SEMANTICAL REPRESENTATION AND RETRIEVAL OF NATURAL PHOTOGRAPHS AND MEDICAL IMAGES USING CONCEPT AND CONTEXT-BASED FEATURE SPACES

Md Mahmudur Rahman

A thesis
in
The Department
of
Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements
For the Degree of Doctor of Philosophy
Concordia University
Montréal, Québec, Canada

March 2008

Your file Votre référence
ISBN: 978-0-494-37748-2
Our file Notre référence
ISBN: 978-0-494-37748-2

Abstract

Semantical Representation and Retrieval of Natural Photographs and Medical Images using Concept and Context-Based Feature Spaces

Md Mahmudur Rahman, Ph.D.
Concordia University, 2008

The growth of image content production and distribution over the world has exploded in recent years. This creates a compelling need for developing innovative tools for managing and retrieving images for many applications, such as digital libraries, web image search engines, medical decision support systems, and so on. Until now, content-based image retrieval (CBIR) addresses the problem of finding images by automatically extracting low-level visual features, such as color, texture, shape, etc. with limited success. The main limitation is due to the large semantic gap that currently exists between the high-level semantic concepts that users naturally associate with images and the low-level visual features that the system is relying upon. Research for the retrieval of images by semantic contents is still in its infancy. A successful solution to bridge or at least narrow the semantic gap requires the investigation of techniques from multiple fields. In addition, specialized retrieval solutions need to emerge, each of which should focus on certain types of image domains, user's search requirements and applications objectivity.

This work is motivated by a multi-disciplinary research effort and focuses on semantic-based image search from a domain perspective with an emphasis on natural photography and biomedical image databases. More precisely, we propose novel image representation and retrieval methods by transforming low-level feature spaces into concept-based feature spaces using statistical learning techniques. To this end, we perform supervised classification for modeling of semantic concepts and unsupervised clustering for constructing codebook of visual concepts to represent images in higher levels of abstraction for effective retrieval. Generalizing upon vector space model of Information Retrieval, we also investigate automatic query expansion techniques from a new perspective to reduce concept mismatch problem by analyzing their correlations information at both local and global levels in a collection. In addition, to perform retrieval in a complete semantic level, we propose an adaptive fusion-based retrieval technique in content and context-based feature spaces based on relevance feedback information from users. We developed a prototype image retrieval system as a part of the CINDI (Concordia INdexing and DIscovery system) digital library project, to perform exhaustive experimental evaluations and show the effectiveness of our retrieval approaches in both narrow and broad domains of application.

Acknowledgments

I would like to take this opportunity to thank all the people who have helped me through the completion of this thesis. First, I would like to express my deepest gratitude to my supervisors Professors Bipin C. Desai and Prabir Bhattacharya. I have been gaining a lot of benefit from them, not only because of the academic advice they have provided, but also because of the encouragement they have given. Whenever there was a problem, they always guided me towards the right direction. Without their support the realization of this research would not have been possible. Moreover, I would like to sincerely thank my examining committee members, Professors Clement Lam, Bill Lynch, and Volker Haarslev for their comments and useful suggestions on this thesis.

I would also like to thank all the talented and hard-working individuals in the CINDI Group for creating a nice work environment and supporting me throughout the process. I would especially thank Varun Sood, who has offered me much support for the last two years to participate in the ImageCLEF cross-language image retrieval track for evaluating some of our retrieval techniques.

My deepest gratitude goes to my wife, Shimi Naurin Ahmad, for the never-ending love, daily patience, incredible support and encouragement, she has given me through out this research. Without her, this research would not have been possible. Last but not least, I thank my family back in my home country for providing me with the education, morals, and confidence that have helped me get through everyday life.

Contents

List of Figures x

List of Tables xiv

1 Introduction and Motivations 1

1.1 Contributions of the thesis 5

1.2 Organization of the thesis 8

1.3 List of publications 10

2 Literature Review 13

2.1 Content-Based Image Retrieval 13

2.1.1 Challenges in CBIR 14

2.2 Building Blocks of a Typical CBIR System 17

2.2.1 Feature Extraction and Representation 19

2.2.2 Similarity Matching 22

2.2.3 Feature Indexing 23

2.2.4 Query Processing 24

2.3 Towards Semantic-Based Image Retrieval 25

2.3.1 Semantic Classification and Annotation 26

2.3.2 Interactive Image Retrieval 27

2.3.3 Multimodal Image Retrieval 29

2.4 Image Retrieval in Medical Domain 30

2.4.1 Retrieval Techniques 31

2.4.2 Application Areas and Projects 33

3 A Global Concept-Based Image Retrieval Framework 36

3.1 Multi-class SVMs 39

3.2 Low-Level Image Feature Representation 42

3.2.1 Color Layout Descriptor (CLD) 43

3.2.2 Edge Histogram Descriptor (EHD) 43

3.2.3 Moment-Based Feature 44

3.2.4 Average Grey Level Feature 45

3.3 Image Representation in Global Concept Space 46

3.3.1 Concept Feature Fusion 47

3.4 Adaptive Statistical Similarity-Based Retrieval 51

3.4.1 Feature Dimension Reduction 53

3.4.2 Parameter Estimation and Similarity Matching 55

3.4.3 Query Parameters Updating by Relevance Feedback 56

3.5 Summary 59

4 Image Representation in Local Visual and Semantic Concept Spaces 60

4.1 Local Visual Concept-Based Image Representation 64

4.1.1 Codebook Generation by SOM 66

4.1.2 Image Encoding and Feature Representation 68

4.1.3 Fuzzy Visual Concept-Based Feature Representation 70

4.1.4 Fuzzy Visual Local Concept Vector 73

4.1.5 Probabilistic Local Semantic Concept-Based Image Representation 76

4.2 Visual and Semantic Structure Descriptors 81

4.3 Summary 84

5 Query Expansion Based on Local and Global Analysis 85

5.1 Query Expansion Based on Local Analysis 89

5.2 Query Expansion Based on Global Analysis 94

5.3 Similarity Matching and Query Expansion in Inverted Index 98

5.4 Summary 102

6 Fusion-Based Retrieval in Content and Context Feature Spaces 103

6.1 Context-based Image Search Approach 106

6.2 Content-Based Image Search Approach 108

6.3 Contextual Query Refinement 113

6.4 Visual Query Refinement 116

6.5 Adaptive Linear Combination of Multiple Evidences 118

6.6 Cross-modal Interaction and Integration 121

6.6.1 Sequential Search with Pre-filtering and Re-ranking 121

6.6.2 Simultaneous Search 123

6.7 Summary 125

7 CBIR as Decision Support System for Dermoscopic Images 126

7.1 Segmentation and Lesion Detection 128

7.1.1 Intensity Image Generation from HVC Color Space 129

7.1.2 Intensity Image Generation from Fuzzy C-Means Clustering . 131

7.1.3 Iterative Thresholding & Post-processing 133

7.2 Lesion-Specific Feature Extraction 135

7.3 Similarity Matching 136

7.4 Summary 138

8 Retrieval Evaluation 139

8.1 Image Collections 139

8.1.1 The General Photographic Collection 139

8.1.2 The Medical Collection 140

8.1.3 The Radiograph Image Collection 142

8.1.4 The Dermoscopic Image Collection 145

8.2 Performance Measure 145

8.3 Performance Evaluation of Global Concept-Based Retrieval 146

8.3.1 Training of SVMs 147

8.4 Performance Evaluation of Local Concept-Based Retrieval 155

8.4.1 Training for Local Visual and Semantic Concepts 156

8.4.2 Retrieval Accuracies (Local Concept) 158

8.5 Performance Evaluation of Query Expansion Techniques 161

8.6 Performance Evaluation of Fusion-Based Retrieval 166

8.6.1 Query Set of General Photographic Images 167

8.6.2 Query set of Medical Images 168

8.6.3 Result Analysis of the Submitted Runs in ImageCLEF'06 . . . 169

8.6.4 Result Analysis of the Submitted Runs in ImageCLEF'07 . . . 171

8.6.5 Extended Result Analysis for Mixed-mode Query Sets 173

8.7 Performance Evaluation for Dermoscopic Image Retrieval 176

8.8 Summary 178

9 Conclusion 179

9.1 Summary of Contributions 179

9.2 Limitations 181

9.3 Future Research Directions 183

Bibliography 186

List of Figures

1 Images returned by AltaVista Image Search 2

2 Ten-level indexing pyramid [27] 17

3 Example images for each level of the visual structure [27]. 18

4 Process flow diagram of a typical CBIR system 19

5 Feature extraction and representation 20

6 Basic Block diagram of an Interactive RF-based Retrieval System . . . 28

7 Block diagram of a CBIR system as a diagnostic aid 32

8 Maximal margin classifier (separable case) 40

9 Maximal margin classifier (non-separable case). 41

10 Filters used for edge detection (a) vertical, (b) horizontal, (c) diagonal 45°, (d) diagonal 135°, (e) non-directional 44

11 Region generation from sub-images 44

12 Feature extraction from a low-resolution image 46

13 Process diagram of concept feature fusion with classifier combination. 48

14 Block diagram of the image representation and retrieval in global concept spaces. 50

15 Block diagram of the statistical similarity-based retrieval approach . . 58

16 Block diagram of the concept-based image representation framework. 64

17 Structure of the SOM 67

18 Codebook generation and Image encoding process 69

19 Topological local neighborhoods 73

20 Vector Generation based on Local Neighborhood 75

21 Intra class perceptual variability of concepts sky and water 76

22 Representation of Images with Concept Occurrences [96]. 79

23 Image encoding with probabilistic membership scores 80

24 Images with different color structures 82

25 Example of Visual Concept Structure Descriptor (VCSD) generation . 82

26 Distance between concepts c_i and c_j in an encoded image based on neighborhood relationship 91

27 Concept c_j as a neighbor of the concept c_i based on a local cluster . . 92

28 Relationships between concepts and virtual query concept in an Euclidean space 98

29 Example of Quadratic similarity matching in Color Histogrms [12] . . 99

30 Example process of Query Expansion in an Inverted File 102

31 Constructed example of querying an image database. (a) using text only, (b) using images only, (c) combining images and text. 104

32 Example chest CT images with micro nodules 104

33 An example medical image and its associated annotation file in ImageCLEFmed collection [58]. 107

34 An example photographic image and its associated annotation file in ImageCLEFphoto (e.g., IAPR) collection [57]. 107

35 Region generation based on segmentation of a photographic image . . 109

36 Region generation based on segmentation of a medical image 110

37 Process flow diagram of the sequential search approach 123

38 Process flow diagram of the simultaneous search 124

39 (a) Original color image (b) Grey level image of the original 130

40 (a) Grey level histogram (b) Iterative thresholded image 130

41 (a) Intensity image generated from HVC space (b) Histogram of the intensity image 131

42 (a) Intensity image generated from FCM (b) Histogram of the intensity image 132

43 (a) Thresholded image from the intensity image of Figure 41(a), (b) Thresholded image from intensity image of Figure 42(a), (c) Final segmented image with lesion mask. 133

44 Block diagram of the CBIR system. 137

45 Example images of the Photographic collection [54] 140

46 Example images of the Medical collection. 141

47 Intra-class variability within the category label 3, (annotated as "X-ray, plain radiography, coronal, upper extremity (arm), hand, musculosceletal system") . . . 143

48 Inter-class similarity between category labels 52 and 5 (Radiograph Collection)) . . . 143

49 Frequency of images in 116 categories (Radiograph Image Collection) 144

50 Sample dermoscopic images of all three categories . . . 144

51 Classification structure of Med-DB. . . . 146

52 Accuracy comparison of global concept and low-level feature spaces . 150

53 A snapshot of the retrieval result based on $\mathbf{f}^{\mathrm{EHD}}$ . . . 150

54 A snapshot of the retrieval result based on $\mathbf{p}^{\mathrm{EHD}}$ . . . 151

55 Accuracy comparison of concept-based feature fusion approaches . . . 151

56 Accuracy comparison of statistical, Euclidean and cosine similarity matching . . . 153

57 A snapshot of the retrieval result based on Euclidean similarity matching154

58 A snapshot of the retrieval result based on Statistical similarity matching154

59 Accuracy comparison with and without RF . . . 155

60 Classification structure of Photo-DB. . . . 156

61 Accuracy comparison of different codebook sizes and image partitions. 158

62 Accuracy comparison of local visual concept feature spaces. . . . 159

63 Accuracy comparison based on global and local membership values . 160

64 Accuracy comparison of local semantic concept-based features. . . . 161

65 Accuracy comparison of different query expansion schemes in local visual concept space. . . . 162

66 Accuracy comparison of different query expansion schemes in local semantic concept space. . . . 162

67 Accuracy comparison with different number of iterations . . . 164

68 Average Precision curves for the Quadratic similarity matching with and without Inverted Index . . . 165

69 Topics for the Photographic (IAPR) collection in ImageCLEF [57] . . 167

70 Topics for the Medical Collection in ImageCLEFmed [58] . . . 168

71 Example of a sematic topic in ImageCLEFmed'07 [58] . . . 173

72 Mixed-mode topics for Photographic collection [57] . . . 174

73 Mixed-mode topics for Medical collection[58] 174

74 (a) Accuracy comparison of three similarity measures (b) Accuracy comparison of three image categories. 177

75 Retrieval result based on Euclidean similarity. 177

76 Retrieval result based on Fusion similarity. 178

List of Tables

1 Correspondense between human color perception and NBS distance [161] 131
2 Individual Databases used in Medical collection 141
3 CV Error Rate (MED-DB training set) 147
4 CV Error Rate (IRMA-DB training set) 147
5 Error Rate of the Individual Classifiers (test sets) 149
6 Error Rate of Different Classifier Combinations (test sets) 149
7 CV Error Rates in PCA subspace (training sets) 153
8 Statistics of the Training Set for Local Semantic Concepts 157
9 Error rates (training set) 157
10 Improvement using expanded queries 164
11 Retrieval time with and without indexing schemes 165
12 Results of the photographic image retrieval (ImageCLEF'06) 170
13 Results of the medical image retrieval (ImageCLEF'06) 170
14 Results of the photographic image retrieval (ImageCLEF'07) 171
15 Results of the medical image retrieval (ImageCLEF'07) 172
16 Retrieval Results of the Photographic Images for Mixed-Mode Topics 175
17 Retrieval Results of the Medical Images for Mixed-Mode Topics . . . 175

List of Abbreviations

CBIR	Content-Based Image Retrieval
CLEF	Cross Language Evaluation Forum
CLD	Color Layout Descriptor
CT	Computed Tomography
CV	Cross Validation
CDA	Canadian Dermatology Association
DCT	Discrete Cosine Transform
DICOM	Digital Imaging and Communications in Medicine
EHD	Edge Histogram Descriptor
FCM	Fuzzy C-Means
FVCV	Fuzzy Visual Concept Vector
FVLCV	Fuzzy Visual Local Concept Vector
GCH	Global Color Histogram
GLA	Generalized Lloyd Algorithm
GLCM	Grey level co-occurrence matrix
HSV	Hue-saturation-value color space
HVC	Hue-value-chroma color space
IR	Information Retrieval
IRMA	Image Retrieval in Medical Applications
JPEG	Joint Photographic Experts Group
LVQ	Learning Vector Quantization
MPEG	Moving Picture Experts Group
PACS	Picture Archiving and Communication Systems
PCA	Principal Component Analysis
PSCV	Probabilistic Semantic Concept vector
PSL	Pigmented Skin lesion
PR	Precision-Recall
LN	Local Neighborhood
SCSD	Semantic Concept Structure Descriptor
SVM	Support Vector Machine
SOM	Self Organizing Map
QBE	Query By Example
QBIC	Query By Image Content
QE	Query Expansion
RBF	Redial Basis Function
RGB	red-green-blue color space
RF	Relevance Feedback
VCSD	Visual Concept Structure Descriptor
VQ	Vector Quantization
VSM	Vector Space Model
WWW	World Wide Web

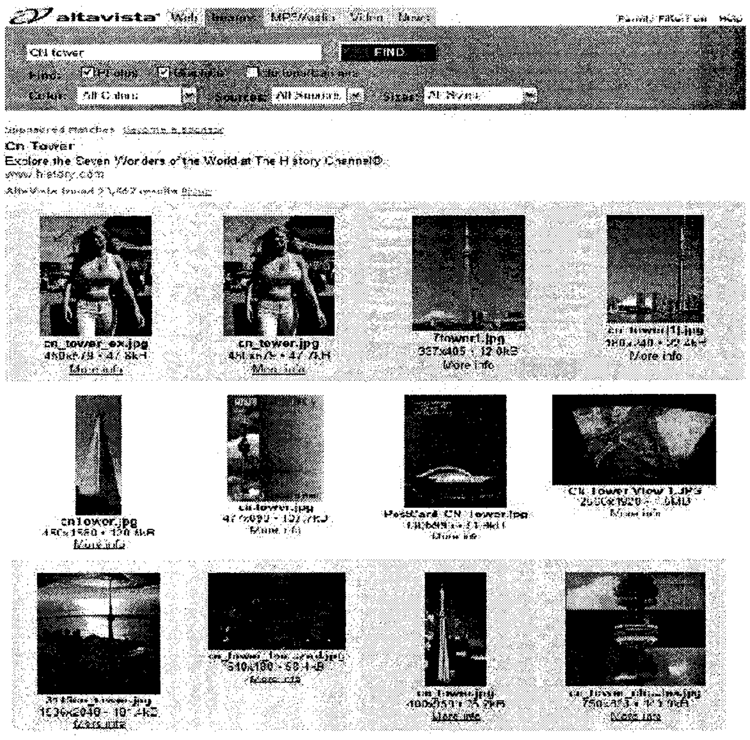
Chapter 1

Introduction and Motivations

We are living in the age of multimedia information technologies. The falling price of storage, wide availability of digital devices, and the World Wide Web (WWW) accelerate the growth of multimedia content production and distribution all over the world. Among the variety of multimedia contents, images are the most prevailing and widely used currently. Technological breakthrough makes it even possible to generate images million of miles away from the Earth by space rovers. Hospitals and medical research centers produce an increasing number of digital images of diverse modalities everyday for clinical decision making and research purposes. The needed technologies, such as digital cameras, multimedia portable phones, and personal computers are becoming available with reasonable price for general use. As result, creating and storing personal digital photos, such as holiday pictures and pictures of friends and family members, is getting easier and more affordable.

The exponential growth of images has created a compelling need for innovative tools for managing, retrieving, and visualizing them for many applications, such as digital libraries, medical decision support systems and teaching applications, Web image search engines, photo journalism, crime prevention, fashion design, trademark registration, and so on. These applications need has led researchers to the consensus that indexing and management is necessary for image data to be valuable in the long term. The existing and widely adopted text-based search methods have proven out to be inadequate for image retrieval purposes. For example, Figure 1 shows the search results of a popular Web image search engines; the AltaVista Image Search [1].

[1] http://www.altavista.com/image/default



CN tower

Cn Tower
Explore the Seven Wonders of the World at The History Channel®
www.history.com

Figure 1: Images returned by AltaVista Image Search

The search engine returned most similar twelve images when we performed a keyword search with *"CN Tower"*; one of the famous landmark in Canada. In this case, images are indexed and returned based on their associated or collateral text in web pages. From the results, we observe that the search returned unwanted images even at the very top positions, i.e., 1st and 2nd ranked images (consider left to right and top to bottom) in Figure 1.

The above mis-hits point to a need for an effective way to search images based on their visual contents commonly known as content-based image retrieval (CBIR) [1, 2, 3, 5, 6, 7, 9, 10, 11]. In CBIR, access to information is performed at a perceptual level based on automatically extracted low-level image features, such as color, texture, and shape. The relevance of image retrieval for many applications makes CBIR research as one of the fastest growing fields in information technology [1]. Unfortunately, even after more then a decade of intensive research, the CBIR systems still lag behind of the today's best text-based search engines, such as Google [2], Yahoo [3], and AltaVista [4].

One of the fundamental problems of CBIR is the *semantic gap*, which is the mismatch between user's search requirements and the capabilities of the systems to represent images [1]. Many studies show that users would like to pose semantic queries in image collections to search images of particular type of objects, activities, location, and events [25]. Though many sophisticated feature extraction methods have been designed during the last decade to represent low-level image features, they cannot adequately depict images at a semantic level. As images with the same semantic content often have variable visual appearances and many perceptually similar images might have different semantic interpretations. Hence, retrieval results based on similarities of pure visual content do not necessarily possess semantic similarities that are of interest to the user.

CBIR exhibits a varying degree of difficulty and complexity depending on the user's search requirements, objectives and extents of application domains. The complexity arises because images are richer in information content than text, and because same image can be interpreted differently in various application contexts. For example, a medical image retrieval system can assist users for different tasks [128]. In particular, a system objective can be for diagnostics (e.g., for case-based reasoning), research (e.g., to support evidence-based medicine), teaching (e.g., for the composition of case collection), and hybrid (e.g., any combination of previously mentioned one). For diagnostic purpose, a physician may search for images with different disease categories in a particular modality, such as search for computed tomography (CT) images of lung with bronchitis or emphysema in ASSERT system [137]. In this case, related images should have similar disease related properties or attributes compared to an unknown query image. Whereas, search requirements can be quite different for teaching and training purposes. Students can search large heterogeneous image repositories of various modalities for important or interesting cases and patterns based on perceptual similarity to see relevant diagnostic and potential problems.

Another important factor in the complexity of CBIR problem is the scope or extent of image domains. A narrow image domain has only a limited and predictable

[2] www.google.com

[3] www.yahoo.com

[4] www.altavista.com

variability in all aspects of appearance whereas a broad domain has unlimited and unpredictable variability of images with little or no domain-specific knowledge available [1]. For example, in a medical image collection of a particular modality, the recording circumstances for all images are almost constant, i.e., same illumination and no occlusion. Although each image has large variability with disease related properties, there are obvious geometrical, physical, and color-texture related constraints governing the domain. For such domain, the gap between image features and their semantic interpretation is usually smaller as domain-specific knowledge can be fairly exploited. The domain would be broader, on the other hand, had the images been part of a larger heterogeneous collection of different modalities [58]. In such a broad domain, semantics of images can not be described by visual only features and search requirements are also always at a higher level. For example, search for images with *"tumors"* in a heterogeneous medical collection is at a higher semantic level then search for images with *"cancer"* in a particular modality. The broader the domain is, the higher the semantic gap due to the less availability of domain knowledge [1]. Similar criteria can be observed for general photographic images. For example, in a narrow domain of personal photo collection, users might search for images of a particular category, such as indoor or outdoor images. Whereas in the broad domain Web images, a user may prefer browsing or open-ended searching instead of any particular category information in his/her mind.

Due to the immense need for effective image retrieval applications, new semantic-based trends for image retrieval using semantic image classification, automatic annotation, and interactive retrieval are being investigated during the past few years [13, 14, 15]. In the majority of cases, however, these new techniques do not focus on domain objectives and real search requirements and try to suggest solutions that are applicable for all retrieval purposes. As we observed, to develop a retrieval system to meet all criteria would be unrealistic due to the varying complexities.

We stress the fact that the *semantic gap* cannot be bridged in a general way, but rather expect that specialized CBIR solutions need to emerge, each of which should focus on certain types of image repositories, application domains, user's needs, and query paradigms. Some of these solutions need to exploit domain knowledge by utilizing learning-based techniques, whereas some may need to rely on contextual information as added semantics and propagate user's perceived semantics to the system in an interactive fashion. Hence, successful solutions to bridge the gap might require a significant paradigm shift, involving techniques originally developed in other fields, such as Machine Learning, Human Computer Interaction, and Information Retrieval (IR). This is the main focus of the work in this thesis. To overcome the limitations of low-level feature based CBIR systems, we use a multi-disciplinary approach to develop domain dependent retrieval frameworks for semantic-based image retrieval. We specially analyze the characteristics of both narrow and broad domains of general photographic and medical images. The available domain knowledge, and the types of use are considered as the prime factors to determine the appropriate image representation and retrieval approaches.

The other major motivation of this work is to create a link between techniques of Machine Learning and IR fields from our image retrieval perspective. In this direction, being inspired from ideas of these fields, we perform statistical semantic modeling to represent images by using both supervised and unsupervised off-line learning and increase retrieval effectiveness by using both automatic and interactive on-line learning techniques in the form of query expansion and relevance feedback [205, 206, 208, 217, 218]. Within this research, we provide both theoretical and practical contributions to the field of multimedia information retrieval in general and CBIR in particular. Promising results have been reported in our papers (provided at the end of this chapter) and will be summarized in this dissertation. For better clarification, we briefly summarize some of the major contributions of our work in following:

1.1 Contributions of the thesis

- We propose and develop a global concept-based image retrieval framework based on exploiting image categorization information in semantically organized databases, such as in a collection of medical images with different modalities, body parts, orientations, and so on [205, 206]. The automatic semantic classification of images is a major trend in semantic-based image retrieval [13, 15]. The majority of the systems, however, did not relate classification to retrieval directly, instead only stressed the usefulness of classification for image pre-filtering purpose [24, 99]. In this framework, we exploit the global classification information directly from a new perspective to transform various low-level image feature spaces into intermediate level semantic feature spaces based on probabilistic outputs and classifier combination of multi-class support vector machines (SVMs) [165]. Further, we present an adaptive statistical similarity matching technique in a low-dimensional feature space by exploiting the category-specific feature distribution information of a collection, on-line classifier's predictions, and feedback information from the users. By empirical analysis in Section 8.3 of Chapter 8, we showed that the proposed techniques achieved retrieval improvements as compared to commonly used low-level feature representations and geometric similarity matching functions.

- A local concept-based image representation framework is proposed for narrow to broad domain of photographic images by performing statistical modeling based on utilization of both supervised classification and unsupervised clustering techniques [209]. In this framework, codebook of visual concept prototypes (e.g., visual entities like dominant color and texture patches) are automatically constructed by utilizing self-organizing map (SOM) [32] based clustering and statistical models are built for local semantic concepts (e.g., semantic entities like water, sand, grass, sky, snow, etc. in local image regions) by using multi-class SVMs [165]. The main limitation of some of the related techniques [96, 95, 126] is that the correspondences between image regions to real objects (local concepts) are always one-to-one. Although, there might be several objects with almost as good match as the one detected for a particular image region. To overcome the limitation, images are represented in this framework in concept-based feature spaces by exploiting fuzzy and probabilistic outputs of the classifiers, topology preserving local neighborhood structure of the codebook, and concept ordering structure in individual images [209, 207]. The proposed representation schemes demonstrate their effectiveness (detailed retrieval evaluations are provided in Section 8.4 of Chapter 8) when compared to low-level features and concept frequency-based features without any feature enhancement.

- Generalizing upon the vector space model of IR, we propose automatic query expansion techniques on the local concept spaces to overcome the concept mismatch (similar to the word mismatch problem in text retrieval domain) problem [209]. Due to the nature of the low-level continuous feature representation, there appears to be less interest for such techniques in CBIR domain. Our query expansion techniques are inspired by ideas from text retrieval domain. The methods are based on local analysis that takes into account metrical constraints based on neighborhood proximity of concepts and global analysis of concept-concept similarities or correlations in a collection. The automatic query expansion approaches have a significant advantage over interactive ones as they require no effort on the part of the user. The improved effectiveness of the proposed approaches compared to original queries will be demonstrated in Section 8.4 of Chapter 8.

- The large *semantic gap* is one of the fundamental problems for CBIR in broad domains. To reduce the gap, additional contextual information, such as associated annotation or collateral text needs to be integrated with CBIR. Recently, some research projects investigated multi-modal retrieval techniques by combining text and images in a single framework towards web-based image retrieval [115, 120, 123]. We observe a lack of systematic evaluation of many of these approaches and it is unclear about how much improvements they achieved and in what context. Motivated by this, we propose interactive and adaptive fusion-based search approaches in single modal content and context-based feature spaces as well as in multimodal feature space by combining both modalities in a simultaneous or sequential way [208, 217, 218]. The proposed retrieval approach can propagate a user's perceived semantics from one modality to another based on a cross-modal multiple query expansions mechanism and dynamically update the inter and intra modality weights based on feedback information. Exhaustive experimental analysis are performed (Section 8.6 of Chapter 8) in broad domain benchmark photographic and medical image collections to show real effectiveness of the search approaches.

Another important contribution of this work is our successful participation in ImageCLEF [57, 58], an evaluation forum for comparing image search techniques, during the last three years (2005, 2006, and 2007). Such a benchmark event is very essential in image retrieval domain as there are very few widely-available image collections for comparative studies. The details of the benchmark collections and different runs submitted by us as well as analysis of the results in the past years are available in our workshop papers [216, 217, 218] and also reported in Chapter 8 of this thesis.

- Users search requirements and systems objectives can be very specific in a narrow domain. For example, in a domain of dermoscopic images, the objective of a retrieval system might be to assist dermatologist as a diagnostic aid for skin cancer or melanoma recognition. Large number of digital dermatological images are regularly being generated in clinics and hospitals. Until now, most of the works in this domain have only focused on the problem of skin cancer detection by classification-based systems and there is no CBIR system (as per our knowledge) yet to be developed in this domain. Motivated by this, we developed a CBIR system for dermoscopic images [210]. To this end, we propose a fast segmentation technique for automated lesion detection by exploiting specific image characteristics and domain knowledge and extract lesion-specific local image features for a fusion-based similarity matching function to compare an unknown query and database images. We demonstrate the effectiveness of our CBIR system based on experimentation on a collection of dermoscopic images in Section 8.7 of Chapter 8.

In summary, from our contributions, we can say that instead of focusing on a single retrieval solution for different application domains (which we assume as unrealistic at this moment), we propose several retrieval techniques where each of them is suitable only for the corresponding application domain. Exploiting ideas from multiple fields and focusing on specialized retrieval solutions are the main motivation of this research. All the proposed techniques are implemented in a prototype retrieval system and validated by experimental evaluations as will be described in Chapter 8.

1.2 Organization of the thesis

The reminder of the thesis is organized as follows:

- **Chapter 2: Literature Review**

We review the state-of-the-art in CBIR and present some ongoing trends and techniques towards semantic-based image retrieval in both general and medical domains.

- **Chapter 3: A Global Concept-based Image Retrieval Framework**
  This chapter presents a retrieval framework for image representation at global concept level and adaptive similarity matching technique based on image categorization and user's feedback information.

- **Chapter 4: Image Representation in Local Concept Spaces**
  We present several image representation schemes in local visual and semantic concept spaces by utilizing both supervised multi-class SVMs and unsupervised SOM-based clustering techniques.

- **Chapter 5: Query Expansion Based on Local and Global Analysis**
  The automatic query expansion techniques in concept-based feature spaces are presented in this chapter based on both local and global analysis of concept correlations information.

- **Chapter 6: Fusion-Based Retrieval in Content and Context Feature Spaces**
  We present the interactive and dynamic fusion-based retrieval approaches by utilizing both content and context information of images in a single framework.

- **Chapter 7: CBIR as a Decision Support System for Dermoscopic Images**
  This chapter presents the image retrieval based decision support system for dermoscopic images based on domain specific specialized image pre-processing and segmentation techniques.

- **Chapter 8: Retrieval Evaluation**
  We present the details about the data sets used for the experiments and exhaustive experimental evaluations of the proposed techniques in different experimental settings.

- **Chapter 9: Conclusion**

Finally, we provide our conclusion of the dissertation in this chapter. We summarize our major contributions as well as limitations of our retrieval approaches and future research directions.

1.3 List of publications

During the course of this work, we published several journal articles and conference papers as listed below as well as provided in the bibliography for appropriate references.

I. Rahman, M. M., Desai, B. C. and Bhattacharya, P. Medical Image Retrieval with Probabilistic Multi-Class Support Vector Machine Classifiers and Adaptive Similarity Fusion, *Computerized Medical Imaging and Graphics*, Vol. 32, pp. 95–108, 2008.

II. Rahman, M. M, Bhattacharya, P. and Desai, B. C. A Framework for Medical Image Retrieval using Machine Learning & Statistical Similarity Matching Techniques with Relevance Feedback, *IEEE Transactions On Information Technology In Biomedicine, (Special Issue on Image Management in Healthcare Enterprises)*, Vol. 11 (1), pp. 59–69, 2007.

III. Rahman, M. M., Desai, B. C. and Bhattacharya, P. Multi-Modal Interactive Approach to ImageCLEF 2007 Photographic and Medical Retrieval Tasks by CINDI, *Working Notes of the 2007 CLEF Workshop*, Sep., 2007, Budapest, Hungary, To appear in LNCS.

IV. Rahman, M. M., Desai, B. C., and Bhattacharya, P. Cross-Modal Interaction and Integration with Relevance Feedback for Medical Image Retrieval, *13th International Multimedia Modeling Conference (MMM 2007)*, Singapore, In *Proceedings of LNCS*, Vol. 4351, pp. 440–449, 2007.

V. Rahman, M. M., Sood, V., Desai, B. C. and Bhattacharya, P. CINDI at ImageCLEF 2006: Image Retrieval & Annotation Tasks for the General Photographic and Medical Image Collections, *7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, Alicante, Spain, September 20-22, 2006, Revised Selected Papers, *Proceedings in LNCS*, Vol. 4730, pp.715–724. 2007.

VI. Rahman, M. M., Desai, B. C. and Bhattacharya, P. Visual Keyword-based Image Retrieval using Correlation-Enhanced Latent Semantic Indexing, Similarity Matching & Query Expansion in Inverted Index. *Tenth International Database Engineering & Applications Symposium (IDEAS06)*, Delhi, India, In *Proceedings of IEEE Computer Society*, pp. 201–208, 2006.

VII. Rahman, M. M., Desai, B. C. and Bhattacharya, P. Image Retrieval-Based Decision Support System for Dermatoscopic Image, *Proceedings of the IEEE Symp. on Computer-Based Medical Systems*, June, 22-23, Salt Lake City, Utah, pp. 285–290, 2006.

VIII. Rahman, M. M., Desai, B. C. and Bhattacharya, P. A Feature Level Fusion in Similarity Matching to Content-Based Image Retrieval, *Proceedings of the 9th International Conference on Information Fusion*, Florence, Italy, 10-13 July, 2006.

IX. Bhattacharya, P., Rahman, M. M. and Desai, B. C. Image Representation and Retrieval Using Support Vector Machine and Fuzzy C-means Clustering Based Semantical Spaces, *Proceedings of the 18th International Conference on Pattern Recognition (ICPR 2006)*, Hong Kong, China, Vol. 2, pp. 1162–1168, 2006.

X. Rahman, M. M., Bhattacharya, P. and Desai, B. C. Probabilistic Similarity Measure in Image Databases with SVM-based Categorization & Relevance Feedback *International Conference on Image Analysis and Recognition (ICIAR'05)*, Toronto, Canada, Sept. 2005. In *Proceedings of LNCS*, Vol. 3656, pp. 601–608, 2006.

XI. Rahman, M. M., Desai, B. C. and Bhattacharya, P. Supervised Machine Learning based Medical Image Annotation and Retrieval in ImageCLEFmed 2005, *6th Workshop on Cross Language Evaluation Forum (CLEF 2005)*, Vienna, Austria, Sept. 2005. *Proceedings in LNCS*, Vol. 4022, pp. 692-701, 2006.

XII. Rahman, M. M., Bhattacharya, P. and Desai, B. C. Similarity Searching in Image Retrieval with Statistical Distance Measure and Supervised Learning, *International Conference on Advances in Pattern Recognition (ICAPR)*, Bath, UK, September 2005. In *Proceedings of LNCS*, Vol. 3686, pp. 315-324, 2005.

XIII. Rahman M. M., Wang T., and Desai B. C. Medical image retrieval and registration: towards computer assisted diagnostic approach. *IDEAS Workshop On Medical Information Systems : The Digital Hospital, IDEAS'04-DH*, Beijing, China, Sept. 2004, *Proceedings of IEEE Computer Society*, pp. 78–89, 2004.

Chapter 2

Literature Review

This chapter provides a brief overview of image retrieval techniques, their limitations, and current trends toward semantic-based image retrieval in general as well as in medical domains.

2.1 Content-Based Image Retrieval

The first generation of image retrieval systems developed in late 70's, was mainly linked to traditional database management or text retrieval systems [3, 4]. In those early systems, manually inserted annotations describing both the contents of the image and other metadata such as the file name, image format, creator, date, and so on were used as indexing terms. Unfortunately, manual assignment of textual attribute is both time consuming and costly. When the database is large and dynamic, such as images available in the World Wide Web, it is almost impossible to manually annotate all the images. In addition, keywords based annotation is inherently subjective in nature as the interpretations of images may vary depending on the context. Another problem is that some visual properties of images, such as certain textures and shapes, are difficult or nearly impossible to describe with text. What is needed in this case is the use of more concrete description of visual contents that are closely related to human perception.

To overcome the limitations of text-based image retrieval, content-based image retrieval (CBIR) systems emerged in the early 1990's [8, 16]. In CBIR, the images are automatically or semi-automatically indexed by features directly derived from their visual content by using image processing techniques. Image indexing differs substantially from keyword-based indexing of associated annotations since the desired attributes of image region or whole image are complex functions in a continuous scale. The common functionalities in CBIR can be summarized as follows [1, 2, 6, 7]

- Image processing and pattern recognition techniques are used to extract low-level features, such as color, texture, shape, etc. from images.
- For a given feature, a representation of the feature in a vector form and a notion of similarity are determined, and image is represented as a collection of features.
- Finally, image retrievals are performed based on computing similarity in feature spaces and results are ranked based on the similarity values computed.

Research in CBIR has gained widespread popularity from different communities during the last decade and has evolved and matured into a distinct research field. As a result, the past decade has witnessed the development of the first commercial CBIR system and many research prototypes from different industrial, research and academic arena. Typical commercial systems are: QBIC [16] from IBM [1], Virage Image Search Engine [17] from Virage Inc. [2], and VisualRetrievalWare [3] from Excalibur Technologies. Well known prototypes of the academic world include Photobook [18] from MIT Media Lab, VisualSeek/WebSeek [19] from Columbia University, NeTra [22] from UCSB, MARS [20] from University of Illinois at Urbana-Champaign, and SIMPLIcity [24] from Stanford University. The interest in CBIR is growing rapidly with an upsurge in publication of different techniques in the last few years.

2.1.1 Challenges in CBIR

Even after a decade of intensive research, the performances of the CBIR systems in reality lag far behind compared to today's text-based retrieval systems or search engines. The most fundamental challenge that CBIR faces is the wider extent of mismatch between user's semantic search requirements and the capabilities of the technology to fulfill the requirements. For a machine, extracting the semantic content from an image is an exceedingly difficult task due to the variability of objects in

[1] http://www.qbic.almaden.ibm.com/

[2] http://www.virage.com/

[3] http:// www.excalib.com/

visual appearances and many semantically different objects, on the other hand, are perceptually similar [9]. We can easily identify different objects in an images or the same object with different variations based on our previous experience or learning and high reasoning ability. Unfortunately, this kind of knowledge is inherently hard to duplicate in a CBIR system. This discrepancy is commonly referred to as *semantic gap* problem, which is *"the lack of coincidence between the information that one can extract from the visual media and the interpretation that the same data have for a user in a given situation"* as quoted from [1]. A user mainly looks for images of particular type of object, phenomenon or event, whereas, image descriptions in a traditional image retrieval system rely on low-level image properties (e.g., color, texture, shape and so on) and the two may be disconnected.

Depending on the search requirements and search types, the CBIR exhibits a varying degree of difficulty. The user search requirements on images can vary considerably as it is a rich and subjective source of information. Eakins in [13], identified three distinct levels of abstraction of search requirement with increasing complexity. Level 1 comprises retrieval by primitive features such as color, texture, shape or the spatial location of image elements. For example, *"find images with 50% red and a uniform texture pattern"*. Queries at level 2 may contain specific objects and scenes. At this level, some degree of object and scene recognition as well as inference about the image content is required. A query example may be as *"find group of people on a sea beach"*. At the highest level of complexity, level 3 comprises retrieval by abstract attributes, involving a significant amount of high-level reasoning about the meaning and purpose of the objects or scenes depicted. This includes retrieval of named events of pictures with emotional or religious significance, etc. For example, queries may contain abstract concept as *"find pictures of a joyful group on a sea beach"*. Here, it is difficult if not impossible to describe the concept *joyful* with low-level image features and without any high-level reasoning. Users formulate queries mostly on levels 2 and 3 and expect the systems to operate at the same levels of complexity and semantics but the current CBIR systems operate mainly at level 1 and partially in level 2.

In general, users like to present a very diverse set of different search scenarios in CBIR, which the system should support. Searches in CBIR can also be distinguished into three major categories [26]: (1) target search (2) category-specific search and (3) open-ended search or browsing. The most precise search task is target search, in which a user tries to find a specific target image which may or may not be actually present in the database and which is the only relevant image for this query. An example situation for a content-based target search takes place when a user is interested about a particular person's face out of all the images in a database of face images that matches to a query image. Category-specific search aims at retrieving an arbitrary image representative of a specific class generally from a narrow domain. In category search, the user may have available a group of images and the search is for additional images of the same class. For example, to find abdomen computed tomography (CT) images with liver blood vessels or chest CT images with micro nodule structures in medical domain, a user might prefer category-specific search based on pre-determined categorization of imaging modalities, body parts, etc. Category searches may be enhanced during the query in a natural way by relevance feedback, i.e. grading the returned images on whether they belong to the class in question and communicating this information back to the retrieval system, thereby providing more information about the class of relevant images and thus guiding the system toward the remaining relevant images in the database. In open-ended search or browsing, the user has a vague or inexact search goal in mind and he/she browses the database for any interesting things. Image searches of this type are highly interactive and often constitute a nonlinear sequence of actions, thus requiring a flexible user interface. A database visualization tool providing an overview of the database as well as relevance feedback techniques are useful to help a user in open-ended searches. This search requirements are therefore in a higher semantical level, i.e., in levels 2 and 3 in [13] and in levels 5 through 10 in [27].

The ten-level visual structure presented in [27] provides an elaborate and systematic way of abstracting images based on syntax and semantics. Syntax refers to the way visual elements are arranged without considering the meaning of such arrangements (e.g., color, texture, etc.). Semantics, on the other hand, deals with the meaning of those elements and of their arrangements (e.g., objects, events, etc.). By analyzing Figure 2 from top to bottom, it is apparent that at the lower levels of the pyramid, more high-level reasoning and domain knowledge is required to perform indexing as represented by the width of each level. Although inter-level dependencies exist, each level can be seen as an independent perspective or dimension when observing an image and the way each level is treated will depend on the nature of

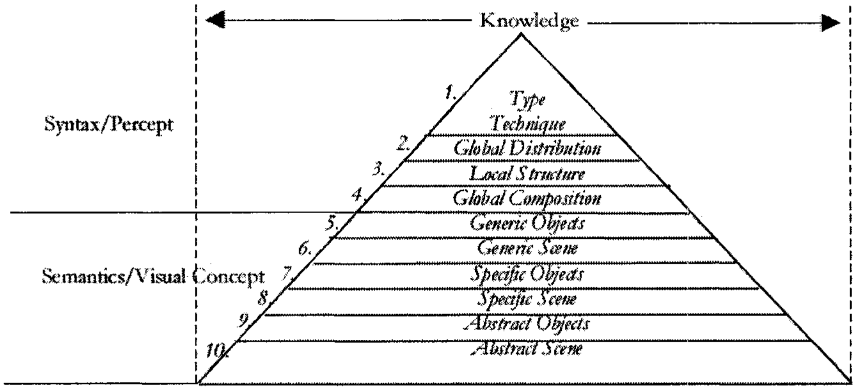


Figure 2: Ten-level indexing pyramid [27]

the database, users and purpose. For clarification, example images for each level of the visual structure is shown in Figure 3 based on [27]. Using structures such as the ones presented, is beneficial not only in terms of understanding the users and their interests, but also in characterizing the limitations of CBIR according to the levels of descriptions used to access visual information.

The most significant gap in CBIR at present lies between levels 1 and 2 in [13] and between the syntax (e.g., levels 1 through 4) and semantics level (e.g., levels 5-10) in [27]. The overwhelming majority of CBIR system offer nothing but level 1 retrieval of [13], and syntax (e.g., levels 1-4) based retrieval of [27]. However, techniques to perform semantic retrieval fully at levels 2 and 3 of [13] and at levels 5 through 10 of [27] are highly desirable. As a result, recent research has been increasingly focusing on moving toward these high levels retrieval [13, 14, 15], which is also the main focus of this research.

The following sections provides a brief overview of the basic components of typical CBIR systems as well as new trends and techniques that are currently prevailing in this domain and the relationships with our work.

2.2 Building Blocks of a Typical CBIR System

The majority of the CBIR systems have some common building blocks or modules although they differ largely in application domains and objectiveness. The building

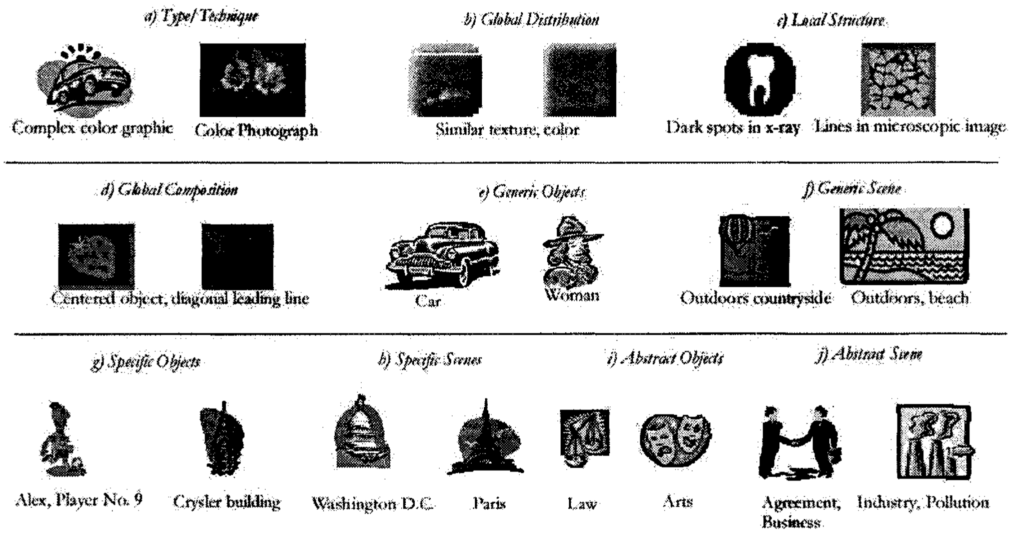


Figure 3: Example images for each level of the visual structure [27].

blocks are [1, 2]:

- A feature extraction module performs extraction of low-level global or region-based local features, such as color, texture, shape, etc. Generally, the feature extraction process is performed off-line for database images due to the large time complexity.

- An index module stores the feature vectors in a file or organize them by applying some multi-dimensional index structure in a logical database. Index mechanism helps filter out irrelevant images compared to a query image based on a nearest-neighbor (NN) search mechanism.

- A similarity matching engine, which compares a query vector and database image feature vectors for rank-based retrieval by applying a distance metric.

- A query processor with an interface in the client side that allows users to formulate queries and to visualize the query results.

Figure 4 shows the functional diagram of the above modules of a typical CBIR system.

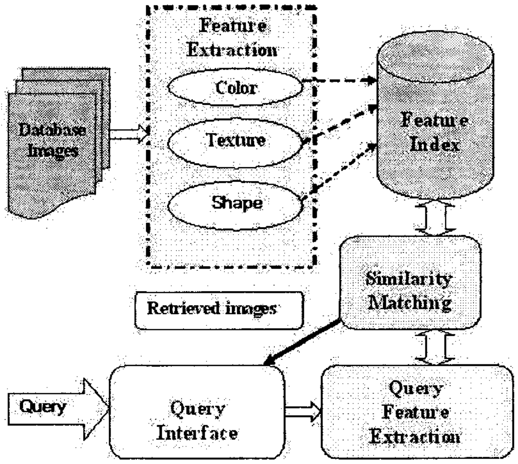


Figure 4: Process flow diagram of a typical CBIR system

2.2.1 Feature Extraction and Representation

Extraction and representation of image features for indexing purpose is the basis of CBIR system [2, 5]. A feature refers to any characteristic which, in some way, describes the content of an image. In feature extraction, each image in the database is transformed with N sets of different feature extraction methods to a set of NM low-dimensional feature descriptors or vectors as shown in Figure 5. Hence, for any given features, such as, color, texture, etc. there exist multiple descriptors (e.g., color histogram, color moment, wavelet-based feature for texture, and so on), which characterize the feature from different perspectives. Because of perception subjectivity, there does not exist a single best descriptor for a given feature. As shown in Figure 5, generally a weight is assigned to each of these features and their descriptors for similarity comparison.

Visual features can be either extracted from the entire image as global feature or from image regions as local feature [2, 5, 10]. Global features are well suited for processing the type of queries that deal with images as single entities during the matching process. In local feature-based retrieval approach, which is generally termed as region-based image retrieval (RBIR) [21, 22], the images are segmented into a collection of homogeneous regions and low-level features are extracted from each region. However, the global features fail to capture enough semantic information due to their limited descriptive power. Although there is a strong correlation between segmented regions

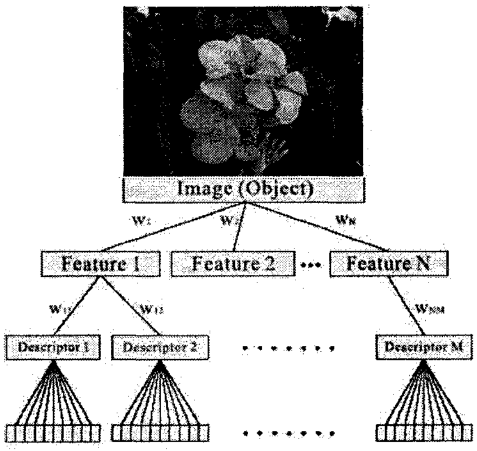


Figure 5: Feature extraction and representation

and real world objects, the accurate automatic segmentation for object detection in broad domain images is still an unsolved problem [24]. Usually, the general-purpose visual features, applicable for a variety of image types, are said to include color, texture, and shape. MPEG-7 [62], a noteworthy standardization initiative for describing multimedia content also follows this categorization, recognizing color, texture, and shape as the three fundamental types of visual features applicable to automated still image content description. MPEG-7 also defines a set of standard visual features or descriptors which have also been used in our work.

Color:

Color is the most widely used feature in CBIR, since it is an important dimension of human visual perception and it is invariant with respect to image scaling, translation and rotation and above all it is computationally least intensive [2, 10]. Histogram is the most commonly used color descriptor, which are obtained by quantizing the color space, such as RGB, LAB, LUV, HSV (HSL), YCrCb, HMMD, etc. and counting how many pixels fall in each discrete color. The drawback of the global histogram representation is the lack of spatial information it retains. To overcome this limitation, many other variants of color representation, such as color moment [59], color coherence vector [60], color sets [12], and color correlogram and autocorrelogram [61]. In MPEG-7 [62], dominant color, scalable color, group of frame or group of pictures, color structure and color layout were defined as a color descriptors as part of the standard. The robustness, effectiveness and efficiency of use of color in image indexing are still open issues.

Texture:

Although there is no strict definition of the image texture, it is easily perceived by humans and is believed to be a rich source of visual information. The existing texture descriptors are classified based on three different approaches as Statistical, Model-based and Transform-based [64, 65, 66, 68]. Haralick in [64], proposed the co-occurrence matrix representation based on statistical approach, which can be used feature to describe spatial relationships between grey-levels in a texture. Many other researchers followed the same line and further proposed enhanced version [65, 66]. Numerous random field models for texture representation [69, 70] were developed based on texture model analysis and a review of some of the recent work can be found in [70]. Transform methods gained its popularity in the late 80's and early 90's. Transform methods based texture representation with Gabor [68] and wavelet transforms [67] are popular now in CBIR. The texture descriptor in MPEG-7 facilitate browsing and similarity retrieval in image and video databases. There are three texture descriptors as homogeneous texture, edge histogram, and texture browsing [63].

Shape:

In many situations, people can recognize an object only by its shape and it is probably the most important property that is perceived about objects. Generally, there are two groups of shape descriptors [71]: boundary or contour-based shape descriptors and region based shape descriptors. Boundary representations emphasize the closed curve that surround the shape. This curve has been described by numerous models, including chain codes, polygons, circular arcs, splines, explicit and implicit polynomials, and boundary fourier descriptors. Region based shape descriptor on the other hand, emphasize the material within the closed boundary or based on the entire shape region. Descriptors of this class include moment invariants, Zernike moments, the morphological descriptors etc [72, 73]. MPEG-7 standard defines three descriptors for shape with different properties: the region-based shape, the contour-based shape, and the 3-D shape spectrum descriptors. It has selected curvature scale space descriptors (CSSD) as contour-based shape descriptors and Zernike moments descriptors (ZMD) as region-based shape descriptors [62].

Several other types of image feature have been proposed as a basis for CBIR [2]. Most of these rely on complex transformations of pixel intensities which have no obvious counterpart in any human description of an image. The most well-researched technique of this kind uses the wavelet transform to model an image at several different resolutions [24].

2.2.2 Similarity Matching

In a traditional database implementation, the user ordinarily makes exact queries and the items matching the query criteria are returned. In CBIR, this kind of exact queries are not that useful as with general images it is difficult to find appropriate matching criteria which would pick only the relevant images. Instead of matching, images are graded using a similarity criterion, resulting in a permutation of all the images in the database sorted according to the used measure of similarity. There are three types of search scheme that are commonly supported by similarity matching schemes [74]: the Range search retrieves all the images within a region of the feature space specified the user, the Nearest Neighbor search finds the k-nearest neighbors to a template (e.g., query) and Within-Distance (or α-cut) find all images with a similarity score better then α with respect to a template, or find all the images at distance less then d from a template.

The above search schemes usually use a distance metric for similarity matching between query and database images [74]. The distance function generally takes two feature vectors as input and outputs a real value as the similarity measurement. In most cases, the smaller of the distance value, the more the similarity between each of the input features. A distance function $d(\mathbf{f}_i, \mathbf{f}_j)$, of feature vectors $\mathbf{f}_i \neq \mathbf{f}_j \neq \mathbf{f}_k$ of images I_i, I_j, and I_k, which is a metric, must satisfy [75](i) reflexivity, i.e., $d(\mathbf{f}_i, \mathbf{f}_i) = 0$, (ii) non-negativity, i.e., $d(\mathbf{f}_i, \mathbf{f}_j) > 0$, (iii) symmetry, i.e., $d(\mathbf{f}_i, \mathbf{f}_j) = d(\mathbf{f}_j, \mathbf{f}_i)$, and (iv) the triangle inequality, i.e., $d(\mathbf{f}_i, \mathbf{f}_j) + d(\mathbf{f}_j, \mathbf{f}_k) \geq d(\mathbf{f}_i, \mathbf{f}_k)$. Some widely used distance measures in CBIR are based on Minkowski-form distance, Quadratic distance, Histogram intersection, and Mahalanobis distance [2, 6, 76, 77]

However, the similarities between images cannot be quantified in an ideal manner. The extent to which the images are similar will change when query requirements are varied. For instance, in the case of two pictures, one of a blue sea with a sunrise and the other of a green mountain with a sunrise, when the sunrise is considered, the similarity between these two images should be high, but if the object of interest is the blue sea, the similarity between these two images should be low. Therefore, when evaluating a CBIR system, one should remember that the retrieval effectiveness depends on the types of query requirements that users make.

2.2.3 Feature Indexing

When the number of images in the database is very large there is a need for indexing to avoid sequential scanning and to support the similarity-based queries. Indexing image databases is much more complex and difficult problem than indexing in traditional databases. Image feature vectors usually have high dimensions. For example, some image feature vectors can have more than 100 dimensions with integrated features of color, texture, shape, etc. High-dimensional spaces lack many intuitive geometric properties we are accustomed to in low-dimensional spaces [11]. This is the well-known *curse of dimensionality* problem [84, 85]. Creating a generalized high-dimensional index that can handle hundreds of dimensions is still an unsolved problem to date. In addition, it may be required to rely on using many features simultaneously in image retrieval as we present such a simultaneous retrieval approach in Chapter 6.

In general, there are two broad categories of index structures for high-dimensional spaces. The first approach is to apply a divide-and-conquer strategy. The data or the feature space is divided into categories (clusters) or subspaces with the intention that only one or a few of these have to be processed in one given query. Some commonly used indexing structure in this category are R-tree [86] and its variants [87], SS-tree [88], agglomerative hierarchical clustering [89], tree-structured vector quantization (VQ) [90], and SOM [36]. Alternatively, we can transform the original feature space into a new space where the operations needed to process a database item are less demanding. This usually means reducing the dimensionality of the original feature space. The mapping from a higher-dimensional to a lower-dimensional space, i.e.

dimensionality reduction, can be accomplished with linear methods such as principal component analysis (PCA) [183, 184] or nonlinear methods such as multidimensional scaling (MDS) [91]. In Chapters 3 and 7, we use PCA-based dimension reduction for effective and efficient similarity matching.

2.2.4 Query Processing

Another important part of the functionality of a CBIR system is the processing of user requests. Current research on CBIR is centered on designing efficient query schemes in order to provide a user with effective mechanisms for image database search. A study conducted by the QBIC group of IBM [16] has shown that users are likely to use the simplest searching interface (i.e., click on an image to find similar images) instead of using more sophisticated user interfaces, as provided in [21, 22].

One of the most popular forms of query in this domain is the *"query by example"* (QBE) image, which requires that the user provides a prototype image to the system as a reference example. In QBE, the image query is based on an example or reference image shown either from the database itself or the user may provide the image externally.The task of the retrieval system is then to return images as similar to the example image as possible. A closely related query type to using an external example is query by sketch, in which the example image is generated by the user on the fly using a sketching tool included in the retrieval interface [10]. The main problem with sketching is that users often find it difficult to produce an adequate sketch of the visual concept they are looking for. In certain restricted domains such as trademark retrieval, *query-by-sketch* can be a valuable search option. *Iconic querying* is a variant of the sketch approach where the user creates an example image by selecting predefined icons, such as human faces, trees, sky etc. They are selected from a predefined set and combined through Boolean connectives so as to create a visual sentence according to a visual language. This approach is also usually infeasible in a general setting, as it requires rather sophisticated object recognition and the queries can only contain visual concepts supported by the system. Keyword-based search techniques of text retrieval domain would also be applicable to image searching if annotation of the contents of the images are available (e.g. commercial image libraries and medical databases) or they could be automatically produced. In this case, traditional query by keywords can be used in conjunction with QBE or other querying methods (such a multimodal query approach will be presented in Chapter 6).

2.3 Towards Semantic-Based Image Retrieval

While the above functionalities behind the CBIR systems is undoubtedly impressive, user take-up of such systems has so far been minimal. Users typically do not think in terms of low-level features, i.e., user queries are typically semantically oriented and most of the CBIR systems have poor performance for these type of queries. By semantics we mean how meaning is embodied in the features and what concepts can infer from the images.

In order to improve the retrieval accuracy and overcome the limitations of CBIR systems, research focus has been shifted from designing sophisticated low-level feature extraction algorithms to reducing the *semantic gap* between the visual features and the richness of human semantics. Although we have witnessed some new trends in this research direction during the last several years [13, 15], research into techniques for semantic retrieval is still in its infancy. Many of the techniques now being applied to the problem have been adapted from other classical fields, such as object recognition, machine learning, information retrieval, and human computer interaction. To develop a complete understanding of image contents at the semantic level for broad domain images is a formidable task, well beyond the capabilities of current technology. Much success has been achieved in recognizing a relatively small set of objects or concepts within specific domains or constraint environment, such as face, finger print, and iris recognition or detecting people in a scene [13].

However, we can be optimistic with the fact that it may not require a complete understanding of images as a human being to perform effective image retrieval at least at level 2 in [13] and levels 5-8 in [27]. Instead, sometimes it only requires to interpret only major objects and their relationships in images. Similar analogy can be made in text retrieval domain where majority of the systems are successful without a complete understanding of the contents in a document. New research also suggests that instead of giving more reasoning and logic to the system or machine, it is effective to involve users to interact with the system. As a result, some successful interactive CBIR systems have emerged [108, 109, 110, 114, 113]. Due to the huge growth of the web, research also focuses on context instead of content based retrieval or a combination of both [117, 120, 123, 125]. By context, it means finding semantics or meaning of images from the related text associated with images. Unifying texts with images is the major trend along this direction, which focuses on an integrated approach to retrieve image data.

In the following sections, we briefly describe some of the new trends and techniques currently prevailing either in a direct or an indirect way to CBIR and also show their relationships with our work.

2.3.1 Semantic Classification and Annotation

Classification means grouping of images into semantic meaningful classes. In a database whose semantic description is reasonably well defined, automatic semantic classification can greatly enhance the performance of CBIR systems. We can observe few research prototypes that successfully classify the following images in photographic and medical domains:

- Natural photographs vs. artificial graphs and textures vs. non-textured images [24].
- Indoor vs. outdoor images and further classified as city (man made) vs. landscape (natural objects) [98, 99].
- Medical images of different modalities (e.g., x-ray, CT, MRI, etc.), body parts (e.g., head, chest, abdomen, etc.), orientations (frontal, sagittal, post anterior, etc.) and so on [145, 144].

In most cases, to derive high-level semantic classification requires the use of statistical learning techniques [15]. Supervised learning-based methods such as SVMs (will be described in Section 3.1 of Chapter 3), Bayesian classifier, and Neural Networks [92, 99, 97, 100, 93, 103] are often used for image classification at a semantic level. The above statistical approaches have the advantage of not requiring the construction of complex and possibly domain-specific models of each type of object to be recognized, relying totally on statistical associations between image semantics and quantifiable low-level properties, learnt in most cases from a training set of a few hundred examples at best. Approaches of classification can be divided as low-level global and region-specific local features depending on how the features are used in the classification tasks. Some approaches mainly used low level global features as their sole classification scheme to classify image collection at a global level [92, 98]. On the other hand, few approaches used region-specific local features to classify objects in individual images instead of classifying the entire collection [100, 24, 126, 96, 102].

These classification-based systems can permit a degree of automatic annotation by assigning keywords at a global level, such as beach, mountain, city scene, etc. or at a local level, such as sky, water, etc. for natural photographic images. Automatic annotation or linguistic indexing of images is essentially important to CBIR, which will lessen the cumbersome manual annotation by an expert. We will focus more on this issue in Chapter 3 and Chapter 4 to show that how effectively retrieval can be performed based on automatic annotation of images both at global and local levels by utilizing machine learning and information retrieval techniques [206, 207].

2.3.2 Interactive Image Retrieval

The approaches of early CBIR systems were isolated, since there was no real human-computer interaction, except only when the user provides a set of weights for different features. Such isolated approach had limitations as a user is an indispensable part of the CBIR system when compared to any pattern recognition systems. To overcome the limitations of the isolated approach, majority of the current retrieval systems have moved to an interactive mechanism that involves users directly as part of the retrieval process. A natural way of getting user in the retrieval loop is to ask him/her to provide feedbacks regarding the current output of the system, which is commonly known as *relevance feedback* (RF) [104, 108, 109] technique. Though this is an idea that was originally generated in text retrieval field [104], it performs better in image domain as it is easier to tell the relevance of an image than that of a document. Compared to the off-line machine learning techniques, RF is an on-line processing which tries to learn the user's intentions on the fly. A typical scenario for RF in CBIR is as follows [109]: a user presents an image query to the system where upon the system retrieves a fixed number of images using a default similarity metric. The user then rates each returned result with respect to how useful the result is for his or her retrieval task at hand. Ratings may be simply *relevant* or *not relevant* or may have finer gradations of relevancy such as *somewhat relevant*, *not sure* and *somewhat irrelevant*. The RF algorithm uses this information to select another set of images to

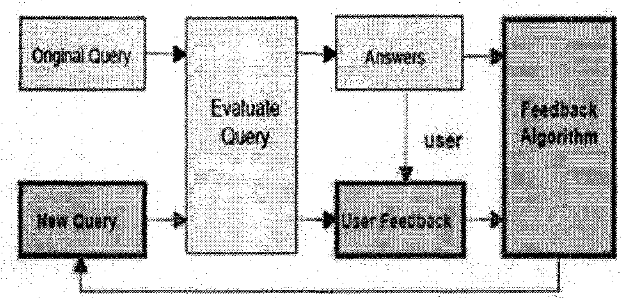


Figure 6: Basic Block diagram of an Interactive RF-based Retrieval System

retrieve for the user; whether the new and previous sets are disjoint depends on the particular system. The goal of a system is to effectively infer which images in the database are of interest to the user based on this feedback. The user could then rate these images in the second set in a similar way and the process may iterate indefinitely in this closed-loop fashion until the user is satisfied with the results. Figure 6 shows a block diagram of a typical interactive RF-based system based on the above mentioned processes.

A number of RF based techniques have been proposed, such as query point movement, feature re-weighting, and active learning [108, 109, 110, 111, 112, 113]. A more detailed survey of these techniques can be found in [109]. The majority of these approaches estimate the ideal query parameters from the low-level global image features [108, 110, 111]. However, for high-level concepts or semantics that cannot be sufficiently represented by low-level features, these systems will not return many relevant results even with a large number of feedback iterations. To address this limitation, recently some systems incorporate RF on both low-level content and high-level context (keyword) based feature spaces [113, 120, 114] for web image retrieval. Being motivated by the success of these systems, we present an interactive cross-modal retrieval framework in Chapter 6. The proposed retrieval technique is highly adaptive in nature, since in addition to perform query point movement and feature re-weighting (e.g., both inter and intra modality weights), it also perform cross-modal multiple query expansions based on user's feedback information.

2.3.3 Multimodal Image Retrieval

Image retrieval based on multi-modal information sources has been recently gaining popularity due the the huge amount of multi-modal information available on the web (e.g., images with collateral texts in image captions, headers, titles and other places in HTML or XML documents) [115, 116, 117, 124, 125, 120, 123]. Retrieval systems of these kind showed improvement in performances by fusing the evidences from textual information and visual image contents in a single framework. The results of the past ImageCLEF tracks [55, 56, 57, 58] also validated this fact that in many cases the combination of visual and text based image search provides better results than using the two different approaches individually. In CBIR systems based on the *query-by-example* strategy it is also hard to find an initial image to give to the system as an example to find similar image, which is called as *page zero problem* [120]. By adding text to images, the user can type some words and find relevant images to start a visual search in next iteration.

In general, there are two main combination techniques currently investigated [115]: (1) The text and image modalities are sequentially used; and (2) The text and image modalities are simultaneously used, combined either linearly or nonlinearly. In a sequential approach, only one modality is used at each iteration for a query. The user can browse a set of images within a category. When both modalities are used simultaneously, a composite query is formed to include both text and an image. During retrieval, the system combines the similarities computed from the text and images [119, 124, 119]. The combined similarity is the weighted sum of similarities computed from both modalities. Different systems differ in the computation of similarity measures and inter-modality weights. There are three general ways to set the inter-modality weight. (1) The weight may be set manually according to prior knowledge. For example, the two modalities are given equal weight to linearly combine similarities from textual features (using dot products) and visual features (using Euclidean distances) in the iFind system [124]. (2) The weight may be learned off-line. For example, to find the optimal weight set for the multiple modalities, a training phase is used to learn a group of optimal weights for a selected set of representative queries in the MMIR system [122]. In the retrieval phase, the individual models for different modalities are linearly combined using the weight set of the representative query which is the most similar to the current user-submitted query. (3) The weight may be adjusted on-line based on user's feedback. For example, relevance feedback (RF) is used to adjust the intra-modality and inter-modality weights in [125, 124, 119, 121].

Text and images may also be seamlessly integrated in a probabilistic model. Several probabilistic learning models used for matching image and text are evaluated in a paper by Barnard et al. [126]. The models can be used for cross-modal information retrieval, for retrieval with a composite query, and for automatic image annotation. In multimodal image retrieval, based on the assumption that semantically related images may be visually similar and vice versa, semantic networks also can be created to link keywords with images or regions of images [125]. Semantic networks may be used to expand a query in various ways: expand a text query using related keywords, expand a text query using visual content related to the keywords, and expand an example based image query using text related to the images.

Many systems combine one or more of these above techniques in a single framework. For example, RF is combined with machine learning techniques [127], some web-based image retrieval systems employed both context features and RF methods to reduce the semantic gap [120, 113, 114] and we also present such an integrated interactive multimodal retrieval appraoch in Chapter 6.

In the following section, we also provide a brief review of image retrieval techniques that are prevailing in medical domain due to the special characteristics of images and significance of this domain in our work.

2.4 Image Retrieval in Medical Domain

The digital imaging revolution in the medical domain over the past three decades has changed the way present-day physicians diagnose and treat diseases. Hospitals and medical research centers produce an increasing number of digital images of diverse modalities everyday [128, 129, 131, 130, 215]. Examples of these modalities are the following: standard radiography (RX), computer tomography (CT), magnetic resonance imaging (MRI), ultrasonography (US), angiography, endoscopy, microscopic pathology, etc.. These images of various modalities are playing an important role in detecting the anatomical and functional information about different body parts for the diagnosis, medical research, and education. Due to the huge growth of the web, medical images also are now available in large numbers in online repositories and atlases [128, 132]. Modern medical information systems need to handle these valuable resources effectively and efficiently.

However, characteristics of medical images differ significantly from general photographic images. Medical images are multi-modal where each modality reveals anatomical and/or functional information of different body parts. Each of the imaging modality has its own set of requirements, such as file format, size, spatial resolution, dimensionality, and image acquisition and production technique. Generally, images are stored as two-dimensional arrays of pixels that can represent calculated x-ray attenuation values, sound intensity, electron density, or various properties of radio waves. The common file format for radiological images is DICOM (Digital Imaging and Communication in Medicine) [134], which contains some additional information regarding image modality, acquisition device, and patient identification in its header along with raw image data. A two-dimensional DICOM image may have a size much larger than other general image formats, such as JPEG, GIF, TIFF etc. Moreover, images of some modalities, such as PET, fMRI contain functional information, which require different kinds of processing. The large image size, high resolution, multimodality, data heterogeneity, structural and functional contexts are the key issues in medical domain and special attention is required when performing image analysis and retrieval techniques [128, 129, 131].

2.4.1 Retrieval Techniques

Currently, the utilization of medical images is limited due to the lack of effective search methods; text-based searches have been the dominating approach for medical image database management [128, 129]. Many hospitals and radiology departments nowadays are equipped with Picture Archiving and Communications Systems (PACS) [133]. In PACS, the images are commonly stored, retrieved and transmitted in the DICOM format [134]. Such systems have many limitations because the search for images is carried out only according to the textual attributes of image headers (such as standardized description of the study, patient, and other technical parameters). The annotations available are generally very brief in the majority of cases as they are filled out automatically by the machine. Moreover, in a web-based environment, medical images are generally stored and accessed in other common formats since they

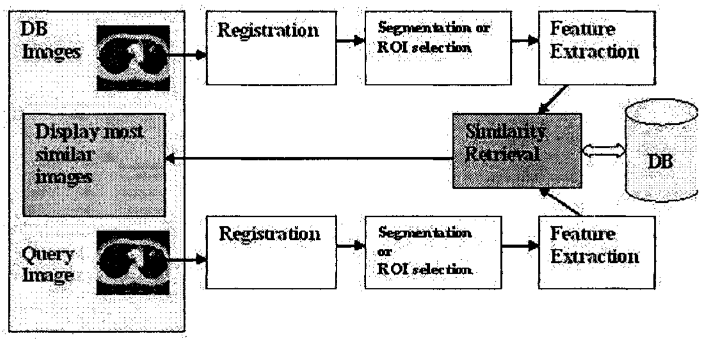


Figure 7: Block diagram of a CBIR system as a diagnostic aid

are easy to store and transmit compared to the large size of images in DICOM format. However, there is an inherent problem with the image formats other than DICOM, since there is no header information attached to the images and thus it is not possible to perform a text-based search without any associated annotation information. The representation and retrieval of clinical images within text-based framework is problematic as medical image data differs in many ways from text based medical data [135]. Pathological and anatomical information contained in medical images are domain specific and need sophisticated computer vision and image processing algorithm to extract and retrieve it.

As a result, CBIR systems have emerged in this domain and one of the main focus of the researchers during the last decade [137, 138, 139, 140, 141, 142, 128]. However, CBIR is more challenging in medical domain than other general-purpose image domains. The main reason is that, important features in biomedical images are often local features rather than global features. Generating local features that can describe fine details of images, is much more complex than global features [135].

The design of a CBIR system will also depend on its application domain. For example a CBIR for teaching and research need to be designed differently than one is required for diagnostic purpose. In a diagnostic based retrieval system, visual features of normal and pathological images are typically separated by only subtle differences in visual appearance, which may not be captured by traditional features such as color, texture or shape [128, 131]. Here, a combination of prior modality specific domain knowledge and image primitive content related to anatomic structure is necessary

[131, 135]. These may include size measurements such as organ volumes, volumes of pathological tissues, relative position of anatomical structure or specific rules such as ABCD rule for detecting melanoma in dermoscopic images [149]. Different image modalities have significantly different visual properties; hence require different types of processing [128]. In radiology images (X-ray, CT, MRI, PET) gray-scale and texture and in microscopic slides and dermoscopic images color and texture features might play more important role. For example, in CT images of lung, pathology bearing regions (PBR) can be better described by a small change in texture of the lung tissues [137]. To extract local features from PBR, segmentation is always a very important step in medical imaging. However, automatic segmentation algorithms for detecting PBR's are not mature enough and manual segmentation is a widely used option at this moment.

Figure 7 shows a block diagram of a typical CBIR as a diagnostic aid. The images are generally registered to a common global coordinate system to ensure that subtle pathologies apparent in the query image are matched to a correct region of a database image of a similar pathology. So registration can up to some extent solve the rotation or transformation variance problem in image retrieval. The pathology bearing region (PBR) is then highlighted with a manual or automatic segmentation scheme and various features and measurements are extracted or computed from this region of interest. These features or measurements are mainly robust to slight misalignment and invariant to various imaging artifacts and highlight the indication of any pathology. Theses features are indexed in the database as an index file along with the original raw images or a pointer is kept to locate the original images stored somewhere else. When a query image is submitted, it must also be aligned to the database global coordinate system and the same way segmentation and feature extraction is performed as shown in the bottom level in Figure 7. The features are then matched in a similarity retrieval subsystem to every image in the database and the top matching images are shown to the query interface according to their ranks along with any additional information such as case or lab report related to them.

2.4.2 Application Areas and Projects

Medical image retrieval systems with advanced browsing and searching capabilities can play an increasingly important role in medical training, research, and diagnosis

[128, 129]. However, requirements for retrieval in this domain differ based on the application areas and need special attention due to the image properties. In a clinical decision-making process or diagnostic purpose, a CBIR system can supply the physicians with cases that offer a similar visual appearance. For instance, in evidence-based medicine, if a doctor encounters a difficult case and needs some supporting evidence to make his diagnosis, the CBIR would allow him to consult similar images associated with a confirmed diagnosis [131]. Medical imaging archives may act in the same way as data warehouse for database management systems. As important patterns are mined in a data warehouse to predict future outcomes, physicians can mine the image repositories for finding similar pattern or images based on the current image under examination. In research and clinical trials, physicians may want to detect the changes recorded in a current image based on the previous images generated for the same patient. This is a very common scenario for detecting the growth of a tumor in a radiological or dermatological images. For example, brain MRI image databases used in clinical trials to detect and track the growth of lesions (multiple sclerosis) automatically [136]. Medical retrieval systems also can be an effective tool for web-based biomedical education [129]. Students can browse large image repositories by their visual content, lecturers can find important or interesting cases based on visual similarity to show relevant diagnostic and potential problems.

During the last decade, several image retrieval prototypes have been implemented in the medical domain [137, 138, 139, 140, 141, 142, 143, 128]. For instance, the ASSERT system [137] is designed for high resolution computed tomography (HRCT) images of the lung, where a rich set of textural features and attributes that measure the perceptual properties of the anatomy are derived from the pathology-bearing regions (PBR). The WebMIRS [4] system [138] is an ongoing research project. The project aims at the retrieval of cervical spinal X-ray images based on automated image segmentation, image feature extraction, and organization along with associated textual data. I-Browse [139] is another prototype, aimed at supporting the intelligent retrieval and browsing of histological images of the gastrointestinal tract. In IGDS system [140], a classification based approach is performed to detect different kind of blood cells in cytopathology images based on the properties of cell nucleus. The majority of these current prototypes or projects concentrate mainly on a specific

[4] http://archive.nlm.nih.gov/proj/webmirs/

imaging modality [128]. The main limitations of these systems is the lack of accurate and semantically valid automatic segmentation of pathology bearing region (PBR) or region of interest (ROI). Most of them are currently relying on manual or semi-automatic segmentation methods.

To date, only a few research projects has as their objective to create CBIR systems for heterogeneous image collections. For example, the IRMA (Image Retrieval in Medical Applications)[5] system [141] is an important project that can handle retrieval from a large set of radiological images obtained from hospitals based on various textural features. The medGIFT [6] project [142] is based on the open source image retrieval engine GNU Image Finding Tool (GIFT). It aims to retrieve diverse medical images where a very high-dimensional feature space of various low-level features is used as visual terms analogous to the use of keywords in a text-based retrieval approach. In other general purpose medical CBIR systems, such as in I^2C [143] or in COBRA [133], the low-level visual. features are extracted either from an entire images or from segmented image regions. However, using only low-level features directly without any semantic interpretation has very limited applicability for these approaches. Many other CBIR systems in medical domain are currently available and a brief introduction of each of the systems are available in [128]. In this research, we propose and develop a classification-driven image retrieval framework for heterogeneous medical image collection by exploiting category-specific information in two different ways (Chapter 3) as well as develop a CBIR system for a specific modality (e.g., dermoscopic images) as a diagnostic aid (Chapter 7) by exploiting the domain specific knowledge and the image characteristic.

[5] http://phobos.imib.rwth-aachen.de/irma/

[6] http://www.dim.hcuge.ch/medgift/

Chapter 3

A Global Concept-Based Image Retrieval Framework

In this chapter, we present a learning-based classification-driven image retrieval framework to bridge the *semantic gap* by transforming low-level feature spaces to high-level generic or global concept-based feature spaces. The concepts are inferred from entire images instead of individual object based semantics and used for both image representation and adaptive similarity matching for effective retrieval in a semantic level [205, 206, 212, 213, 214]. Current CBIR systems can be distinguished as either global or region-specific in nature based on their feature representation and employed similarity matching functions [2, 5]. Therefore, the similarity comparison between a query and target images in CBIR is performed either globally based on visual features from entire images or locally based on features derived from automatically segmented image regions. As we mentioned earlier in Chapter 1, images with high feature similarities might be different from a query image in terms of the semantics at a global context or semantics as perceived by users. The main reason is that both global and local features fail to capture enough semantic information due to their limited descriptive power. Although there is a strong correlation between segmented regions and real world objects, the accurate automatic segmentation for object detection is still an unsolved problem in computer vision.

There exist some image domains (generally narrow domains) where images can be organized with prior semantic groupings in a generic way that can depict the semantics of images as a whole instead of individual object-based semantics in images.

For example, in a heterogeneous medical image collection for teaching and training applications, images are generally organized at different levels of abstraction with imaging modalities, body regions, orientations, biological system, and so on [145]. A collection of natural photographic images or consumer photos can be classified as indoor and outdoor at the very top level and further they can be classified as natural scenery (e.g., mountain, lake, etc.) and man-made objects (e.g., city, road, etc.) for outdoor images and it can go further deeper levels [98, 99].

To enable effective search in such semantically organized collections, it might be advantageous for a retrieval system to be able to recognize the current image class prior to any kinds of post-processing or similarity matching [144, 145]. A successful categorization of images would greatly enhance the performance of CBIR systems by filtering out irrelevant images and thereby reducing the search space. For example, for a query like *"Find posteroanterior (PA) chest X-rays with an enlarged heart"* in a medical collection, images at first can be pre-filtered with automatic categorization according to modality (e.g., X-ray), body part (e.g., chest), and orientation (e.g., PA). The latter search could be performed on the pre-filtered set to find the enlarged heart as a distinct visual property. The automatic classification will also allow the labeling or annotation of unknown images up to certain axes. For example, a category could denote a code corresponding to an imaging modality, a body part, a direction, and a biological system, in order to organize images in a general way without limiting them to a specific modality, such as the IRMA code [145] in medical domain. Based on image categorization and subsequent annotation, semantical retrieval can be performed by applying techniques analogous to commonly used ones in CBIR domain. This simple yet relatively effective solution has not been investigated adequately in current CBIR systems. Many approaches have been explored recently to classify image collections into multiple semantic categories in general photographic and medical domains [98, 92, 93, 145, 99, 13, 15]. However, these approaches did not relate directly the usefulness of classification information towards semantic-based image retrieval. For example, images are classified as indoor/outdoor and further classified as city (man made)/landscape (natural scenery) in general photographic domains [98, 99, 92] without any direct relation to image retrieval. An automatic soft image annotation approach is investigated recently in [93] by using Bayes Point Machines (BPMs)-*One against the others* ensemble. Only, the approach in [93] showed some promising direction towards semantic-based retrieval by performing keyword-based search on annotated images at a global level. In medical domain, the automatic categorization of radiological images is examined in [145] by utilizing a combination of low-level global texture features with low-resolution scaled images and a K-nearest-neighbors (KNN) classifier. In [132], the performances of two medical image categorization architectures with and without a learning scheme are evaluated on images based on modality, body part, and orientations. These approaches demonstrated promising results for medical image classification at a global level. We will extend the above approaches further for semantical image representation, feature level fusion, and adaptive similarity matching by exploiting category-specific information of a collection.

The aim of our retrieval framework is to reach some important global or generic semantic concepts that are inferred from entire images by utilizing low-level image features at different levels of abstraction. In this framework, support vector machines (SVMs) [165], a popular supervised learning technique is utilized for semantic modeling of images at the global concept level. Instead of hard classifying an image to only a single category in a mutually exclusive way, we perform a soft classification approach based on probabilistic outputs of the classifier. In this approach, various low-level global, semi-global and low-resolution scale-specific image features are extracted to represent different aspects of images. The SVMs are utilized to associate these low-level features with their high-level global semantic categories. The utilization of the probabilistic outputs of multi-class SVMs and the classifier combination rules derived from Bayes's theory [173] are explored for representation of images in global concept-based feature spaces with a successive semantic level of information abstraction. In addition, to perform retrieval effectively in low-dimensional feature spaces without transforming them to the concept spaces, we exploit the category-specific feature distribution information in an adaptive similarity matching function [205]. In this approach, a classifier is trained based on low-dimensional image features to predict the global categories of unknown query and database images on-line. Based on the on-line prediction, pre-computed category specific first and second order statistical parameters are utilized in a similarity matching function on the assumption that distributions are multivariate Gaussian. The proposed image representation and similarity matching schemes have proved to be effective compared to commonly used low-level image features and L-norm or geometric similarity measures in CBIR domain, as demonstrated experimentally in Chapter 8. In the following sections, the classification technique based on multi-class SVMs is presented at first and then we will describe the proposed image representation and similarity matching approaches.

3.1 Multi-class SVMs

Recently, SVMs have become a standard tool for supervised learning-based classification and the subject of intense research in both theory and application [165, 167, 168, 166]. In supervised learning, a semantic concept is defined at first by a sufficient number of training samples. A classifier creates a function from the training data, where instances in the training set contain category or class specific labels along with their feature descriptors [174, 173]. The task of the supervised learner or classifier is to predict the label of a newly encountered unknown data after having seen only a small number of training examples. To achieve this, the learner has to generalize from the presented data to unseen situations in a reasonable way. Hence, the success of a classifier is highly dependent on its generalization ability [173]. The empirical success or good generalization ability of SVMs has already been proved to detect and analyze complex patterns in data, such as in text categorization [169], hand-written character recognition [171], face recognition [170], general photographic and medical image classification [132, 92], and so on.

Briefly, SVMs constructs a decision surface between the samples of two classes, maximizing the margin between them differing from the other classifiers [165, 167]. The basic training for SVMs involves finding a function which optimizes a bound on the generalization capability, i.e., performance on unseen data. We are given N observations $\{\mathbf{x}_1, \ldots \mathbf{x}_i, \ldots, \mathbf{x}_N\}$ that are vectors in space $\mathbf{x}_i \in \Re^d$ with associated labels $y_i \in (+1, -1)^N$. The objective is to train a mapping $\mathbf{x} \mapsto y = f_\alpha(x)$ and to find the vectors α of parameter that balances empirical risk and generalization error. The set of observations or training vectors are linearly separable if there exists a hyperplane $(\mathbf{w}^t, b) \in \Re^{d+1}$ where $\mathbf{w}^t \in \Re^d$ is its normal and $b \in \Re$ is its canonical distance to the origin, for which the positive examples lie on one side and the negative examples on the other. Provided that all of the observations are linearly separable, the goal is then to find the parameters $\mathbf{w}$ and b for the optimal separating hyperplane

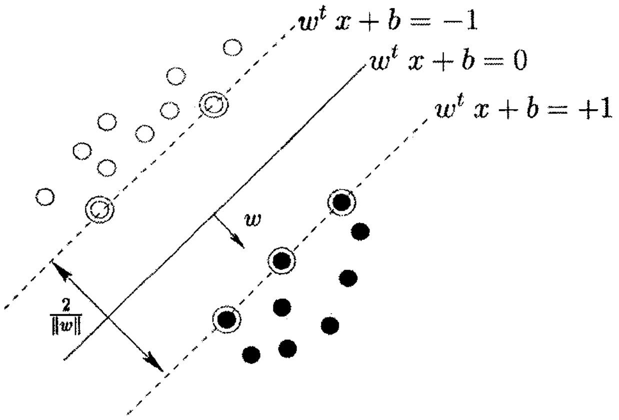


Figure 8: Maximal margin classifier (separable case) .

(OSH) to maximize the geometric margin $\frac{2}{\|\mathbf{w}\|}$ between the hyper planes as shown in Figure 8. The OSH is found by minimizing the L_2 norm of $\mathbf{w}$ by solving the following constrained minimization problem [166]:

$$\begin{array}{ll} \min\limits_{\mathbf{w},b} & \frac{1}{2}\mathbf{w}^t\mathbf{w} \\ \text{subject} & \text{to, } y_i\ (\mathbf{w}^t\mathbf{x}_i + b) - 1 \geq 0,\ \forall\ i. \end{array} \tag{1}$$

Using Lagrangian multipliers, the primal form of the objective function is:

$$\begin{array}{ll} L(\alpha, w, b) = & \frac{1}{2}\mathbf{w}^t\mathbf{w} - \alpha_i(y_i\ (\mathbf{w}^t\mathbf{x}_i + b) - 1) \\ \text{subject} & \text{to}, \alpha_i \geq 0,\ \forall\ i. \end{array} \tag{2}$$

The function $L(\alpha, w, b)$ is minimized with respect to w, b and maximized with respect to α. A local minimum of $L(\alpha, w, b)$ is found when its underlying gradient vanishes with respect to w and b and can be specified using only the parameter α. Now, the dual form of constraint optimization problem is changed to a constraint quadratic problem. Again, when solving this problem, training data for which the coefficients α_i are different from zero values are called support vectors (e.g., circles with two rings in Figure 8) and are the closest data elements to the OSH. The general form of the binary linear classification function is

$$f(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + b \tag{3}$$

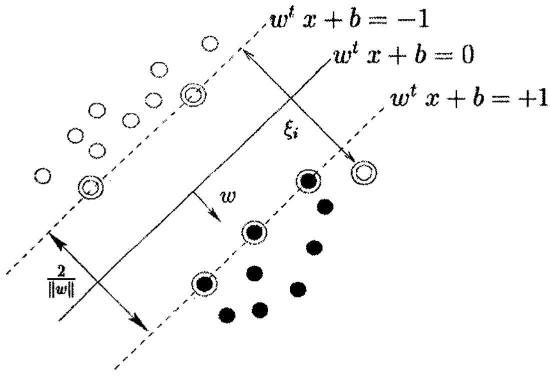


Figure 9: Maximal margin classifier (non-separable case).

In the case when the training set is not linearly separable, slack variables ξ_i are defined as the amount by which each $\mathbf{x}_i$ violates the constraint $y_i\ (\mathbf{w}^t\mathbf{x}_i + b) \geq 1$. Using the slack variables as shown in Figure 9, the constraints in (1) are relaxed, if necessary, and the new constrained minimization problem becomes:

$$\begin{aligned} &\min_{\mathbf{w}, b, \xi} \quad \frac{1}{2}\mathbf{w}^t\mathbf{w} + C\sum_{i=1}^{N}\xi_i \\ &\text{subject to, } y_i(\mathbf{w}^t\mathbf{x}_i + b) \geq 1 - \xi_i \\ &\xi_i \geq 0,\ \forall\ i \end{aligned} \tag{4}$$

Here C is a penalty term related to misclassification errors.

In SVM training, the global framework for the non-linear case consists in mapping the training data into a high dimensional space where linear separability will be possible. Here training vectors $\mathbf{x}_i$ are mapped into a high dimensional Euclidean space by the non linear mapping function $\phi : \Re^d \rightarrow \Re^h$, where $h > d$ or h could even be infinite. Both the optimization problem and its solution can be represented by the inner product. Hence,

$$\mathbf{x}_i \cdot \mathbf{x}_j \rightarrow \phi(\mathbf{x}_i)^t\phi(\mathbf{x}_j) = K(\mathbf{x}_i, \mathbf{x}_j) \tag{5}$$

where the symmetric function K is referred to as a kernel under Mercer's condition

[167]. Under non-linear case, the SVM classification function is given by [165]:

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{N} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b\right) \tag{6}$$

Thus the membership of a test element $\mathbf{x}$ is given by the sign of $f(\mathbf{x})$. Hence, input $\mathbf{x}$ is classified as 1, if $f(\mathbf{x}) \geq 0$, and as -1 otherwise.

The SVMs were originally designed for binary classification problems. However, when dealing with several classes, as in general medical image classification, one need an appropriate multi-class method. As two-class or binary classification problems are much easier to solve, a number of methods have been proposed for its extension to multi-class problems [177, 93]. They essentially separate L mutually exclusive classes by solving many two-class problems and combining their predictions in various ways. For example, *One against one* or pairwise coupling (PWC) method [176, 175] constructs binary SVMs between all possible pairs of classes. Hence, this method uses $L * (L - 1)/2$ binary classifiers for L number of classes, each of which provides a partial decision for classifying a data point. During the testing of a feature $\mathbf{x}$, each of the $L * (L - 1)/2$ classifiers votes for one class. The winning class is the one with the largest number of accumulated votes. On the other hand, *One against the others* method compares a given class with all the others put together. It basically constructs L hyperplanes where each hyperplane separates one class from the other classes. In this way, it generates L decision functions and an observation $\mathbf{x}$ is mapped to a class with the largest decision function. In [177], it was shown that the *One against one* or PWC method is more suitable for practical use than the other methods, such as *One against the others.* Hence, we use the *One against one* multi-class classification method by combining all pairwise comparisons of binary SVMs [175].

3.2 Low-Level Image Feature Representation

The success of an image classification system depends on the underlying image representation, usually in the form of a feature vector. In this framework, our aim is to reach global semantic concepts which can be inferred from the entire image instead of individual object based semantics. Therefore, low-level features are extracted from entire images at a global or semi-global levels. In recent years, numerous low-level feature representation schemes based on color, texture, and shape are proposed [2, 10, 11]. We characterize images by color layout, edge distribution, color and texture moments, and average grey-level features to generate the feature vectors as initial inputs to the classifiers. These features are selected so that they can represent images from different aspects according to their compatibility with the global semantic concepts. Following sections briefly describes each of the feature extraction and representation approaches.

3.2.1 Color Layout Descriptor (CLD)

To represent the spatial structure of images, we utilize the Color Layout Descriptor (CLD) of MPEG-7 [62]. CLD is a compact and resolution invariant representation of spatial distribution of colors in an image [63]. It is especially recommended for applications that need to be fast and are based on spatial-structure of color. It is obtained by applying the discrete cosine transformation (DCT) on the 2-D array of local representative colors in the $YCbCr$ color space where Y is the luma component and Cb and Cr are the blue and red chroma components. Each channel is represented by 8 bits and each of the 3 channels is averaged separately for the 8×8 image blocks. The scalable representation of CLD is allowed in the standard meaning that one can select the number of coefficients to use from each channel's DCT output. For each channel, 3, 6, 10, 15, 21, 28 or 64 coefficients can be used. The coefficients are taken from 8×8 arrays in zigzag scan order. Hence, for different collections, the coefficient can be set differently to form a feature vector $\mathbf{f}^{\mathrm{CLD}}$, whose dimension depends on the total number of coefficients.

3.2.2 Edge Histogram Descriptor (EHD)

In this work, spatial distribution of edges are utilized for image classification by using MPEG-7 Edge Histogram Descriptor (EHD) [62]. The EHD represents local edge distribution in an image by dividing the image into 4×4 sub-images and generating a histogram from the edges present in each of these sub-images. Edge detection is performed inside each of the sub-images. Edges in the image are categorized into five types, namely vertical, horizontal, 45° diagonal, 135° diagonal and non-directional

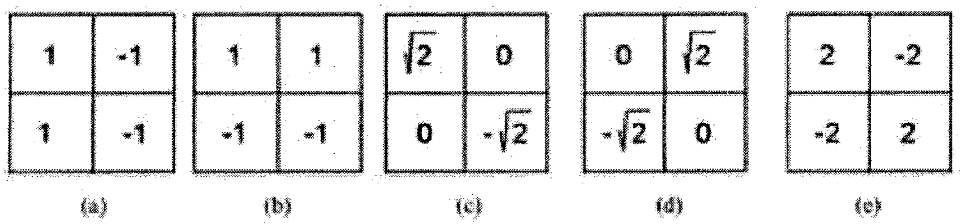


Figure 10: Filters used for edge detection (a) vertical, (b) horizontal, (c) diagonal 45°, (d) diagonal 135°, (e) non-directional

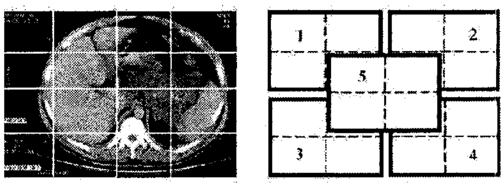


Figure 11: Region generation from sub-images

edges. The filters for edge detection are shown in Fig. 10. After applying the filters, if maximum result obtained by the filters exceeds a threshold, an edge with the type of the filter is reported to be found and the corresponding histogram bin is incremented. The histogram, constructed by the result of this process, is then normalized according to the size of the image. Finally, for 16 sub-images, a histogram with $16 \times 5 = 80$ bins or an 80-dimensional feature vector is obtained as $\mathbf{f}^{\mathrm{EHD}}$. Both the EHD and CLD are selected by MPEG-7 because of their success and reliability as a result of some experiments. Using these features as classifier's inputs give us the chance to gain experience about a newly emerged standard that will probably constitute the core of future multimedia applications.

3.2.3 Moment-Based Feature

A simple grid-based approach is used to divide the images into five overlapping sub-images [211]. These sub-images are obtained by first dividing the entire image space into 16 non overlapping sub-images. From there, we cluster four connected sub-images to generate five different clusters of overlapping sub-images as shown in Figure 11. The first (mean), second (standard deviation) and third (skewness) central moments of each color channel are extracted in HSV (Hue, Saturation, and Value) color space from each cluster. We also extract texture features from the grey level co-occurrence matrix (GLCM) [64] of each sub-image. A GLCM is defined as a sample of the joint probability density of the gray levels of two pixels separated by a given displacement d and angle θ. Typically, the information stored in a GLCM is sparse and it is often useful to consider a number of GLCM's, one for each relative position of interest. In order to obtain efficient descriptors, the information contained in GLCM is traditionally condensed in a few statistical features. Four GLCM's for four different orientations (horizontal 0°,vertical 90°, and two diagonals 45° and 135°) are obtained and normalized to the entries [0,1] by dividing each entry by total number of pixels. Haralick [64] has proposed a number of useful texture features that can be computed from the GLCM. Higher order features, such as energy, entropy, contrast, homogeneity and maximum probability are measured based on averaging features in GLCMs to form a 5-dimensional feature vector. Finally, color and texture feature vectors are normalized and combined to form a joint feature vector of 14-dimensions (9 for color and 5 for texture) for each sub-image and a 70-dimensional moment based feature vector $\mathbf{f}^{\mathrm{Moment}}$ for an entire image.

3.2.4 Average Grey Level Feature

Images in a collection may vary in size for different categories or within the same category and may undergo translations. Resizing them into a thumbnail of a fixed size can reduce the translational error and some of the noise due to the artifacts present in the images, specially for images in medical domain. This approach is extensively used in face or fingerprint recognition and has proven to be effective. We use a similar approach for feature extraction from low-resolution scaled images where each image is converted to a gray-level image (one channel only) and scaled down to the size 64×64 regardless of the original aspect ratio. Next, the down-scaled image is partitioned further with a 16×16 grid to form small blocks of (4×4) pixels. The average gray value of each block is measured and concatenated to form a 256-dimensional feature vector, $\mathbf{f}^{\mathrm{Grey}}$. An example of this approach is shown in Figure 12 where the left image is the original one, the middle image is the down-scaled version (64×64 pixels), and the right image shows the average gray values of each block (4×4 pixels). By measuring the average gray value of each block we can partially

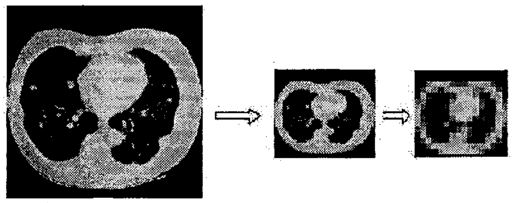


Figure 12: Feature extraction from a low-resolution image

cope with global or local image deformations and can add robustness with respect to translations and intensity changes.

3.3 Image Representation in Global Concept Space

This section presents our approach of converting the low-level feature vectors described above to an intermediate level global concept-based feature vectors based on probabilistic outputs of the multi-class SVMs and classifier combination strategies that are derived from Bayes's theory [173]. For training of the SVMs, the initial inputs are the feature vectors (e.g., EHD, CLD, Moment, and Avg. Grey) of sample images in which each vector is associated with a single category label selected out of all categories. Let, $\{C_1, \cdots, C_k, \cdots, C_M\}$ is a set of M global categories where each C_k characterizes the representative global semantic concept of a collection. In the testing stage, each image without a label is classified against the M categories. The output of the classification produces a ranking of the M category labels with probability or confidence scores. The confidence scores represent the weight of the category (concept) labels in the overall description of an image. The probability or confidence scores of the categories form an M-dimensional global concept label vector as follows

$$\mathbf{p}_j^m = [p_{1_j}^m \cdots p_{k_j}^m \cdots p_{M_j}^m]^{\mathrm{T}} \tag{7}$$

for a test image I_j with feature vector $\mathbf{f}_j^{\mathrm{m}}$ where $m \in F$ and $F = \{\mathrm{EHD}, \mathrm{CLD}, \mathrm{Moment}, \mathrm{Grey}\}$. Here, $p_{k_j}^m, 1 \leq k \leq M$, denotes the *posterior* probability that an image I_j belongs to category C_k in terms of input feature vector $\mathbf{f}_j^{\mathrm{m}}$. Finally, an image I_j belongs to a category $C_l, l \in \{1, \cdots, M\}$ is determined by

$$l = \arg\max_k [p_{k_j}^m] \tag{8}$$

that is the label of the category with the maximum probability score. In this context, given a feature vector $\mathbf{f}_j^{\mathrm{m}}$, the goal is to estimate

$$p_{k_j}^m = P(y = k \mid \mathbf{f}_j^{\mathrm{m}}), k = 1, \cdots, M \tag{9}$$

Although the voting procedure for multi-class classification based on *One against one* ensemble [176, 175] requires just pairwise decisions, it predicts only a class label. To generate the class prediction as probabilistic output, we use a probability estimation approach as described in [175]. Following the setting of the *one-against-one* approach in [175], the pairwise class probabilities r_{kl} are estimated as an approximation of μ_{kl} as

$$r_{kl} \approx P(y = k \mid y = k \text{ or } l, \mathbf{f}_j^{\mathrm{m}}) \approx \frac{1}{1 + e^{A\hat{f}+B}} \tag{10}$$

where A and B are the parameters estimated by minimizing the negative log-likelihood function, and $\hat{f}$ are the decision values of the training data based on v-fold cross-validation (CV) to form an unbiased training set. Finally, $p_{k_j}^m$ is obtained from all these r_{kl}'s by solving the following optimization problem based on the second approach in [175]:

$$\min_{\mathbf{p}_j^m} \frac{1}{2} \sum_{k=1}^{M} \sum_{l:l \neq k} (r_{lk}p_k - r_{kl}p_l)^2 \quad \text{subject to } \sum_{k=1}^{M} p_k = 1, p_k \geq 0, \forall k. \tag{11}$$

where $\mathbf{p}_j^m$ is the M-dimensional global concept vector as presented in (7), for image I_j based on feature vector $\mathbf{f}_j^{\mathrm{m}}$. The detailed implementation of the solution of (11) is given in [175].

3.3.1 Concept Feature Fusion

The feature descriptors, as described in Section 3.2, are in diversified forms and often complementary in nature. Since, the features represent image data from different viewpoints; the simultaneous use of different features might lead to a better classification result. A traditional method is to concatenate different feature vectors together into a single composite feature vector. However, it is rather unwise to concatenate them together since the dimension of a composite feature vector becomes much higher than any of individual feature vectors. Hence, multiple classifiers are needed to deal

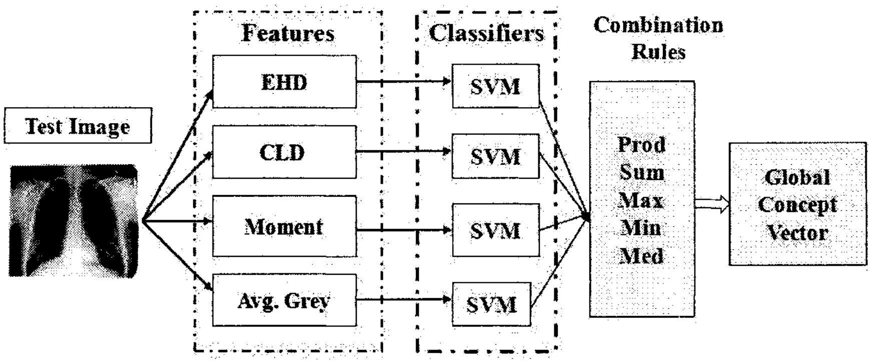


Figure 13: Process diagram of concept feature fusion with classifier combination.

with different features resulting in a general problem of combining those classifiers to yield improved performance. The combination of ensembles of classifiers has been studied intensively and evaluated on various image classification data sets involving the classification of digits, faces, photographs, etc. [178, 179, 180, 181]. In general, a classifier combination is defined as the instances of the classifiers with different structures trained on the distinct feature spaces [178, 179]. It has been realized that combination approaches can be more robust and more accurate than the systems using a single classifier alone.

In this framework, we consider combination strategies of the SVMs with different low-level features as inputs, based on five fundamental classifier combination rules derived from Bayes's theory [178]. These combination rules, namely product, sum, max, min, and median, and the relations among them have been theoretically analyzed in depth in [178]. These rules are simple to use but require that the classifiers output posterior probabilities of classification. This is exactly the kind of output the SVMs classifiers produce as described in Section 3.3.

Each classifier is trained with a particular input feature $m \in F$ and
$F = \{\text{EHD}, \text{CLD}, \text{Moment}, \text{Grey}\}$ from the training set and measures the posterior probability $p(C_k|\mathbf{f}_j^m)$ of a test image I_j belonging to class $C_k, k \in \{1, \cdots, M\}$ using feature vector $\mathbf{f}_j^m$. In these combination rules, a priori probabilities $P(C_k)$ are assumed to be equal. The decision rules for the product, sum, max, min and median are made by using the following formulas in terms of the *a posteriori* probabilities yielded by the respective classifiers for image I_j as

$$l = \arg\max_{k} \; [p^{\mathrm{r}}_{k_j}], \; r \in \{\mathrm{prod}, \mathrm{sum}, \mathrm{max}, \mathrm{min}, \mathrm{med}\} \tag{12}$$

where for the product rule

$$p^{\mathrm{prod}}_{k_j} = P^{-(|F|-1)}(C_k) \prod_{m \in F} p(C_k|\mathbf{f}^{\mathrm{m}}_j) \tag{13}$$

Similarly, for the sum, max, min and median rules

$$p^{\mathrm{sum}}_{k_j} = (1 - |F|)P(C_k) + \sum_{m \in F} p(C_k|\mathbf{f}^{\mathrm{m}}_j) \tag{14}$$

$$p^{\mathrm{max}}_{k_j} = (1 - |F|)P(C_k) + \max_{m \in F} p(C_k|\mathbf{f}^{\mathrm{m}}_j) \tag{15}$$

$$p^{\mathrm{min}}_{k_j} = P^{-(|F|-1)}(C_k) + \min_{m \in F} p(C_k|\mathbf{f}^{\mathrm{m}}_j) \tag{16}$$

and

$$p^{\mathrm{med}}_{k_j} = \frac{1}{|F|} \sum_{m \in F} p(C_k|\mathbf{f}^{\mathrm{m}}_j) \tag{17}$$

In the product rule, it is assumed that the representations used are conditionally statistically independent. In addition to the conditional independence assumption of the product rule, the sum rule assumes that the probability distribution will not deviate significantly from the a priori probabilities [178]. Classifier combination based on these two rules often performs better then the other rules, such as min, max and median [178, 178, 217]. The probabilistic outputs of SVMs with different test feature descriptors m are combined with the above rules based on equations (13)-(17) and finally represent a test image I_j as an M-dimensional global concept vector as

$$\mathbf{p}^{\mathrm{r}}_j = [p^{\mathrm{r}}_{1_j} \cdots p^{\mathrm{r}}_{k_j} \cdots p^{\mathrm{r}}_{M_j}]^{\mathrm{T}} \tag{18}$$

Here, the element $p^{\mathrm{r}}_{k}, 1 \leq k \leq M$ denotes the probability or membership score according to which an image I_j belongs to class C_k in terms of the combination rule $r \in \{\mathrm{prod}, \mathrm{sum}, \mathrm{max}, \mathrm{min}, \mathrm{med}\}$. Figure 13 shows the process diagram of the classifier combination and concept vector generation process.

The above concept vectors are computed off-line for database images (e.g., test images) based on the prediction of the SVMs and application of the combination rules.

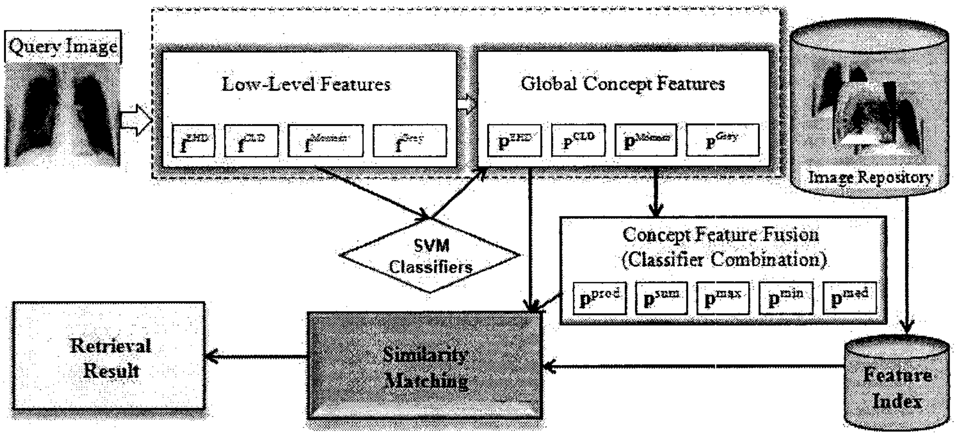


Figure 14: Block diagram of the image representation and retrieval in global concept spaces.

Similarly, the global concept vector of a query image I_q can be computed on-line as

$$\mathbf{p}_q^r = [p_{1_q}^r \cdots p_{k_q}^r \cdots p_{M_q}^r]^{\mathrm{T}} \tag{19}$$

Based on the global concept-based image representation, we can now compare a query and database image for rank-based retrieval. One common measure of similarity is the cosine of the angle between the query and document vectors in IR [44]. In many cases, the direction or angle of the vectors are a more reliable indication of the semantic similarities of the objects than the Euclidean distance between the objects in the term-document space. The proposed feature representation scheme closely resembles the document representation where a category or global concept label may be considered as analogous to a keyword in a document. Hence, we adopt the cosine similarity measure between feature vectors of query image I_q and database image I_j for a particular combination rule r as follows:

$$\mathrm{S}_r(I_q, I_j) = \frac{\sum_{k=1}^{M} p_{k_q}^r * p_{k_j}^r}{\sqrt{\sum_{k=1}^{M}(p_{k_q}^r)^2} * \sqrt{\sum_{k=1}^{M}(p_{k_j}^r)^2}} \tag{20}$$

This approach of semantic image categorization and representation may not attain high accuracy due the present state of the computer vision and pattern recognition technologies. However, the reliability of the global concept vector representations based on probability or membership scores is significantly better then the low-level image feature representations due to the exploitation of domain knowledge with supervised learning. The block diagram of image representation and similarity matching in the proposed feature spaces is shown in Figure 14 from a query image perspective. The features for database images are calculated off-line and stored in a logical database as feature indices for later on-line comparison with a query image feature.

In the following sections, we investigate an image retrieval approach from a different perspective where the image categorization information is utilized indirectly in a similarity matching function on a low-dimensional feature space. The major advantage of this approach is that due to the adaptive nature of the proposed similarity measure, it can effectively exploit category specific distribution information as well as captures correlations or variations between feature attributes in on-line matching.

3.4 Adaptive Statistical Similarity-Based Retrieval

In general, the majority of CBIR systems are similarity-based, where similarity between query and target images in a database is measured by some form of distance metrics in feature spaces [1]. These systems generally conduct the similarity matching on a high-dimensional feature space without any semantic interpretation or without paying enough attention about the underlying distribution of feature spaces [78, 79]. For example, the Euclidean distance is the most commonly used metric in CBIR whose effectiveness depends on the assumption of a sphere shape distribution of similar images around the query image point in feature space [80]. However, this assumption is not always true in reality. In addition, high-dimensional feature vectors not only increase the computational complexity in similarity matching but also increase the logical database size. Similarity measures based on empirical estimates of the distributions of features have also been proposed in CBIR [79]. However, the comparison is most often point-wise or statistics of the first order (i.e., mean vector) of the distribution is considered only with limited success [82]. To overcome the limitation and consider both first and second order statistics, several statistical distance measures are proposed based on multivariate Gaussian assumption of underlying probabilistic distribution of feature space [82, 78]. This assumption is a reasonable approximation for many image databases as distribution of feature vectors are sums of many random variables, where central limit theorem can be applied [172].

Statistical distance measure, defined as the distances between two probability distributions captures correlations or variations between attributes of the feature vectors [172, 173]. In this scheme, query image I_q and target image I_t in the database are assumed to be in different classes and their respective probability density functions (pdf) are $p_{\mathrm{q}}(\mathbf{f}_q)$ and $p_{\mathrm{t}}(\mathbf{f}_t)$. When these densities are multivariate normal, they can be approximated by mean vector $\boldsymbol{\mu}$ and covariance matrix Σ as $p_{\mathrm{q}}(\mathbf{f}_q) = N(\mathbf{f}_q; \boldsymbol{\mu}_{\mathrm{q}}, \Sigma_{\mathrm{q}})$ & $p_{\mathrm{t}}(\mathbf{f}_t) = N(\mathbf{f}_t; \boldsymbol{\mu}_{\mathrm{t}}, \Sigma_{\mathrm{t}})$ where,

$$N(\mathbf{f}; \boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{(2\pi)d|\Sigma|}} \exp^{-\frac{1}{2}(\mathbf{f}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{f}-\boldsymbol{\mu})} \tag{21}$$

here, $\mathbf{f} \in \Re^d$ and $|\cdot|$ is matrix determinant [172].

A popular measure of similarity between two Gaussian distributions is the Bhattacharyya distance [81], which is equivalent to an upper bound of the optimal Bayesian classification error probability. The Bhattacharyya distance between query image I_q and target image I_t in the database is given by [81, 172]:

$$D_{\text{Bhattacharyya}}(I_q, I_t) = \frac{1}{8}(\boldsymbol{\mu}_{\mathrm{q}} - \boldsymbol{\mu}_{\mathrm{t}})^T \left[\frac{(\Sigma_{\mathrm{q}} + \Sigma_{\mathrm{t}})}{2}\right]^{-1} (\boldsymbol{\mu}_{\mathrm{q}} - \boldsymbol{\mu}_{\mathrm{t}}) + \frac{1}{2} \ln \frac{\left|\frac{(\Sigma_{\mathrm{q}} + \Sigma_{\mathrm{t}})}{2}\right|}{\sqrt{|\Sigma_{\mathrm{q}}||\Sigma_{\mathrm{t}}|}} \tag{22}$$

where $\boldsymbol{\mu}_{\mathrm{q}}$ and $\boldsymbol{\mu}_{\mathrm{t}}$ are the mean vectors, and Σ_{q} and Σ_{t} are the covariance matrices of the feature distributions of I_q and I_t respectively. Equation (22) is composed of two terms, the first one being the distance between mean vectors of images, while the second term gives the class separability due to the difference between class covariance matrices. When all classes have the same covariance matrices, the Bhattacharyya distance reduces to the Mahalanobis distance, which is also a widely used similarity measure in CBIR [82].

$$D_{\text{Mahalanobis}}(I_q, I_t) = (\boldsymbol{\mu}_{\mathrm{q}} - \boldsymbol{\mu}_{\mathrm{t}})^T \Sigma^{-1} (\boldsymbol{\mu}_{\mathrm{q}} - \boldsymbol{\mu}_{\mathrm{t}}) \tag{23}$$

However, if inclusion of both query and target covariance matrices is useful, Bhattacharyya distance will outperform Mahalanobis distance. In general, the visual features such as texture or color are often defined at the output of a window operator or pixel-wise computations [78, 82]. From the ensemble of outputs, statistical information about the variation of the feature in an image are computed by a mean vector and a covariance matrix and can be utilized in the distance measures as described in equations (22) and (23).

For many frequently used visual features in CBIR, often their category specific distributions are also available in a semantically organized image database. In this case, feature descriptors may vary substantially from one category to another. Here, an image can be best characterized with its feature vector and by exploiting the information of feature distribution of its semantic category. Another observation is that most of the energy of a multivariate feature is often contained in a low dimensional subspace. Often the feature vectors belong to a subspace and the complexity of the retrieval process can be decreased if the effective dimension of the feature is taken into account [82]. To consider these properties into account, we propose an adaptive similarity matching scheme between features based on the utilization of the Bhattacharyya distance in (22). In this approach, training samples in the form of low-dimensional feature vectors of known categories are used to estimate the statistical parameters and train the SVMs. We assume that, features in each category follow a multivariate Gaussian distribution and based on this assumption, images are characterized with the first and second order statistical parameters. These category specific parameters are utilized by a statistical similarity matching function based on the on-line category prediction of SVMs. This work is motivated by the fact that the statistical similarity measures have not been investigated so far by exploiting category specific distribution information in semantically organized collections. Similarity measure based on the parameter estimation of category specific distribution would perform better if the right categories of query and database images are predicted in real time. Hence, we utilize on-line predictions of the SVMs, so that the proposed similarity measure function can be adjusted with category specific parameters for query and database images.

3.4.1 Feature Dimension Reduction

A high dimensional combined feature vector based on the color, edge and texture features would increase the computational complexity of the similarity matching as well as increase the logical database size. Often the feature vectors belong to a subspace where most of the energy of a multivariate feature is contained. Hence, we perform a dimension reduction by projecting a combined vector in a low-level feature space to a low-dimensional subspace based on principal component analysis (PCA) [183, 184]. PCA reduces the dimensionality of the feature to a basis set of prototypes that best describes the images. Each image feature vector is described by its projection on the basis set and similarity matching is performed on the projected feature vector of query and target images in the database. The basic idea of PCA is to find n linearly transformed components so that they explain the maximum amount of variances in the input data and mathematical steps used to describe the method is as follows.

Given a set of N feature vectors of training samples images, let the vector of image I_i in the training set be represented as $\mathbf{f}_i \in \Re^d | i = (1 \cdots N)$ in a combined feature space. The composite feature vector $\mathbf{f}_i$ is formed by simple concatenation of each individual feature vectors $\mathbf{f}_i^{\text{CLD}}$, $\mathbf{f}_i^{\text{EHD}}$, $\mathbf{f}_i^{\text{Moment}}$, and $\mathbf{f}_i^{\text{Grey}}$ where d is the sum of individual feature vector dimensions. The mean vector $(\boldsymbol{\mu})$ and covariance matrix (Σ) of the training samples are estimated as

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{f}_i \quad \& \quad \Sigma = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{f}_i - \boldsymbol{\mu})(\mathbf{f}_i - \boldsymbol{\mu})^T \tag{24}$$

Let $\boldsymbol{\nu}_i$ and λ_i be the eigenvectors and the eigenvalues of covariance matrix Σ, then they satisfy the following:

$$\lambda_i = \sum_{i=1}^{N}(\boldsymbol{\nu}_i^T(\mathbf{f}_i - \boldsymbol{\mu}))^2 \tag{25}$$

Here, $\sum_{i=1}^{N}\lambda_i$ accounts for the total variance of the original feature vectors set.

The PCA method tries to approximate the original feature space using an n dimensional feature vector, that is using n largest eigenvalues account for a large percentage of variance, where typically $n << \min(d, N)$. These n eigenvectors span a subspace, where $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_n]$ is the $d \times n$-dimensional matrix that contains orthogonal basis vectors of the feature space in its columns. The $n \times d$ transformation $\mathbf{S}^T$ transforms the original feature vector from $\Re^d \rightarrow \Re^n$ ones. That is

$$\mathbf{S}^T(\mathbf{f}_i - \boldsymbol{\mu}) = \hat{\mathbf{f}}_i, i = 1 \cdots N \tag{26}$$

where $\hat{\mathbf{f}}_i \in \Re^n$ and kth component of the $\hat{\mathbf{f}}_i$ vector is called the k-th principal component (PC) of the original feature vector $\mathbf{f}_i$. So, the feature vector in the original $\Re^d$ space for query and database images can be projected on to the $\Re^n$ space via the transformation of $\mathbf{S}^T$ [183]. After generating the reduced dimensional feature vectors for both training and database images, they will be utilized in consequent parameter estimation and similarity matching functions as described in the following sections.

3.4.2 Parameter Estimation and Similarity Matching

Computing the statistical distance measures between two multivariate Gaussian (normal) distributions requires first (mean) and second order (covariance matrix) statistical parameters of the distributions as depicted in (22). Let us consider, we have M different semantic categories in the database as $\{C_1, \cdots, C_i, \cdots, C_M\}$, where features in each C_i assumed to have a multivariate normal distribution. The true values of the parameters in each category are not known in advance and must be estimated from the set of training samples [172]. To estimate the parameters of each category specific distribution as well as to train the multi-class SVMs, a set of N images with M categories is selected as a training set. Now, for each category $C_i, i \in \{1, \cdots, M\}$, the mean ($\boldsymbol{\mu}_{C_i}$) and covariance matrix (Σ_{C_i}) are estimated by following the maximum likelihood estimation (MLE) [173] as

$$\boldsymbol{\mu}_{C_i} = \frac{1}{N_i} \sum_{j=1}^{N_i} \hat{\mathbf{f}}_{i,j} \tag{27}$$

$$\Sigma_{C_i} = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (\hat{\mathbf{f}}_{i,j} - \boldsymbol{\mu}_{C_i})(\hat{\mathbf{f}}_{i,j} - \boldsymbol{\mu}_{C_i})^T \tag{28}$$

where $\hat{\mathbf{f}}_{i,j}$ is a sample vector in PCA subspace of image I_j from category C_i, N_i is the number of training samples from category C_i and $N = (N_1 + N_2 + \ldots + N_\mathrm{M})$. These parameters are estimated off-line during the training of the SVMs and stored in a logical database. During similarity matching between a query and database image, the corresponding parameters will be accessed and utilized based on the online predictions of the SVMs for the query and target image categories. The detail steps required for the statistical similarity-based retrieval are given in Algorithm 1.

Hence, the main difference between the Bhattacharyya distance measure in (22) as

Algorithm 1 Adaptive Statistical Similarity Matching

```
(Initialize): Consider, there are $M$ different semantic categories in the database as $\{C_1, \cdots, C_i, \cdots, C_M\}$. Extract the feature vector $\hat{\mathbf{f}}$ of database images in a PCA-based subspace.
(Off-line): Select a set of $N$ training feature vectors of $M$ categories with associated category label to train the multi-class SVMs and calculate the parameters $\boldsymbol{\mu}_{C_i}$ and $\Sigma_{C_i}$ for $i \in \{1, \cdots, M\}$ using equation (27) and (28) and store these in a logical database.
(On-line): For on-line similarity matching of query image $I_q$ and database image $I_j$, get the category prediction of SVMs as $C_i(q)$ and $C_i(j)$, where $C_i(q), C_i(j) \in \{C_1, \cdots, C_M\}$.
Lookup the corresponding category-specific covariance matrices $\Sigma_{C_i(q)}$ and $\Sigma_{C_i(j)}$ from the logical database.
Utilize the covariance matrices in statistical distance measures as $D(I_q, I_j) = \frac{1}{8}(\hat{\mathbf{f}}_q - \hat{\mathbf{f}}_j)^T \left[\frac{(\Sigma_{C_i(q)}+\Sigma_{C_i(j)})}{2}\right]^{-1} (\hat{\mathbf{f}}_q - \hat{\mathbf{f}}_j) + \frac{1}{2}\ln \frac{\left|\frac{(\Sigma_{C_i(q)}+\Sigma_{C_i(j)})}{2}\right|}{\sqrt{|\Sigma_{C_i(q)}||\Sigma_{C_i(j)}|}}$
Convert the distance measure function into a similarity measures as $S(I_q, I_j) = \exp^{-D(I_q,I_j)/\sigma}$, where $\sigma^2$ is the distance variance computed separately over the training image set.
Finally return the images based on the similarity matching values in descending order.
```

commonly used for similarity matching of images in CBIR [82, 23] and the proposed one is the estimation of the statistical parameters and the dynamic nature of the matching function where parameters are adjusted on-line based on predictions of the SVMs. Instead of estimating parameters for each image and store them in a logical database, our approach requires only to estimate parameters from each category. The number of image categories are generally far less then number of images in a database. Thus, our approach decreases the logical database size by saving spaces. The adaptive nature of the overall similarity matching technique demonstrates its effectiveness empirically for category-specific searches, which will be shown in the experimental Section 8.3 of Chapter 8.

3.4.3 Query Parameters Updating by Relevance Feedback

A user might have a different interpretation of the semantic description in his/her mind or the prediction of the classifier might go wrong. Hence, it may be advantageous to have the option to interact with the system to refine the search process, which is commonly known as relevance feedback (RF) [108, 109]. It is an iterative and supervised learning process used to improve the performance of information retrieval systems [104]. A number of RF techniques have been proposed in CBIR [108, 110, 112, 111, 120] as described in Section 2.3.2 of Chapter 2. We now present a RF-based approach in this section where user selected positive or relevant images are used for calculating new query point and updating the statistical parameters of query image in each iteration.

Let, $N_{P(t)}$ be the number of positive feedbacks from a user which are selected from top retrieved k images compared to a query image I_q. $\hat{\mathbf{f}}_{j(t)} \in \Re^n$ is a feature vector in PCA subspace that represents j-th image for $j \in \{1, \cdots, N_{P(t)}\}$ at a particular feedback iteration t. The new query vector at the next iteration is estimated as

$$\hat{\mathbf{f}}_{q(t+1)} = \frac{1}{N_{P(t)}} \sum_{j=1}^{N_{P(t)}} \hat{\mathbf{f}}_{j(t)} \tag{29}$$

as the mean vector of positive images. The covariance matrix of the positive feature vectors is estimated as

$$\hat{\Sigma}_{q(t+1)} = \frac{1}{N_{P(t)} - 1} \sum_{j=1}^{N_{P(t)}} (\hat{\mathbf{f}}_{j(t)} - \hat{\mathbf{f}}_{q(t)})(\hat{\mathbf{f}}_{j(t)} - \hat{\mathbf{f}}_{q(t)})^T \tag{30}$$

Based on the above parameters, we use two different approaches of statistical similarity matching. In the first approach, for the updated query vector $\hat{\mathbf{f}}_{q(t+1)}$, the query image category $C_i(q(t+1)), i \in \{1, \cdots, M\}$ is determined based on the on-line prediction of the SVMs at feedback iteration $(t+1)$. After finding the query category, corresponding pre-determined category-specific covariance matrix $\Sigma_{C_i(q(t+1))}$ is used at iteration $t+1$ for the statistical distance measure between I_q and I_j as follows:

$$D_{\mathrm{RF1}}(I_q, I_j) = \frac{1}{8}(\hat{\mathbf{f}}_{q(t+1)} - \hat{\mathbf{f}}_j)^T \left[\frac{(\Sigma_{C_i(q(t+1))} + \Sigma_{C_i(j)})}{2}\right]^{-1} (\hat{\mathbf{f}}_{q(t+1)} - \hat{\mathbf{f}}_j) + \frac{1}{2} \ln \frac{\left|\frac{(\Sigma_{C_i(q(t+1))} + \Sigma_{C_i(j)})}{2}\right|}{\sqrt{|\Sigma_{C_i(q(t+1))}||\Sigma_{C_i(j)}|}} \tag{31}$$

In the second approach of similarity matching, instead of using the pre-determined covariance matrix from a logical database, we utilize the one estimated by the positive feedback images as $\hat{\Sigma}_{q(t+1)}$ by using (30). Hence, the adjusted distance matching

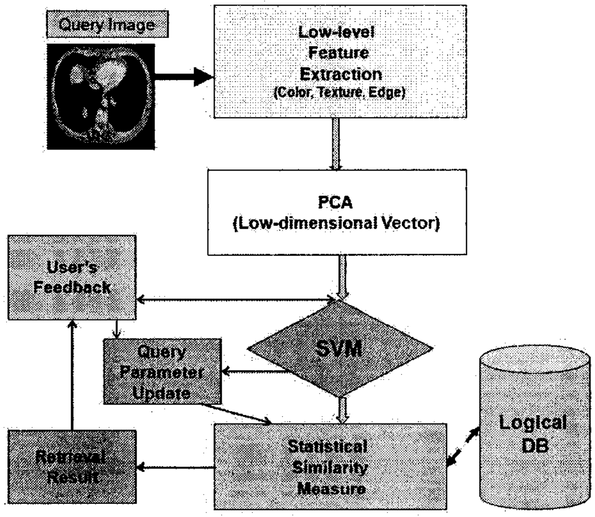

Query Image
Low-level Feature Extraction (Color, Texture, Edge)
PCA (Low-dimensional Vector)
User's Feedback
Query Parameter Update
SVM
Logical DB
Retrieval Result
Statistical Similarity Measure

Figure 15: Block diagram of the statistical similarity-based retrieval approach

function is:

$$D_{\mathrm{RF2}}(I_q; I_j) = \frac{1}{8}(\hat{\mathbf{f}}_{q(t+1)} - \hat{\mathbf{f}}_j)^T \left[\frac{(\hat{\Sigma}_{q(t+1)} + \Sigma_{C_i(j)})}{2}\right]^{-1} (\hat{\mathbf{f}}_{q(t+1)} - \hat{\mathbf{f}}_j) + \frac{1}{2} \ln \frac{\left|\frac{(\hat{\Sigma}_{q(t+1)} + \Sigma_{C_i(j)})}{2}\right|}{\sqrt{|\hat{\Sigma}_{q(t+1)}||\Sigma_{C_i(j)}|}} \tag{32}$$

Based on the above distance based similarity matching, system will present top k images to the user and he/she may continue to provide feedback till the system converges, i.e., no changes are noticed. The MindReader [111] retrieval system formulates a minimization problem on the parameter estimating process where the distance function is not necessarily aligned with the coordinate axis and allows for correlations between feature attributes. It was proved in [111] that, when using positive feedback (scores) and the Mahalanobis distance, the optimal query point is a weighted average based on available set of good results. In some way, our approach is related to the approach in [111] to update the parameters related to the query image based on user's positive feedback information. Though our RF method differs in the way in which query parameters are used in distance measures based on the equations (31) and (32). Figure 15 shows the block diagram of the proposed image retrieval approach based on adaptive statistical similarity matching and query parameters updating by user's positive feedback information from a query image perspective.

3.5 Summary

In this chapter, a novel image retrieval framework is presented based on global categorization of images in semantically organized databases. This is a first step to narrow the semantic gap in CBIR. The main drawback of the proposed framework is that it is suitable only when the categorization information at a global level is available in a particular image collection. This framework is not extensible when there is little or no domain knowledge available, such as images in the Web or any other broad domains. Another fact is that in majority case, queries are based on more complex concepts than the ones inferred from entire images at a global level. In following chapters, we will present image retrieval techniques by considering this fact.

Chapter 4

Image Representation in Local Visual and Semantic Concept Spaces

In this chapter, we propose a local concept-based image retrieval framework for general purpose (broad domain) to domain-specific (narrow domain) image collections [209]. In this framework, local visual and semantic concepts are modeled by employing both supervised classification and unsupervised clustering techniques [174, 173]. Based on the concept modeling, database images are segmented and encoded with visual and semantic concept labels. We propose effective feature extraction and representation methods that are robust against classification and quantization errors based on a soft image encoding scheme and effective to capture spatial ordering information of concepts in encoded images. Before investigating the proposed techniques, the central questions that are raised:

- What are the local *"visual"* and *"semantic"* concepts?
- How can we generate or model the concepts from images?

By the term *"local visual concept"*, we refer to perceptually distinguishable color and/or texture patches that are identified locally in image regions. Although things which are similar perceptually may not be similar semantically. For example, a predominant red color patch with smooth texture can describe either a red apple or a red car in images. Based on the color and texture feature, both apple and car are visually similar, semantically we know they are quite different. This is one of the major limitations in CBIR as we mentioned earlier in Chapter 2. The local visual concepts can be modeled by applying any statistical clustering techniques [174] to generate a codebook of concept prototypes. When images are represented with such visual concept prototypes based on their frequency of occurrences, it has proved to be effective and efficient in retrieval compared to low-level features calculated from pixel-level statistics [28, 209].

On the other hand, by the term *"local semantic concept"*, we refer to semantically distinguishable local patches/regions in individual images. For example, in an image collection of natural scenery, specific local patches such as, water, sand, grass, sky, snow, and so on can be identified with different semantic interpretations (i.e., what is the image about). However, being similar semantically does not always conform perceptual or visual similarity. For example, water and sky can have different sheds of color and texture patterns depending on the location and other contexts. Sea water might be perceptually different from river or lake water and color of the sky might be different at different times of the day. Although as a human being, we can easily distinguish such semantic concepts with many variations in different perspective due to our long accumulated knowledge, it is always difficult for a machine to do the same. Based on a supervised learning mechanism, a system might be able to model these concepts with limited variabilities and can detect them in unknown images in a closely similar way we identify objects based on our previous learning experiences.

Some recent approaches also investigate in this line to manually or automatically generate local concepts from image regions and finally represent or annotate images with the concept labels [28, 30, 31, 96, 97, 126]. In most cases, to model the high-level concepts require the use of formal tools such as supervised or unsupervised machine learning techniques. For example, a framework is proposed in [28] by applying Generalized Lloyd Algorithm (GLA) [34] and Learning Vector Quantization (LVQ) [193] based clustering techniques to generate codebooks of different visual concepts (termed as *"keyblock"*). In this framework, images are encoded with the codebook indices and finally they are represented as a generalization of vector space and n-gram models of IR [44]. The work in [31] manually identifies local semantic concepts (termed as *"visual keyword"*) as a set of pixel blocks where each set containing similar semantic concepts, such as face, crowd, building, and so on. Here, each image is indexed as a spatial tessellation of the semantic concept distributions to form a core signature of its visual content for image retrieval. Supervised learning based classification approaches are investigated in [96, 97, 94], where non-overlapping image regions are classified into local semantic concept categories. Many other statistical modeling based image classification and annotation approaches at local image level are investigated recently in CBIR. Eakins in [13] and Ying in [15] provide an overview of some of the techniques.

We observe some limitations in majority of these approaches to different extents. One of the main limitations is that during the image encoding process, a region is classified or matched to a single concept (object) only and rest of the concepts are overlooked or ignored [28, 30, 96, 94, 126, 127, 102]. Hence, the correspondence of an image region to a local concept is in general *one-to-one* due to the nature of hard classification or clustering schemes utilized by these approaches. In reality, there are usually several concepts with almost as closely match as the one detected for a particular image region. Considering only a single concept and ignoring others for image encoding and annotation processes might be problematic. For example, with a *one-to-one* matching approach, an encoded image can be represented as a set of labels where each region is associated with one concept label. However, this kind of representation is very sensitive to quantization or classification errors. Two regions will be considered totally different if they match to different concepts or objects even though they might be very similar or correlated to each other. Another drawback is that images are annotated and represented without considering the relative ordering of the associated concept labels in encoded images [30, 96]. However, the quantization errors and spatial relations of the local visual or semantic concepts are not negligible, which need to be exploited further for effective image annotation and representation purposes.

Motivated by this, we propose a local concept-based image representation framework to overcome the above limitations [209, 207]. In this framework, a visual concept vocabulary (codebook) is automatically constructed by utilizing Self-Organizing Map (SOM) [32] based clustering or vector quantization (VQ) and statistical models are built for local semantic concepts using probabilistic multi-class SVMs [165] (described in Chapter 3). The SOM has a topology preserving capability, which is a useful property lacking in basic clustering algorithms and the SVM has good generalization ability compared to other classifiers. These properties are helpful for the concept modeling and consequent feature generation processes as will be described in the following sections. The main functionalities of this framework are described as

- Selection of a training image set that closely represents the corresponding image collection.
- Segmentation of sample images from the training set to automatically and/or manually generate a set of image blocks as local regions.
- Generation of a codebook of local visual concept prototypes by unsupervised learning of SOM and modeling of local semantic concepts by supervised learning of multi-class SVMs based on input feature vectors of the image regions in the training set.
- Segmentation of database images into regions and encoding of individual images by classifying their corresponding region vectors with both soft (e.g., fuzzy and probabilistic) and hard or crisp concept labeling approaches.
- Represent images in local visual and semantic concept-based feature spaces by exploiting the soft annotation scheme, local neighborhood structure of the codebook, and relative ordering information of the concept labels in encoded images.

The block diagram of the retrieval framework is shown in Figure 16. As can be seen from the top portion of the figure, to model the local visual and semantic concepts, i.e., to generate codebook and SVM model file, learning of SOM and SVMs are conducted by using input region vectors of a set of training images. Based on the model generation, several feature vectors are constructed to represent images where the details of vector generation processes will be described in the following sections. The training and feature extraction of database images are conducted off-line, whereas the feature vectors generation of a query image (which is not present in a collection) and similarity matching are conducted on-line as shown in Figure 16.

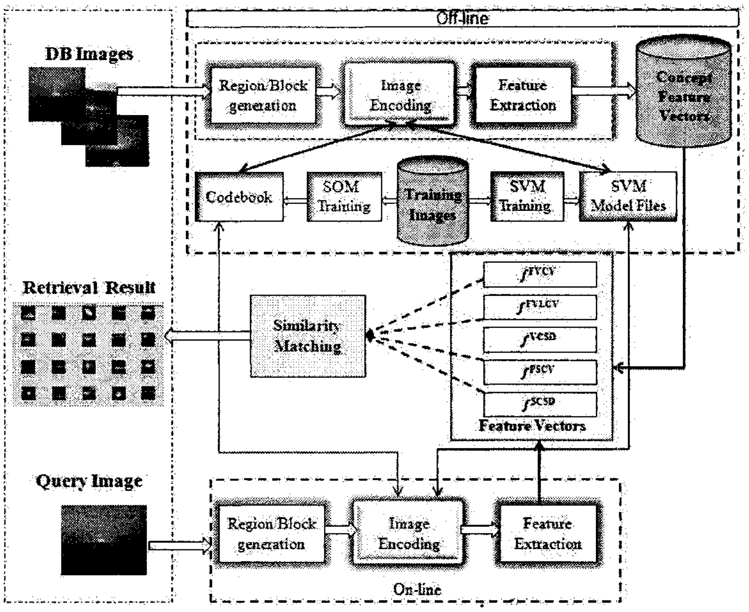


Figure 16: Block diagram of the concept-based image representation framework.

4.1 Local Visual Concept-Based Image Representation

This section presents our visual concept modeling and consequent image encoding and feature generation processes. The major issue of this approach is to construct an optimal codebook or visual concept vocabulary that will be representative enough of all possible visual concepts in a particular image collection. In this context, a codebook is defined as follows

Definition 1 *A codebook $C = \{c_1, \cdots, c_j, \cdots, c_N\}$ is a set of prototype visual concepts and N is its size. Each prototype visual concept c_j is associated with a label j and a vector $\mathbf{c}_j = [c_{j1} \;\cdots c_{j2} \cdots c_{jd}]^{\mathrm{T}}$ of dimension d in an Euclidean space.*

To generate the codebook based on SOM clustering, a reasonable training set of images needs to be selected either manually or in a random manner. Let $\mathcal{D}$ be an image database and a subset of this database $\hat{\mathcal{D}} = \{I_1, \cdots, I_j, \cdots, I_m\} \subset \mathcal{D}$ forms a training image set of m images where I_j is an image in the training set. Generally, for a narrow domain, we need to group images into different categories and pick few images from each category to make a reasonable training set. When a collection is very large in size without any category information (e.g., broad domain), the way of choosing a training set is to randomly pick a suitable percentage of the database images and use them as the training images.

After forming the training set, the next step is to segment the training images into regions and extract low-level image features from each region as a representative of initial visual concept vectors. Since, automatic segmentation schemes usually offer only an unreliable object description and regions are encoded with soft concept labels in our framework to make it robust against week segmentation, we use a fixed partitioning scheme. Let an image $I_j \in \hat{\mathcal{D}}$ be partitioned into a $(r \times r)$ grid of l blocks as segmented regions to generate region vectors as $\{\mathbf{x}_{1_j}, \cdots, \mathbf{x}_{k_j}, \cdots, \mathbf{x}_{l_j}\}$, where each $\mathbf{x}_{k_j} \in \Re^d$ is a vector of d-dimension in low-level feature space. In this work, to represent each region as a feature vector $\mathbf{x}_i$, the mean and standard deviation of each channel in the HSV (Hue, Saturation, and Value) color space as 6-dimensional color feature vector and second order moments (such as, energy, maximum probability, entropy, contrast, and inverse difference moment) as 5-dimensional texture feature vector are extracted from a grey level co-occurrence matrix (GLCM) [64]. Finally the color and texture vectors are combined as a single region vector after re-scaling the feature attributes with zero mean and unit variance [80].

There are in total m number of training images. So, finally the partition scheme will generate $n = (l \times m)$ region vectors for all the training images and collectively we can refer to them as a set of vectors $\mathbf{X} = \{\mathbf{x}_1, \cdots, , \mathbf{x}_i, \cdots \mathbf{x}_n\}$, where each $\mathbf{x}_i = [x_{i_1} \ x_{i_2} \cdots x_{i_d}]^{\mathrm{T}}$ is a vector of d-dimension. Since, features from blocks rather than individual pixels are used as vectors, some information on the spatial relationship among the neighboring pixels in the images are already retained. In general, there might be several similar regions in terms of image features in an individual image as well as in different images in the same training set. Since our visual system should tolerate some small errors, if the difference between two regions is below a certain preset threshold, they are deemed as the same. Hence, a subset of these representative vectors needs to be selected as a codebook of visual concept prototype by applying a clustering or VQ algorithm [174, 192].

4.1.1 Codebook Generation by SOM

SOM is a competitive learning-based clustering algorithm, which maps the high dimensional input vectors to a low-dimensional (usually two-dimensional) regular lattice or grid of map units and preserve the topological relationships between the vectors [32, 33]. In other words, similar input vectors are mapped to neighboring regions. Most clustering algorithms are based on iterative square error partitioning techniques [174, 192]. Square-error partitioning algorithms attempt to obtain the partition which minimizes the within-cluster scatter or maximizes the between-cluster scatter [174]. For example, the GLA [34], an iterative square error clustering algorithm, is used to generate codebooks in [28, 30]. On the other hand, even without exploiting its topology preserving properties, SOM has been shown to yield codebooks for image compression that are better than those generated by GLA [195]. It has been successfully applied to various problem domains such as data visualization, text retrieval, speech recognition, image compression, segmentation and so on [41, 42, 38, 195, 39, 197, 197]. Due to the topology preserving property, SOM is also successfully utilized in CBIR for indexing structure and browsing purposes [35, 36, 37]. In this framework, we also exploit this property from other perspectives as for feature enhancement process based on a local neighborhood structure and for an automatic query expansion process (will be presented in Chapter 5).

The basic structure of the SOM consists of two layers [32]: an input layer and a competitive output layer as a map. The input layer consists of a set of input node vectors $X = \{\mathbf{x}_1, \cdots \mathbf{x}_i, \cdots \mathbf{x}_n\}$, where each $\mathbf{x}_i = [x_{i_1}\ x_{i_2} \cdots x_{i_d}]^{\mathrm{T}}$ is a vector of d-dimension. The output map consists of a set of N units organized into either a one or two-dimensional lattice structure, where each unit m_j is associated with a weight vector $\mathbf{w}_j \in \Re^d$. The architecture of a SOM based on an input layer of n vectors and two-dimensional output map with $N = 25(5 \times 5)$ units is shown in Figure 17. The map attempts to represent all the available observations in X with optimal accuracy by using the map units as a restricted set of models. During the training phase, the set of input vectors is presented to the map multiple times and the weight vectors stored in the map units are modified to match the distribution and topological ordering of the feature vector space. The first step of the learning process is to initialize the weight vectors of the output map. Then, for each input vector $\mathbf{x}_i \in \Re^d$, the distances

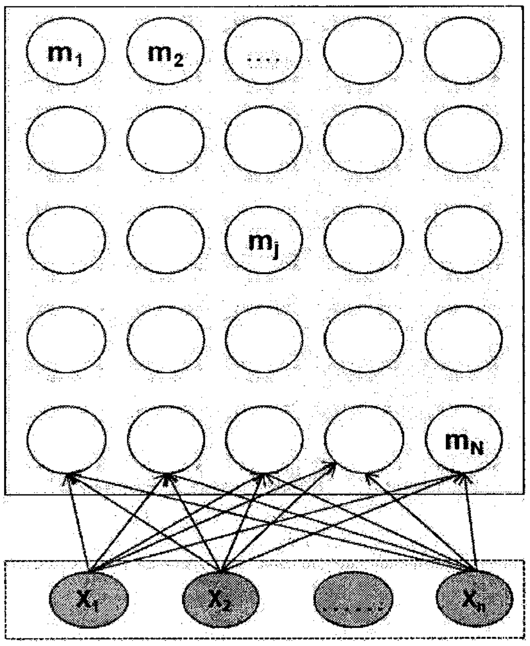


Figure 17: Structure of the SOM

between the $\mathbf{x}_i$ and weight vectors of all map units are calculated as

$$\|\mathbf{x}_i - \mathbf{w}_c\|^2 = \min_j \ \|\mathbf{x}_i - \mathbf{w}_j\|^2, \quad \text{for } j = 1, 2, \cdots, N \tag{33}$$

where $\|.\|^2$ is a distance measure in Euclidean norm. The unit, which has the smallest distance is called the best-matching unit (BMU) or winning node. The next step is to update the weight vectors associated with the BMU, m_c as

$$\mathbf{w}_j(t+1) = \mathbf{w}_j(t) + \alpha(t).\theta_{cj}(t)(\mathbf{x}_i(t) - \mathbf{w}_j(t)) \tag{34}$$

Here, t is the current iteration, $\mathbf{w}_j(t)$ and $\mathbf{x}_i(t)$ are the weight vector and target input vector at iteration t, whereas $\theta(t)$ and $\alpha(t)$ are the smooth neighborhood function and time-dependent learning rate. The $\alpha(t) = \exp^{-at} + b$, is shrunk in each training cycle (iteration),where a and b are the parameters which control the decay. The neighboring function $\theta_{cj}(t)$ is usually a time-decreasing Gaussian function of the coordinate or position vector $\mathbf{p}_c$ and $\mathbf{p}_j$ of these two map units c and j, given by

$$\theta_{cj} = \exp(-\frac{\|\mathbf{p}_j - \mathbf{p}_c\|^2}{2\sigma^2}) \tag{35}$$

where σ is the maximum neighborhood range. The major steps of codebook generation process are summarized in Algorithm 2.

Algorithm 2 Codebook Generation Process by SOM Clustering

```
Set the size of the two-dimensional output map as $N$.
for $j = 1, 2, \cdots, N$ do
    Initialize each weight vector $\mathbf{w}_j$ of the map units with random values.
end for
repeat
    Present the set $\mathbf{X}$ of input block vectors to the map in batch mode.
    for each $\mathbf{x}_i \in \mathbf{X}$ do
        Traverse each unit in the map and find the BMU, $m_c$ with associated weight
        vector $\mathbf{w}_c$ based on equation (33).
        Update the weight vectors in the neighborhood of $m_c$ by pulling them closer
        to the input vector based on equation (34).
    end for
until $\alpha(t)$ and $\theta_{cj}(t)$ are less then a threshold or the $\mathbf{w}_j$'s are not changed
```

Due to the process of self-organization, the initially chosen $\mathbf{w}_j$ gradually attain new values such that the output space acquires appropriate topological ordering. After the learning phase, the map can be acted as a codebook, where the map units represent the prototype visual concepts and their associated weight vectors represent the prototype concept vectors. Hence, a weight vector $\mathbf{w}_j$ of unit m_j resembles a visual concept vector $\mathbf{c}_j$ of concept c_j in the codebook C based on Definition 1. In general, the visual concept prototypes in the resulting codebook represent the most general structures extracted from all the training input vectors.

4.1.2 Image Encoding and Feature Representation

The codebook can be effectively utilized as a simple image compression as well as image representation scheme [197, 197, 28]. To encode an image with visual concept prototype labels or indices of the codebook, it is also decomposed into an even gird-based $(r \times r)$ partition, where similar low-level color and texture features are extracted from each region as is performed for training images. Let an image I_j be partitioned

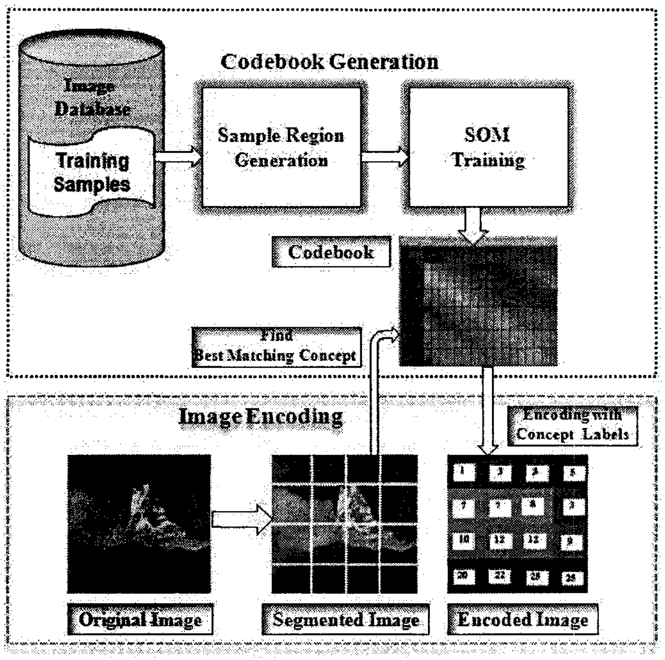


Figure 18: Codebook generation and Image encoding process

into $l = (r \times r)$ blocks or regions to generate vectors: $\{\mathbf{x}_{1_j}, \cdots, \mathbf{x}_{k_j}, \cdots, \mathbf{x}_{l_j}\}$; where each $\mathbf{x}_{k_j} \in \Re^d$. For each vector $\mathbf{x}_{k_j}$ in I_j , the codebook is searched to find the best match concept prototype (e.g., BMU in the map) $c_k, 1 \leq k \leq N$ as

$$c_k = \arg\ \min_{1 \leq l \leq N}\ \left\|\mathbf{x}_{k_j} - \mathbf{c}_l\right\|^2 \tag{36}$$

where k denotes the label of c_k and $\|.\|^2$ denotes the Euclidean distance metric between the region vectors of I_j and the concept prototype vectors.

After this encoding process, each image is represented as a two-dimensional grid of concept prototype labels where image blocks are linked to the corresponding best matching concept prototypes in the codebook. Figure 18 shows the codebook generation and image encoding processes. The codebook generation process is performed in the top portion of the figure and the bottom portion shows how an example image is encoded with the indices (e.g., prototype concept labels) of the codebook. Based on this encoding scheme, an image I_j can be represented as a histogram or feature vector

$$\mathbf{f}_j^{\text{V-concept}} = [f_{1_j}\ f_{2_j} \cdots f_{i_j} \cdots f_{N_j}]^{\text{T}} \tag{37}$$

where each dimension of the vector (e.g., bin of the histogram) corresponds to a prototype concept label of the codebook. The element f_{i_j} represents the normalized frequency of occurrences of visual concept label i of c_i appearing in encoded I_j. This feature representation captures only a coarse distribution of the visual concepts that is analogous to the distribution of quantized color in a global color histogram. As we already mentioned, image representation based on the above hard encoding scheme (e.g., to find only the best concept prototype for each region) is very sensitive to quantization error and lacks any spatial relationships information. Two regions in an encoded image will be considered totally different if their corresponding labels fall into two different bins even though they might be very similar or correlated to each other. In the following subsections, we propose image encoding and consequent feature representation schemes, which overcome these limitations.

4.1.3 Fuzzy Visual Concept-Based Feature Representation

This section presents a feature representation scheme based on the observation that there are usually several concept prototypes in the codebook with almost as good match as the best matching one for a particular image region. This scheme considers this fact by spreading each region's membership values through a global fuzzy membership function to all the concept prototypes in the codebook during the encoding and consequent feature extraction process.

The vector or histogram $\mathbf{f}^{\text{V-concept}}$ in (37) is viewed as a visual concept distribution from the probability viewpoint. Given a codebook of size N, each element f_{i_j} of the concept vector $\mathbf{f}_j^{\text{V-concept}}$ for an image I_j is calculated as $f_{i_j} = l_i/l$. It is the probability of a region in the image encoded with label i of visual concept $c_i \in C$, and l_i is the total number of regions that map to c_i. According to the total probability theory [172, 190], f_{i_j} can be defined as follows

$$f_{i_j} = \sum_{k_j=1}^{l} P_{i|k_j} P_k \; = \frac{1}{l} \sum_{k_j=1}^{l} P_{i|k_j} \tag{38}$$

where P_k is the probability of a region selected from image I_j being the k_jth region, which is $1/l$, and $P_{i|k_j}$ is the conditional probability of the selected k_jth region in I_j maps to the concept prototype c_i. In the context of visual concept vector $\mathbf{f}^{\text{V-concept}}$, the value of $P_{i|k_j}$ is 1 if the k_jth region is mapped to the c_i concept prototype or 0 otherwise. Due to the crisp membership value, this feature representation is sensitive to quantization errors as already mentioned.

In such a case, fuzzy set-theoretic techniques can be very useful to solve uncertainty problem in classification tasks [186, 187, 188]. This technique assigns an observation (input vector) to more than one class with different degrees instead of a definite class by crisp classification. In traditional two-state classifiers, an input vector $\mathbf{x}$ either belongs or does not belong to a given class A; thus, the characteristic function is expressed as

$$\mu_A(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in A \\ 0 & \text{otherwise.} \end{cases}$$

In a fuzzy context, the input vector $\mathbf{x}$, belonging to the universe X, may be assigned a characteristic function value or grade of membership value $\mu_A(\mathbf{x})$ $(0 \leq \mu_A(\mathbf{x}) \leq 1)$ which represents its degree of membership in the fuzzy set A [186].

Many methods could be adapted to generate membership from input observations. These include the histogram method, transformation of probability distributions to possibility distributions, and methods based on clustering [186, 187]. For example, fuzzy-c-means (FCM) [187] is a popular clustering method, which embeds the generation of fuzzy membership function while clustering. Few schemes have been proposed to generate fuzzy membership functions using SOM [199, 200], where the main idea is to augment the input feature vector with the class labeling information. However, without any class label information (as in our case), it might be difficult to generate such fuzzy membership functions. Due to this, we perform a two-step procedure, where in the first step we generate the proper clusters (e.g., concept prototypes in codebook) based on the SOM clustering and next the fuzzy membership values are generated according to the clusters (concept prototypes) in the first step as follows [188]:

Definition 2 *The membership degree μ_{ik_j} of a region vector $\mathbf{x}_{k_j} \in \Re^d, k = 1, 2, \cdots, l$, of the k_jth region in I_j to concept prototype vectors $\mathbf{c}_i, i = 1, 2, \cdots, N$ is:*

$$\mu_{ik_j} = \frac{{\frac{1}{\|\mathbf{x}_{k_j} - \mathbf{c}_i\|^2}}^{\frac{2}{m-1}}}{\sum_{n=1}^{N} {\frac{1}{\|\mathbf{x}_{k_j} - \mathbf{c}_n\|^2}}^{\frac{2}{m-1}}} \tag{39}$$

The higher the distance of an input vector from a concept prototype vector, the lower is its membership value to that concept prototype based on (39). It is to be noted that when the distance is zero, the membership value is one (maximum) and when the distance is infinite, the membership value is zero (minimum). The values of μ_{ik_j} lies in the interval $[0, 1]$. The fuzziness exponent $\frac{2}{m-1}$ controls the extent or spread of membership shared among the concept prototypes.

In this approach, during the image encoding process, the fuzzy membership values of each region to all concept prototypes are computed for an image I_j based on (39), instead of finding the best matching concept only. Based on the fuzzy membership values of each region in I_j, the *fuzzy visual concept vector* (FVCV) is represented as $\mathbf{f}_j^{\text{FVCV}} = [\hat{f}_{1_j}, \cdots, \hat{f}_{i_j}, \cdots \hat{f}_{N_j}]^{\text{T}}$, where

$$\hat{f}_{i_j} = \sum_{k=1}^{l} \mu_{ik_j} \; P_k = \frac{1}{l} \sum_{k=1}^{l} \mu_{ik_j}; \quad \text{for } i = 1, 2, \cdots, N \tag{40}$$

The proposed vector essentially modifies probability as follows. Instead of using the probability $P_{i|k_j}$, we consider each of the regions in an image being related to all the visual concept prototypes in the codebook via fuzzy-set membership function such that the degree of association of the k_j-th region in I_j to the c_i concept prototype is determined by distributing the membership degree of the μ_{ik_j} to the corresponding index of the vector. In contrast to the simple visual concept vector (V-concept), the proposed vector representation (FVCV) considers not only the similarity of different region vectors from different concept prototypes but also the dissimilarity of those region vectors mapped to the same concept prototype in the codebook. A similar approach to represent color histogram in a fuzzy space is proposed in [190]. However, our approach is computationally efficient due the less number of image regions/blocks and smaller codebook sizes as generated when compared to the number of image pixels and their colors. This representation is also in a higher level due to the exploitation of both color and texture features. In addition, there would be only few concepts prototypes as similar to the best matching one for a particular image regions. So, instead of considering the membership values based on global fuzzy membership function in (39), we can effectively estimate the values by considering only those few closely matched concepts by exploiting the topology preserving property of the codebook (e.g., output map). To address this issue, we present an efficient

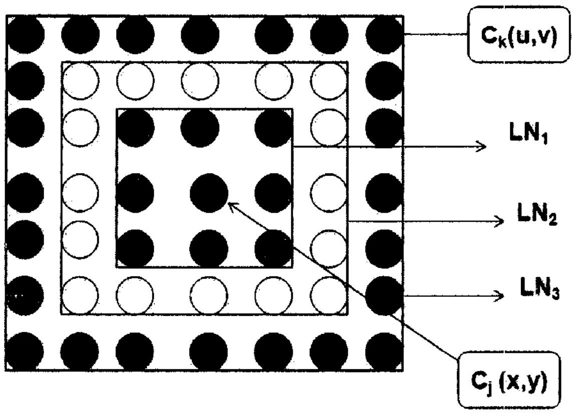


Figure 19: Topological local neighborhoods

method in the following section to compute membership values based on an adaptive local membership function and consequent feature representation.

4.1.4 Fuzzy Visual Local Concept Vector

Since, the codebook is generated by SOM where the trained map represents the codebook and map units represents the visual concept prototypes, we can exploit the topology preserving property to generate a local membership function. Based on the topology preserving structure of the SOM, the distance between two units on the map indicates the degree of similarity of the input vectors represented by the units. The basic idea is based on this fact that if two similar blocks map to two different concept prototypes in a codebook then the distance between the corresponding prototype vectors should be small and should be located close to each other in the codebook. In a similar aspect, we can say that there are usually several related visual concept prototypes in the codebook that are situated in a local neighborhood of the best matching one for a particular block. The local topological neighborhood in a two-dimensional codebook is defined as follows:

Definition 3 *Each visual concept prototype $c_j(x,y) \in C$ has a local γ-neighborhood $LN_\gamma(x,y)$ in a two-dimensional grid of codebook as depicted in Figure 19. We have*

$$LN_\gamma(x,y) = \{c_k(u,v) \,|\, \max\{|u-x|, |v-y|\} = \gamma\} \tag{41}$$

Here, the coordinate (x,y) and (u,v) denote the row and column wise position of any *two concept prototypes c_j and c_k respectively, where $x, u \in \{1, \cdots, M\}$ and $y, v \in \{1, \cdots, M\}$ for a codebook of size $N = (M \times M$* units*). The value of γ can be from 1 up to a maximum of $M - 1$.*

For example, Figure 19 shows the local neighborhood structure of a concept prototype in a two-dimensional codebook based on the Definition 3. Here, each concept prototype is visualized as a circle on the grid and the black circle in the middle denotes a particular concept prototype $c_j(x, y)$. Here, the concept prototype $c_k(u, v)$ is three (e.g., LN_3) neighborhood level apart from $c_j(x, y)$ based on the Definition 3 as the maximum distance between them (coordinate wise) either in horizontal or vertical direction is three. Basically, all the gray circles within the square are positioned in LN_1 neighborhood, the gray and yellow circles are positioned up to LN_2 and gray, yellow and blue circles in combine are positioned up to LN_3 neighborhoods of c_j as shown in the Figure 19. As the value of γ increases, the number of neighboring concept prototypes increases for c_j.

Based on the above neighborhood structure, we calculate the local fuzzy membership values in a set of visual concept prototypes S_γ, which contains the best matching visual concept prototype c_m for a particular input vector and all other closely related visual concepts prototypes located up to LN_γ-neighborhoods of c_m. Let us consider, there are n_γ concept prototypes located only in LN_γ, so the total number of concept prototypes of S_γ or the size be $|S_\gamma| = (1 + n_1 + \cdots + n_\gamma)$, e.g., $|S_1| = 9 = (1 + n_1) = (1 + 8)$ and $|S_2| = 25 = (1 + n_1 + n_2) = (1 + 8 + 16)$. For example, in Figure 19 n_1 counts all the grey color circles in LN_1 and n_2 counts all the yellow color circles in LN_2.

Now, the membership degree μ_{ik_j} of a k_jth region vector $\mathbf{x}_{k_j} \in \Re^d, k = 1, 2, \cdots, l$ in I_j to a concept prototype c_i is calculated as following

$$\mu_{ik_j} = \frac{\frac{1}{\|\mathbf{x}_{k_j} - \mathbf{c}_i\|^2}^{\frac{2}{m-1}}}{\sum_{n=1}^{|S_\gamma|} \frac{1}{\|\mathbf{x}_{k_j} - \mathbf{c}_n\|^2}^{\frac{2}{m-1}}.} \tag{42}$$

where $c_i, c_n \in S_\gamma$. Hence, the fuzzy membership values are calculated locally in a neighborhood of a best matching concept prototype for a particular input vector $\mathbf{x}_{k_j}$ during the image encoding process instead of considering the entire codebook. Due to the nature of the membership value computation from local neighborhood, the

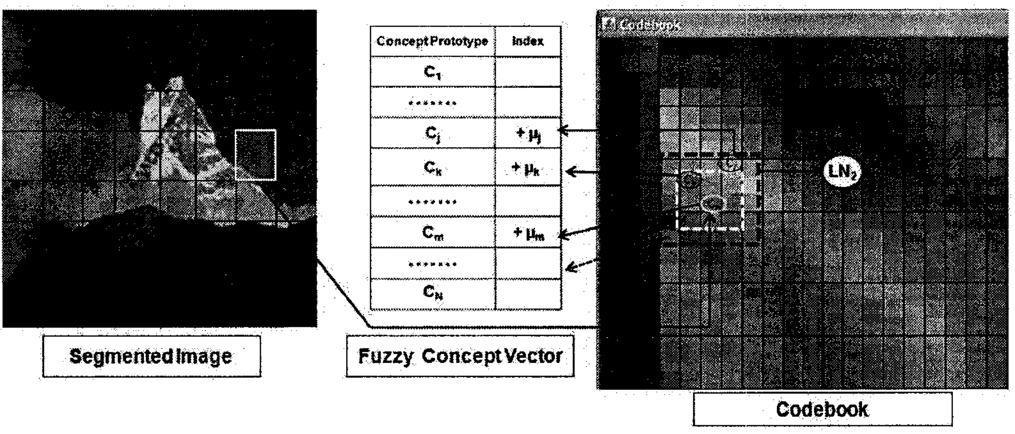


Figure 20: Vector Generation based on Local Neighborhood

proposed feature vector is termed here as *fuzzy visual local concept vector* (FVLCV). Summarizing, the steps involved in the feature vector computation are given in Algorithm 3.

Figure 20 shows a sample example as a process diagram of the above feature vector (histogram) generation approach. The left side of the figure shows a partitioned image with a particular block r_i mapped to the best matching concept $c_m \in C$. Based on S_2 (e.g., up to LN_2 neighborhood of c_m), the elements in the corresponding indices (such as labels m, j, k for concepts c_m, c_j and c_k) in the feature vector are incremented with fuzzy membership values by performing the above steps of the Algorithm 3. Due to the space limitation, we illustrate only three connections in the example instead of all twenty-five in $|S_2|$.

In order to calculate the global membership degree of a region vector, we need to compute the distances between the vector to all the visual concept prototype vectors with extra computation to determine fuzzy membership values. The distance computation is also essential to find the best matching concept prototype for local neighborhood determination or in case of image encoding with a single concept index for a frequency based image representation approach (e.g., V-concept). However, the local neighborhood based approach is comparatively efficient to the global one as only few computations are required to determine the membership values due to

Algorithm 3 Feature Generation based on Local Fuzzy Membership values

```
Initialize the *fuzzy visual local concept vector* of image $I_j$ as $\mathbf{f}_j^{\mathrm{FVLCV}} = [\bar{f}_{1j}\ \bar{f}_{2j} \cdots \bar{f}_{ij} \cdots \bar{f}_{Nj}]^{\mathrm{T}}$ with each $\bar{f}_{ij} = 0$.
Decompose $I_j$ into a $(r \times r)$ grid of $l$ blocks and generate the vectors as $\{\mathbf{x}_{1_j}, \cdots, \mathbf{x}_{k_j}, \cdots, \mathbf{x}_{l_j}\}$, where each $\mathbf{x}_{k_j} \in \Re^d$.
for $k = 1, 2, \cdots, l$ do
    Find the best matching concept prototype $c_m \in C$ for vector $\mathbf{x}_{k_j}$
    Consider up to local neighborhood $LN_\gamma$
    for each concept prototype $c_i \in S_\gamma$ do
        Calculate the membership values $\mu_{ik_j}$ based equation (42)
        Increase the value of the corresponding element $\bar{f}_{ij}$ at an index position $i$ of vector $\mathbf{f}_j^{\mathrm{FVLCV}}$ as $\bar{f}_{ij} += \mu_{ik_j}$.
    end for
end for
for $i = 1, 2, \cdots, N$ do
    Normalize each element of the vector as $\bar{f}_{ij} = \bar{f}_{ij}/l$.
end for
Finally, obtain the $\mathbf{f}_j^{\mathrm{FVLCV}}$ of image $I_j$.
```

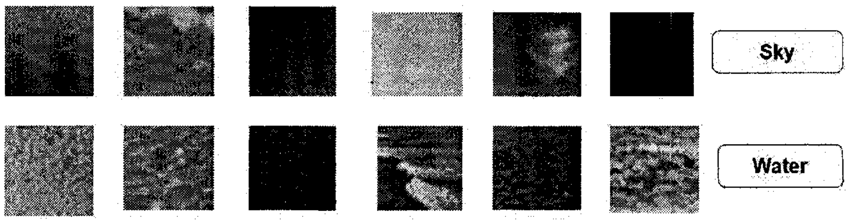


Figure 21: Intra class perceptual variability of concepts sky and water

the presence of much less number of visual concept prototypes in a neighborhood compared to considering all them in a codebook.

4.1.5 Probabilistic Local Semantic Concept-Based Image Representation

For some narrow domains, such as natural scenery or geographic image collections, various local semantic concepts at the region level are instantly available. For example, we can easily identify specific local patches, such as, water, sand, grass, sky, cloud, soil, snow, etc. that are semantically distinguishable from each other in these collections. Figure 21 shows several patches of sky (top row) and water (bottom row) that are cropped from a fixed partition-based image segmentation. Although the images here depict only two different concepts (e.g., sky and water), we can notice large variability in appearances in terms of color and texture features. This makes them distinguishable from *local visual concepts* and we called them as *local semantic concepts* in the beginning of this chapter.

In Section 4.1, we showed how images can be represented in visual concept-based feature spaces by concept modeling in the from of a codebook generation. Due to the unsupervised nature of the codebook generation process by SOM clustering [32], we could not incorporate enough domain specific knowledge of the collection in the feature representation schemes. Although for broad domain, an unsupervised learning is the only viable option available. When the domain knowledge is available, it would be effective to exploit the knowledge by utilizing any supervised learning-based techniques. A supervised learner can create a model to capture the variabilities of local semantic concepts with sufficient training samples. In this context, an instance (e.g., local patch) in the training set can be represented by a feature vector along with its local category specific labels. In this section, we present a local semantic concept-based feature representation scheme based on semantic modeling of local concepts by using the probabilistic multi-class SVMs [175] (described in Chapter 3). The proposed image representation scheme is closely related to the fuzzy visual concept -based image representation described in Section 4.1.3. The main difference is that instead of using fuzzy membership values, here the probabilistic membership values are utilized for image representation as a natural outcome of the class predictions of SVMs. This approach is effective due to the soft semantic labeling of image regions with category membership scores that reduces the effect of classification error in feature representation.

The main step of this approach is the training of multi-class SVMs for model generation of local semantic concepts. For this, we need to construct a training set of local semantic patches from individual image regions. Since, there exist large differences in visual appearances of patches from the same local concept category as shown in Figure 21, an effective training is essential to model such concepts. We consider a semi-automatic approach for the semantic patch generation. In this approach, training images are at first equally fixed partitioned into a $(r \times r)$ grid of non-overlapping regions as described previously in the case of visual concept modeling. Due to the fixed partitioning scheme, some regions or patches would contain multiple local concepts. Hence, instead of considering all the patches from fixed partitioning, a subset of these are manually selected based on the criteria that the area of each region should correspond to a particular concept by at least 70-80%. Although this will create a reasonable training set, problem will arise in later encoding or annotation process. What will happen when a region corresponds to two concepts where each of them covers half of the region? In case of a hard classification scheme, the region will be classified to either one of the concepts and totally ignore the other one. This will lead to quantization error in image encoding due to the obvious reason. However, instead of such a hard classification, it would be more effective if we can generate probabilistic output so that each region can be associated with the concept categories based on probability or confidence or membership scores. In that way, we will be able to generate probabilistic semantic concept vectors in lieu of fuzzy visual concept vectors as described in section 4.1.3. Only difference is that instead of using the membership values generated by global fuzzy function in case of fuzzy visual concept based representation, here the membership values are used from probabilistic outputs of the multi-class SVMs.

In order to perform the multi-class SVMs training based on the local concept categories, a set of L labels are assigned as $SC = \{sc_1, sc_2, \cdots, sc_L\}$, where each $sc_i \in SC$ characterizes a local concept category. A training set of local patches are annotated manually with the above concept labels in a mutually exclusive way. Hence, each patch is labeled with only one local concept category. The patches are represented as a combined vector of color and texture moment-based features as described in Section 4.1. For the SVMs training, the initial input to the system is the feature vector set of the patches along with their manually assigned corresponding concept labels. The details of the local concept extraction process and training of SVMs will be described in experimental Section 8.4 of Chapter 8 for a particular image collection.

To encode database images with local semantic concept labels, each image I_j is also partitioned into a $(r \times r)$ grid of l vectors as $\{\mathbf{x}_{1_j}, \cdots, \mathbf{x}_{k_j}, \cdots, \mathbf{x}_{l_j}\}$, where each $\mathbf{x}_{k_j} \in \Re^d$ is a d-dimensional combined color and texture feature vector in an Euclidean space. For each $\mathbf{x}_{k_j}$, the local concept category probabilities are determined by the

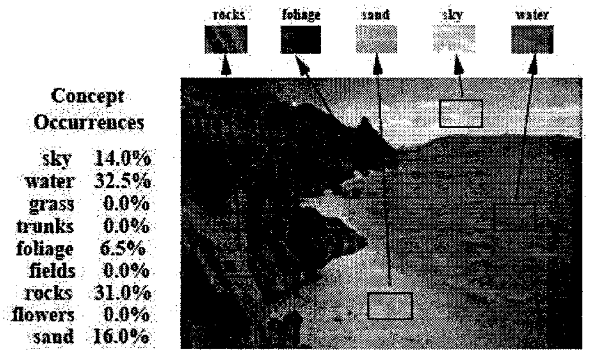


Figure 22: Representation of Images with Concept Occurrences [96].

prediction of the multi-class SVMs [175] as

$$p_{ik_j} = P(y = i \mid \mathbf{x}_{k_j}), \; i = 1, \cdots, L \tag{43}$$

for a L number of local semantic concept categories. The probability or confidence scores of all categories now form a L-dimensional vector for a region x_{k_j} of I_j as

$$\mathbf{p}_{k_j} = [p_{1k_j} \; p_{2k_j} \cdots p_{ik_j} \cdots p_{Lk_j}]^{\mathrm{T}} \tag{44}$$

Here each $p_{ik_j}, 1 \leq i \leq L$, denotes the probability that a region x_{k_j} belongs to category sc_i. Based on the probability scores, the category label of x_{k_j} is determined by

$$m = \arg\max_{1 \leq m \leq L} [p_{m_k}] \tag{45}$$

that is the label of the category sc_m with the maximum probability score. This way, the entire image is thus encoded as a two-dimensional indices linked to the local semantic concept labels assigned for each region. Based on this encoding, an image I_j can be represented as a vector in a local semantic concept space as

$$\mathbf{f}_j^{\mathrm{S-concept}} = \; < h_{1_j}, \cdots, h_{i_j}, \cdots h_{L_j} > \tag{46}$$

where each element h_{i_j} corresponds to the normalized frequency of a concept label $sc_i, 1 \leq i \leq L$ in image I_j. Similar to $\mathbf{f}^{\mathrm{V-concept}}$, the $\mathbf{f}^{\mathrm{S-concept}}$ can be viewed as a

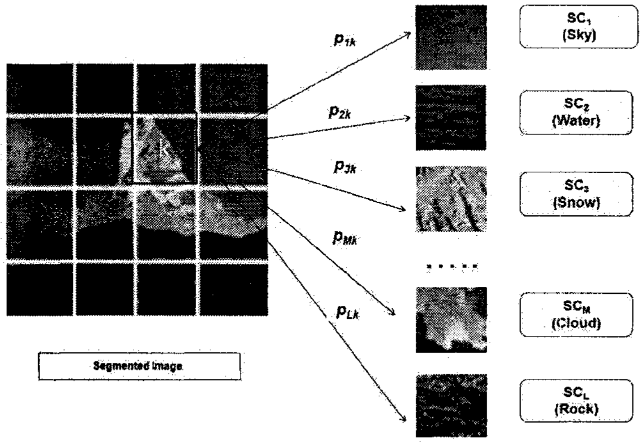


Figure 23: Image encoding with probabilistic membership scores

local semantic concept distribution from the probability viewpoint. Each element h_{i_j} of the concept vector $\mathbf{f}_j^{\text{S-concept}}$ is calculated as $h_{i_j} = n_i/l$. It is the probability of regions in the image belonging to the sc_i-th concept label, and n_i is the total number of regions that are categorized to $sc_i \in SC$. Each h_{i_j} is defined as follows:

$$h_{i_j} = \sum_{k=1}^{l} P_{i|k_j} P_k \; = \frac{1}{l} \sum_{k=1}^{l} P_{i|k_j} \tag{47}$$

where P_k is the probability of a region selected from image I_j being the k_jth region, which is $1/l$, and $P_{i|k_j}$ is the conditional probability of the selected k_j-th region belonging to the concept category of label sc_i. In the context of $\mathbf{f}^{\text{S-concept}}$, the value of $P_{i|k_j}$ is 1 if the k_jth region is categorized to sc_i-th concept category or 0 otherwise.

This representation is closely related to the $\mathbf{f}^{\text{V-concept}}$ in Section 4.1 and *concept occurrence vector (COV)* in [96], where only the frequency of occurrences of the semantic concepts is considered. For example, Figure 22 shows an example image from [96], where it presents the percentages of different concepts as normalized frequency of occurrences for that image. However, this feature representation approach has similar limitation as we mentioned for local visual concept vector (e.g., V-concept) based representation.

As we have already mentioned, a region might belong to more then one concept category or whose category membership might be very fuzzy in nature. Using a hard classification for such regions based on (45) would introduce only quantization error in image encoding process. Thereby it might be more effective, if we can utilize the probabilistic output as membership or confidence scores of each region in image encoding and consequent feature representation processes. Therefore, the proposed feature representation scheme distributes each region's class confidence values based on the semantic categories in SC to the corresponding vector indices. For example, Figure 23 shows a particular region in a segmented image and its probabilistic membership scores to different local semantic categories. Finally, the *probabilistic semantic concept vector* (PSCV) of an image I_j is represented as $\mathbf{f}_j^{\mathrm{PSCV}} = [\hat{h}_{1_j} \cdots \hat{h}_{i_j} \cdots \hat{h}_{L_j}]^{\mathrm{T}}$, where

$$\hat{h}_{i_j} = \sum_{k=1}^{l} p_{ik_j} P_k = \frac{1}{l} \sum_{k=1}^{l} p_{ik_j}; \quad \text{for } i = 1, 2, \cdots, L \tag{48}$$

where p_{ik_j} is determined by applying (43). Instead of using the probability $P_{i|k_j} = \{0, 1\}$, here we consider each of the regions in image being related to all the concept categories via probabilistic membership values. So that the degree of association of the k_j-th region of I_j to the sc_i-th concept category is determined by distributing the confidence scores p_{ik_j} to the i-th concept index in the vector. Due to the exploitation of domain knowledge in a supervised manner and encoding images with soft category membership scores, this representation scheme is robust compared to the representation based on only frequency of semantic concept occurrences (e.g., S-concept).

4.2 Visual and Semantic Structure Descriptors

One of the main drawbacks of the above visual and semantic concept based feature representation schemes is that no spatial relationship information between concepts are exploited. However, there might be instances, where spatial ordering information is required. As an example, if we partition the three images in Figure 24 as $r = 4 \times r = 4$ and encode it with a codebook, then they will generate identical visual concept vectors, although the images will be different perceptually based on the ordering of the concept labels. Due to this limitation, we present a spatial relationship-enhanced feature representation scheme on top of the visual and semantic concept-based representation. It captures both the concept frequencies and information about

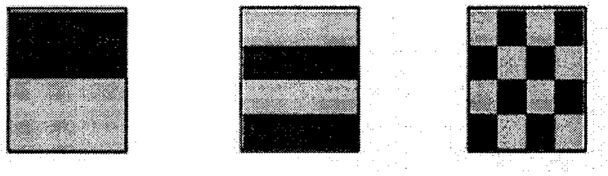


Figure 24: Images with different color structures

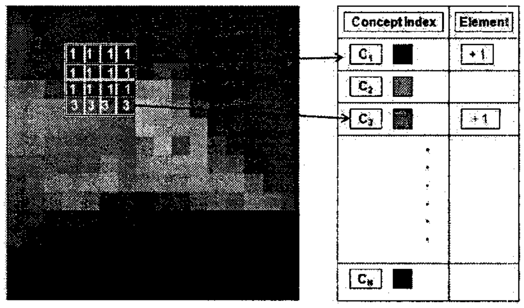


Figure 25: Example of Visual Concept Structure Descriptor (VCSD) generation

the local spatial relationships of the concepts. Specifically, it is a histogram where each bin counts the number of times a visual or semantic concept prototype label is present in a windowed neighborhood determined by a small square structuring element. The size of the structuring element is $(b \times b)$ units where each block represents a unit and $(b < r)$. For example, the left side of Figure 25 shows an encoded image with concept prototype labels which is partitioned into $(r = 16 \times r = 16) = 256$ blocks. If we consider each block as a unit, then the encoded image has $m = 16$ rows and $n = 16$ columns. When the structuring element progresses over the rows and columns of the encoded image, it enables to distinguish, for example, between an image in which the local concepts labels are distributed uniformly and an image in which the concept labels are occurred in same proportions, but are located in distinct block units. The feature extraction method embeds local concept structure information into the feature vector by taking into account all the concept labels in the structuring element that slides over the encoded image, instead of considering concept label of each block separately.

The accumulation process is illustrated in Figure 25 for an encoded image with visual concept labels. For each unique concept label that falls inside the structuring element at a particular position in encoded image, the corresponding histogram bin or vector index is only incremented once. As shown in Figure 25, in this case of two different labels of visual concepts c_1 and c_2 inside the structuring element, the corresponding histogram bins or vector indices are incremented once.

The proposed technique is closely related to the MPEG-7 color structure descriptor (CSD) [62] where instead of considering image block as a unit it considers each pixel unit and quantization is performed only in a color space. As already mentioned, color based histogram has many limitations and pixel level statistics always require time consuming calculation. Unlike the original CSD [62], we also need not require to measure the spatial extent of the structuring element since the images are equally partitioned into blocks and after encoding, each image contains the same number of blocks in total. As the basic mechanism is closely related, we call the proposed representation as *visual concept structure descriptor* (VCSD). Hence, an encoded image I_j is represented as a vector as $\mathbf{f}_j^{\mathrm{VCSD}} = [f_{1j}^v \cdots f_{ij}^v \cdots f_{Nj}^v]^{\mathrm{T}}$, where each dimension corresponds to a visual concept label. The element f_{ij}^v represents the number of structuring elements in the image containing one or more region with the visual concept c_i and is normalized by the number of locations of the structuring element, which lie in the range $[0, 1]$. The origin of the structure element is defined by its top-left sample and the locations of the structure element over which the elements of the vector are accumulated are defined by the position of the block (e.g., the smallest unit in the encoded image) that contains the visual concept index.

In a similar fashion, for semantic concept based representation, it is termed as *semantic concept structure descriptor* (SCSD), where an image is represented as $\mathbf{f}_j^{\mathrm{SCSD}} = [f_{1j}^s \cdots f_{ij}^s \cdots f_{Nj}^s]^{\mathrm{T}}$. Here, each dimension of the vector corresponds to a local semantic concept category label or index. The element f_{ij}^s represents the the number of structuring elements in the image containing one or more region with the local semantic category sc_i and is normalized. Hence, both VCSD and SCSD express local concept structure by visiting all the location in encoded images with the structuring element. The size of the structuring element may vary between visual and semantic concept-based representation due to the variations in number of regions as generated by image partitioning scheme to encode an image.

4.3 Summary

In this chapter, we have explored some alternatives for improving accuracy of image retrieval by representing images in various intermediate local visual and semantic concept levels. In summary, the proposed image representation schemes realize semantic abstraction via prior learning when compared to representation based on low-level features. Experimental results (will be described in Chapter 8) validate this assumption and show that the proposed representation schemes improve retrieval accuracy. Hence, we feel that the computational resources devoted to prior learning and index generation are good trade-off for effective retrieval performance. The proposed approaches are applicable in different domains where visual concept based representations are suitable for both broad and narrow domains and semantic concept-based representation schemes are more suitable for narrow domain as domain knowledge needs to be exploited for effective semantic modeling based on supervised learning.

Chapter 5

Query Expansion Based on Local and Global Analysis

There is a strong similarity between the keyword-based representation of documents in vector space model (VSM) of IR [44, 45] and the local visual and semantic concept-based image representations described in Chapter 4. This model relies on the premise that the meaning of a document (image) can be derived from its constituent keywords (concepts). Depending on the context, a keyword has an amount of information in individual documents as well as in a collection as a whole. For example, in an information retrieval archive about *"computer science"*, the word *"computer"* does not add much information since it is is very likely to appear in many documents in the collection [44]. In our image domain, the visual and semantic concepts also contain information from both perceptual and semantical perspectives. Some concepts are very likely to appear in all images in a collection, whereas other concepts occur often in few images but relatively rare in the entire collection of images. For example, a semantic concept *"sky"* would occur in many outdoor images, whereas a concept *"snow"* might be found only images of a mountain category in an image collection of natural scenery.

The rationale is that concepts that occur frequently in the entire collection have low information content. However, if a single image contains many occurrences of a concept then it is probably valuable to distinguish that image from other images in a collection. To consider this effect into account, several term weighting schemes have been developed in IR, in which the *term-frequency - inverse document frequency* (tf-idf) weighting is the most commonly used one [44, 45]. This weighting scheme is based on a combination of both global and local weights. A global weight indicates the overall importance of a keyword across the entire collection, whereas a local weight of a keyword indicates its importance in individual documents. Generalizing upon the VSM in our image retrieval domain, a concept also can be given a weight, depending on its information content. In this context, for the *tf-idf* based scheme, the local weight is denoted as $L_{i,j} = \frac{f_{i,j}}{\max_l f_{l,j}}$, where $f_{i,j}$ be the raw frequency of occurrence of a local visual concept c_i or a local semantic concept sc_i in image I_j and the maximum is computed over all the concepts which are presented in I_j. The global weight G_i is denoted as inverse image frequency (e.g., inverse document frequency in text) as $G_i = log(M/M_i) + 1, i = (1, \cdots, , N)$, where M_i be the number of images in which concept c_i (sc_i) is found [44, 45]. Finally, an element w_{ij} is expressed as the product of local and global weight as $w_{ij} = L_{i,j} * G_i$. Based on the above weighting scheme, an image I_j can be represented as a vector in a concept space as

$$\mathbf{f}_j^{\mathrm{VM}} = [w_{1j} \; w_{2j} \cdots w_{ij} \cdots w_{Nj}]^{\mathrm{T}} \tag{49}$$

where w_{ij} be the weight of concept c_i (sc_i). In case of visual concept-based representation, the vector dimension N equals the size of the codebook and for semantic concept-based representation, it is equal to the number of local semantic categories. In order to distinguish them in the following sections, the vectors will be referred as $\mathbf{f}^{\mathrm{V-VM}}$ and $\mathbf{f}^{\mathrm{S-VM}}$ for visual and semantic concepts respectively.

In the above representation scheme, images are modeled using a *bag-of-concepts* (e.g., *bag-of-words* in text) approach [44, 45]. Besides the loss of all ordering structure, each concept is considered independent of all the other concepts in this model. Although this simple assumption has proved to be effective for keyword-based representation in text retrieval domain, it is found that the keywords are related by use, for example in phrases, and their similarity of occurrence in documents can reflect underlying semantic relations between them. In reality, there are often strong patterns of dependence between keywords used to describe similar topics in documents. For example, in a text retrieval system about tourist information, the occurrence of a word *accommodation* in a text will give the word *hotel* a higher probability of co-occurring than any random word. Another problem with this model is that the words used in a query are often not the same as the words used to represent relevant documents, which is commonly termed as *word mismatch* problem [44, 48, 49]. Similar words can have multiple meanings (polysemy) and different words can have the same meaning (synonymy). For example the words *car* and *automobile* are synonyms where the polysemous word *jaguar* can either mean a large *cat* or can denote a model of *car* [119].

Similarly, in image domain, the same concept might represent different things and different concepts might depict the same thing depending on the context. For example, in an outdoor image under broad day light, an yellow color concept probably depicts a sun, whereas in an image of a garden the same color may represent a flower. Here, the concept of *yellow color* has different meaning (polysemy) in different context. In another example, different shades of water color with varying texture as different visual concepts might actually represent the same thing (synonymy), i.e., *water*, in a visual concept-based image representation. Similar to text retrieval, we can usually find several correlated or co-occurring concepts for a particular one. For example, in the outdoor image in the previous example, there is a higher probability of occurrence of a blue sky around the sun. Whereas, a yellow color flower has more probability to co-occur with green leaves in the garden image. Hence, there is indeed a need to expand the visual and semantic concepts in the original query image by exploiting the information on correlation or co-occurrence patterns to improve retrieval effectiveness.

To increase retrieval effectiveness and reduce ambiguity due to the the word mismatch problem in IR, a variety of query modification and reformulation strategies have been investigated during the last four decades [104, 46, 48, 105, 44]. There are mainly three approaches of query reformulation [44]:

1. Interactive approaches based on feedback information from the user as commonly known as *relevance feedback* (RF);

2. Automatic approaches based on global information derived from the entire collection or corpus as commonly known as *thesaurus based query expansion*, and

3. Automatic (might be interactive also in some cases) approaches based on local information from top retrieved results, which is commonly known as *local feedback.*

Among these approaches, RF is the most popular query reformulation strategy which prompts the user for a feedback on the retrieval results and then use that information on subsequent iterations with a goal of increasing retrieval performance [44, 45, 105]. Although the RF approach initially developed for text retrieval domain [45], the ease with which the relevance of an image can be evaluated accelerates the development of it into CBIR systems [108, 109] (reviewed in Section 2.3.2 of Chapter 2). We also presented a RF-based similarity matching approach in Section 3.4.3 of Chapter 3. Although RF-based approaches have been shown to provide dramatic performance improvement in retrieval systems [108, 109, 110, 111, 112, 113], one of the major drawbacks is that the users are not always able or motivated to provide feedback information to make the methods perform effectively. It has been often proved to be a complex mechanism with different levels of feedback (e.g., relevance levels in MARS [110], goodness scores in Mindreader [111]) and a time consuming aspect for users.

The automatic query reformulation based on term co-occurrence or term similarity has been investigated for more then four decades in text retrieval domain with varying successes [46, 48, 49, 47, 50, 52]. These approaches of query refinement exploit term (keyword) dependency as term clustering in document collection, e.g., grouping sets of related terms with a view to select query expansion terms from these sets [46, 48, 105]. Approaches of these kinds have a significant advantage over interactive relevance feedback as they require no effort on the part of the user. The techniques that have so far been investigated can be described as being based on either global or local analysis [46, 48]. Both local and global analysis are highly dependent on clustering algorithms based on the term-term correlations in feature space. In a global analysis, all documents in the collection are analyzed to determine a global thesaurus-like structure which defines term relationships. This structure is then utilized to select additional terms for query expansion. In local analysis, the top retrieved documents for a query are examined (without any assistant from the user in general) at query time to determine the terms for query expansion. There is also an underlying notion of clustering that supports the RF strategy [44]. In this particularity, known relevant documents contain terms that can be used to describe a larger cluster of relevant documents with user interaction. However, the global and local analysis techniques attempt to obtain a description for a larger cluster of relevant documents automatically.

Due to the nature of the low-level continuous feature representation in majority of the CBIR systems [2, 5], the automatic query reformulation techniques based on term co-occurrence or term similarity have not been investigated yet in CBIR domain. In a continuous feature space, each feature attribute has a value on a continuous scale. Hence, there is no notion about whether an attribute (term) is occurred in an image feature or not. Since, the local concept-based feature representations in VSM is closely related to the keyword-based representation of documents in text retrieval domain, we explore analogous query expansion techniques in image domain from a new perspective to investigate whether they can improve retrieval effectiveness when compared to search without using any query expansion. Hence, inspired by ideas from text domain, we propose automatic query expansion approaches for image domain based on both local and global analysis. For automatic query expansion based on local analysis, we exploit the concept-concept correlations by analyzing a local clustering method which takes into account metrical constraints based on neighborhood proximity between concepts in encoded images. For global analysis, we construct a global structure or thesauruses in the form of a similarity matrix, which consider the similarities between visual concept prototypes in a codebook. Finally, we also propose an efficient query expansion and similarity matching technique by combining both local and global analysis in a single process.

5.1 Query Expansion Based on Local Analysis

Query expansion based on local feedback and cluster analysis (commonly known as local analysis) has been found to be one of the most effective methods for expanding queries in text retrieval domain [46, 48]. Generally, the techniques based on local analysis expand a query based on the information from the top retrieved items (without any assistant from the user) for that query. In most expansion methods making use of local analysis, there are five key stages. First, the original query is used to rank an initial set of documents. This set is then retrieved from disk and all terms are extracted from those documents. Correlated terms are identified and ranked in order of their potential contribution to the query. The top ranked terms are re-weighted and appended to the query, and finally the reformulated query is reissued and a final set of documents is ranked [48, 46]. Techniques based on local feedback analysis are interesting because they take advantages of the local context provided with each query, hence it aims at optimizing the current search. Before presenting our query expansion method in image domain, some basic terminologies are defined as follows:

Definition 4 *For a given query image I_q, the set S_l of images retrieved is called the local image set. Further, the set $C_l \subseteq C$ of all distinct concepts $c_i \in C_l$ in the local image set S_l is called the local vocabulary of concepts. Here, for simplicity c_i represents either a visual concept prototype or a local semantic concept category (e.g., sc_i) depending on the feature representation scheme.*

Since, correlated terms for expansion are those present in the local cluster, we first need to generate such a cluster from S_l and thereby from C_l. To generate the cluster, we rely on a local *correlation matrix* that is built based on the co-occurrence of concepts inside images. The matrix is defined as follows:

Definition 5 *Let, $\mathbf{A}_{l|C_l| \times |C_l|} = [a_{uv}]$ be a local concept-concept correlation matrix in which the rows and columns are associated with the concepts in C_l. Each element a_{uv} expresses a normalized correlation factor between concepts c_u and c_v as*

$$a_{uv} = \frac{n_{uv}}{n_u + n_v - n_{uv}} \tag{50}$$

where n_u be the number of images in S_l which contain concept c_u, n_v be the number of images which contain concept c_v, and n_{uv} be the number of top retrieved images in S_l which contain both concepts. The a_{uv} measures the ratio between the number of images where both c_u and c_v appear and the total number of images in S_l where either c_u or c_v appear and its value ranges to $0 \leq a_{uv} \leq 1$. If c_u and c_v have many co-occurrences in images, then the value of a_{uv} increases, and the images are considered to be more correlated.

The global version of this matrix, which is termed as a *connection matrix*, is successfully utilized in a fuzzy information retrieval approach in [53]. The local correlation (connection) matrix $\mathbf{A}_l$ is created based on the frequency of co-occurrence of pairs of concepts in images and does not take into account where the concepts are located in an encoded image. Since, two concepts which occur relatively close in a neighborhood seem more correlated than the two concepts which occur far apart in an image. We provided such an example in the previous section for the concepts of

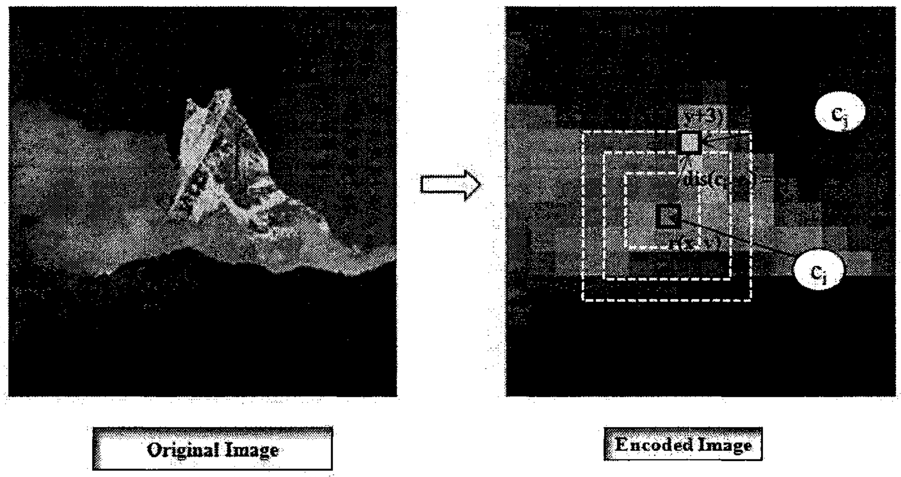


Figure 26: Distance between concepts c_i and c_j in an encoded image based on neighborhood relationship

sun and sky for outdoor images. Hence, it would be worthwhile to factor the distance between two concepts in the computation of their correlation factor.

To resolve this issue, metric cluster was introduced in text retrieval domain a long time ago in [46]. In this cluster generation process, correlation factors are computed by considering inverse of a distance between two terms. A document can be viewed as a one dimensional ordered list of terms where the distance between two terms are calculated by counting the number of words between them in a document. For our image domain, concepts are organized in a two-dimensional grid in encoded images. To measure the distance between two concepts, we need to rely on some sort of neighboring relationship to count how many block units (maximally) they are apart by considering both horizontal and vertical directions. Such a neighborhood relationship is defined as follows:

Definition 6 *Let, a two dimensional encoded image I_j is represented as a grid of P columns and Q rows of block units. Each block $r(x, y)$, in a co-ordinate position $x, y, |0 \leq x \leq P - 1, 0 \leq y \leq Q - 1$, has a topological d-neighborhood and is mapped to a concept $c_i \in C_l$ of the local codebook as shown in Figure 26. The value of d can go up to a maximum of $P - 1$ or $Q - 1$ for horizontal or vertical directions. If we consider two blocks $r(x, y)$ and $r(u, v)$ for $0 \leq x, u \leq P - 1, 0 \leq y, v \leq Q - 1$ are*

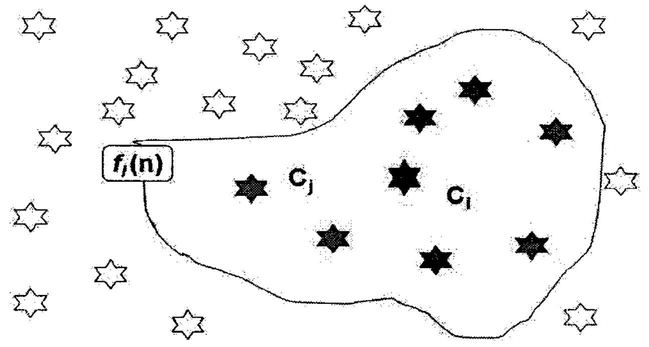


Figure 27: Concept c_j as a neighbor of the concept c_i based on a local cluster

mapped to concepts c_i and c_j respectively. The distance $dis(c_i, c_j)$ between these two concepts is measured as

$$dis(c_i, c_j) = dis(r(x, y), r(u, v)) = \max\{|u - x|, |v - y|\} = d \tag{51}$$

Hence, if c_i and c_j are 1 neighborhood apart then $dis(c_i, c_j) = 1$ and if c_i and c_j are in distinct images then, we consider $dis(c_i, c_j) = \infty$.

For example, Figure 26 shows the distance between concepts c_i and c_j as $dis(c_i, c_j) = 3$ based on the definition above. Now, a local *concept-concept* metric correlation matrix is defined as follows [46]:

Definition 7 $\mathbf{M}_{l_{|C_l| \times |C_l|}} = [m_{uv}]$ *be a local concept-concept metric correlation matrix. Each element m_{uv} of $\mathbf{M}_l$ expresses a normalized metric correlation factor between concepts c_u and c_v as*

$$m_{uv} = \frac{c_{uv}}{|S(c_u)| \times |S(c_v)|} \tag{52}$$

where $S(c_u)$ and $S(c_v)$ indicate the sets of the concepts c_u and c_v based on coordinate positions that are mapped at different regions in encoded images of local image set S_l. The metric correlation c_{uv} is calculated as

$$c_{uv} = \sum_{c_u \in \ S(c_u)} \sum_{c_v \in \ S(c_v)} \frac{1}{dis(c_u, c_v)} \tag{53}$$

Given the above definitions of local connection matrix $\mathbf{A}_l$ and metric matrix $\mathbf{M}_l$, we can use them to build local association and metric clusters respectively for query expansion. For a concept c_i, that occurred in a query image I_q, we consider the i-th row in matrix $\mathbf{A}_l$ or $\mathbf{M}_l$. Let $f_i(n)$ be a function which takes the i-th row and return the ordered set of n largest values a_{ij} from $\mathbf{A}_l$ or m_{ij} from $\mathbf{M}_l$, where j varies over the set of local concepts and $i \neq j$. Then $f_i(n)$ defines a local correlation or metric cluster around the concept c_i as a blue star as shown in Figure 27. Here, the concept c_j (e.g., red star)is located within a neighborhood $f_i(n)$ associated with the concept c_i.

Now, concepts that belong to clusters associated to the query concepts can be used to expand the original query. Often these neighbor concepts correlated by the current query context [44]. The steps of the query expansion process based on metric cluster are given in Algorithm 4. Similar steps also can be applied for query expansion based on the correlation cluster, where we just need to replace $\mathbf{M}_l$ with $\mathbf{A}_l$.

Algorithm 4 Query Expansion through Metric Clusters

```
1: Initialize a temporary expanded query vector that need to be added to the original
   vector of query image $I_q$, as $\mathbf{f}_q^{\mathrm{e}} = [\hat{w}_{1q}\ \hat{w}_{2q} \cdots \hat{w}_{iq} \cdots \hat{w}_{Nq}]^{\mathrm{T}}$ where each $\hat{w}_{iq} = 0$.
2: For an original query vector $\mathbf{f}_q^{\mathrm{o}} = [w_{1q}\ w_{2q} \cdots w_{iq} \cdots w_{Nq}]^{\mathrm{T}}$ of $I_q$, perform initial
   retrieval by applying a similarity (e.g. cosine) matching function.
3: Consider, the top ranked $K$ images as the local image set $S_l$ and determine the
   local codebook $C_l$.
4: Construct, the local metric matrix $\mathbf{M}_l$ based on equations (52) and (53) at query
   time.
5: for $i = 1$ to $N$ do
6:   if $w_{iq} > 0$ then
7:     Consider the $i$-th row in the metric matrix $\mathbf{M}_l$ for concept $c_i$.
8:     Return $f_i(n)$, the ordered set of $n$ largest values $m_{ij}$, where $i \neq j$, therefore
       $c_j \in C_l - \{c_i\}$.
9:     for each $c_j$ do
10:      Add and re-weight the corresponding element in query vector as $\hat{w}_{jq} += $
         $w_{iq} - ((w_{iq} - 0.1) \times k/n)$, where $k$ is the position of $c_j$ in the rank order.
11:    end for
12:  end if
13: end for
14: Obtain the re-formulated or modified query vector as $\mathbf{f}_q^{\mathrm{m}} = \mathbf{f}_q^{\mathrm{e}} + \mathbf{f}_q^{\mathrm{o}}$.
15: Perform the retrieval with the modified query vector $\mathbf{f}_q^{\mathrm{m}}$.
16: Continue the process, i.e., steps 3 to 15 until the system converges or no more
    changes are noticed.
```

Based on the step 10 of the Algorithm 4, weights are given in such a way that a top ranked concept in a ordered set gets the largest weight value and the next one gets the second largest value and so on. After this expansion process, new concepts may have been added to the original query based on the step 10 and the weight of a original query concept may have been modified had the concept belonged to the top ranked concepts based on the step 14 of the algorithm. The main drawback of the above algorithm is that it might take too long to construct the local metric matrix $\mathbf{M}_l$ in the step 4 of the algorithm. As a result, the query modification process might reduce drastically the interactive nature of a system. However, this query expansion approach would be of great assistance for application domain where effectiveness is the major issue then efficiency. As it empirically demonstrates the effectiveness in image domain, which will be shown in the experimental Section 8.5 of Chapter 8.

5.2 Query Expansion Based on Global Analysis

In this section, we present another query expansion approach in the visual concept space based on global analysis of a collection. This approach is basically a modified version of a query expansion model in text retrieval domain as proposed by Qiu and Frei in [47]. In global analysis, the query is expanded using information from the entire collection [49, 47, 50]. The basic idea is almost similar to local analysis based approaches as presented in previous section. The main difference is that in global analysis a thesaurus (matrix) is constructed based on a *term-term* (*concept-concept*) correlation or similarity relationship in entire collection instead of a local set in IR. The global analysis techniques are computationally intensive, but the computations are done off-line once per database. The only component done on a per query basis is the actual query expansion itself. The early days of research based on global analysis did not yield good improvements in text retrieval performance [50, 51, 52]. The query expansion methods tend to add a term from a thesaurus when it is strongly related in terms of similarities or correlations to one of the query terms [51]. However, query expansion need not to be limited to terms that cluster together with query terms.

Based on this observation, a query expansion model in text retrieval using a similarity thesaurus is presented by Qiu and Frei in [47]. They obtain a relatively large improvement (15-30%) in performance by adding terms that have the largest similarity to the entire query concept, rather than individual terms. To determine the ranking of each of the selected terms, the approach in [47] relies on a term-term similarity matrix. It is based on how the terms in the collection are indexed by documents, i.e., for each term there is document vector space. The query expansion terms are weighted according to the similarity between these terms and the query. If we consider the approach in [47], then each concept (term) $c_i, i \in \{1, \cdots, N\}$ is associated with a vector $\mathbf{c}_i =< w_{i1}, \cdots, w_{ij}, \cdots, w_{iM} >$ where M is the number of images (document) in the collection. The element w_{ij} is a weight for concept c_i in image I_j, which is computed (54) in a rather distinct form as [47]:

$$w_{ij} = \frac{(\frac{f_{ij}}{max_j \ (f_{ij})}) \ ikf_j}{\sqrt{\sum_{l=1}^{M}(\frac{f_{il}}{max_l \ (f_{il})})^2 \ ikf_l^2}} \tag{54}$$

where f_{ij} be the frequency of occurrence of the concept c_i in the image I_j and $max_j \ (f_{ij})$ computes the maximum frequency of c_i under all images in the collection. Further, the inverse concept frequency ikf_j for I_j, (e.g., analogous to the inverse image (document) frequency), is computed as $ikf_j = log\frac{N}{k_j}$, where k_j be the number of distinct concepts in the I_j. After generating the concept vectors, a similarity matrix $\mathbf{S}_{N \times N} = [s_{u,v}]$ is built through the computation of each element $s_{u,v}$ as the normalized cosine relationship or dot product between two concept vectors $\mathbf{c}_u$ and $\mathbf{c}_v$ as

$$s_{u,v} = \mathbf{c}_u \cdot \mathbf{c}_v = \sum_{j=1}^{M} w_{uj} * w_{vj} \tag{55}$$

Unfortunately, construction of the matrix $\mathbf{S}$ is prohibitively difficult for large collections. Many collections are available now-a-days, with several hundred thousand images or documents. Although the matrix need to be computed only once and can be computed off-line, still the approach in [47] is not computationally feasible for the on-line query expansion process. This is due to the fact that a virtual query (query concept) is constructed for query expansion as the union of all the terms that it contains and terms are represented in a feature space whose dimension is the same as the number of images, which can be very large.

To overcome this limitation, we propose a modified query expansion approach based on a global similarity matrix, which is constructed by considering the similarities of visual concept prototype vectors in a much lower dimensional feature space compared to the above one. Since, these prototype vectors are already represented in a feature space based on color and texture feature, we can use them directly to generate a *concept-concept* similarity matrix.

Definition 8 *The* concept-concept *similarity matrix* $\mathbf{S}^*_{N\times N} = [s_{u,v}]$ *is built through the computation of each element* $s_{u,v}$ *as the Euclidean similarity values between two vectors* $\mathbf{c}_u$ *and* $\mathbf{c}_v$ *of concept prototypes* c_u *and* c_v *as*

$$s_{u,v} = sim(\mathbf{c}_u, \mathbf{c}_v) = \frac{1}{1 + dis(\mathbf{c}_u, \mathbf{c}_v)} \tag{56}$$

where

$$dis(\mathbf{c}_u, \mathbf{c}_v) = [\sum_{i=1}^{d}(c_{u_i} - c_{v_i})^2]^{1/2} \tag{57}$$

Here, $\mathbf{c}_u$ *and* $\mathbf{c}_v$ *are d-dimensional vector in a combined color and texture feature space and* $c_u, c_v \in C$ *where* N *is the size of the codebook* C.

Given the global similarity matrix $\mathbf{S}^*$, query expansion is performed in the following four main steps [47]:

- First, we map an original query vector $\mathbf{f}^{\text{o}}_q = [w_{1q}\ w_{2q} \cdots w_{iq} \cdots w_{Nq}]^{\text{T}}$ from the visual concept space to a space as virtual query concept vector $\mathbf{c}^*_q$ that is used for representation of concept vectors in an Euclidean-based combined feature space based on color and texture features. The $\mathbf{c}^*_q$ is computed as

  $$\mathbf{c}^*_q = \sum_{c_u \in I_q} w_{uq} \cdot \mathbf{c}_u \tag{58}$$

  where w_{uq} is the weight of concept c_u in I_q or in other words, $w_{uq} \cdot \mathbf{c}_u$ expresses the importance of concept vector $\mathbf{c}_u$ for the query [47].

  A temporary expanded query vector is also initialized with each element has a zero value as $\mathbf{f}^{\text{e}}_q = [\hat{w}_{1q}\ \hat{w}_{2q} \cdots \hat{w}_{iq} \cdots \hat{w}_{Nq}]^{\text{T}}$, that need to be added to the original vector $\mathbf{f}^{\text{o}}_q$.

- Second, based on the similarity matrix $\mathbf{S}^*$, a similarity between each concept $c_v \in C$ and query image I_q is computed as

  $$sim(I_q, c_v) = \mathbf{c^*_q}^{\text{T}} \cdot \mathbf{c}_v = (\sum_{c_u \in I_q} w_{uq} \cdot \mathbf{c}_u)^{\text{T}} \cdot \mathbf{c}_v = \sum_{c_u \in I_q} w_{uq} \cdot (\mathbf{c}_u^{\text{T}} \mathbf{c}_v) \tag{59}$$

where $(\mathbf{c}_u^{\mathrm{T}}\mathbf{c}_v) = s_{u,v}$ is the similarity between two concept c_u and c_v that is pre-computed and defined in (56).

- Third, based on the $sim(I_q, c_v)$ value, each concept $c_v \in C$ is ranked and top n ranked concepts are selected for query expansion. After selecting the additional search concepts, we need to weight them so that they can be added to the original query vector. The weight of each concept c_v is calculated as [47]:

$$weight(I_q, c_v) = \frac{sim(I_q, c_v)}{\sum_{c_u \in I_q} w_{uq}} \tag{60}$$

  where $0 \leq weight(I_q, c_v) \leq 1$.

- Finally, we obtain the re-formulated or modified query vector as $\mathbf{f}_q^{\mathrm{m}} = (\mathbf{f}_q^{\mathrm{e}} + \mathbf{f}_q^{\mathrm{o}})$. For $\mathbf{f}_q^{\mathrm{e}} = [\hat{w}_{1q}\ \hat{w}_{2q} \cdots \hat{w}_{iq} \cdots \hat{w}_{Nq}]^{\mathrm{T}}$,

$$\hat{w}_{vq} = \begin{cases} weight(I_q, c_v) \text{ from (60)}, & \text{if } c_v \text{ belongs to top } n \text{ ranked concepts;} \\ 0, & \text{otherwise;} \end{cases}$$

  Finally, $\mathbf{f}_q^{\mathrm{m}}$ is used to retrieve images to the user.

With the above steps, the query and the concepts most similar to the virtual query concept are classified in the same cluster. After this expansion process, new concepts may have been added to the original query and the weight of an original query concept may have been modified had the concept belonged to the top ranked concept based on (60) [47].

To illustrate the above process, Figure 28 shows a virtual query concept vector $\mathbf{c}_q^*$ (e.g., the red star) and its relationships with other close concept vectors in an Euclidean space. The closeness is defined by the similarity calculation based on (56) and the closer two linked concepts (based on solid or dashed line) are to each other, the more similar they are. From the Figure 28, it is clear that concepts c_c and c_d are closer to individual query concepts $c_u, c_v \in I_q$ (e.g., the gray stars) respectively. However, in terms of the whole query concept (i.e., the virtual query vector c_q^*), concepts c_a and c_b are qualified as the closer one, if we consider only two concepts out of all other concepts in the vocabulary. It implies that, the concepts selected here for query expansion might be distinct from the ordinary global or local analysis method.

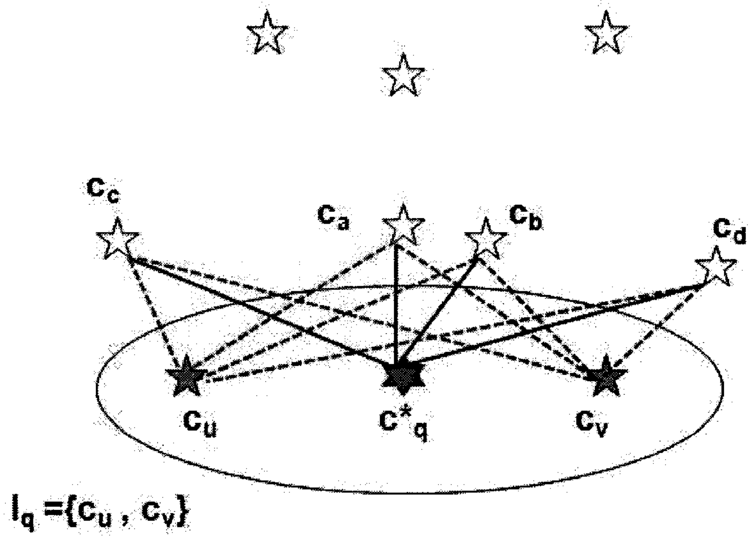
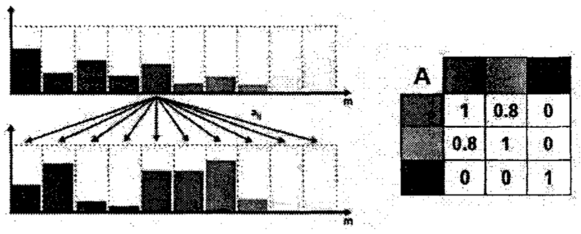


Figure 28: Relationships between concepts and virtual query concept in an Euclidean space

5.3 Similarity Matching and Query Expansion in Inverted Index

In this section, we propose a retrieval approach that combines both local and global analysis from a different perspective. In this approach, a global matrix as generated in the previous section is utilized in a Quadratic form of distance measure [16] to compare a query and database images. However, due to its quadratic nature, this distance measure is computationally expensive. To overcome this, only a subset of images from the entire collection is compared based on a local neighborhood analysis in an inverted index built on top of a codebook of visual concept prototypes.

The Quadratic distance measure is first implemented in the QBIC [16] system for the color histogram-based matching. It overcomes the shortcomings of L-norm distance functions by comparing not only the same bins but multiple bins between color histograms. Due to this property, it performs better compared to the Euclidean and histogram intersection-based distance measures for color-based image retrieval [16]. For example, Figure 29 shows the concept of the quadratic distance measure function, where instead of a *one-to-one* matching of the bins between two histograms,



A	▩	▨	■
▩	1	0.8	0
▨	0.8	1	0
■	0	0	1

Figure 29: Example of Quadratic similarity matching in Color Histogrms [12]

each bin (e.g., red here) in one histogram is matched with all other bins in the other histogram. The matching is performed with a weighted similarity score based on the matrix **A** on the right side of Figure 29, where each element a_{ij} of the matrix depicts the similarity between two colors i and j in some color space. However, as mentioned earlier, similarity based on only color feature does not always indicate semantic similarities between images due to the *semantic gap* problem.

The visual concept-based feature representation is at a higher level then the simple pixel-based color feature representation due to the incorporation of both color and texture features in a region level. Hence, instead of using a matrix based on similarities in a color space, we can effectively utilize the global visual *concept-concept* similarity matrix $\mathbf{S}^*$ as defined in Definition 8 for the distance measure computation as follows

$$Dis_{\mathbf{S}^*}(I_q, I_j) = \sqrt{(\mathbf{f}_q^{\text{V-VM}} - \mathbf{f}_j^{\text{V-VM}})^T \mathbf{S}^* (\mathbf{f}_q^{\text{V-VM}} - \mathbf{f}_j^{\text{V-VM}})} \tag{61}$$

Since, the above distance measure computes the cross similarities between concept prototype vectors, it requires longer computational time compared to the cosine and L-norm based distance measures. One solution is to compare only a subset of images from the entire collection by utilizing some indexing or pre-filtering techniques [84, 85]. In large database applications, indexing or pre-filtering techniques are essential to avoid exhaustive search in the entire collection. Some multi-dimensional tree or clustering based indexing structure has been proposed recently [2, 84, 85]. However, the accuracy of these algorithms largely depends on the feature dimension, which degrades rapidly as the dimension increases and commonly termed as the *curse of dimensionality* problem [84].

Inverted file is a very popular indexing technique in IR [44]. In text retrieval, feature vectors of both query and database documents are sparse, where they have only a small subset of all possible features or terms. The search is thus can be restricted to a subspace spanned by terms of the query. An inverted file contains an entry for every possible terms and each term contains a list of the documents if the documents have at least one occurrence of that particular term. In CBIR domain, an inverted index is used in a suitable sparse set of color and texture feature space of dimension more then ten thousands in [29]. We now present an enhanced inverted index [209] that considers the correlations or similarities between visual concept prototypes by exploiting the topology preserving property of the codebook. Our goal is to efficiently decrease the response time, where the codebook is acted as an inverted file to store the mapping from concepts to images. In this index, for each visual concept prototype in a codebook, a list of pointers or references to images that have at least one region map to this concept is stored in a list. Hence, an image in the collection is a candidate for further distance measure calculation if it contains at least one region that corresponds to a concept c_i in a query image.

Now to consider the correlation or similarity factor between concepts, this simple lookup strategy in inverted index is slightly modified. In this approach, for each concept prototype $c_i \in I_q$ with a weight (e.g., *tf-idf* based weighting) w_{iq}, we expand it to other $\lfloor w_{iq} \times (|S_\gamma| - 1) \rfloor$ concept prototypes based on the topology preserving ordering in a codebook. Here, S_γ contains all the concept prototypes including c_i up to a local neighborhood level LN_γ (we defined it in Section 3 of Chapter 4). So, for expansion, we only consider the concepts other then c_i by subtracting it from S_γ. After the expansion, those images that appear in the list of expanded concepts are deemed as candidate for further distance measure calculation, while the other images are ignored. A larger γ will lead to more expanded concepts, which means more images need to be compared with the query. This might lead to more accurate retrieval results in trade off larger computational time. After finding the $|S_\gamma| - 1$ concept prototypes, they are ranked based on their similarity values with c_i by looking up the corresponding entry in the matrix $\mathbf{S}^*$. This way relationship between two concepts are actually determined by both their closeness in the topology preserving codebook and their similarity obtained from the global matrix. Finally, the top $\lfloor w_{iq} \times (|S_\gamma| - 1) \rfloor$ concepts are selected as expanded concepts for c_i. Hence, a concept with more weighage in a query vector will be expanded to more closely related concepts and as a result will have more influence to retrieve candidate images.

Therefore, the enhanced inverted index contains an entry for a concept which consists of a list of images as well as images from closely related concepts based on the local neighborhood property. The steps of the algorithm are described below;

Algorithm 5 Query Expansion and Similarity Matching in Inverted File

```
Initially compute the global visual concept similarity matrix $\mathbf{S}^*$ off-line. Let, the feature vector of a query image $I_q$ be $\mathbf{f}_q^{\mathrm{V-VM}} = [w_{1q} \cdots w_{iq} \cdots w_{Nq}]^{\mathrm{T}}$ in a visual concept-based feature space. Initialize the list of candidate image as $L = \phi$.
for $i = 1$ to $N$ do
  if $w_{iq} > 0$ (i.e., $c_i \in I_q$) then
    Locate the corresponding concept prototype $c_i$ in the two-dimensional codebook $C$.
    Read the corresponding list $L_{c_i}$ of images from the inverted file and add it to $L$ as $L \leftarrow L \cup L_{c_i}$.
    Consider up to $LN_\gamma$ neighborhoods of $c_i$ to find related $|S_\gamma| - 1$ concept prototypes.
    For each $c_j \in S_\gamma - \{c_i\}$, determine its ranking based on the similarity values by looking up corresponding entry $s_{ij}$ in matrix $\mathbf{S}^*$.
    Consider the top $k = \lfloor w_{iq} \times (|S_\gamma| - 1) \rfloor$ ranked concept prototypes in set $S^k$ for further expansion.
    for each $c_k \in S^k$ do
      Read the corresponding list as $L(c_k)$ and add to $L$ as $L \leftarrow L \cup L_{c_k}$ after removing the duplicates.
    end for
  end if
end for
for each $I_j \in L$ do
  Apply the distance matching functions of Equation (61) between $I_q$ and $I_j$.
end for
Finally, return the top $K$ images by sorting the distance measure values in ascending order (e.g., a value of 0 indicates closest match).
```

Figure 30 shows a sample example of the above processing steps. Here, for a particular concept c_j with associated weight in vector as w_{jq} that is presented in query image I_q, the corresponding location of the concept in the codebook is found out. Suppose, based on LN_1 neighborhood of the above algorithm, only two concepts c_k and c_m are further selected for expansion. After finding the expanded concept prototypes, images in their inverted lists are merged with original set of images and considered for further distance measure calculation for ranked-based retrieval. Therefore, in addition to consider all images in the inverted list of c_j (images under black

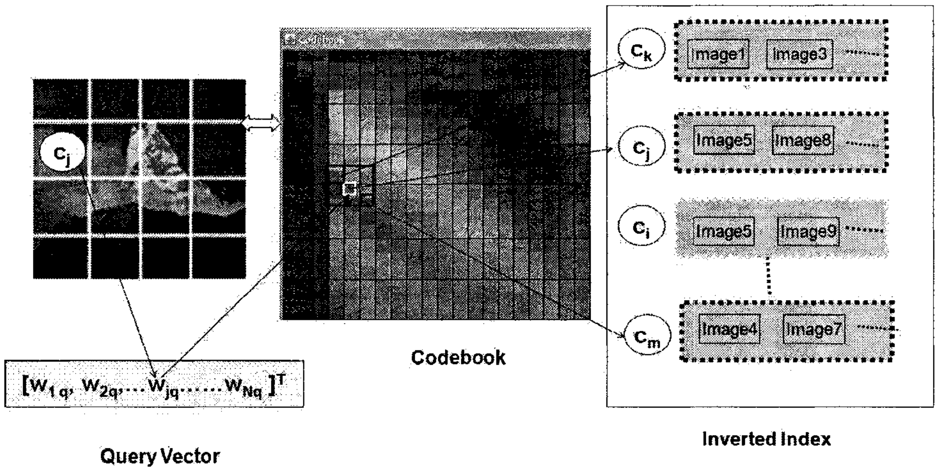


Figure 30: Example process of Query Expansion in an Inverted File

dotted rectangle), we also need to consider images in the list of c_k and c_m (under the blue dotted rectangle) as candidate images. Due to the space limitation, all actual links are not shown in Figure 30. In this way, the response time is reduced while the retrieval accuracy is still maintained.

5.4 Summary

In this chapter, we investigate automatic query expansion techniques in image domain inspired from the ideas of text retrieval in IR. These approaches exploit the correlations between concepts in different ways based on local and global analysis of an image collection. Due to the nature of the image representation schemes in concept-based feature spaces, there always exists enough correlations and/or similarities between the concepts. Hence, exploiting this property has proved to be effective in retrieval performances as will be demonstrated in the experimental section of Chapter 8.

Chapter 6

Fusion-Based Retrieval in Content and Context Feature Spaces

In a broad domain application, such as retrieval of web images, it is simply not possible to exploit domain knowledge of any kinds due to large variability among images and huge size of the collection. Many studies also suggest that a user in general would prefer browsing and posing semantic queries in broad domains to find images of particular activities, event, geographic constraints, proper names, or more abstract concepts at a much higher level (e.g., level 3 in [13] or levels 8-10 in [27]) [25, 26]. Hence, the large *semantic gap* is one of the fundamental problems in CBIR for broad domains. To reduce the gap, additional associated information is required to complement visual features of images. Adding visual information to the keyword-based search also might help to distinguish specific visual only features. For example, Figure 31 shows the differences in results (constructed manually) that we might obtain by using only keyword, only visual, and both keyword and visual example to search for a *"red bike"* in an image collection. It is clear from the Figure 31 that better result, i.e., what the user is actually looking for, can be achieved when a combined keyword and visual example based search is performed. This is because, an annotator tend to leave out what is visually obvious in the image (e.g. the color of the bike) and mention properties that are very difficult to infer using vision (e.g. the concept of the bike as a keyword). Another example from a medical domain, could involve a physician may searching for chest CT images with certain kinds of micro nodule structures. By using merely keyword search in this case will only succeed if relevant images are sufficiently

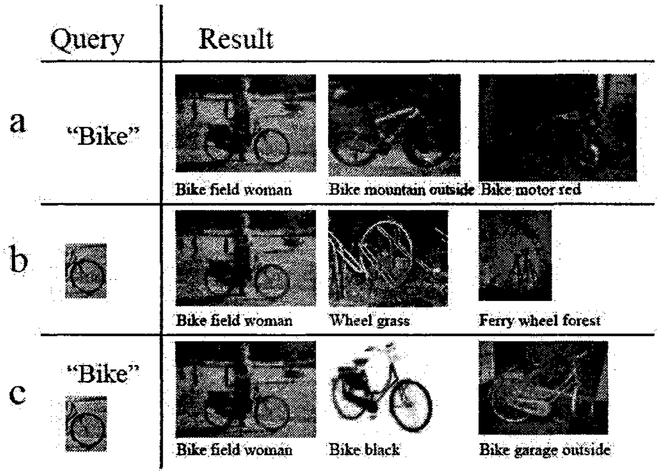


Figure 31: Constructed example of querying an image database. (a) using text only, (b) using images only, (c) combining images and text.

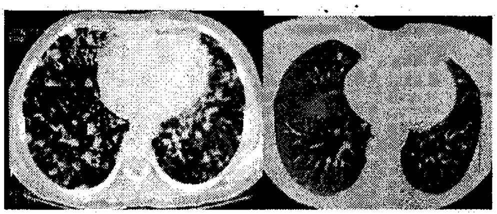


Figure 32: Example chest CT images with micro nodules

annotated with different kinds of nodule structures in addition to the term *"chest CT"*. Here, users would like to see chest CT images with certain textural properties such as, small scattered nodules and irregular sharp objects as shown in Figure 32. However, manual annotation or description of such visual properties are difficult to achieve as mentioned in [120]. Hence, in this case, it would be more appropriate to refine a keyword-based search (e.g., using the keyword *"chest CT"*) result by later applying a visual example-based search (e.g., considering some texture-based feature). This might significantly improve the initial text-based result. The above examples draw the conclusion that single-modality information retrieval, either using text as contextual information or images as visual feature, has limitations. Therefore, integration of textual information in a CBIR system or image content information to a text retrieval system might improve the retrieval performance [115, 116, 122, 119].

While there is a substantial amount of completed and ongoing research in both text retrieval as well as in CBIR, much remains to be done to see how effectively these approaches can complement each other, and how the text and image modalities can be seamlessly integrated in a single framework. Image retrieval based on multimodal information sources is recently gaining popularity due the the huge amount of multimodal information available on the web [115, 116, 117, 124, 125, 120, 123]. Majority of these systems basically perform retrieval by combining collateral text of images from web pages and visual features of images (e.g., color, texture, etc.) towards web-based image retrieval. However, we have found a lack of in-depth research and systematic evaluation along this direction in CBIR as well as in multi-modal image retrieval domains. To achieve real retrieval effectiveness, more experimental evaluation in benchmark collections is also required as it is often done in text retrieval domain, such as large-scale evaluation of text retrieval methodologies in Text REtrieval Conference (TREC) [1].

Motivated by this, we present an interactive multimodal fusion-based retrieval framework to search images in broad domain photographic and medical image collections [208, 211, 217, 218]. In this framework, for a text-based image search, keywords from the annotated files are extracted and indexed by employing the vector space model of IR. For a content-based image search, various global, semi-global, region-specific local and concept-based features are extracted at different levels of image representation. The main contribution of this work can be summarized as follows:

- Propose a relevance feedback (RF) based data fusion strategy that in addition to query reformulation in both context and content-based feature spaces, also dynamically adjusts the similarity matching functions, and inter and intra modality weights in a linear combination of similarity matching.

- Propose a function to measure the effectiveness of the features based on considering both precision and ranked order information in their corresponding result lists.

- Investigate how information from one modality can be propagated to another with a cross-modal multiple query expansion mechanism.

- Investigate the effect of retrieval performances based on both sequential and simultaneous retrieval approaches in a single retrieval framework.

[1] http://trec.nist.gov/overview.html

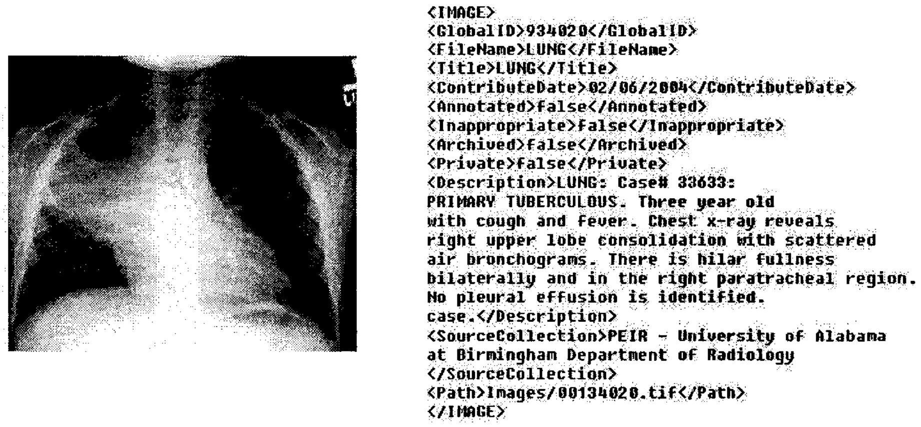
- Experiment and evaluate our retrieval approaches in benchmark photographic and medical image collections under ImageCLEF track [57, 58]. The details of the data sets will be described in Chapter 8.

6.1 Context-based Image Search Approach

This section presents the context-based image retrieval approach based on indexing of the associated caption or annotation files of database images under ImageCLEF [2] benchmark [57, 58]. In this approach, a user submits a textual query with few keywords and expects the system to return a set of relevant images associated with the top retrieved annotation files that conform to the user's query. Each annotation file in the collections is linked to the image(s) either in a one-to-one or one-to-many relationships.

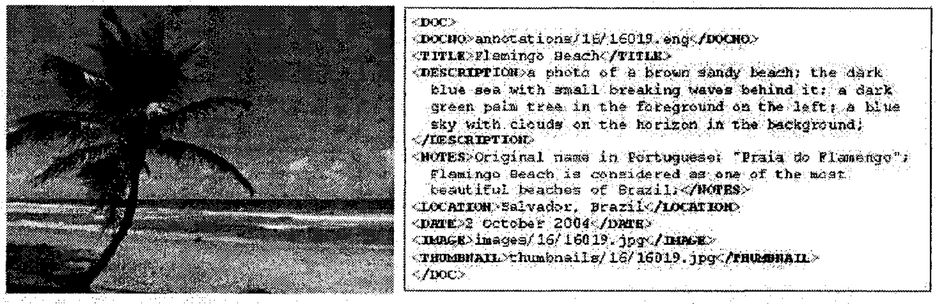
For example, Figure 33 shows an example of a lung X-ray image from a medical collection [58] with tuberculosis and its annotation with several XML tags on the right side. Figure 34 shows another image from a photographic collection [57] and its annotation with several SGML tags on the right side. As we can notice, the most important information in the above annotation files is mainly contained inside the *description* tag. For context-based image search, it is necessary to transform the annotation files into an easily accessible representation known as indexing. Indexing can be used to facilitate the location of those files and thereby corresponding images that are most likely to be relevant. There are a variety of indexing techniques which mostly rely on keywords or terms to represent the information content of documents [44]. In our case, information from only relevant tags are extracted and preprocessed by removing stop words that are considered to be of no importance for the actual retrieval process. Subsequently, the remaining words are reduced to their stems, which finally form the index terms or keywords of the annotation files. Next, the annotation files (document) are modeled as a vector of words based on the popular vector space model of IR [44, 45]. Let $T = \{t_1, t_2, \cdots, t_N\}$ denote the set of terms in the collection. Then it can represent a document or annotation file D_j as vector in a N-dimensional spaces $\mathbf{f}_{D_j} = [w_{j1}\ w_{j2} \cdots w_{jk} \cdots w_{jN}]^{\mathrm{T}}$ [45]. The element w_{jk} denotes the weight of term t_k in document D_j, depending on its information content. We used

[2] http://ir.shef.ac.uk/imageclef/



```
<IMAGE>
<GlobalID>934020</GlobalID>
<FileName>LUNG</FileName>
<Title>LUNG</Title>
<ContributeDate>02/06/2004</ContributeDate>
<Annotated>false</Annotated>
<Inappropriate>false</Inappropriate>
<Archived>false</Archived>
<Private>false</Private>
<Description>LUNG: Case# 33633:
PRIMARY TUBERCULOUS. Three year old
with cough and fever. Chest x-ray reveals
right upper lobe consolidation with scattered
air bronchograms. There is hilar fullness
bilaterally and in the right paratracheal region.
No pleural effusion is identified.
case.</Description>
<SourceCollection>PEIR - University of Alabama
at Birmingham Department of Radiology
</SourceCollection>
<Path>Images/00134020.tif</Path>
</IMAGE>
```

Figure 33: An example medical image and its associated annotation file in ImageCLEFmed collection [58].



```
<DOC>
<DOCNO>annotations/16/16019.eng</DOCNO>
<TITLE>Flamingo Beach</TITLE>
<DESCRIPTION>a photo of a brown sandy beach; the dark
  blue sea with small breaking waves behind it; a dark
  green palm tree in the foreground on the left; a blue
  sky with clouds on the horizon in the background;
</DESCRIPTION>
<NOTES>Original name in Portuguese: "Praia do Flamengo";
  Flamingo Beach is considered as one of the most
  beautiful beaches of Brazil;</NOTES>
<LOCATION>Salvador, Brazil</LOCATION>
<DATE>2 October 2004</DATE>
<IMAGE>images/16/16019.jpg</IMAGE>
<THUMBNAIL>thumbnails/16/16019.jpg</THUMBNAIL>
</DOC>
```

Figure 34: An example photographic image and its associated annotation file in ImageCLEFphoto (e.g., IAPR) collection [57].

the popular *tf-idf* term-weighting scheme [45] as described in Chapter 5. A query D_q is also represented as a vector of length N as $\mathbf{f}_{D_q} = [w_{q1}\ w_{q2} \cdots w_{qk} \cdots w_{qN}]^{\mathrm{T}}$. To compare a query and document vector, the cosine similarity measure is applied as follows [44]

$$\mathrm{Sim}_{\mathrm{text}}(D_q, D_j) = \cos(\mathbf{f}_{D_q}, \mathbf{f}_{D_j}) = \frac{\sum_{i=1}^{N} w_{qi} * w_{ji}}{\sqrt{\sum_{i=1}^{N}(w_{qi})^2} * \sqrt{\sum_{i=1}^{N}(w_{ji})^2}} \tag{62}$$

where, w_{iq} and w_{ij} are the weights of the term $\mathbf{t}_i$ in D_q and D_j respectively. The main advantages of this representation model include a ranked result of the retrieved documents, the possibility to enter free text and in exact matching of the documents [44, 45].

6.2 Content-Based Image Search Approach

The performance of an image only or content-based search mainly depends on the underlying image representation scheme [1]. Based on our previous experiments in [211], we found that the low-level image features at different levels of abstraction are complementary in nature and together they might contribute effectively to distinguish images of different visual and/or semantic categories. With this assumption, various low-level global, semi-global, region specific local, and visual concept-based image features are extracted from images and distance matching functions are defined for the features as follows.

Global feature:

For image representation at a global level, the MPEG-7 based Edge Histogram Descriptor (EHD) and Color Layout Descriptor (CLD) [62, 63] are extracted as described in Section 3.2 of Chapter 3. Let EHD and CLD be represented as vectors $\mathbf{f}_{I_j}^{\mathrm{ehd}}$ and $\mathbf{f}_{I_j}^{\mathrm{cld}}$ respectively of image I_j. The CLD and EHD descriptors are in combine termed as ***global*** feature. The overall distance measure between global feature vectors of query image I_q and database image I_j is defined as a linear combination of individual distance measures as

$$\mathrm{Dis}_{\mathrm{global}}(I_q, I_j) = \omega_{\mathrm{cld}}\mathrm{Dis}_{\mathrm{cld}}(\mathbf{f}_{I_q}^{\mathrm{cld}}, \mathbf{f}_{I_j}^{\mathrm{cld}}) + \omega_{\mathrm{ehd}}\mathrm{Dis}_{\mathrm{ehd}}(\mathbf{f}_{I_q}^{\mathrm{ehd}}, \mathbf{f}_{I_j}^{\mathrm{ehd}}), \tag{63}$$

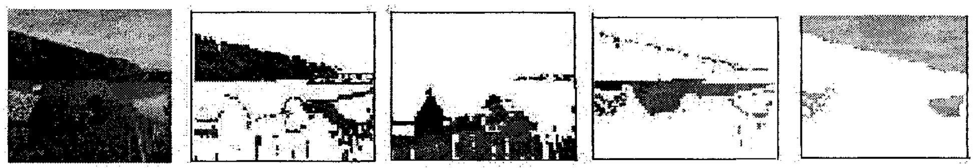


Figure 35: Region generation based on segmentation of a photographic image

where, $\text{Dis}_{\text{cld}}(\mathbf{f}^{\text{cld}}_{I_q}, \mathbf{f}^{\text{cld}}_{I_j})$ and $\text{Dis}_{\text{ehd}}(\mathbf{f}^{\text{ehd}}_{I_q}, \mathbf{f}^{\text{ehd}}_{I_j})$ are the Euclidean distance measures for CLD and EHD respectively and ω_{cld} and ω_{ehd} are weights for each feature distance measure subject to $0 \leq \omega_{\text{cld}}, \omega_{\text{ehd}} \leq 1$ and $\omega_{\text{cld}} + \omega_{\text{ehd}} = 1$. Initially these are taken to be equal weights with $\omega_{\text{cld}} = 0.5$ and $\omega_{\text{ehd}} = 0.5$.

Semi-Global Feature:

For image representation at a semi-global level, we use the grid-based decomposition and feature extraction approach as described in Section 3.2.3 of Chapter 3. In this approach, images are divided into five overlapping sub-images and color and texture moment-based features are extracted from each sub-images. Color and texture feature vectors are normalized and combined to form a joint feature vector $\mathbf{f}^{\text{sg}}_{r_j}$ of each sub-image $r_j \in I_j$ and finally they are combined as the semi-global feature vector for the entire image as $\mathbf{f}^{\text{sg}}_{I_j}$. The semi-global distance measure between I_q and I_j is defined as

$$\text{Dis}_{\text{s-global}}(I_q, I_j) = \text{D}_{\text{sg}}(\mathbf{f}^{\text{sg}}_{I_q}, \mathbf{f}^{\text{sg}}_{I_j}) = \frac{1}{5}\sum_{r=1}^{5} \omega_r \text{Dis}_r(\mathbf{f}^{\text{sg}}_{r_q}, \mathbf{f}^{\text{sg}}_{r_j}) \tag{64}$$

where, $\text{Dis}_r(\mathbf{f}^{\text{sg}}_{r_q}, \mathbf{f}^{\text{sg}}_{r_j})$ is the Euclidean distance measure of the feature vector of a region $r \in I_q, I_j$ and ω_r are the weights for the regions, which are initially set as equal.

Region-Specific Local Feature:

The above fixed partitioning scheme is comparatively simpler and has limitations as it might not match with the actual semantic partitioning of the objects. Region-based image retrieval aims to overcome this limitation by fragmenting an image automatically into a set of homogeneous regions based on color and/or texture properties [21, 22, 24, 23]. Representation of images at region level is proved to be more close to human visual system where each region can be described by means of local features.

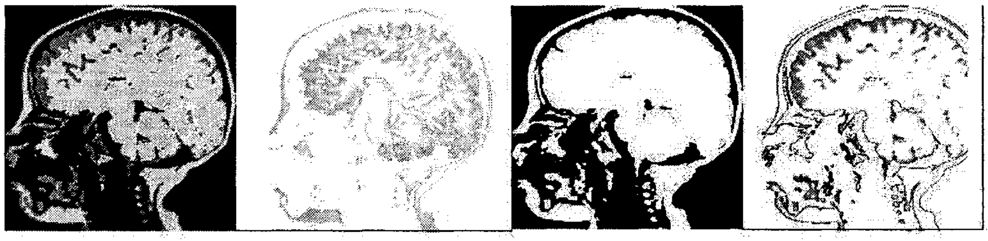


Figure 36: Region generation based on segmentation of a medical image

To perform region-based retrieval, the first step is to implement image segmentation. Then, low-level features such as color, texture, shape or spatial location can be extracted from the segmented regions. The overall similarity between two images can be defined based on all the corresponding region-based local features.

We use a clustering-based image segmentation technique [24] and a robust image to image level similarity matching function based on clusters or regions generated from automatic segmentation [211]. For images in broad domains, currently there is no image segmentation algorithm that can perform at the level of the human visual system. The segmentation accuracy of our system is not crucial because we use a more robust similarity matching scheme which is insensitive to inaccurate segmentation. To automatically segment the images with distinct regions, a fast k-means clustering method is utilized [43, 24]. The clustering method is fully automatic and unsupervised in nature that can adaptively generate the number of regions as an iterative process since the number of region is unknown before the segmentation. To segment an image, we partition the image into non-overlapping blocks with (2 × 2) pixels. We choose block-wise segmentation since it has little effect in retrieval with the benefit of 4 times faster segmentation. The average color components of each image-block are extracted as feature vector in RGB color space. Suppose there is a set $X = \{x_i, \cdots, x_L\}$ of L blocks for each image to be segmented. The goal of the k-means algorithm is to partition the blocks into k clusters with means $V = \{\mu_i, \cdots, \mu_k\}$ such that

$$D(k) = \sum_{i=1}^{k} \sum_{x_j \in C_i} (\mathbf{x}_j - \boldsymbol{\mu}_i)^2 \qquad (65)$$

is minimized, where $\boldsymbol{\mu}_i$ is the centroid or mean vector of the i-th cluster C_i and $\mathbf{x}_j$ is the block vector (e.g., average color) of a particular block $x_j \in X$. We primarily select the number of clusters with $k = 2$ and gradually increase k until either the the number k exceeds an upper bound or the distortion $D(k)$ is below a threshold. A low $D(k)$ indicates high purity in the clustering process. For experiments, we set maximum $k = 10$, therefore allow an image to be segmented into a maximum of 10 regions and threshold value as 10^{-3}.

After segmentation, every cluster is corresponding to one region in the segmented image. Since clustering is performed in the feature space, blocks in each cluster do not necessarily form a connected region in the image space. Figure 35 shows an example of segmentation result of a photographic image of mountain with four regions where it clearly separates the major objects, such as sky, water, hill and grasses. Whereas, Figure 36 shows the segmentation result with separation of brain, background, and skull in different regions of a MRI-head image in a medical collection. We do not apply any post-processing methods to smooth region boundaries or to delete small isolated regions because these errors are rarely significant. The non-connected property of some regions preserves the natural clustering of an object that is good enough for our proposed image level similarity matching function.

To represent each region with local features, we consider information on weight and color and texture related feature as in [23]. Let, w_{r_j} is the weight of a region $r_j \in I_j$ and is defined as $w_{r_j} = \frac{N_r}{N_{I_j}}$, where N_r is the number of blocks in region r and N_{I_j} is the total number of blocks in I_j. The average color feature vector $\mathbf{f}^{\mathrm{c}}_{r_j}$ of each region r is represented by the k-means cluster center, i.e., the average value for each of the three color channels in RGB space of all the image blocks in this region. Texture feature of each region is measured in an indirect way by considering the cross-correlation among color channels due to the off diagonal of the 3×3 covariance matrix C_{r_j} of region r of I_j and is estimated as

$$C_{r_j} = \frac{1}{N_r - 1} \sum_{k=1}^{N_r} (\mathbf{f}^{\mathrm{c}}_{x_k} - \mathbf{f}^{\mathrm{c}}_{r_j})(\mathbf{f}^{\mathrm{c}}_{x_k} - \mathbf{f}^{\mathrm{c}}_{r_j})^T \tag{66}$$

where $\mathbf{f}^{\mathrm{c}}_{x_k}$ is the average color vector of a block $x_k \in r$. Finally, the mean color vector $\mathbf{f}^{\mathrm{c}}_{r_j}$ and color covariance C_{r_j} of each region r of I_j are combined or concatenated to form a region-based local feature vector $\mathbf{f}^{\mathrm{local}}_{I_j}$ of variable lengths.

In region-based image retrieval systems, image similarity is measured first at region level and after that at image level. That is to measure the overall similarity of two images which might contain different number of regions [15]. In a *"one-to-one"* matching scheme, each region in the query image is only allowed to match one region in the target image [23]. Whereas, each region in the query image is allowed to match more than one region in the target image and vise versa in a *"many-to-many"* matching scheme, for example the *Integrated Region Matching (IRM)* in [24]. We propose an image level *"one-to-one"* matching scheme [211] between I_q and I_j by integrating properties of all the regions in both query and database images. This matching scheme is closely related to the one proposed in [23]. The main difference is that, instead of only one way matching in [23], i.e., from query image regions to database image regions, here we allow both ways of matching. One region of an image is to be matched to several regions of another image and vice versa by considering only the best matching pair for final distance calculation and finally averaging out the distance scores obtained from both ways. This scheme is more robust against poor segmentation as it considers properties of all regions in both query and database images. Suppose, there are M regions in image I_q and N regions in image I_j. Now, the image-level distance is defined as

$$\mathrm{Dis}_{\mathrm{local}}(I_q, I_j) = \frac{\sum_{i=1}^{M} w_{r_{i_q}} \mathrm{Dis}_{r_{i_q}}(q, j) + \sum_{k=1}^{N} w_{r_{k_j}} \mathrm{Dis}_{r_{k_j}}(j, q)}{2} \tag{67}$$

where $w_{r_{i_q}}$ and $w_{r_{k_j}}$ are the weights for region r_{i_q} and region r_{k_j} of image I_q and I_j respectively. For each region $r_{i_q} \in I_q$, $\mathrm{Dis}_{r_{i_q}}(q, j)$ is defined as the minimum distance between this region and any region $r_{k_j} \in I_j$

$$\mathrm{Dis}_{r_{i_q}}(q, j) = \min(\mathrm{Dis}(r_{i_q}, r_{1_j}), \cdots, \mathrm{Dis}(r_{i_q}, r_{N_j})) \tag{68}$$

Similarly,

$$\mathrm{Dis}_{r_{k_j}}(j, q) = \min(\mathrm{Dis}(r_{k_j}, r_{1_q}), \cdots, \mathrm{Dis}(r_{k_j}, r_{M_q})) \tag{69}$$

Now, to compute the distance between any two regions r_{i_q} and r_{k_j} of I_q and I_j respectively, we apply the Bhattacharyya distance measure [172] as follows:

$$\mathrm{Dis}(r_{i_q}, r_{k_j}) = \frac{1}{8}(\mathbf{f}^{\mathrm{c}}_{r_{i_q}} - \mathbf{f}^{\mathrm{c}}_{r_{k_j}})^T \left[\frac{(C_{r_{i_q}} + C_{r_{k_j}})}{2}\right]^{-1}$$

$$(\mathbf{f}^{\mathrm{c}}_{r_{i_q}} - \mathbf{f}^{\mathrm{c}}_{r_{k_j}}) + \frac{1}{2}\ln\frac{\left|\frac{(C_{r_{i_q}}+C_{r_{k_j}})}{2}\right|}{\sqrt{|C_{r_{i_q}}||C_{r_{k_j}}|}} \tag{70}$$

where $\mathbf{f}^{\mathrm{c}}_{r_{i_q}}$ and $\mathbf{f}^{\mathrm{c}}_{r_{k_j}}$ are the mean color vectors and $C_{r_{i_q}}$ and $C_{r_{k_j}}$ are the covariance matrices of region r_{i_q} and r_{k_j} respectively. Equation (89) is composed of two terms, the first one being the distance between feature vectors of image regions, while the second term gives the class separability due to the difference between covariance matrices. This definition of image-level distance between two images captured by the overall distance between the region sets of two images is a balanced scheme in distance measure between regional and global matching.

Visual Concept-Based Feature:

In addition, we extract the normal frequency-based visual concept-based feature vector from images as described in subsection 4.1.2 of Chapter 4. Here, an image I_j can be represented as

$$\mathbf{f}_{I_j}^{\mathrm{V-concept}} = [f_{1_j}\ f_{2j} \cdots f_{i_j} \cdots f_{N_j}]^{\mathrm{T}} \tag{71}$$

where each element f_{ij} represents the normalized frequency of a concept $c_i \in C$ of I_j and N is the size of the codebook C. Since, the concept-based feature representation is closely related to the keyword-based representation of documents, we apply the similar cosine measure to compare image I_q and I_j as described in (62). For the experiments in Chapter 8, codebooks of size of 400 (e.g.,20 × 20 units) are constructed for the photographic and medical collections by manually selecting 2% images from each collection as training sets.

In the following sections, from a data fusion perspective, we present several query reformulation methods in both contextual and visual domains and show how they can be dynamically integrated in an interactive retrieval framework for a better retrieval accuracy.

6.3 Contextual Query Refinement

For the context-based retrieval approach, users might often find it difficult to express their information need in the form of a short query with few keywords. Moreover, query terms might not exactly match the terms appearing in the annotation files with only few keywords or caption. Hence, a query reformulation process is essential to add additional query terms and re-weight the original query vector. Here, we investigate interactive ways to generate multiple contextual query representations by applying few well known RF-based techniques in text retrieval domain [106, 107, 105]. The main motivation is that different RF methods might have different properties and generate quite different modified query vectors [202, 203]. Multiple query representations can provide different interpretations of a user's underlying information need, or provide more detail about how the user is making relevance assessments. It is well known that different query representation could retrieve different set of documents in text retrieval domain [203]. From data fusion research perspective in IR [201, 202, 203], it also has been shown that if two search retrieve different sets of documents, significant improvement can be achieved by combining the retrieval results.

In order to investigate the effectiveness of a fusion approach in the context-based retrieval approach, we at first generate an initial query vector for a given query topic and perform the retrieval to rank an initial set of annotation files and thereby rank the images linked to the files. Based on the user's feedback information from the top retrieved images and thereby from associated annotation files, we next generate multiple query vectors by applying several RF-based methods [106, 107, 105]. The main idea of RF in text retrieval is to refine the original query by adding new terms from relevant documents (i.e. query expansion) and enhance the importance of query terms appearing in relevant documents (i.e. term re-weighting). One of the best known query refinement approaches is the *Rocchio* algorithm [106]. In this approach, the problem of retrieval is defined as that of defining an optimal query; one that maximizes the difference between the average vector of the relevant documents and the average vector of the non-relevant documents. In most systems, the following improved version of the original Rocchio's formula is utilized [106]:

$$\mathbf{f}^{m}_{D_q}(\mathit{Rocchio}) = \alpha\ \mathbf{f}^{o}_{D_q} + \beta \frac{1}{|R|} \sum_{\mathbf{f}_{D_j} \in R} \mathbf{f}_{D_j} - \gamma \frac{1}{|\hat{R}|} \sum_{\mathbf{f}_{\hat{D}_j} \in \hat{R}} \mathbf{f}_{\hat{D}_j} \tag{72}$$

where, $\mathbf{f}^{m}_{D_q}$ and $\mathbf{f}^{o}_{D_q}$ are the modified and original query vectors, R and $\hat{R}$ are the sets of relevant and irrelevant document vectors and α, β, and γ are weights attached to each term. These control the balance between trusting the judged document set versus the query: if we have a lot of judged documents, we would like a higher β and γ. This algorithm generally moves a new query point toward relevant documents and away from irrelevant documents in feature space [106]. The original Rocchio's formula also modified in [107] by eliminating the normalization for the number of relevant and non-relevant documents and allowing limited negative feedback from only the top-ranked non-relevant document. Although this technique, known as *"Ide-dec-hi"* formula, did not improve results greatly it was more consistent; improving the performance of more queries.

$$\mathbf{f}^{m}_{D_q}(Ide) = \alpha\, \mathbf{f}^{o}_{D_q} + \beta \sum_{\mathbf{f}_{D_j} \in R} \mathbf{f}_{D_j} - \gamma \max_{\hat{R}}(\mathbf{f}_{D_j}) \tag{73}$$

where $\max_{\hat{R}}(\mathbf{f}_{D_j})$ is a vector to the highest ranked non-relevant document.

In addition to utilizing the above two RF approaches, we also perform two different query modification based on local analysis as described in section 5.1 of Chapter 5. Generally, the local analysis approaches consider the top K most highly ranked documents for query expansion without any assistance from the user. However, we consider only the user selected relevant images for further analysis as that information is already available from the RF-based approaches.

At first, a simpler approach of query expansion is considered based on identifying useful terms or keywords from the associated annotation files for the relevant images. This approach extracts the keywords after removing the stop words from the relevant annotation files based on user's positive feedback information. After extracting the keywords, the method determines the most frequently occurring k keywords and add them to the original query. The query vector is modified as $\mathbf{f}^{m}_{D_q}(Local1)$ by re-weighting its keywords (e.g., original plus added keywords) based on the *tf-idf* weighting scheme and is re-submitted to the system as the new query. We also utilize a local cluster-based query reformulation approach based on expanding the query with terms correlated to the query terms. Such correlated terms are those present in local clusters built from the relevant documents as indicated by the user. There are many ways to build the local cluster before performing any query expansion [46, 51, 52]. We generate similar local cluster based on the local correlation matrix as defined in Section 1.2 of Chapter 3. However, instead of using local concepts, here we use the extracted keywords of relevant annotation files as indicated by a user. Let us consider, for a given query D_q, the set of all distinct terms as T_l called the local vocabulary in a set of relevant documents. A correlation matrix $A_{(|T_l| \times |T_l|)}$ is constructed in which the rows and columns are associated with the terms in the local vocabulary. The element of this matrix $a_{u,v}$, is defined as

$$a_{u,v} = \frac{n_{u,v}}{n_u + n_v - n_{u,v}} \tag{74}$$

where, n_u is the number of local documents which contain term $t_u \in T_l$, n_v is the number of local documents which contain term $t_v \in T_l$, and $n_{u,v}$ is the number of local documents which contain both terms. Now, given the correlation matrix A_l, we can use it to build the local correlation cluster as follows. Consider the u-th row in matrix A_l (i.e., the row with all the correlations for the keyword t_u). From each row, it returns the set of n largest correlation values $a_{u,l}$, where $l \neq i$. Now, for a query D_q, we are normally interested in finding clusters only for the $|D_q|$ query terms. After extracting there additional n terms for each query term, the query vector is updated as $\mathbf{f}^m_{D_q}(Local2)$ by re-weighting its keywords based on the *tf-idf* weighting scheme.

Hence, we obtain for different query vectors (e.g., $\mathbf{f}^m_{D_q}(Rocchio), \mathbf{f}^m_{D_q}(Ide), \mathbf{f}^m_{D_q}(Local1)$, and $\mathbf{f}^m_{D_q}(Local2)$) by applying the above methods. In Section 6.5, we will present an elaborate methods to dynamically weight each of the representations for on-line retrieval, so that they can contribute effectively according to their importance for a particular query.

6.4 Visual Query Refinement

This section presents our visual query refinement approach based on RF that not only performs query point movement but also adjusts the distance matching functions and the feature weights for different image representations [208, 217, 217]. In this approach, it is assumed that, all positive feedback images at some particular iteration belong to a user perceived semantic space and obey the Gaussian distribution to form a cluster in that space. We consider the rest of the images as irrelevant and they may belong to different semantic categories. However, only relevant images are considered for query refinement in this case. The modified query vector at iteration

k is represented as the mean of the relevant image vectors as follows

$$\mathbf{f}^{m}_{I^{x}_{q}(k)} = \frac{1}{|R_k|} \sum_{\mathbf{f}_{I^{x}_{l}} \in R_k} \mathbf{f}_{I^{x}_{l}} \tag{75}$$

where, R_k be the set of relevant image vectors at iteration k and x individually represents the CLD or EHD features for global, and color or texture moment-based features for semi-global feature vectors. Since, there might exist significant correlations between the feature attributes of the positive individual feature vectors of relevant images, we capture this information in the form of covariance matrices and finally use them in Mahalanobis distance measure functions [172]. The covariance matrices of the positive feature vectors are estimated as

$$\mathbf{C}^{x}_{(k)} = \frac{1}{|R_k| - 1} \sum_{l=1}^{|R_k|} (\mathbf{f}_{I^{x}_{l}} - \mathbf{f}^{m}_{I^{x}_{q}(k)})(\mathbf{f}_{I^{x}_{l}} - \mathbf{f}^{m}_{I^{x}_{q}(k)})^{T} \tag{76}$$

However, singularity issue might arise for the inverse computation of covariance matrices in distance measure functions if fewer training samples (e.g., positive feedback images) are available compared to the feature dimension. This will be the most probable case as users tend to provide only few positive or negative feedbacks. So, we add the following regularization to avoid singularity in matrices as follows[185]:

$$\hat{\mathbf{C}}^{x}{}_{(k)} = \alpha \mathbf{C}^{x}_{(k)} + (1 - \alpha)\mathbf{I} \tag{77}$$

for some $0 \leq \alpha \leq 1$ and $\mathbf{I}$ is the identity matrix. After generating the mean vector and covariance matrix for a feature x, the individual Euclidean-based distance measure functions in equations (63) and (64) are replaced with the following Mahalanobis distance measure [172] for query image I_q and database image I_j at iteration k as

$$\mathrm{Dis}_{\mathrm{x}}(I_q, I_j) = (\mathbf{f}^{m}_{I^{x}_{q}(k)} - \mathbf{f}_{I^{x}_{j}})^{T} \hat{\mathbf{C}}^{x}{}^{-1}_{(k)} (\mathbf{f}^{m}_{I^{x}_{q}(k)} - \mathbf{f}_{I^{x}_{j}}) \tag{78}$$

The Mahalanobis distance differs from the Euclidean distance in that it takes into account the correlations of the data set and is scale-invariant, i.e. not dependent on the scale of measurements [172].

We do not perform any query refinement for region-specific feature at this moment due to its variable feature dimension for variable number of regions in each image. Since the visual concept-based image feature is closely related to the keyword-based feature of text retrieval, the *"Ide-dec-hi"* formula in equation (73) is used for its better performances for query refinement. In this case, the modified query vector is defined as

$$\mathbf{f}_{I_q}^{m} = \alpha\ \mathbf{f}_{I_q}^{o} + \beta \sum_{\mathbf{f}_{I_j} \in R} \mathbf{f}_{I_j} - \gamma \max_{\hat{R}}(\mathbf{f}_{I_j}) \tag{79}$$

where $\max_{\hat{R}}(\mathbf{f}_{I_j})$ is a vector to the highest ranked non-relevant image.

The modified query vectors of both contextual and visual feature spaces are submitted to the system for the next iteration. In the following section, we propose a dynamically weighted linear combination of similarity fusion technique based on relevance feedback information. This technique updates both inter and intra modality feature weights in similarity matching for the next iteration to obtain a final ranked list of images either from a individual modality (e.g., text or image) or combination of both in a single search.

6.5 Adaptive Linear Combination of Multiple Evidences

It is difficult to find a unique representation to compare documents and images accurately for all types of queries. In other words, each feature representation along with its similarity matching function might be complementary in nature and will have its own limitations. In IR, the category of work is known as *data fusion*. Data fusion or multiple-evidence combination describes a range of techniques in IR whereby multiple pieces of information are combined to achieve improvements in retrieval effectiveness [201, 202, 203]. The information can take many forms including different query representations, different document representations, and different retrieval strategies used to obtain a measure of relationship between a query and a document. Recently, some multimodal image retrieval approaches also adopts some of the ideas from data fusion research, where the most commonly used approach is a linear combination of text and image-based similarity scores based on pre-determined or adaptive feature weights [125, 124, 119, 121]. This section presents an improved adaptive linear combination scheme based on user's feedback information. It considers the importance of a feature based on both precision and rank order information of top retrieved images in its result list. By considering two criteria at a time, it provides us a better measurement of importance or weightage of the features.

Before performing any linear combination, the distance measure scores of each representation are normalized and converted to the similarity scores with a range of $[0, 1]$ as

$$\mathrm{Sim}(\mathbf{f}_q, \mathbf{f}_j) = 1 - \frac{\mathrm{Dis}(\mathbf{f}_q, \mathbf{f}_j) - \min(\mathrm{Dis}(\mathbf{f}_q, \mathbf{f}_j))}{\max(\mathrm{Dis}(\mathbf{f}_q, \mathbf{f}_j)) - \min(\mathrm{Dis}(\mathbf{f}_q, \mathbf{f}_j))} \tag{80}$$

where $\min(\cdot)$ and $\max(\cdot)$ are the minimum and maximum distance scores between query and database images (documents) for a particular feature vector $\mathbf{f}$. Generally, a similarity score is the converse of a distance score. So, when the similarity score is one (i.e. exactly similar), the distance score is zero and vice versa.

For the multi-modal retrieval purpose, let us consider q as a multi-modal query which has an image part as I_q and a context part as annotation file as D_q. In a linear combination scheme, the similarity between q and a multi-modal item j, which also has two parts (e.g., image I_j and context D_j), is defined as

$$\mathrm{Sim}(q, j) = \omega_I \mathrm{Sim_I}(I_q, I_j) + \omega_D \mathrm{Sim_D}(D_q, D_j) \tag{81}$$

where ω_I and ω_D are inter-modality weights within the context and image feature spaces, which subject to $0 \leq \omega_I, \omega_D \leq 1$ and $\omega_I + \omega_D = 1$.

In this framework, the image-based similarity, $\mathrm{Sim_I}(I_q, I_j)$ is again defined as the linear combination of individual similarity measures in different level of image representation as

$$\mathrm{Sim_I}(I_q, I_j) = \sum_{IF} \omega_I^{IF}\ \mathrm{Sim}_I^{IF}(I_q, I_j) \tag{82}$$

where $IF \in \{\mathrm{global, semi-global, region, visual-concept}\}$ and ω^{IF} are the weights within the different image representation schemes (e.g., intra-modality weights). Whereas, the context based similarity, $\mathrm{Sim_D}(D_q, D_j)$ is defined as the linear combination of similarity matching based on different query representation schemes (as obtained by query reformulation approaches in section 6.3) as

$$\mathrm{Sim_D}(D_q, D_j) = \sum_{QF} \omega_D^{QF}\ \mathrm{Sim}_I^{QF}(D_q, D_j) \tag{83}$$

where $QF \in \{\text{Rocchio}, \text{Ide}, \text{Local1}, \text{Local2}\}$ and ω^{QF} are the intra-modality weights within the different query representation schemes.

The effectiveness of the linear combination depends mainly on the choice of different inter and intra-modality weights. We provide an equal emphasis by providing equal weights to all the features along with their similarity matching functions initially. However, the weights are updated dynamically during the subsequent iterations by incorporating the feedback information from the previous round. To update the inter-modality weights (e.g., ω_I and ω_D), we at first need to perform the multi-modal similarity matching based on equation (81). After the initial retrieval result with a linear combination of equal weights (e.g., $\omega_I = 0.5$ and $\omega_D = 0.5$), a user needs to provide feedback about the relevant images from the top K returned images. For each ranked result list based on individual similarity matching, we also consider the top K images and measure the effectiveness of a feature by applying the following function

$$\mathrm{E}(D \text{ or } \mathrm{I}) = \frac{\sum_{\mathrm{i}=1}^{\mathrm{K}} \mathrm{Rank(i)}}{\mathrm{K}/2} * \mathrm{P(K)} \tag{84}$$

where $\mathrm{Rank(i)} = 0$ if image in the rank position i is not relevant based on user's feedback and $\mathrm{Rank(i)} = (\mathrm{K} - \mathrm{i})/(\mathrm{K} - 1)$ for the relevant images. Hence, the function Rank(i) monotonically decreasing from one (if the image at rank position 1 is relevant) down to zero (e.g., for a relevant image at rank position K). On the other hand, $P(K) = R_K/K$ is the precision at top K, where R_k be the number of relevant images in the top K retrieved result. Hence, the equation (84) is basically the product of two factors, rank order and precision. The rank order factor takes into account the position in the retrieval set of the relevant images, whereas the precision is a measure of the retrieval accuracy, regardless of the position. Generally, the rank order factor is heavily biased for the position in the ranked list over the total number of relevant images and the precision value totally ignores the rank order of the images. To balance both the criteria, we use a performance measure that is the product of the rank order factor and precision. If there is more overlap between the relevant images of a particular result set and the final one from which the user provides the relevant feedback information, then the performance score will be higher. Both terms on the right side of equation (84) will be 1, if all the top K returned images are considered as relevant. The raw performance scores obtained by the above procedure are then normalized by the total score as $\hat{E}(D) = \hat{\omega}_D = \frac{E(D)}{E(D)+E(I)}$ and $\hat{E}(I) = \hat{\omega}_I = \frac{E(I)}{E(D)+E(I)}$

to generate the updated context and content feature weights respectively. For the next iteration of retrieval with the same query, these modified weights are utilized for the multi-modal similarity matching function as

$$\mathrm{Sim}(q,j) = \hat{\omega}_I \mathrm{Sim_I}(I_q, I_j) + \hat{\omega}_D \mathrm{Sim_D}(D_q, D_j) \tag{85}$$

This feedback and therefore weight updating process can be continued as long as no changes are noticed in final retrieval result due to the convergence of the system.

In a similar fashion, to update the intra-modality weights (e.g., ω_D^{QF} and ω_I^{IF}), we consider the top K images in individual result lists. So, for image-based similarity in equation (82), we consider the result lists of different image features, $IF \in \{\mathrm{global}, \mathrm{semi-global}, \mathrm{region}, \mathrm{visual-concept}\}$ and measure their weights by using equation (84) for the next retrieval iteration. For text-based similarity in equation (83), the top K images in result lists of different query features, $QF \in \{\mathrm{Rocchio}, \mathrm{Ide}, \mathrm{Local1}, \mathrm{Local2}\}$ are considered and text-level weights are determined in a similar way by applying equation (84).

6.6 Cross-modal Interaction and Integration

This section presents both sequential and simultaneous search approaches by considering query refinement and dynamic weight updating approaches in a single framework. The approaches can be used to expand a contextual query using related keywords obtained from top retrieved relevant (based on user's feedback) annotation files based on a context or content search in previous iteration. In a similar fashion, a visual example-based query can be reformulated using image features from top retrieved relevant images based on a content or context search in previous iteration. Hence, the flexibility of the search processes can implicitly creates a semantic network to link keywords with image features or vice versa.

6.6.1 Sequential Search with Pre-filtering and Re-ranking

Since a query can be represented with both keywords and visual features, it can be initiated either by a keyword-based search or by a visual example image search. Here, we consider a pre-filtering and re-ranking approach based on the image search in the filtered image set that is obtained previously by the context-based search. This approach would be more appropriate and effective for searching images when the queries are in totally abstract level, i.e., level 3 in [13]. Hence, a context-based search can be performed at first to locate and filter images with high-level semantic contents and latter we can perform an image-based search to refine or re-rank the filtered images. In this process, combining the results of the context and content-based retrieval is a matter of re-ranking or re-ordering of the images in a context-based pre-filtered result set. The steps involved in this entire search process based on user interaction are described in Algorithm 6.

Algorithm 6 Sequential Search Approach

```
Initially, for a multi-modal query $q$ with a document part as $D_q$, perform a textual search with vector $\mathbf{f}_{D_q}$ and rank the images based on the ranking of the associated annotation files by applying equation (62).
repeat
    Obtain user's feedback from top retrieved $K$ images about relevant images for the textual query refinement.
    Calculate the reformulated query vectors as $\mathbf{f}^m_{D_q}(Rocchio), \mathbf{f}^m_{D_q}(Ide), \mathbf{f}^m_{D_q}(Local1)$ and $\mathbf{f}^m_{D_q}(Local2)$.
    Re-submit the modified query vectors for context-based search.
    Calculate the query weights $\omega^{QF}$ by using (84) based on individual result lists.
    Use the linear combination of similarity matching scores based on (83) with updated weights to obtain a final ranked list.
until user switches to visual only search
For the image part as $I_q$ of $q$, extract different query features as $\mathbf{f}^{\text{global}}_{I_q}, \mathbf{f}^{\text{sg}}_{I_q}, \mathbf{f}^{\text{local}}_{I_q}$, and $\mathbf{f}^{\text{V-concept}}_{I_q}$.
Perform visual only search in filtered $L$ images retrieved by previous context-based search and rank them based on the similarity values by applying equation (82) with an equal feature weighting.
repeat
    Obtain user's feedback about relevant images from the top retrieved $K$ images.
    Perform visual query refinements as $\mathbf{f}^m_{I_q}$, based on Equations (75) for global and semi-global and (79) for visual concept-based feature respectively.
    Re-submit the modified query vectors for content-based search, where each individual Euclidean distances are replaced by Mahalanobis distance based on Equation (78).
    Calculate the feature weights $\omega^{IF}$ by using (84) based on individual result lists.
    Use the linear combination of similarity matching scores based on (82) with updated weights to obtain a final ranked list.
until user is satisfied or no more improvement
```

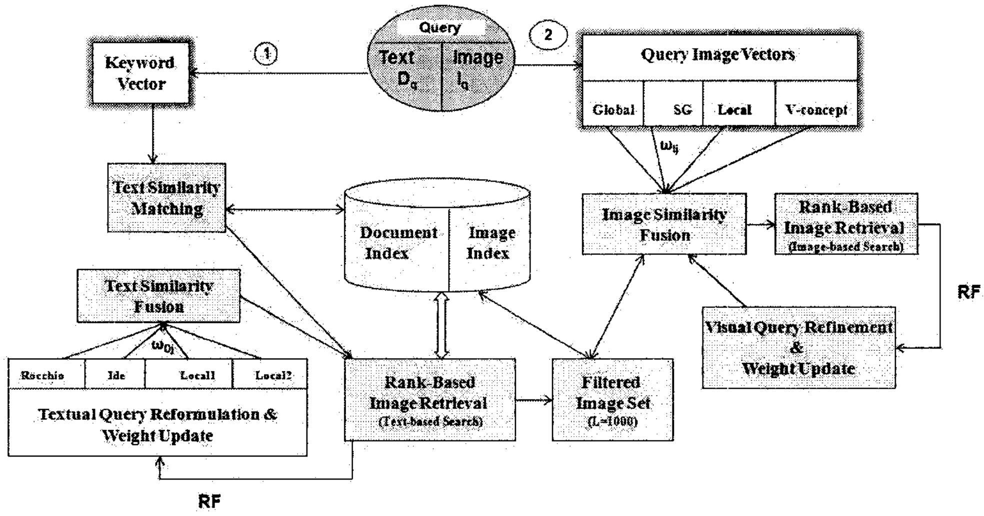


Figure 37: Process flow diagram of the sequential search approach

Figure 37 shows the process flow diagram of the above multi-modal sequential and filtering based search approach. Based on the algorithm, the context-based search with query reformulation is performed first as shown in the left portion (1) of the Figure and content-based search is performed in the filtered image set as shown in the right portion (2) of the Figure 37. However, we can perform the textual and visual searches in different order also depending on the relative importance of visual and semantic features in a query, which make the whole process flexible enough to search for any particular order.

6.6.2 Simultaneous Search

This section presents our simultaneous multi-modal search approach, where contextual and content-based searches are performed simultaneously from the beginning and the results are combined with an adaptive linear combination scheme as described in Section 6.5. Due to the simultaneous nature of the search, user's feedback information can be propagated easily from one modality to another for query refinement and dynamic weight updating. The steps involved in this approach are depicted in Algorithm 7.

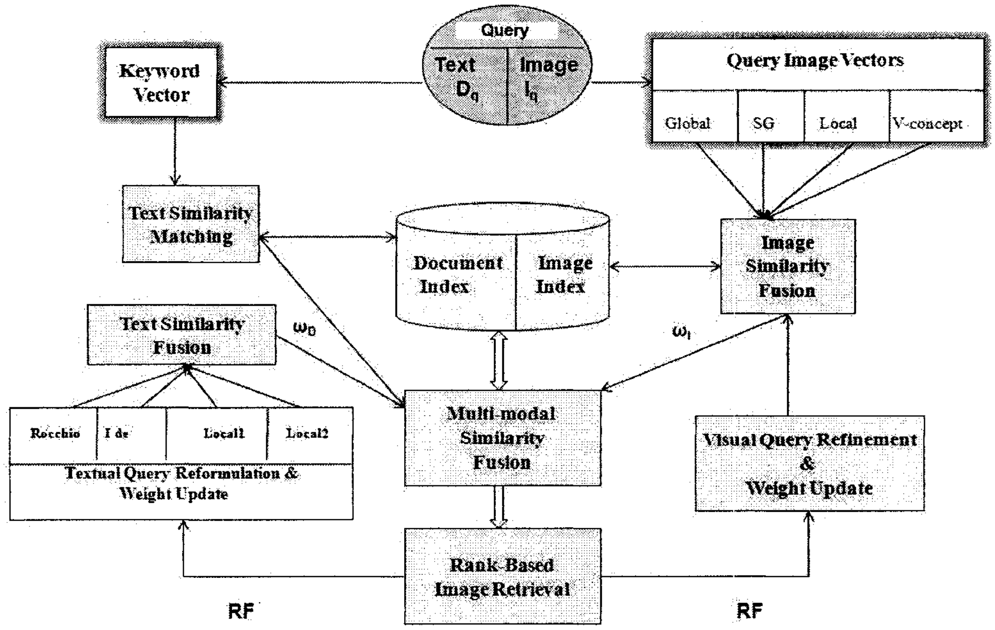


Figure 38: Process flow diagram of the simultaneous search

Algorithm 7 Simultaneous Search Approach

```
Initially, for a multi-modal query $q$ with a document part as $D_q$ and an image part as $I_q$, extract textual query vector as $\mathbf{f}_{D_q}$ and different image feature vectors as $\mathbf{f}_{I_q}^{\text{global}}, \mathbf{f}_{I_q}^{\text{sg}}, \mathbf{f}_{I_q}^{\text{local}}$, and $\mathbf{f}_{I_q}^{\text{V-concept}}$.
Perform a multi-modal search to rank the images based on equation (81), where $\text{Sim}_{\text{D}}(D_q, D_j)$ is initially performed through $\text{Sim}_{\text{text}}(D_q, D_j)$ equation (62) and $\text{Sim}_{\text{I}}(I_q, I_j)$ is performed through equation (82) with initially equal weighting in both inter and intra-modality weights.
repeat
    Obtain user's feedback about relevant images from the top retrieved $K$ images.
    Calculate the modified textual query vectors as $\mathbf{f}_{D_q}^{m}$ and image query vectors as $\mathbf{f}_{I_q}^{m}$ for different query representations.
    Perform the individual context and content-based searches based on Equations (83) and (82).
    Update the inter (e.g., $\omega_I$ and $\omega_D$) and intra-modality (e.g., $\omega^{IF}$ and $\omega^{QF}$) weights based on equation (84) by analyzing individual result lists.
    Finally, combine the similarity scores with the updated weights based on Equation (85) to obtain a final ranked list of images.
until user is satisfied or no more improvement
```

Figure 38 shows the process flow diagram of the above multi-modal simultaneous search approach. Here, the top-middle portion shows how a search is initiated simultaneously based on both text and image parts of a multimodal query and later the individual results are combined for a final ranked result list. Both sequential and simultaneous search processes would be effective when there is a direct relation between annotated keywords and image features. The majority of the multimodal retrieval systems are based on this assumption [115]. Although in reality, we can find many situation where search requirements and image annotations are so abstract that there is practically no relation in between context and content (e.g., text and image) information. In that case, a text-based search might be the only viable option. The positive side is that there also exists many situations where text and images are correlated and we can exploit that information in a better way by performing multimodal searching, either sequentially or simultaneously as described in the previous sections.

6.7 Summary

One of the main drawbacks of the fusion-based (content and context) or multimodal image retrieval research is that it lacks standard benchmarks to evaluate the results. In literature, almost all the systems [125, 124, 119, 123, 121] report better performance for multimodal search compared to using only a single modality, i.e., either text or image. However, these systems use different data sets, different query sets, and there is no standard comparison metrics. A standard image database with a query set and corresponding performance measure model is needed for objective performance evaluation of multimodal image retrieval systems. For a solution, the ImageCLEF [3] retrieval benchmark was established in 2003 with the aim of evaluating multimodal and multi-lingual image retrieval techniques. ImageCLEF provides tasks for both system-centered and user-centered retrieval evaluation within two main areas: retrieval of images from photographic collections and retrieval of images from medical collections. This chapter mainly presents our works in ImageCLEF'06 and ImageCLEF'07 workshops [217, 218] for ad hoc image retrieval of photographic and medical images in benchmark collections. A detailed description of the image collections, query sets, and the results we achieved will be described in Chapter 8.

[3] http://ir.shef.ac.uk/imageclef/

Chapter 7

CBIR as Decision Support System for Dermoscopic Images

In previous chapters, we have presented several image representation and retrieval approaches for both narrow and broad image domains. There also exist some very narrow application domains with specific retrieval objectives. In a decision support based medical retrieval system, a physician mainly search for images according to different disease categories (e.g., category search) in a particular modality, such as looking for CT images of lung with bronchitis or emphysema in ASSERT system [137]. Due to the specific search requirements and very narrow extent of these application domains, specialized image processing and pattern recognition techniques need to be applied and domain knowledge needs to be exploited as much as possible for effective retrieval. This chapter presents a CBIR approach in a specific narrow domain of dermoscopic images as a diagnostic aid to dermatologists for automated melanoma recognition [210]. Malignant melanoma is one of the most common skin cancers in human being in the world [146, 147]. In Canada, an estimated 76,000 new cases of common skin cancers were expected in year 2004 compared to 58,500 new cases for 1994, which is up by 30%, as it was mentioned in the webpage of the Canadian Dermatology Association (CDA) [1]. Detection of malignant melanoma in its early stages considerably reduces mortality, hence this a crucial issue for dermatologists.

In CBIR context, here the ultimate aim is to support decision making by retrieving and displaying relevant images with past cases, either benign or malignant, compared

[1] http://www.dermatology.ca/

to an unknown query image. Dermoscopic images are produced by a technique that allows in vivo microscopic examination of skin lesions. The technique is interchangeably named as dermoscopy, dermatoscopy, epiluminescence microscopy (ELM) and skin surface microscopy [147, 148]. Dermoscopic images have proved to be very effective for early detection of malignant melanoma. Recently, digital imaging and pattern analysis has been found to produce objective and reliable patterns of dermoscopic images of the pigmented skin lesions (PSL) [155, 156, 157]. In clinical practice, several scoring systems and algorithms such as the ABCD rule (ie, asymmetry, border, color, and differential structures), the seven-point checklist, and the Menzies method have been proposed to improve the diagnostic performance of the less experienced clinicians [149, 150]. However, these techniques require formal training and skills in image interpretation and are highly dependent on the subjective judgment. The identification of specific diagnostic patterns related to the distribution of colors and differential dermoscopy structures can better suggest a malignant or benign PSL and demonstrates that computer aided diagnosis can be a very helpful tool, particularly in areas which lack experienced specialists [154, 155].

Todate, most of the work in the dermatology area have focused on the problem of melanoma recognition, in which the likelihood of malignancy is computed based on some feature extraction and classification schemes. A variety of statistical and machine learning approaches to classification of dermoscopic images to melanoma, benign or common and dysplastic nevi are currently available [151, 152, 153, 155, 157]. Digitization of the dermoscopic images after the initial visual assessment permits the storage and use for the comparison when a lesion is being followed over time [156]. As a result, it also makes them a suitable candidate for the application area of CBIR, where it could be used to presents cases that are not only similar in diagnosis, but also similar in appearance and in cases with visual similarity but different diagnoses for better understanding the diseases [128]. In the medical domain, there are already some successful implementation of CBIR as decision support systems for different modalities, such as CT images of lung in ASSERT [137], X-ray images of cervical spine in WebMIRS [138], histological images in I-Browse system [139], or pathological images in IGDS [140] (reviewed in Chapter 2). Although, most of the systems currently available are based on the radiological or pathological domain [128], there is yet no CBIR system developed (as far we know) in the dermatological domain.

Motivated by this, we developed an initial CBIR system for digital dermoscopic images with the aim to make it a decision support system in dermatology domain. By observing the specific characteristics of the dermoscopic images, we proposed a fast segmentation method for the automatic lesion detection as an image pre-processing step [210]. Lesion specific various local color and texture related features are extracted for the image representation and a fusion-based similarity matching function is proposed as a weighted combination of the Bhattacharyya and Euclidean distance measures in color and PCA-based feature spaces.

7.1 Segmentation and Lesion Detection

Detection of the lesion is a difficult problem in dermoscopic images as the transition between the lesion and the surrounding skin is smooth and often it is difficult to notice accurately even for the trained dermatologist [154, 155]. Various image segmentation methods have been proposed in the literature to delineate lesion boundaries from the skin cancer images [158, 159, 154, 155]. In [158], an elaborate method was proposed by reducing the color image to an intensity image and using a double thresholding and a elastic curve fitting techniques to finally detect lesion boundary. The assumption is based on the fact that the majority of the dermoscopic images are captured in a way so the lesion is generally situated close to the center of the image and the background healthy skin surrounds the lesion and more likely to be visible along the periphery of the images. Color changes from the background to a lesion or from a lesion to the background is more important then the color variations within a lesion or in the background. However, their proposed method has three parameters, which might require user interaction to tune the parameters and also require longer processing time. Six different color image segmentation techniques for skin cancer images were compared in [159]. It was found that lowest average error could be achieved by adaptive thresholding and when two or more techniques are combined, the accuracy can be improved further. In accordance with the above observations, we utilize an iterative thresholding based segmentation method [162] by first transforming an image in *RGB* color space to several intensity images and later combine them to detect the lesion as described in following sections.

7.1.1 Intensity Image Generation from HVC Color Space

Many color spaces have been designed to facilitate the color specification and the key issue being considered in selection of a perceptually uniform color space. *HVC* color space comes from Munsell color coordinate system, which is considered for its successful imitation of human color perception [160]. It represents a color in terms of hue (H), which indicates the types of the color; value (V), which tells the total amount of light; and chroma (C) that describes how much white light is mixed with the color (purity). There are several ways to mathematically transform the RGB to the HVC color space. Because $CIEL^*a^*b^*$ color space is known for its good perceptual correspondence and simple computation, RGB values are first transformed into $CIEXYZ$, and then changed to $CIEL^*a^*b^*$, and then altered to HVC values using the following formulas [160]:

$X = 0.607R + 0.174G + 0.201B$
$Y = 0.299R + 0.587G + 0.114B$
$Z = 0.000R + 0.066G + 1.117B$

$L^* = 116(Y/Y_0)^{1/3} - 16, (Y/Y_0) > 0.008856$
$L^* = 903.3(Y/Y_0), (Y/Y_0) \leq 0.008856$
$a^* = 500((X/X_0)^{1/3} - (Y/Y_0)^{1/3})$
$b^* = 200((Y/Y_0)^{1/3} - (Z/Z_0)^{1/3})$

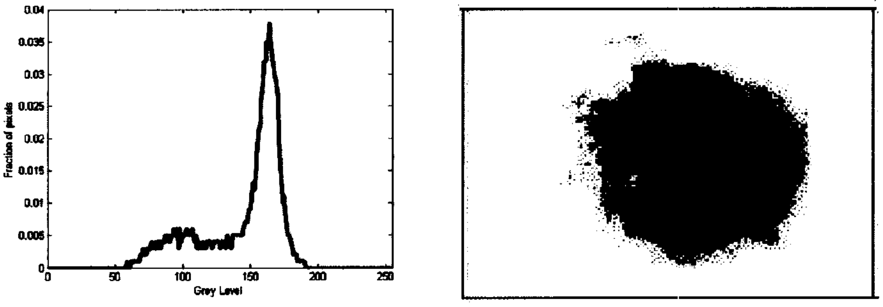
$H = arctan(b^*/a^*)$
$V = L^*$
$C = (a^2 + b^2)^{1/2}$

X, Y, Z are the primaries in XYZ color system, while X_0, Y_0, Z_0 are values of a nominally white object-color stimulus, which are usually chosen to be 0.9642, 1.0 and 0.8249 respectively. If $\Delta H, \Delta V$ and ΔC be the differences of H, V, C color components of an image pixel $A = (H_1, V_1, C_1)$ and its background $B = (H_2, V_2, C_2)$, then *NBS* (National Bureau of Standards) color distance between A and B is defined as follows [161]:

$$E_{\mathrm{NBS}}(A, B) = 1.2 * \sqrt{2C_1C_2\{1 - cos(\frac{2\pi}{100}\Delta H)\} + (\Delta C)^2 + (4\Delta V)^2} \qquad (86)$$

(a) (b)

Figure 39: (a) Original color image (b) Grey level image of the original



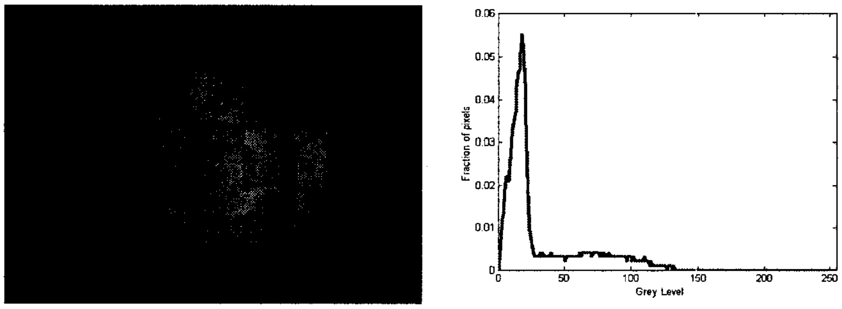
(a) (b)

Figure 40: (a) Grey level histogram (b) Iterative thresholded image

There is a close relation between the human color perception and the NBS color distance, which is shown in Table 1 [161]. Taking advantage of the above properties of the *HVC* space and the *NBS* distance, we transform the image from original *RGB* space to *HVC* space and determine the mean *HVC* values of the pixels from the border (2 pixels wide from each side) of an image. Next, a color image in *HVC* space is transformed into an intensity image in such a way that the intensity at a pixel $f(x, y), 1 \leq x \leq M, 1 \leq y \leq N$ for an image of size $M \times N$ shows the *NBS* color distance of that pixel with the color of the background (e.g., mean of the border). Hence, we obtain an intensity image in which it has higher grey level values in the lesions and lower values in the background after re-scaling as shown in Figure 41(a). We can observe the differences between the grey level histogram in Figure 40(a) of an original image in Figure 39(a) and that of the intensity image in Figure 41(b) generated by the above approach. The later has a more clear separation between the background and foreground pixel intensities with a threshold value of around 27.

Table 1: Correspondense between human color perception and NBS distance [161]

NBS Value	Human Perception
$0 \sim 1.5$	Almost the same
$1.5 \sim 3.0$	Slightly different
$3.0 \sim 6.0$	Remarkably different
$6.0 \sim 12.0$	Very different
$12.0 \sim$	Different color



(a)

(b)

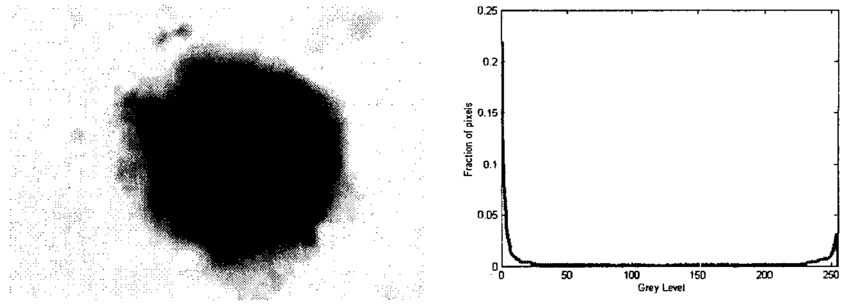
Figure 41: (a) Intensity image generated from HVC space (b) Histogram of the intensity image

Hence, by this transformation, lesion can be distinguished better from the background skin by applying any thresholding methods.

7.1.2 Intensity Image Generation from Fuzzy C-Means Clustering

To take spatial properties of images into account, we also consider another approach to generate an intensity image by utilizing the fuzzy c-means (FCM) clustering [187, 188] in a multi-dimensional feature space. FCM [187] is the most widely used fuzzy clustering algorithm which assigns degrees of membership in several clusters to each input pattern. This algorithm is based on an iterative optimization of a fuzzy objective function [189]:

$$J_{FCM}(\mathbf{U}, V, X) = \sum_{i=1}^{N} \sum_{j=1}^{c} (\mu_{ji}^{m}) D(\mathbf{x}_i, \mathbf{v}_j) \qquad (87)$$



(a) (b)

Figure 42: (a) Intensity image generated from FCM (b) Histogram of the intensity image

where $X = \{\mathbf{x}_1, \cdots, \mathbf{x}_i, \cdots, \mathbf{x}_N\}$ be a finite set of N unlabeled feature vectors, c is the number of clusters and $V = \{\mathbf{v}_1, \cdots, \mathbf{v}_j, \cdots, \mathbf{v}_c\}$ represents the unknown prototype vectors (centroid), where $\mathbf{x}_i, \mathbf{v}_j \in \Re^d$. The fuzzy c-partition is defined by a $c \times N$ matrix $\mathbf{U} = [\mu_{ji}]$ where, μ_{ji} is the membership degree of vector $\mathbf{x}_i$ to the $\mathbf{v}_j$th prototype and satisfies $\mu_{ji} \in [0, 1], \forall j, i$. The distance measurement $D(\mathbf{x}_i, \mathbf{v}_j)$ is in general an Euclidean distance. The parameter m controls the fuzziness of membership of each datum and is usually set larger than 1. The properties $\sum_{i=1}^{c} \mu_{ji} = 1, \forall i \quad and \quad 0 < \sum_{i=1}^{N} \mu_{ji} < N, \forall j$ must be true for $\mathbf{U}$ to be a non degenerate fuzzy c-partition. This iteration will stop when certain termination criterion is met. That is, $\max_{ji}\{|\mu_{ji}^{(k+1)} - \mu_{ji}^{k}\} < \epsilon$ where ϵ satisfies $0 < \epsilon < 1$ and the superscript k denotes iteration number. For clustering, the parameters are set to be $m = 2$ and $\epsilon = 10^{-5}$.

In [189], two eigen-based FCM clustering algorithms are proposed to accurately segment images, which have the same color as the pre-selected pixels. Here, we utilize the FCM with number of clusters is set to 2 to mainly generate an intensity image for latter processing. In this approach, at each pixel, two membership values are determined, where one representing the degree of certainty of a pixel belonging to background normal skin and the other representing the degree of certainty of a pixel belonging to foreground lesion. For input to the FCM, we extract the mean and standard deviation of RGB values as a 6-dimensional feature vector $\mathbf{x}$ in a neighborhood of 5 by 5 pixels around each pixel of an image. After generating the cluster membership values of each pixel based on applying FCM, we only consider the background

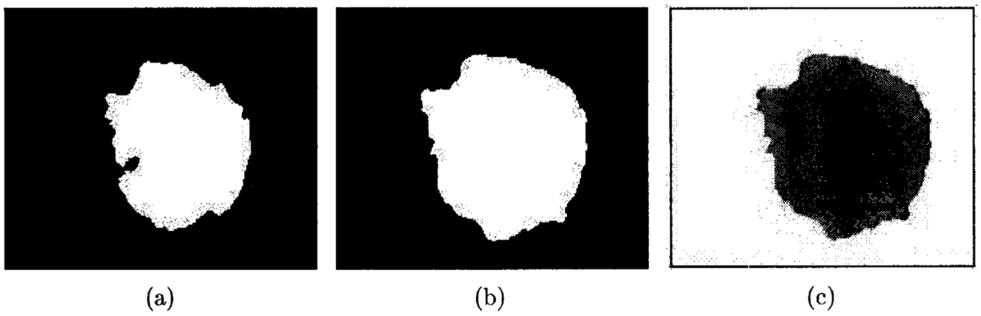

(a) (b) (c)

Figure 43: (a) Thresholded image from the intensity image of Figure 41(a), (b) Thresholded image from intensity image of Figure 42(a), (c) Final segmented image with lesion mask.

membership values, which produces the intensity image after re-scaling as shown in Figure 42(a). Here also we can observe the differences between the grey level histograms of the original image (as shown in Figure 40(a)) and that of the intensity image in Figure 42(b) generated by the above approach. The later has a much deeper and wider valley, which would be more reliable for using any thresholding method. The following section describes the thresholding operation to be performed on the pre-processed intensity images.

7.1.3 Iterative Thresholding & Post-processing

Thresholding is a computationally inexpensive and fast technique for image segmentation, which is suitable for real time application, such as the on-line retrieval in this CBIR approach. However, thresholding simply based on gray-level dermoscopic images is not enough as the transition between lesion and background skin is often very smooth and not clear. Therefore, the above pre-processing operations are performed to increase the distinguishing ability between the lesion and the background skin. Correct threshold detection is crucial for successful segmentation. We utilize the iterative thresholding algorithm as described in [162] by performing the following steps of Algorithm 8. The above algorithm basically uses an iterative clustering approach. An initial estimate of the threshold is made (e.g. mean image intensity). Pixels above and below the threshold are assigned to the foreground and background classes respectively. The threshold value is iteratively re-estimated as the mean of the two class means [162]. After computing the threshold values for the two intensity

Algorithm 8 Segmentation by iterative thresohlding

```
Compute $\mu_B$, the mean intensity level of the border pixels as background
Compute $\mu_F$, the mean intensity level of the center pixels as foreground lesion
$T_{old} = 0$
$T_{\text{new}} = (\mu_B + \mu_F)/2$
while $T_{new} \neq T_{old}$ do
    $\mu_B = \frac{\sum f(x,y)}{N_B}$    for $f(x,y) < T_{new}$ and $N_B =$ Number of background (healthy skin)-pixels
    $\mu_F = \frac{\sum f(x,y)}{N_F}$    for $f(x,y) \geq T_{new}$ and $N_F =$ Number of foreground (lesion)-pixels
    $T_{old} = T_{new}$
    $T_{\text{new}} = (\mu_B + \mu_F)/2$
end while
```

images generated in the pre-processing stage, we obtain the binary images as shown in Figures 43(a) and 43(b). It clearly shows the better separation of the lesion from the background healthy skin in both cases as compared to the image in Figure 40(b). In some cases, the segmentation produces several skin lesion candidates due to the presence of small non-lesion objects. So, a post-processing operation is applied on the binary segmented images to reduce the number of objects based on the morphological operation of opening and closing [163]. Specially, for a binary image, let the lesion be represented by the set X and its background by the set complement X^c. Both the opening $X \circ B \triangleq (X \ominus B) \oplus B$ and closing $X \bullet B \triangleq (X \oplus B) \ominus B$ are derived from the fundamental operations of erosion $\ominus$ and dilation $\oplus$ [163]. Here B is usually called a structuring element and has a simple geometrical shape and a size smaller than the image X. The size of the structuring element in both the cases is chosen as a square of 5×5 pixels. This size is good enough for removing small isolated objects or filling the holes in this context. Both opening and closing operations act as nonlinear filters that smooth the contours or lesion border of the input image and tiny artifacts or holes are removed from background and lesion images. After detecting the lesion masks from the segmented images, a simple union (OR) operation is applied to obtain the final lesion mask as shown in Figure 43(c). Usually the largest object is the skin lesion and is thus selected for further processing for feature extraction.

7.2 Lesion-Specific Feature Extraction

Feature extraction from skin lesion is another very important step to develop an effective retrieval system for dermoscopic images. Many feature selection and extraction strategies have been proposed [155, 156, 157] from the perspective of classification of images as malignant or benign. These features are calculated, which attempt to reflect the parameters used in medical diagnosis, such as *ABCD* rule or more advanced features. These features are certainly effective for the classification purposes, as seen from the performances of some classification-based systems in this domain [154, 155, 157]. However, features good for classification or distinguishing one disease from another, may not be suitable for the presentation. In a retrieval system, we are looking for similar images in terms of color, texture, shape etc. By selecting and extracting good representative features, we may be able to identify images similar to an unknown query image, whether it belongs to the same disease group or not.

In this direction, suitable local color feature in the form of a feature vector is extracted by considering the mean or average color of the lesion in HVC color space and variance-covariances of the color channels by estimating the covariance matrix. If average color feature $\mathbf{f}^{\mathrm{avg}}$ of a lesion is represented as $\mathbf{f}^{\mathrm{avg}} = [\mu_{\mathrm{H}}, \mu_{\mathrm{V}}, \mu_{\mathrm{C}}]^{\mathrm{T}}$, where μ_{H}, μ_{V} and μ_{C} is the average H, V and C values in HVC space, then the cross-correlation among color channels due to the off diagonal of the 3×3 covariance matrix Σ_j of the lesion of I_j is estimated as

$$\Sigma_j = \frac{1}{N_j - 1} \sum_{k=1}^{N_j} (\mathbf{f}^{x_j} - \mathbf{f}^{\mathrm{avg}})(\mathbf{f}^{x_k} - \mathbf{f}^{\mathrm{avg}})^T \tag{88}$$

where $\mathbf{f}^{x_k}$ is the color vector of pixel x_k and N_j is the number of pixels of the lesion of I_j. Since the covariance matrix is symmetric, only 6 values of it need to be stored in the feature vector for later similarity matching based on a Bhattacharyya distance measure [81].

In addition, local texture features are extracted from the grey level co-occurrence matrix (GLCM) [64] (similar feature extraction is described in Section 3.2.3 of Chapter 3). Higher order features, such as energy, maximum probability, entropy, contrast and inverse difference moment are measured based on each GLCM to form a five dimensional feature vector and finally obtained a twenty dimensional feature vector

$\mathbf{f}^{\text{texture}}$, by concatenating the feature vector of each GLCM. We also uniformly quantized the HVC space into 12 bins for hue (each bin consisting of a range of 30°), 3 bins for the value and 3 bins for the chroma, to generate a 108-dimensional color histogram feature vector $\mathbf{f}^{\text{color}}$. Finally, the color histogram and texture moment based feature vectors are normalized to the zero mean and unit variance and combined or concatenated to form a single vector. Since, the dimension of the combined feature vector is large enough (e.g., 108 for color and 20 for texture for a total of 128) to contain redundant information, we applied principal component analysis (PCA) [183, 184] based dimension reduction technique (we also used the same technique in Section 3.4.1 of Chapter 3) to reduce the feature dimension. The feature vector with reduced dimension is called as $\mathbf{f}_j^{\text{pca}} \in \Re^m$ for an image I_j. For the experimental purpose, the dimensionality of the combined color and texture feature vector is reduced to $m = 10$ from $d = 128$ dimension by applying the PCA, where the first 10 eigenvalues related to the 10 principal components (PC's) account for 99.9% of the total variances. The advantage of using a smaller subset of eigenvectors is that it could increase the retrieval speed when the large image databases are searched for.

7.3 Similarity Matching

For similarity matching based on color covariance-based feature, the distance between the regions or segmented lesions of query image I_q and database image I_j is computed by way of the Bhattacharyya distance metric as follows [172]:

$$\mathrm{D}_{\mathrm{Bhatt}}(I_q, I_j) = \frac{1}{8}(\mathbf{f}_q^{\text{avg}} - \mathbf{f}_j^{\text{avg}})^T \left[\frac{(\Sigma_q + \Sigma_j)}{2}\right]^{-1}$$

$$(\mathbf{f}_q^{\text{avg}} - \mathbf{f}_j^{\text{avg}}) + \frac{1}{2} \ln \frac{\left|\frac{(\Sigma_Q + \Sigma_I)}{2}\right|}{\sqrt{|\Sigma_q||\Sigma_j|}} \tag{89}$$

where $\mathbf{f}_q^{\text{avg}}$ and $\mathbf{f}_j^{\text{avg}}$ are the average color feature vectors, and Σ_{q} and Σ_{j} are the covariance matrices of the lesions of image I_q and I_j respectively. Equation (89) is composed of two terms, the first one being the distance between feature vectors of image regions, while the second term gives the class separability due to the difference between covariance matrices.

Euclidean distance measure is used for comparing feature vectors of I_q and I_j in

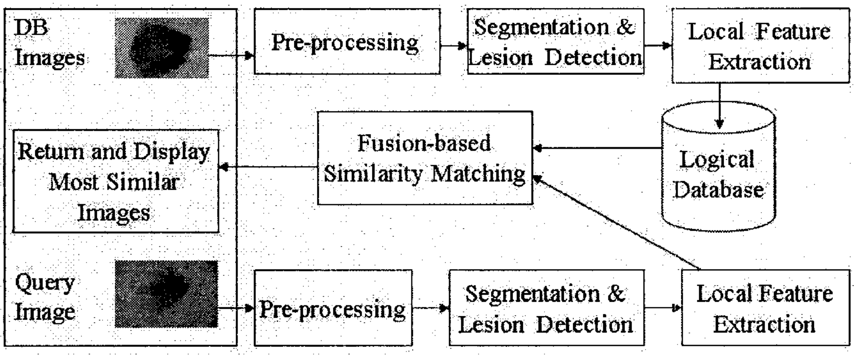

DB Images
Pre-processing
Segmentation & Lesion Detection
Local Feature Extraction
Return and Display Most Similar Images
Fusion-based Similarity Matching
Logical Database
Query Image
Pre-processing
Segmentation & Lesion Detection
Local Feature Extraction

Figure 44: Block diagram of the CBIR system.

PCA sub-space as

$$\mathrm{D}_{\mathrm{Euclidean}}(I_q, I_j) = ||\mathbf{f}_q^{\mathrm{pca}} - \mathbf{f}_j^{\mathrm{pca}}|| = \sqrt{\sum_{i=1}^{m} (f_q^{\mathrm{pca}_i} - f_j^{\mathrm{pca}_i})} \tag{90}$$

For the above distances, the following function is used to transform the distance measures into a similarity measures as $\mathrm{S}(I_q, I_j) = \exp^{-\mathrm{D}(I_q,I_j)/\sigma_{\mathrm{D}}}$, where σ_{D}^2 is the distance variance computed for each distance measure separately over a sample image set. After the similarity measures of color and feature in PCA-subspace are determined as $\mathrm{S}_{\mathrm{Bhatt}}(I_q, I_j)$ and $\mathrm{S}_{\mathrm{Euclidean}}(I_q, I_j)$, we aggregate them into a single similarity matching function as follows:

$$\mathrm{S}(I_q, I_j) = w_{\mathrm{Bhatt}}\mathrm{S}_{\mathrm{Bhatt}}(I_q, I_j) + w_{\mathrm{Euclidean}}\mathrm{S}_{\mathrm{Euclidean}}(I_q, I_j) \tag{91}$$

Here, w_{Bhatt} and $w_{\mathrm{Euclidean}}$ are non-negative weighting factors with normalization $(w_{\mathrm{Bhatt}} + w_{\mathrm{Euclidean}} = 1)$, which needs to be selected experimentally. Figure 44 shows the block diagram of the proposed CBIR approach. When a query image is submitted, pre-processing, segmentation and feature extraction are performed the same way as database images as shown in the bottom level of Figure 44. The features of query and database images are then matched in a similarity retrieval subsystem. The match score is compared and sorted, where the top K matches are shown to the query interface according to their ranks.

7.4 Summary

In this chapter, we proposed and developed a CBIR based system as a diagnostic aid for melanoma recognition. The system is evaluated for the retrieval of dermoscopic images of pigmented skin lesions. The experimental results, as provided in Chapter 8, indicate that the proposed CBIR approach is effective to retrieve visually similar lesions from a database compared to an unknown query image. From the image retrieval context, we conjecture that by presenting images with known pathology that are visually similar to the image being evaluated, it may provide a more intuitive aid to the dermatologist, potentially leading to improvement in their diagnostic accuracy. However, it is recognized that many other advanced image-based features and features from other sources in the form of a case or lab reports would be necessary towards a complete decision support system. The presence of an expert dermatologist is considered necessary for the overall visual assessment of skin lesions and for the final diagnosis based on objective evaluation suggested by the system and contextual information from lab report, such as histopathological tests.

Chapter 8

Retrieval Evaluation

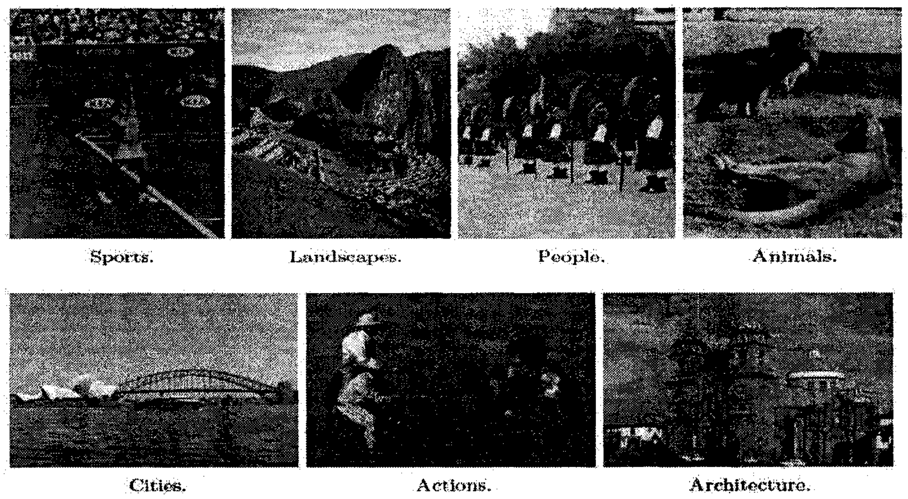
This chapter presents the detailed empirical analysis of the proposed retrieval techniques described in this thesis. Specially, we present the data sets used for the experiments, experimental settings, accuracy comparisons of classification and retrieval, and analysis of results. This chapter is organized in such a way that in different sections we present the experiments and result analysis that correspond to the techniques proposed in different chapters (Chapters 3-7) of this thesis. In the following section, we at first present the different image collections used for the experiments.

8.1 Image Collections

Several image collections from natural photographic and medical domains have been used to test the retrieval effectiveness from different perspectives and compare the techniques with existing well known image representation and retrieval methods.

8.1.1 The General Photographic Collection

The photographic image collection is the IAPR TC-12 benchmark created under Technical Committee 12 (TC-12) of the International Association of Pattern Recognition (IAPR) [54]. This collection is publicly available for research purposes and currently contains around 20,000 photos taken from locations around the world that comprises a varying cross-section of still natural images. It is currently used for ad-hoc photographic retrieval task in ImageCLEF [57]. Few example images of this collection are shown in Figure 45. The size of the images are in 360 × 480 or 480 × 360 pixels in



Sports. Landscapes. People. Animals.

Cities. Actions. Architecture.

Figure 45: Example images of the Photographic collection [54]

this collection. The domain of the images in this collection is very generic that covers a wide ranges of daily life situations. Unlike the commonly used COREL [1] images, this collection is very general in content with many different images of similar visual content, but varying illumination, viewing angle and background. Thus making it more challenging for successful application of image retrieval techniques. Due to the growth of the desktop search market and popularity of tools such as FlickR [2], this type of collection is likely to become of increasing interest to researchers. Each image in this collection has a corresponding semi-structured caption consisting of the fields as a unique identifier, a title, a free-text description of the semantic and visual contents of the image, notes for additional information, the provider of the photo and fields describing where and when the photo was taken. This annotation information helps us to perform experimental evaluation of our multimodal retrieval approach as presented in Chapter 6.

8.1.2 The Medical Collection

The medical collection contains around 66,000 images of different modalities with annotations in XML format in three different languages. This collection consists of six

[1] http://www.corel.com
[2] http://www.flickr.com

Table 2: Individual Databases used in Medical collection

Collection	#Images	#Cases	#Annotation
Casimage	8725	2076	2076
MIR	1177	407	407
PEIR	32,319	32,319	32,319
PathoPIC	7805	7805	15,610
myPACS	15,140	3577	3577
Endoscopic	1496	1496	1496
Total	66,662	47,680	55,485

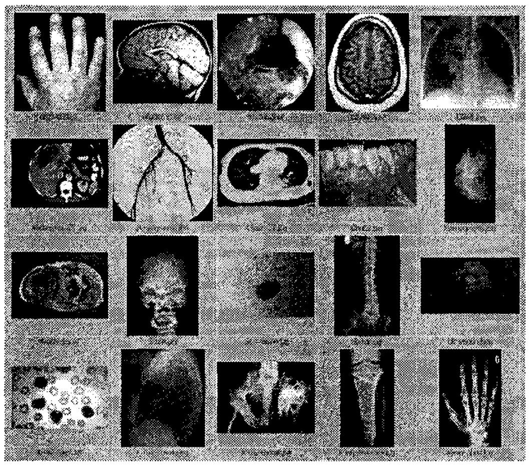


Figure 46: Example images of the Medical collection.

different data sets and used for medical image retrieval task in ImageCLEF [57, 58]. Hence, all the data sets are made available by the organizers of the CLEF [3]. The detailed statistics of each individual collection is shown in Table 2. The Casimage [4] data set represents images of mostly of radiology modalities, along with some pathological images, photographs, and illustrations. The PEIR [5] (Pathology Education Instructional Resource) data set contains pathology and radiology images and each image has an associated annotation file. The nuclear medicine data set of Mallinkrodt Institute of Radiology (MIR) [6] contains images mainly from nuclear medicine with

[3] http://www.clef-campaign.org/

[4] http://www.casimage.com/

[5] http://peir.path.uab.edu/

[6] http://gamma.wustl.edu/home.html

annotations provided per case basis. The PathoPic [7] collection contains only pathology images with extensive annotation on a per image basis in English and German. The myPACS [8] data set contains radiology images and the Endoscopic [9] data set consists of only endoscopic images with an English annotation per image basis [58]. As the entire collection contains variety of image data sets, imaging modalities, image sizes, and resolutions, it makes really difficult to perform semantic search based on current CBIR techniques. Few example images of different modalities in the medical collection are shown in Figure 46.

8.1.3 The Radiograph Image Collection

This collection contains 10,000 radiographs grouped into 116 categories, which is made available by the IRMA (Image Retrieval in Medical Applications) group from the University Hospital, Aachen, Germany [141]. The images are in grey level and PNG (Portable Network Graphics) format. All the images are classified manually by reference coding with respect to a mono-hierarchical coding scheme, which makes this collection distinctive from all other collections. In this scheme, the technical code (T) describes the imaging modality or technique used, the directional code (D) denotes the body orientation, the anatomical code (A) refers to the body region examined, and the biological code (B) describes the biological system examined. The entire code results in a character string of not more than 13 characters (TTTT-DDD-AAA-BBB). Based on this code, 116 distinct categories are defined [141]. The images have a high intra-class variability and inter-class similarity, which make the classification and retrieval task more difficult. For example, Figure 47 shows that a great deal of intra-class variability exists in images of category label 3, mostly due to the illumination changes, small amounts of position and scale differences, and noise. On the other hand, Figure 48 exhibits an example of inter-class similarities between two different categories. Images in the upper row of Figure 48 belong to category label 52 ("1121-127-700-500" in IRMA code), whereas images in the lower row belong to category label 5 ("1121-115-700-400" in IRMA code). Although the images in both categories are hard to distinguish with an untrained eye, they differ in orientations (anteroposterior vs. posteroanterior) and biological systems (uropoietic vs. gastrointestinal). The

[7] http://alf3.urz.unibas.ch/pathopic/intro.htm

[8] http://www.mypacs.net/

[9] http://www.cori.org

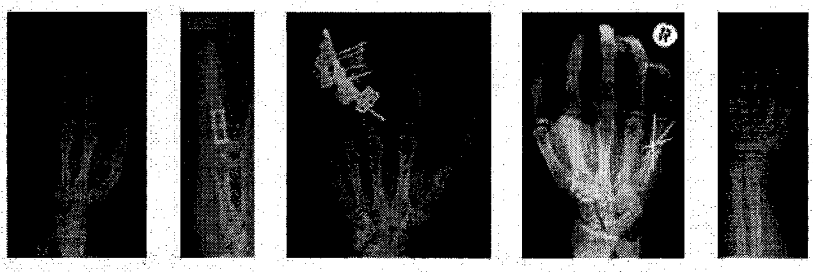


Figure 47: Intra-class variability within the category label 3, (annotated as "X-ray, plain radiography, coronal, upper extremity (arm), hand, musculosceletal system")

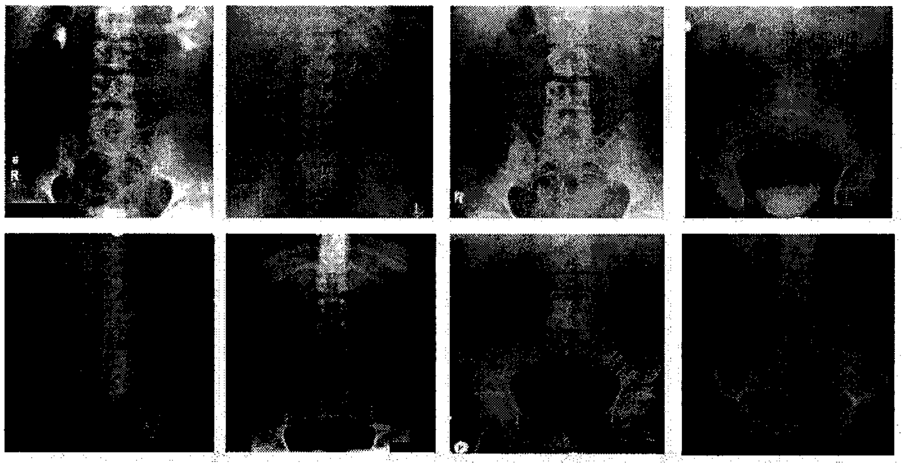


Figure 48: Inter-class similarity between category labels 52 and 5 (Radiograph Collection))

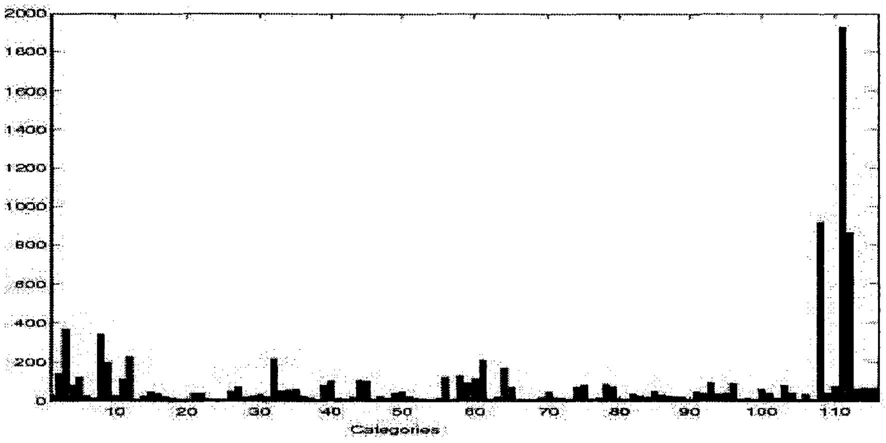


Figure 49: Frequency of images in 116 categories (Radiograph Image Collection)

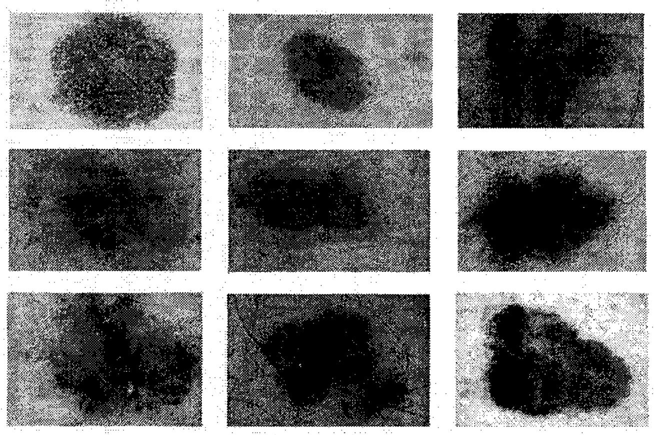


Figure 50: Sample dermoscopic images of all three categories

detailed classification scheme of this collection can be found in [141]. Figure 49 shows the number of images in each category in this collection. The images in the categories are not uniformly distributed, such as category 111 has 1927 images, whereas four categories only have 10 images. It makes both the classification and retrieval tasks more difficult compared to many experimental collections with less number of image categories and more or less uniform distribution of images [128]. Currently, this collection is utilized for medical image annotation task in ImageCLEF [58].

8.1.4 The Dermoscopic Image Collection

This collection contains 358 dermoscopic images pigmented skin lesions that are obtained from two dermatology image atlases [148, 164]. The images are classified to three different categories: benign or common nevi (106 images), dysplastic nevi (118 images) and melanoma (134 images). In this collection, the images are captured in such a way that majority of the pigmented skin lesions are located in the central portion as shown in Figure 50. Since, images are collected from two different data sets and are captured by different devices under different conditions, it makes both retrieval and classification tasks even harder. Figure 50 shows example images of all three categories from the collection, where image in the first, second and last rows belong to benign nevi, dysplastic nevi, and melanoma respectively. As can be seen from the figure, it is hard to distinguish with untrained eyes whether an image is benign or malignant.

8.2 Performance Measure

A major problem in CBIR is the lack of a common performance measure which allows the researcher to compare different image retrieval systems in a quantitative and objective manner. Several measures for evaluating the performance of CBIR have been proposed in [204]. For example, the precision-recall graph, the rank of the first retrieved relevant image, the average normalized rank, the precision after 20, 50, and N images retrieved, the recall at the point where precision is 0.5, the recall after 100 images retrieved, and so on. The performance measures assume a ground truth notion of relevancy, i.e., every image should known to be either relevant or non-relevant to a particular query. The best-known and most widely used measures of retrieval efficiency in text retrieval as well as in CBIR are *precision* and *recall* [204, 44]. Precision is the ratio of the number of relevant images returned to the total number of images returned and *Recall* is the ratio of the number of relevant images returned to the total number of relevant images. So, when the top N images are considered and there are Q relevant images, the precision within top N images is defined to be $Precision(N) = Q/N$, whereas recall be $Recall(N) = Q/R$. Here, R be the number of all images that are relevant to the query image. Both precision and recall are insufficient measures when used alone. Precision is always higher if we

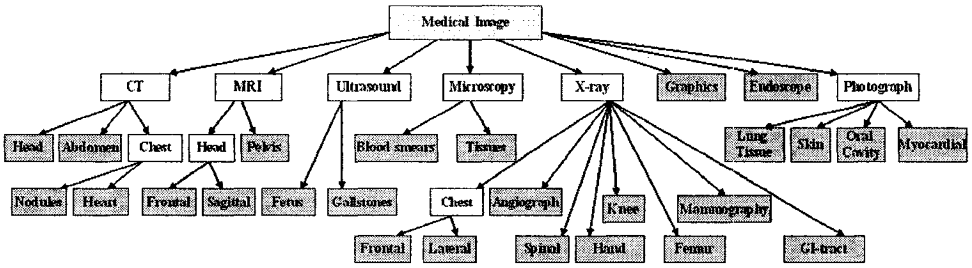


Figure 51: Classification structure of Med-DB.

consider only few retrieved images. On the other hand, we can make always recall one by retrieving all the images in a database. As a result, precision and recall are used together and represented as a precision-recall (PR) graph, in which precision values are plotted against different values of recall. In our experiments, the precision-recall graph are used as the main performance evaluation measure for the retrieval techniques.

8.3 Performance Evaluation of Global Concept-Based Retrieval

This section presents the experiments and results analysis of the global concept-based image retrieval framework described in Chapter 3. The proposed retrieval techniques of the framework are classification-driven at a global label. Hence, we need an application domain or image collection where images are semantically organized and where domain knowledge can be exploited by applying learning-based techniques. We performed exhaustive experiments in such two different medical image collections with known categories to evaluate the retrieval effectiveness. The performance measures on two different collections would provide us a fair analysis of the proposed approaches as the characteristic of images varies in between the collections. The first collection (we call it *MED-DB*) contains around 4000 biomedical images of 26 disjoint categories. This collection is basically a subset of the larger medical collection [58] as described in Section 8.1.2. In this collection, the images are manually classified into three levels based on their annotation information. In the first level, images are categorized according to the imaging modalities (e.g., X-ray, CT, MRI, etc.), at the next level,

Table 3: CV Error Rate (MED-DB training set)

Image Feature	Kernel	C	γ	Error Rate (%)
EHD	RBF	200	.005	11.22
CLD	RBF	200	.007	12.52
Avg.-Grey	RBF	100	.05	14.76
Moment	RBF	100	.09	13.16

Table 4: CV Error Rate (IRMA-DB training set)

Image Feature	Kernel	C	γ	Error Rate (%)
EHD	RBF	100	.0015	23.06
CLD	RBF	10	.0025	25.75
Avg.-Grey	RBF	20	.05	26.49
Moment	RBF	100	.0015	24.60

images in each of the modalities are further classified according to the examined body parts (e.g., head, chest, knee, etc.) and finally it is further classified by orientation (e.g., frontal, coronal, sagittal, etc.) and/or distinct visual observations (e.g. ultra sound with gallstones, CT images with nodules) as shown in Figure 51. The disjoint categories as global concepts are selected from the leaf nodes (grey in color). Images in this collection are in both grey-level (e.g., X-ray, CT) and color (e.g., microscopic slides, photographs) with different sizes and resolutions. We used the Radiograph collection of 10,000 grey level images (we call it *IRMA-DB*) as described in Section 8.1.3 for our second experimental data set. This data set is suitable for experimental evaluation, since the images are already pre-classified into 116 distinct categories or global concepts.

8.3.1 Training of SVMs

To generate the global concept models based on classifier learning, we need to create a training set of images with all categories that can reasonably represent the corresponding collection. Hence, for training of the multi-class SVMs, 40% images of each collection are manually and carefully selected as training sets such that they contain images of all categories of the actual image sets. The remaining of the 60% images are considered as test sets for both collection for measuring classification and retrieval accuracies.

We used the radial basis function (RBF), $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma||\mathbf{x}_i - \mathbf{x}_j||^2), \gamma > 0$ as the kernel, since recent work shows that it works well when the relation between class labels and feature attributes is nonlinear [92]. There are two tunable parameters while using RBF kernels : C and γ. The kernel parameter γ controls the shape of the kernel and regularization parameter C controls the trade-offs between margin maximization and error minimization. Increasing C may decrease training error, but it can also lead to poor generalization. It is not known beforehand which C and γ are the best for the classification problem at hand and are selected by cross-validation (CV). In the training stage, the goal is to identify the best (C and γ), so that the classifier can accurately predict testing data. For training set, a 10-fold cross-validation (CV) is conducted, where we first divide the training set into 10 subsets of equal size. Sequentially one subset is tested using the classifier trained on the remaining 9 subsets. Thus, each instance of the whole training set is predicted once so the cross-validation accuracy is the percentage of data which are correctly classified. Basically pairs of (C,γ) are tried and the one with the best cross-validation accuracy or the lowest error rate is picked. The best values of the parameters C and γ that are obtained for the different feature representations are shown in Tables 3 and 4 for the MED-DB and the IRMA-DB training sets respectively in terms of lowest error rates (e.g. converse of accuracy). The error rates are higher for the IRMA-DB as shown in Table 4 due to the presence of large number of categories, intra-class variability, and inter-class similarity of images as described in Section 8.1.3. After finding the best values of the parameters C and γ for each classifier with distinct input, we utilize them to generate the final SVMs model files for latter prediction purpose. For SVMs-based classification, we utilized the LIBSVM tool [182].

The test sets are used to generate the concept-based feature vectors as described in Section 3.3 of Chapter 3 for retrieval evaluation. In following, we at first present the error rates in the test sets of individual classifiers as well as error rates when they are combined with different combination rules. After that, we show the analysis of retrieval accuracies in terms of PR graphs.

Classification Accuracies (Global Concept):

The accuracies of the classifiers are measured in terms of error rate, which is the proportion of number of images misclassified to total number of images in a test set. The error rates of the test sets for the individual classifiers with low-level features as inputs is shown in Table 5. We observe that classifiers with EHD-based feature of

Table 5: Error Rate of the Individual Classifiers (test sets)

Test Set	EHD	CLD	Avg.-Grey	Moment
MED-DB	13.20%	14.35%	16.09%	14.41%
IRMA-DB	24.40%	26.25%	28.05%	25.90%

Table 6: Error Rate of Different Classifier Combinations (test sets)

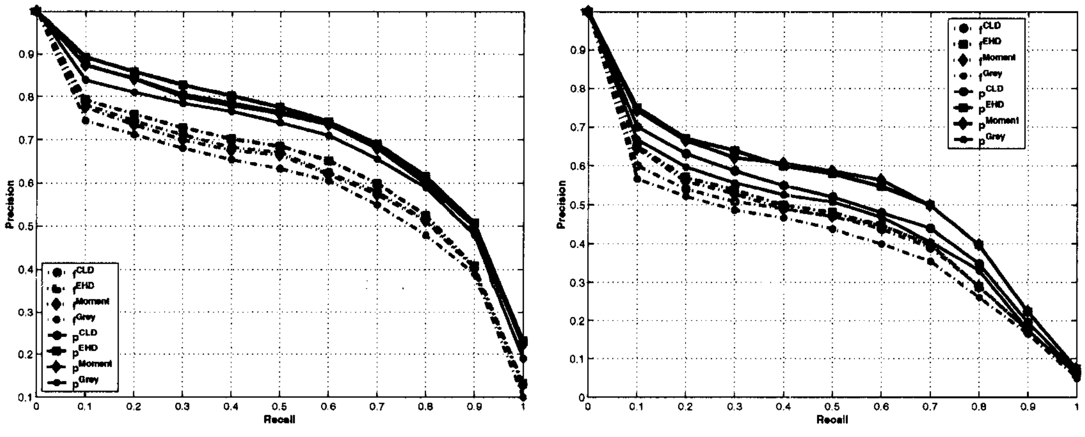
Test Set	Prod	Sum	Max	Min	Med
MED-DB	10.70%	10.05%	12.06%	14.05 %	13.75%
IRMA-DB	21.3%	20.20%	24.05%	26.45%	23.50%

MPEG-7 [62] as inputs comparatively performed better in both test sets, since they achieved the lowest error rates (e.g., 13.20 % for MED-DB and 24.40% for IRMA-DB). It conforms the importance of edge-based features to distinguish images in diverse medical collections.

The classification performances of the test sets show significant improvements as shown in Table 6, when the classifier combination rules are applied as described in Section 3.3.1 of Chapter 3. We can observe that there is an improvement in accuracy around 3% for MED-DB test set, when the SVMs with the lowest error rate (e.g., 13.20%) is compared to the lowest error rate among the combination rules (e.g., 10.05%). For IRMA-DB, we achieved around 4% accuracy improvement in a similar fashion. The classifier combination improves the results because each of the single SVM classifier evaluates different aspects of the image representation. In general, the product and sum rules showed better effectiveness in both collections compared to other rules. This conforms to the fact that representations used are conditionally statistically independent. The classifier results are therefore rather uncorrelated and complementary in nature.

Retrieval Accuracies (Global Concept):

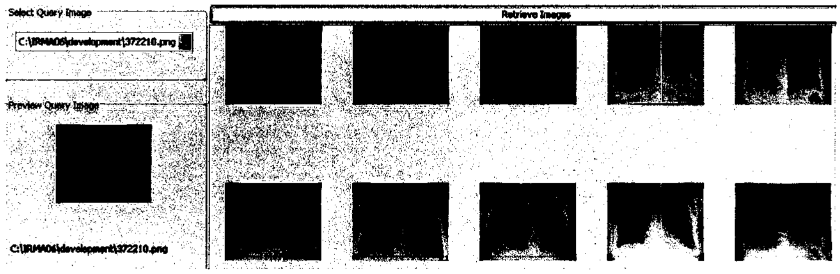
For a quantitative evaluation of the retrieval results, we selected all the images in the test sets as query images and used *"query-by-example"* as the search method where a query is specified by providing an example image to the system. A retrieved image is considered to be a correct match if it belongs to the same query image category. We at first evaluated, whether the proposed global concept-based feature representation schemes can achieve any improvements when compared to the low-level features in terms of retrieval accuracy. Figure 52 shows the precision-recall (PR) graphs based



(a) MED-DB

(b) IRMA-DB

Figure 52: Accuracy comparison of global concept and low-level feature spaces



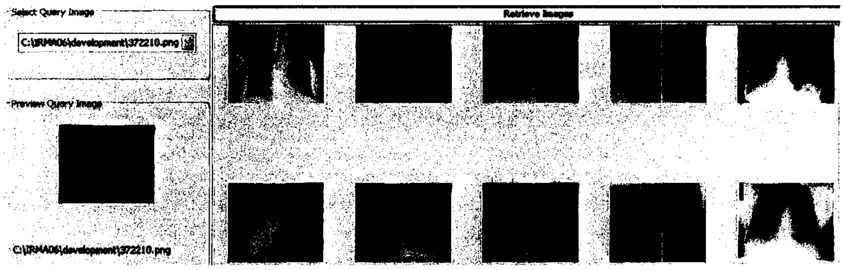
C:\IRMA06\development\372210.png

C:\IRMA06\development\372210.png

Figure 53: A snapshot of the retrieval result based on $\mathbf{f}^{\mathrm{EHD}}$

on retrieval results of different concept and low-level feature spaces for both MED-DB and IRMA-DB collections. As shown in Figure 52, there are clearly visible large gaps in performances in between low-level and concept-based feature spaces. For example, by comparing the curves in between the low-level EHD ($\mathbf{f}^{\mathrm{EHD}}$) and the concept-based feature that is transformed from EHD by SVMs prediction ($\mathbf{p}^{\mathrm{EHD}}$), we can see there are around 5-8% increase in precision at each recall level for both collections. Here, $\mathbf{f}^{\mathrm{EHD}}$ and $\mathbf{p}^{\mathrm{EHD}}$ are represented by dashed and solid blue color curves as shown in Figure 52. A similar trend of improvements are observed in Figure 52 for other concept-based feature spaces when compared to their corresponding low-level feature representations. Such results are expected as the proposed concept-based features retain better semantic categorization information in their representations when compared to their counterpart low-level features.

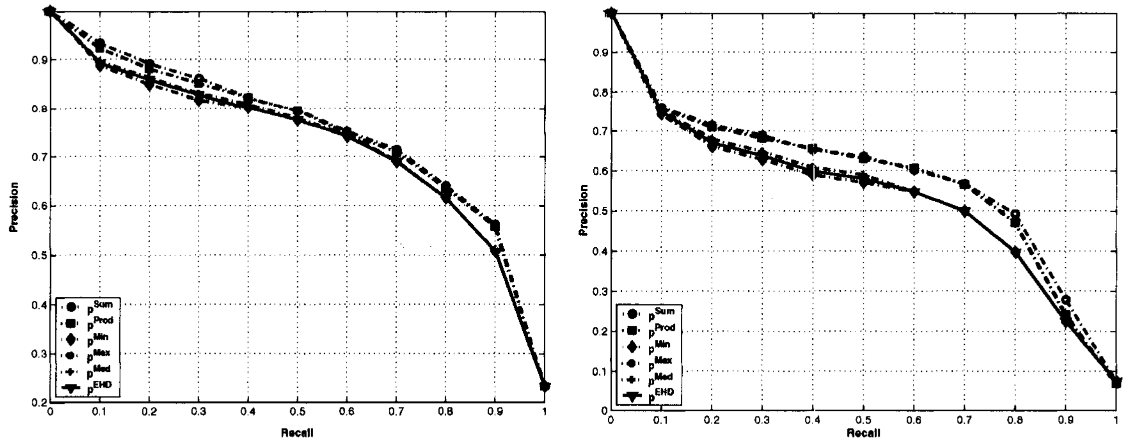
For a qualitative evaluation of the performances, Figure 53 and Figure 54 show



C:\IRMA06\development\372210.png

C:\IRMA06\development\372210.png

Figure 54: A snapshot of the retrieval result based on $\mathbf{p}^{\mathrm{EHD}}$



(a) MED-DB

(b) IRMA-DB

Figure 55: Accuracy comparison of concept-based feature fusion approaches

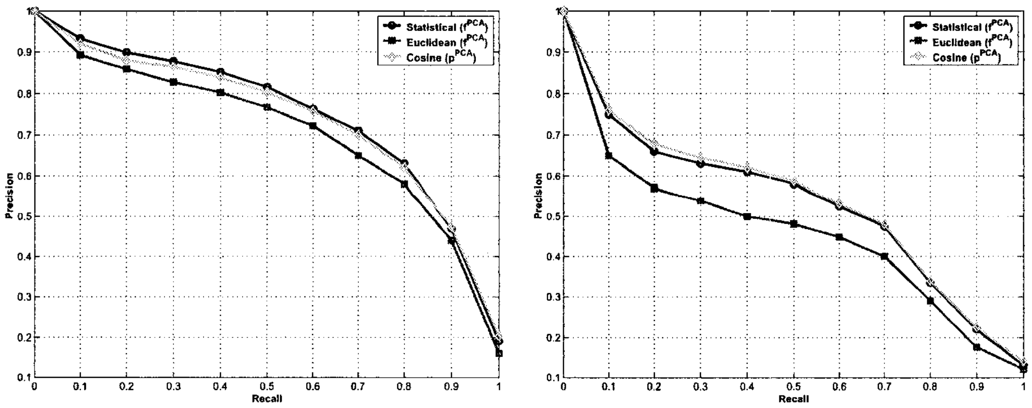
the snapshots of the retrieval result for a query image in IRMA-DB. In Figure 53, for a query image of chest X-ray (the image in the left panel) that belongs to the category label 111, the system returns 5 images of the same category out of the top 10 images by applying the Euclidean similarity measure on $\mathbf{f}^{\mathrm{EHD}}$. The 5 relevant images are located in rank position 1, 3, 5, 7 and 9 where ranking goes from left to right and from top to bottom. On the other hand, as shown in Figure 54, the system returns 7 (in position 1-4, 6, 8, and 9) images from the same category based on the global concept-based feature, $\mathbf{p}^{\mathrm{EHD}}$. In both cases, the irrelevant returned images are from category label 108, where the main difference between these two categories is in the orientation (e.g., anteroposterior (AP) vs. posteroanterior (PA)). There is thus clear improvement in performance in the concept feature space for this particular query in terms of finding images of the correct categories.

Figure 55 shows the PR graphs for both collections based on the concept-based feature fusion approaches as described in Section 3.3.1 of Chapter 3. Like the improved classification accuracies as achieved by majority of the combination rules, here also we can observe improved performances in terms of precision-recall in the corresponding feature spaces. For example, Figure 55 shows around 3-5 % increase in precisions at majority of the recall levels for the fusion based concept feature spaces, $\mathbf{p}^{\text{Prod}}$ and $\mathbf{p}^{\text{Sum}}$ when they are compared to $\mathbf{p}^{\text{EHD}}$, i.e., the best performed concept feature without any fusion. This improvement in performances conforms to the improvement in classification accuracies based on applying combination rules as shown in Table 6. Hence, the retrieval performances are closely related to the classification accuracies in general. To perform statistical similarity matching-based retrieval as described in Section 3.4 of Chapter 3, category-specific parameters (e.g., mean and covariance matrix) were estimated from low-dimensional image features of the same training sets used for training the SVMs (Section 8.3.1). Before performing dimension reduction based on PCA, each low-level image features were normalized with zero mean and unit variance to transform feature attributes to the same scale and after that features are concatenated to form a combined feature vector [80]. For the retrieval evaluation, we consider a composite feature vector based on the combination of CLD, EHD, and Moment-based feature descriptors as described in Section 3.2 of Chapter 3. Here, CLD with 10 Y, 3 Cb and 3 Cr coefficients were extracted to form a 16-dimensional feature vector for the MED-DB collection and CLD with only 64 Y is extracted to form a 64-dimensional feature vector for the IRMA-DB collection as it contains only grey level images. CLD with these dimensions was also used for the experiments described previously. The dimensionality of the combined feature vectors in both collections are reduced from $\Re^d \rightarrow \Re^n$ in such a way that the n largest eigenvalues account for 99.0 % of total variance. In this way, we obtained 16-dimensional feature vectors for images in MED-DB and 20-dimensional vector in IRMA-DB images in a PCA subspaces as $\mathbf{f}^{PCA}$.

For SVMs training, we used the low-dimensional feature vectors as inputs with corresponding category labels in both training sets. The RBF kernel is also utilized here and after 10-fold cross validation, the best values of C and γ were determined as shown in Table 7. These values were finally used to generate the SVMs model files for latter on-line predictions. Form Table 7, we can observe that CV error rates in low-dimensional PCA subspaces for both training sets are close to the lowest error

Table 7: CV Error Rates in PCA subspace (training sets)

data set	Kernel	C	γ	Error rate
MED-DB	RBF	100	.0025	11.96%
IRMA-DB	RBF	200	.002	23.90%



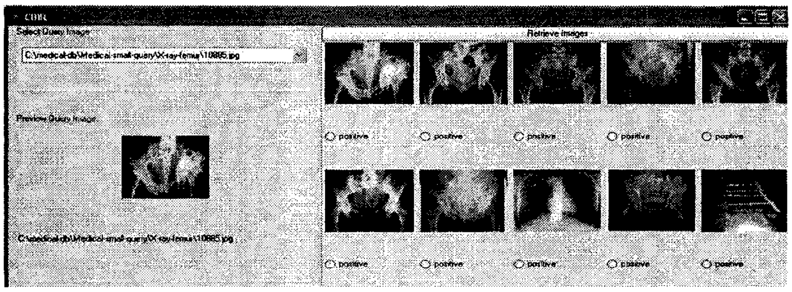
(a) MED-DB

(b) IRMA-DB

Figure 56: Accuracy comparison of statistical, Euclidean and cosine similarity matching

rates in Table 3 and Table 4. It demonstrates that there were enough redundancies among feature attributes in the original feature spaces. Moreover, the low dimensional feature vectors are computationally efficient for similarity matching as well reducing logical database size.

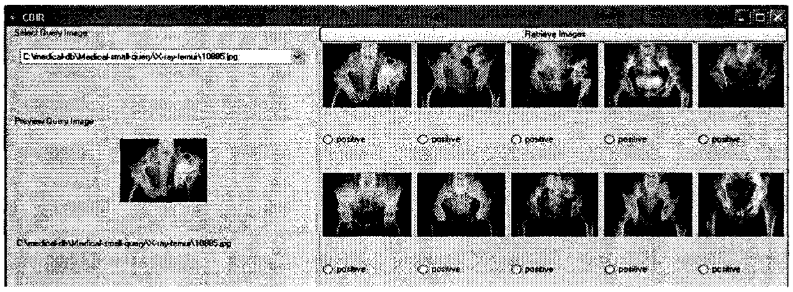
Figure 56 shows PR graphs of the proposed statistical similarity measure based retrieval in low-dimensional feature spaces ($\mathbf{f}^{\mathrm{PCA}}$) for both collection. The retrieval accuracies were also compared with an Euclidean similarity matching in $\mathbf{f}^{\mathrm{PCA}}$ and cosine similarity matching in a concept space ($\mathbf{p}^{\mathrm{PCA}}$) that is generated by probabilistic output of SVMs based on $\mathbf{f}^{\mathrm{PCA}}$ as input. From Figure 56, it is clear that, the statistical similarity matching performed significantly better when compared to Euclidean similarity matching in the same feature space. Such a result is expected as the Euclidean distance does not consider the correlations or variations of its feature attributes in a feature space. This justifies the assumption that search by exploiting



C:\medical-db\Medical-small-query\X-ray-femur\10885.jpg

C:\medical-db\Medical-small-query\X-ray-femur\10885.jpg

Figure 57: A snapshot of the retrieval result based on Euclidean similarity matching



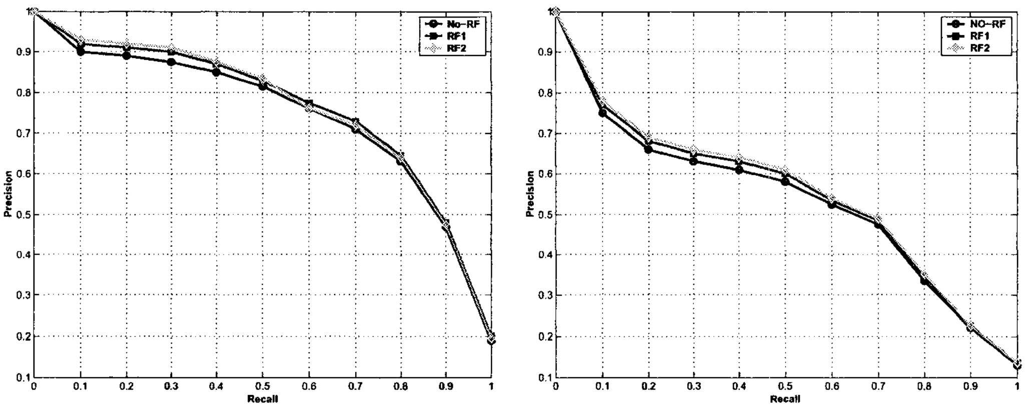
C:\medical-db\Medical-small-query\X-ray-femur\10885.jpg

C:\medical-db\Medical-small-query\X-ray-femur\10885.jpg

Figure 58: A snapshot of the retrieval result based on Statistical similarity matching

the category-specific feature distribution information is more appropriate in a semantically organized database. Similarity between two images based solely on Euclidean or any other geometric distance measures might not conform to their semantic similarity. One more observation is that the performances were almost equal in between cosine measure in global concept-based feature space and statistical measure in PCA based subspace. Although there is no direct relation, both techniques utilizes category specific information either in a feature space or in a similarity matching function. Hence, in semantically organized collections, it is always effective to exploit category information as much as possible. For a qualitative evaluation of the performances, Figure 57 and Figure 58 show the snapshots of the retrieval result for a query image of *"X-ray-Fumer"* category in MED-DB data set. In Figure 57, for the query image, system returns 8 images of the same category out of the top ten images by applying the Euclidean similarity measure on low-dimensional feature space. Whereas, it returns all relevant images in the top ten positions when the statistical similarity matching is performed in the entire data set as shown in Figure 58.

Finally, the retrieval accuracy of the statistical similarity measure is evaluated by considering first two iterations of relevance feedback-based method. We manually



(a) MED-DB

(b) IRMA-DB

Figure 59: Accuracy comparison with and without RF

selected one image from each category to form query image sets for both collections and provide judgement about relevant and non relevant images from top retrieved 20 images at each iteration. Figure 67 shows PR graphs for statistical similarity based retrieval with and without RF-based method. We can observe that performances are slightly improved when both query parameter updating approaches (e.g., RF1 and RF2) were utilized as described in Section 3.4.3 of Chapter 3. Some of the retrieval results were already good without providing any feedback information. The RF methods showed good performances only when images at the top ranked positions were not in the same category as the query image. Due to the averaging out of the results for all queries, the PR graph only show slight improvements in accuracies in both collections.

8.4 Performance Evaluation of Local Concept-Based Retrieval

In this section, we evaluate our proposed local concept-based image retrieval approaches (described in Chapter 4) on a collection of 4000 natural photographic images. The collection (we call it PHOTO-DB) is basically a subset of the larger IAPR photographic collection as described in Section 8.1.1. The images in this collection are manually classified into 18 disjoint semantical categories at a global level as shown

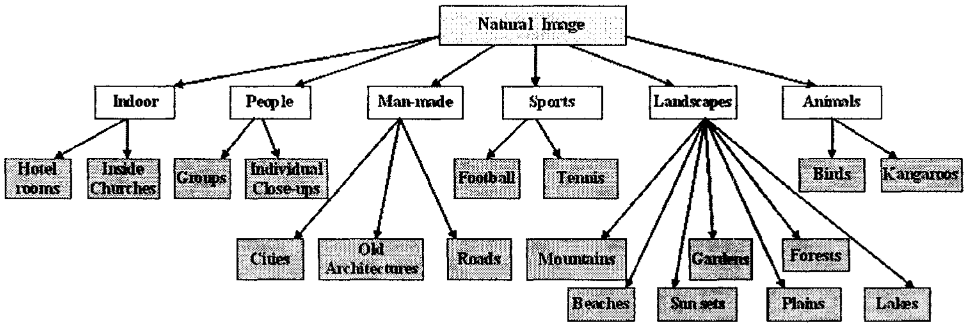


Figure 60: Classification structure of Photo-DB.

in Figure 60 with grey level leaf nodes. The categories include people such as groups and closeup of individual persons; natural landscapes, such as mountains and beaches; man made objects, such as cities and old architectures; and so on. This category information will serve as ground truths for experimental purposes. The images in all categories are chosen in such a way that we can extract both visual and semantic concepts at local image level for indexing as described in Chapter 4. The collection can be termed somewhere in between a narrow and broad domain due to the variabilities of the images.

8.4.1 Training for Local Visual and Semantic Concepts

To generate the codebook of visual concept prototypes based on SOM clustering and SVMs models for semantic concepts, we need a training set of images beforehand for the learning processes. The training set used for this purpose consist of 10% images of the entire data set (4000 images) resulting in a total of 400 images. Images in all categories are selected as equal portions and the remaining 90% images of the data set are used for retrieval evaluation. The reduced size of the training set compared to the global classification-based approach is reasonable as less images are required due to their partition to sub-images to generate local patches.

To find an optimal codebook for image encoding and representation, the training images are partitioned into 64, 144, and 256 sub-images in (8 × 8), (12 × 12), and (16 × 16) grids respectively. After the feature vector generation from the sub-images or blocks of each partition scheme, the SOM is trained to generate two-dimensional codebook of four different sizes as 256 (16 × 16), 400 (20 × 20), 900 (30 × 30),

Table 8: Statistics of the Training Set for Local Semantic Concepts

Concept	Number of Regions	Concept	Number of Regions
Water	300	Sky-Blue	320
Sky-Cloud	250	Snow	250
Sand	200	Grass	250
Rock	200	Floor	200
Brick	250	Pavement	200
Sun	220	Dark Background	270
Skin-Light	230	Skin-Dark	210
Cloth-Plain	260	Cloth-Textured	200
Faces	245	Leaf	200
Green Foliage	290	Floral	280

Table 9: Error rates (training set)

Kernel	C	γ	Degree	Error rate (%)
RBF	200	0.05		17.10
Polynomial	100		1	21.22
Polynomial	100		2	22.09

and 1600 (40 × 40) units. For retrieval based on visual concepts, the testing is therefore conducted with twelve (3 different partitions × 4 codebook sizes) different configurations. After the codebook construction process, all images in the test sets are encoded with the indices of concept prototypes of a particular codebook and visual concept frequency based feature vectors (e.g., V-Concept) are generated as described in Section 4.1.2 of Chapter 4. The effective combination of image partition and codebook size will be determined based on retrieval accuracies of their corresponding vectors in the test set. For training of the SOM, we set the initial learning rate as $\alpha = 0.07$,

For modeling of local semantic concepts, we manually defined 20 local categories based on image regions of the training set. To generate the local patches, each image in the same training set are at first partitioned into an even grid of 8 × 8 sub-images. Only sub-images that conform to at least 70-80% of a particular category (concept) out of the 20 pre-defined categories are selected and labeled with the corresponding category label.

Table 8 shows the statistics of each local semantic concept with number of regions to present them in the training set for SVMs. For training of SVMs, we experimented with both RBF and Polynomial kernels [165]. The error rate of training is measured by a 10-fold CV. The lowest error rate is achieved with RBF kernel (17.10) as shown in Table 9. Hence, after finding the best values of parameters C and γ of the RBF

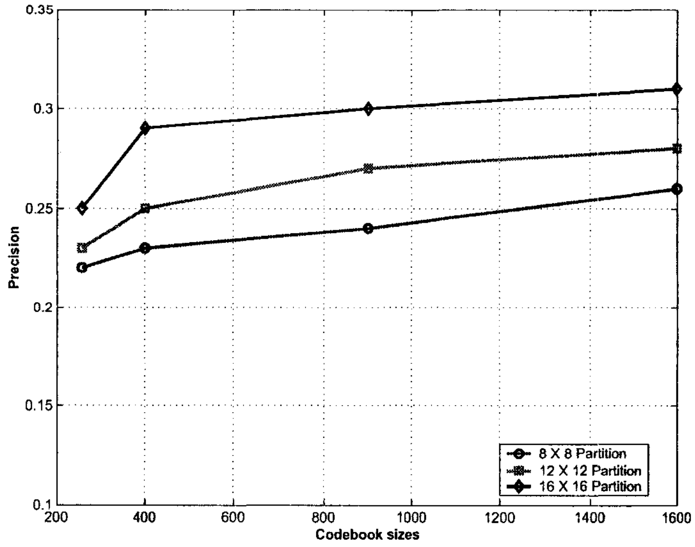


Figure 61: Accuracy comparison of different codebook sizes and image partitions.

kernel, these are utilized to generate the final SVMs model file for later prediction and annotation of images in the test set.

8.4.2 Retrieval Accuracies (Local Concept)

For a quantitative evaluation of the retrieval results, all the images in the test set are selected as query images. A retrieved image is considered a match if it belongs to the same category as the query image out of the 18 disjoint semantical categories at global level.

Figure 61 shows the average precisions within top 20 images (P(20)) for visual concept-based retrieval (e.g., V-concept) on four codebook sizes. It is clear from the Figure 61 that a larger codebook size generally leads to higher precision and a partition of images into 16 × 16 grid (black solid curve in the figure) achieved better precisions for all different codebook sizes in this collection. The general trend here is that larger codebook size with smaller blocks leads to higher retrieval accuracy and in the same time more storage and computation requirements. Hence, we choose a codebook of size 400 (20 × 20) units, which corresponds to the first turning point in Figure 61 and consider the 16 × 16 partition scheme for the generation of all the proposed feature representations and consequent retrieval evaluation.

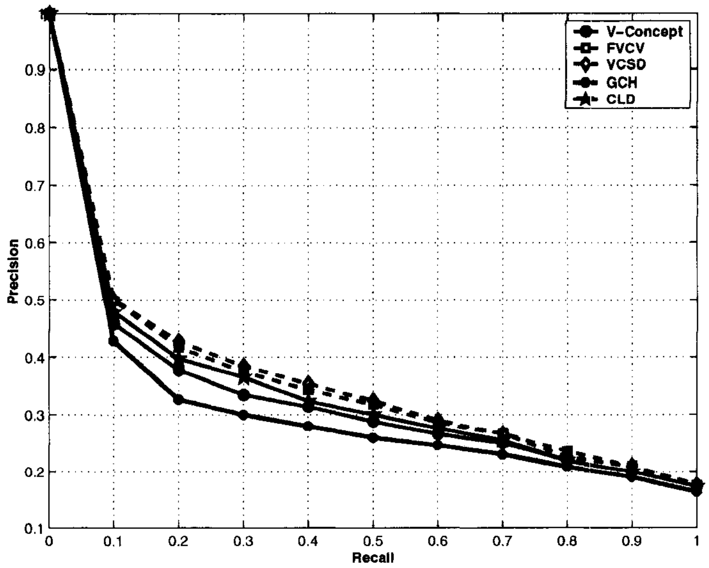


Figure 62: Accuracy comparison of local visual concept feature spaces.

To show the effectiveness of the proposed fuzzy visual concept vector (FVCV) and visual concept structure descriptor (VCSD) as described in Sections 4.1.3 and 25 of Chapter 4, they are compared with simple frequency-based visual concept vector (V-concept). In addition, the performances are compared to a global color histogram (GCH), which is quantized to a 64 bin (4 × 4 × 4) in RGB color space and 16-dimensional MPEG-7 color layout descriptor (CLD) [62] by considering 10 Y, 3 Cb and 3 Cr coefficients for each image. The reason of choosing these two low-level feature descriptors is that they present different aspects of images, where GCH simply counts the frequency of pixels in each bin of a histogram and CLD considers the spatial layout of colors in images. For fuzzy visual concept-based representations, we consider the value of m of the fuzziness exponent as 2 and an (8 × 8) structuring element for the visual concept structure descriptor. In all representations, we apply a $L1$-norm (e.g., city block or Manhattan) based distance measure to compare a query and database images. Figure 62 presents the PR graph of all the feature representation schemes. We can observe that the performance of the V-concept is in between the performances of GCH and CLD. However, both the proposed feature representation schemes (e.g., FVCV and e VCSD) performed significantly better then GCH and V-concept and slightly better then CLD. The results indicate that the proposed feature representations better capture the information based on fuzzy and spatial relations

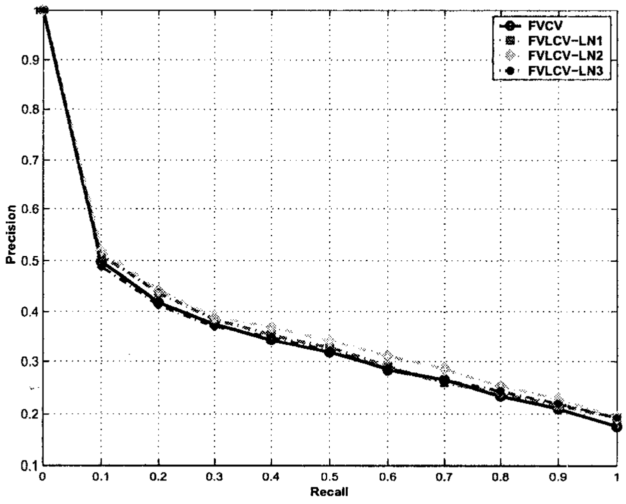


Figure 63: Accuracy comparison based on global and local membership values

between concepts.

Figure 63 shows the PR graph of the proposed visual concept representation schemes based on both global and local fuzzy membership values. For the local fuzzy visual concept vector (FVLCV), retrieval were performed by considering up to local neighborhood levels LN_1, LN_2, and LN_3. The retrieval accuracies are almost similar in all these cases as shown in Figure 63. However, we achieved slightly better performance when the membership values were computed from a neighborhood level LN_2. By increasing the level further, the performance was decreasing to a point which matches to the performance based on global membership values. It demonstrate that, it is effective if we consider membership values only from one or two level of local neighborhood. Since the majority of correlated or similar concepts prototypes are located close to each other due to the topology preserving structure of the codebook. Moreover, it is also better from an efficiency viewpoint as less time is required to compute the membership values from local neighborhood to generate the final feature vector.

To show the effectiveness of the probabilistic semantic concept vector (PSCV) and semantic concept structure descriptor (SCSD) they are compared with the simple frequency-based semantic concept vector (S-concept) as well as with GCH and CLD. For SCSD, we used a small structuring element of size (4×4) units due to the 8×8 partition of images. We can observe that precisions of both PSCV and SCSD are

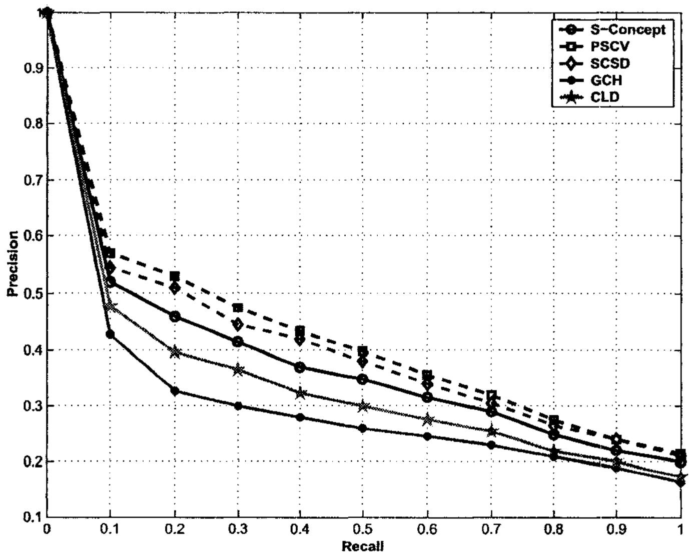


Figure 64: Accuracy comparison of local semantic concept-based features.

always higher at all recall levels when compared to the other three representation schemes. Overall, the precision of the semantic concept-based vectors are better compared to the visual concept-based representations. The reason behind this is the exploitation of domain knowledge in a supervised manner for concept modeling and consequent feature extraction process. Finally, from both PR graphs in Figure 62 and Figure 64, we can conclude that our feature representation methods better capture the semantic information compared to the low-level features commonly used in CBIR.

8.5 Performance Evaluation of Query Expansion Techniques

This section evaluates the retrieval effectiveness of the proposed automatic query expansion techniques described in Chapter 5. For this, we performed experiments on the same PHOTO-DB data set as described in previous section.

Figure 65 and Figure 66 show the PR graphs of the automatic query expansion approaches in local visual and semantic concept spaces after two iterations of feedback information. The performances are compared without any query expansion and with a relevance (RF) method based on Rocchio algorithm [106] as described in Section

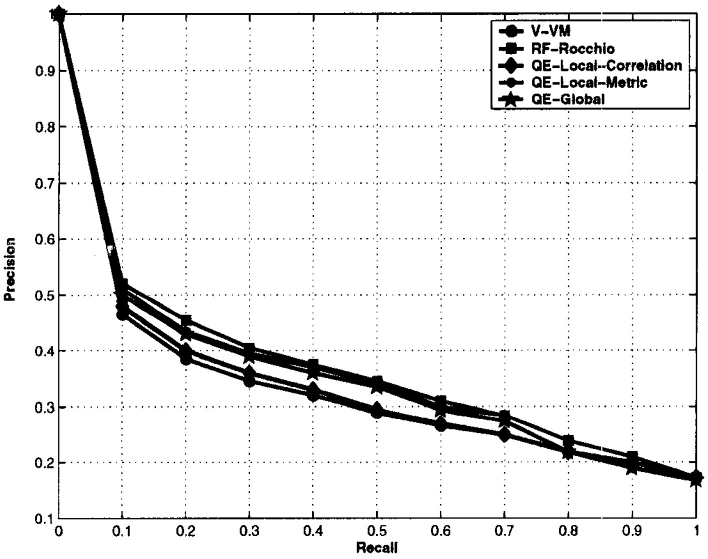


Figure 65: Accuracy comparison of different query expansion schemes in local visual concept space.

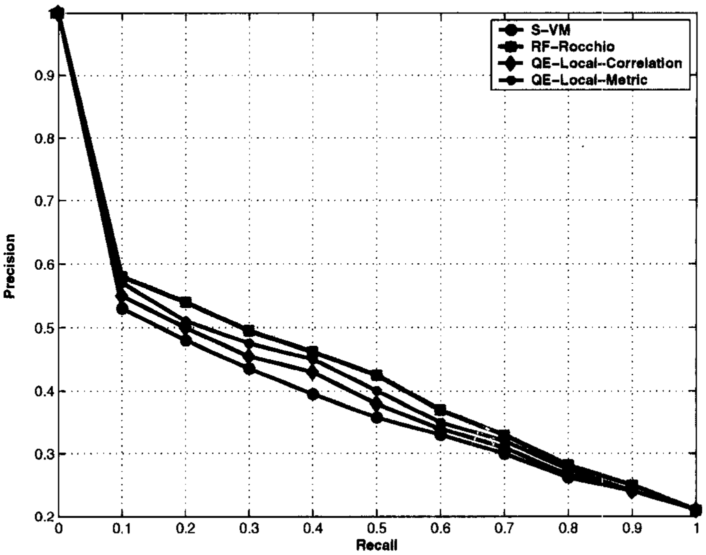
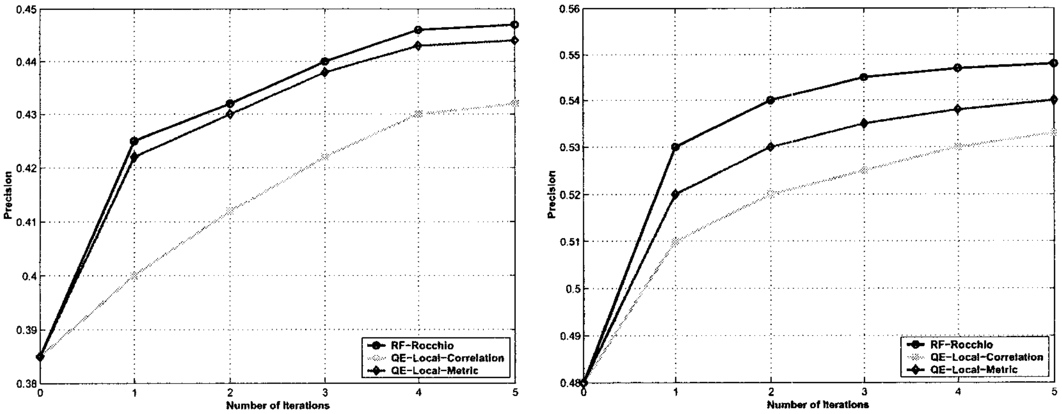


Figure 66: Accuracy comparison of different query expansion schemes in local semantic concept space.

6.3 of Chapter 6. For query expansion based on local analysis and RF, we considered top 20 retrieved images from previous iteration as local feedback for the next iteration. The Rocchio's RF method is interactive in nature, which relies on both positive and negative feedbacks from users. Due to the automatic simulation of the feedback information, we consider the first two irrelevant images as negative feedback at each iteration. For query expansion approach based on global analysis in visual concept space, we only need to perform one extra retrieval iteration to exploit the information from global cluster and do not require any local or user's feedback information. For local analysis based approaches, we selected 3 additional concepts from local clusters for each query concepts and for global analysis, 10 concepts are selected from codebook for a query image.

It is clear from both Figure 65 and Figure 66 that performances were improved for all the query expansion and RF-based approaches compared to the case when no query expansion is utilized in vector space model based representation of visual and semantic concepts (e.g., V-VM and S-VM). Although the performances of RF-based approach (RF-Rocchio) were best in both visual and semantic concept spaces, there were no significant differences when we compared with our query expansion approaches, specially in visual concept based feature space. The performances of query expansion based on local analysis of metric cluster (QE-Local-Metric) is comparatively better then the local analysis based on correlation cluster (QE-Local-Correltion) in both visual and semantic concept spaces. The most probable reason is that it is difficult to capture enough concept correlation information from a local image set without considering their relative ordering and distances in a neighborhood. Therefore, when we encode that information in a metric cluster, we achieved better performances as shown in both PR graphs. To also check the consistency of the query expansion approaches in based on local feedback, we also considered 5 iterations of feedback and compared the performances by considering the average precision within top 20 (P(20)) retrieved images as shown in Figure 67. From Figures 67(a) and 67(b), we can observe one common trend that the precision increases rapidly at the first 2-3 iterations and after that the improvement subsides or the system converges. Hence, we can say that the performances of the query expansion approaches are consistent from one iteration to another and query expansion based on local analysis of metric cluster (QE-Local-Metric) performed comparatively well in both feature spaces.



(a) Visual Concept Space

(b) Semantic Concept Space

Figure 67: Accuracy comparison with different number of iterations

Table 10: Improvement using expanded queries

Number of additional concepts	10	15	20	30
Without Query Expansion	0.3532	0.3532	0.3532	0.3532
With Query Expansion	0.3766	0.3845	0.3706	0.3478
Improvement	+ 6.62%	+ 8.86 %	+ 4.90 %	-1.50%

Moreover, the query expansion approach based on a global analysis (e.g., QE-Global) also performed equally well compared to the RF-based approach as shown in Figure 65 and Figure 66. This validates the initial assumption that instead of expanding the terms (concepts) based on each query term, it is effective to select expanded terms based on a single virtual query vector only. The results were also evaluated by comparing the overall average precisions among the average precision on top 20, 30, 50, and 100 retrieved images with and without query expansion with different number of added concepts. From Table 10, it can be observed easily that the improvement by expanded queries increases when the number of additional concepts increases initially. But after adding certain number of concepts, the performance started to decrease and at a point it even became lower then the original query. This could be explained by the fact that fewer selected concepts contains better information and when the number is getting larger, the expansion approach might add the concepts that are not related to relevant images.

Figure 68 shows the effectiveness of Quadratic distance measure as described in Section 5.3 of Chapter 5, it is compared with a cosine distance in the same visual

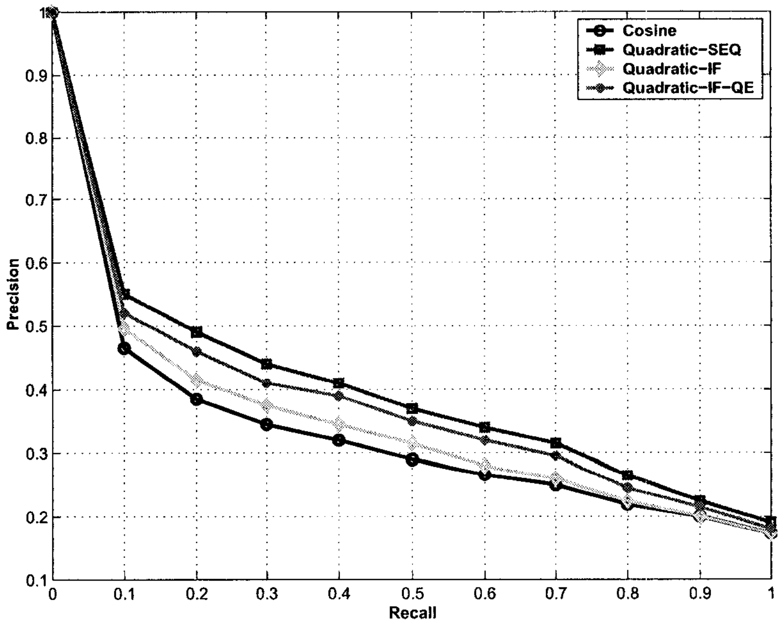


Figure 68: Average Precision curves for the Quadratic similarity matching with and without Inverted Index

Table 11: Retrieval time with and without indexing schemes

Linear Search (ms)	Inverted Index (ms)	QE-Inverted Index (ms)
445	119	236

concept-based feature space (V-VM) as well as compared with and without query expansion in inverted indexing scheme. For query expansion, we consider up to two level of neighborhoods (e.g.,LN_2). The quadratic distance is much better then L_1 distance and the result indicate that the correlation among the concepts is not negligible. In addition, the result shows that the performances of the Quadratic distance measure based on the sequential search in the whole collection and based on query expansion in inverted index are comparatively close. The performance in inverted index without query expansion degrades to the extent that can not be ignored.

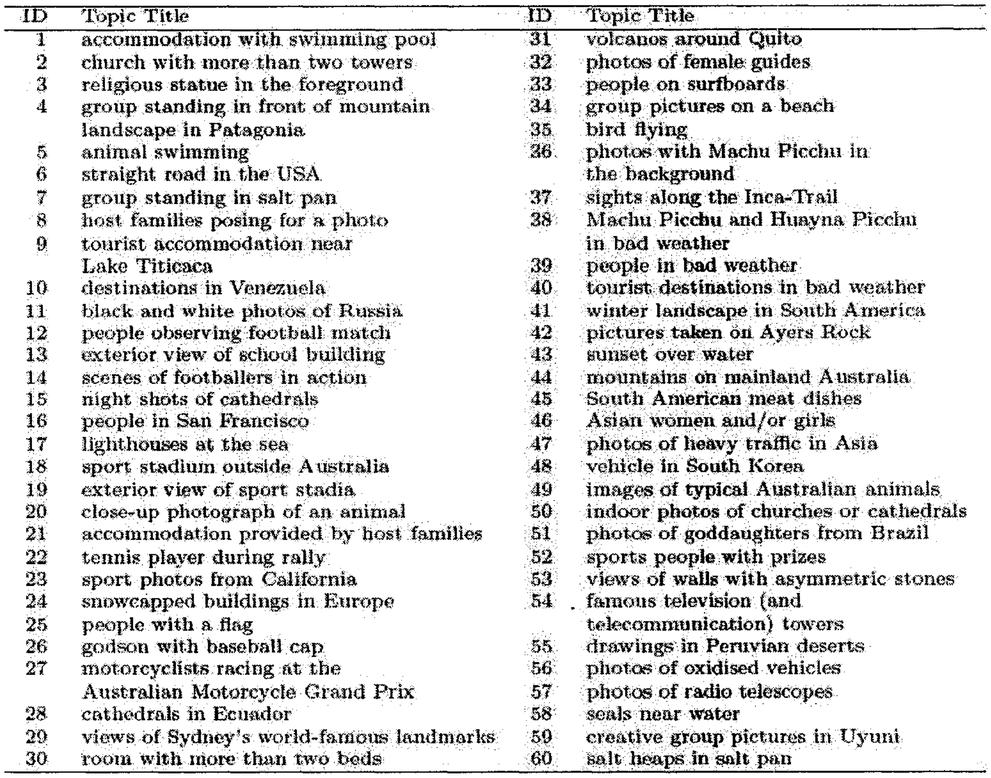
To test the efficiency of the search schemes, we also compared the average retrieval time (in milliseconds) with and without indexing schemes as well as with and without query expansion (in an Intel Pentium 4 processor with Windows XP as the operating system and 1 Gb memory) for the query set. From the results in Table 11, it is clear that the search in inverted index is about four to five times faster compared to the linear search in the entire test set. Whereas, the search in inverted index with query expansion is about two times faster then liner search with much improved retrieval accuracy. Hence, the Quadratic distance matching in inverted index with query expansion has proved to be both effective and efficient.

In summary, we can conclude that automatic query expansion approaches based on correlation information both at local and global levels improve retrieval performances of initial queries and also seems to be comparative when compared to a well known RF-based algorithm. Hence, query expansion techniques need to be explored further in CBIR domain in a similar manner it was investigated in text-retrieval domain.

8.6 Performance Evaluation of Fusion-Based Retrieval

This section presents the experimental results of the fusion based retrieval approaches in both context and content-based feature spaces either using a single modality or combining both modalities described in Chapter 6. We utilized the entire photographic collection of 20,000 images and the medical collection of 66,662 images described in Sections 8.1.1 and 8.1.2 respectively. Since, we performed the experiments in benchmark collections under CLEF [57, 58], the results are generated based on the query topics (e.g., a short sentence or phrase describing the search request in a few words with three relevant images) provided by the organizers of CLEF. The performances of different methods/approaches are described here based on our submission of different runs in ImageCLEF competition for the year of 2006 and 2007 [217, 218]. We also extended our experiments based on two modified query sets for both collections to show the limitation and real effectiveness of the proposed methods.

There are some differences among the data sets and the query topics used for the evaluation in ImageCLEF'06 [55, 56] and ImageCLEF'07 [57, 58]. For example, the medical retrieval task added two news data sets in 2007 and the query topics are also different from the ones used in 2006. The photographic images were fully annotated in 2006 with a "Description" tag, whereas in 2007, they were lightly annotated with a "Title" tag only. Although, query topics and image collection were the same in both years for photographic image retrieval. In following, we mainly describe about



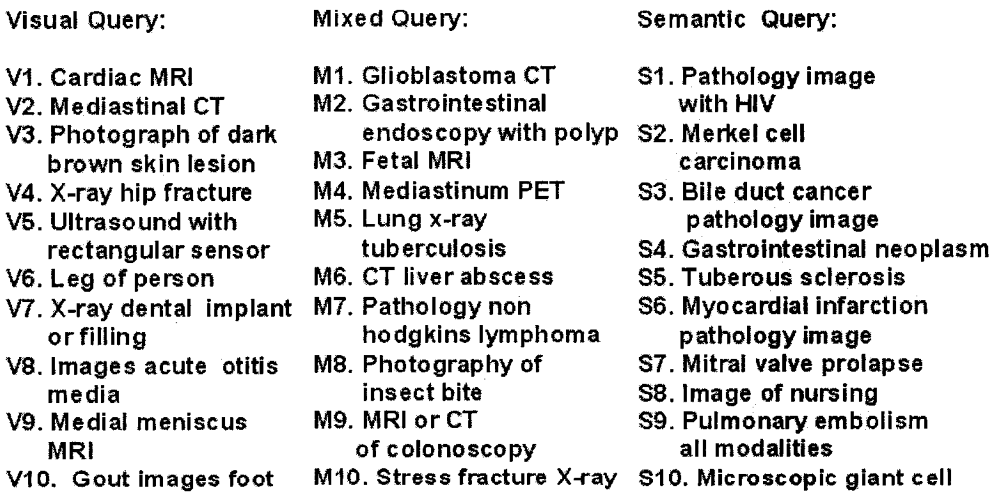
ID	Topic Title	ID	Topic Title
1	accommodation with swimming pool	31	volcanos around Quito
2	church with more than two towers	32	photos of female guides
3	religious statue in the foreground	33	people on surfboards
4	group standing in front of mountain landscape in Patagonia	34	group pictures on a beach
5	animal swimming	35	bird flying
6	straight road in the USA	36	photos with Machu Picchu in the background
7	group standing in salt pan	37	sights along the Inca-Trail
8	host families posing for a photo	38	Machu Picchu and Huayna Picchu in bad weather
9	tourist accommodation near Lake Titicaca	39	people in bad weather
10	destinations in Venezuela	40	tourist destinations in bad weather
11	black and white photos of Russia	41	winter landscape in South America
12	people observing football match	42	pictures taken on Ayers Rock
13	exterior view of school building	43	sunset over water
14	scenes of footballers in action	44	mountains on mainland Australia
15	night shots of cathedrals	45	South American meat dishes
16	people in San Francisco	46	Asian women and/or girls
17	lighthouses at the sea	47	photos of heavy traffic in Asia
18	sport stadium outside Australia	48	vehicle in South Korea
19	exterior view of sport stadia	49	images of typical Australian animals
20	close-up photograph of an animal	50	indoor photos of churches or cathedrals
21	accommodation provided by host families	51	photos of goddaughters from Brazil
22	tennis player during rally	52	sports people with prizes
23	sport photos from California	53	views of walls with asymmetric stones
24	snowcapped buildings in Europe	54	famous television (and telecommunication) towers
25	people with a flag	55	drawings in Peruvian deserts
26	godson with baseball cap	56	photos of oxidised vehicles
27	motorcyclists racing at the Australian Motorcycle Grand Prix	57	photos of radio telescopes
28	cathedrals in Ecuador	58	seals near water
29	views of Sydney's world-famous landmarks	59	creative group pictures in Uyuni
30	room with more than two beds	60	salt heaps in salt pan

Figure 69: Topics for the Photographic (IAPR) collection in ImageCLEF [57]

the topics in ImageCLEF'07 [57, 58] as well as the modified mixed mode query sets, whereas the details of the topics for 2006 evaluation are provided in overview papers [55, 56].

8.6.1 Query Set of General Photographic Images

A total of 60 topics were provided by the ImageCLEF'07 [57] for ad-hoc retrieval of general photographic images as shown in Figure 69 with a short description of each topic. In order to increase the reliability of results and to make the task realistic, we can see that many topics are highly semantic in nature with geographic constraints or named entities, which makes the content-based image search quite difficult.



Visual Query:

V1. Cardiac MRI
V2. Mediastinal CT
V3. Photograph of dark brown skin lesion
V4. X-ray hip fracture
V5. Ultrasound with rectangular sensor
V6. Leg of person
V7. X-ray dental implant or filling
V8. Images acute otitis media
V9. Medial meniscus MRI
V10. Gout images foot

Mixed Query:

M1. Glioblastoma CT
M2. Gastrointestinal endoscopy with polyp
M3. Fetal MRI
M4. Mediastinum PET
M5. Lung x-ray tuberculosis
M6. CT liver abscess
M7. Pathology non hodgkins lymphoma
M8. Photography of insect bite
M9. MRI or CT of colonoscopy
M10. Stress fracture X-ray

Semantic Query:

S1. Pathology image with HIV
S2. Merkel cell carcinoma
S3. Bile duct cancer pathology image
S4. Gastrointestinal neoplasm
S5. Tuberous sclerosis
S6. Myocardial infarction pathology image
S7. Mitral valve prolapse
S8. Image of nursing
S9. Pulmonary embolism all modalities
S10. Microscopic giant cell

Figure 70: Topics for the Medical Collection in ImageCLEFmed [58]

8.6.2 Query set of Medical Images

For ad-hoc medical image retrieval, a total of 30 query topics were provided [58] that were initially generated based on a log file of Pubmed [10]. All topics were categorized with respect to the retrieval approaches expected to perform best, i.e., visual topics for CBIR, semantic topics for text retrieval and mixed topics for multi-modal retrieval. Each topic consisted of the query itself in three languages (English, German, French) and 2 to 3 example images for the visual part of the topic. Figure 70 shows all three types of topics based on a short description in English.

Performance Measures used in CLEF:

The relevant sets of all topics were crated by the CLEF organizers by considering top retrieval results of all submitted runs of the participating groups. Results for submitted runs were computed using the latest version of TREC-EVAL [11] software. Submissions were evaluated using un interpolated (arithmetic) Mean Average Precisions (MAP) and Precision at rank 20 (P20) because most online image retrieval engines like Google, Yahoo, and Altavista display 20 images by default. Further

[10] http://www.pubmed.gov

[11] http://trec.nist.gov/*trec* − *eval*/

measures considered include Geometric Mean Average Precision (GMAP) to test robustness, and the Binary Preference (BPREF) measure which is a good indicator for the completeness of relevance judgments [57].

8.6.3 Result Analysis of the Submitted Runs in ImageCLEF'06

This section presents the techniques used and the analysis of different runs submitted by us (CINDI group) in ImageCLEF'06 [217, 55, 56]. For the ad-hoc image retrieval from both photographic and medical image collections, we experimented with cross-modal interaction and integration approaches described in Chapter 6, based on the relevance feedback in the form of textual query expansion and visual query point movement with adaptive similarity matching functions. We submitted three different runs for the ad-hoc retrieval of the photographic collection and three runs for the ad-hoc retrieval of the medical collection as shown in Table 12 and Table 13. In all these runs, only English is used as the source and target language without any translation. For RF-based method, a user provided feedback from top retrieved 30 images. For content based search, the overall image level similarity is measured by fusion of manually weighted combination of individual similarity measures of different image representation as described in Section 6.2 of Chapter 6 and in our workshop paper in [217].

As shown in Table 12, for photographic image retrieval with run ID *"Cindi-Text-Eng"*, we performed only an automatic text-based search without any feedback as our base run. For the second run with ID *"Cindi-TXT-EXP"*, we performed manual feedback in the text only modality. For textual query expansion, we used an approach similar to the "Local1" described in Section 6.3 of Chapter 6. In this case, we considered 10 additional keywords for query expansion. For the third run with ID *"Cindi-Exp-RF"*, both text and image modalities are interactively combined in a single search process (with only one or two iterations of feedback for each modality). For image-based query refinement, a similar approach was used as described in Section 6.4 of Chapter 6 and for final merging of the result lists, the weights are selected as $\omega_D = 0.7$ and $\omega_I = 0.3$ for the text and image-based search as described in [217]. From Table 12, it is clear that the MAP scores are almost doubled for the last two runs compared to the base run and the best performance is obtained with feedback and integration of text and image search. In fact, these two runs ranked first and

Table 12: Results of the photographic image retrieval (ImageCLEF'06)

Run ID	Modality	A/M	RF	QE	MAP
Cindi-Text-Eng	Text	Automatic	Without	No	0.1995
Cindi-TXT-EXP	Text	Manual	With	Yes	0.3749
Cindi-Exp-RF	Text+Image	Manual	With	Yes	0.3850

Table 13: Results of the medical image retrieval (ImageCLEF'06)

Run ID	Modality	RF	MAP	R-prec	B-pref
CINDI-Fusion-Visual	Image	Without	0.0753	0.1311	0.166
CINDI-Visual-RF	Image	With	0.0957	0.1347	0.1796
CINDI-Text-Visual-RF	Image+Text	With	0.1513	0.1969	0.2397

second in terms of the MAP score among the 157 submissions in the photographic retrieval task in ImageCLEF'06 [57]. We performed manual submissions using relevance judgement from the user. This along with the integration of both modalities was the main reason for such good results.

For the medical retrieval task, the results are shown in Table 13. We performed a content-based search without feedback for the first run with ID *"CINDI-Fusion-Visual"*. For this search, the overall image level similarity is measured by fusion of manually weighted combination of individual similarity measures of different global, semi-global, local, and low-resolution image representations as described in [217]. We ranked first in this run in the category (automatic+visual) based on the MAP score (0.0753) out of 11 different submissions. For the second run with ID *"CINDI-Visual-RF"*, we performed manual feedback in the image only modality. For this category (e.g., visual only run with RF), only our group has participated this year and achieved a better MAP score (0.0957) than without RF as shown in Table 13. For the third run with ID *"CINDI-Text-Visual-RF"*, we performed manual feedback in both modalities and merged the result lists in a similar way as we did for the photographic collection. For this search with a MAP score of 0.1513, it is clear that combining both modalities is far better then using a single one and validates our initial assumption for multimodal retrieval.

Overall, our participation was successful in ImageCLEF'06 as we achieved some

Table 14: Results of the photographic image retrieval (ImageCLEF'07)

Run ID	Modality	MAP	BPREF
CINDI-TXT-ENG-PHOTO	Text	0.1529	0.1426
CINDI-TXT-QE-PHOTO	Text	0.2637	0.2515
CINDI-TXT-QE-IMG-RF-RERANK	Image+Text	0.2336	0.2398
CINDI-TXTIMG-FUSION-PHOTO	Image+Text	0.1483	0.1620
CINDI-TXTIMG-RF-PHOTO	Image+Text	0.1363	0.1576

very good results by applying our fusion-based retrieval approaches in both single and multimodal search processes as described in Chapter 6.

8.6.4 Result Analysis of the Submitted Runs in ImageCLEF'07

This section presents the techniques used and the analysis of different runs submitted by us (CINDI group) in ImageCLEF'07 [218, 57, 58]. For the ad-hoc image retrieval from both photographic and medical image collections, we performed experiments based on the multiple query reformulations and dynamic fusion-based approaches described in Chapter 6.

The descriptions and performances of the different official runs are shown in Table 14 and Table 15 for ad-hoc retrieval tasks of the photographic and the medical collection respectively. We submitted five different runs for the ad-hoc retrieval of the photographic collection in ImageCLEF'07 [57, 58], where first two runs are based on text only search approaches and last three runs are based on multi-modal searches. For the first run *"CINDI-TXT-ENG-PHOTO"*, we performed only a manual text-based search without any query expansion as our base run. This run achieved a MAP score of 0.1529 and ranked within the top 30% out of all 476 submitted runs. Our second run *"CINDI-TXT-QE-PHOTO"* achieved the best MAP score (0.2637) among all our submitted runs and ranked 21st ImageCLEF'07. In this run, we performed two iterations of manual feedback for textual query expansion and combination based on dynamic weight update schemes for text only retrieval as described in Section 6.3 of Chapter 6. The rest of the runs are based on multi-modal approach, where for the third run *"CINDI-TXT-QE-IMG-RF-RERANK"*, we performed the sequential

Table 15: Results of the medical image retrieval (ImageCLEF'07)

Run ID	Modality	MAP	R-prec
CINDI-IMG-FUSION	Image	0.0355	0.0566
CINDI-IMG-FUSION-RF	Image	0.0396	0.0574
CINDI-TXT-IMAGE-LINEAR	Image+Text	0.1906	0.2366
CINDI-TXT-IMG-RF-LINEAR	Image+Text	0.1227	0.1545

approach (Section 6.6.1 of Chapter 6) with pre-filtering and re-ordering with two iterations of manual feedback in both text and image-based searches. However, the re-ordering approach did not improve the result as a whole (e.g., ranked 32nd) in terms of MAP score (0.2336) as compared to the text only query expansion approach. For the fourth run *"CINDI-TXTIMG-FUSION-PHOTO"*, we performed a simultaneous retrieval approach without any feedback information with a linear combination of weights as $\omega_D = 0.7$ and $\omega_I = 0.3$ and for the fifth run *"CINDI-TXTIMG-RF-PHOTO"*, two iterations of manual relevance feedback are performed. However, these two runs did not perform well in terms of the MAP score as compared to the sequential approach due to early combination and semantical nature of majority of the topics.

Table 15 shows the results of the four submitted runs for the image retrieval task in the medical collections. In the first run *"CINDI-IMG-FUSION"*, we performed only a visual only search based on various image feature representation schemes (Section 6.2 of Chapter 6) without any feedback information and with a linear combination of equal feature weights. For the second run *"CINDI-IMG-FUSION-RF"*, we performed only one iteration of manual feedback for visual query refinement and combined the similarity matching functions based on the dynamic weight updating scheme. For this run we achieved a MAP score of 0.0396, which is slightly better then the score (0.0355) achieved by the first run without any RF. These two runs ranked among the top five results based on pure visual only run in ImageCLEF'07. For the third run *"CINDI-TXT-IMAGE-LINEAR"*, we performed a simultaneous retrieval approach without any feedback information with a linear combination of weights as $\omega_D = 0.7$ and $\omega_I = 0.3$ and for the fourth run *"CINDI-TXT-IMG-RF-LINEAR"*, two iterations of manual relevance feedback are performed similar to the last two runs of photographic retrieval task. From Table 15, it is clear that combining both modalities for the

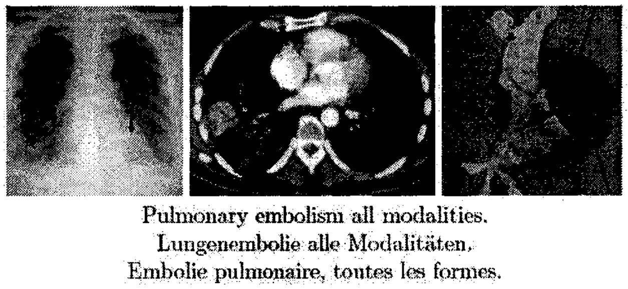

Pulmonary embolism all modalities.
Lungenembolie alle Modalitäten.
Embolie pulmonaire, toutes les formes.

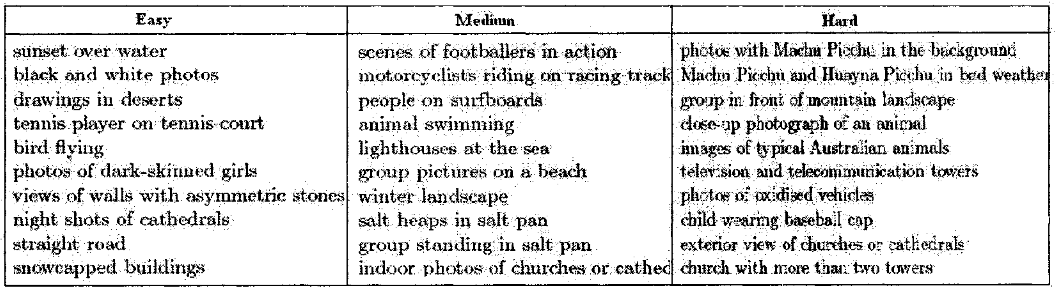
Figure 71: Example of a sematic topic in ImageCLEFmed'07 [58]

medical retrieval task is far better then using only a single modality based on the MAP scores.

Overall, the performances of content-based search approaches are very low compared to the the text-based and multimodal searches as observed in Table 13 and Table 15. The main reason is the high-level semantic contents in query topics. It strongly validates the point that for semantic retrieval of images in broad domain, associated or contextual information largely improve retrieval results. The similar trend were also observed in ImageCLEF results for the last few years. From the results of Table 14 and Table 15, we can also observe that our adaptive multimodal runs did not improve the result as we expected. The reason for this might be the system did not get enough feedback information to make the weight update algorithm perform effectively as described in Section 6.5 of Chapter 6. In some cases, combining image and text-based searches might also have negative effect in final retrieval result. Figure 71 shows a example semantic topic for medical retrieval with perceptually very different relevant images. This topic should be well suited for textual retrieval approach only and by refining or integrating it with a content-based search would only decrease the performance of a final retrieval result set.

8.6.5 Extended Result Analysis for Mixed-mode Query Sets

In the introductory part of Chapter 6, we showed with an example that cross-modal or multi-modal retrieval approaches would be only effective, when a search requirement is in mixed-mode nature. Therefore, the contents of texts and images should have some internal relationship. To show the real effectiveness of the proposed multimodal approaches, we performed additional experiments and evaluate the results



Easy	Medium	Hard
sunset over water	scenes of footballers in action	photos with Machu Picchu in the background
black and white photos	motorcyclists riding on racing track	Machu Picchu and Huayna Picchu in bed weathe
drawings in deserts	people on surfboards	group in front of mountain landscape
tennis player on tennis court	animal swimming	close-up photograph of an animal
bird flying	lighthouses at the sea	images of typical Australian animals
photos of dark-skinned girls	group pictures on a beach	television and telecommunication towers
views of walls with asymmetric stones	winter landscape	photos of oxidised vehicles
night shots of cathedrals	salt heaps in salt pan	child wearing baseball cap
straight road	group standing in salt pan	exterior view of churches or cathedrals
snowcapped buildings	indoor photos of churches or cathed	church with more than two towers

Figure 72: Mixed-mode topics for Photographic collection [57]

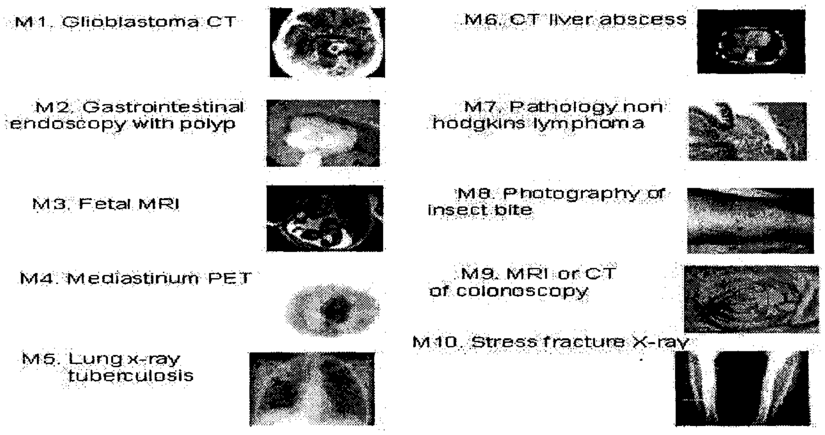

M1. Glioblastoma CT
M6. CT liver abscess
M2. Gastrointestinal endoscopy with polyp
M7. Pathology non hodgkins lymphoma
M3. Fetal MRI
M8. Photography of insect bite
M4. Mediastinum PET
M9. MRI or CT of colonoscopy
M5. Lung x-ray tuberculosis
M10. Stress fracture X-ray

Figure 73: Mixed-mode topics for Medical collection[58]

based on mixed topics for both collections. For the IAPR collection, 30 mixed topics are selected as shown in Figure 72, which were initially considered for visual only retrieval task. For the medical collection, 10 mixed topics were already provided as shown in Figure 73 or the topics in the middle column of Figure 70.

Table 16 and Table 17 show the mixed topic based retrieval results of photographic and medical collection respectively. Only the MAP scores of different result sets that are generated by different combination of methods as described in Chapter 6 are shown here. This time, we provided feedback information from top retrieved 50 images so that dynamic weight update method can perform effectively. By observing the MAP scores, the effectiveness of multi-modal retrieval searches (e.g., Sequential and Simultaneous methods) are evident now. The proposed weight update method performed well now in all cases compared to the corresponding equal weighting approach, i.e.,

Table 16: Retrieval Results of the Photographic Images for Mixed-Mode Topics

Method	Modality	MAP
Visual-Fusion (Equal Weight)	Image	0.0675
Visual-RF-Fusion (Dynamic Weight)	Image	0.0832
Keyword	Text	0.1656
Keyword-RF-Fusion (Equal weight)	Text	0.2542
Keyword-RF-Fusion (Dynamic weight)	Text	0.2785
Sequential (Equal weight)	Image+Text	0.3051
Sequential (Dynamic weight)	Image+Text	0.3120
Simultaneous (Equal weight)	Image+Text	0.2875
Simultaneous (Dynamic weight)	Image+Text	0.2956

Table 17: Retrieval Results of the Medical Images for Mixed-Mode Topics

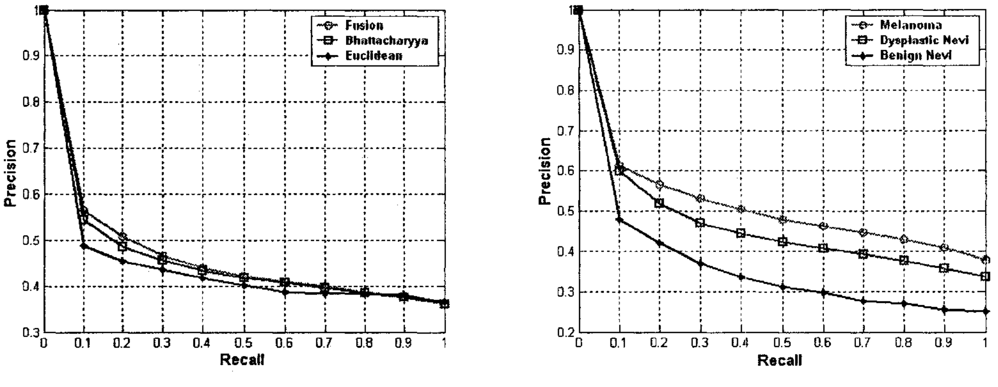
Method	Modality	MAP
Visual-Fusion (Equal Weight)	Image	0.0445
Visual-RF-Fusion (Dynamic Weight)	Image	0.0567
Keyword	Text	0.1329
Keyword-RF-Fusion (Equal weight)	Text	0.1845
Keyword-RF-Fusion (Dynamic weight)	Text	0.1975
Sequential (Equal weight)	Image+Text	0.2251
Sequential (Dynamic weight)	Image+Text	0.2467
Simultaneous (Equal weight)	Image+Text	0.2066
Simultaneous (Dynamic weight)	Image+Text	0.2170

whether it was applied to a single modality (e.g., Visual-RF-Fusion or Keyword-RF-Fusion) or was applied in multi-modal (e.g., Sequential or Simultaneous) searches. In general, we achieved around 4-5% increases in MAP scores for all dynamic weight update based search approaches compared to equal weight based searches for both photographic and medical collections. In addition to the query topic characteristics, feedback information from an increased number of images also contributed to better retrieval effectiveness. Another observation is that the performances of the sequential searches are always better then the simultaneous ones as observed in both Table 16 and Table 17. Hence, it can be concluded that for mixed-mode queries it might be more effective to refine a text-based search result with content-based search. Finally, the performances of the text-based and multi-modal searches are much more better then the content-based searches even for the mixed-mode queries. This justifies the initial prediction that for ad-hoc retrieval in a broad domain, we need additional contextual information in addition to image content information to perform effective retrieval. Finally, we can conclude that combining content and context-based feature as well as using relevance feedback and multiple query expansion techniques can significantly improve retrieval performance.

8.7 Performance Evaluation for Dermoscopic Image Retrieval

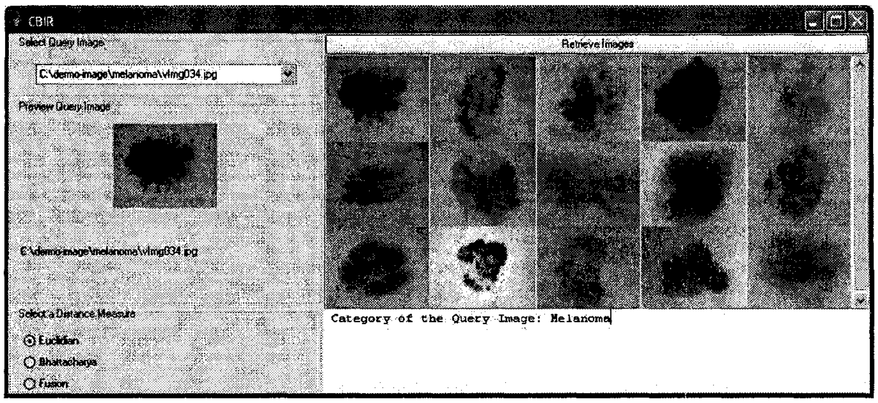
This section presents the experiments and results analysis of the CBIR framework for dermoscopic images described in Chapter 7. We experimented with the database described in Section 8.1.4. Every image in the database is served as a query image. A retrieved image is considered to be a correct match if it belongs to the same category as the query image. The performances of the three distance measures (e.g., Euclidean, Bhattacharyya and Fusion-based distance) and three image categories (e.g., benign, dysplastic and melanoma) are compared based on PR graphs. We have experimented with different weighting combinations and have found out the best combination as $w_{\text{Bhatt}} = 0.7$ and $w_{\text{Euclidean}} = 0.3$ for the the fusion-based similarity matching scheme. More weight is assigned to the Bhattacharyya based similarity measure as it performed better due to the consideration of the cross-correlation among the color channels.

Figure 74(a) presents the precision-recall curves for the Euclidean, Bhattacharyya



(a) (b)

Figure 74: (a) Accuracy comparison of three similarity measures (b) Accuracy comparison of three image categories.



C:\demo-image\melanoma\vlmg034.jpg

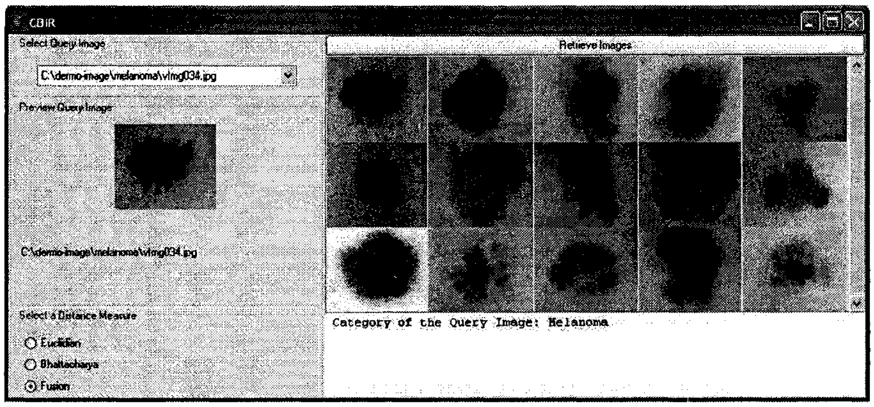
C:\demo-image\melanoma\vlmg034.jpg

● Euclidean
○ Bhattacharya
○ Fusion

Category of the Query Image: Melanoma

Figure 75: Retrieval result based on Euclidean similarity.

and the proposed fusion based distance measures. From Figure 74(a), it is clear that the best accuracy is achieved when other two distances are fused together as discussed in Section 7.3 of Chapter 7, whereas the performance of Euclidean distance measure is significantly lower then the other two. Based on the above observation, we can conclude that similarity measures which utilize cross correlation between feature attributes and fuse distances in different spaces perform better in CBIR. Figure 74(b) presents the average precision curves for three different image categories (melanoma, benign and dysplastic nevi). From Figure 74(b), it is clear that best accuracy is achieved by melanoma category, which is more important from diagnostic point of view as it can distinguish images better from other categories. Figures 75 and 76 show



Select Query Image

C:\dermo-image\melanoma\vlmg034.jpg

Preview Query Image

C:\dermo-image\melanoma\vlmg034.jpg

Select a Distance Measure

○ Euclidian
○ Bhattacharya
● Fusion

Category of the Query Image: Melanoma

Figure 76: Retrieval result based on Fusion similarity.

the snapshots of the retrieval results based on the Euclidean and the Fusion-based similarity matching functions for a query image belongs to the melanoma category. Here, the system returned 11 melanoma images out of 15 from the database by applying the proposed fusion-based similarity matching function, whereas the Euclidean similarity measure returns only 7 images of melanoma category excluding the query image in the first ranked position.

8.8 Summary

In this chapter, we provide the experimental details of different retrieval approaches as we proposed for different image domains and different contexts. We introduce the image collections, the query sets, the performance measures, and the effectiveness of the results. We validate our techniques by showing their improved accuracies compared to other commonly used feature representation, similarity matching, and retrieval techniques. Distinct experiments were carried out and described in our publications also [205, 206, 209, 217, 218, 210]. Overall, we can say that the performances of our proposed approaches in different narrow to broad domains of photographic and medical images are encouraging.

Chapter 9

Conclusion

This chapter provides an overview of our work in image retrieval domain presented in this thesis. In Section 9.1 we state the main themes and scientific achievements of this research. We mention some of the limitations of our approaches in Section 9.2, which currently persist both in domain and technical (e.g., computational) perspectives. Finally, in Section 9.3, we suggest some promising research directions and future work to address these limitations.

9.1 Summary of Contributions

The main theme of this research is originated from the idea that in order to bridge or at least narrow the *semantic gap* in CBIR, we have to look at the problem from a domain perspective. Instead of focusing on the universality of the retrieval method and finding a single retrieval solution that is applicable for all domains, we focused on domain specific solution to fulfill the users varying search requirements. It is simply not possible with current technology to develop a single retrieval solution to solve all problems in a multitude of application domains. Although, many of the proposed techniques in this field try to achieve this unrealistic goal. As a result, we can see the limited success of the CBIR systems even after a decade of intensive research [13].

Depending on the scope of image domains (e.g., narrow or broad), users search requirements at different levels, and the amount of domain knowledge available, the CBIR exhibits a varying degree of difficulty as discussed in Chapter 1 and Chapter 2 of this thesis. To overcome the difficulties and find specific domain dependent retrieval solutions, we focused on a multi-disciplinary research approach by incorporating ideas originally developed in other fields, such as Machine Learning, IR, and Human Computer Interaction. This is the main driving force for this research. Some of our retrieval solutions utilized various off-line learning techniques to represent images in both global or local concept levels, some exploited on-line learning based on human computer interaction, and in some other cases, we relied on contextual information, i.e., associated annotation of images. This work primarily focused on incorporation and integration of these solutions for the semantic retrieval of natural photographs and medical images. We present our work in this thesis in such a way that the proposed techniques are easily distinguishable based on their domain dependent criteria. In summary, we have made the following contributions:

- For the purpose of searching images in semantically organized collection, such as medical images with different modalities, we proposed a classification-driven image retrieval framework at global concept level based on statistical modeling by utilizing probabilistic multi-class SVMs [205, 206].

- We proposed local concept-based image representation approaches for searching both broad and narrow domain images by utilizing both supervised SVMs based classification and unsupervised SOM based clustering techniques. Images are represented in both local visual and semantic concept spaces such that the representations are robust against classification and quantization errors and in a higher semantic level then the commonly used low-level feature representations in CBIR [209].

- Inspired from ideas of text retrieval domain, we proposed automatic query expansion techniques in CBIR domain to reduce the concept mismatch problem and remove the burden from users to provide feedback information to a system. The proposed techniques are based on the concept correlation and concept similarity analysis at both local feedback and global collection levels. The decision to rely on techniques from text retrieval domain and generalize them to our concept feature-based image domain has proved to be effective [209].

- For the purpose of searching images in a higher semantic level in broad domains, we proposed and developed fusion-based multi-modal image retrieval framework that facilitates both visual and contextual querying [217, 218, 208]. The interactive nature of the framework allows the user's perceived semantics to propagate from one modality to another as well as dynamic fusion of context (text) and content (image) based modalities. The evaluation of this framework in a broad domain general photographic and medical collection showed promising results [217, 217].

- To investigate retrieval effectiveness in a single modality medical image domain and aid dermatologist as a decision support system, we developed a CBIR system in the domain of dermoscopic images [210]. To this end, we proposed a fast and automatic segmentation algorithm for lesion detection by exploiting the domain knowledge. Lesion specific color and texture related features are extracted and finally they are combined in a fusion-based similarity matching function to retrieve database images. The initial retrieval results to find similar images based on an unknown query image were promising [210].

- For empirical analysis of the proposed techniques in different domains, we implemented a prototype retrieval system and conducted exhaustive experiments with different performance measures on a variety of image collections. In addition, by participating in ImageCLEF [57, 58] campaign during the last three years, we were able to perform some of our experiments in standard benchmark collections and compare our performances with different participating groups. Our results were encouraging for the past two years (2006 and 2007) based on both visual and multi-modal ad-hoc retrieval approaches in general photographic and medical image collections as reported in Chapter 8 and in our papers [217, 218].

9.2 Limitations

Due to the main theme of this research, therefore developing retrieval solutions from domain perspective, we already have some form of limitations in our proposed techniques. For example, techniques for narrow domain, such as global and local semantic concept-based retrieval approaches are not extendible in broad domain due to their nature of exploitation of the domain knowledge by supervised learning. In addition, from technical and computational perspectives, we have some limitations in different approaches, which are as follows.

- One of the main limitations of our global concept-based retrieval framework is that we utilized only low-level global features as input to the classifiers to transform them to intermediate level semantic feature spaces. Global features, however, have limitations as they can not adequately capture the subtle details of images, specially in medical domain. Although, it has proved to be quite difficult in medical domain to automatically extract features from local region of interest (ROI) of diagnostic relevance as perceived by physicians [128]. Another limitation is that, for classifier training, a large number of labeled training samples are needed and the training set is fixed during the learning and application stages. If images of a new category are encountered or inserted into the database, then we might need to re-train the whole system with the new category label. So for a dynamic database with many insertions or deletions, this approach is not flexible enough to deal with that issue.

- For the local concept-based feature representation, we limited our approaches by modeling only intermediate level visual concepts in broad and narrow domains and semantic concepts in specific narrow domains (e.g., natural scenery images). Although this limitation is obvious due to the current state of object recognition in broad domain images. It would be more effective, if specific objects, such as person, animal, house, car, etc. can be identified in large collections irrespective to their variations and occlusions. Many researchers from computer vision community are investigating along this line with limited success up to now [126, 103]. However, the main focus of our retrieval approach is to represent images with a soft labeling approach that can exploit concept correlations and overcome the quantization error due to inaccurate object identification. In future, when object recognition techniques will be mature enough to a certain level, our approaches would be easily extendible with higher level semantic concepts.

- The proposed automatic query expansion and interactive RF-based techniques are effective only for a particular query session. The main limitation is that they can not memorize each session and users behavior for effective retrieval in the long run. Another drawback is that, for query expansion techniques based on local analysis, it requires large on-line computation which in some cases may hamper the interactive nature of the system. And for RF-based technique for multi-modal retrieval, users need to provide enough feedback information to make the system work effectively. As already mentioned in Chapter 5, it may create unnecessary burden to the reluctant users who are not willing to provide reasonable feedback at each round or iteration of RF.

- Finally, in majority of our approaches, we more or less focused on improving retrieval effectiveness and in some cases in trade of computational efficiency aspect. As a result, some of our feature extraction methods are computationally expensive and features are high-dimensional. Although, it might not be a problem for the off-line feature computation of database images, it would be problematic for on-line query feature extraction and similarity matching. For example, feature extraction for the fuzzy visual concept vector (described in Section 4.1.3 of Chapter 3) requires large computation to generate the fuzzy membership values of each region in images and the feature dimension is generally higher that depends on the codebook size. A large size codebook provides better feature effectiveness at the expense of higher feature dimension as shown in Section 8.4 of Chapter 8. Although, due to the sparse representation of some of the feature vectors, we were able to exploit inverted file-based indexing technique as described in Section 5.3 of Chapter 5.

9.3 Future Research Directions

Due to the multi-disciplinary and multi-perspective nature of this thesis, we have a good opportunity to expand our work in several directions. There are still many open research issues that need to be solved before the current image retrieval systems in both general photographic and medical domain can be of practical use. In following, we have identified some of the promising directions, which can be extended from our present work in the near future.

- Although some work has been done in developing intermediate level semantic feature representation methods in this thesis as presented in Chapter 3 and

Chapter 4, there is still room for considerable improvement. A successful semantic level retrieval should involve images of complex objects in real life and the system itself has to be able to adapt with the variations of the same objects or scenes. We need to perform more research to automatically identify specific objects in general broad domain images for retrieval in a higher semantic level and pathology bearing region (PBR) in medical images for retrieval based on diagnostic relevance. To achieve these goals, we need to investigate closely the recent advancement in object recognition and image understanding fields.

- For the multimodal retrieval approach described in Chapter 6, we have shown by experimental evaluation in Chapter 8 that one modality can make another modality more informative and more precise, and the combination of content and context information usually improve performances of mixed-mode queries. However, in future we need to focus on constructing a model or formalism to show how much the inclusion of text can contribute to the improvement of image retrieval or vice versa. Another major issue is the scalability and efficiency. Since we use different query and image representation for the dynamic fusion of content and context feature spaces, there is a large computational overhead currently persisting. To overcome this, we need to concentrate more on multidimensional and specially multi feature indexing approach, which might be an interesting topic in CBIR research.

- In our proposed retrieval approaches, the actual data and feature vectors are typically stored in files addressed by names. However, this approach is not efficient and scalable due to the high disk access time and large memory requirement for large image collections. Hence, in future, we need to connect our image retrieval approaches with database research as instead of what to index, database is more concerned with how to index the features. Making image retrieval as a plug-in module in an existing DBMS will not only solve the image data integrity problem and allows dynamic updates, but also it will provide natural integration with features derived from other sources.

- Another future work is the full integration of our image retrieval system with CINDI (Concordia INdexing and DIscovery System) [1], a system for cataloguing,

[1] http://dumbo.encs.concordia.ca:8080/cindimg2

searching, and annotating electronic documents in a digital library, the library being distributed over a computer communication network. The user of CINDI is helped by an expert system that mimics the basic expertise of professional librarians. We are currently working to make CINDI as an integrated digital library with functionalities of searching images as well as documents for varied collections.

- Last of all, we want to continue our participation in ImageCLEF [57, 58, 55] in upcoming years by extending our multimodal retrieval approaches from other perspectives, such as addition of long term learning in RF, multi-lingual search approaches, and use domain related thesaurus for query expansion. The continuing participation in ImageCLEF is very essential as we can compare our image retrieval techniques easily with other participating groups and investigate the real retrieval effectiveness in benchmark collections.

Bibliography

[1] Smeulder, A., Worring, M., Santini, S., Gupta, A. and Jain, R., Content-Based Image Retrieval at the End of the Early Years, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**: 1349–1380, 2000.

[2] Rui, Y., Huang, T. S. and Chang S. F. Image Retrieval: Current Techniques, Promising Directions and Open Issues, *Journal of Visual Communication and Image Representation* **10 (4)**: 39-62, 1999.

[3] Tamura, H. and Yokoya, N. Image database systems: A survey, *Pattern Recognition* **17 (1)**: 29-43, 1984.

[4] Fidel, R., Hahn, T. B., Rasmussen, E. M. and Smith, P. J. (eds) *Challenges in Indexing Electronic Text and Images*, ASIS Monograph Series, Learned Information, Inc., 1994.

[5] Aigrain, D., Zhang, H. and Petkovic, D. Content-Based Representation and Retrieval of Visual Media: A State of the Art Review, *Multimedia Tools and Applications* **3**: 178–202, 1996.

[6] Antani, S., Kasturi, R. and Jain, R. A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video, *Pattern Recognition* **35 (4)**: 945-965, 2002.

[7] Yoshitaka, A. and Ichikawa, T. A survey on content-based retrieval for multimedia databases, *IEEE Transactions on Knowledge and Data Engineering* **11 (1)**: 81-93, 1999.

[8] Kato, T. Database architecture for content-based image retrieval, in I. K. Sethi and R. J. Jain (eds), Image Storage and Retrieval Systems, *Proceedings of SPIE*, San Jose, CA, USA, Vol. 1662, pp. 112-123, 1982.

[9] Gupta, A. and Jain, R. Visual information retrieval, *Communications of the ACM* **40 (5)**: 7079, 1997.

[10] Del Bimbo A. *Visual Information Retrieval*, Morgan Kaufmann Publishers, San Francisco, Calif., 1999.

[11] Castelli, V. and Bergman, L. D. *Image Databases: Search and Retrieval of Digital Imagery*, John Wiley & Sons, Inc., 2002.

[12] Smith, J. R. *Integrated Spatial and Feature Image Systems: Retrieval, Analysis and Compression*, PhD thesis, Graduate School of Arts and Sciences, Columbia University, 1997.

[13] Eakins John, P. Towards Intelligent image retrieval, *Pattern Recognition* **35**: 3–14, 2002.

[14] Naphade, M. R. and Huang, T. S. Extracting semantics from audiovisual content: The final frontier in multimedia retrieval, *IEEE Transactions on Neural Networks* **13 (4)**: 793-810, 2002.

[15] Liua, Y., Zhang, D., Lu, G. and Ma W. Y. A survey of content-based image retrieval with high-level semantics, *Pattern Recognition* **40**: 262–282, 2007.

[16] Faloutsos, C., Barber, R., Flickner, M., Hafner, J., Niblack, W., Petkovic, D. and Equitz, W. Efficient and effective querying by image content, *Journal of Intelligent Information System* **3 (34)**: 231-262, 1994.

[17] Gupta, A. and Jain, R. Visual information retrieval, *Communications of the ACM* **40 (5)**: 70-79, 1997.

[18] Pentland, A., Picard, R. W. and Sclaroff, S. Photobook: Tools for contentbased manipulation of image databases, *Storage and Retrieval for Image and Video Databases II*, Vol. 2185 of *Proceedings of SPIE*, San Jose, CA, USA, pp. 34-47, 1994.

[19] Smith, J. R. and Chang, S. F. VisualSEEk: A fully automated content-based image query system, *Proceedings of the 4th International ACM Multimedia Conference (ACM MM 96)*, Boston, MA, USA, pp. 87-98, 1996.

[20] Huang, T. S., Mehrotra, S. and Ramchandran, K. Multimedia analysis and retrieval system (MARS) project, *Proceedings of 33rd Annual Clinic on Library Application on Data Processing - Digital Image Access and Retrieval*, Urbana-Champaign, IL, USA, 1996.

[21] Carson, C., Belongie, S., Greenspan, H. and Malik, J. Blobworld: Image segmentation using expectation-maximization and its application to image querying, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24 (8)**: 1026-1038, 2002.

[22] Ma, W. Y. and Manjunath, B. S. NETRA: A toolbox for navigating large image databases, *Proceedings of IEEE International Conference on Image Processing (ICIP 97)*, Vol. 1, Santa Barbara, CA, USA, pp. 568571. 1997.

[23] Ardizzoni, S., Bartolini, I. and Patella, M. Windsurf: Region-based image retrieval using wavelets, *in DEXA Workshop*, pp. 167–173, 1999.

[24] Wang, J. Z., Liu, J. and Wiederhold, G. SIMPLIcity: Semantics-sensitive integrated matching for picture libraries, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23 (9)**: 947963, 2001.

[25] Markkula, M. and Sormunen, E. End-user searching challenges indexing practices in the digital newspaper photo archive, *Information Retrieval* **1 (4)**: 259-285, 2000.

[26] Cox, I. J., Miller, M. L., Omohundro, S. M. and Yianilos, P. N. Target testing and the PicHunter bayesian multimedia retrieval system, *Proceedings of 3rd Forum on Research and Technology Advances in Digital Libraries (ADL96)*, Washington DC, USA, pp. 66-75, 1996.

[27] Jaimes A. and Chang, S.F. A Conceptual Framework for Indexing Visual Information at Multiple Levels, *IS&T/SPIE Internet Imaging* Vol. 3964, San Jose, CA January, 2000.

[28] Zhu, L., Zhang, A., Rao, A. and Srihari, R. Theory of keyblock-based image retrieval, *ACM Transactions on Information Systems* **20 (2)**: 224–257, 2002.

[29] Müller, H., Squire, D., Müller M, Pun, W. and Thierry. Efficient access methods for content-based image retrieval with inverted files, *Proc. SPIE*, Vol. 3846, pp. 461–472, 1999.

[30] Jing, F., Li, M., Zhang, H.J. and Zhang, B. An efficient and effective region-based image retrieval framework, *IEEE Transaction on Image Processing*, **13**: 699–709, 2004.

[31] Lim, J. H. Building visual vocabulary for image indexation and query formulation, *Pattern Analysis and Applications (Special Issue on Image Indexation)*, **4 (2/3)**:125-139, 2001.

[32] Kohonen, T. *Self-Organizing Maps*, Vol. 30 of *Springer Series in Information Sciences*, third edn, Springer-Verlag. 2001.

[33] Kohonen, T. Self-organizing formation of topologically correct feature maps, *Biological Cybernetics* **43 (1)**: 59-69. 1982.

[34] Gersho, A. and Gray, R. M. *Vector Quantization and Signal Compression*, Norwell, MA: Kluwer, 1992.

[35] Zhang, H. and Zhong, D. A scheme for visual feature based image indexing, *Storage and Retrieval for Image and Video Databases III (SPIE)*, Vol. 2420 of *SPIE Proceedings Series*, San Jose, CA, USA, 1995.

[36] Laaksonen, J., Koskela, M. and Oja, E. PicSOM: Self-Organizing Image Retrieval With MPEG-7 Content Descriptors, *IEEE Transection Neural Networks*, **13 (4)**: 841–853, 2002.

[37] Suganthan, P. N. Shape indexing using self-organizing maps, *IEEE Transactions on Neural Networks* **13 (4)**: 835840, 2002.

[38] Honkela, T., Kaski, S., Lagus, K. and Kohonen, T. WEBSOMself-organizing maps of document collections, *Proceedings of WSOM97, Workshop on Self-Organizing Maps*, Espoo, Finland, pp. 310-315, 1997.

[39] Ong, S. H., Yeo, N. C., Lee, K. H., Venkatesh, Y. V. and Cao, D. M. Segmentation of color images using a two-stage self-organizing network, *Image and Vision Computing* **20 (4)**: 279-289, 2002.

[40] Koskela, M., Laaksonen, J. and Oja, E. Implementing Relevance Feedback as Convolutions of Local Neighborhoods on Self-Organizing Maps, *Proceedings of the International Conference on Artificial Neural Networks (ICANN 2002)*, Madrid, Spain, pp. 981-986, 2002.

[41] Vesanto, J. SOM-Based Data Visualization Methods, *Intelligent Data Analysis* **3 (2)**: 111–126, 1999.

[42] Yen, G.G. and Zheng, W. Ranked centroid projection: a data visualization approach for self-organizing maps, *Proc. IEEE International Joint Conference on Neural Networks, (IJCNN'05)*, pp. 1587–1592, 2005.

[43] Hartigan, J. A. and Wong, M.A. Algorithm AS136: A kmeans Clustering Algorithm, *Applied Statistics,* **28**: 100–108, 1979.

[44] Baeza-Yates, R. and Ribeiro-Neto, B. *Modern·Information Retrieval*, Addison-Wesley, 1999.

[45] Salton, G. and McGill, M. J. *Introduction to Modern Information Retrieval*, Computer Science Series, McGraw-Hill, 1983.

[46] Attar, R. and Fraenkel, A. S. Local feedback in full-text retrieval systems, *Journal ACM* **24 (3)**: 397–417, 1977.

[47] Qiu, Y. and Frei, H. P. Concept Based Query Expansion, *Proc. of the 16th Int. ACM SIGIR Conf. on R&D in Information Retrieval*, Pittsburgh, SIGIR Forum, ACM Press, June, 1993.

[48] Xu, J. and Croft, B. Improving the effectiveness of information retrieval with local context analysis, *ACM Transection on Information System* **18 (1)**: 79–112, 2000.

[49] Crouch, C. J. An approach to the automatic construction of global thesauri, *Information Processing & Management* **26 (5)**: 629–640, 1990.

[50] Jack, M., Gerald W. and Barbara Z. An evaluation of query expansion by the addition of clustered terms for a document retrieval system. *Information Storage and Retrieval*, **8**: 329–348, 1972.

[51] Lesk, M.E., Word-word association in document retrieval systems, *American Documentation*,**20 (1)**: 27–38, 1969.

[52] Peat, H. J., Willett, P., The limitations of term co-occurrence data for query expansion in document retrieval systems, *Journal of the ASIS*, **41 (5)**: 378–83, 1991.

[53] Yasushi O., Tetsuya M. and Kiyohiko K. A fuzzy document retrieval system using the keyword connection matrix and a learning method, *Fuzzy Sets and Systems archive* **39 (2)**: 163–179, 1991.

[54] Grubinger M., Clough P., Müller, H. and Deselears, T. The IAPR-TC12 benchmark: A new evaluation resource for visual information systems. *In International Workshop OntoImage2006 Language Resources for Content-Based Image Retrieval, held in conjunction with LREC06*, Genoa, Italy, pp. 13–23, May 2006.

[55] Clough, P., Grubinger, M., Deselaers, T., Hanbury, A., Müller, H. : Overview of the ImageCLEF 2006 photo retrieval and object annotation tasks. *Seventh Workshop of the Cross-Language Evaluation Forum.* Alicante, Spain, *Proceedings of LNCS*, Vol. 4730, pp. 579–594, 2006.

[56] Müller, H., Deselaers, T., Lehmann, T., Clough, P., Hersh, W. : Overview of the ImageCLEFmed 2006 medical retrieval and annotation tasks. *Seventh Workshop of the Cross-Language Evaluation Forum.* Alicante, Spain, *Proceedings of LNCS*, Vol. 4730, pp. 595–608, 2006.

[57] Grubinger, M., Clough, P., Hanbury, A. and Müller, H. Overview of the ImageCLEF 2007 Photographic Retrieval Task, *Working Notes of the 2007 CLEF Workshop*, Sep., 2007, Budapest, Hungary. To appear in Proc. of LNCS.

[58] Müller, H., Deselaers, T., Kim, E., Kalpathy C., Jayashree D., Thomas M., Clough, P. and Hersh, W. Overview of the ImageCLEFmed 2007 Medical Retrieval and Annotation Tasks, *Working Notes of the 2007 CLEF Workshop*, Sep., 2007, Budapest, Hungary. To appear in Proc. of LNCS.

[59] Stricker, M. and Orengo, M. Similarity of color images *in Storage and Retrieval for Image and Video Databases III* (Niblack, W R and Jain, R C, eds), *Proc SPIE*, Vol. 2420, pp. 381–392, 1995.

[60] Pass, G. and Zabih, R. Histogram refinement for content-based image retrieval. *IEEE Workshop on Applications of Computer Vision*, pp. 96-102, December, 1996.

[61] Huang, J., Kumar, S. R., Mitra, M. Zhu, W. J. and Zabih. R. Image indexing using color correlograms, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 762-768, 1997.

[62] Manjunath, B. S., Salembier, P., Sikora, T. (eds.) *Introduction to MPEG-7-Multimedia Content Description Interface*, John Wiley & Sons Ltd. pp. 187–212, 2002.

[63] ISO/IEC JTC1/SC29/WG11/W3703 *MPEG-7 Multimedia Content Description Interface Part 3 Visual*, October, 2000.

[64] Haralick, R. M. Statistical and Structural Approaches to Texture, *Proceedings of the IEEE*, Vol. 67 (5), pp. 786–804, 1979.

[65] Tamura, H., Mori, S., Yamawaki, T. Textural Features Corresponding to Visual Perception, *IEEE Trans. on Systems, Man and Cybernetics* **8 (6)**, 1978.

[66] Picard, R. W. and Minka, T. P. Vision Texture for Annotation, *Technical Report, MIT Media Laboratory*, No. 302. Available at http://citeseer.nj.nec.com/picard95vision.html

[67] Voulgaris, G. and Jiang, J. Texture-based image retrieval in wavelets compressed domain, *Proceedings International Conference on Image Processing*, Vol. 2, pp. 125–128, 2001.

[68] Manjunath, B. S. and Ma, W. Y. Texture features for browsing and retrieval of image data, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **18 (9)**: 837–842, 1996.

[69] Chellappa, R. and Chatterjee, S. Classification of Textures Using Gaussian Markov Random Fields, *IEEE Trans. Acoustic, Speech and Signal Processing*, **33 (4)**: 959–963, 1985.

[70] Chellappa, R., Kashyap, R. L. and Manjunath. B. S. Model-based texture segmentationand classification, *In Handbook of Pattern Recognition and Computer Vision*, pp. 277–310. World Scientific, Singapore, 1993.

[71] Mehrotra, R. and Gary, J. E. Similar-shape retrieval in shape data management, *IEEE Compututer* **28 (9)**: 5762, 1995.

[72] Hoffman, M. and Wong, E. Content-based image retrieval by scale-space object boundary shape representation, *In Storage and Retrieval for Image and Video Databases*, pp. 86–97, 2000.

[73] Zhang D., Lu, G. Evaluation of MPEG-7 shape descriptors against other shape descriptors, *Multimedia Systems*, **9 (1)**: 15–30, 2003.

[74] Santini, S. and Jain, R. Similarity measures, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **21 (9)**: 871–883, 1999.

[75] Tversky, A. and Gati, I. Similarity, Separability, and the Triangle Inequality, *Phychol Review* **89**: 123–154, 1982.

[76] Swain, M. J. and Ballard, D. H. Color indexing *International Journal of Computer Vision*, **7 (1)**: 11–32, 1991.

[77] Hafner, J., Sawhney, H. S., Equitz, W., Flickner, M. and Niblack, W. Efficient color histogram indexing for quadratic form distance functions, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **17 (7)**: 729-736, 1995.

[78] Vasconcelos, N. and Lippman, A. A Unifying View of Image Similarity, *Proceedings of 15th International Conference on Pattern Recognition*, Barcelona, Spain, Vol. 1, pp. 38–41, September 2000.

[79] Puzicha, J., Buhmann, J., Rubner, Y., Tomasi, C. Empirical evaluation of dissimilarity measures for color and texture, *Proceedings of the Seventh IEEE International Conference on*, Vol.2, pp. 1165–1172, 1999.

[80] Aksoy, S. and Haralick, R.M. Probabilistic vs. geometric similarity measures for image retrieval, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Vol.2, pp. 357–362, 2000.

[81] Kailath, T. The divergence and Bhattacharyya distance measures in signal selection, *IEEE Transection on Communication*, **15 (1)**: 52–60, 1967.

[82] Comaniciu, D., Meer, P. and Tyler, D. Dissimilarity computation through low rank corrections, *Pattern Recognition Letters* **24**: 227-236, 2003.

[83] Zhou, S. K. and Chellappa, R. From sample similarity to ensemble similarity: probabilistic distance measures in reproducing kernel hilbert space, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28 (6)**: 917-929, 2006.

[84] Gaede, V. and Gunther, O. Multidimensional Access Methods, *ACM Computing Surveys*, **30 (2)**: 170–231, 1998.

[85] Böhm C., Berchtold, S., Keim, D. A. Searching in High-dimensional Spaces: Index Structures for Improving the Performance of Multimedia Databases, *ACM Computing Surveys* **30 (3)**: 322–373, 2001.

[86] Guttman, A. R-trees: A dynamic index structure for spatial searching, *Proceedings of the 1984 ACM SIGMOD International Conference on Management of Data*, Boston, MA, USA, pp. 47-57, 1984.

[87] Kriegel, N. B. H.-P., Scheider, R. and Seeger, B. The R*-tree: an efficient and robust access method for points and rectangles, *Proceedings of ACM SIGMOD International Conference on the Management of Data*, Atlantic City, NJ, USA, pp. 322-331, 1990.

[88] White, D. A. and Jain, R. Similarity indexing with the SS-tree, *Proceedings of 12th IEEE International Conference on Data Engineering,* New Orleans, LA, USA, pp. 516-523, 1996.

[89] Duffing, G. and Smaïl, M. A novel approach for accessing partially indexed image corpora, *Proceedings of Fourth International Conference on Visual Information Systems (VISual 2000)*, Lyon, France, pp. 244-256, 2000.

[90] Chen, J. Y., Bouman, C. A. and Allebach, J. P. Fast image database search using tree-structured VQ, *Proceedings of IEEE International Conference on Image Processing (ICIP 97)*, Vol. 2, Santa Barbara, CA, USA, pp. 827-830, 1997.

[91] Kruskal, J. B. Multidimensional scaling, *Psychometrika* **29 (1)**: 1-27, 1964.

[92] Chapelle, O., Haffner, P., Vapnik, V. SVMs for histogram-based image classification, *IEEE Transaction on Neural Networks*, **10 (5)**: 1055–1064, 1999.

[93] Chang, E., Kingshy, G., Sychay, G. and Gang, W. CBSA: Content-based Soft Annotation for Multimodal Image Retrieval Using Bayes Point Machines, *IEEE Transactions on Circuits and Systems for Video Technology* **13**: 26–38, 2003.

[94] Shi, R., Feng, H., Chua, T. S. and Lee, C. H. An adaptive image content representation and segmentation approach to automatic image annotation, *Proceedings of International Conference on Image and Video Retrieval (CIVR)*, pp. 545-554, 2004.

[95] Jin, W., Shi, R. and Chua, T. S. A semi-naive bayesian method incorporating clustering with pair-wise constraints for auto image annotation, *Proceedings of the ACM Multimedia*, pp. 336–339, 2004.

[96] Vogel, J. and Schiele, B. Semantic Modeling of Natural Scenes for Content-Based Image Retrieval, *International Journal of Computer Vission* **72 (2)**: 133–157, 2007.

[97] Town, C. and Sinclair, D. Content-based image retrieval using semantic visual categories, *Technical report*, 2000.14, AT&T Research, Cambridge, 2000.

[98] Szummer, M. and Picard, R.W. Indoor-Outdoor Image Classification, *Proceedings of IEEE International Workshop on Content-based Access of Image and Video Databases*, pp. 42–51, 1998.

[99] Vailaya, A., Figueiredo, M., Jain, A. and Zhang, H.J. Image Classification for Content-Based Indexing, *IEEE Transaction on Image Processing* **10 (1)**: 117–130, 2001.

[100] Campbell, N. W., Mackeown, W. P. J., Thomas B. T. and Troscianko T. Interpreting image databases by region classification, *Pattern Recognition* **30 (4)**: 555–567, 1997.

[101] Duygulu, P., Barnard, K., Freitas, N., Forsyth, D. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary, *Proceedings of Seventh European Conference on Computer Vision*, pp. 97–112, 2002.

[102] Li, Y., Bilmes, J. and Shapiro, L. G. Object class recognition using images of abstract regions, *Proceedings of 17th International Conference on Pattern Recognition*, Vol. 1, pp. 40–43, August 2004.

[103] Oliva, A., Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope, *International Journal of Computer Vision* **42 (3)**: 145–175, 2001.

[104] Salton, G., Buckley, C., Improving Retrieval Performance by Relevance Feedback, *Journal of the American Society for Information Science* **41 (4)**: 288–297, 1999.

[105] Ruthven, I. and Lalmas, M. A survey on the use of relevance feedback for information access systems, *Knowledge Engineering Review* **18 (2)**: 95–145, 2003.

[106] Rocchio, J. J. Relevance feedback in information retrieval. *In the SMART retrieval system -experiments in automatic document processing,*(G. Salton ed). pp. 313–323, Englewood Cliffs, NJ, Prentice Hall, Inc., 1971.

[107] Ide, E. New experiments in relevance feedback. *In The SMART retrieval system - experiments in automatic document processing* (G. Salton ed). Chapter 16, pp. 337–354. 1971.

[108] Rui, Y. and Huang, T. S. Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval, *IEEE Transactions on Circuits and Systems for Video Technology* **8 (5)**: 644–655, 1999.

[109] Zhou, X. S. and Huang, T. S. Relevance feedback for image retrieval: a comprehensive review, *Multimedia Systems* **8 (6)**: 536–544, 2003.

[110] Rui, Y., Huang, T. S. and Mehrotra, S. Content-based image retrieval with relevance feedback in MARS, *Proceedings of IEEE International Conference on Image Processing*, Vol. 2, pp. 815–818, 1997.

[111] Ishikawa, Y., Subramanya, R. and Faloutsos, C. MindReader: Querying Databases Through Multiple Examples, *Proceedings of 24th International Conference on Very Large Databases*, New York, pp. 24–27, 1998.

[112] Tong, S. and Chang, E. Support vector machine active learning for image retrieval, *Proceedings of the ninth ACM international conference on Multimedia,* Vol. 9, pp. 107–118, 2001.

[113] Zhang, H., Chen, Z., Li, M. and Su, Z. Relevance Feedback and Learning in Content-Based Image Search, *World Wide Web: Internet and Web Information Systems*, **6**: 131–155, 2003.

[114] Hu, Y. C., Zhu, X., Zhang, H., Yang, Q. A unified framework for semantics and feature based relevance feedback in image retrieval systems, *Proceedings of ACM International Conference on Multimedia*, pp. 31-37, 2000.

[115] Chen, N. A Survey of Indexing and Retrieval of Multimodal Documents: Text and Images, *Technical Report*, 2006-505, Queen's University.

[116] Santini, S. Multimodal search in collections of images and text. *Journal of Electronic Imaging*, **11 (4)**: 455–468, 2002.

[117] Westerveld, T. Image retrieval: Content versus context. *Proceedings of Computer- Assisted Information Retrieval*, Vol. 1, Paris, France, pp. 276–284, 2000.

[118] Paek, S., Sable, C., Hatzivassiloglou, V., Jaimes, A., Schiman, B., Chang, S. F. and McKeown, K. Integration of visual and text-based approaches for the content labeling and classification of photographs, *Proceedings of the ACM SIGIR Workshop on Multimedia Indexing and Retrieval (SIGIR-99)*, 1999.

[119] Zhou, X. S. and Huang, T. S. Unifying keywords and visual contents in image retrieval, *IEEE Multimedia*, **4 (2)**: 23–33, June 2002.

[120] Sclaroff, S., La Cascia, M., Sethi, S. and Taycher, L. Unifying Textual and Visual Cues for Content-Based Image Retrieval on the World Wide Web, *Computer Vision and Image Understanding*, **4 (2)**: 86-98, 1999.

[121] Sclaroff, S., Taycher, L. and La Cascia, M. ImageRover: A content-based image browser for the world wide web, *Proceedings of IEEE International Workshop on Content-Based Access of Image and Video Libraries*, 1997.

[122] Srihari. R. K. Intelligent indexing and semantic retrieval of multimodal documents, *Information Retrieval*, **2**: 245–275, 2000.

[123] Rong Zhao and William I. Grosky, Narrowing the Semantic GapImproved Text-Based Web Document Retrieval Using Visual Features, *IEEE Transaction on Multimeda*, **4 (2)**: 189–200, 2002.

[124] Chen, Z., Liu, W. Y., Zhang, F., Li, M. J. and Zhang, H. J. Web mining for web image retrieval, *Journal of the American Society for Information Science and Technology*, **52 (10)**: 831–839, 2001.

[125] Lu, Y., Zhang, H., Liu, W. Y. and Hu. C. Joint semantics and feature based image retrieval using relevance feedback, *IEEE Transactions on Multimedia*, **5 (3)**: 339–347, 2003.

[126] Barnard, K., Duygulu, P., Forsyth, D., Freitas, N., Blei, D. M. and Jordan, M. I. Matching words and pictures, *Journal of Machine Learning Research (Special Issue on Machine Learning Methods for Text and Images)*, **3**: 1107-1135, 2003.

[127] Shi, R., Feng, H., Chua, T. S. and Lee, C. H. An adaptive image content representation and segmentation approach to automatic image annotation, *Proceedings of International Conference on Image and Video Retrieval (CIVR)*, pp. 545-554, 2004.

[128] Müller, H., Michoux, N., Bandon, D. and Geissbuhler, A. A Review of Content-Based Image Retrieval Systems in Medical Applications Clinical Benefits and Future Directions, *International Journal of Medical Informatics* **73**: 1–23, 2004.

[129] Wong, T. C. *Medical Image Databases.* New York, LLC: Springer Verlag, 1998.

[130] Tang, L. Hanka, R. and Ip, H. A review of intelligent content-based indexing and browsing of medical images, *Journal oh Health Informatics* **5**: 40–49, 1999.

[131] Tagare, H. D. and Jaffe, C. C., Duncan, J. Medical Image Databases: a content-based retrieval approach, *Journal of the American Medical Informatics Association* **4**: 184–98, 1997.

[132] Florea, F. Müller, H. Rogozan, A., Geissbuhler, A. and Darmoni, S. Medical image categorization with MedIC and MedGIFT, *Proceedings of Medical Informatics Europe (MIE 2006)*, Maastricht, Netherlands, pp. 3–11, 2006.

[133] Lehmann, T. M., Güld, M.O., Thies, C., Fischer, B., Keysers, M., Kohnen, D., Schubert, H. and Wein, B. B. Content-based image retrieval in medical applications for picture archiving and communication systems, *Proceedings of SPIE*, Vol. 5033, pp. 109–117, 2003.

[134] Güld, M. O., Kohnen, M., Schubert, H., Wein, B. B. and Lehmann, T. M. Quality of DICOM header information for image categorization. *Proceedings of SPIE*, Vol. 4685, pp. 280–287, 2002.

[135] Lowe, H. J., Antipov, I., Hersh, W. and Smith, C. A. Towards knowledge-based retrieval of medical images: the role of semantic indexing,image content representation and knowledge-based retrieval, *Proceedings of the Annual Symposium of the American Society for Medical Informatics (AMIA)*, pp.882–886, 1998.

[136] Liu, Y. and Dellaert, F. Classification-driven medical image retrieval, *Proceedings of the ARPA Image Understanding Workshop*, 1997.

[137] Shyu, C. R., Brodley, C. E., Kak, A. C., Kosaka, A., Aisen, A. M. and Broderick, L. S. ASSERT: a physician-in-the-loop content-based image retrieval system for HRCT image databases, *Comput Vision and Image Understanding*, **75**: 111–132, 1999.

[138] Long, L. R. and Thoma, G. R. Landmarking and feature localization in spine x-rays, *Journal of Electronic Imaging* **10 (4)**: 939–956, 2001.

[139] Tang, L. H., Hanka, R., Ip, H. H. S. and Lam, R. Extraction of semantic features of histological images for content-based retrieval of images, *Proceedings of SPIE*, Vol. 3662, pp. 360–368, 2000.

[140] Comaniciu, D., Meer, P. and Foran, D. Image Guided Decision Support System for Pathology, *Machine Vision and Applications*, **11**: 213–224, 1999.

[141] Lehmann, T. M., Wein, B. B., Dahmen, J., Bredno, J., Vogelsang, F. and Kohnen, M. Content–based image retrieval in medical applications-A novel multistep approach. *Proceedings of SPIE*, Vol. 3972, pp. 312–320, 2000.

[142] Müller H., Rosset, A., Vallee, J. and Geissbuhler, A. Integrating content-based visual access methods into a medical case database, *Proceedings of Medical Informatics Europe (MIE 2003)*, St Malo, France, pp. 480–485, 2003.

[143] Orphanoudakis, S. C., Chornaki, C. and Kostomanolakis, S. I^2C–a system for the indexing, storage and retrieval of medical images by content, *Medical Informatics*, **19 (2)**: 109–122, 1994.

[144] Mojsilovic, A. and Gomes, J. Semantic based image categorization, browsing and retrieval in medical image databases, *Proceedings of IEEE International Conference of Image Processing* **3**: 145–148, 2002.

[145] Lehmann, T. M., Güld, M. O., Deselaers, T., Keysers, D., Schubert, H., Spitzer, K., Ney, H. and Wein, B. B. Automatic categorization of medical images for content-based retrieval and data mining, *Computerized Medical Imaging and Graphics* **29**: 143-155, 2005.

[146] Rigel, D. S. and Carucci, J. A. Malignant melanoma: prevention, early detection, and treatment in the 21st century, *A Cancer Journal for Clinicians* **50**: 215-236, 2000.

[147] Binder, M., Schwarz, M., Winkler, A., Steiner, A., Kaider, A., Wolff, K. and Pehamberger, H. Epiluminescence microscopy. A useful tool for the diagnosis of pigmented skin lesions for formally trained dermatologists, *Archives of Dermatology* **132 (3)**: 286–291, 1995.

[148] Menzies, S. W., Crotty, K., Ingvar, C. and McCarthy, W. *An atlas of surface microscopy of pigmented skin lesions dermoscopy*, 2nd Edition & CDROM quiz, Sydney, McGraw-Hill Book Co., 2003.

[149] Stolz, W., Riemann, A., Cognetta, A. B., Pillet, L. et al, ABCD rule of dermatoscopy: a new practical method for early recognition of malignant melanoma, *European Journal of Dermatolgy* **4**: 521–527, 1994.

[150] Johr, R. H. Dermoscopy: alternative melanocytic algorithms-the ABCD rule of dermatoscopy, Menzies scoring method, and 7-point checklist, *Clinical Dermatology* **20 (3)**: 240–247, 2002.

[151] Binder, M., Kittler, H., Seeber, A. and Steiner, A. Epiluminescence microscopy-based classification of pigmented skin lesions using computerized image analysis and an artificial neural network, *Melanoma Reseasrch* **8**: 261-266, 1998.

[152] Dreiseitl, S., Ohno-Machado, L., Kittler, H., Vinterbo, S., Billhardt, H. and Binder, M. A comparison of machine learning methods for the diagnosis of pigmented skin lesions, *Journal of Biomedical Informatics* **34** **(1)**: 28–36, 2001.

[153] Maglogiannis, I. G. and Zafiropoulos, E. P. Characterization of digital medical images utilizing support vector machines, *BMC Medical Informatics and Decision Making* **4** **(4)**, 2004.

[154] Saugeon, P. S., Guillod, J. and Thiran, J. P. Towards a computer-aided diagnosis system for pigmented skin lesions *Computerized Medical Imaging and Graphics*, **27**: 65-78, 2003.

[155] Ganster, H., Pinz, A., Rhrer, R., Wildling, E., Binder, M. and Kittler, H. Automated Melanoma Recognition *IEEE Transaction on Medical Imaging*, **20** **(3)**: 233-239, 2001.

[156] Pompl, R., Bunk, W., Horsch, A., Abmayr, W., Morfill, G., Brauer, W. and Stolz, W. Computer vision of melanocytic lesions using MELDOQ, *Proceedings of 6th Congress Int. Soc. Skin Imaging*, London, Skin Research and Technology, Vol. 5, pp. 150, 1999.

[157] Binder, M., Kittler, H., Seeber, A. and Steiner, A. Epiluminescence microscopy-based classification of pigmented skin lesions using computerized image analysis and an artificial neural network, *Melanoma Research* **8**: 261-266, 1998.

[158] Xu, L., Jackowski, M., Goshtasby, A., Roseman, D., Bines, S., Yu, C., Dhawan, A. and Huntley, A. Segmentation of skin cancer images, *Image and Vision Computing*, **17**: 65-74, 1999.

[159] Umbaugh, S. E., Moss, R. H., Stoecker, W. V. and Hance, G. A. Automatic color segmentation algorithms: With application to skin tumor feature identification, *IEEE Engineering in Medicine and Biology Magazine* **12** **(3)**: 75-82, 1993.

[160] Miyahara, M. and Yoshida, Y. Mathematical transform of (r,g,b) color data to Munsell(h,v,c) color data, *SPIE Proceedings in Visual Communication and Image Processing*, Vol. 1001, pp. 650–657, 1988.

[161] Gong, Y., Proietti, G., Faloutsos, C. Image Indexing and Retrieval Based on Human Perceptual Color Clustering, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, p. 578, 1998.

[162] Ridler, T. W. and Calvard, S. Picture thresholding using an iterative selection method, *IEEE Transactions on Systems, Man and Cybernetics* **SMC-8**: 630-632, 1978.

[163] Gonzalez, R. C. and Woods, R. E. *Digital Image Processing*, 2nd ed., Prentice Hall, 2002.

[164] "Dermnet: the dermtologist's image resource", Dermatology Image Atlas, available at http://www.dermnet.com/

[165] Vapnik, V. *Statistical Learning Theory*, New York, Wiley, 1998.

[166] Burges, C. A tutorial on support vector machines for pattern recognition, *Data Mining and Knowledge Discovery* **2 (2)**: 121–167, 1998.

[167] Christianini, N. and Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, UK, 2000.

[168] Pontil, M. and Verri, A. Properties of Support Vector Machines, *MIT Artificial Intelligence Laboratory Report*, A. I. Memo No. 1612, August 1997.

[169] Joachims, T. Text categorization with support vector machines: Learning with many relevant features, *Proceedings of the European Conference on Machine Learning*, pp. 137–142, 1998.

[170] Phillips, P. J. Support Vector Machines Applied to Face Recognition, *Advances in Neural Image Information Processing Systems* Vol. 11, MIT Press, 1999.

[171] Cortes, C. and Vapnik, V. Support vector networks, *Machine Learning* **20 (3)**: 273–297, 1995.

[172] Fukunaga, K. *Introduction to Statistical Pattern Recognition.*, Second ed. Academic Press, 1990.

[173] Duda, R. O., Hart, P. E. and Stork, D. G. *Pattern Classification*, 2nd ed. Canada: John Wiley & Sons Ltd., 2001.

[174] Jain, A. K., Murty, M. N. and Flynn, P. J. Data clustering: a review, *ACM Computing Surveys*, **31 (3)**: 264–323, 1999.

[175] Wu, T. F., Lin, C. J. and Weng, R. C. Probability Estimates for Multi-class Classification by Pairwise Coupling, *J. of Machine Learning Research* , **5**: 975–1005, 2004.

[176] Krebel, U. Pairwise classification and support vector machines. *Adv in kernel methods: support vector learning*, Cambridge, MA: MIT Press, pp. 255–268, 1999.

[177] Hsu, C. W. and Lin, C. J. A comparison of methods for multi-class support vector machines, *IEEE Transaction on Neural Network* **13 (2)**: 415–425, 2002.

[178] Kittler, J., Hatef, M., Duin, R.P.W. and Matas, J. On combining classifiers, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20 (3)**: 226–239, 1998.

[179] Xu, L., Krzyzak, A. and Suen C. Y. Methods of combining multiple classifiers and their applications to handwriting recognition, *IEEE Transactions on System, Man, and Cybernatics* **23 (3)**: 418–435, 1992.

[180] Hansen, L. K. and Salamon, P. Neural Network Ensembles, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **12 (10)**: 993–1001, 1990.

[181] Cho, S.B. and Kim, J. H. Combining Multiple Neural Networks by Fuzzy Integral for Robust Classification, *IEEE Trans Syst Man Cybernetics* **25 (2)**: 380–384, 1995.

[182] Chang, C. C and Lin C. J. LIBSVM : a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm, 2001.

[183] Hotelling, H. Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology* **24**: 498-520, 1933.

[184] Jain, A. K. and Bhandrasekaran, B. Dimensionality and sample size considerations in pattern recognition practice, *Handbook of Statistics*, Vol. 2, pp. 835–855, 1987.

[185] Friedman, J. Regularized Discriminant Analysis, *Journal of American Statistical Association*, **84**: 165–175, 2002.

[186] Bezdek, J. C and Pal, S. K. eds., *Fuzzy Models for Pattern Recognition: Methods that Search for Structures in Data*, NY: IEEE Press, 1992.

[187] Bezdek, J. C., Pal, M. R., Keller J. and Krisnapuram R. *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*, Kluwer Academic Publishers, Boston, 1999.

[188] Bezdek, J. C. *Pattern Recognition with Fuzzy Objective Function Algorithms*, New York: Plenum, 1981.

[189] Yang, J. F., Hao, S. S., Chung, P. C. Color image segmentation using fuzzy C-means and eigenspace projections, *Signal Processing*, **82**: 461-472, 2002.

[190] Han J. and Ma, K. K. Fuzzy Color Histogram and Its Use in Color Image Retrieval, *IEEE Transactions on Image Processing* **11 (8)**: 944–952, 2002.

[191] Ahalt, S.C., Krishnamurthy, A.K.,. Chen, P., Melton, D.E., Competitive learning algorithms for vector quantization, *Neural Networks* **3 (3)**: 277-290, 1990.

[192] Gersho, A. and Gray, R. M. *Vector Quantization and Signal Compression*, Norwell, MA: Kluwer, 1992.

[193] Kohonen, T., Hynninen, J., Kangas, J., Laaksonen, J. and Torkkola, K. Lvq pak: The learning vector quantization program package, 1995.

[194] Nasrabadi, N. M. and Feng, Y. Vector quantization of images based upon Kohonen self-organizing feature maps, *Proceedings of IEEE International Conference on Neural Networks*, pp. 101-108, 1988.

[195] Chiueh, T., Tang, T. and Chen, L. Vector quantization using treestructured self-organizing feature maps, *IEEE Journal on Selected Areas in Communications*, **12**: 1594-1599, 1994.

[196] Godfrey, K. R. L. and Y. Attikiouzel. Applying neural networks to colour image data compression. *Proceedings of IEEE Region 10 Conference, Tencon 92*, Melbourne, Australia, 1992.

[197] Pei, S. C. and Lo, Y. S., Color image compression and limited display using self-organization Kohonen map, *IEEE Transactions on Circuits and Systems for Video Technology* **8 (2)**: 191-205, 1998.

[198] Chang, C. H., Xu, P.F., Xiao, R., Srikanthan, T., New adaptive color quantization method based on self-organizing maps, *IEEE Transactions on Neural Networks* **16 (1)**: 237-249, 2005.

[199] Mitra, S., Pal, S. K., Self-organizing neural network as a fuzzy classifier, *IEEE Transactions on Systems Man Cybernetics* **24 (3)**: 385-399, 1994.

[200] Yang, C. C., Bose, N. K. Generating fuzzy membership function with self-organizing feature map, *Pattern Recognition Letters* **27 (5)**: 356–365, 2006.

[201] Fox, E. A. and Shaw, J. A. Combination of Multiple Searches, *Proceedings of the 2nd Text Retrieval Conference (TREC-2)*, NIST Special Publication 500-215, pp. 243–252, 1994.

[202] Belkin, N. J., Cool, C., Croft, W. B. and Callan, J. P. The effect of multiple query representations on information retrieval performance, *Proceedings of the 16th Annual ACMSIGIR*, pp. 339–346, 1993.

[203] Lee, J. H. Combining Multiple Evidence from Different Properties of Weighting Schemes, *Proceedings of the 18th Annual ACM-SIGIR*, pp. 180–188, 1995.

[204] Müller, H., Müller, W., Squire, D. M., Marchand-Maillet, S. and Pun. T. Performance Evaluation in Content-based Image Retrieval: Overview and Proposals, *Pattern Recognition Letters*, **22 (5)**: 593-601, 2001.

[205] Rahman, M. M, Bhattacharya, P. and Desai, B. C. A Framework for Medical Image Retrieval using Machine Learning & Statistical Similarity Matching Techniques with Relevance Feedback, *IEEE Transactions On Information Technology In Biomedicine, (Special Issue on Image Management in Healthcare Enterprises)* **11 (1)**: 59–69, 2007.

[206] Rahman, M. M., Desai, B. C. and Bhattacharya, P. Medical Image Retrieval with Probabilistic Multi-Class Support Vector Machine Classifiers and Adaptive

Similarity Fusion, *Computerized Medical Imaging and Graphics*, **32**: 95–108, 2008.

[207] Rahman, M. M., Desai, B. C. and Bhattacharya, P. A Unified Image Retrieval Framework on Local Visual and Semantic Concept-based Feature Spaces. *Computer Vision and Image Understanding*, Ref. No.: CVIU-07-12, Submitted on 04, July, 2007, under review.

[208] Rahman, M. M., Desai, B. C., and Bhattacharya, P. Cross-Modal Interaction and Integration with Relevance Feedback for Medical Image Retrieval, *13th International Multimedia Modeling Conference (MMM 2007)*, Singapore, In *Proceedings of LNCS*, Vol. 4351, pp. 440–449, 2007.

[209] Rahman, M. M., Desai, B. C. and Bhattacharya, P. Visual Keyword-based Image Retrieval using Correlation-Enhanced Latent Semantic Indexing, Similarity Matching & Query Expansion in Inverted Index. *Tenth International Database Engineering & Applications Symposium (IDEAS06)*, Delhi, India, In *Proceedings of IEEE Computer Society*, pp. 201–208, 2006.

[210] Rahman, M. M., Desai, B. C. and Bhattacharya, P. Image Retrieval-Based Decision Support System for Dermatoscopic Image, *Proceedings of the IEEE Symp. on Computer-Based Medical Systems*, June, 22-23, Salt Lake City, Utah, pp. 285–290, 2006.

[211] Rahman, M. M., Desai, B. C. and Bhattacharya, P. A Feature Level Fusion in Similarity Matching to Content-Based Image Retrieval, *Proceedings of the 9th International Conference on Information Fusion*, Florence, Italy, 10-13 July, 2006.

[212] Bhattacharya, P., Rahman, M. M. and Desai, B. C. Image Representation and Retrieval Using Support Vector Machine and Fuzzy C-means Clustering Based Semantical Spaces, *Proceedings of the 18th International Conference on Pattern Recognition (ICPR 2006)*, Hong Kong, China, Vol. 2, pp. 1162–1168, 2006.

[213] Rahman, M. M., Bhattacharya, P. and Desai, B. C. Probabilistic Similarity Measure in Image Databases with SVM-based Categorization & Relevance Feedback *International Conference on Image Analysis and Recognition (ICIAR'05)*,

Toronto, Canada, Sept. 2005. In *Proceedings of LNCS*, Vol. 3656, pp. 601–608, 2006.

[214] Rahman, M. M., Bhattacharya, P. and Desai, B. C. Similarity Searching in Image Retrieval with Statistical Distance Measure and Supervised Learning, *International Conference on Advances in Pattern Recognition (ICAPR)*, Bath, UK, September 2005. In *Proceedings of LNCS*, Vol. 3686, pp. 315-324, 2005.

[215] Rahman M. M., Wang T., and Desai B. C. Medical image retrieval and registration: towards computer assisted diagnostic approach. *IDEAS Workshop On Medical Information Systems : The Digital Hospital, IDEAS'04-DH*, Beijing, China, Sept. 2004, *Proceedings of IEEE Computer Society*, pp. 78–89, 2004.

[216] Rahman, M. M., Desai, B. C. and Bhattacharya, P. Supervised Machine Learning based Medical Image Annotation and Retrieval in ImageCLEFmed 2005, *6th Workshop on Cross Language Evaluation Forum (CLEF 2005)*, Vienna, Austria, Sept. 2005. *Proceedings in LNCS*, Vol. 4022, pp. 692-701, 2006.

[217] Rahman, M. M., Sood, V., Desai, B. C. and Bhattacharya, P. CINDI at ImageCLEF 2006: Image Retrieval & Annotation Tasks for the General Photographic and Medical Image Collections, *7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, Alicante, Spain, September 20-22, 2006, Revised Selected Papers, *Proceedings in LNCS*, Vol. 4730 ,pp.715–724. 2007.

[218] Rahman, M. M., Desai, B. C. and Bhattacharya, P. Multi-Modal Interactive Approach to ImageCLEF 2007 Photographic and Medical Retrieval Tasks by CINDI, *Working Notes of the 2007 CLEF Workshop*, Sep., 2007, Budapest, Hungary, To appear in LNCS.