] Q. Liu, X. Tang, H. Lu, and S. Ma, “Face recognition using kernel scatter-difference-based discriminant analysis,” *IEEE Transactions on Neural Networks*, vol. 17, no. 4, pp. 1081–1085, 2006.
- [140] X. Li, S. Lin, S. Yan, and D. Xu, “Discriminant locally linear embedding with high-order tensor data,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 38, no. 2, pp. 342–352, 2008.
- [141] E. Kokiopoulou and Y. Saad, “Orthogonal neighborhood preserving projections: A projection-based dimensionality reduction technique,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2143–2156, 2007.
- [142] T. Zhang, K. Huang, X. Li, J. Yang, and D. Tao, “Discriminative orthogonal neighborhood-preserving projections for classification,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 40, no. 1, pp. 253–263, 2010.
- [143] M. Sugiyama, “Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis,” *Journal of Machine Learning Research*, vol. 8, pp. 1027–1061, 2007.

- [144] X. He and P. Niyogi, “Locality preserving projections,” in *Advances in Neural Information Processing Systems*, 2004.
- [145] J. Z. Wang, J. Li, and G. Wiederhold, “Simplicity: semantics-sensitive integrated matching for picture libraries,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 947–963, Sept. 2001.
- [146] J. Chen, S. Shan, C. He, G. Zhao, P. Matti, X. Chen, and W. Gao, “Wld: A robust local image descriptor,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1705–1720, 2010.
- [147] J. Yang, K. Yu, and T. Huang, “Supervised translation-invariant sparse coding,” in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 3517–3524, 2010.
- [148] J. Peng, “Multi-class relevance feedback content-based image retrieval,” *Computer Vision and Image Understanding*, vol. 90, no. 1, pp. 42–67, 2003.
- [149] D. Tao, X. Tang, and X. Li, “Which components are important for interactive image searching?,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 1, pp. 3–11, 2008.
- [150] X. He, D. Cai, S. Yan, and H. Zhang, “Neighborhood preserving embedding,” in *Proceedings of the 10th IEEE International Conference on Computer Vision*, vol. 2, pp. 1208–1213, 2005.
- [151] D. Cai, X. He, and J. Han, “Isometric projection,” in *Proceedings of the 22nd National Conference on Artificial Intelligence*, pp. 528–533, 2007.
- [152] T. Zhang, D. Tao, and J. Yang, “Discriminative locality alignment,” in *Proceedings of the 10th European Conference on Computer Vision*, pp. 725–738, 2008.
- [153] L. Zhang, L. Wang, and W. Lin, “Semisupervised biased maximum margin analysis for interactive image retrieval,” *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 2294–2308, 2012.

- [154] K. Yu, J. Bi, and V. Tresp, “Active learning via transductive experimental design,” in *Proceedings of the 23rd International Conference on Machine Learning*, vol. 23, pp. 1081–1088, 2006.
- [155] K. Yu, S. Zhu, W. Xu, and Y. Gong, “Non-greedy active learning for text categorization using convex ansductive experimental design,” in *Proceedings of the 31st International Conference on Research and Development in Information Retrieval*, pp. 635–642, 2008.
- [156] K. Bennett and A. Demiriz, “Semi-supervised support vector machines,” in *Advances in Neural Information Processing Systems*, pp. 368–374, 1999.
- [157] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schölkopf, “Learning with local and global consistency,” in *Advances in Neural Information Processing Systems*, vol. 16, pp. 321–328, 2004.
- [158] X. He, “Laplacian regularized d-optimal design for active learning and its application to image retrieval,” *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 254–263, 2010.
- [159] L. Zhang, C. Chen, W. Chen, J. Bu, D. Cai, and X. He, “Convex experimental design using manifold structure for image retrieval,” in *Proceedings of the 17th ACM International Conference on Multimedia*, pp. 45–54, 2009.
- [160] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, “Relevance feedback: a power tool for interactive content-based image retrieval,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, pp. 644–655, Sept. 1998.
- [161] L. Wang, K. L. Chan, and P. Xue, “A criterion for optimizing kernel parameters in kbda for image retrieval,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 35, no. 3, pp. 556–562, 2005.
- [162] C. Hoi, R. Jin, and M. Lyu, “Batch mode active learning with applications to text categorization and image retrieval,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1233–1248, 2009.

- [163] X. Tian, D. Tao, X. Hua, and X. Wu, “Active reranking for web image search,” *IEEE Transactions on Image Processing*, vol. 19, pp. 805–820, Mar. 2010.
- [164] F. Li, R. Fergus, and P. Perona, “Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories,” in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition Workshop*, pp. 178–178, 2004.
- [165] K. Levi, M. Fink, and Y. Weiss, “Learning from a small number of training examples by exploiting object categories,” in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition Workshop*, pp. 96–96, 2004.
- [166] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.
- [167] B. Scholkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.
- [168] C. H. Hoi, W. Liu, and S. F. Chang, “Semi-supervised distance metric learning for collaborative image retrieval and clustering,” *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 6, no. 3, pp. 1–26, 2010.
- [169] L. Zhang, L. Wang, and W. Lin, “Conjunctive patches subspace learning with side information for collaborative image retrieval,” *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3707–3720, 2012.
- [170] C. H. Hoi and M. R. Lyu, “A novel log-based relevance feedback technique in content-based image retrieval,” in *Proceedings of the 12th ACM International Conference on Multimedia*, pp. 24–31, 2004.
- [171] L. Si, R. Jin, C. H. Hoi, and M. Lyu, “Collaborative image retrieval via regularized metric learning,” *Multimedia Systems*, vol. 12, pp. 34–44, 2006.

- [172] H. Hotelling, “Analysis of a complex of statistical variables into principal components,” *Journal of Educational Psychology*, vol. 24, no. 6, p. 417, 1933.
- [173] X. He, M. Ji, and H. Bao, “Graph embedding with constraints,” in *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, pp. 1065–1070, Morgan Kaufmann Publishers Inc., 2009.
- [174] S. Si, D. Tao, and K. Chan, “Evolutionary cross-domain discriminative hessian eigenmaps,” *IEEE Transactions on Image Processing*, vol. 19, no. 4, pp. 1075–1086, 2010.
- [175] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall, “Learning a mahalanobis metric from equivalence constraints,” *Journal of Machine Learning Research*, vol. 6, no. 1, pp. 937–965, 2006.
- [176] M. A. A. Cox and T. F. Cox, “Multidimensional scaling,” *Handbook of Data Visualization*, pp. 315–347, 2008.
- [177] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, “Neighbourhood components analysis,” in *Advances in Neural Information Processing Systems*, 2004.
- [178] K. Q. Weinberger and L. K. Saul, “Distance metric learning for large margin nearest neighbor classification,” *Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.
- [179] A. Globerson and S. Roweis, “Metric learning by collapsing classes,” in *Advances in Neural Information Processing Systems*, 2006.
- [180] A. Ghodsi, D. Wilkinson, and F. Southey, “Improving embeddings by flexible exploitation of side information,” in *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pp. 810–816, 2007.
- [181] F. Girosi, M. Jones, and T. Poggio, “Regularization theory and neural networks architectures,” *Neural Computation*, vol. 7, no. 2, pp. 219–269, 1995.
- [182] L. Wang and X. Fu, *Data Mining with Computational Intelligence*. Springer Berlin, 2005.

- [183] T. Zhang, D. Tao, X. Li, and J. Yang, "Patch alignment for dimensionality reduction," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1299–1313, 2009.
- [184] Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimension reduction via local tangent space alignment," *SIAM Journal of Scientific Computing*, vol. 26, pp. 313–338, 2002.
- [185] H. Ye and G. Xu, "Fast search in large-scale image database using vector quantization," in *Proceedings of the 2nd International Conference on Image and Video Retrieval*, pp. 477–487, 2003.
- [186] M. Worring, G. P. Nguyen, L. Hollink, J. C. Van Gemert, and D. C. Koelma, "Accessing video archives using interactive search," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, vol. 1, pp. 297–300, 2004.
- [187] H. D. Tagare, C. C. Jaffe, and J. Duncan, "Medical image databases a content-based retrieval approach," *Journal of the American Medical Informatics Association*, vol. 4, no. 3, pp. 184–198, 1997.
- [188] H. J. Lowe, I. Antipov, W. Hersh, and C. A. Smith, "Towards knowledge-based retrieval of medical images. the role of semantic indexing, image content representation and knowledge-based retrieval," in *Proceedings of the American Medical Informatics Association Symposium*, 1998.
- [189] W. D. Bidgood Jr, B. Bray, N. Brown, A. R. Mori, K. A. Spackman, A. Golichowski, R. H. Jones, L. Korman, B. Dove, L. Hildebrand, *et al.*, "Image acquisition context," *Journal of the American Medical Informatics Association*, vol. 6, no. 1, pp. 61–75, 1999.
- [190] A. Joly, O. Buisson, and C. Frelicot, "Content-based copy retrieval using distortion-based probabilistic similarity search," *IEEE Transactions on Multimedia*, vol. 9, no. 2, pp. 293–306, 2007.

- [191] A. Joly, C. Frélicot, and O. Buisson, “Feature statistical retrieval applied to content based copy identification,” in *Proceedings of the IEEE International Conference on Image Processing*, vol. 1, pp. 681–684, 2004.
- [192] C. Kim, “Content-based image copy detection,” *Signal Processing: Image Communication*, vol. 18, no. 3, pp. 169–184, 2003.
- [193] J. Hong, H. Chen, and J. Hsiang, “A digital museum of taiwanese butterflies,” in *Proceedings of the 5th ACM Conference on Digital libraries*, pp. 260–261, 2000.
- [194] B. Zhu, M. Ramsey, and H. Chen, “Creating a large-scale content-based airphoto image digital library,” *IEEE Transactions on Image Processing*, vol. 9, pp. 163–167, Jan. 2000.
- [195] J. C. French, A. C. Chapin, and W. N. Martin, “An application of multiple viewpoints to content-based image retrieval,” in *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 128–130, 2003.
- [196] M. Pearce, A. Goel, I. Kolodner, C. Zimring, L. Sentosa, and R. Billington, “Case-based design support: A case study in architectural design,” *IEEE Expert*, vol. 7, no. 5, pp. 14–20, 1992.
- [197] G. Yu, J. Yuan, and Z. Liu, “Unsupervised random forest indexing for fast action search,” in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 865–872, 2011.
- [198] Y. Liu, D. Xu, I. Tsang, and J. Luo, “Textual query of personal photos facilitated by large-scale web data,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 1022–1036, 2011.
- [199] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, “Adapting visual category models to new domains,” in *Proceedings of the 11th European Conference on Computer Vision*, pp. 213–226, 2010.
- [200] L. Duan, D. Xu, I. Tsang, and J. Luo, “Visual event recognition in videos by learning from web data,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1667–1680, 2012.