Bridging Semantic Gap: Learning and Integrating Semantics for Content-Based Image Retrieval

Joo Hwee Lim, *B.Sc.(Hons I), M.Sc.*

A dissertation submitted to the
School of Computer Science and Engineering
The University of New South Wales
Sydney, NSW 2052, Australia
in fulfillment of the requirements
for the degree of Doctor of Philosophy

June 2004

Abstract

Digital cameras have entered ordinary homes and produced incredibly large number of photos. As a typical example of broad image domain, unconstrained consumer photos vary significantly. Unlike professional or domain-specific images, the objects in the photos are ill-posed, occluded, and cluttered with poor lighting, focus, and exposure. Content-based image retrieval research has yet to bridge the semantic gap between computable low-level information and high-level user interpretation.

In this thesis, we address the issue of semantic gap with a structured learning framework to allow modular extraction of visual semantics. Semantic image regions (e.g. face, building, sky etc) are learned statistically, detected directly from image without segmentation, reconciled across multiple scales, and aggregated spatially to form compact semantic index. To circumvent the ambiguity and subjectivity in a query, a new query method that allows spatial arrangement of visual semantics is proposed. A query is represented as a disjunctive normal form of visual query terms and processed using fuzzy set operators.

A drawback of supervised learning is the manual labeling of regions as training samples. In this thesis, a new learning framework to discover local semantic patterns and to generate their samples for training with minimal human intervention has been developed. The discovered patterns can be visualized and used in semantic indexing.

In addition, three new class-based indexing schemes are explored. The winner-take-all scheme supports class-based image retrieval. The class relative scheme and the local classification scheme compute inter-class memberships and local class patterns as indexes for similarity matching respectively. A Bayesian formulation is proposed to unify local and global indexes in image comparison and ranking that resulted in superior image retrieval performance over those of single indexes.

Query-by-example experiments on 2400 consumer photos with 16 semantic queries show that the proposed approaches have significantly better (18% to 55%) average precisions than a high-dimension feature fusion approach. The thesis has paved two promising research directions, namely the semantics design approach and the semantics discovery approach. They form elegant dual frameworks that exploits pattern classifiers in learning and integrating local and global image semantics.

Acknowledgments

I am grateful to my supervisor, Prof. Jesse S. Jin, for accepting me as his student when I first met him in Sydney in December 2001. I would like to thank him for his trust, guidance and support during my Ph.D. study from March 2002 to June 2004.

I sincerely thank the School of Computer Science and Engineering (UNSW) for awarding me the SCSE International External Tuition Scholarship. I am deeply impressed with the school's administrative efficiency and flexibility. I would also like to thank Dr. Albert Nymeyer for approving my request for waiver of course requirement.

I am also grateful to the management of the Institute for Infocomm Research (I^2R), Singapore, especially Dr. Changsheng Xu and Dr. MunKew Leong, for their understanding and support of my study while working. Special thanks also go to Dr. Mohan S. Kankanhalli of the National University of Singapore for supporting my request for course waiver as well as Dr. Jiankang Wu and Dr. Qi Tian for being my referees of my application to UNSW in 2001.

I also owe a debt to Jean-Luc Lebrun for lending his 2400 family photos and T. Joachims for his SVM^{light} software for the experiments in the thesis.

Finally, I would like to thank my parents, my wife, and our two daughters for their unconditional love and encouragement. They are my source of motivation.

June 2004

Joo Hwee Lim

Contents

Abstract ii

Acknowledgments iii

List of Tables ix

List of Figures xi

1 Introduction 1

1.1 Motivations 1

1.1.1 Broad Consumer Images 1

1.1.2 "Keywords" in Visual Data 6

1.1.3 Semantic Gap 8

1.1.4 Research Challenges 12

1.2 Background 13

1.3 Scope and Contributions 17

1.4 Thesis Organization 19

2 Related Work 21

2.1 From Classification to Retrieval 21

2.2 Text-Based Retrieval 23

2.3 Feature-Based Retrieval 24

2.4 Region-Based Retrieval 25

2.5 Object-Based Retrieval 28

2.6 Probabilistic Retrieval 31

2.7 Image Classification 33

2.8 Query Formulation 34
2.9 Feature Fusion 37
2.10 Automatic Annotation 40

3 Semantics Design **43**
3.1 Semantic Support Regions 43
3.2 Features 46
3.2.1 Color 47
3.2.2 Texture 48
3.2.3 Normalization 49
3.2.4 Distance and Similarity 51
3.3 Learning 52
3.3.1 Support Vector Machines 52
3.3.2 SSR Learning 55
3.3.3 Feature Fusion 57
3.3.4 Learning Evaluation on Consumer Images 60
3.4 Detection 65
3.5 Multi-Scale Reconciliation 66
3.6 Spatial Aggregation 68
3.7 Abstraction Hierarchy 74
3.8 Incremental Learning 76
3.9 Object Segmentation 78
3.10 Discussion 80

4 Semantics Discovery **83**
4.1 Overview 83
4.2 Learning of Local Class Semantics 90
4.3 Learning of Typical Semantic Partitions 92
4.4 Learning of Discovered Semantic Regions 92
4.5 Image Indexing 96
4.6 Discussion 99

5 Class-Based Image Semantics 103

5.1 Overview . . . 103
5.2 Event-Based Retrieval . . . 104
5.3 Semantic Support Classes . . . 109
5.4 Local Class Patterns . . . 113
5.5 Discussion . . . 117

6 Integrated Similarity Matching 119

6.1 Overview . . . 119
6.2 Similarity Matching . . . 120
6.2.1 Local Index . . . 120
6.2.2 Global Index . . . 121
6.3 Combining Local and Global Similarities . . . 122
6.4 Dual Frameworks . . . 125

7 Query and Retrieval 127

7.1 Overview . . . 127
7.2 Test Collection . . . 128
7.3 Query by Class/Event (QBCE) . . . 129
7.3.1 Query Processing . . . 129
7.3.2 Queries and Ground Truth . . . 131
7.3.3 Experimental Results . . . 133
7.4 Query by Spatial Icons (QBSI) . . . 136
7.4.1 Query Processing . . . 136
7.4.2 Queries and Ground Truth . . . 140
7.4.3 Experimental Results . . . 141
7.5 Query by Multiple Examples (QBME) . . . 144
7.5.1 Query Processing . . . 144
7.5.2 Queries and Ground Truth . . . 146
7.5.3 Scope of Comparison . . . 148
7.5.4 Indexing based on Fusion of Color and Texture . . . 149
7.5.5 Indexing based on SSRs . . . 151
7.5.6 Similarity Integration . . . 154

7.5.7 Quantitative Comparison 155
7.5.8 Qualitative Comparison 160

8 Conclusion **165**
8.1 Contributions 165
8.2 Related Collaborations and Extensions 172
8.2.1 Fusion with Conceptual Graph Image Representation 172
8.2.2 Mapping Mid-Level Representation to Video Events 174
8.2.3 Photo Summarization for Visual Communication 175
8.2.4 Snap2Tell: Mobile Scene-based Information Retrieval 176
8.2.5 Roadmap 177
8.3 Future Directions 178

A List of Publications **183**
A.1 Book Chapters 184
A.2 Journal Papers 184
A.3 Conference Papers 185
A.4 Patent 188

Bibliography **189**

List of Tables

3.1 Features and fused features for SSR learning and classification 59
3.2 SSR classes grouped into 8 superclasses 60
3.3 Compare features and fused features on SSR generalization 61
3.4 Compare polynomial SVM classifiers on SSR generalization 62
3.5 Compare RBF SVM classifiers on SSR generalization 62
3.6 Training statistics for 26 SSR classes 63
3.7 Training statistics for each SSR class 64
3.8 Key SSRs in the index for the image shown in Figure 3.8 69
3.9 Key SSRs in the index for the image shown in Figure 3.9 70
3.10 Key SSRs in the index for the image shown in Figure 3.10 71
3.11 Key SSRs in the index for the image shown in Figure 3.11 72
3.12 Key SSRs in the index for the image shown in Figure 3.12 73

4.1 Training statistics for image semantics discovery 93
4.2 Semantic labels for DSRs shown in Figure 4.7 95
4.3 Key DSRs in the index for the image shown in Figure 4.9 97
4.4 Key DSRs in the index for the image shown in Figure 4.10 98
4.5 Key DSRs in the index for the image shown in Figure 4.11 98

5.1 Training statistics for SSCs 112
5.2 Key SSCs in the indexes for the images shown in Figure 5.4 112
5.3 Training statistics of classes learned for LCP-bsaed indexing 114
5.4 Key LCPs in the index for the image shown in Figure 5.6 115
5.5 Key LCPs in the index for the image shown in Figure 5.7 116
5.6 Key LCPs in the index for the image shown in Figure 5.8 117

5.7 Comparison of indexing schemes based on SSR, DSR, SSC, and LCP 118

7.1 Events and ground truth (G.T.) sizes 131
7.2 Categories and ground truth (G.T.) sizes 132
7.3 Average precisions at top numbers of photos 134
7.4 Average precisions at top numbers of photos 135
7.5 Precisions at top retrieved images for QBSI experiment 142
7.6 The 16 semantic queries used in QBME experiments 147
7.7 Average precisions by global and local color histograms 150
7.8 Average precisions by global and local texture histograms 150
7.9 Average precisions by fusion of global/local color/texture similarities 151
7.10 Average precisions by different matching functions and spatial aggregation templates 152
7.11 Average precisions by different classifiers 153
7.12 Average precisions by integration of SSR and SSC similarities 154
7.13 Average precisions by integration of DSR and LCP similarities 155
7.14 Average precisions at top retrieved images (CTO, SSR, SSC, Dsgn) . 156
7.15 Average precisions at top retrieved images (CTO, DSR, LCP, Dscv) . 156
7.16 Average precisions at top retrieved images (CTO, Dsgn, Dscv) 158
7.17 Average precisions for each of the 16 queries 159

List of Figures

1.1 Some photographs from each natural scene class (column). 7
1.2 A coast image and its scrambled version 8
1.3 Semantic gap between visual data and user interpretation 10
1.4 The methodology of visual keywords 14
1.5 Visual keywords as soft cluster centers 16
1.6 Visual keywords as neural network pattern classifiers 16

3.1 A structured learning framework for indexing and query 46
3.2 The idea of SVM learning for pattern classification 53
3.3 An example of SVM learning with mapping from $\Re^2$ to $\Re^3$ 54
3.4 An example of support vector classifier to separate two classes 55
3.5 Examples of semantic support regions 61
3.6 A visual information processing architecture for image indexing . . . 65
3.7 Reconciling multi-scale SSR detection maps 67
3.8 A sample image of park to illustrate SSR-based image index 69
3.9 A sample image of street scene to illustrate SSR-based image index . 70
3.10 A sample image of indoor to illustrate SSR-based image index 71
3.11 A sample image of street scene to illustrate SSR-based image index . 72
3.12 A sample image of indoor to illustrate SSR-based image index 73
3.13 Transforming from primitive feature space to semantic feature space . 75
3.14 Supervised Incremental Clustering Architecture 77
3.15 The right image shows three dominent objects segmented from the left image: sky, building, and ground 80
3.16 The right image shows three dominent objects segmented from the left image: water, face, and ground 80

4.1 The problem of image semantics discovery 85
4.2 Automatic image annotation approaches 86
4.3 Discovering typical local patterns 89
4.4 Flow of image semantics discovery 90
4.5 Proposed consumer image taxonomy 91
4.6 Training set of 105 images 94
4.7 Most typical image blocks of the DSRs 95
4.8 A schematic digram of image indexing based on DSRs 96
4.9 A sample image of park to illustrate DSR-based image index 96
4.10 A sample image of street scene to illustrate DSR-based image index . 97
4.11 A sample image of indoor to illustrate DSR-based image index 98
4.12 A spectrum of proposed semantic learning and indexing approaches . 101

5.1 Event taxonomy for consumer photos 106
5.2 Learning event models for retrieval 107
5.3 Proposed consumer image taxonomy 110
5.4 3 image examples to illustrate SSC-based image indexes 112
5.5 A schematic digram of image indexing based on LCPs 114
5.6 A sample image of park to illustrate LCP-based image index 115
5.7 A sample image of street scene ito illustrate LCP-based image index . 116
5.8 A sample image of indoor to illustrate LCP-based image index 117

6.1 System flow of indexing and retrieval with similarity integration . . . 124
6.2 Dual cascading image indexing and matching frameworks 125

7.1 Sample consumer photos in the 2400 test collection 130
7.2 Some consumer images of bad quality 130
7.3 Sample photos of each (column) event in Table 7.1 131
7.4 A hierarchy of consumer image categories 132
7.5 Two sample photos for each category listed in Table 7.2 133
7.6 A screen shot for QBSI interface 138
7.7 QBSI queries Q01 to Q04 140
7.8 QBSI queries Q05 to Q07 140
7.9 QBSI queries Q08 to Q10 140

7.10 QBSI queries Q11 and Q13 141
7.11 QBSI queries Q14 and Q15 141
7.12 Top 18 retrieved images for QBSI query Q02 142
7.13 Top 18 retrieved images for QBSI query Q05 143
7.14 Top 18 retrieved images for QBSI query Q07 143
7.15 Sample consumer photos associated with queries Q01 to Q08 147
7.16 Sample consumer photos associated with queries Q09 to Q16 148
7.17 A spatial aggregation template that focuses at image center 152
7.18 Precision/Recall curves for CTO, SSR, SSC, and Dsgn 156
7.19 Precision/Recall curves for CTO, DSR, LCP, and Dscv 157
7.20 Precision/Recall curves for CTO, Dsgn, and Dscv 158
7.21 Average precisions of each query for RND, CTO, Dsgn, and Dscv . . 159
7.22 Query Q08 "at a swimming pool" 161
7.23 Top 18 retrieved images for query Q08 by CTO 161
7.24 Top 18 retrieved images for query Q08 by Dsgn 161
7.25 Query Q10 "along waterside" 162
7.26 Top 18 retrieved images for query Q10 by CTO 162
7.27 Top 18 retrieved images for query Q10 by Dsgn 162
7.28 Query Q14 "people close-up indoor" 163
7.29 Top 18 retrieved images for query Q14 by CTO 163
7.30 Top 18 retrieved images for query Q14 by Dsgn 164

8.1 Proposed solutions for bridging the semantic gap 172
8.2 An example of conceptual graph representation of image (img0623) . 173
8.3 Abstraction levels of indexing 173
8.4 A mid-level mapping approach to video event detection 174
8.5 Flow of video event detection via audio-visual keywords 175
8.6 A photo summarization framework for visual communication 176
8.7 A Snap2Tell framework for scene-based information retrieval 177
8.8 A summary of past and current research efforts 178
8.9 An illustration of image code and basis images for color images 181

Chapter 1

Introduction

One picture is worth a thousand words.
Fred R. Barnard

1.1 Motivations

1.1.1 Broad Consumer Images

We live in a 3D analog world. Our activities fade with time. Digital cameras allow us to preserve them as pictorial memories. We can then recall and share these pictorial memories with family and friends.

Digital cameras have entered ordinary homes and produced incredibly large number of photos. According to Photo Marketing Association (PMA) (www.pmai.org) marketing research, the sales of digital cameras grew 64% in 2003 to 50 million units worldwide. In U.S. alone, digital camera sales will rise to 15.7 million in 2004, from 12.5 million in 2003. In fact, in 2003, U.S. sales of digital cameras surpassed those of traditional cameras for the first time. At the end of 2003, 31% of U.S. households owned digital cameras and the ownership of digital cameras in U.S. homes will reach 40% by end of 2004. InfoTrends Research Group, Inc. (www.infotrends-rgi.com) even predicted that by 2008, digital cameras would replace film cameras.

The irreversible trend of digital photo-taking is getting a boost with the growth

of camera phones and multimedia messaging service (MMS) that enable a new experience of visual communication. Indeed camera phones shipments, which already outnumber digital cameras worldwide, are expected to reach 298 million in 2007, according to a forecast released by IDC (www.idc.com) in October 2003. In another market forecast, by 2008, 366 million of the 680 million (i.e. 53.8%) mobile phones sold will have cameras inside.

With digital cameras, consumers are much more proliferate in taking photos as it costs next to nothing to take a digital photo (especially with the option to delete), and many more images can be stored in flash memory than on a film [Rodden and Wood, 2003]. Hence it is not difficult at all to foresee that in the near future, average consumers would face the genuine problem of organizing and accessing tens of thousands of photos, reckoning that most consumers are reluctant to spend too much manual effort in annotation and manipulation.

User studies on the behaviour of users of image collection is limited. The most comprehensive effort in understanding what a user wants to do with an image collection is Enser's work on image [Enser, 1993] [Enser, 1995] (and also video [Armitage and Enser, 1997]) libraries for media professionals. Other user studies have focused on art images [Frost et al., 2000], medical image archive [Keister, 1994], and newspaper photo archives [Ornager, 1996] [Markkula and Sormunen, 2000]. Typically, knowledgeable users searched and casual users browsed. But all users found that both searching and browsing are useful.

The most relevant findings on how consumers manage their personal digital photos come from the user studies by Rodden [Rodden and Wood, 2003] [Rodden, 1999]. A key objective in her more recent study with Wood [Rodden and Wood, 2003], that is relevant to the research undertaken in this thesis, seeks to evaluate the usefulness of speech annotation and content-based image retrieval in the context of personal photo collections based on the ShoeBox system [Mills et al., 2000]. Besides conventional thumbnail-based browsing tool for organizing, labeling, and viewing photos, the Shoebox system also allows users to perform audio annotation which can be transcribed automatically using speech recognition for subsequent text-based retrieval.

On the visual side, the Shoebox system segments images into regions based on color and texture and indexes the regions. A user can search for photos visually

similar to a selected photo or photos that contain regions similar to one or more highlighted regions within a photo. There are 13 participants (8 males and 5 females aged 24 to 38) with an average collection of 1000 photos (ranges from 300 to 3000 pictures) in the study [Rodden and Wood, 2003].

There are many interesting findings from the user study [Rodden and Wood, 2003]. We only highlight those related to image indexing and retrieval here:

- With digital cameras, participants were more willing to take "risky" photos and "everyday" photos. Hence they tended to have taken a larger number of bad photos to obtain the good ones, which were what they would later look for to show to other people and to print;

- Organizing photos by specific events (e.g. holidays) into albums or folders and sorting photos within each album or folder in chronological order make browsing easier. But chronological ordering or classification by event does not help much in finding photos matching a more general requirement;

- Annotation only becomes important after the photos have been taken for quite some time, when many of the details have already been forgotten. Most of the participants would only want to annotate some of their photos. For those photos that are annotated, only very few annotations (either typed or spoken) were made;

- It is difficult for people to make comprehensive annotation (either typed or spoken). So even if all the photos are annotated (or spoken annotations transcribed with high accuracy), it is unlikely that all of the photos relevant to a query will be retrieved. The problem of preparing a good text description of an image has also been reported in other user studies [Armitage and Enser, 1997] [Markkula and Sormunen, 2000];

- Spoken annotations save typing effort. But some participants dislike the idea as they would feel self-conscious about speaking to a computer, and would first have to plan what to say. For those who used spoken annotation, the inaccuracy of speech recognition was unacceptably high. Names of people and

places, which are usually the most important elements of annotations, are often wrongly transcribed as they may not even in the vocabulary;

- Visual queries can be used to specify more general requirements, especially common visual properties shared by a set of photos taken at different events. However the participants expressed little interest. The authors of the study believe that queries would become important as a collection grows, and the photos get older and less familiar;

- Users had unrealistically high expectations on visual queries (for example, finding all photos of a particular person). Even those who have tried visual queries with more realistic expectations (based primarily on color) were disappointed with the results. The authors of the study felt that reliable object recognition, if available, would take some of the effort out of manual annotation.

Based on the findings from the user study [Rodden and Wood, 2003] presented above, we strongly feel that research in content-based image retrieval will play a role in the tools for managing consumer images for the following reasons:

- When the accumulation of photos becomes voluminous, search will be regarded as a useful function to complement browsing. In particular, while organizing photos based on meta-data such as time stamps [Graham et al., 2002] is useful for navigation, query and retrieval of photos across existing groupings is also important;

- Speech and text annotations are still not reliable and comprehensive even for the willing users. There are also users who are uneasy with spoken annotation and find text annotation tedious. Although speech annotation in a constrained format seemed to provide a viable alternative for content-based indexing [Chen et al., 2001] [Chen and Tan, 2003], speaking freely and timely to the digital camera with microphone in an open environment (e.g. outdoor) is still a challenge for speech recognition;

- As current image indexing based on global features and region-based features fail to meet users' high expectations, a more semantic representation based on

objects would be necessary. The object-based information extracted from the images would also be useful for automatic annotation.

In fact, consumer images satisfy the following practical criteria for content-based analysis to produce high impact as suggested by S.F. Chang [Chang, 2002],

- Generating meta-data not available from production;
- Providing meta-data that humans are not good at generating;
- Focusing on content with large volume and low individual value;
- Adopting well-defined tasks and performance metric.

That is, consumer images are created in large volumes with low individual values. Though they have personal values to individual consumer and the owner is the best person to describe the images, it is not likely that the content owner would invest enough resources such as time for manual annotation of each image. With the time stamps from digital cameras and the location information from built-in positioning devices which are meta-data available from image capturing, research can concentrate on the automatic creation of meta-data from the image content (i.e. not available from image production). As for the last item in the criteria, the query methods should be relatively simple and unambiguous. Simple and practical performance measure such as average precisions at top number of retrieved images should be adopted.

Indeed unconstrained consumer images pose great technical challenges for content-based image retrieval research. Unlike professional images, which are well defined, carefully taken and clearly layered, or domain-specific images such as medical images, which have a clear classification and are usually attached with semantic annotation, consumer images vary significantly due to the spontaneous and casual nature during image capturing. More often than not, the objects in the photos are ill-posed, occluded, and cluttered with poor lighting, focus, and exposure. We will elaborate the technical challenges related to broad domain images later in this chapter. For a feel of the visual complexity of real consumer images, readers are referred to the sample images in our test collection in Section 7.2.

1.1.2 "Keywords" in Visual Data

In the past few decades, successful text retrieval models (e.g. vector space model [Salton, 1971], probabilistic model [Robertson and Sparck Jones, 1976]) and systems (e.g. text search engines available on the World Wide Web) have been developed based on matching of keywords (or terms) between those specified in a query and those extracted from text documents in the database. Despite their conceptual simplicity, keywords are natural and yet powerful means for indexing and retrieval of text documents. Similarly, keyword-based retrieval is an intuitive and effective method to retrieve visual data if the visual data are annotated with comprehensive text labels.

However, comprehensive manual annotations are costly if not impossible. Automatic annotation based on content alone for visual data is difficult because visual data are very different in content representation from text documents. Texts are conceptual and symbolic in nature. Text keywords, which are relatively well-defined and well-segmented entities, convey meaningful semantics to human querants. Visual data are perceptual and pattern-based. Interpreting visual data is underconstrained in general. There are multiple world interpretations consistent with the visual data. Visual variations such as pose, scale, skew, translation, perspective, illumination, occlusion, clutter etc further complicate visual perception and understanding.

For instance, look at the photographs of natural scene shown in Figure 1.1. Each column of the photographs constitutes a semantic class of images perceived as similar by human users, although images in the same class could vary significantly in color, texture, and spatial configuration. The classes are (from left to right), namely, coasts, fields, trees, snowy mountains, and streams/waterfalls respectively. We are interested in the answers to the following questions:

- How would a computer perform retrieval and classification based on the visual contents of these images?
- What would be the natural sets of features for indexing and retrieval of visual data?
- Can we describe and compare visual contents beyond primitive perceptual features such as color, texture, shapes etc specific to their contents?

Figure 1.1: Some photographs from each natural scene class (column).

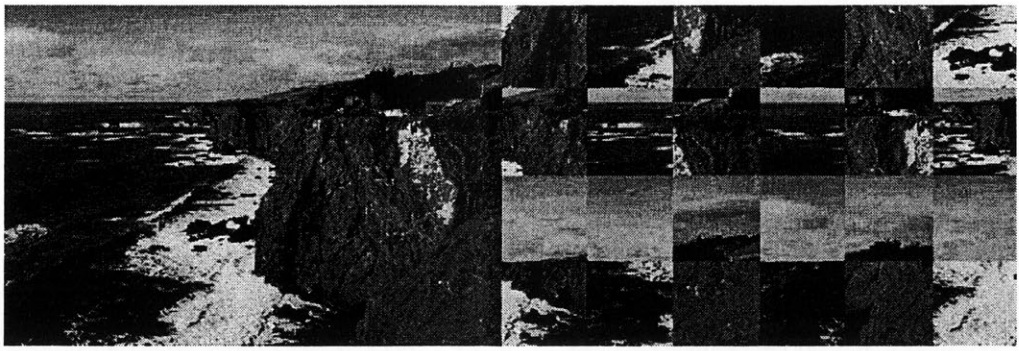

Figure 1.2: A coast image and its scrambled version

- Are there corresponding "keywords" that are inherent and consistent in a visual domain?

Considering Figure 1.2. The left half (say I_0) shows a perceptually coherent view of a coast and the right half of the same figure is its scrambled version (say I_1). Based on distributions of color or other low level features solely, I_0 and I_1 will be considered very similar (if not identical) though they are perceptually dissimilar. Scrambling I_0 in different ways can easily produce perceptually incoherent images $I_2, I_3 \cdots$ etc to fool a search engine that relies only on distribution of low level features and make its performance looks bad for comparison.

How would one describe visual content such as the coast image given in (left of) Figure 1.2? An intuitive and reasonable textual description could be: "there is cloudy blue sky at the top, dark blue sea at bottom left, brownish rocky highland (or mountain) at bottom right, and white bubbly waves along the bottom middle". The latter textual description utilizes visual features (color, texture) that characterize *types* of visual objects ('sky','sea' etc) as well as *spatial configuration* ('top','bottom right' etc). Hence our quest for "keywords" for visual data must capture these two aspects of information in the image content.

1.1.3 Semantic Gap

Users usually query images based on semantics. For example, in a recent paper by Enser, he gave a typical request to a stock photo library [Enser, 2000],

> *"Pretty girl doing something active, sporty in a summery setting, beach - not wearing lycra, exercise clothes - more relaxed in tee-shirt. Feature is about deodorant so girl should look active - not sweaty but happy, healthy, carefree - nothing too posed or set up - nice and natural looking"*

that used broad and abstract semantics to describe the images one is looking for.

Using existing image processing and computer vision techniques, low-level features such as color, texture, and shape can be easily extracted from images. However, they are inconsistent with human visual perception, let alone the incapability to capture broad and abstract semantics as illustrated by the above example. Hence low-level features cannot provide sufficient descriptive information for meaningful retrieval.

High-level semantic information is useful and effective in retrieval. However, semantic information is heavily depending on semantic image regions and beyond, which are difficult to obtain themselves. Between low-level features and high-level semantic information, there is a so called "semantic gap". Content-based image retrieval research has yet to bridge this "gap between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation" [Smeulders et al., 2000].

More precisely, based on the user studies [Rodden and Wood, 2003] [Enser, 2000], D.A. Forsyth [Forsyth, 2001] considered the following points important from the perspective of semantic gap for content-based image retrieval:

- Users request images both by object kinds (i.e. a princess) and identities (i.e. the princess of Wales);
- User request images both by what they depict (i.e. things visible in the picture) and by what they are about (i.e. concepts evoked by what is visible in the picture);
- Queries based on image histograms, texture, overall appearance, etc are vanishingly uncommon;
- text associated with images is extremely useful in practice – for example, newspaper archivists index largely on captions.

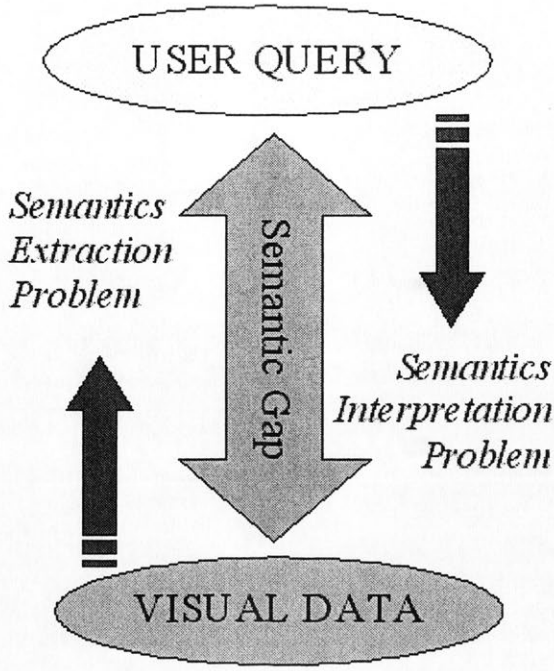

Figure 1.3: Semantic gap between visual data and user interpretation

While the goals of detection of objects and recognition of exact object identities visible in a picture are attainable, though not perfectly reliable yet, the task of characterizing concepts beyond what is visible in the picture seems only possible if there is relevant associated text. Note that in the case of consumer images used in the experiments of the thesis, text annotation and other meta-data such as time and locationis are not available. Hence looking beyond image content for semantic information is not viable.

In our opinion, the semantic gap is due to two inherent problems. One problem is that the extraction of complete semantics from image data is extremely hard as it demands general object recognition and scene understanding. This is called the *semantics extraction problem.* The other problem is the complexity, ambiguity and subjectivity in user interpretation i.e. the *semantics interpretation problem.* They are illustrated in Figure 1.3. We think that these two problems are manifestation of two one-to-many relations.

In the first one-to-many relation that makes the *semantics extraction problem* difficult, a real world object, say a face, can be presented in various appearances in an image. This could be due to the illumination condition when the image of the

face is being recorded; the parameters associated with the image capturing device (focus, zooming, angle, distance etc); the pose of the person; the facial expression; artifacts such as spectacles and hats; variations due to moustache, aging etc. Hence the same real world object may not have consistent color, texture, and shape as far as computer vision is concerned.

Indeed, highly accurate segmentation of objects is a major bottleneck except for selected narrow domains when few dominant objects are recorded against a clear background. The challenge of object segmentation is acute for polysemic images in broad domains such as general consumer images. In particular, a challenge for computer vision in broad image domain is the usually very large number of object classes in polysemic images. Moreover, the interpretation of such scenes is usually not unique as it may have numerous conspicuous objects, for which some of them have unknown object classes. Though there is promising progress in specific object recognition such as face [Zhao et al., 2000], general object recognition is still an open problem.

The other one-to-many relation is related to the *semantics interpretation problem.* Given an image, there are usually many possible interpretations due to several factors. One factor is task-related. Different regions or objects of interest might be focused upon depending on the task or need at hand. For instance, a user looking for beautiful scenic images as wallpaper for his or her desktop computer would emphasize on the aesthetic aspect of the images (besides additional requirement of very high resolution). On the other hand, a journalist working on a news story related to a celebrity would focus on images in which the celebrity appears. Thus different user needs can introduce ambiguity into a query if the requirements (scenic image, name of celebrity) associated with the needs cannot be expressed explicitly in the query. In consequence, a scenic image with the presence of the celebrity may satisfy both the requirements but the interpretations of relevance are different to the users.

Furthermore differences in culture, education background, gender etc would also inject subjectivity into user interpretation of an image, not to mention that perception and judgement are not time-invariant. For example, a Chinese user may look for red-dominant images in designing greeting cards for auspicious events but these images may not have special appeal to a European user. Indeed a recent survey

[Enser and Sandom, 2003] has suggested a framework based on the classifications of types of images and of types of user for evaluating future image retrieval systems.

1.1.4 Research Challenges

Based on the motivations described above, we face the following research issues and challenges in content-based image retrieval that we would like to investigate:

- Broad domain images have very high content variations;
- There are very large number of object classes in polysemic images;
- General object segmentation is not robust;
- General object recognition is difficult;
- Text annotation is incomprehensive and tedious;
- User interpretation is ambiguous and subjective.

Hence a systematic, modular, and adaptive framework is necessary to deal with content diversity. Modularity is required to handle a plurality of semantic entities independently. Training pattern classifiers from examples allows the system to abstract semantic entities from their instances and to adapt to new semantic requirement easily. Much like structured design and programming in software engineering, the framework should provide guiding principles to construct content-based image indexing and retrieval systems for a given content domain.

The image indexes generated should support semantic interpretation and query. In particular, one way to reduce ambiguity in a query is to allow explicit formulation of query in terms of semantic entities by the users. Since object segmentation is usually a means to extract object identity and robust object segmentation is still an open problem, dependency on object segmentation should be minimized or even removed. As object recognition for numerous object classes is in general unsolved, the framework should accommodate imperfection and uncertainty in object detection and recognition. In other words, the image index representation should incorporate soft classification decisions instead of hard decisions. In short, in contrast to the

conventional segmentation and recognition paradigm that tend to accumulate errors in the computational pathway, the framework should retain semantic certainty information as much as possible.

Another important research objective should also aim to minimize the manual effort required to label training region samples. This is an ambitious goal to further automate statistical learning for computer vision. One should also look beyond semantics in a single image and exploit recurrent (intra-class) and discriminative (inter-class) semantics in classes of images.

1.2 Background

Some of the desirable features such as semantic interpretation and query, soft detection, and segmentation-free indexing as discussed above have been explored in our previous research on *Visual Keywords* (VK) [Lim, 1999a] [Lim, 1999c] [Lim, 1999b] [Lim, 2000a] [Wu et al., 2000a] [Lim, 2000d] [Lim, 2000b] [Lim, 2000c] [Lim, 2001b] [Lim, 2001a] as part of two international research collaboration projects. The research presented in the thesis is a substantial extension of the VK framework.

The Real World Computing Partnership project (RWCP) (Phase 2, from Apr. 1997 to Dec. 2001) was funded by the Ministry of Economics, Trade and Industry (METI) of Japan. The research theme of the project was to explore novel functions for flexible organization of information bases. The author of the thesis was the acting head of the RWCP Information-Base Functions KRDL Lab in the Kent Ridge Digital Laboratories of Singapore.

The other research project, Digital Image/Video Album (DIVA), was a collaboration among CNRS (France), School of Computing of the National University of Singapore, and Kent Ridge Digital Laboratories (Singapore) (now Institute for Infocomm Research) from Jan. 2000 to Jun. 2003. The objective of the research project was to develop new image and video indexing and retrieval techniques for home users.

The conceptual framework of VK was conceived in late 1998. Figure 1.4 recapitulates the methodology of VK.

In the VK methodology, a *visual document* is defined as a complete unit of

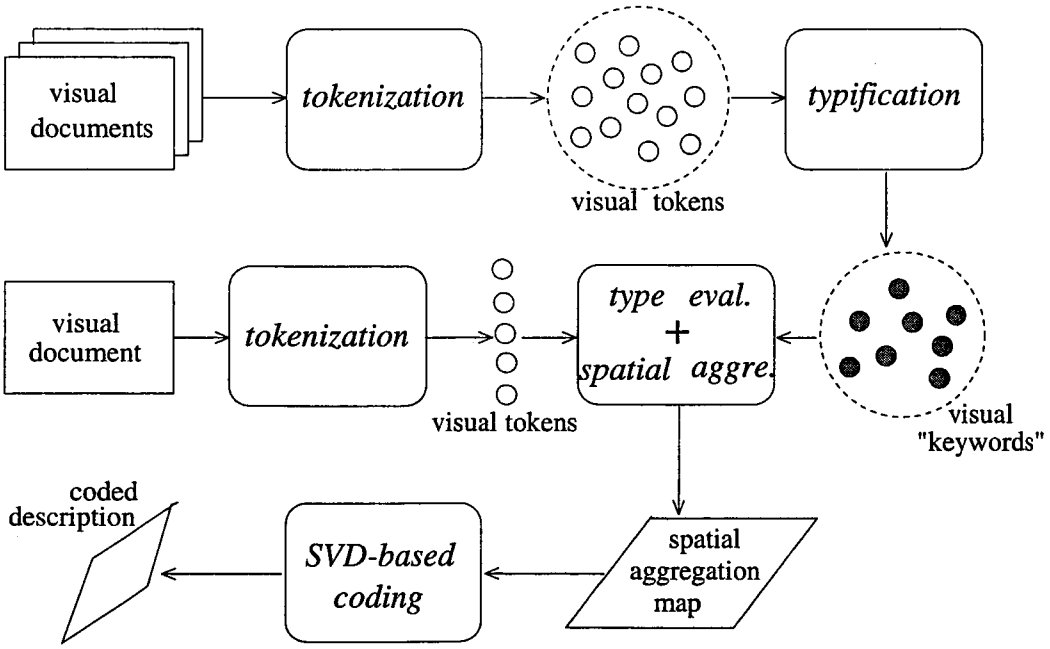

Figure 1.4: The methodology of visual keywords

visual data. Examples include a digital image, a video shot represented by some key frame(s) etc. A coherent unit in a visual document, such as a region of pixels in an image, is called a *visual token*. There are prototypical visual tokens present in a given distribution of visual documents. Using soft computing techniques, these visual keywords can be extracted and abstracted from a sufficiently large sample of visual tokens of a visual content domain.

Visual keywords could correspond to "things" like faces, pedestrians etc and "stuffs" like foliage, water etc in visual contents, represented by suitable visual characteristics. They are called "keywords" as in text documents for the following reasons. First of all, they represent unique *types* (or classes) of visual tokens occurring in a visual content domain. Next, a visual content is described by the presence or absence of these typed visual entities at a spatial abstraction, rather than directly by the visual entities or primitive features. Last but not least, the higher-order semantic structure implicit in the association of these typed visual entities with the visual documents are exploited to develop a coding scheme.

Figure 1.4 summarizes the methodology in a flow diagram. The top row depicts the extraction of visual keywords. A systematic and automatic component called

tokenization extracts visual tokens from visual documents. A *typification* component creates visual keywords from the set of visual tokens. The visual keywords are visual representation resulting from supervised or/and unsupervised learning.

The middle row of Figure 1.4 shows the steps to produce visual content signature based on extracted visual keywords. During indexing (or retrieval), a visual document (or a query sample), is subjected to tokenization to produce visual tokens. The location-specific visual tokens are evaluated against the visual keywords and their soft occurrences aggregated spatially (*type evaluation + spatial aggregation*) to form a *Spatial Aggregation Map* (SAM) as visual content signature for the visual document. With appropriate similarity measure, the SAMs of visual documents can be used in similarity matching for image retrieval and categorization applications.

Last but not least, the bottom row illustrates a coding process based on singular value decomposition to reduce the dimensionality and noise in SAMs. This is similar to the *Latent Semantic Analysis* (LSA) [Deerwester et al., 1990] technique in text retrieval that exploits higher-order semantic structure implicit in the association of terms with documents. Using singular value decomposition with truncation and cell transformation as given in [Landauer et al., 1998], LSA captures most of the essential underlying structure in the association of terms and documents, yet at the same time removes the noise or variability in word usage that plagues word-based retrieval methods. The derived coded description achieves a reduction in dimensionality while preserving structural similarity in term-document association for good discriminating power in similarity matching.

The typification component in Figure 1.4 aims to induce the types (or classes) of visual tokens from sufficiently large number of examples in a visual content domain. Both supervised and unsupervised learning methods can be employed. Thus, while visual keywords are visual content domain-dependent, the framework allows them to be customized for the domain via learning.

Unsupervised learning methods such as Self-Organizing Maps (SOM) neural networks [Kohonen, 1997], Fuzzy C-Means (FCM) algorithm [Bezdek, 1981], and the Expectation-Maximization (EM) algorithm [Mitchell, 1997] can be used to discover regularities in the visual tokens in visual documents. Soft clusters (visual keywords) that represent prototypical visual tokens are formed from a training set of visual

tokens sampled from visual documents of a given visual content domain (Figure 1.5).

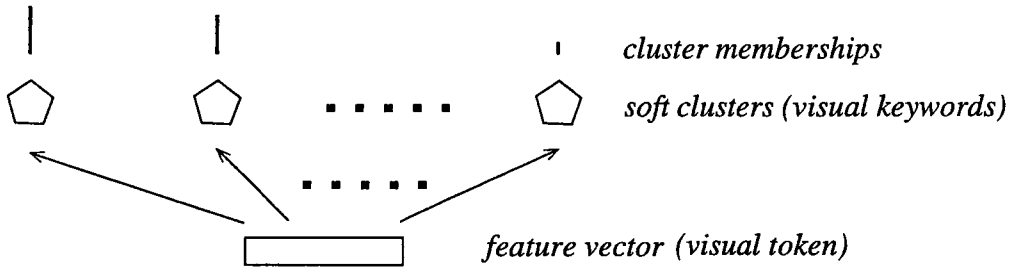

Figure 1.5: Visual keywords as soft cluster centers

For supervised learning, view-based detectors such as neural network recognizers for salient objects such as human faces, pedestrians, foliage, clouds etc can be induced from a training set of positive and negative examples of visual tokens collected from visual documents of a given content domain (e.g. [Papageorgiou et al., 1998]). Suppose the domain is natural scene images and we employ neural networks as object detectors. Then we need to design neural network object detectors for foliage, skies, sea waves, snowy mountains etc and train them using positive and negative examples of these objects represented in suitable feature vectors (e.g. color, texture). Detectors may be further specialized for different views (e.g. different types of foliage, skies of cloudy and clear days etc) to improve the accuracies of the view-based neural network object detectors. In this supervised paradigm, a visual keyword is a neural network trained on a class of visual objects (Figure 1.6).

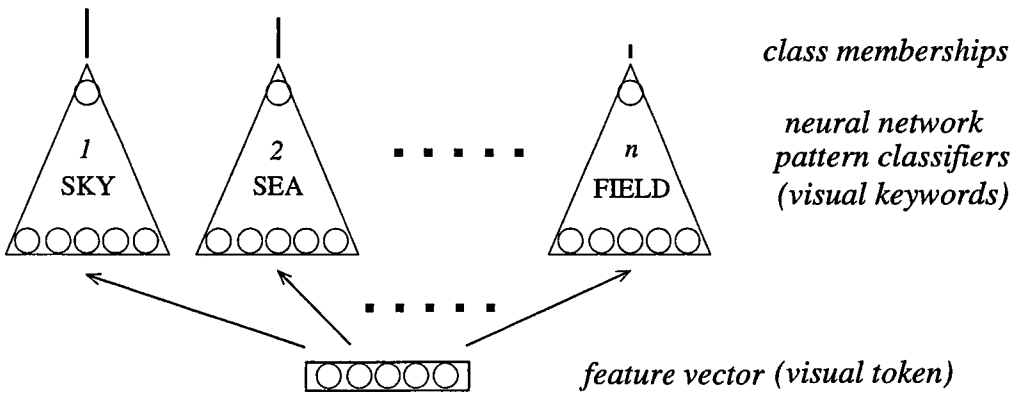

Figure 1.6: Visual keywords as neural network pattern classifiers

While the unsupervised learning approach may produce visual keywords without

clear semantics, the supervised learning approach generally requires many examples for training neural network classifiers. As another alternative approach, the visual keywords are explicitly taught to the system by a user. That is, visual keywords are visual prototypes manually specified by a user (i.e. handcrafted). Using an appropriate visual tool, the user crops domain-relevant regions from sample images and assigns sub-labels and labels to form vocabulary and thesaurus respectively.

Before embarking on the research documented in this thesis, preliminary explorations were pursued along the unsupervised approach [Lim, 1999a] [Lim, 1999c] [Lim, 1999b] [Lim, 2000d] [Lim, 2000c] and the handcrafted approach [Lim, 2000a] [Wu et al., 2000a] [Lim, 2000b][Lim, 2001b] [Lim, 2001a] only. The process of creating image signature based on the VK methodology has been patented [Lim, 2003].

1.3 Scope and Contributions

There are many important and interesting problems related to content-based image retrieval. This thesis focuses on the semantic gap problem. We study visual semantics that can be directly extracted from image content (without using associated text) with computer vision and pattern recognition techniques. As a typical example of broad domains, unconstrained consumer images are used as the test collection to address the practical need due to the explosive growth in personal digital images.

In this thesis, we address the issue of high content diversity with a structured learning framework to allow modular design and extraction of visual semantics called *Semantic Support Regions* (SSRs) [Lim and Jin, 2002b] [Lim and Jin, 2002a] [Lim and Jin, 2004e] [Lim and Jin, 2004h] [Lim and Jin, 2004d]. They are semantic image regions learned statistically, detected directly from image content without segmentation, reconciled across multiple resolutions, and aggregated spatially to form compact semantic index. They can be used to bridge the semantic gap.

To circumvent the complexity, ambiguity and subjectivity in user interpretation during query, a new query method called *Query by Spatial Icons* (QBSI) that allows spatial arrangement of visual semantics (e.g. face, sky, building etc) is proposed in the thesis [Lim, 2000a] [Lim, 2001a] [Lim and Jin, 2004e] [Lim and Jin, 2004h]. Unlike existing query methods that expect the retrieval system to guess a user's

intention expressed implicitly in the query, QBSI lets user specify a query explicitly using higher level of visual semantics. A QBSI query is expressed as a disjunctive normal form of visual query terms and processed based on fuzzy set operators. As spatial information is retained in the image index based on SSRs, QBSI can be applied naturally and efficiently.

However, a drawback of the supervised learning approach is the human effort to provide labeled regions as training samples. In this thesis, a new hybrid learning framework to discover local semantic patterns and generate their samples for training with minimal human intervention has been developed. Different from existing approaches in unsupervised semantics learning, we do not make use of associated text description nor define the object classes to be recognized. The *Discovered Semantic Regions* (DSRs) can be visualized and used as SSRs to form local semantic histograms for image indexing and retrieval [Lim and Jin, 2004g] [Lim and Jin, 2004c] [Lim and Jin, 2004d].

On the use of global class information, the thesis has explored three new indexing schemes. The winner-take-all scheme supports retrieval by events or classes [Lim and Jin, 2002c] [Lim et al., 2003b] [Lim and Jin, 2003a] [Lim and Jin, 2003b] [Lim et al., 2003c]. The class relative scheme computes inter-class memberships from SSR-based index for similarity-based retrieval [Lim and Jin, 2004d]. The local classification scheme embeds local class patterns as index for similarity matching.

In this thesis, we propose a Bayesian formulation to unify both local and global semantic indexes in similarity matching. The SSR-based indexes and the indexes from class relative scheme are combined [Lim and Jin, 2004a] [Lim and Jin, 2004i] [Lim and Jin, 2004b]. On the other hand, the indexes based on the local classification scheme and DSRs are complementary [Lim and Jin, 2004c]. Both combined similarity measures have resulted in superior performance over those of individual index in the image retrieval experiments.

All the proposed indexing schemes in the thesis are evaluated against a typical feature fusion approach that combines color and texture in a lineary optimal way. The query-by-example experiments on 2400 genuine consumer images with 16 semantic queries show that the proposed compact indexes have significantly better average precisions and precisions at top retrieved images over the feature-based ap-

proach that requires very high dimension index to attain reasonble performance for the challenging dataset. The improvement in overall average precision ranges from 18.4% to 55.3%. The SSR and winner-take-all indexes are also evaluated on QBSI and class-based retrieval experiments with promising results respectively.

The thesis has paved two promising directions of research, namely the semantics design approach and the semantics discovery approach, for content-based image retrieval. Indeed, they form elegant dual frameworks that exploit pattern classifiers in local and global image semantic learning and matching [Lim and Jin, 2003c] [Lim and Jin, 2004f]. While the semantics design approach goes from local SSR index to class relative index, the semantics discovery approach starts with local classification to bootstrap DSRs.

1.4 Thesis Organization

This thesis has 8 chapters in total. The motivations, background, scope and contributions of the research described in this thesis have been presented in this first chapter.

In the second chapter, the key developments in content-based image indexing and retrieval and relevant computer vision techniques are reviewed.

The research contributions of the thesis are described in the next five chapters.

Chapter 3 describes a semantics design approach to image indexing based on structured learning and extraction of local semantic regions without segmentation. Next, in Chapter 4, a semantics discovery approach is proposed to alleviate the region labeling problem of supervised learning. The experimental results of semantic region learning and discovery are included in both chapters respectively.

While the previous chapters focus on local semantics, Chapter 5 discusses three different class-based indexing schemes. A Bayesian formulation is then proposed in Chapter 6 to unify both local and global semantic indexes for image matching and ranking. The dual frameworks of learning and integration for image indexing and matching are also discussed in Chapter 6.

Chapter 7 is devoted to image query and retrieval based on the indexing schemes proposed in Chapters 3 to 5. The test collection of 2400 genuine family photos is

described and three query methods are presented. For each query method, the query processing, queries and associated ground truths, and experimental results are described. In particular, a comprehensive comparison, both quantitative and qualitative, against a feature fusion approach is given for the query by examples experiments in Section 7.5.

In the last chapter, the thesis is concluded with a list of contributions. The proposed frameworks in this thesis have been extended in a few directions with other collaborators. They are briefly summarized in this chapter. Last but not least, direction for future work is discussed.

In Appendix A, the list of publications related to the research presented in this thesis is given.

Chapter 2

Related Work

To find a fault is easy; to do better may be difficult.
Plutarch (46 AD - 120 AD)

2.1 From Classification to Retrieval

As a spin-off from the fields of pattern recognition and computer vision more than a decade ago [Smeulders et al., 2000], content-based image retrieval research focuses on a different problem from pattern classification though they are closely related. In pattern classification, according to the Bayes decision theory, we should select class C_i with the maximum a posteriori probability $P(C_i|x)$ for a given pattern x in order to minimize the average probability of classification error ([Duda and Hart, 1973], pp. 17). When the construction of pattern classifiers relies on statistical learning from observed data, the models for the pattern classifiers could be parametric or non-parametric.

When the patterns concerned are images, pattern classification could become an image classification problem (e.g. [Vailaya et al., 2001]) or an object recognition problem (e.g. [Papageorgiou et al., 1998]). While the former deals with the entire image as a pattern, the latter attempts to extract useful local semantics, in the form of objects, in the image to enhance image understanding. Needless to say, the

success of accurate object recognition would result in better scene understanding and hence more effective image classification.

In content-based image retrieval, the objective of a user is to find images *relevant* to his or her information need, expressed in some form of query input to an image retrieval system. Given an image retrieval system with a database of N images (assuming N is large and stable for a query session), the hidden information need of a user cast over the N images can be modeled as the posterior probability of the class of relevant images R given an expression of the information need in the form of query specification q and an image x in the current database, $P(R|q,x)$. This formulation follows the formalism of probabilistic text information retrieval [Robertson and Sparck Jones, 1976]. Here we assume that the image retrieval system can compute $P(R|q,x)$ for each x in the database. The objective of the system is to rank and return the images in descending order of probability of relevance to the user.

Certainly, the image classification and object recognition problems are related to the image retrieval problem as their solutions would provide better image semantics to an image retrieval system to boost its performance. However, the image retrieval problem is inherently user-centric or query-centric (i.e. $P(R|q,x)$ versus $P(C_i|x)$)). There is no predefined class and the number of object classes to be recognized to support queries is huge [Smeulders et al., 2000] in unconstrained or broad domains.

Content-based image retrieval research has progressed from the pioneering feature-based approach (e.g. [Bach et al., 1996] [Flickner et al., 1995] [Pentland et al., 1995]) to the region-based approach (e.g. [Smith and Chang, 1996] [Carson et al., 1997] [Li et al., 2000]). However, a desired feature and hence a key research challenge is to extract semantics to support meaningful queries.

In this chapter, we will review several key developments in content-based image retrieval (text-based, feature-based, region-based, object-based, probabilistic). For a comprehensive coverage and understanding of these approaches, readers are referred to the survey paper [Smeulders et al., 2000] and individual papers mentioned in our review. We will also review image classification and query formulation methods. Feature fusion is an important issue when multiple cues such as color and texture have to be combined. We will discuss feature fusion in a broader context of fusion of

multiple modalities for image and video indexing. The new trends in object recognition and text-image association are included in this chapter too as they provide promising means for automatic annotation.

For a review on semantic video indexing, we refer readers to a recent survey [Snoek and Worring, 2002] and some of the representative developments in different domains (e.g. Films [Sundaram and Chang, 2000] [Vendrig and Worring, 2003], Soccer [Xie et al., 2004] [Kang et al., 2004a], Medical [Ebadollahi et al., 2002], News [Hsu and Chang, 2004] [Amir et al., 2003], Documentary [Haering et al., 2000], and General [Naphade et al., 2002]).

2.2 Text-Based Retrieval

Text retrieval based on keywords has been the main stream in the field of information retrieval [Sparck Jones and Willett, 1997]. Many existing visual retrieval systems (e.g. [Rowe and Eads, 1994]) extract and annotate the data objects in the visual content manually, often with some assistance of user interfaces. It is assumed that once keywords are associated with the visual content, text retrieval techniques can be deployed easily, though articulating a comprehensive set of keywords for an image is not an easy task [Armitage and Enser, 1997] [Markkula and Sormunen, 2000] [Rodden and Wood, 2003].

For certain image collections such as personal photos (c.f. Section 1.1.1), very few people are willing to spend time in annotation and when they do, only very few annotations are given [Rodden and Wood, 2003]. Furthermore, comprehensive annotation becomes more difficult after a photo has been taken for quite some time as the memory of many of the details has faded [Rodden and Wood, 2003]. Annotation at image capturing time is most effective when the context is available but natural input interface such as voice recording is preferred. Unfortunately, except for recording in a controlled environment in a constrained format [Chen et al., 2001] [Chen and Tan, 2003], recovering the keywords from voice annotation still pose great challenge for existing speech recognition technology.

On the other hand, although text descriptions are certainly important to reflect the (largely conceptual) semantics of multimedia data, they may result in excessive

amount of keywords in the attempt of annotation due to the ambiguous and variational nature of multimedia data. Inherently there is a limit to how much semantic information the textual attributes can provide to convey the meanings of a piece of multimedia data [Bolle et al., 1998].

Moreover, as user interpretation of multimedia data is often ambiguous and subjective (c.f. the semantics interpretation problem discussed in Section 1.1.3), annotations of the same multimedia data can vary with different information needs (tasks), different users (gender, age, education background, experience, culture, etc), and at different times.

In short, manual annotation, as a means of pre-query indexing, is usually incomplete, inconsistent, and context sensitive. The process is tedious and yet not always effective. While Query by Keywords (QBK) does allow information need to be described in high-level meaningful terms, the semantic gap between articulated expectation and image indexes is large unless the image indexes cover comprehensive labels. This is not achievable with the current automatic content-based image indexing systems.

As a significant step towards bridging this semantic gap, this thesis proposes a structured and modular learning framework to capture semantic labels with location information from the image which supports query by spatial arrangement of semantic icons.

2.3 Feature-Based Retrieval

In the early days, primitive visual features such as color, texture, and shapes are used to index and retrieve images (e.g. [Flickner et al., 1995] [Pentland et al., 1995] [Bach et al., 1996]). These pioneering systems have mainly relied on aggregate measures (e.g. histograms) of primitive features for describing and comparing visual contents. However, this approach often produces results that are incongruent with human expectations [Lipson et al., 1997] because it does not consider spatial localities and higher-level perceptive cues.

For example, images sharing similar overall color distribution can differ greatly in semantic content (c.f. Section 1.1.2). This paradigm roughly corresponds to

pre-attentive similarity matching which is a low-level function in human visual perception. Nevertheless, new low-level features such as banded color correlograms [Huang et al., 1998], joint histograms [Pass and Zabih, 1999] etc are still being proposed to improve the approach on aggregate measures of low-level features.

With technological advances in digital cameras, we can now easily recover the time stamps of image creation. Industrial players have been looking into the standardization of the file format (e.g. Exchangeable Image File Format, version 2.2, [JEITA, 2002]) that contains this information. Similarly with the advances in global positioning systems (GPS) technology, the location at which a photo is taken can also be automatically obtained from the camera (e.g. the Kodak Digital Science 420 GPS camera). Hence time and location information can serve as additional indexing axes for consumer images.

In particular, time information is considered very useful for clustering photos into events and for ordering photos within an event to facilitate browsing of personal image collection [Rodden and Wood, 2003]. For example, under the Stanford's Personal Digital Library project [Graham et al., 2002], photo creation time has been heavily exploited to allow efficient browsing of personal photos over simple file folder mechanism. Algorithms have been proposed to determine the number and selection of photos to be presented in the browser interface. There are also interesting efforts that combine both time-based and content-based (feature-based) analysis to enhance the organization of photos for browsing [Cooper et al., 2003] [Platt, 2000] [Platt et al., 2003] and summarization [Li et al., 2003b] [Lim et al., 2003a].

In this thesis, the index generated for an image can be viewed as a set of *local* histograms, though the bins of the histograms correspond to semantic labels instead of low-level features (i.e. *semantic bins*). Hence our framework improves upon the feature-based approach by capturing both the locality and semantics in an image index.

2.4 Region-Based Retrieval

In contrast to the feature-based approach that tends to focus on global measures of low-level visual features, region-based methods (e.g. [Smith and Chang, 1996]

[Carson et al., 1997] [Ma and Manjunath, 1997b] [Li et al., 2000]) pre-segment an image by color (or both color and texture) into cohesive regions of pixels and compute the similarity between two images in terms of the features (and spatial relationships [Smith and Chang, 1996]) of these segmented regions. But image segmentation is generally unreliable. A poor segmentation can result in incongruent regions for further similarity matching.

The VisualSEEk system [Smith and Chang, 1996] [Smith and Chang, 1999] was first to consider the spatial relationships among segmented regions extensively and to combine them with primitive features of regions for image retrieval. The matching algorithm merges lists of image candidates, resulting from region-based matching between query and database images, with respect to some thresholds and hence tends to be rather complex in realization. Segmentation of regions is based on color only and no object or type information is obtained from the segmented regions.

The descendent of VisualSEEk, WebSEEk [Smith and Chang, 1997], is an image and video catalog and search tool for the World Wide Web. It collects the images and videos using a few autonomous Web agents. Among the new features, WebSEEk utilizes both text and visual information synergistically and supports query modification with relevance feedback.

A different research project, MetaSEEk, deals with issues involved with efficiently querying large and distributed online image repositories [Chang et al., 1997b] as well as exploiting user feedback in previous searches for recommending target search engines and integrating the results from different search engines in future queries [Benitez et al., 1998].

Going beyond global primitive features, a new image representation called *blobs*, which are coherent clusters segmented in combined color and texture space based on the Expectation-Maximization algorithm, has been developed [Carson et al., 1997] [Carson et al., 2002]. In particular, a new approach to texture description and scale selection was introduced. By finding image regions which roughly correspond to objects, the authors hope that image querying can be done at the level of objects. In addition, a unique feature of the Blobworld system allows the user to view the internal representation of the submitted image and the query results. Similarity matching is based on the features of the segmented regions.

For image classification, all the blobs from the categories in the training data are clustered into "canonical" blobs using Gaussian models with diagonal variance. A decision-tree classifier is trained on the distance vectors that measure the nearest distance of each canonical blob to the images.

The NETRA project [Ma and Manjunath, 1997b] [Ma and Manjunath, 1999] also uses color, texture, shape and spatial location information in segmented image regions to retrieve similar regions from the database. Robust image segmentation algorithm is the key research effort of the NETRA project. While the initial system has focused on texture features [Ma and Manjunath, 1997a] [Manjunath and Ma, 1996], the new version of NETRA emphasizes on color image segmentation and local color feature [Deng and Manjunath, 1999].

Moving away from pixelwise segmentation to blockwise segmentation based on wavelet-based feature extraction and simple k-means clustering algorithm to reduce computational cost, the SIMPLIcity system [Li et al., 2000] [Wang et al., 2001] assumes that blocky boundary has little effect on retrieval when the block size is small (4×4 as adopted in their system). In addition, the authors argued that inaccurate image segmentation can be tolerated with the proposed integrated region matching (IRM) scheme that measures the overall similarity between images by integrating properties of all the regions in the image. Last but not least, the authors proposed to pre-classify images into semantic categories based on segmented regions, such as textured-nontextured, objectionable-benign, or graph-photograph, so as to reduce the search space.

More recently, the IRM scheme is extended to fuzzy feature matching to incorporate segmentation-related uncertainties more naturally to further reduce the effect of poor image segmentation [Chen and Wang, 2002]. A new graph-theoretic clustering algorithm has also been designed to retrieve image clusters by dynamically partitioning a collection of images in the vicinity of the query to enhance user interaction [Chen et al., 2004].

Since robust image segmentation is still a very hard problem, almost as difficult as automatic image semantic understanding [Wang et al., 2001], there are also attempts to bypass the segmentation step. In particular, motivated as an analogy of "keywords" of an image, the theory of Keyblocks [Zhu et al., 2002] and Visual

Keywords [Lim, 2000d] [Lim, 2001a] also build image index from image regions. However, the regions are extracted from multi-resolution image blocks without segmentation. The generation of Keyblocks or Visual Keywords are based on either clustering [Zhu et al., 2002] [Lim, 1999b] [Lim, 1999c] [Lim, 2000d] or manual selection [Zhu et al., 2002] [Lim, 2001a]. While in general the semantics obtained from unsupervised learning is not strong, the manual selection approach requires intensive human expert labor. Although automatic selection was proposed as an alternative for Keyblock generation [Zhu et al., 2002], the codebook-based process is primarily cluster-based and hence may not be discriminative enough for semantic detection.

The research in this thesis extends the segmentation-free Visual Keywords approach substantially with structured learning and discovery, multi-scale reconciliation, similarity integration etc with extensive experimentation.

2.5 Object-Based Retrieval

The Visual Apprentice (VA) system [Jaimes and Chang, 2001] [Jaimes, 2003] is a dynamic and flexible system for learning visual object detectors using examples from images or video provided from a user. Compared to the interactive FourEyes system [Minka and Picard, 1997] that also learns from labels assigned by a user using multiple feature models, the VA system allows the user to define a much more comprehensive multiple-level object definition hierarchy and automates the tasks of feature and classifier selection using k-fold cross-validation over the training set. To accommodate subjectivity in user perception, a user defines visual scene and object detectors in a hierarchical model according to his interests.

The VA system performs automatic image region segmentation while the user manually labels and maps segmented regions to various nodes in the object definition hierarchy. Optimal features and classifiers are then learned for each node in the hierarchy. Given a new test image or video, regions are segmented and propagated bottom up the hierarchy to arrive at the final scene-level decision by fusing classification decisions at the nodes of various levels.

In the similar spirit, the Semantic Visual Template (SVT) approach associates each semantics with a set of exemplar queries [Chang et al., 1998a]. That is, instead

of labeling ground truth data in the database, the SVT approach relies on a two-way interaction between the user and the system (returned results and relevant feedback) to converge on small set of queries that provide maximal recall for the user's concept. With direct access and manipulation to any SVT in the library, new and complex SVT can be composed graphically from the combination of existing templates.

While these interactive systems [Minka and Picard, 1997] [Chang et al., 1998a] [Jaimes, 2003] believe that end users should design the semantics and provides the training samples dynamically to reflect their subjective preferences, the issues of competence in design and manual effort in labeling for average users is not addressed, let alone the problem of sustainable discrimination and scalability in a dynamic classifier learning environment. A new concept visually similar to an existing concept learned may make the existing features and classifiers inadequate. In fact, these systems have only been demonstrated on learning and classification of limited number of concepts.

Town and Sinclair [Town and Sinclair, 2000] adopted an off-line semantic design and labeling approach. An image is segmented into non-overlapping regions grown from seed points generated from the peaks in the distance transform of the edge image. Each region is classified into one of the 11 predefined visual categories of outdoor scenes by neural networks. The best classification results were achieved by multi-layer perceptrons neural networks with 3 hidden layers of up to 2000 neurons. Similarity between a query and an image is computed as either the sum over all grids of the Euclidean distance between classification vectors, or their cosine of correlation. Retrieval evaluation was carried out on over 1000 Corel Photo Library images and about 500 home photos, with better classification and retrieval results obtained for the professional Corel images.

A probabilistic generative approach to segment and label image regions was given in [Kumar et al., 2002]. While generative models offer modular framework for learning the semantic classes, it may not work well when the classes have close multimodal distributions and the data near the discriminative boundary will not be emphasized. The method was only tested on 130 real images with 5 semantic labels (sky, water, skin, sand/soil, and grass/tree). Based on a much bigger test collection of news video, the experiments reported on visual semantic concept re-

trieval [Adams et al., 2003] shown that the Gaussian Mixtures [Bishop, 1995] classifiers have lower test set accurarcy than the SVM classifers. Similar empirical evidence has also been reported for the task of news story segmentation on large news video corpus whereby the generative approach based on maximum entropy is found to be less effective than the discriminative approach based on support vector machines [Hsu and Chang, 2004].

In a leading effort by the IBM research group to design and detect 34 visual concepts (both objects and sites) in the TREC 2002 benchmark corpus (www-nlpir.nist.gov/projects/trecvid/), support vector machines are trained on segmented regions in key frames using various color and texture features [Naphade et al., 2003] [Naphade and Smith, 2003]. Recently the vocabulary has been extended to include 64 visual concepts for the TREC 2003 news video corpus [Amir et al., 2003]. Several months of effort were devoted to the manual labeling of the training samples using their VideoAnnEx annotation tool [Lin et al., 2003] contributed by the TREC participants. We would return to their work when we discuss the issue of feature fusion later in this chapter (Section 2.9) and in Chapter 3.

In the domain of consumer images, people identification such as face recognition in still images will be useful in image indexing and query since people are one of the key subjects in these images. We reckon that general face recognition [Zhao et al., 2000] in still images is a hard problem when it has to deal with small faces (20×20 pixels or less), varying poses and lighting conditions, facial expressions, occlusions etc. In fact, our preliminary face recognition experiment [Li et al., 2003b] for 9 family members in 2400 photos using a state-of-the-art public domain face detector [Rowley et al., 1998] and face recognizer [Nefian and Hayes III, 1999] produced results that are far from satisfaction.

In this thesis, we advocate the semantic design approach to learn and detect segmentation-free regions in images. To reduce the manual annotation effort, a semantics discovery approach is also proposed. In our approaches, objects and scenes are handled separately and image similarities based on their detection are integrated in a principled way to improve retrieval performance.

2.6 Probabilistic Retrieval

The CANDID project [Kelly et al., 1996] is one of the early works that employed probability density functions (PDFs) of local features for representation and matching of image contents in image retrieval. As a tradeoff between accurate representation and manipulation efficiency, typically Gaussian mixture was adopted to represent each PDF and L_2 distance measure or a normalized inner-product were used to compare two PDFs [Kelly et al., 1996]. More recently, as an enhancement of the Blobworld approch [Carson et al., 1997] [Carson et al., 2002], the Kullback-Leibler (KL) distance (or relative entropy) [Kapur and Kesava, 1992] was proposed as a distance measure for comparing the Gaussian mixture distributions that represent the segmented homogeneous regions [Greenspan et al., 2001]. Furthermore, the KL distance was also extended for matching image categories.

In [Moghaddam et al., 1998], intra-personal and extra-personal classes of variation between two facial images were modeled. Then, the similarity between the image intensity of two facial images was expressed as a probabilistic measure in terms of the intra-personal and extra-personal class likelihoods and priors using a Bayesian formulation.

At the retrieval level, a natural and useful insight is to formulate image retrieval as a classification problem i.e. class-based retrieval. In very general terms, the goal of image retrieval is to return images of a class C that the user has in mind based on a set of features x computed for each image in the database. In probabilistic sense, the system should return images ranked in the descending return status value of $P(C|x)$, whatever C may be defined as desirable. Under this general formulation, several approaches have emerged.

In [Vasconcelos and Lippman, 2000], a Bayesian formulation to minimize the probability of retrieval error (i.e. the probability of wrong classification) had been proposed to drive the selection of color and texture features and to unify similarity measures with the maximum likelihood criteria. Similarly, in an attempt to classify indoor/outdoor and natural/man-made images, a Bayesian approach was used to combine class likelihoods resulted from multi-resolution probabilistic class labels [Bradshaw, 2000]. The class likelihoods were estimated based on local average color information and complex wavelet transform cofficients.

In a different way, [Aksoy and Haralick, 2002] and [Wu et al., 2000b] considered a two-class problem with only the relevance class and the irrelevance class. A two-level classification framework was proposed in [Aksoy and Haralick, 2002]. Image feature vectors were first mapped to two-dimensional class-conditional probabilities based on simple parametric models. Linear classifiers were then trained on these probabilities and their classification outputs were combined to rank images for retrieval.

From a different motivation, the image retrieval problem was cast as a transductive learning problem in [Wu et al., 2000b] to include an unlabeled data set for training the image classifier. In particular, a new discriminant-EM algorithm was proposed to generalize the mapping function learned from the labeled training data to a specific unlabeled data set. The algorithm was evaluated on a small database (134 images) of 7 classes using 12 labeled images in the form of relevance feedback.

Naphade [Naphade and Huang, 2001] [Naphade et al., 2002] proposed a probabilistic framework for mapping audio-visual features to high-level semantics in terms of concepts and context. Semantic concepts consisting of objects, sites, and events are represented as probabilistic multimedia objects called multijects using audio and visual features. Contextual constraints are modeled as inter-relationships among the multiject nodes using probabilistic graphical methods in an explicit network form, known as multinet, to enhance the detection of multijects.

Compared with the multinet framework [Naphade et al., 2002], the semantic concept and context modeling approach in this thesis is simpler as both the local semantics and their implicit co-occurrence context are trained separately and their complementary indexes integrated at similarity matching, hence simplifying the learning problem. In addition, segmented objects and sites (e.g. outdoor scene) are treated as equal entities as multijects [Naphade et al., 2002]. In our case, segmentation-free image regions and image classes are represented at different levels of semantics as content and context respectively.

2.7 Image Classification

Categorization is a powerful divide-and-conquer metaphor to organize and access information such as text [Larkey and Croft, 1996] [Lewis and Ringuette, 1994] and images. Once the images are sorted into semantic classes, searching and browsing can be carried out in more effective and efficient way by focusing only at relevant classes and subclasses. Moreover the classes provide context for other tasks. For example, for medical images, the context could be the pathological classes for diagnostic purpose [Brodley et al., 1999] or imaging modalities for visualization purpose [Mojsilovic and Gomes, 2002].

Image classification is considered as another approach to bridge the semantic gap as class labels convey higher semantic meanings. Hence it has received more attention lately [Bradshaw, 2000] [Lipson et al., 1997] [Szummer and Picard, 1998] [Vailaya et al., 2001].

On the approach that advocates the use of configuration, the work reported in [Lipson et al., 1997] hand-crafted relational model templates that encode the common global scene configuration structure for each category, based on qualitative measurements of color, luminance and spatial properties of examples from the categories. Classification is performed by deformable template matching which involves heavy computation. The manual construction of relational model templates is time consuming and incomprehensive. To avoid this problem, a learning scheme that automatically computes scene templates from a few examples [Ratan and Grimson, 1997] is proposed and tested on a smaller scene classification problem with promising results.

The attempts to classify photos based on contents have been devoted to: indoor versus outdoor [Bradshaw, 2000] [Szummer and Picard, 1998], natural versus man-made [Bradshaw, 2000] [Vailaya et al., 2001], and categories of natural scenes [Lipson et al., 1997] [Vailaya et al., 2001]. In general, the classifications were made based on low-level features such as color, edge directions etc. The work by Vailaya et al. [Vailaya et al., 2001] has one of the most comprehensive coverage of the problem by dealing with a hierarchy of 8 categories (plus 3 "others") progressively using specifically designed features for different classes. The vacation photos used in their experiments are a mixture of Corel photos, personal photos, video key frames, and

photos from the web.

Image classification or class-based retrieval approaches (such as those class-based probabilistic retrieval frameworks reviewed above) are adequate for query by *predefined* image class. However, the set of relevant images R may not correspond to any predefined class C in general. In the research presented in this thesis, image classification is not the end by itself but a means to provide discriminative image indexes for similarity-based matching and retrieval. It is used to bootstrap the recurrent local semantic regions that discriminate classes of images. Image classification is also used to support event-based retrieval, to compute relative inter-class semantic image indexes, and to embed as local class patterns in image indexes.

2.8 Query Formulation

The call for user interpretation in an image indexing and retrieval system can occur at three stages, namely pre-query, query, and post-query interventions. We have discussed text annotation as a form of manual indexing related to pre-query interpretation. Post-query intervention is required when the user is asked to feedback the relevance of the retrieved images to the system.

In fact, relevance feedback is regarded as a promising technique to bridge the semantic gap in image retrieval [Cox et al., 2000] [Rui et al., 1997]. However the correctness of user's feedback may not be statistically reflected due to the small sampling problem. Although innovative techniques have been proposed to increase the number of training examples with relevance feedback, the experimental results are not conclusive yet [Wu et al., 2000b] [Tieu and Viola, 2000].

The VISMap system [Chang and Chen, 2001] replaces the relevance feedback model with principles from information visualization and concept representation. A rich set of tools are provided for users to construct personal views of the video database and directly visualize and manipulate various views and comprehend effects of individual query criteria on the final search results.

An interesting interface model based on guided exploration has also been explored [Santini et al., 2001]. The interface expects the user to feedback positive and negative examples and to manipulate the image space directly by moving images

around to reflect their perceived similarities. The semantics of an image is emergent as the users learn what the image database has to offer and redefine their goals based on what they have seen. However, this paradigm requires complex database organization to support arbitrary (or almost arbitrary) similarity measures and users' understanding of the interaction metaphor. Though a novel and promising way for image retrieval, it has not been evaluated on a systematic basis yet.

An inevitable situation that requires user interpretation is during query specification when the user has to express his or her information need as some query input to an image retrieval system. In this thesis, we focus on the *semantic interpretation problem* (c.f. Section 1.1.3) related to query specification, rather than pre-query and post-query user intervention. Below we review existing query formulation methods (QBK has been discussed above).

Query By Example (QBE) is an intuitive query formulation metaphor for image retrieval (e.g. QBIC [Flickner et al., 1995], Photobook [Pentland et al., 1995]). A user selects or submits an image as a query example and requests the system to look for images that are visually similar to the query image. However it suffers from the bootstrapping problem. That is, it requires a relevant image to be visible or available as a query example to start with the search. Different methods have been proposed to solve the bootstrapping problem. For examples, the ImageRover [Taycher et al., 1997] and the WebSEEk [Smith and Chang, 1997] systems deploy text-based queries to obtain an initial set of images, and the PicToSeek [Gevers and Smeulders, 1997] approach allows the user to supply a query image. As an enhancement to QBE, the Query By Multiple Regions (QBMR) approach [Moghaddam et al., 2001] allows a composition of query from multiple "regions-of-interest" from example images with or without spatial layout.

Query By Canvas (QBC) allows a user to compose a visual query using geometrical shapes, colors and textures in the drawing canvas of a graphical editor (e.g. QBIC [Flickner et al., 1995], Virage [Bach et al., 1996]). The user expects the system to understand the semantics that is represented by the drawn graphics-based query. However, this approach inherently tends to specify things/stuff of interest in an indirect way using primitive features. For example, one would draw an orange circle and expect the system to know that it represents the sun, though it can also

represent an orange, an orange balloon etc. Moreover the similarity matching between query and images relies on effective pre-segmentation of regions in the images which is complex and difficult in general.

Query By Sketches (QBS) is another interesting visual query method whereby a user outlines the shape of an object as query (e.g. [Del Bimbo and Pala, 1997] [Daoudi and Matusiak, 2000]). A difficulty in this method is that a shape does not have a mathematical definition that exactly matches what the user perceives as a shape [Daoudi and Matusiak, 2000]. And it may not be easy for some users to articulate a shape precisely nor for any user to draw the shapes of certain real-life objects without ambiguity (e.g. tree, sitting person, mountain etc). Since automatic object shape extraction from images (especially in cluttered scenes) is an open problem, applications of QBS have been limited to images with dominant objects on uniform background [Daoudi and Matusiak, 2000].

Since automatic region segmentation and shape extraction are in general very difficult problems, researchers have also proposed to allow a user to guide the query process. For example, as described in [Cinque et al., 2000], the query image is presented to the user at several stages of segmentation and the user is allowed to select the best segmentation, adjust a segmentation, and assign importance values to regions.

A pioneering effort related to the QBC and QBS paradigms for video retrieval is the VideoQ system [Chang et al., 1997a] [Chang et al., 1998b]. It allows video query by animated sketches. Automatic video object segmentation and tracking are performed for the videos in the database. For each segmented object, visual features such as color, texture, shape, and motion as well as spatio-temporal relationships are extracted as indexes to support queries that involve spatio-temporal arrangements of multiple objects, specified using trajectories of shapes of different colors and textures.

A relatively new query paradigm that allows explicit placement of visual semantic icons (e.g. face, sky, building etc) on a canvas has been proposed independently [Lew, 2000] [Lim, 2000a] [Lim, 2001a]. Unlike the discussed query formulation methods that expect the retrieval system to guess a user's intention expressed implicitly in the query (i.e. by example(s), by a composition of graphical primitives, by a

sketch of shape), Query by Spatial Icons (QBSI) [Lim, 2000a] [Lim, 2001a] lets user specify a query explicitly using higher level of visual semantics represented by visual icons with spatial constraints in a Boolean expression. For example, a user can specify pool water, sunflowers, or crowd if they are part of the visual vocabulary of the system. In the case of implicit query expression, specifying pool water, sunflowers, or crowd is unnatural, if not impossible.

The QBSI approach and its comparison with related query formulation methods will be elaborated in Section 7.4.

2.9 Feature Fusion

The problem of combining information from multiple sources to make a better classification decision has always been an active research area in pattern recognition and statistical learning. In fact, a series of international workshops on Multiple Classifier Systems (http://www.diee.unica.it/mcs/) have been organized to bring together researchers of the diverse communities working in the field of multiple classifier systems. Recently, multiple classifier systems have also gained attention in the multimedia analysis research community to exploit multi-modal cues and improve system performance [Smith et al., 2001] [Lin and Hauptmann, 2002] [Amir et al., 2003] [Li et al., 2003a] [Hsu and Chang, 2004] [Snoek et al., 2004].

In this thesis, we focus on the need of feature fusion in content-based indexing and retrieval rather than general multiple classifier systems. From the perspective of feature fusion, the IBM research group divides the feature fusion approaches into Early Feature Fusion and Late Feature Fusion [Smith et al., 2001]. In the Early Feature Fusion approach, various features are processed and integrated into a single feature vector for the pattern classifier. In the Late Feature Fusion approach, the outputs of pattern classifiers based on separate feature vectors are processed and combined to obtain a final classification decision.

The IBM team has experimented with both the Early Feature Fusion approach [Naphade et al., 2003, Naphade and Smith, 2003] and the Late Feature Fusion approach [Tseng et al., 2003] [Iyengar et al., 2003] using the TRECVID benchmark video corpus for visual concept detection tasks (more details can be found at the

website http://www-nlpir.nist.gov/projects/trecvid/).

In one set of experiments on Early Feature Fusion, various color, texture, and shape features at both global and region levels are extracted from the key-frame of a segmented shot and concatenated into a 232-dimension feature vector for SVM learning [Naphade et al., 2003] [Naphade and Smith, 2003].

In another set of experiments using a Late Feature Fusion approach called *normalized ensemble fusion*, separate SVM models are learned for each feature. The confidence scores from each classifier are normalized, aggregated, and optimized in a three-stage process with different data sets to improve classification performance [Tseng et al., 2003].

In yet another exploration effort, a meta-level SVM classifier is trained on the new feature space of classifier scores for classifier fusion [Iyengar et al., 2003]. As a whole, based on the latest slides for the TRECVID 2003 experiments [Amir et al., 2003] and a private communication with one of the IBM authors, the Late Feature Fusion approach is preferred as it has delivered better experimental results and the high dimensionality and normalization issues associated with the Early Feature Fusion approach has no elegant solution.

In particular, for the data set and experiments in this thesis, we focus on the fusion of color and texture features. Color texture discrimination for image segmentation, classification, and retrieval tasks is a challenging problem in image processing and computer vision [Maenpaa et al., 2002]. There are two approaches proposed for color texture discrimination.

In the approach of joint color texture features, spatial interactions within or/and between color bands are considered. For example, a multiscale Gabor representation that includes both unichrome features computed from each spectral band independently and opponent features that captured the spatial correlation between spectral bands has been propsoed [Jain and Healey, 1998]. The opponent features are modeled after the opponent processes in the human visual system and are found to improve the classification of 80 color texture images over the unichrome features empirically.

In another approach for color texture analysis is to divide the color signal into luminance and chrominance components, and process them separately. Interestingly,

there is also biological evidence showing that the image signal in human eye is composed of a luminance and a chrominance component, both of which processed by separate pathways, although there are some secondary interactions between the pathways [Pietikainen et al., 2002]. Psychophysical studies [Poirson and Wandell, 1996] also suggest that color and pattern information are processed separately.

In the experiments conducted on small texture databases (54 VisTex images and 68 Outex images) [Maenpaa et al., 2002], joint color texture features are not the best ones in the classification tasks. Although color histograms are very discriminative in the experiments, they are rather sensitive to changes in illumination. On the other hand, texture features provide fairly robust performance regardless of illumination. Two methods of combining color and texture features have been proposed. In one method, separate dissimilarity measures are used for color and texture feature vectors and summed up to produce an overall dissimilarity during classification. This method requires the normalization of the dissimilarities to reduce the effects of incompatible dissimilarity value ranges.

In another method, the classification results (class rankings) based on separate color and texture feature vectors are combined using the Borda count decision criterion. Both methods have achieved better results than the joint color texture approach though it is not conclusive from the experimental results that which method is superior. The authors concluded [Pietikainen et al., 2002] [Maenpaa et al., 2002] that color and texture have complementary roles. Hence they should be processed independently to allow optimization of color and texture measures separately.

In this thesis, we have adopted the Early Feature Fusion approach for the learning and indexing of local semantic regions (with justification given in Section 3.3.3). In particular, we have studied and compared different Early Feature Fusion methods for color and texture features in SVM learning in Section 3.3.3. Among the methods attempted, the proposed distance and similarity fusion method that resolves the high dimensionality and normalization issues has achieved the best generalization performance.

2.10 Automatic Annotation

While supervised pattern classifiers allow the design of image semantics, either local object classes (c.f. Section 2.5) or global scene classes (c.f. Section 2.7), a major drawback of the supervised learning paradigm is the human effort required to provide labeled training samples, especially at the image region level.

In the field of computer vision, researchers have been pushing the limit of learning by developing object recognition systems from unlabeled and unsegmented images [Fergus et al., 2003] [Selinger and Nelson, 2001] [Weber et al., 2000]. For the purpose of image retrieval, unsupervised models based on "generic" texture-like descriptors without explicit object semantics can also be earned from images without manual extraction of objects or features [Schmid, 2001]. As a representative of the state-of-the-art, sophiscated generative and probabilistic model has been proposed to represent, learn, and detect object parts, locations, scales, and appearances from fairly cluttered scenes with promising results [Fergus et al., 2003].

In the context of relevance feedback, unlabeled images have also been used to boost the learning from very limited labeled examples (e.g. [Wang et al., 2003] [Wu et al., 2000b]). In particular, the MiAlbum system exploits relevance feedback method [Lu et al., 2000] to automatically produce annotation for consumer photos [Liu et al., 2000]. The text keywords in a query are assigned to positive feedback examples (i.e. retrieved images that are considered relevant by the user who issues the query). This would require constant user intervention (in the form of relevance feedback) and the keywords issued in a query might not necessarily correspond to what is considered relevant in the positive examples.

In the Intelligent Multimedia Knowledge Application (IMKA) project, Benitez and Chang proposes a framework for representing and discovering knowledge from multimedia content to enhance the classification, navigation and retrieval of multimedia [Benitez and Chang, 2003a]. The MediaNet knowledge representation unifies both perceptual and semantic concepts and relationships exemplified by media [Benitez et al., 2000].

Using a collection of 3624 annotated nature and news images, perceptual and semantic knowledge are automatically discovered by integrating both the processing of images and text. Perceptual knowledge is constructed by clustering the images

based on both visual and text feature descriptors, and by discovering statistical and similarity relationships between the clusters [Benitez and Chang, 2002a]. Using WordNet and the image clusters, semantic knowledge is further constructed by disambiguating the senses of words in annotations, and by finding semantic relations between the detected senses in WordNet [Benitez and Chang, 2002b]. More recently, interdependence among discovered concepts are used to construct Bayesian networks for probabilistic inferencing in image classification task with promising results [Benitez and Chang, 2003b].

Motivated from a machine translation perspective, object recognition is posed as a lexicon learning problem to translate image regions to corresponding words [Duygulu et al., 2002]. More generally, the joint distribution of meaningful text descriptions and entire or local image contents are learned from images or categories of images labeled with a few words [Barnard and Forsyth, 2001] [Barnard et al., 2003b] [Kutics et al., 2003] [Li and Wang, 2003]. The lexicon learning metaphor offers a new way of looking at object recognition [Duygulu et al., 2002] and a powerful means to annotate entire images with concepts evoked by what is visible in the image and specific words (e.g. fitness, holiday, Paris) [Li and Wang, 2003]. While the results for the annotation problem on entire images look promising [Li and Wang, 2003], the correspondence problem of associating words with segmented image regions remains challenging [Barnard et al., 2003b] as segmentation, feature selection, and shape representation are critical and non-trivial [Barnard et al., 2003a].

While the approaches described above attempt to automate image annotation by using content-based analysis with or without associated text information, another approach exemplified by the Google Image Search tool (www.google.com/imghp) is to index images based on the text that describes a given image (e.g. filename, URL etc) and possibly other non-content-based information (e.g. citation-based). Hence it is not surprising that the images returned by this approach may have content irrelevant to the intended query. For instance, a search with the keyword 'Paris' to look for images of the French capital Paris may return portrait images of people with the name 'Paris'. On 25 March 2004, the 39^{th} image returned by Google Image Search using keyword 'Paris' shows a man *Jon Paris plays "Born to Be Wild" to a crowd that understands* (www.jsonline.com/general/harley95/images/paris.jp).

In this thesis, we address the issue of minimal supervision in a different direction. We do not assume availability of text descriptions for image or image classes as in [Barnard et al., 2003b] [Li and Wang, 2003] [Benitez and Chang, 2003a]. Neither do we know the object classes to be recognized as in [Fergus et al., 2003]. A novel semi-supervised framework is proposed to discover and associate local unsegmented regions with semantics and generate their samples so as to construct semantic models for content-based image retrieval, all with minimal human intervention.

Chapter 3

Semantics Design

Few things are harder to put up with than the annoyance of a good example.

Mark Twain (1835 - 1910)

3.1 Semantic Support Regions

In this chapter, we address the issue of high content diversity with a structured learning framework to allow modular design and extraction of domain-relevant visual semantics in building content-based image retrieval systems. To realize strong semantic interpretation of content, we propose the use of salient image regions, known as *Semantic Support Regions* (SSRs), that exhibit semantic meanings to human users to support image indexing. These are similar to the *signs* designed for domain-specific applications ([Smeulders et al., 2000], pp. 1359) and the *Visual Keywords* handcrafted for explicit query specification [Lim, 2000a] [Lim, 2001a].

In a nutshell, the proposed SSR framework incorporates modular view-based object detectors to generate spatial semantic signatures for similarity-based and fuzzy logic-based query processing without region segmentation. Hence our approach is not restricted to images that have the main area of attention, which are assumed by other approaches that attempted object-based indexing and retrieval [Martinez and Serra, 2000] [Tao and Grosky, 2000].

SSRs are salient image patches that have semantic meanings to us and that can be learned statistically to span a new indexing space. A cropped face region, a typical grass patch, and a patch of swimming pool water etc can all be treated as their instances. SSRs can be linked to both specific and more general concepts by textual labels in a vocabulary and thesaurus to provide relevant semantics to an image domain. Without loss of generality, we consider two levels of concept hierarchy here for simplicity. Suppose a visual concept with text label C can have different appearances each associated with text label S_i,

$$C : S_1, S_2, \cdots, S_i, \cdots \tag{3.1}$$

For instance, the concept 'Sky' in a given image domain may appear as 'Clear', 'Blue', and 'Cloudy' (i.e. Equation (3.1) becomes Sky: Clear, Blue, Cloudy). And each SSR with text label S_i is represented as a set of instances of that particular kind of sky ('Sky:Clear', 'Sky:Blue', 'Sky:Cloudy'),

$$S_i : (s_{i1}, z_{i1}), (s_{i2}, z_{i2}), \cdots, (s_{ij}, z_{ij}) \cdots \tag{3.2}$$

where s_{ij} are optional text labels associated with an instance, and z_{ij} denote some computable representation such as feature vectors of the instances. This concept-oriented visual thesaurus is different from the visual relations proposed by Picard [Picard, 1995], which are founded on similarities between low-level visual features. Thus SSRs are highly flexible visual knowledge that can be customized according to a content domain.

Different from the unsupervised Visual Keywords [Lim, 2000d] and the manually selected Visual Keywords [Lim, 2001a], the SSRs are learned a priori and detected during image indexing from multi-scale block-based image regions, as inspired by multi-resolution view-based object recognition framework [Papageorgiou et al., 1998] [Sung and Poggio, 1998] [Rowley et al., 1998], hence without a region segmentation step. The key in image indexing here is not to record the primitive feature vectors themselves but to project them into a classification space spanned by semantic labels and use the soft classification decisions as the local indexes for further spatial aggregation.

Indeed the late K.K. Sung also constructed six face clusters and six non-face clusters and used the distance between the feature vector of a local image block and these clusters as the input to the trained face detector rather than using the feature vector directly [Sung and Poggio, 1998].

Figure 3.1 summarizes our proposed framework in a schematic diagram. In the figure, arrows with solid heads denote processing steps and arrows with empty heads represent matching. Given an image to be indexed, multi-scale view-based detection against the learned SSRs is first carried out. The detection results are reconciled into a fine-grained common representation for spatial aggregation. The compact coarse-grained image index, shown as a 3×3 grid of SSR histograms (in the middle of the top row), can then be used to support similarity-based matching for Query by Example (QBE) and fuzzy query processing for Query by Spatial Icons (QBSI) (both to be detailed in Chapter 7).

The SSR framework is applied to indexing and retrieval of consumer images which contain highly varied contents, diverse resolutions and inconsistent quality in this thesis. To bridge the semantic gap, SSRs possess the following properties:

- SSRs are designed with strong semantics in a concept hierarchy;
- SSRs are built upon modular learning from examples;
- SSRs are extracted directly from images without segmentation;
- SSRs are detected from multi-scale tessellated image blocks and reconciled to account for translation and scale variances;
- Spatial information is retained in the index based on SSRs, so Query by Spatial Icons (QBSI) (Section 7.4) can be naturally and efficiently applied;
- SSRs provide a mid-level representation: indexing and matching are performed in a higher level classification space, rather than low-level feature space;
- SSRs are not specific to particular feature, classifier, and tessellation.

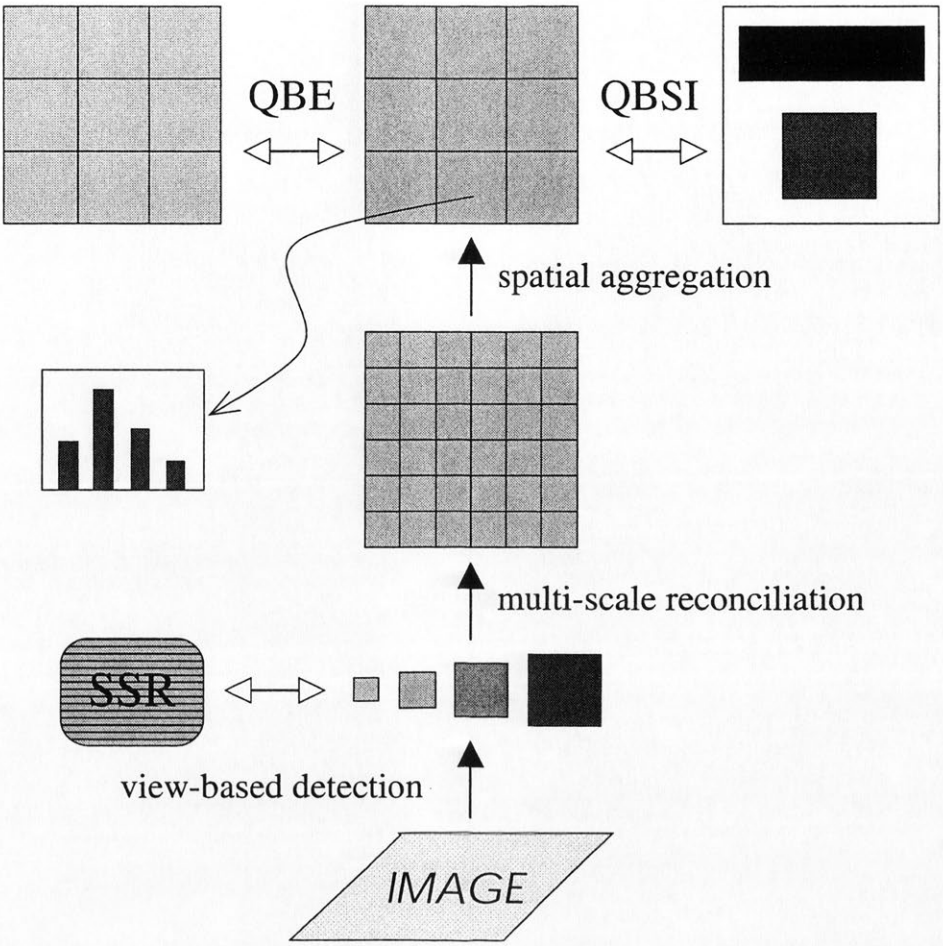

Figure 3.1: A structured learning framework for indexing and query

3.2 Features

As we mentioned above, SSRs are mid-level semantics grounded on low-level visual features but not specific to a particular feature. Given an image domain, a set of SSRs (i.e. a visual vocabulary) considered useful for query and retrieval is determined. Then based on the set of SSRs for learning and detection, appropriate visual features related to color, texture, and shape are designed.

Since we are dealing with unconstrained consumer images in this thesis, color and texture features are considered important and shape information is not used. For instance, while color is useful to characterize sky, water, face etc, texture will play a role in discriminating buildings, crowd, trees etc. Due to the fact that there is usually no dominant object with clear background and objects such as sky, mountain, water,

building etc do not have consistent shapes, shape feature is not computed. Hence for the implementation of the methods presented in this thesis, a SSR is characterized using both color and texture features. A feature vector z has two parts, namely, a color feature vector z^c and a texture feature vector z^t.

3.2.1 Color

Color has been considered a powerful descriptor that often simplifies object idenification and extraction from an image [Gonzalez and Woods, 1992]. For example, local color histogram has been demonstrated in locating an "object" in a color image [Ennesser and Medioni, 1995]. On the other hand, color is an important perceptual cue as human eye can discern thousands of color shades and intensities [Gonzalez and Woods, 1992].

Perceptually uniform spaces such as L*a*b and approximately-uniform color spaces, such as HSV, have been touted as preferred color spaces for color-based image retrieval as measured color differences in these spaces are proportional to the human perception of such differences. A recent empirical evaluation on image classification using a small data set with color texture features based on Gabor filters also seemed to confirm their advantages over the non-uniform RGB color space [Paschos, 2001]. However, in our experiments reported in Chapter 7, we have not found much difference in performance among the different color models attempted (RGB, YIQ, HSV, L*a*b, L*u*v). Since the conversion from RGB space to YIQ space requires a simple matrix multiplication (see below), we have adopted the YIQ color space in our experiments reported in this thesis.

The YIQ model is used in commercial color TV broadcasting. The luminance (Y) component is decoupled from the color information (I and Q) and can be used directly for texture feature extraction. The computation of YIQ values from the raw RGB values ($\in [0,1]$) is a simple transformation [Gonzalez and Woods, 1992],

$$\begin{bmatrix} Y \\ I \\ Q \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ 0.596 & -0.275 & -0.321 \\ 0.212 & -0.523 & 0.311 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}, \tag{3.3}$$

resulting in Y, I, and Q in the ranges $[0, 1]$, $[-0.596, -0.596]$, and $[-0.523, 0.523]$ respectively.

For the color feature vector z^c, as the image patch for training and detection is relatively small (20×20 to 60×60), the mean and standard deviation of each color channel is deemed sufficient (i.e. z^c has six dimensions). Hence the color information of an image region or block is represented as

$$z^c = [\mu_Y, \sigma_Y, \mu_I, \sigma_I, \mu_Q, \sigma_Q]. \tag{3.4}$$

We have also tested local color histograms [Ennesser and Medioni, 1995] with histogram intersection as similarity measure [Swain and Ballard, 1991]. But as it requires more feature dimenions and yet does not outperform the simple second-order statistical feature of Equation (3.4) probably due to quantization errors, we have adopted mean and standard deviation in our experiments.

3.2.2 Texture

Image texture, defined as a function of the spatial variation in pixel intensities (gray values), has been studied extensively in many computer vision problems such as texture segmentation and classification [Tuceryan and Jain, 1998]. Many analysis methods ranging from statistical, geometrical, model-based, to multi-resolution filtering techniques have been proposed.

In particular, pattern retrieval using a simple multi-resolution representation based on Gabor filters has shown promising performance [Manjunath and Ma, 1996]. Besides the motivation in biological modeling of the receptive fields of simple cells in the visual cortex of some mammals [Daugman, 1980], the Gabor representation has also been shown to be optimal in the sense of minimizing the joint two-dimensional uncertainty in space and frequency [Daugman, 1988]. For the purpose of feature representation, Gabor filters can be considered as orientation and scale turnable edge and line (bar) detectors, and the statistics of these microfeatures in a given image region are often used to characterize the underlying texture information.

In this thesis, the texture feature proposed in [Manjunath and Ma, 1996] is adopted. We shall not repeat the details on Gabor function and filter bank de-

sign here but focus only on the texture feature extraction. Given an image $I(x, y)$, its Gabor wavelet transform is defined as

$$W_{mn}(x, y) = \int I(x_1, y_1) g_{mn} * (x - x_1, y - y_1) dx_1 dy_1 \tag{3.5}$$

where $*$ denotes the complex conjugate. The assumption that the local texture regions are spatially homogeneous is valid in our case as the image patch for training and detection is relatively small (20×20 to 60×60).

Hence for the texture feature vector z^t, the mean μ_{mn} and the standard deviation σ_{mn} of the magnitude of the transform coefficients are used to represent an image region. In our experiments, five scales and six orientations are used, resulting in a feature vector,

$$z^t = [\mu_{00}, \sigma_{00}, \mu_{01}, \sigma_{01}, \cdots, \mu_{45}, \sigma_{45}]. \tag{3.6}$$

where μ_{mn} and σ_{mn} are computed as,

$$\mu_{mn} = \int\int |W_{mn}(x, y)| dx dy, \tag{3.7}$$

$$\sigma_{mn} = \sqrt{\int\int (|W_{mn}(x, y)| - \mu_{mn})^2 dx dy}. \tag{3.8}$$

3.2.3 Normalization

As the feature elements of the color and texture feature vectors have different ranges, it is important to normalize them into the same range so that each feature element contributes equal weight in a distance or similarity function for a feature vector. There are two common normalization schemes in the pattern recognition and neural network research community. Both schemes require a representative set of feature vectors of images drawn from the domain to determine the parameters related to the distribution of the feature elements. Ideally all the possible image regions in the database should be used to compute these parameters. In this thesis, we have used the feature vectors of the image regions in the training set for SSR learning to estimate the normalization parameters since they cover key semantic regions for the domain and it is more practical to deal with a sample set than all possible image regions.

Suppose the parameter estimation set has M feature vectors (i.e. image regions) and a feature vector has N feature elements. Then each feature element is denoted as z_{ij} ($i = 1$ to M, $j = 1$ to N), either from z_i^c or z_i^t. The first normalization method computes the smallest and largest feature values (over M feature vectors) in each feature dimension as min_j and max_j respectively. A feature element z_{ij} is normalized to $[0, 1]$ as,

$$z'_{ij} = \frac{z_{ij} - min_j}{max_j - min_j}. \tag{3.9}$$

However this normalization schemes suffers from the problem of outlier i.e. if there is some extreme value (e.g. large number) in a feature element, the other feature values will be warped into a very narrow range.

Hence in this thesis, we have adopted the zero-mean normalization scheme, also known as the Gaussian normalization [Ortega et al., 1997]. In this scheme, the mean m_j and standard deviation s_j are computed for each feature dimension based on the M feature vectors in the parameter estimation set. Then a feature element z_{ij} is transformed into,

$$z'_{ij} = \frac{z_{ij} - m_j}{s_j}. \tag{3.10}$$

Based on the properties of Gaussian distribution, the normalized feature values will fall in the range of $[-1, 1]$ with a probability of 0.68. If $3s_j$ is used as the denominator in Equation (3.10), appromixmately 99% of the normalized feature values will be in $[-1, 1]$. As we do not require all the feature values to be strictly in the range of $[-1, 1]$, we have implemented Equation (3.10) in this thesis.

Note that while the μ and σ in Equations (3.4) and (3.6) are sample (i.e. pixels in an image region) mean and standard deviation respectively, m_j and s_j are the estimated population mean and standard deviation for the feature elements (i.e. μ and σ) respectively. As we have 6 color feature elements and 60 texture feature elements, we have a total of $66 \times 2 = 132$ normalization parameters. They are computed only once from the SSR training set and utilized in both SSR learning (Section 3.3) and SSR detection during image indexing (Section 3.4). For simplicity, we shall drop the 'prime' superscript in z'_{ij} when we discuss feature elements in the rest of the thesis i.e. a feature vector z or feature element z_k refer to their normalized versions.

3.2.4 Distance and Similarity

In this thesis, we have experimented with both distance (i.e. dissimlarity) and similarity measures for comparing two feature vectors (color or texture). Although many sophiscated dissimilarity measures have been proposed and evaluated empirically [Puzicha et al., 1999], we opt for measures that require simple computation for practical reasons.

Since a simple city block distance (L_1-norm) has better learning and detection performance than other distance measures such as Euclidean distance (L_2-norm) etc, we have adopted it in our experiments. The distance between two feature vectors (color or texture) y and z is computed as,

$$\Delta(y,z) = \sum_j |y_j - z_j|. \tag{3.11}$$

In fact, in the case of texture feature vectors, this city block distance on the normalized feature elements turns out to be equivalent to the distance measure used in [Manjunath and Ma, 1996] (reproduced here)

$$d(i,j) = \sum_m \sum_n (|\frac{\mu_{mn}^{(i)} - \mu_{mn}^{(j)}}{\alpha(\mu_{mn})}| + |\frac{\sigma_{mn}^{(i)} - \sigma_{mn}^{(j)}}{\alpha(\sigma_{mn})}|), \tag{3.12}$$

where i and j are image patterns, μ_{mn} and σ_{mn} are the mean and standard deviation with m scales and n orientations, and $\alpha(\mu_{mn})$ and $\alpha(\sigma_{mn})$ are the standard deviations of the respective features over the entire database, and are used to normalize the individual feature elements. The substraction of the term m_j in Equation (3.10) gets cancelled off and simplified into the numerators in Equation (3.12) and the $\alpha(.)$ are the s_j in Equation (3.10).

As for the similarity measure, the cosine measure (i.e. normalized dot product) popular in the information retrieval research community [Salton, 1971] is used, i.e. for vectors y and z,

$$\Omega(y,z) = \frac{y \cdot z}{|y||z|}, \tag{3.13}$$

where $\cdot$ indicates a dot product.

3.3 Learning

3.3.1 Support Vector Machines

Though the theory of statistical learning [Vapnik, 1979, Vapnik, 1995, Vapnik, 1998] behind Support Vector Machines (SVM) has been proposed many year ago, SVM only became a popular machine learning tool since the mid nineties. One reason being that efficient implementations of support vector learning were proposed (e.g. [Joachims, 1999] [Platt, 1999a]) and made available at the website for Kernel Machines resources (www.kernel-machines.org).

The other reason that SVM has received much attention is their superior generalization performance (i.e. small error rates) in many applications such as handwritten digit recognition [LeCun et al., 1995] [Burges and Scholkopf, 1997], object recognition [Blanz et al., 1996], speaker identification [Schmid, 1996], face detection [Osuna et al., 1997], time series prediction [Muller et al., 1997], text categorization [Joachims, 1998] etc when compared to other competing methods.

In this thesis, we have decided to adopt SVM as the key statistical learning technique. Since SVM is heavily used in our experiments, we shall present a brief overview on the key concepts of SVM for pattern classification here. For a detailed introduction on the subject, readers are referred to the excellent tutorials on SVM [Burges, 1998] [Scholkopf, 2000] [Cristianni and Shawe-Taylor, 2000]. For a broader treatment and review on kernel-based learning algorithms that include applications in classification, regression, and unsupervised learning, the article by Muller et al. [Muller et al., 2001] is a good starting point.

In pattern classification, the objective is to learn the mapping,

$$\begin{aligned} f: \ \mathcal{X} &\rightarrow \mathcal{Y} \\ x &\mapsto \{\pm 1\}, \end{aligned} \tag{3.14}$$

from examples

$$(x_1, y_1), \cdots, (x_m, y_m) \in \mathcal{X} \times \{\pm 1\}, \tag{3.15}$$

and be able to generalize to unseen data points.

The key idea of SVM learning for pattern classification is to map the training

vectors x ($\in \mathcal{X}$) into a higher-dimensional feature space $\mathcal{F}$ via a mapping function Φ, and construct a separating hyperplane with maximum margin in $\mathcal{F}$ (Figure 3.2),

$$\begin{aligned} \Phi: \mathcal{X} &\to \mathcal{F} \\ x &\mapsto \Phi(x), \end{aligned} \tag{3.16}$$

where $\mathcal{F}$ is a dot product space.

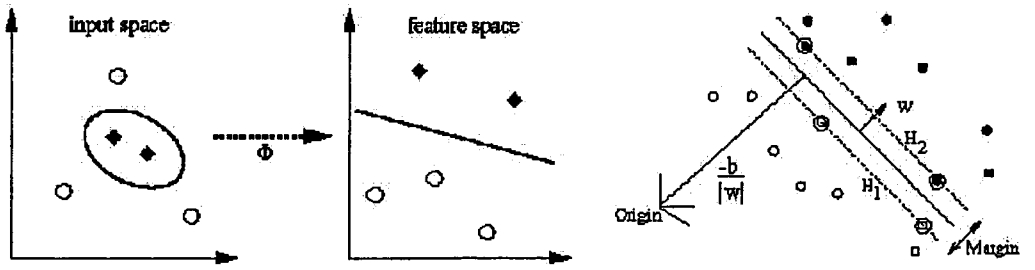

Figure 3.2: The idea of SVM learning for pattern classification

The beauty of SVM is that by using a kernel function, the computation of the separating hyperplane can be performed implicitly, without explicitly carrying out the map into the feature space. That is, the computation of a scalar product between two feature space vectors can be readily reformulated in terms of a kernel function k,

$$(\Phi(x) \cdot \Phi(x')) =: k(x, x'). \tag{3.17}$$

A simple example to illustrate the power of mapping to a dot product space is given in Figure 3.3 (duplicated from [Muller et al., 2001]),

$$\begin{aligned} \Phi : \Re^2 &\to \Re^3 \\ (x_1, x_2) \mapsto (z_1, z_2, z_3) &:= (x_1^2, \sqrt{2}x_1x_2, x_2^2) \end{aligned} \tag{3.18}$$

In the original two-dimensional data space, a rather complicated nonlinear decision surface is necessary to separate the classes. But in a feature space of second order monomials, a linear hyperplane is sufficient to separate the classes. With the notion of kernel function, the computation of a scalar product between two feature

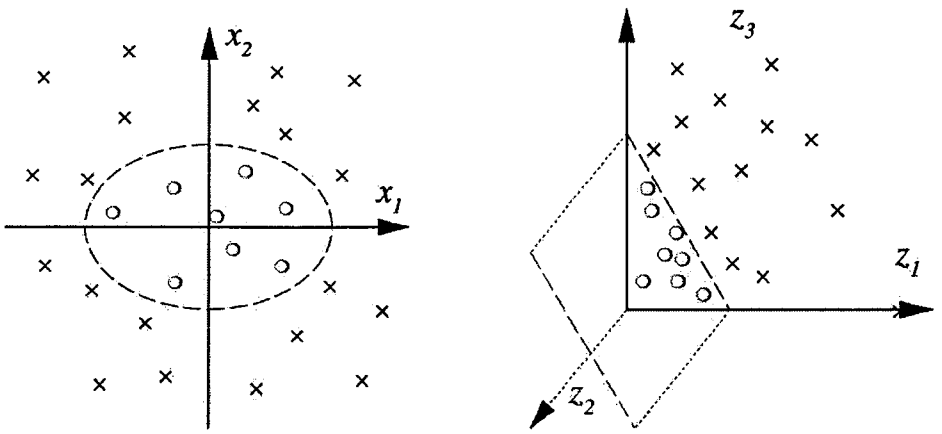

Figure 3.3: An example of SVM learning with mapping from $\Re^2$ to $\Re^3$

space vectors can be reformulated in terms of a kernel function,

$$\begin{aligned}(\Phi(x)\cdot\Phi(x')) &= (x_1^2,\sqrt{2}x_1x_2,x_2^2)(x_1'^2,\sqrt{2}x_1'x_2',x_2'^2)^T \\ &= ((x_1,x_2)(x_1',x_2')^T)^2 \\ &= (x,x')^2 \\ &=: k(x,x') \end{aligned} \tag{3.19}$$

i.e. the dot product in $\Re^3$ can be computed in $\Re^2$.

Among many kernel functions studied, the following kernel functions, namely *Polynomial*, *Sigmoidal*, and *Gaussian RBF* (radial basis function), respectively are commonly used

$$k(x,x') = ((x\cdot x')+\theta)^d, d\in\aleph,\theta\in\Re \tag{3.20}$$

$$k(x,x') = tanh(\kappa(x,x')+\theta),\kappa,\theta\in\Re \tag{3.21}$$

$$k(x,x') = exp(\frac{-\|x-x'\|^2}{c}),c\in\Re \tag{3.22}$$

The task of supervised learning in SVM is to find the support vectors x_i, a subset of training patterns from the training data, that are closest to the separating hyperplane with non-zero weights α_i. These support vectors are marked by extra circles in Figure 3.4 that illustrates a support vector classifier found by using a Gaussian RBF kernel function (Equation (3.22)). In the same figure, the two classes

of training examples are denoted by circles and disks respectively and the middle line is the decision surface.

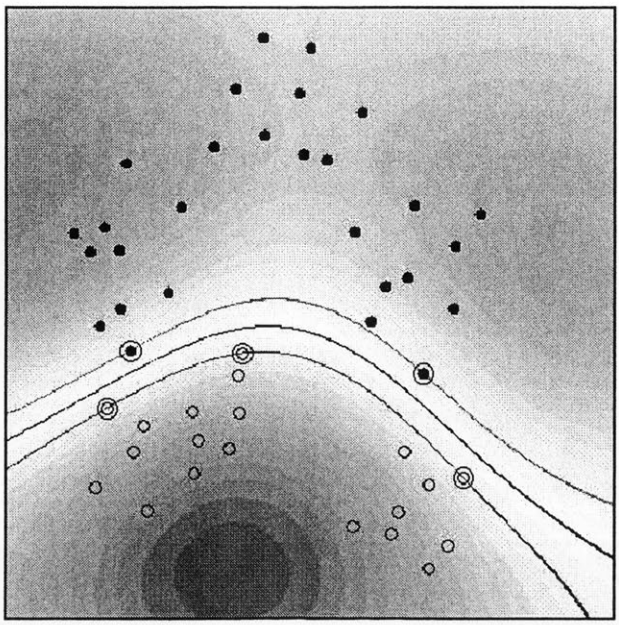

Figure 3.4: An example of support vector classifier to separate two classes

Once the support vectors have been found, via some optimization procedure, a given pattern x can be classified based on the following hyperplane decision function (upon m support vectors x_i and weights α_i, and a threshold b),

$$\begin{aligned} f(x) &= sgn(\sum_{i=1}^{m} y_i \alpha_i \cdot (\Phi(x) \cdot \Phi(x_i)) + b) \\ &= sgn(\sum_{i=1}^{m} y_i \alpha_i \cdot k(x \cdot x_i) + b) \end{aligned} \tag{3.23}$$

3.3.2 SSR Learning

The key component in the SSR framework is statistical learning of the SSRs from examples. In this thesis, Support Vector Machines (SVM) [Vapnik, 1979] [Vapnik, 1995] [Vapnik, 1998] [Cristianni and Shawe-Taylor, 2000], a popular and powerful discriminative learning method, is adopted for this purpose. SVM is preferred to the probabilistic generative models [Bishop, 1995] as the latter does not emphasize data close to the discriminative boundary. This will affect the classification accuracy especially

when the classes have close multimodal distributions [Kumar et al., 2002].

In particular, as mentioned in Section 2.5, empirical evaluations based on large news video collections [Adams et al., 2003] [Hsu and Chang, 2004] have reported better results using SVM classifers when compared to generative models on visual semantic concept retrieval and news story segmentation tasks respectively. An explanation for the better performance of the SVM classifiers [Adams et al., 2003] is that the SVM classifiers need to model less information in terms of what differentiates a positive example from a negative example and hence requires less data to estimate parameters reliably. This advantage of SVM is important when we are dealing with many visual semantic classes and we would like to minimize the effort of labeling of training and validation samples for image regions.

A local image region is represented by its feature vector z, composed from z^c (Equation (3.4)) and z^t (Equation (3.6)). A support vector classifier $\mathcal{S}_i$ devoted to a SSR class S_i (c.f. Equation (3.2)) is treated as a function on z, $\mathcal{S}_i(z) \in (-\infty, +\infty)$. Then the classification vector T for image region with feature vector z can be computed via the softmax function [Bishop, 1995] [Bridle, 1990] as

$$T_i(z) = \frac{\exp^{\mathcal{S}_i(z)}}{\sum_j \exp^{\mathcal{S}_j(z)}}. \tag{3.24}$$

The softmax function is also known as the normalized exponential activation function in the neural networks and pattern recognition community. It represents a smooth version of the winner-take-all activation model in which the unit with the largest input has output +1 while all other units have output 0. It can also be regarded as a generalization of the logistic activation function [Bishop, 1995]. As the output values of the softmax function lie in the range $(0, 1)$ and they sum to unity, it provides a simple way to interpret $T_i(z)$ as the posterior probability of SSR class S_i given an image region with feature vector z,

$$P(S_i|z) = \frac{\exp^{\mathcal{S}_i(z)}}{\sum_j \exp^{\mathcal{S}_j(z)}}. \tag{3.25}$$

More sophiscated ways to estimate posterior probability from SVM outputs have also been proposed and experimented [Platt, 1999b] but we shall not delve into this subject here.

Alternatively as each SVM classifier can be regarded as an expert on a SSR class, we can adopt a hybrid winner-take-all and softmax scheme. That is, if some SVM classifier(s) $\mathcal{S}_i$ has positive output(s), then the outputs of the other SVM classifiers $\mathcal{S}_j, j \neq i$ are set to zeroes. After this step, the softmax function in Equation (3.24) is applied. More specifically, if there is only one SVM classifier $\mathcal{S}_i$ having positive output, then $T_i(z) = 1$ (and $T_j(z) = 0, j \neq i$). If more than one SVM classifier $\mathcal{S}_i$ has positive outputs, then $T_i(z)$ will be positive values determined by the softmax function while the other $T_j(z) = 0, j \neq i$. Finally if all SVM classifers $\mathcal{S}_i$ $\forall i$ have non-positive outputs, then the values of $T_i(z)$ will be computed as in Equation (3.24). We shall compare the retrieval results based on these two different classification normalization schemes in Chapter 7.

3.3.3 Feature Fusion

As mentioned above (Section 3.2), color and texture features are used to characterize local image regions for the unconstrained consumer images in our experiments. In another word, a feature vector z for an image region has two parts, the color feature vector z^c based on the mean and standard deviation in YIQ color channels (Equation (3.4)), and the texture feature vector z^t based on the mean and standard deviation of output of Gabor filters with 5 scales and 6 orientations (Equation (3.6)).

In general, the choice of combining multiple features for pattern classification is non-trivial. Borrowing the terms coined in [Smith et al., 2001], there are two paradigms in combining features, namely Early Feature Fusion and Late Feature Fusion. In the Early Feature Fusion paradigm, different features are combined into a final feature vector and used as an input to the pattern classifier. In the Late Feature Fusion paradigm, feature vectors associated with different modalities are fed into independent pattern classifiers whose classification outputs are then combined.

As discussed in Section 2.9, fusion of multiple modalities to make a better classification decision is a promising research trend for video indexing and retrieval. Although exhaustive efforts have been attempted with the Late Feature Fusion paradigm by the IBM team and the experimental results also shown a performance improvement of classification based on fusion of classifiers [Smith et al., 2001] [Amir et al., 2003], we have decided not to pursue this line of investigation for the

following drawbacks of this paradigm:

- For each visual concept to be learnt, multiple classifiers have to be trained on different features or feature sets. After which, model and feature selection are necessary to optimize the weights of the first level classifiers in fusion. If cascaded classifers (instead of voting schemes) are adopted to fuse the outputs of the trained classifiers, then additional training on these fusion classifiers have to be carried out after the first level classifiers have been trained;

- Additional validation set is required to test the fusion classifier or to search for the best model parameters and feature set. This will either require more labeled training samples or reduce the number of training samples for the first level classifiers;

- Since fusion of outputs from multiple classifiers is a sequential processing step even though the first level classifiers can be executed concurrently, the recognition process will require longer computation time;

- Training multiple classifiers on individual features may not be viable at all as single feature does not provide sufficient discriminative power, hence may result in many poor classifiers for fusion. In particular, in the case of SSR learning and detection for consumer images in this thesis, training on color and texture features separately have resulted in asymmetric classifiers whose fusion has not outperformed the result of the Early Feature Fusion method proposed below. The experimental results on image retrieval will be presented in Section 7.5.

In this thesis, we have evaluated different approaches in the Early Feature Fusion paradigm for SSR classification empirically in the next subsection. The features and fused features investigated are listed in Table 3.1. The first two rows consider only single modality (i.e. z^c and z^t). The third row combines the color and texture feature vectors in a distance or similarity measure (to be explained below) [Maenpaa et al., 2002], denoted symbolically as z. The fourth row considers simple concatenation of the color and texture feature vectors as z^{c+t} [Naphade and Smith, 2003]. The fifth row computes texture feature vectors for each

Table 3.1: Features and fused features for SSR learning and classification

Feature Vector	Symbol	#dim
Equation (3.4)	z^c	6
Equation (3.6)	z^t	60
Fused matching of z^c and z^t	z	66
Concatenation of z^c and z^t	z^{c+t}	66
z^t from separate RGB channels	$z^{c\times t}$	180
principal components of $z^{c\times t}$	z_{pca}	10

of the RGB channels separately and concatenates the three vectors together as $z^{c\times t}$ [Maenpaa et al., 2002]. The last row retains the 10 most important components of $z^{c\times t}$ via principal component analysis [Johnson and Wichern, 1988] which accounted for 98% percent of the total variance in the data set, denoted as z_{pca}. Note that z is identical in physical form as z^{c+t} but they differ in the computation of distance or similarity function.

The distance or similarity function depends on the kernel adopted for the SVM classifiers. For the features and fused features shown in Table 3.1 (except z in the third row), we have adopted the city block distance Δ (Equation (3.11)) as the distance function for the radial basis function (RBF) kernels and the cosine similarity Ω (Equation (3.13)) as the similarity function for the polynomial kernels.

For z (third row), in order to balance the contributions of the color and texture features, we have modified the distance function between two feature vectors y and z for the RBF kernels as,

$$\Delta(y,z) = \frac{1}{2}\left(\frac{\Delta(y^c,z^c)}{N_c} + \frac{\Delta(y^t,z^t)}{N_t}\right) \tag{3.26}$$

where N_c and N_t are the numbers of dimensions of the color and texture feature vectors (i.e. 6 and 60) respectively.

As the feature elements in both color and texture feature vectors have been normalized (Section 3.2.3) to fall mainly within $[-1,1]$, we need not perform the inter-feature normalization procedure that requires computation of pairwise distances of all feature vectors to estimate the mean and standard deviation of the distance values [Ortega et al., 1997]. We simply divide the distances between two color feature

Table 3.2: SSR classes grouped into 8 superclasses

SSR Superclass	SSR Classes
People	Face, Figure, Crowd, Skin
Sky	Clear, Cloudy, Blue
Ground	Floor, Sand, Grass
Water	Pool, Pond, River
Foliage	Green, Floral, Branch
Mountain	Far, Rocky
Building	Old, City, Far
Interior	Wall, Wooden, China, Fabric, Light

vectors and two texture feature vectors by their feature dimensions respectively as given in Equation (3.26). Note that we have assumed equal importance for both the color and texture features without any prior knowledge. One could also assign different weights to the color and texture distances if necessary.

In a similar manner, the similarity function between two feature vectors y and z for the polynomial kernels is modified as,

$$\Omega(y, z) = \frac{1}{2}(\Omega(y^c, z^c) + \Omega(y^t, z^t)) \tag{3.27}$$

Again, we have assumed equal contributions from both color and texture features.

3.3.4 Learning Evaluation on Consumer Images

For the consumer image data and experiments reported in Chapter 7 of this thesis, 26 classes of SSRs (i.e. $S_i, i = 1, 2, \cdots, 26$ in Equation (3.24)) are designed after studying the test collection. They are organized into 8 superclasses, namely `People`, `Sky`, `Ground`, `Water`, `Foliage`, `Mountain`, `Building`, and `Interior`. Each of them is further divided into several classes as listed in Table 3.2. Figure 3.5 shows, in top-down and left-to-right order, single examples of these 26 classes of SSRs as listed in Table 3.2.

For the learning of these 26 SSR classes and evaluation of different Early Feature Fusion approaches, SVM kernels, and kernel parameters, we have cropped 554 image regions from 138 images in our 2400 consumer image collection and used 375 (i.e.

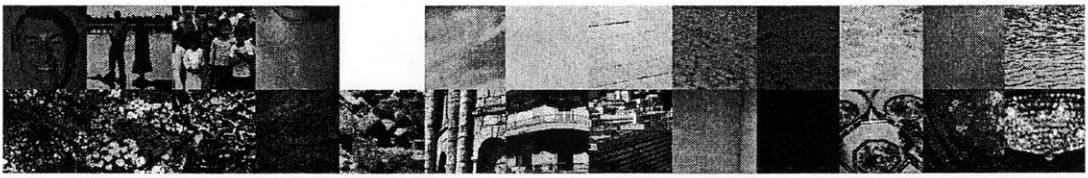

Figure 3.5: Examples of semantic support regions

Table 3.3: Compare features and fused features on SSR generalization

Feat. Vec.	# Err. on D_{tst}	Avg. # Err.
z^c	214	8.2
z^t	278	10.7
z	149	5.7
z^{c+t}	201	7.7
$z^{c\times t}$	267	10.3
z_{pca}	303	11.7

two-third) of them (from 105 images) as training set D_{trg} for SVM learning to compute the support vectors of the SSRs. The remaining one-third (i.e. 179 regions) are used as test set D_{tst} for generalization performance evaluation. In other words, both the training and test data for SSRs utilize only a very small percentage (5.8%) of the 2400 collection.

First we compare the generalization performances of six features and fused features extracted for an image region (Table 3.1) using average numbers of classification errors (over 26 SSR classes) on test set D_{tst}. For the SVM classifiers, polynomial kernels with degree 2 and constant 1 ($C = 100$) [Joachims, 1999] (i.e. similarity measures based on Equations (3.13) and (3.27)) are used.

Clearly, from Table 3.3, the feature fusion method based on modified cosine measure has the best result. These generalization results confirm that using both color and texture features (z, z^{c+t}) is necessary to achieve better performance than using a single feature (z^c, z^t) though some feature may be more discriminative for certain SSRs. For example, color is more useful than texture to classify SSR `Sky:Blue`. However, we prefer not to handcraft feature specificity into the kernel functions. Concatenated feature z^{c+t} is less effective than z as the feature vector is dominated by z^t with 60 dimensions. A long color texture feature vector (i.e. texture from each color channel) $z^{c\times t}$ does not work well probably due to the texture

Table 3.4: Compare polynomial SVM classifiers on SSR generalization

d	# Err. on D_{tst}	Avg. # Err.
2	149	5.73
3	146	5.62
4	145	5.58
5	133	5.12
6	131	5.04
7	134	5.15
8	136	5.23
9	139	5.35
10	142	5.46
20	164	6.31

Table 3.5: Compare RBF SVM classifiers on SSR generalization

α	σ	# Err. on D_{tst}	Avg. # Err.
10	0.22	178	6.85
5	0.32	146	5.62
2	0.50	123	4.73
1	0.71	111	4.27
0.5	1.00	115	4.42
0.1	2.24	126	4.85

feature redundency in all three color channels. However, the feature vector based on its principal components z_{pca} does not improve in performance too. In conclusion, feature z that fuses color and texture feature in a kernel function is adopted for the rest of the experiments in this thesis.

Next we compare the generalization performances of different SVM kernels and kernel parameters based on feature vector z. We have experimented with the polynomial and RBF kernels.

For the polynomial kernel based on Equation (3.27), we fixed the constant as 1 ($C = 100$) [Joachims, 1999] and varied the degree $d = 2, 3, \cdots 10, 20$. Table 3.4 shows the average numbers of classification errors (over 26 SSR classes) on test set D_{tst} for these SVM classifiers.

For the RBF kernel based on Equation (3.26), we varied the α parameter with fixed $C = 100$ [Joachims, 1999]. The α parameter is related to the standard devia-

Table 3.6: Training statistics for 26 SSR classes

	min.	max.	avg.
num. pos. trg.	5	26	14.4
num. sup. vec.	9	66	33.3
num. pos. test	3	13	6.9
num. errors	0	14	5.7
error (%)	0	7.8	3.2

tion σ of the RBF function as follows,

$$\alpha = \frac{1}{2\sigma^2} \tag{3.28}$$

Table 3.5 shows the average numbers of classification errors (over 26 SSR classes) on test set D_{tst} for different α and associated σ values.

From Table 3.4, we see that polynomial kernel of degree 6 (denoted as $Poly_6$) has optimal generalization performance. Similarly, the best generalization result was obtained by the RBF kernel with $\alpha = 1$ ($\sigma = 0.71$) (denoted as RBF_1) as shown in Table 3.5. Hence these two kernels plus polynomial kernel of degree 2 (most efficient in computation) (denoted as $Poly_2$) are adopted for the indexing and retrieval experiments in the thesis.

Lastly in this subsection, we report the training statistics of the SVM classifiers. As it turned out that the polynomial kernel of degree 2, $Poly_2$, has outperformed the other two kernels ($Poly_6$, RBF_1) in our experiments (Chapter 7), we shall report the training statistics for the $Poly_2$ SVM classifiers. Table 3.6 summarizes training statistics for $Poly_2$ in terms of the minimum, maximum, and average numbers (columns) of positive training examples, support vectors, positive test examples, misclassifications, error rates (rows).

Table 3.7 lists the SVM training details for each SSR class. From left to right, the columns list SSR class labels, the numbers of positive training examples from a total of 375 (p-train), numbers of positive test examples from a total of 179 (p-test), numbers of support vectors computed (sv), and the numbers of misclassified examples on the test set (err). The negative training (test) examples for a SSR class are the union of positive training (test) examples of the other 25 classes. The minimum number of

Table 3.7: Training statistics for each SSR class

Semantic Support Regions	p-train	p-test	sv	err
People:Face	26	13	36	2
People:Figure	22	11	49	10
People:Crowd	14	7	27	2
People:Skin	14	6	24	2
Sky:Clear	7	3	9	1
Sky:Cloudy	15	8	29	11
Sky:Blue	7	3	18	2
Ground:Floor	20	9	35	13
Ground:Sand	12	5	22	4
Ground:Grass	9	4	23	6
Water:Pool	14	7	16	6
Water:Pond	11	5	35	7
Water:River	14	6	32	8
Foliage:Green	20	9	42	4
Foliage:Floral	14	7	37	4
Foliage:Branch	13	6	40	9
Mountain:Far	10	5	18	6
Mountain:Rocky	9	4	41	8
Building:Old	23	12	66	14
Building:City	24	13	64	7
Building:Far	20	9	58	7
Interior:Wall	20	10	34	7
Interior:Wooden	5	3	15	0
Interior:China	14	6	41	4
Interior:Fabric	9	4	29	3
Interior:Light	9	4	27	2

positive training and test examples are from the Interior:Wooden SSR while their maximum numbers are from the People:Face class. The mininum and maximum numbers of support vectors are associated with the Sky:Clear and Building:Old SSRs respectively. The SSR with the best generalization is the Interior:Wooden class while the worst test error belongs to the Building:Old class.

3.4 Detection

Once a vocabulary of domain-relevant SSRs has been learned in the form of binary SVMs, an image can be indexed automatically against the SSRs. Figure 3.6 depicts a three-layer visual information processing architecture for image indexing. The bottom layer denotes the pixel-feature maps computed for feature extraction. In our experiments, conceptually there are 3 color maps (i.e. YIQ channels) and 30 texture maps (i.e. Gabor coefficients of 5 scales and 6 orientations). From these maps, feature vectors z^c and z^t compatible with those adopted for SSR learning are extracted.

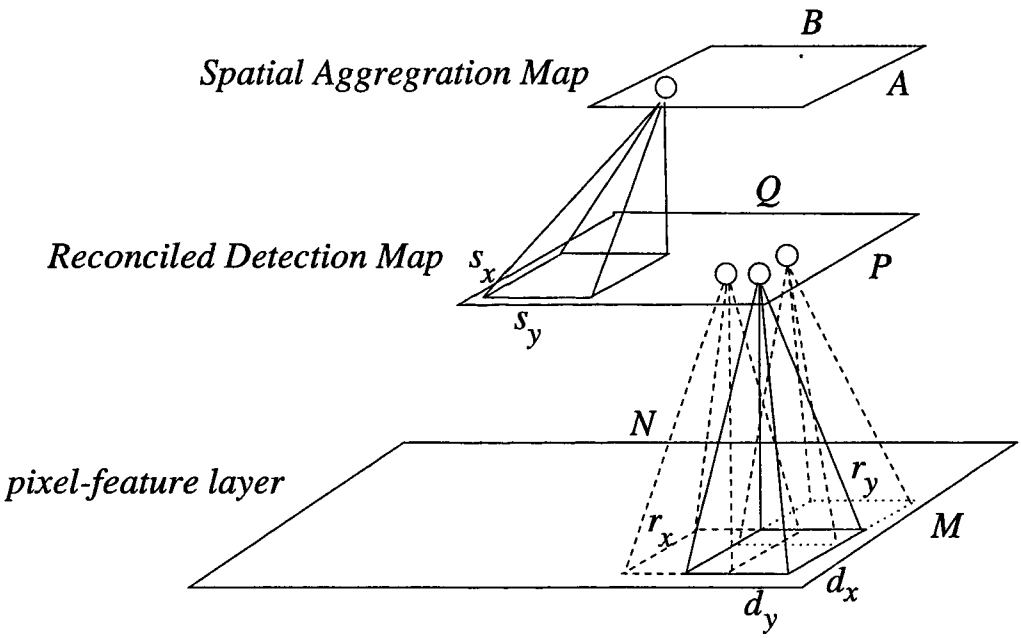

Figure 3.6: A visual information processing architecture for image indexing

To detect SSRs with translation and scale invariance in an image to be indexed, the image is scanned with windows of different scales, similar to the strategy in view-based object detection [Papageorgiou et al., 1998]. More precisely, given an image I with resolution $M \times N$, the middle layer (Figure 3.6), Reconciled Detection Map (RDM), has a lower resolution of $P \times Q, P \leq M, Q \leq N$. Each pixel (p, q) in RDM corresponds to a two-dimensional region of size $r_x \times r_y$ in I. We further allow tessellation displacements $d_x, d_y > 0$ in X, Y directions respectively such that adjacent pixels in RDM along X direction (along Y direction) have receptive fields

in I which are displaced by d_x pixels along X direction (d_y pixels along Y direction) in I. At the end of scanning an image, each pixel (p, q) that covers a region z in the pixel-feature layer will consolidate the SSR classification vector $T_i(z)$ (Equation (3.24)).

In our experiments, we progressively increase the window size $r_x \times r_y$ from 20×20 to 60×60 at a displacement (d_x, d_y) of $(10, 10)$ pixels, on a 240×360 size-normalized image. That is, after the detection step, we have 5 maps of detection of dimensions 23×35 to 19×31, which are reconciled into a common RDM to be explained below.

Using larger images may allow more accurate features for SVM learning and classification, but the computation requirement is higher. In fact, the strategy adopted in view-based object detection [Sung and Poggio, 1998] [Papageorgiou et al., 1998] is to fix the window size and resize the image smaller to achieve multi-scale detection. Hence the number of pixels available for object detection is constant. To alleviate the effect of feature extraction on small window size, we fix the image size (after size normalization) and increase the window size instead. As our features z^c and z^t are second order statistical features (i.e. mean and standard deviation), we do not see any problem with the window sizes we adopted as can be seen from the generalization performance shown in Table 3.6.

3.5 Multi-Scale Reconciliation

In the case of object detection [Sung and Poggio, 1998] [Papageorgiou et al., 1998], the system only needs to output the bounding box of an object detected at any location at any image scale attempted. In our case of image indexing, we seek a common representation of multiple SSRs detected from various image scales attempted. Hence we need to devise a new way to fuse multi-scale SSR detection outcomes.

To reconcile the detection maps across different resolutions onto a common basis, we adopt the following principle: If the most confident classification of a region at resolution r is less than that of a larger region (at resolution $r + 1$) that subsumes the region, then the classification output of the region should be replaced by those of the larger region at resolution $r + 1$. For instance, if the detection of a face is

more confident than that of a building at the nose region (assuming nose is not in the SSR vocabulary), then the entire region covered by the face, which subsumes the nose region, should be labeled as face.

To illustrate the point, suppose a region at resolution r is covered by 4 larger regions at resolution $r+1$ as shown in Figure 3.7. Let $\rho = max_k max_i T_i(z_k^{r+1})$ where k refers to one of the 4 larger regions in the case of the example shown in Figure 3.7. Then the principle of reconciliation says that if $max_i T_i(z^r) < \rho$, the classification vector $T_i(z^r)$ $\forall i$ should be replaced by the classification vector $T_i(z_m^{r+1})$ $\forall i$ where $max_i T_i(z_m^{r+1}) = \rho$.

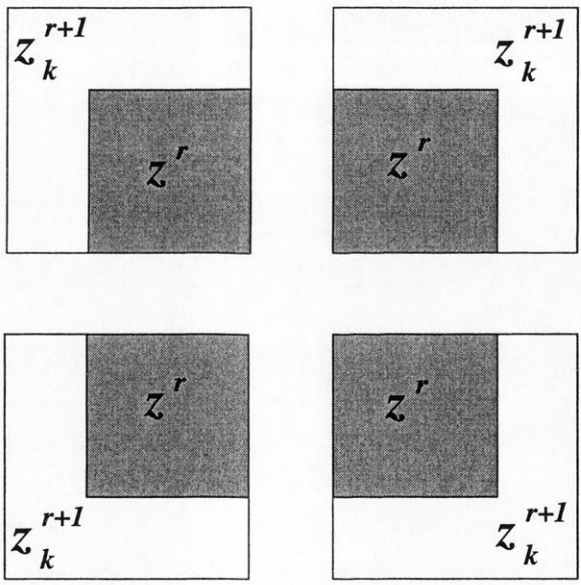

Figure 3.7: Reconciling multi-scale SSR detection maps

Using this principle, we compare detection maps of two consecutive resolutions at a time, in descending window sizes (i.e. from windows of 60×60 and 50×50 to windows of 30×30 and 20×20). After 4 cycles of reconciliation, the detection map that is based on the smallest scan window (20×20) would have consolidated the detection decisions obtained at other resolutions as the RDM (Figure 3.6) for further spatial aggregation.

3.6 Spatial Aggregation

The purpose of spatial aggregation is to summarize the reconciled detection outcome in a larger spatial region. Suppose a region Z comprises of n small equal regions with feature vectors $z_1, z_2, \cdots, z_n$ respectively. To account for the relative proportion of detected SSRs in the spatial area Z, the SSR detection vectors of the RDM is aggregated as

$$T_i(Z) = \frac{1}{n} \sum_k T_i(z_k). \tag{3.29}$$

If one interpretes $T_i(z_k)$ as the posterior probability of SSR class S_i given region z_k, $P(S_i|z_k)$, and assumes that the scan windows are non-overlapping, the posterior probability of SSR class S_i of a larger region Z, that comprises of n small equal regions with feature vectors $z_1, z_2, \cdots, z_n$ respectively, could be computed as

$$P(S_i|Z) = 1 - \prod_k (1 - P(S_i|z_k)) \tag{3.30}$$

However, this probability has lower discrimination power as it focuses on the existential aspect of S_i in Z and hence it fails to capture the quantitative occurrence or spatial extent of S_i. For instance, a single small face in an image region will result in almost identical posterior probability value as a large face or many small faces in aother image region assuming that the faces are detected reliably. This undesirable phenomenon has been confirmed in our experiments. Hence it is more appropriate to replace $P(S_i|Z)$ by the expected value of $P(S_i|z_k)$ over z_k that takes into account the occurrences and sizes of S_i in a region Z i.e.

$$\hat{P}(S_i|Z) = \sum_k P(S_i|z_k)P(z_k) = \frac{1}{n} \sum_k P(S_i|z_k) \tag{3.31}$$

The last step is possible since we can regard $P(z_k)$ as equal for all small regions z_k of equal areas. This outcome is equivalent to that of Equation (3.29).

The spatial aggregation process is illustrated in Figure 3.6 where a Spatial Aggregation Map (SAM) further tessellates over RDM with $A \times B, A \leq P, B \leq Q$ pixels. This form of spatial aggregation does not encode spatial relation explicity. But the design flexibility of s_x, s_y allows us to specify the location and extent in the

content to be focused and indexed. We can choose to ignore unimportant areas (e.g. margins) and emphasize certain areas with overlapping tessellation. We can even have different weights attached to the areas during similarity matching (see Chapter 7).

The SAM has similar representation scheme as local color histograms, except that the bins refer to proportions of SSRs instead of proportions of colors. They are invariant to translation and rotation about the viewing axis and change only slowly under change of angle of view, change of scale, and occlusion [Swain and Ballard, 1991]. The effect of averaging in Equation (3.29) will not dilute $T_i(Z)$ into a flat histogram. As an illustration, we show the $T_i(Z) \geq 0.1$ of SSRs for 3 sample image indexes.

Figure 3.8: A sample image of park to illustrate SSR-based image index

Table 3.8: Key SSRs in the index for the image shown in Figure 3.8

Image Block	Key SSR Aggregated	$T_i(Z)$
top	`Foliage:Green`	0.78
top	`Foliage:Branch`	0.11
center	`People:Crowd`	0.52
center	`Foliage:Green`	0.20
right	`People:Crowd`	0.36
right	`Building:Old`	0.32

Tables 3.8, 3.9, and 3.10 list the dominant SSRs detected, reconciled, and aggregated in 3 tessellated blocks (outlined in red bounding boxes) in Figures 3.8, 3.9, and 3.10 respectively.

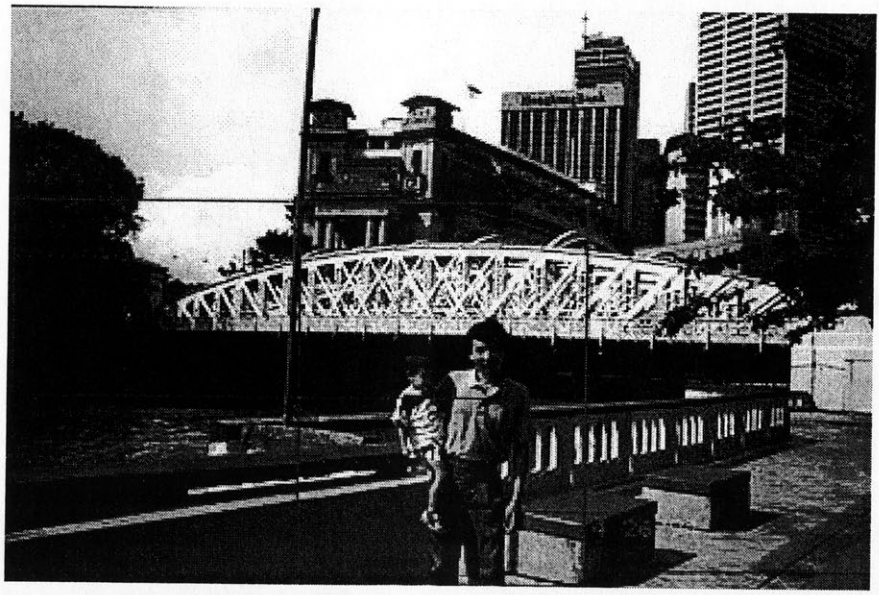

Figure 3.9: A sample image of street scene to illustrate SSR-based image index

Table 3.9: Key SSRs in the index for the image shown in Figure 3.9

Image Block	Key SSR Aggregated	$T_i(Z)$
left	`Building:City`	0.30
left	`Foliage:Green`	0.16
left	`Interior:Wall`	0.14
left	`Building:Old`	0.13
center	`Building:City`	0.75
bottom	`Building:Old`	0.29
bottom	`Building:City`	0.23
bottom	`Ground:Floor`	0.17
bottom	`People:Figure`	0.16

For Figure 3.8, the key SSRs listed in Table 3.8 capture the dominant semantic meanings of the tessellated blocks. In the case of Figure 3.9, some noise has been introduced into its index due to detection error. For example, the bright sky and the dark shadow areas in the left image block were probably detected wrongly as `Interior:Wall` (0.14) and `Building:Old` (0.13) respectively. Conversely the two

faces in the center image block have been missed during detection. There are also detection errors for Figure 3.10. In particular, the SSRs `Sky:Cloudy`, `Sky:Blue`, and `Foliage:Floral` were mistaken for the wall and shorts appeared in the left image block. Similarly, part of the dress in the center image block was detected as `Foliage:Floral` and the sofa in the right image block has more resemblance to `Interior:Wall` and some resemblance to `Sky:Cloudy` and `Ground:Floor`.

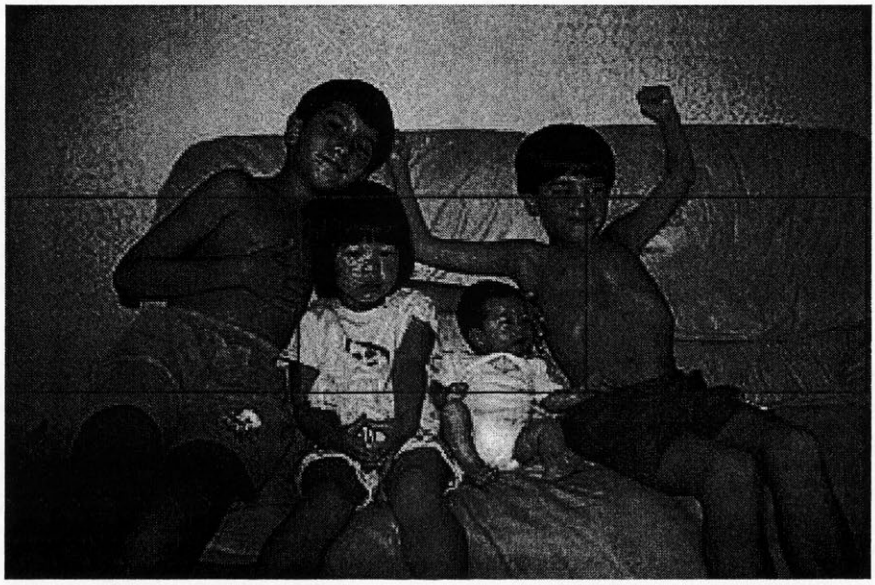

Figure 3.10: A sample image of indoor to illustrate SSR-based image index

Table 3.10: Key SSRs in the index for the image shown in Figure 3.10

Image Block	Key SSR Aggregated	$T_i(Z)$
left	`People:Skin`	0.22
left	`Sky:Cloudy`	0.18
left	`Sky:Blue`	0.17
left	`Foliage:Floral`	0.17
left	`Interior:Wall`	0.11
center	`People:Face`	0.47
center	`Foliage:Floral`	0.17
right	`Interior:Wall`	0.42
right	`Sky:Cloudy`	0.20
right	`People:Skin`	0.18
right	`Ground:Floor`	0.10

One may wonder how the SSR detection errors affect the similarity matching

of images. Certainly, accurate SSR detection is desirable. However since we are dealing with heterogenous images, robust detection for all SSR classes is in general not possible. If conventional segmentation-then-recognition framework that records only the most probable object label detected in a segmented region in image index is adopted, the errors accumulated in both the segmentation and recognition stages could result in high mismatch between the indexes of two images of similar semantics but of different visual appearances.

The proposed SSR approach minimizes the mismatch errors with segmentation-free multi-scale detection and reconciliation as well as with the preservation of soft detection result during spatial aggregation. That is, entries in a SSR-based image histogram have better chances of matching similar values in corresponding entries of the index of image of similar visual content.

Figure 3.11: A sample image of street scene to illustrate SSR-based image index

Table 3.11: Key SSRs in the index for the image shown in Figure 3.11

Image Block	Key SSR Aggregated	$T_i(Z)$
center	`Building:City`	0.34
center	`People:Crowd`	0.19
center	`Building:Old`	0.17

The detection errors as described above also raise the following interesting question: what is the value of $T_i(Z)$ if Z contains visual entities that are not part of the pre-defined SSR vocabulary? The answer is that the $T_i(Z)$ value for an unknown object appearing in Z will be spread across SSR classes that are visually similar to the object. That is, an unknown object will be represented as a distribution of detection values (i.e. detection vector) of visually similar SSR classes. Hence two visually similar instances of an unknown object class will have similar detection vectors for good similarity matching.

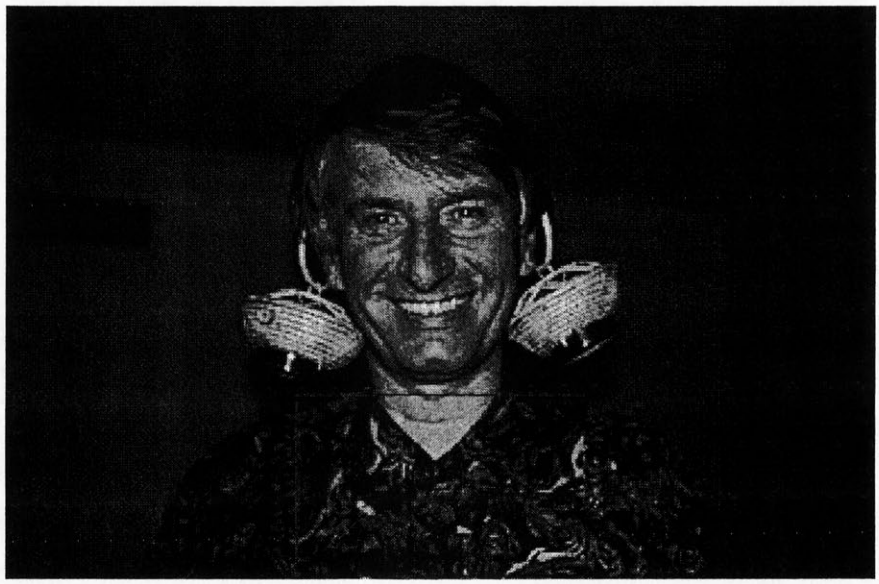

Figure 3.12: A sample image of indoor to illustrate SSR-based image index

Table 3.12: Key SSRs in the index for the image shown in Figure 3.12

Image Block	Key SSR Aggregated	$T_i(Z)$
bottom	`Foliage:Floral`	0.46
bottom	`People:Crowd`	0.34

The two image blocks (i.e. red bounding boxes) in Figures 3.11 and 3.12 refer to vehicles (part of a bus and a trishaw) with several people and a colorful shirt respectively. Tables 3.11 and 3.12 list the SSR-based interpretation for these image blocks respectively.

For Figure 3.11, bus and trishaw are not part of the SSR vocabulary, the most

similar man-made artifact in the SSR vocabulary detected is `Building:City` followed by `Building:Old`. The three people on the trishaw are detected as SSR `People:Crowd`. In the case of Figure 3.12, the closest match for the colourful shirt during SSR detection is `Foliage:Floral` followed by `People:Crowd`.

3.7 Abstraction Hierarchy

As described in Section 3.1, the SSRs can be structured into an abstraction hierarchy. In particular, two types of abstraction hierarchy are useful, namely, IS-A hierarchy, and Part-Whole hierarchy. For ease of comprehension by users in applications, we feel that two levels of hierarchy are usually adequate and useful.

For the consumer image collection used in our experiments, a simple two-level IS-A hierarchy has been designed and implemented as shown in Table 3.2. The learning and detection of SSR classes are based on the 26 more specific SSR classes such as `People:Face`, `Sky:Clear`, and `Building:City` etc and the detection value D_k of a more general concept C_k (e.g. `People`, `Sky`, `Building`) within an image region Z can be derived from the detection values $T_i(Z)$ of those SSR classes S_i that are subclasses of C_k as

$$D_k(Z) = \max_i T_i(Z), \tag{3.32}$$

since the subclasses S_i under C_k are assumed to be disjoint.

On the other hand, a complex visual object can be represented in terms of its parts, i.e. a Part-Whole hierarchy. For instance, a human figure can be represented and detected by the presence of a face and a body. Indeed interesting approaches to recognize objects by their components have been proposed and applied to people detection based on adaptive combination of classifiers (e.g. [Mohan et al., 2001]). This approach is especially useful when a 3D object has no consistent shape representation in a 2D image. The detection of multiple parts of a complex object can help to enhance the detection accuracy although not every part of an object is good candidate for detection (e.g. besides the wheels, the other parts of a car may not possess consistent color, texture, or shape feature for reliable detection).

Similar to the detection in IS-A hierarchy, the detection value D_k of a multi-part object C_k within an image region Z can be inferred from the detection values $T_i(Z)$

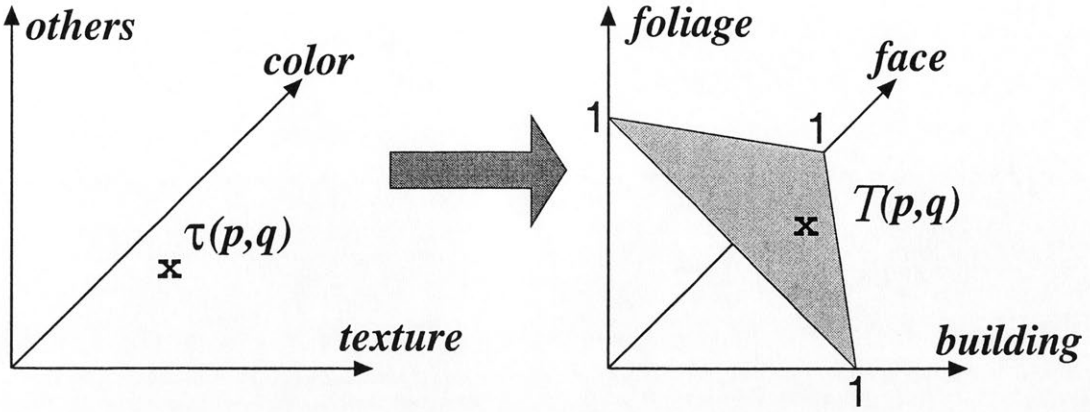

Figure 3.13: Transforming from primitive feature space to semantic feature space

of those SSR classes S_i that correspond to the parts of C_k as

$$C_k(Z) = \sum_i T_i(Z), \tag{3.33}$$

since the parts S_i of C_k can co-occur and they occupy spatial areas.

Note that in order to ensure that $D_k(Z)$ $\forall k$ within an image region Z sum up to unity, the $D_k(Z)$ is normalized by dividing each of them with their sum. As we are dealing with heterogeneous consumer images in which many objects (e.g. sky, foliage, buildings) do not have well-defined Part-Whole structure, we have not designed any Part-Whole hierarchy for the image collection experimented in the thesis.

From the perspective of pattern recognition, SSRs, which are detected against tessellated image regions based on color and texture features, span a new semantic feature space in which spatial aggregation is computed. Each SSR S_i denotes a dimension in this new feature space with feature value $T_i(z)$ in $[0, 1]$ to represent its presence in a scan window z. For any scan window z in the image, $T_i(z)$ $\forall i$ can be viewed as a feature vector whose feature values sum to unity. Geometrically, $T_i(z)$ is a point within the constrained hyperplane (i.e. $\sum_i T_i(z) = 1$) as shown schematically in Figure 3.13 ($\tau(p, q)$ is a feature vector and (p, q) denotes the $x - y$ coordinate of region z).

Collectively, an aggregate measure such as SSR histogram $T_i(Z)$ described above is computed over a spatial tessellation to represent the distribution of SSRs in the

image region Z. This semantics-rich description is clearly beyond a simple feature-based content representation (e.g. color histogram). The construction of a visual vocabulary corresponds to feature selection and the view-based object detection with spatial aggregation are indeed feature extraction to arrive at a *content-based* image representation for similarity matching.

3.8 Incremental Learning

To allow design of new visual semantics and addition of new training samples to refine existing visual semantics, incremental and rapid learning without revisiting all the training samples, similar to those interactive image indexing systems designed to capture user preferences [Minka and Picard, 1997] [Chang et al., 1998a] [Jaimes, 2003], is desired. In this thesis, we have also explored the possibility of incremental learning. In particular, we have adopted a Supervised Incremental Clustering Architecture (SICA) [Lim, 1993] [Lim, 1996] to learn SSR classes from examples.

SICA is a 3-layer feedforward neural network with dynamic node creation capability (Fig.3.14). Each input node corresponds to a feature and each output node is a class. The only hidden layer, which grows prototypes from scratch, captures the regularity of input examples through learning. Each hidden node (or prototype) receives full connections from the input layer, with a weight vector representing the position of the prototype in the input space. Prototypes of the same class are joined to the output node denoting their class with weight values '1', thus giving an 'OR' (union) operation. Learning involves the modification of the weight vectors to the prototypes as well as the recruitment and initialization of new prototypes.

When an input vector z (i.e. such as the composite color and texture feature vector described above) is presented, the closest prototype m_k from among the existing prototypes, m_i, is first determined as follows

$$\Omega(z, m_k) \geq \Omega(z, m_i) \quad \forall i, \tag{3.34}$$

where $\Omega(y, z) \in [\Omega_{min}, \Omega_{max}], \Omega_{min}, \Omega_{max} \in R$, is some similarity function between vectors y and z.

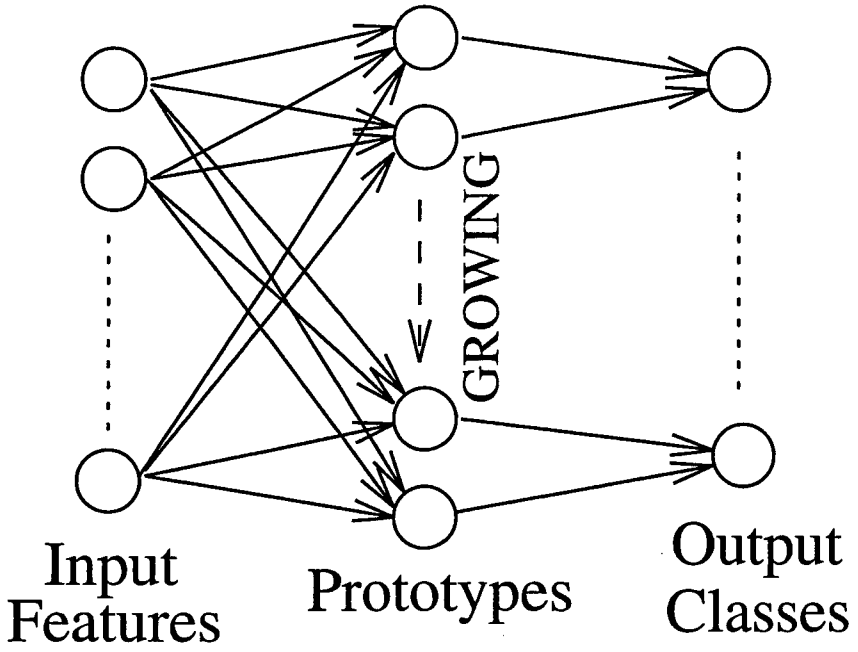

Figure 3.14: Supervised Incremental Clustering Architecture

If the following conditions are fulfilled

$$class(m_k) = class(z) \ \vee \ \Omega(z, m_k) > \alpha, \tag{3.35}$$

where $class(z)$ returns the class label of z and α is a *Prototype Creation Threshold (PCT)*, we adapt m_k towards z

$$m_k \leftarrow \frac{N_k.m_k + z}{N_k + 1}, \tag{3.36}$$

$$N_k \leftarrow N_k + 1, \tag{3.37}$$

where N_k is the number of examples that have been 'won' by (i.e. assigned to) m_k.

This update rule ensures that the prototypes are indeed the mean of all examples that have been assigned to them. In this way, similar cases are generalized to their statistical average (i.e. *local generalization*). When N_k goes to infinity, the movement of the winners will diminish asymptotically. Therefore, it implements some form of decaying learning rate automatically.

Otherwise (i.e. if Equation (3.35) is not satisfied), we have a wrong classification.

We memorize z as a new prototype

$$m_{new} \leftarrow z, \tag{3.38}$$

$$N_{new} \leftarrow 1. \tag{3.39}$$

where m_{new} is a dynamically created prototype.

When SICA is adopted for the learning of SSR classes, each SSR is represented by a number of prototypes (i.e. hidden nodes) dynamically created during learning. The similarity function $\Omega(z, m_i)$ follows that of Equation (3.27). If distance function Δ is preferred instead, then all the similarity function Ω in Equations (3.34) and (3.35) are replaced by Δ as defined by Equation (3.26) and all the $>$ comparative operators are changed to $<$.

We will compare the SICA-based learning with SVM-based learning in indexing and retrieval of consumer images in Chapter 7.

3.9 Object Segmentation

Although the SSR framework performs image indexing and retrieval based on semantics detection without region (or object) segmentation, an unconventional post-detection approach to object segmentation has been explored in this thesis.

More often than not, image or region segmentation algorithms aim to produce disjoint coherent regions based on pixel-based properties such as color or/and texture that correspond to different objects. Unfortunately the resulting segmented regions could be either over- or under-segmented, especially in complex heterogeneous images such as unconstrained consumer images. In the former case, the pixels of an object (e.g. a face) are grouped into different regions (e.g. due to shadow cast on the face). In the latter case, pixels that belonged to more than one object (e.g. face and the wooden furniture in the background) are considered a single region. Indeed object segmentation is an ill-posed and difficult problem as image segmentation without a priori knowledge of objects is underconstrained.

Though segmentation-then-recognition paradigm is dominant in computer vision systems, it is still unclear that segmentation always precedes recognition in human

vision system. Although perceptual groupings in image understanding seems plausible and logical, object recognition does facilitate object segmentation (e.g. recognizing a face does help in separating the face region from the background). It is likely that segmentation and recognition are intertwined in an iterative process.

In this thesis, we do not claim to have solved the object segmentation problem. We have only proposed a reverse detection-segmentation algorithm to extract objects based on soft detection decisions.

We propose to cluster the detection vectors $T_i(z)$ (Equation (3.24)) of the reconciled 20×20 tessellations of an image incrementally after the multi-scale reconciliation step as described in Section 3.5. That is, the clustering is carried out in the new feature space (right-hand-side of Fig. 3.13) in which the detection vectors reside. The clustering is coarse-grained as each detection vector corresponds to a 20×20 image block rather than pixels.

The steps of the clustering algorithm is as follows. The detection vectors $T_i(z)$ are examined from top-down, left to right manner. The first (top-left) detection vector starts as a new cluster center. When a current detection vector is considered close enough (i.e. distance measure such as Δ in Equation (3.26) less than some threshold or similarity measure such as Ω in Equation (3.27) more than some threshold) to its adjacent cluster center, the detection vector is absorbed (i.e. averaging similar to SICA Equations (3.36) and (3.37)) into the cluster. Otherwise a new cluster is formed with cluster center being initialized to the current detection vector (similar to the SICA Equations (3.38) and (3.39)).

After one pass through the tessellation of detection vectors, we obtain a reduced set of larger tessellated blocks (i.e. block-based regions). To further reduce the number of clusters, the adjacent clusters whose largest detection values share the same class label (i.e. likely that they share the same semantic label) are merged into a larger cluster with the new cluster center being the average of the two cluser centers. Lastly, small clusters that occupy insignificant areas and uncertain clusters with low detection values can also be removed.

Figures 3.15 and 3.16 illustrate two examples of object segmentation based on the incremental clustering algorithm. Both images are segmented with 3 dominant objects. In Figure 3.15, the key objects, sky, building, and ground, have been

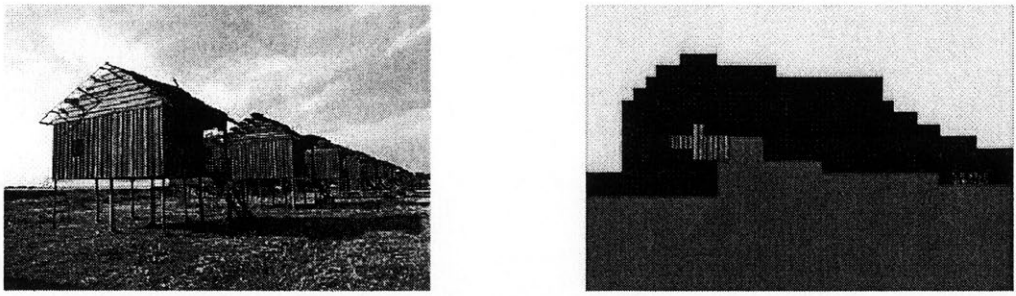

Figure 3.15: The right image shows three dominent objects segmented from the left image: sky, building, and ground

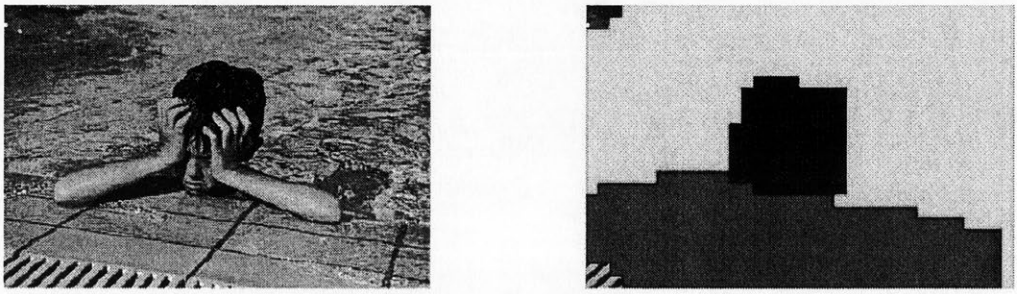

Figure 3.16: The right image shows three dominent objects segmented from the left image: water, face, and ground

properly detected and segmented. Similarly, the dominant regions that correspond to water, face, and ground, are also given the correct semantic labels and extracted.

3.10 Discussion

First, we touch on the issue of computational efficiency. The experiments of SSR learning and indexing were conducted on a Pentium 4 PC (1.4 GHz, 256 MB memory). The learning of 26 SSRs on 375 training samples was very fast (less than a minute). The indexing of one image with the SSR approach required about 20 seconds (without any code optimization). However, the small footprint of a SSR-based image index is highly efficient in storage space and retrieval.

Suppose a 4-byte floating point number is required for each $T_i(Z)$. Then a SSR-based image index requires less than 2 kilobytes ($26 \times 16 \times 4$ if a regular 4×4 grid is used for spatial tessellation) of storage and simple operations on small number of vectors. This would have great advantage over the need to represent and process

very high dimensions of color and texture features and yet not achieving the same level of retrieval performance as we shall see in Chapter 7.

In short, the image signatures based on SSRs realize semantic abstraction via prior learning and detection of visual classes when compared to direct indexing based on low-level features. As the performance comparison in Chapter 7 show, the computational resources devoted to prior learning of SSRs and to their detection during indexing are good trade-off for concise semantic representation as well as effective and efficient retrieval performance.

Nevertheless, there are also various possibilities to improve indexing efficiency. Thanks to the modular nature (binary detectors, tessellations, and multiple scales) of the SSR framework, it is straight forward to parallelize the learning, detection, and aggregation tasks. That is, we can train the binary detectors independently. During SSR detection, we compute the feature maps for the pixel-feature layer (Figure 3.6) in parallel, and feed the combined feature vector to the binary detectors which can perform classification concurrently.

Further parallelization can be achieved by performing SSR detection on different parts of an image (i.e. firing the nodes in RDM simultaneously) and along different scales. After the reconciliation process which is a sequential process, the spatial aggregation by different nodes in SAM can be carried out concurrently. In short, the indexing process as depicted by Figure 3.6 is inherently parallel.

In the current implementation, since we are using two-class SVMs that require both positive and negative examples, re-training of the SVMs is necessary when a new SSR class is added. If we replace two-class SVMs with one-class SVMs [Manevitz and Yousef, 2001] or generative models [Kumar et al., 2002], we can train only the new SSR detector based on new positive examples. The performance of one-class SVMs has been shown to be reasonable when compared to other two-class classifiers though they are rather sensitive to the choice of parameters [Manevitz and Yousef, 2001]. The potential problem with generative models has been discussed before.

In general, re-indexing is desirable when the number of SSRs (say s) has been expanded. This is applicable to other indexing methods as well when new feature dimensions are added (e.g. more bins for color histograms, new feature vector for

region segmentation or recognition). However, suppose re-training of existing detectors is not required in the case of one-class SVMs, when a new SSR class $s+1$ has been trained or a better detector becomes available to replace the detector of an existing SSR class j, an efficient re-indexing procedure can be executed as follows.

First, SSR detection is performed on all images to be indexed with the new detector ($s+1$ or j) only. The detection outcome ($T_{s+1}(z)$ or $T_j(z)$) is set to either 1 or 0 using a threshold. Next the same reconciliation step can be used to compute the RDM nodes to have either value 1 or 0. Lastly, for each SAM node with a tessellated area Z (size denoted as $|Z|$) in RDM, we count the number (i.e. area) of RDM nodes with value 1 within Z as $|X|$. The new index $T'(Z)$ that includes new SSR detector $s+1$ is computed as

$$T'_{s+1}(Z) = \frac{|X|}{|Z|}, T'_i(Z) = T_i(Z) \cdot (1 - \frac{|X|}{|Z|}) \tag{3.40}$$

and the new index $T'(Z)$ with replacement of better SSR detector j is revised as

$$T'_j(Z) = \frac{|X|}{|Z|}, T'_{i \neq j}(Z) = \frac{T_i(Z)}{\sum_{i \neq j} T_i(Z)} \cdot (1 - \frac{|X|}{|Z|}) \tag{3.41}$$

Chapter 4

Semantics Discovery

All truths are easy to understand once they are discovered; the point is to discover them.

Galileo Galilei (1564 - 1642)

4.1 Overview

Using supervised pattern classifiers to learn image semantics and ensemble of pattern classifiers to enhance system performance have become an active trend in content-based analysis research [Hsu and Chang, 2004] [Li et al., 2003a] [Snoek et al., 2004] [Tseng et al., 2003]. One of the most notable efforts by the IBM research group [Amir et al., 2003] [Tseng et al., 2003] deployed numerous SVM classifiers in multi-stage optimization for learning and detection of visual concepts in the TRECVID news video corpus. While the semantics design process and the computation involved to train and validate the SVM classifiers are certainly non-trivial, they are relatively insignificant when compared to the several months of manual annotation effort for the training, validation, and test samples by the TREC participants, with the comprehensive VideoAnnEx annotation tool [Lin et al., 2003] developed by the IBM team.

In short, supervised learning requires labeled data. Ensemble learning with multiple classifiers demands more data for feature and classifier selection. In particular,

probabilistic generative models usually require more data than discriminative models to estimate parameters reliably [Adams et al., 2003]. Hence the bottleneck for a supervised learning approach to multimedia semantic analysis is the manual effort of data labeling.

On the other hand, supervised learning of multimedia semantics is primarily design-oriented. The designers must possess knowledge about the content domain (e.g. sports, news, medical etc) in order to design the ontology and relevant features and classifiers for the domain before data annotation can take place. While this design framework is useful for many applications, there are situations (e.g. images from planet Mars, unmanned robots and vehicles in unexplored areas, unexpected behaviors in open surveillance applications) whereby limited prior knowledge is available about the multimedia data source and a complete design approach is infeasible or ineffective.

Hence an alternative semantics discovery approach is desired, for alleviating the manual annotation effort and for dealing with exploratory content domains. In this chapter, we focus on image semantics discovery. The framework proposed can be extended to other modality in future. Image indexing and retrieval task will be used for evaluation in Chapter 7.

We define the problem of image semantics discovery (ISD) (Figure 4.1) as follows. Given a number of classes of images, the task is to discover the local semantic regions (e.g. faces and foliage in bounding boxes as shown in Figure 4.1) that are recurrent within each class and discriminative against other classes. Note that the only prior knowledge we have here is the prior groupings of the image samples i.e. some form of global knowledge about the images. The emphasis here is on local image semantics discovery based on global image grouping information.

The problem of ISD is a relatively new one. However we can position ISD in the context of automatic image annotation (AIA) and review existing works related to AIA. In general, the several AIA approaches discussed here can be placed on a two-dimensional grid (Figure 4.2). The x-axis denotes the extent of the exploitation of text information associated with the images (if they are available) and the y-axis indicates the extent of content-based analysis on the images. Note that manual effort is required at some point in time to produce the associated text information though

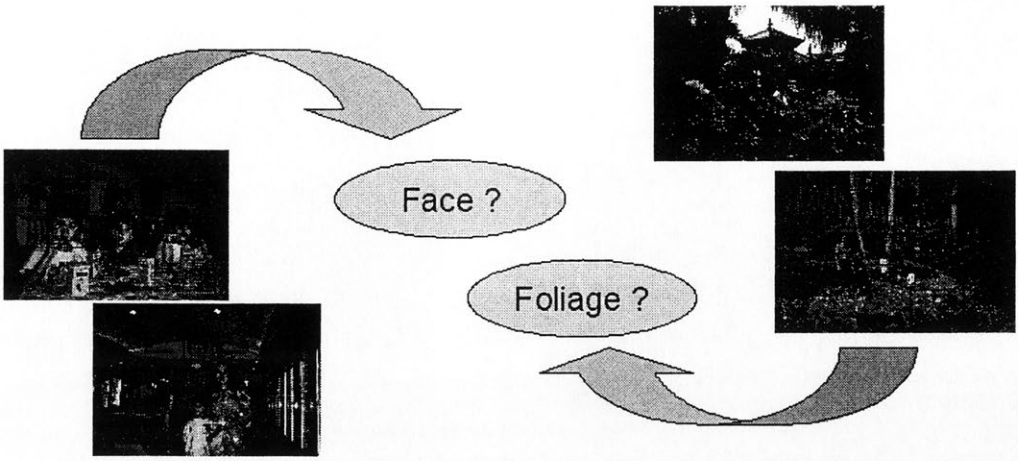

Figure 4.1: The problem of image semantics discovery

the text might be generated for other purpose and is treated as free information source to aid image annotation.

On the x-axis of Figure 4.2, the coordinate $(1, 0)$ represents AIA approaches based on the text that describes a given image (e.g. filename, URL etc) and possibly other non-content-based information (e.g. citation-based). This approach is exemplified by the Google Image Search engine on the Web (www.google.com/imghp). Since it does not analyse the image content, it is not surprised that the images returned by this approach may have content irrelevant to the intended query. For instance, a search with the keyword 'Paris' to look for images of the French capital Paris may return portrait images of people with the name 'Paris'. On 25 March 2004, the 39^{th} image returned by Google Image Search using keyword 'Paris' shows a man *Jon Paris plays "Born to Be Wild" to a crowd that understands* (www.jsonline.com/general/harley95/images/paris.jp).

In the context of relevance feedback, unlabeled images have also been used to boost the learning from very limited labeled examples (e.g. [Wang et al., 2003] [Wu et al., 2000b]). In particular, the MiAlbum system uses relevance feedback technique [Lu et al., 2000] to automatically generate text annotation for consumer photos [Liu et al., 2000]. The text keywords in a query are assigned to positive feedback examples (i.e. retrieved images that are considered relevant by the user who issues the query). This would require constant user intervention (in the form of relevance feedback) and the keywords issued in a query might not necessarily

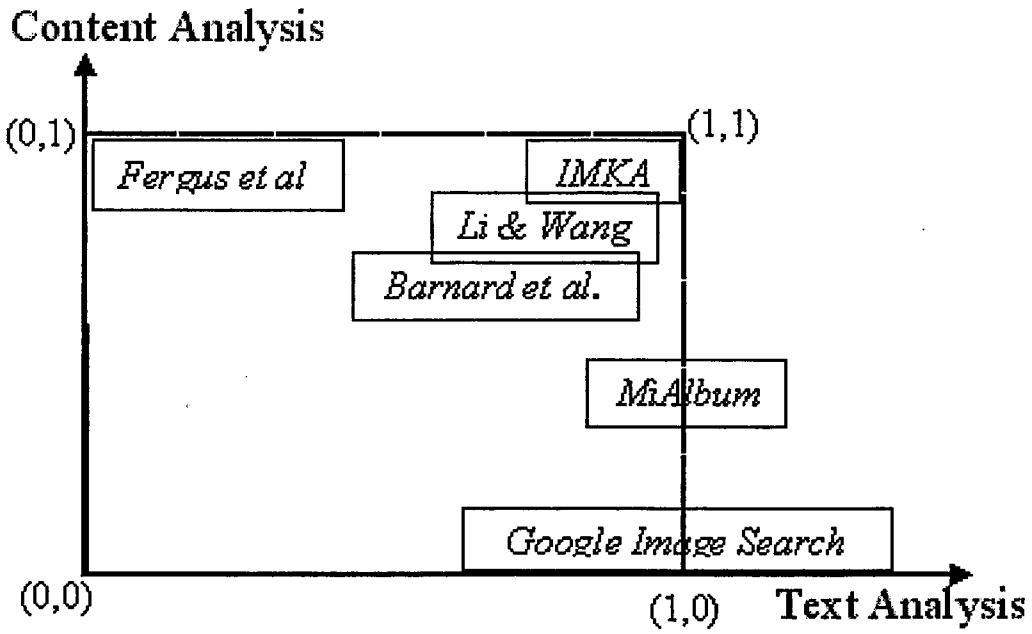

Figure 4.2: Automatic image annotation approaches

correspond to what is considered relevant in the positive examples.

Moving upwards from the x-axis, the regions towards the $(1, 1)$ coordinate in Figure 4.2 covers AIA approaches that exploit both image content and text information. Several methods have emerged in the past few years.

In the Intelligent Multimedia Knowledge Application (IMKA) project, Benitez and Chang proposes a framework for representing and discovering knowledge from multimedia content to enhance the classification, navigation and retrieval of multimedia [Benitez and Chang, 2003a]. The MediaNet knowledge representation unifies both perceptual and semantic concepts and relationships exemplified by media [Benitez et al., 2000]. Using a collection of 3624 annotated nature and news images, perceptual and semantic knowledge are automatically discovered by integrating both the processing of images and text. Perceptual knowledge is constructed by clustering the images based on both visual and text feature descriptors, and by discovering statistical and similarity relationships between the clusters. Using WordNet and the image clusters, semantic knowledge is further constructed by disambiguating the senses of words in annotations, and by finding semantic relations between the detected senses in WordNet. More recently, interdependence among discovered con-

cepts are used to construct Bayesian networks for probabilistic inferencing in image classification task with promising results [Benitez and Chang, 2003b].

Motivated from a machine translation perspective, object recognition is posed as a lexicon learning problem to translate image regions to corresponding words [Duygulu et al., 2002]. More generally, the joint distribution of meaningful text descriptions and entire or local image contents are learned from images or categories of images labeled with a few words [Barnard et al., 2003b] [Kutics et al., 2003] [Li and Wang, 2003]. The lexicon learning metaphor offers a new way of looking at object recognition [Duygulu et al., 2002] and a powerful means to annotate entire images with concepts evoked by what is visible in the image and specific words (e.g. fitness, holiday, Paris etc [Li and Wang, 2003]). While the results for the annotation problem on entire images look promising [Li and Wang, 2003], the correspondence problem of associating words with segmented image regions remains very challenging [Barnard et al., 2003b] as segmentation, feature selection, and shape representation are critical and non-trivial choices [Barnard et al., 2003a].

Without assuming the availability of associated text information (i.e. represented by the $(0, 1)$ coordinate in Figure 4.2), researchers in the field of computer vision have been pushing the limit of learning by developing object recognition systems from unlabeled and unsegmented images [Fergus et al., 2003] [Selinger and Nelson, 2001] [Weber et al., 2000]. For the purpose of image retrieval, unsupervised models based on "generic" texture-like descriptors without explicit object semantics can also be earned from images without manual extraction of objects or features [Schmid, 2001]. As a representative of the state-of-the-art, sophiscated generative and probabilistic model has been proposed to represent, learn, and detect object parts, locations, scales, and appearances from fairly cluttered scenes with promising results [Fergus et al., 2003].

In this chapter, we address the issue of minimal supervision differently. We do not assume availability of text descriptions for image or image classes as in [Barnard et al., 2003b] [Li and Wang, 2003] [Benitez and Chang, 2003a]. Neither do we know the object classes to be recognized as in [Fergus et al., 2003]. We wish to answer three basic questions related to ISD: what are the local image regions that are unique to a class of images? How do we extract this intra-class recur-

rent and inter-class discriminative image regions? How can these regions be used in image indexing and matching? The answer to these questions is a novel semi-supervised framework proposed to discover and associate local unsegmented regions with semantics and generate their samples so as to construct semantic models for content-based image retrieval, all with minimal manual intervention.

The proposed generic framework of image semantics discovery (ISD) consists of three learning steps:

- Supervised learning of class discrimination;
- Unsupervised learning of recurrent patterns;
- Supervised learning of discovered semantics regions.

In this chapter, Support Vector Machines (SVMs) [Vapnik, 1998] and Fuzzy C-Means clustering (FCM) [Bezdek, 1981] are adopted for the supervised and unsupervised learning steps respectively.

We first describe the key ideas of the ISD framework (Figure 4.3) as follows before presenting the technical details. We assume that a set of representative images, grouped into K distinct classes, of a content domain is available. Each image is tessellated into possibly overlapping small image blocks with features appropriate for the domain extracted. That is, each image class is now represented by the collective local image blocks of the images from the same class.

In the first supervised learning step, the class boundaries are computed based on the feature vectors of the tessellated blocks. Using binary SVMs in this paper, this step is performed K times, each time using samples of one of the classes as positive examples against the samples of all the other classes as negative examples. Figure 4.3 depicts an example of inter-discriminative class boundaries separating two classes of local patterns, denoted as shapes of diamond and triangle respectively. The darken diamond and triangle shapes on the boundaries represent the support vectors derived from support vector learning [Vapnik, 1998].

While the support vectors are important parameters in the classification decision function for discrimination [Vapnik, 1998], they may not refer to local visual patterns unique to a class of images. Conversely, input patterns that result in high SVM classification outputs, denoted by diamond shapes further away from the class

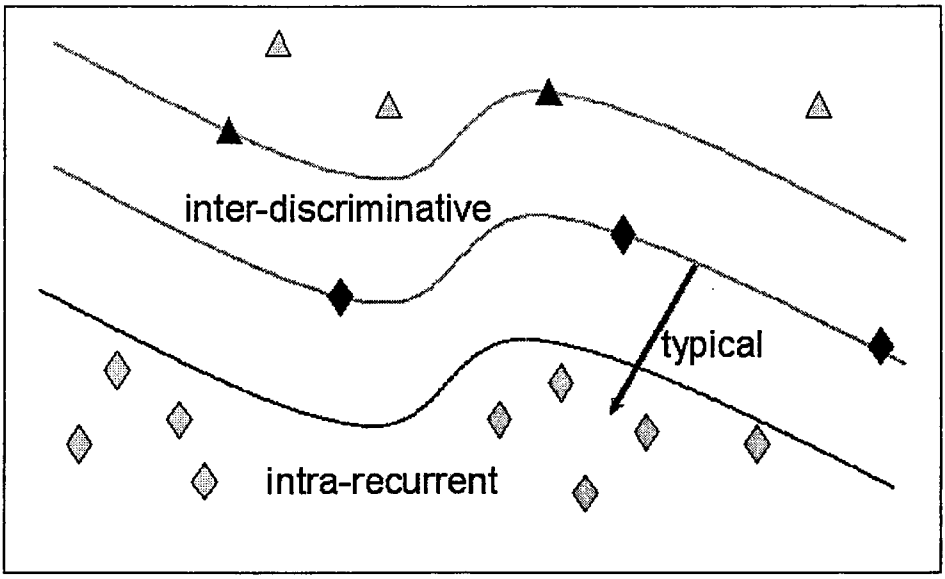

Figure 4.3: Discovering typical local patterns

decision boundary, may refer to local visual patterns that are *typical* in that image class, hence capturing intra-class recurrent patterns.

The second learning step in the ISD framework identifies these typical training patterns in each class by examing the SVM output for each training pattern. Unsupervised learning algorithm such as FCM is applied to these identified typical patterns in each of the K classes in turn to discover their multi-mode groupings, shown using different colors for two groups of diamond shapes in Figure 4.3. The clusters of local patterns are called *Discovered Semantic Regions* (DSRs).

The last step of the ISD framework is to generate the positive and negative training samples from the clusters formed in the previous unsupervised step for the modeling of DSRs. In this chapter, we also adopt binary SVM classifiers to learn the DSRs. That is, using Figure 4.3 as an illustration, the task is to discriminate the diamond shapes of the same color from the diamond shapes of different colors and the triangle shapes. The local patterns that are nearest to the respective cluster centers can be computed and their visual appearances in the original images can be extracted as a means to visualize the DSRs.

The flow of learning in the proposed ISD framework is summarized in Figure 4.4. We now describe the steps in more details.

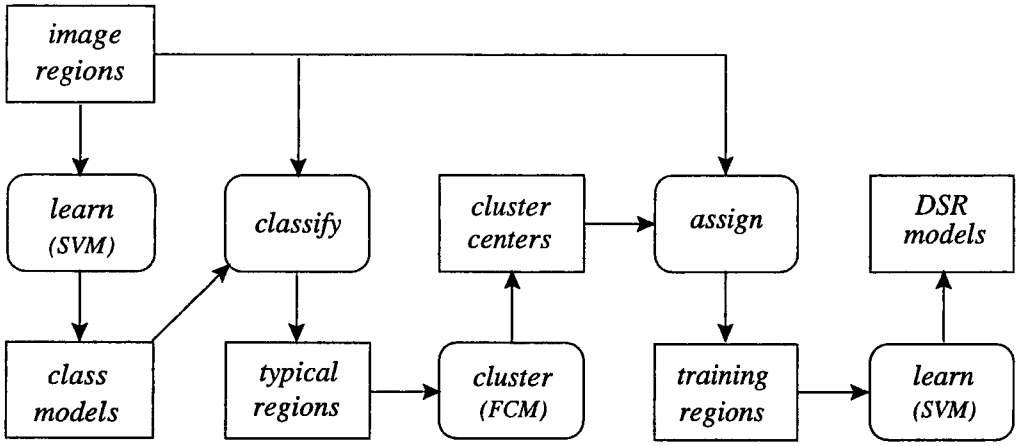

Figure 4.4: Flow of image semantics discovery

4.2 Learning of Local Class Semantics

Given an application domain, some typical classes C_k with their image samples are identified. For consumer images used in our experiments, a taxonomy as shown in Figure 4.5 has been designed. This hierarchy of 11 categories is more comprehensive than the 8 categories addressed in [Vailaya et al., 2001]. We trained 7 binary SVMs on the following categories (leaf nodes of Figure 4.5 except `miscellaneous`): interior or objects indoor (`inob`), people indoor (`inpp`), mountain and rocky area (`mtrk`), parks or gardens (`park`), swimming pool (`pool`), street scene (`strt`), and waterside (`wtsd`).

The training samples are tessellated image blocks, each represented as suitable feature vector z, from the class samples. After learning, the class models would have captured the local class semantics and a high SVM output (i.e. $C_k(z) \gg 0$) would suggest that the local region z is typical to the semantics of class C_k.

In this chapter, as our test data are heterogeneous consumer images, we extract color and texture features for a local image block and denote this feature vector as z. Hence a feature vector z has two parts, namely, a color feature vector z^c and a texture feature vector z^t. For the color feature, as the image patch for training and detection is relatively small, the mean and standard deviation of each color channel in the YIQ color model is deemed sufficient (i.e. z^c has 6 dimensions).

For the texture feature, we adopted the Gabor coefficients which have been shown to provide excellent pattern retrieval results [Manjunath and Ma, 1996]. Similarly,

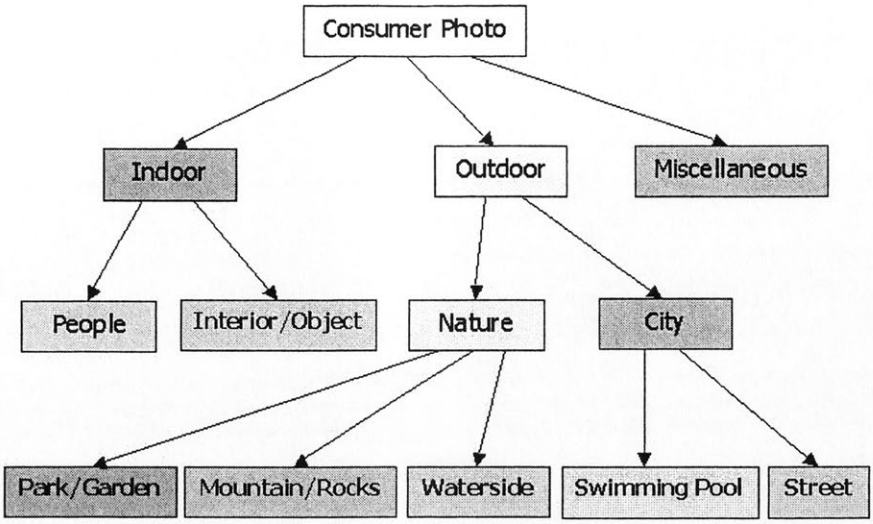

Figure 4.5: Proposed consumer image taxonomy

the mean and standard deviation of the Gabor coefficients (5 scales and 6 orientations) in an image block are computed as z^t which has 60 dimensions. To normalize both the color and texture features, we use the Gaussian (i.e. zero-mean) normalization. This composite feature vector and the normalization step are the same as those described in Chapter 3 for learning and detection of SSRs.

As mentioned in Chapter 3 before, the distance or similarity measure depends on the kernel adopted for the SVMs. For the experimental results reported in this chapter and later in Chapter 7, we have adopted polynomial kernels with the modified dot product similarity measure defined as Equation (3.27) in Chapter 3 and shown below for convenience of reader. That is, the similarity function between two feature vectors y and z for the polynomial kernels is computed as

$$\Omega(y, z) = \frac{1}{2}(\Omega(y^c, z^c) + \Omega(y^t, z^t)) \tag{4.1}$$

where $\Omega(v, w)$ is defined as

$$\Omega(v, w) = \frac{v \cdot w}{|v||w|}, \tag{4.2}$$

where $\cdot$ indicates a dot product.

4.3 Learning of Typical Semantic Partitions

With the help of the learned class models C_k, we can generate sets of local image regions that characterize the class semantics (which in turn captures the semantic of the content domain) $\mathcal{X}_k$ as

$$\mathcal{X}_k = \{z | C_k(z) > \rho\} \ (\rho \geq 0) \tag{4.3}$$

However, the local semantics hidden in each $\mathcal{X}_k$ is opague and possibly multi-mode. We would like to discover the multiple groupings in each class by unsupervised learning such as Gaussian mixture modeling and fuzzy c-means clustering. The result of the clustering is a collection of partitions m_{kj}, $j = 1, 2, \cdots, N_k$ in the space of local semantics for each class, where m_{kj} are usually represented as cluster centers and N_k are the numbers of partitions for each class.

4.4 Learning of Discovered Semantic Regions

Once we have obtained the typical semantic partitions for each class, we can learn the models of Discovered Semantic Regions (DSRs) S_i $i = 1, 2, \cdots, N$ where $N = \Sigma_k N_k$ (i.e. we linearize the ordering of m_{kj} as m_i). We label a local image block ($x \in \cup_k \mathcal{X}_k$) as positive example for S_i if it is closest to m_i and as negative example for S_j $j \neq i$,

$$X_i^+ = \{x | i = \arg\min_t |x - m_t|\} \tag{4.4}$$

$$X_i^- = \{x | i \neq \arg\min_t |x - m_t|\} \tag{4.5}$$

where $|.|$ is some distance measure. Now we can perform supervised learning again on X_i^+ and X_i^- using SVMs $\mathcal{S}_i(x)$ as DSR models.

To visualize a DSR S_i, we can display the image block s_i that is most typical among those assigned to cluster m_i that belonged to class C_k,

$$C_k(s_i) = \max_{x \in X_i^+} C_k(x) \tag{4.6}$$

As mentioned, we trained the 7 binary SVMs with polynomial kernels (degree

Table 4.1: Training statistics for image semantics discovery

Class	size	#trg	#SV	#data	#clus
inob	134	15	1905	1429	4
inpp	840	20	2249	936	5
mtrk	67	10	1090	1550	2
park	304	15	955	728	4
pool	52	10	1138	1357	2
strt	645	20	2424	735	5
wtsd	150	15	2454	732	4

2, $C = 100$ [Joachims, 1999]) for the leaf-node categories (except `miscellaneous`) on color and texture features (Equation (4.1)) of 60×60 image blocks (tessellated with 20 pixels in both directions) from 105 sample images. Hence each SVM C_k was trained on 16,800 image blocks z.

Table 4.1 lists the training statistics of the semantic classes C_k for bootstrapping local semantics. The columns (from left to right) list the class labels, the number of images of each class in the 2400 collection, the number of training images, the number of support vectors learned, the number of typical image blocks subject to clustering ($C_k(z) > 2$), and the number of clusters assigned. The 105 training images are shown in Figure 4.6. Their top-down, left-to-right order (and the number of images in each class) corresponds to the classes (and #trg) as listed in Table 4.1.

After training, the samples from each class is fed into classifier C_k to test their typicalities. Those samples with SVM output $C_k(z) > 2$ (Equation (4.3)) are subject to fuzzy c-means clustering. The number of clusters assigned to each class is roughly proportional to the number of training images in each class as shown in Table 4.1. Hence we have 26 DSRs in total.

To build the DSR models, we trained 26 binary SVMs with polynomial kernels (degree 2, $C = 100$ [Joachims, 1999]), each on 7467 positive and negative examples (Equation (4.4) and (4.5)) (i.e. sum of column 5 of Table 4.1).

To visualize the 26 DSRs that have been learned, we compute the most typical image block for each cluster (Equation (4.6)) and concatenate their appearances as shown in Figure 4.7 (from left to right): 4 for the `inob` class; 5 for the `inpp` class; 2 for the `mtrk` class; 4 for the `park` class; 2 for the `pool` class; 5 for the `strt` class;

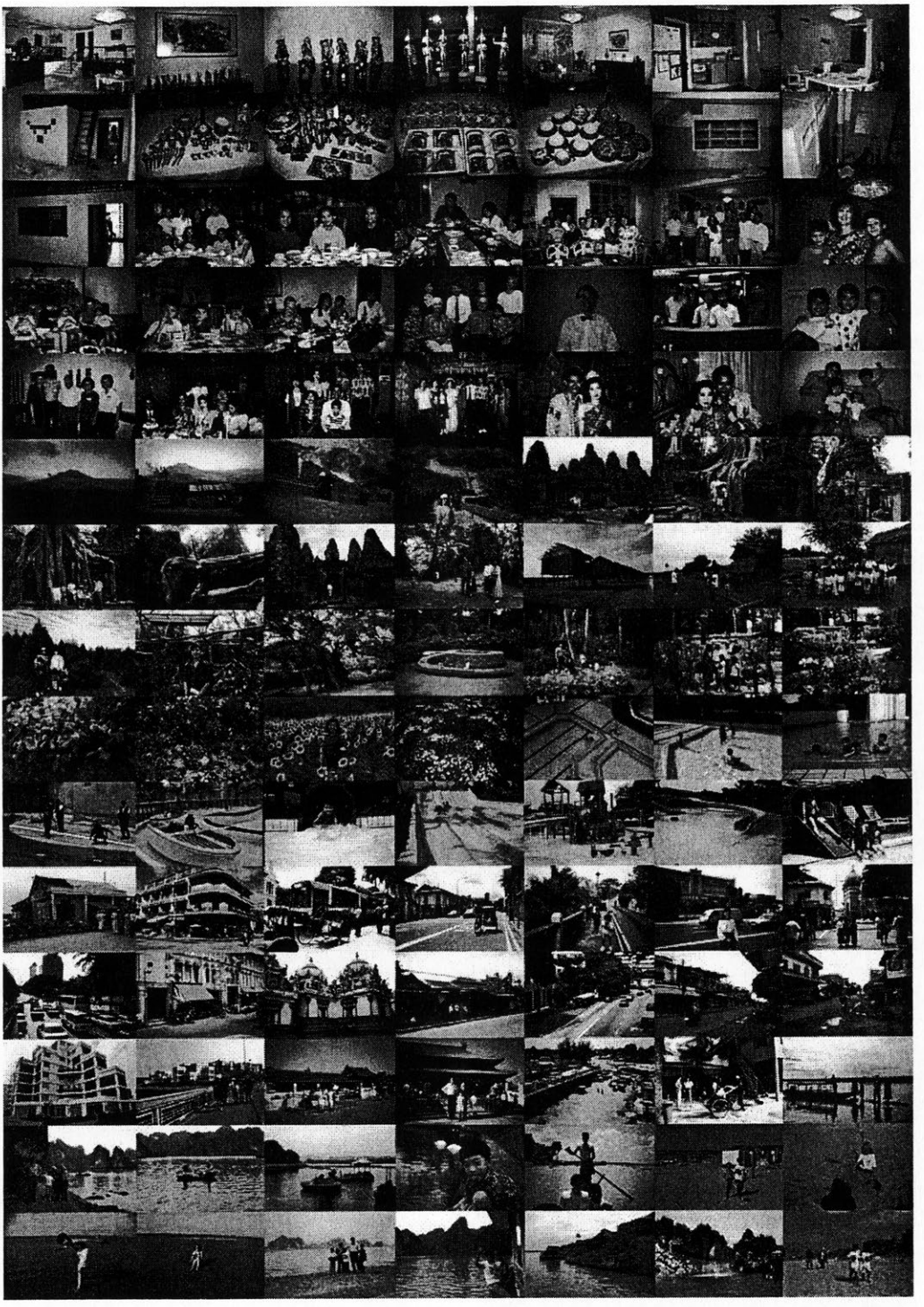

Figure 4.6: Training set of 105 images

4 for the wtsd class. Their semantic labels are list in the same order in Table 4.2.

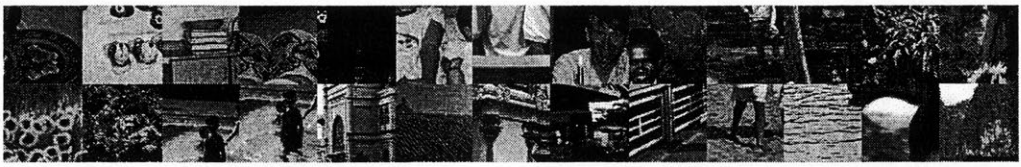

Figure 4.7: Most typical image blocks of the DSRs

Table 4.2: Semantic labels for DSRs shown in Figure 4.7

Class	Semantic Label
inob	china-1, china-2, furniture-1, china-3
inpp	body-1, body-2, body-3, face-1, face-2
mtrk	rocky-1, rocky-2
park	foliage-1, foliage-2, flower-1, foliage-3
pool	water-1, water-2
strt	man-made-1, man-made-2, man-made-3, man-made-4, man-made-5
wtsd	sand-1, river-1, pond-1, mountain-1

In Table 4.2, labels china-1/2/3 refer to the different types of china utensils in the indoor objects/interior (inob) images. The label furniture-1 is assigned to the image block that shows a cupboard top. For the indoor people class (inpp), three different appearances of body parts and faces with two kinds of background are labeled as body-1/2/3 and face-1/2 respectively. Similarly, two different rocky textures, labeled as rocky-1/2, are available for the class of mountain/rocks (mtrk). For the park class, three kinds of foliage plus a flower type are identified and labeled as foliage-1/2/3 and flower-1 accordingly. The label water-1 refers to the swimming pool side and the label water-2 denotes pool water for the (swimming) pool class of images. In the case of the street class (strt), five different appearances of man-made structures such as part of buidling, roof top etc are given the labels of man-made-1/2/3/4/5. Lastly, the DSRs from the waterside class (wtsd) can be interpreted as sand-1 (from the beach), river-1 (river water), pond-1 (pond water), and mountain (far view of mountain) respectively.

Note that the semantic interpretation for DSRs (Figure 4.7 and Table 4.2) as described above is indicative only. This is because each of the visualized image

blocks in Figure 4.7 is only a typical instance for a DSR class. For instance, the label `inpp:body-1` shows the lower body part in black. It could also refer to body part such as black hair.

4.5 Image Indexing

DSRs are local semantics patterns learned from examples. They can be employed in image indexing similar to the detection-based indexing of SSRs (c.f. Chapter 3). To recapitulate, the indexing process consists of three steps, namely view-based detection (Section 3.4), multi-scale reconciliation (Section 3.5), and spatial aggregation (Section 3.6), as summarized in Figure 4.8. That is, in these steps of indexing, the SSRs are replaced by DSRs.

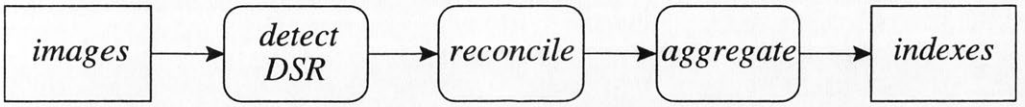

Figure 4.8: A schematic digram of image indexing based on DSRs

Similar to the illustration of the SSR-based indexes in the previous chapter, we show the $T_i(Z) \geq 0.1$ of DSRs in Tables 4.3, 4.4, and 4.5 that are detected, reconciled, and aggregated in 3 tessellated blocks (outlined in red bounding boxes) in Figures 4.9, 4.10, and 4.11 respectively.

Figure 4.9: A sample image of park to illustrate DSR-based image index

Table 4.3: Key DSRs in the index for the image shown in Figure 4.9

Image Block	Key DSR Aggregated	$T_i(Z)$
top	park:foliage-1	0.30
top	park:foliage-2	0.26
top	park:foliage-3	0.23
center	park:foliage-3	0.33
center	inpp:face-2	0.22
center	inpp:body-2	0.14
right	inpp:face-2	0.23
right	mtrk:rocky-2	0.22
right	inob:china-2	0.18

For Figure 4.9, the key DSRs listed in Table 4.3 are appropriate except that the brown hut in the right image block is detected as DSR `mtrk:rocky-2` (0.22) since there is no "building" DSR that is visually similar to hut. Also some of the white clothing in the same image block is wrongly detected as DSR `inob:china-2` (0.18).

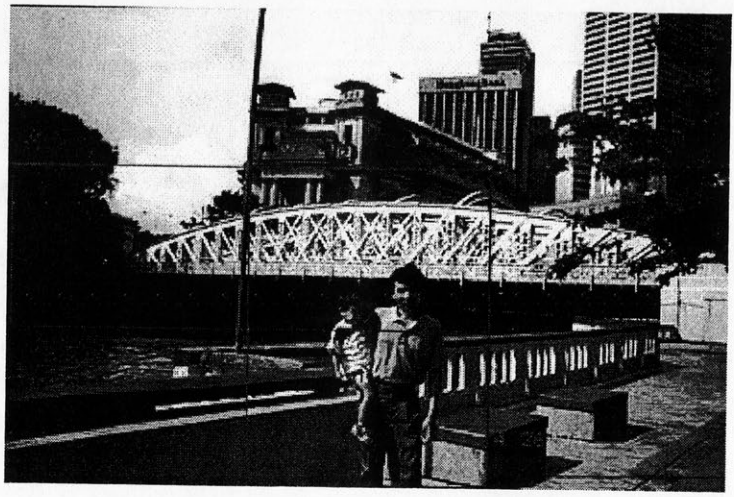

Figure 4.10: A sample image of street scene to illustrate DSR-based image index

In the case of Figure 4.10, some noise such as `mtrk:rocky-2` (0.30), as shown in Table 4.4, has been introduced into the index for the left image block.

In the image shown in Figure 4.11, the DSRs related to face, body, and furniture have been reasonably detected for the image blocks in red bounding boxes.

The effectiveness of the DSR-based image indexes will be evaluated using query by example experiments in Chapter 7.

Table 4.4: Key DSRs in the index for the image shown in Figure 4.10

Image Block	Key DSR Aggregated	$T_i(Z)$
left	`strt:man-made-5`	0.36
left	`mtrk:rocky-2`	0.30
center	`strt:man-made-5`	0.65
center	`inpp:face-1`	0.11
bottom	`inpp:body-2`	0.36
bottom	`strt:man-made-2`	0.21

Figure 4.11: A sample image of indoor to illustrate DSR-based image index

Table 4.5: Key DSRs in the index for the image shown in Figure 4.11

Image Block	Key DSR Aggregated	$T_i(Z)$
left	`inpp:body-2`	0.30
left	`inpp:body-3`	0.20
center	`inpp:body-2`	0.30
center	`inpp:face-1`	0.21
center	`inpp:body-3`	0.18
right	`inob:furniture-1`	0.45
right	`inpp:body-2`	0.29

4.6 Discussion

For the current implementation of our ISD framework, there are still several issues to be addressed.

The first issue is related to the sampling of training data. In order not to miss out interesting local region semantics from the images, the training data z for the learning of SVM classifiers $\mathcal{C}_k(z)$ (Section 4.2) should ideally be as dense and varied as possible. That is, z should cover tessellated image blocks of multiple resolutions with maximum overlaps from a comprehensive set of images.

For instance, we can improve the sampling of image blocks for semantic class learning by randomly selecting say 20% of the ground truth images in each class as positive samples (and as negative samples for all other classes) as well as by tessellating image blocks with different sizes (e.g. $20 \times 20, 30 \times 30$ etc) and displacements (e.g. 10 pixels) to generate a more complete and denser coverage of the local semantic space. But these attempts have turned out to be too ambitious for practical training sessions in our experiments. Hence as a trade-off between sampling coverage and training time, we have only used the 60×60 image blocks (tessellated with 20 pixels in both directions) from 105 sample images as reported above.

Another issue is regarding the usefulness of the discriminative class learning and typicality check in the proposed ISD framework. As an alternative, can we either perform clustering of image blocks from all training images (regardless of classes) or clustering of image blocks in each class directly and separately (i.e. without worrying about training of $\mathcal{C}_k(z)$ and selection of image blocks z such that $\mathcal{C}_k(z) > \rho$)?

We have indeed explored these alternatives. The average precisions of retrieval in the query-by-example experiments (Chapter 7) turned out to be inferior when compared to the proposed ISD framework. Hence we believe that class discrimination and typicality checks are important to constrain the clustering on relevant data points that hide the local semantic regions for discovery. Without these constraints, the unsupervised learning process would tend to converge to local optimal states.

Cluster validity is a tricky issue. We have tried fixed number of clusters (e.g. $3, 4, 5, 7$) and retained large clusters as DSRs. Alternatively we relied on human inspection to select perceptually distinctive clusters (as visualized using Equation (4.6)) as DSRs. However the current way of assigning number of clusters roughly

proportional to the number of training images has produced the best performance in our experiments. In future, we would explore other ways to model DSRs (e.g. Gaussian mixture) and to determine the value of ρ. We would also like to verify our approach on other content domains such as art images, medical images etc to see if the DSRs make sense to the domain experts.

Although our attempt to alleviate the supervised learning requirement of labeled images and regions differs from the current trends of unsupervised object recognition and matching words with pictures, the methods do share some common techniques. For instance, similar to those of Schmid [Schmid, 2001] and Fergus et al. [Fergus et al., 2003], our approach computes local region features based on tessellation instead of segmentation though [Fergus et al., 2003] used an interest detector and kept the number of features below 30 for practical implementation.

While Schmid focused on "Gabor-like" features [Schmid, 2001] and Fergus et al. worked on monochrome information only [Fergus et al., 2003], we have incorporated both color and texture information. As the clusters in [Schmid, 2001] were generated by unsupervised learning only, they may not correspond to well-perceived semantics when compared to our DSRs.

As we are dealing with cluttered and heterogeneous scenes, we did not model object parts as in the comprehensive case of [Fergus et al., 2003]. On the other hand, we handle scale invariance with multi-scale detection and reconciliation of DSRs during image indexing. Last but not least, while the generative and probabilistic approaches [Fergus et al., 2003] [Li and Wang, 2003] may enjoy modularity and scalability in learning, they do not exploit inter-class discrimination to compute features unique to classes as in our case.

To put indexing solutions based on local semantics in perspective, Figure 4.12 shows a spectrum of semantic learning and indexing approaches that we have investigated. The extreme left points towards knowledge-based approaches that require more human intervention. The opposite direction (i.e. right) represents discovery-oriented approaches that need less human involvement. In our earlier research effort, we have explored both extreme directions.

A completely unsupervised approach [Lim, 2000d] [Lim, 1999b] [Lim, 1999c] that corresponds to the label "CLUS" on the extreme left direction of Figure 4.12 has

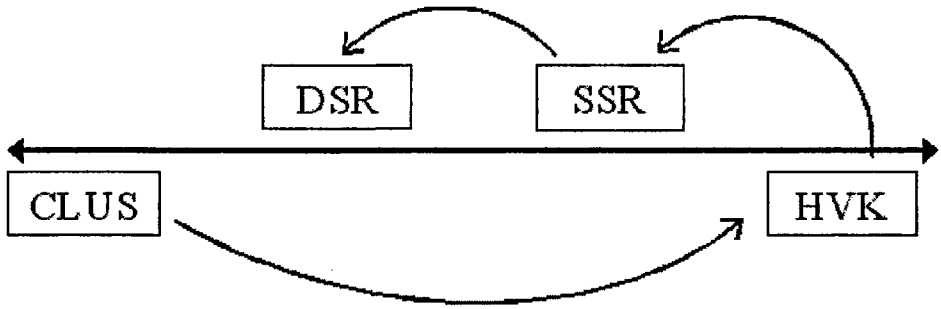

Figure 4.12: A spectrum of proposed semantic learning and indexing approaches

been explored. In this approach, tessellated image blocks from many training images are subjected to unsupervised learning such as fuzzy c-means clustering directly. As thousands of cluster centers are necessary to represent the high variations of visual semantics inherent in the images and to achieve reasonable retrieval performance, singular value decomposition (SVD) is applied to reduce the dimensionality of the index (i.e. number of cluster centers). While enjoying high degree of automation, this unsupervised approach suffers from weak interpretation of visual semantics. That is, it is not clear what visual semantics are represented by the cluster centers and their SVD-transformed counterparts.

On the extreme right in Figure 4.12, a handcrafted approach (denoted as "HVK" in the figure) [Lim, 2000b] [Lim, 2001a] that requires a human subject to design the set of Visual Keywords for an application domain and crop highly representative image blocks from the images. No statistical learning is performed and the handpicked image blocks are directly used for indexing. Rather than visual concept detection, the feature vector of an image block is matched against the feature vectors of those handpicked image blocks. The relative distances to the handpicked image blocks are used to compute fuzzy memberships as semantic histograms. This pioneering approach is not practical as it demands high precision effort from the human designer in the construction of the visual vocabulary.

Moving away from the extreme right, the image semantics design and learning approach based on SSRs, presented in Chapter 3 and denoted as "SSR" in Figure 4.12, still requires a human designer to determine the visual vocabulary for an appli-

cation domain. But the precision requirement in cropping training instances for the designed visual vocabulary is relaxed. This approach, as a natural enhancement of the handcrafted approach ("HVK"), is first published in 2002 [Lim and Jin, 2002b], though the statistical learning was based on SICA instead of SVM.

The image semantics discovery approach (labeled as "DSR" in Figure 4.12) described in this chapter represents an attempt towards the unsupervised learning paradigm. However, instead of complete hands-free automation, both supervised and unsupervised learning steps are applied to tessellated image blocks extracted from class-labeled images to infer the semantic regions. Although minimum human intervention is required as compared to the SSR approach, there is no direct control over the visual vocabulary to be discovered yet, let alone the few computational issues discussed above.

With reference to Figure 4.12, we started off with the "CLUS" approach (extreme left) and swung to the extreme right with the "HVK" approach. We then moved towards the left with less and less human annotation effort, but also less control over the visual semantics. Perhaps further innovation is necessary to achieve an optimal balance of human labeling and semantics control. This will be part of our future research endeavor.

Chapter 5

Class-Based Image Semantics

The more alternatives, the more difficult the choice.
Abbe 'D'Allanival

5.1 Overview

The previous two chapters focus on local semantics extracted from image content based on design and discovery approaches respectively. We switch attention in this chapter to focus on global semantics related to a coherent set of images, i.e. forming an image class.

However, as explained in Chapter 2, the problem of image retrieval is different from object recognition and image classification. In this chapter, we explore three different indexing and retrieval schemes based on image class semantics.

In the next section, an *Event-Based Retrieval* paradigm, especially useful in the context of consumer images, is developed. An event taxonomy for consumer images and a winner-take-all approach to compute the relevance score of an image for a query event are proposed.

In the section that follows, an *Inter-Class Indexing* scheme that exploits the relative memberships of an image to prototypical classes as image index is proposed. That is, instead of making a hard decision on which class an image belongs

to, its memberships to the image classes are normalized as an inter-class semantic histogram for similarity-based matching and retrieval.

Last but not least, in the section before the discussion section, an indexing scheme based on embeding of *Local Class Patterns* is proposed. A local image region is represented as a membership histogram of image classes. Hence an image index is a collection of histograms of class memberships, each for an image region, suitable for similarity-based matching and retrieval.

5.2 Event-Based Retrieval

The notion of Event conveys rich semantics to consumers in their collection of photos. From previous user study [Rodden and Wood, 2003] and a user survey we have conducted recently, we confirm that consumers prefer to organize and access photos along semantic axes such as Event (e.g. wedding party, fun time at swimming pool, at the park etc), People (e.g. myself, my daughter, Mary etc), Time (e.g. last Summer, this year, 2000 etc), and Place (e.g. at home, Disneyland, Paris etc). However consumers are reluctant to annotate all their photos manually as the process is too tedious and time-consuming.

As a cognitively convenient semantic unit, an Event actually encompasses other semantic axes as part of its 4 Ws:

- *who* takes part in the event (e.g. my family, John and his wife);
- *when* is the event taking place (e.g. last week, Dec. 2003);
- *where* is the event taking place (e.g. my house, at the beach); and
- *what* activity is involved (e.g. having meal, my birthday).

In this thesis, we define consumer photos (or home photos, family photos) as typical digital photos taken by average consumers to record their lives as digital memory as opposed to those taken by professional photographers for commercial purposes (e.g. stock photos like the Corel collection and many others of which previews are available at www.fotosearch.com).

At webshots (community.webshots.com), a typical website dedicated for consumers to upload and share their home photos, we notice that users apparently prefer occasions or activities as a broad descriptor for photos to other characteristics (like objects present in the photo or the location at which a photo was taken). In particular, the classification (as at the 3rd April 2003) contains many more photos under the category "Family and Friends" (more than 9 millions) than the sum of other categories (which add up to around 5 millions). Furthermore, categories such as "Scenery & Nature", "Sports", "Travel" etc are the outcome of activities.

Although Vailaya et al. [Vailaya et al., 2001] have presented a hierarchy of 8 (plus 3 "Others") categories for vacation photos, they are skewed towards scenery classification. Hence an event-based taxonomy is what consumers need. Figure 1 depicts our proposed Event Taxonomy for home photos. A typical event could be a gathering, a family activity, or a visit to some place during holidays for instance. These correspond to the purposes of meeting with someone(s), performing some activity, and going to some place respectively.

For the gathering event, it could be in the form of parties, which we include here common occasions birthday parties and wedding parties, or having meals together. The family activities event refers to activity that involves family members. We keep it simple and general by dividing it into indoor and outdoor family activities. Examples of indoor activities can include kids playing, dancing, chatting etc.

Outdoor activities may include sports, kids at playground, picnic etc. The third major type of events is known as visits to places. It could be either people-centric or not. By people-centric, we mean the family members are the focus of a photo. In the case of non-people-centric photos, the family members are not the subjects of the photos. They are either absent or not clearly visible in the photos. In this latter case, we divide it into photos of natural scenes (nature) and urban scenes (man-made). For the nature event, the photos can be taken at mountain area, along riverside or lakeside (waterside), at a beach, and in a park (also garden, field, forest etc). As for the man-made event, we include photos taken at a swimming pool, along roadside (or street) and photos of interior.

Using the visual content alone (i.e. without using information from People, Time, and Place), we would not be able to model all the 4 Ws of an Event. For example,

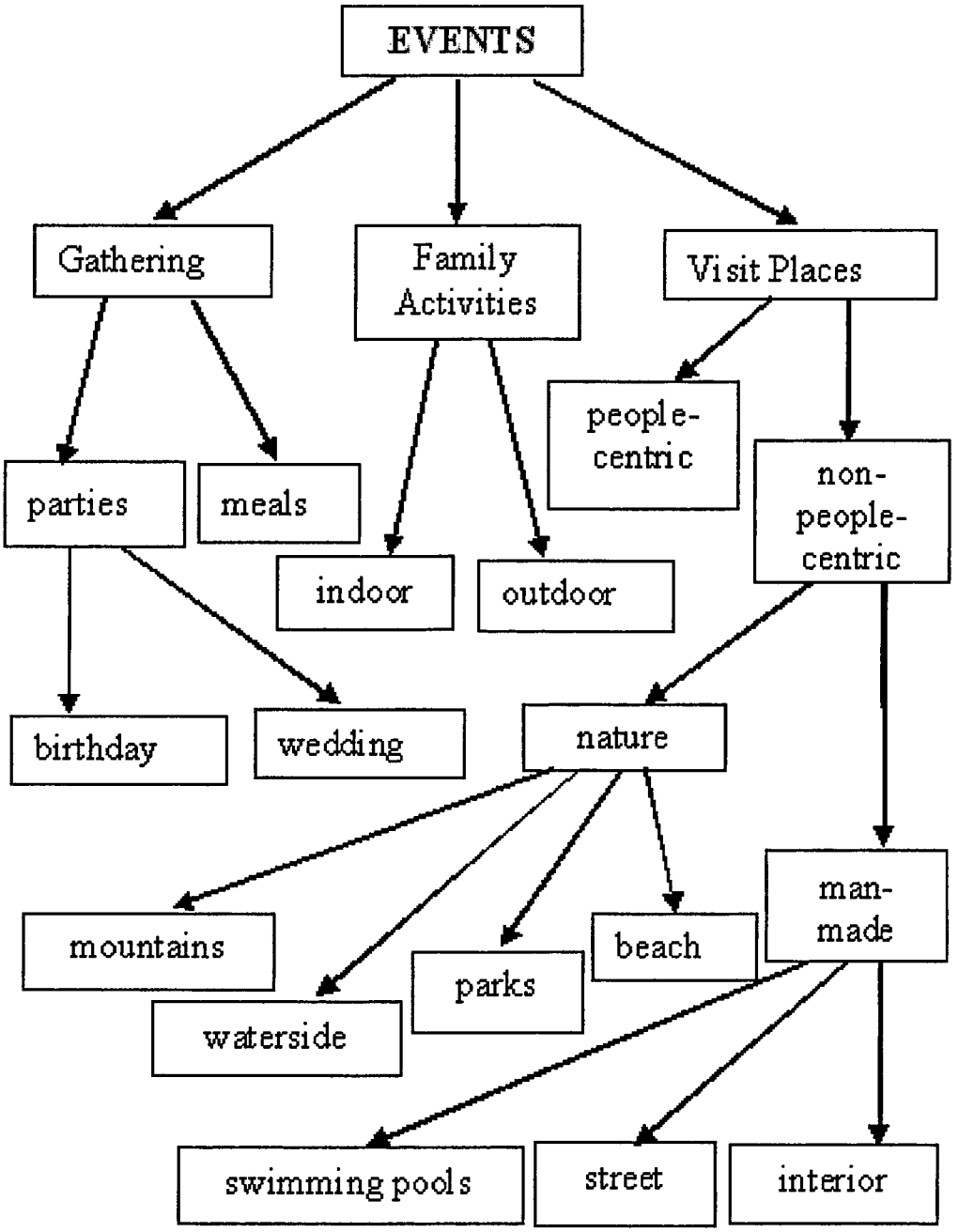

Figure 5.1: Event taxonomy for consumer photos

it is not feasible to further differentiate breakfast, lunch, and dinner for the meals event if we do not make use of the Time information. In this thesis, we approximate Event by Visual Event, defined as an event that is based on the visual content of photos (i.e. the"what" aspect).

We propose a computational learning framework to model visual semantics of consumer photo events from sample photos at 2 levels. Figure 5.2 shows a schematic diagram of this framework.

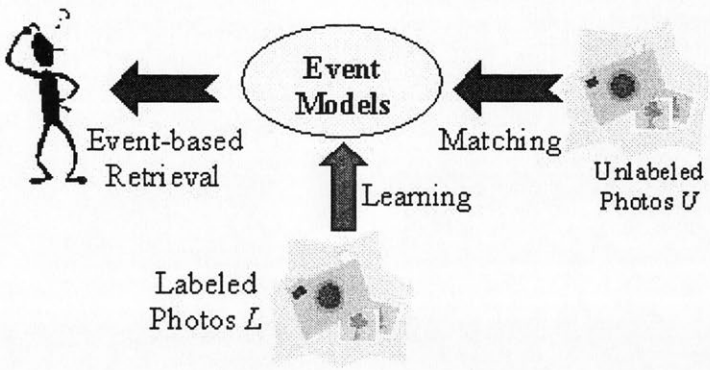

Figure 5.2: Learning event models for retrieval

At the single image level, salient image regions that exhibit semantic meanings to human users are adopted as training examples to construct semantic support regions (SSRs) that span a new indexing space. Local image regions of a photo is projected into this space as linear combinations of the semantic support regions and further aggregated spatially to form image content signature for similarity matching. The SSR framework has been described in Chapter 3.

At the image set level, we assume that for each Event E_i, there is an associated computational model M_i that allows us to compute the relevance measure $R(M_i, Z)$ of a photo Z to E_i. To minimize manual annotation effort, event models M_i are learned from a small set of labeled photos $\mathcal{L}$ and the relevance measures of unlabeled photos $\mathcal{U}$ are computed in a winner-take-all approach to the event models (Figure 5.2).

In this thesis, an Event model M_i is also learned statistically using support vector machines (SVMs), denoted as $\mathcal{M}_i$. The input patterns X to M_i are the SSR-based (or DSR-based) indexes of the images, $T_i(Z_j)$, where i indexes the SSR (or DSR) classes and j refers to the spatial regions.

The following similarity measure Ω between image index X with m local regions X_j and image index Y with m local regions Y_j is defined when SVMs with polynomial kernels are used,

$$\Omega(X,Y) = \frac{1}{m}\sum_j \frac{\sum_i T_i(X_j)T_i(Y_j)}{\sqrt{\sum_k T_k(X_j)^2}\sqrt{\sum_k T_k(Y_j)^2}}. \tag{5.1}$$

If RBF kernels are preferred for the SVM learning and classification, the following L_1-norm city block distance measure Δ can be adopted,

$$\Delta(X,Y) = \frac{1}{m}\sum_j \sum_i |T_i(X_j) - T_i(Y_j)| \tag{5.2}$$

The SVM learning will compute the support images for the events from a set of labeled photos. Given an unlabeled photo of index Z, the output of an Event model M_i, denoted as $S(M_i, Z)$, can be computed via the softmax function [Bishop, 1995] [Bridle, 1990] as

$$S(M_i, Z) = \frac{\exp^{\mathcal{M}_i(Z)}}{\sum_j \exp^{\mathcal{M}_j(Z)}}, \tag{5.3}$$

where $\mathcal{M}_i(Z)$ denotes a SVM classifier output.

In the winner-take-all approach, we compute the winner k as

$$k = argmax_i S(M_i, Z). \tag{5.4}$$

Then the relevance measure of Z to Event model M_i is defined as

$$R(M_i, Z) = \begin{cases} S(M_k, Z), & \text{if } i = k \\ 0, & \text{otherwise} \end{cases} \tag{5.5}$$

That is, the relevance measure is the outcome of "competition" among the Event models M_i on the unlabeled photo Z. The unlabeled photo is "assigned" to the winner event with relevance measure being the maximum similarity matching score. The relevance measures to all other losers are simply defined as zeroes.

5.3 Semantic Support Classes

The context of image retrieval is related to the application domain. More often than not, there exists prior semantic groupings of the images in the application domain that can be exploited for more effective and efficient retrieval and browsing. For example, for medical images, the context could be the pathological classes used for diagnostic purpose [Brodley et al., 1999] or the imaging modalities required for appropriate visualization [Mojsilovic and Gomes, 2002]. Stock image library such as the Corel image database is also organized into hierarchies of categories to facilitate access. For more general collection of images, the SIMPLIcity system [Li et al., 2000] [Wang et al., 2001] pre-classifies images into semantic categories based on segmented regions, such as textured-nontextured, objectionable-benign, or graph-photograph, to support effective browsing and to reduce the search space.

While modeling of image collection as semantic categories to facilitate browsing and filtering as well as a preprocessing step to reduce the search space in retrieval are popular, the information of semantic categories has not been used in a more direct form in similarity-based retrieval. In this thesis, besides looking at the local semantic regions of the query and database images as proposed in the previous chapters, we observe that the *categorical context* of an image with respect to prior semantic categories can also be used for indexing and matching.

In the next chapter, we shall elaborate on the role of categorical context in a probabilistic Bayesian formulation as a means to probe the relevance class for a given query and to allow contextual similarity matching. In this chapter, we focus on the indexing aspect.

Given an application domain, semantic classes $C_k, k = 1, \cdots, M$ are first identified. Then the class distribution is modeled using statistical learning. In this thesis, support vector learning is adopted for the learning of C_k. To echo the SSRs (Chapter 3) that are local semantic regions designed to support image indexing and retrieval, we refer to C_k as *Semantic Support Classes* (SSCs) that are also used for the same purpose.

As our experimental evaluation is carried out on consumer images, we have designed a taxonomy for consumer images as used in Chapter 4 before and shown here again as Figure 5.3 for ease of reference. This hierarchy of 11 categories is

more comprehensive than the 8 categories (plus 3 more "Others") addressed in [Vailaya et al., 2001]. In particular, we consider sub-categories for `indoor` and `city` as well as more common sub-categories for `nature`.

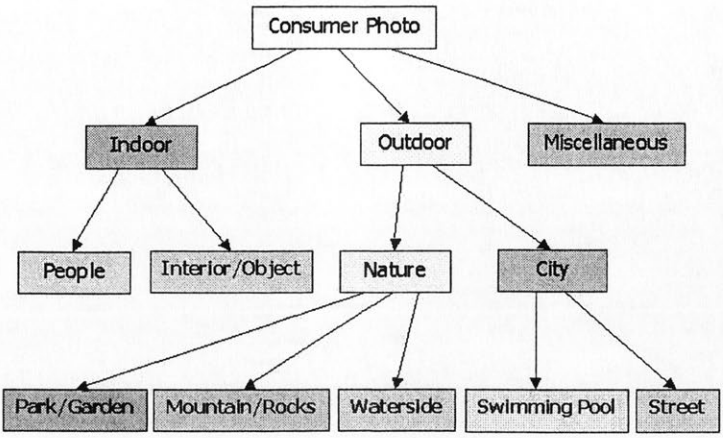

Figure 5.3: Proposed consumer image taxonomy

A support vector classifier $\mathcal{C}_k, k = 1, \cdots, 7$ is trained to differentiate each category from other categories. Using the softmax function [Bishop, 1995] [Bridle, 1990], the output of classification C_k given an image x is computed as,

$$R_k(x) = \frac{\exp^{\mathcal{C}_k(x)}}{\sum_j \exp^{\mathcal{C}_j(x)}}, \tag{5.6}$$

and is used as an inter-class index for the image to capture the categorical context.

The use of relative memberships to classes or clusters as some form of image representation for object detection and recognition (but not for the purpose of content-based retrieval) has been reported. For example, for the purpose of view-based face detection, K.K. Sung has constructed 6 face clusters and 6 non-face clusters and used the distance between the feature vector of a local image block and these clusters as the input to the trained face detector rather than using the feature vector directly [Sung and Poggio, 1998].

As we shall see in the following chapters, SSRs and SSCs are complementary image indexes that can be combined in matching and ranking of images for similarity-based retrieval. The integrated matching can be derived from a principled Bayesian formulation (Chapter 6) and its resulting average precisions in the query-by-example experiments outperforms individual SSR-based and SSC-based indexes as well as a

feature fusion approach based on color and texture features (Chapter 7).

Similar to event modeling and retrieval described in the previous section, the detection-based index using SSRs (or DSRs) is viewed as a feature vector for image classification. That is, $T_i(Z_j)\ \forall i,j$ as described before is treated as an input vector for SVMs based on the following similarity measure Ω between image index x with m local regions X_j and image index y with m local regions Y_j for polynomial kernels,

$$\Omega(x,y) = \frac{1}{m}\sum_j \frac{\sum_i T_i(X_j)T_i(Y_j)}{\sqrt{\sum_k T_k(X_j)^2}\sqrt{\sum_k T_k(Y_j)^2}}. \tag{5.7}$$

Note that this similarity measure is also used as the matching function in similarity-based retrieval using SSR-based indexes (to be described in Chapter 7). The following L_1-norm city block distance measure Δ has also been considered, for the case of RBF kernels (or as an alternative matching function for similarity-based retrieval),

$$\Delta(x,y) = \frac{1}{m}\sum_j \sum_i |T_i(X_j) - T_i(Y_j)| \tag{5.8}$$

In our experiments, we trained support vector machines on 7 classes of images (i.e. $C_k, k = 1,2,\cdots,7$ in Equation (5.6)) for the modeling of categorical context. Similar to the SSR training, the support vector machines were trained using a polynomial kernel with degree 2 and constant 1 ($C = 100$) [Joachims, 1999]. For each class, a human subject was asked to define the list of ground truth images from the 2400 collection and 20% of the lists was used for training. To ensure unbiased training samples, we generated 10 different sets of positive training samples from the ground truth list for each class based on uniform random distribution. The negative training (test) examples for a class are the union of positive training (test) examples of the other 6 classes and the `miscellaneous` class. The classifier training for each class was carried out 10 times on these different training sets and the support vector classifier of the best run was retained.

Table 5.1 lists the statistics related to the training of the SSC classes (left-to-right): SSC class labels, numbers of positive training examples (p-train), numbers of positive test examples (p-test), numbers of support vectors computed (sv), and the classification rate (rate) on the entire 2400 collection. The `miscellaneous` class

Table 5.1: Training statistics for SSCs

Semantic Support Classes	p-train	p-test	sv	rate
`Indoor:People inpp`	172	688	234	85.1
`Indoor:Objects inob`	27	107	136	95.7
`Nature:Park park`	61	243	158	92.4
`Nature:Mountain mtrk`	13	54	116	98.0
`Nature:Waterside wtsd`	30	120	151	95.3
`City:Pool pool`	10	42	72	98.7
`City:Street strt`	129	516	259	84.4

(not shown in the table) has 188 images that include images of dark scene and bad quality.

Similar to the illustration of the SSR-based and DSR-based indexes in the previous chapters, we show the $R_k(x) \geq 0.1$ of SSCs in Table 5.2 that correspond to the 3 images shown in Figure 5.4.

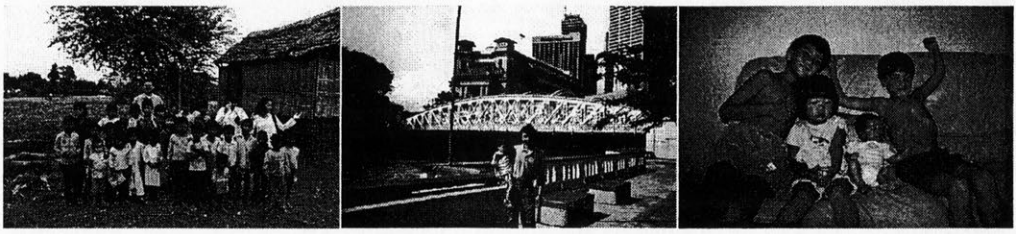

Figure 5.4: 3 image examples to illustrate SSC-based image indexes

Table 5.2: Key SSCs in the indexes for the images shown in Figure 5.4

Image x	Class C_k	$R_k(x)$
left	`park`	0.48
left	`strt`	0.20
left	`wtsd`	0.11
center	`strt`	0.76
right	`inpp`	0.48
right	`pool`	0.14
right	`wtsd`	0.12
right	`strt`	0.10

As can be seen from Table 5.2, the leftmost image in Figure 5.4 is classified mainly into the `park` class (0.48). But it has also resulted in some classification

outcome in SSRc `strt` (0.20) and `wtsd` (0.11) because images of these two classes also contain foliage and people. The single dominant class for the center image in Figure 5.4 is the `strt` class with a high confidence of 0.76. For the rightmost image in Figure 5.4, the SSC `inpp` (indoor people) has the highest classification output of 0.48, although there are also some prediction for the `pool` (0.14) and `wtsd` (0.12) classes due to the presence of large area of skin and for the `strt` class (0.10) due to the presence of people.

The effectiveness of the SSC-based image indexes will be evaluated using query by example experiments in Chapter 7.

5.4 Local Class Patterns

While the SSC indexing scheme described above focuses on the use of entire image representation (i.e. SSR detection-based image index) x for learning, classification, and indexing, this section looks at image classification decisions on local image regions z and the use of these *Local Class Patterns* (LCPs) for image indexing.

Recall that in the image semantics discovery framework described in Chapter 4, the classifiers C_k are trained on the feature vectors of local image blocks z to derive intra-class recurrent and inter-class discriminative patterns as DSRs. LCPs can be computed using the softmax function [Bishop, 1995] [Bridle, 1990] as,

$$T_k(z) = \frac{\exp^{C_k(z)}}{\sum_j \exp^{C_j(z)}}, \tag{5.9}$$

and embedded as image index, similar to the detection-based indexing of SSRs (Chapter 3) and DSRs (Chapter 4)), for similarity matching and retrieval.

To recapitulate, the indexing process consists of three steps, namely view-based detection (Section 3.4), multi-scale reconciliation (Section 3.5), and spatial aggregation (Section 3.6), as summarized in Figure 5.5. That is, in these steps of indexing, the SSRs and DSRs are replaced by LCPs.

In [Szummer and Picard, 1998], classification decisions on image blocks have been used as binary patterns for indoor/outdoor image classification. Our aim here is not image classification but image indexing based on local class patterns.

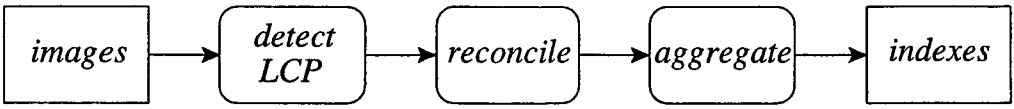

Figure 5.5: A schematic digram of image indexing based on LCPs

Table 5.3: Training statistics of classes learned for LCP-bsaed indexing

Class	size	#trg	#SV
inob	134	15	1905
inpp	840	20	2249
mtrk	67	10	1090
park	304	15	955
pool	52	10	1138
strt	645	20	2424
wtsd	150	15	2454

Moreover, we preserve the soft classification decision vectors and allow fine-grained tessellated blocks with multi-scale reconciliation.

As mentioned in the previous chapter, we trained 7 SVMs with polynomial kernels (degree 2, $C = 100$ [Joachims, 1999]) for the leaf-node categories (except `miscellaneous`) in Figure 5.3 on color and texture features (Equation (4.1)) of 60×60 image blocks (tessellated with 20 pixels in both directions) from 105 sample images. Hence each SVM $\mathcal{C}_k$ was trained on 16,800 image blocks z.

Table 5.3 lists the training statistics of the semantic classes C_k for LCP-based indexing (Equation (5.9)). The columns (from left to right) list the class labels, the number of images of each class in the 2400 collection, the number of training images, and the number of support vectors learned. The 105 training images are shown in Figure 4.6. Their top-down, left-to-right order (and the number of training images in each class) corresponds to the classes (and #trg) as listed in Table 5.3.

Similar to the illustration of the SSR-based and DSR-based indexes in the previous chapters, we show the $T_i(Z) \geq 0.1$ of LCPs in Tables 5.4, 5.5, and 5.6 that are detected, reconciled, and aggregated in 3 tessellated blocks (outlined in red bounding boxes) in Figures 5.6, 5.7, and 5.8 respectively.

For Figure 5.6, the key LCPs listed in Table 5.4 are reasonable. The top image block is classified as `park` with very high confidence of 0.81. For the center image

Figure 5.6: A sample image of park to illustrate LCP-based image index

Table 5.4: Key LCPs in the index for the image shown in Figure 5.6

Image Block	Key LCP Aggregated	$T_k(Z)$
top	park	0.81
center	park	0.48
center	inob	0.22
center	inpp	0.18
right	strt	0.23
right	inpp	0.22
right	park	0.19
right	wtsd	0.10

block, park is still the top choice (0.48). Indoor people (inpp) (0.18) is also detected due to the presence of crowd. There is also some false detection of indoor objects/interior (inob) (0.22). For the right image block, street strt and indoor people inpp have relatively higher classification output (0.23 and 0.22 respectively) due to the appearances of hut and people, although park with 0.19 is also detected. There is also some classification error for the waterside wtsd class (0.10).

In the case of Figure 5.7 (c.f. Table 5.5), the left block is classified mainly as street (strt) (0.36). It is also mistaken as mtrk class (0.32) due to the dark shadow area. For the center image block, the strt class has a strong confidence value of 0.66 due to the dominant presence of man-made structure in the image block. For the bottom image block, indoor people (inpp) (0.38) has a higher classification

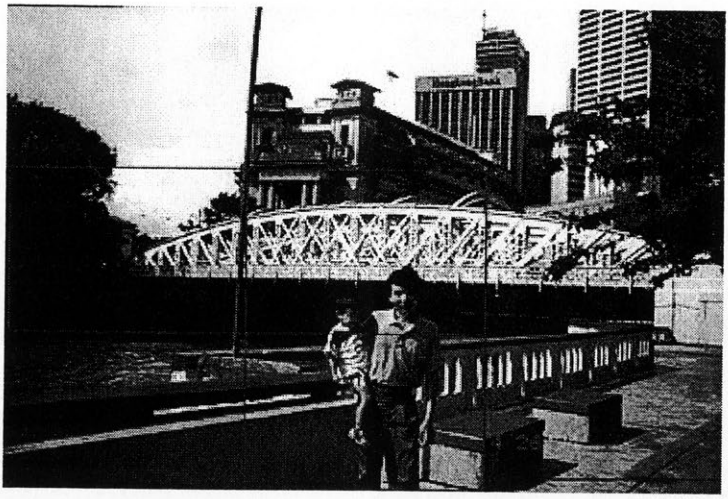

Figure 5.7: A sample image of street scene ito illustrate LCP-based image index

Table 5.5: Key LCPs in the index for the image shown in Figure 5.7

Image Block	Key LCP Aggregated	$T_k(Z)$
left	`strt`	0.36
left	`mtrk`	0.32
center	`strt`	0.66
center	`inpp`	0.16
center	`mtrk`	0.10
bottom	`inpp`	0.38
bottom	`strt`	0.29
bottom	`inob`	0.19

output than that of the `strt` class (0.29) as the appearance of human figure is more dominant. There is also false detection of indoor objects/interior (`inob`) (0.19).

In the image shown in Figure 5.8, the indoor people (`inopp`) class is dominant in all three image blocks considered (Table 5.6) with classification outputs of 0.77, 0.73, and 0.76 respectively. For the left image block, there is some classification error of 0.13 for the waterside (`wtsd`) class due to the background wall color that is similar to that of river water. Indoor objects/interior (`inob`) is detected for both center and right image blocks with same value of 0.15.

The effectiveness of the LCP-based image indexes will be evaluated using query by example experiments in Chapter 7.

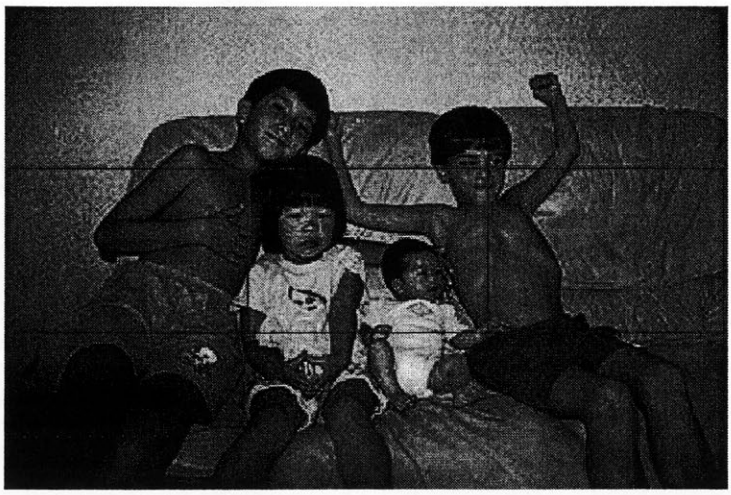

Figure 5.8: A sample image of indoor to illustrate LCP-based image index

Table 5.6: Key LCPs in the index for the image shown in Figure 5.8

Image Block	Key LCP Aggregated	$T_k(Z)$
left	`inpp`	0.77
left	`wtsd`	0.13
center	`inpp`	0.73
center	`inob`	0.15
right	`inpp`	0.76
right	`inob`	0.15

5.5 Discussion

We have proposed two local region semantics learning and indexing schemes (i.e. SSRs and DSRs) in the previous two chapters and three class-based semantics learning and indexing schemes in this chapter, one for event-based retrieval and two for similarity-based retrieval. We compare the four learning and indexing schemes designed for similarity-based retrieval, denoted as "SSR", "DSR", "SSC", and "LCP" as Table 5.7.

While both the SSR and DSR schemes refer to local visual semantics such as faces, foliage, water, buildings etc, the SSC and LCP schemes are based on global class meanings such as indoor, outdoor, city, nature etc (first row of Table 5.7). All the schemes compared except SSC focus at local regions for indexing. The SSC scheme looks at the entire image to compute its index (second row). Hence

Table 5.7: Comparison of indexing schemes based on SSR, DSR, SSC, and LCP

	SSR	DSR	SSC	LCP
semantics	local	local	global	global
index area	local	local	global	local
multi-scale	yes	yes	no	yes
trg. vectors	z	z	$T_i(Z_j)$	z
# trg. vectors	375	7467	613	16800
# trg. images	105	105	613	105
avg. # sup. vec.	33	253	161	1745
# semantic dim.	26	26	7	7

multi-scale view-based detection and reconciliation is only applicable to SSC, DSR, and LCP schemes but not SSC scheme (third row). The schemes that index on local regions use composite feature vector z (i.e. z^c and z^t) for SVM learning and detection while semantic histograms $T_i(Z_j)$ for a set of regions Z_j are necessary for SSC learning and classification (fourth row).

In terms of the size of training data for SVM learning, the semantics design approaches (SSR and SSC) require fewer samples that are provided by human designer (fifth row of Table 5.7). On the other hand, the semantics discovery approaches (DSR and LCP) use many more training samples but they can be generated with very little human effort. However, in terms of number of images used for SVM training (sixth row), we have deliberately kept it identical as 105 for local semantic schemes (SSR, DSR, and LCP) so that their retrieval performances for the query-by-example experiments (Chapter 7) are comparable.

The average numbers of support vectors computed from the SVM learning are listed in the seventh row of Table 5.7. They are more or less proportional to the number of training vectors (fifth row). Last but not least, as shown in the last row, indexing schemes based on local semantics (SSR and DSR) span an indexing space of 26 semantic axes while indexing based on global semantics (SSC, LCP) only has 7 index dimensions (i.e. 7 disjoint consumer image classes).

Chapter 6

Integrated Similarity Matching

Two paradoxes are better than one, they may even suggest a solution.

Edward Teller (1908 - 2003)

6.1 Overview

To bridge the semantic gap in content-based image indexing and retrieval, several indexing schemes based on local and global semantics have been proposed in previous chapters. A key emphasis in these proposed indexing schemes is the use of statistical learning to train pattern classifers for semantics detection. Except for event-based image retrieval (Section 5.2), the other schemes (SSR, DSR, SSC, LCP) represent an image index in the form of normalized histograms that can be used in similarity-based retrieval.

In this chapter, we will focus on the distance (i.e. dissimilarity) and similarity measures for these indexing schemes. Most importantly, we present a unified principle based on Bayesian probability theory to combine local and global semantic indexes in similarity matching, as another key idea in this thesis is to bridge the semantic gap. Interestingly, the integrated similarity matching schemes involving different pairs of image indexes form dual frameworks corresponding to semantic design and discovery approaches respectively.

6.2 Similarity Matching

Many distance and similarity measures have been studied empirically for similarity-based retrieval in the literature [Puzicha et al., 1999]. We believe that the effectiveness of a similarity (or distance) measure depends on the application domain and the index representation. In this thesis, we have only considered several more commonly used similarity (or distance) measures in our similarity-based retrieval experiments. We shall discuss these measures below and compare their effectiveness in the next chapter.

6.2.1 Local Index

Suppose we wish to compare two image indexes x and y, each consists of m detection vectors T_i for local blocks X_j and Y_j respectively. Their similarity (or distance) can be computed in terms of the similarities (or distances) between their corresponding local blocks. The detection vector T_i could be either SSR-based, DSR-based, or LCP-based.

A popular similarity measure used by the information retrieval community is the cosine similarity (i.e. normalized dot product),

$$\Omega(x,y) = \frac{1}{m}\sum_j \frac{\sum_i T_i(X_j)T_i(Y_j)}{\sqrt{\sum_k T_k(X_j)^2}\sqrt{\sum_k T_k(Y_j)^2}}. \tag{6.1}$$

One of the simplest distance measures is the L_1-norm city block distance,

$$\Delta(x,y) = \frac{1}{m}\sum_j \sum_i |T_i(X_j) - T_i(Y_j)|. \tag{6.2}$$

If L_2-norm Euclidean distance is used, then the distance measure becomes,

$$\Delta(x,y) = \frac{1}{m}\sum_j \sqrt{\sum_i (T_i(X_j) - T_i(Y_j))^2}. \tag{6.3}$$

Another distance measure considered is the Kullback-Leibler (KL) distance (or cross-entropy) [Kapur and Kesava, 1992] for comparing two probability distribu-

tions x and y,

$$D(x||y) = \sum_i x_i \ln \frac{x_i}{y_i}. \tag{6.4}$$

As $D(x||y)$ is not symmetric in general, a symmetric cross-entropy can also be used [Kapur and Kesava, 1992],

$$J(x||y) = D(x||y) + D(y||x). \tag{6.5}$$

For comparing local semantic image indexes, $D(x||y)$ is modified as

$$\Delta(x,y) = \frac{1}{m} \sum_j \sum_i T_i(X_j) \ln \frac{T_i(X_j)}{T_i(Y_j)}. \tag{6.6}$$

6.2.2 Global Index

In the case of inter-class indexes based on SSCs, the definitions of the distance and similarity measures are similar to those described above but simpler. Suppose x and y denote the detection-based indexes for two images. Then their SSC-based indexes are represented as $R_k(x)$ and $R_k(y)$ respectively.

The cosine similarity is defined as

$$\Omega_{ssc}(x,y) = \frac{\sum_k R_k(x) R_k(y)}{\sqrt{\sum_i R_i(x)^2} \sqrt{\sum_i R_i(y)^2}}. \tag{6.7}$$

The city block distance is computed as

$$\Delta_{ssc}(x,y) = \sum_k |R_k(x) - R_k(y)|. \tag{6.8}$$

The Euclidean distance is defined as

$$\Delta_{ssc}(x,y) = \sqrt{\sum_k (R_k(x) - R_k(y))^2}. \tag{6.9}$$

Last but not least, the KL-distance is computed as

$$\Delta_{ssc}(x,y) = \sum_k R_k(x) \ln \frac{R_k(x)}{R_k(y)}. \tag{6.10}$$

6.3 Combining Local and Global Similarities

The task of looking for images to satisfy some information need is complex because it may range from targetting at a specific image, searching for a group of images that share some common properties, to browsing without a well-defined objective other than finding interesting images [Cox et al., 2000].

Suppose every database image is represented as a data point in the indexing space. While target-specific search is equivalent to finding a particular data point in this space, browsing is like wandering in this space until some images that match the user's current dynamic information need are found (or until the user decides to terminate browsing). Between these two extreme modes of image information seeking, the more common form of category-based image search aims to look for a group of related images, represented as a distribution of data points in the indexing space, that share some common properties entailed by the information need. Since target-specific search is too rigid and aimless browsing is ill-defined, we shall focus on the category-based search that finds images similar to some query image(s) in this thesis.

Given an image retrieval system with a database of N images (assume N is stable within a query session), the hidden information need of a user over the N images can be modeled as the posterior probability of the set of relevant images R given an expression of the information need in the form of query specification q and an image x in the current database, $P(R|q,x)$. We assume that the image retrieval system can compute $P(R|q,x)$ for each x in the database. The objective of the system is to return images with high probabilities of relevance to the user.

When the query is in the form of image examples, query processing becomes an underconstrained density estimation problem i.e. to compute the probability distribution of relevance based on very few query examples. Though there are innovative techniques proposed to increase the number of training examples with relevance feedback technique [Wu et al., 2000b] [Tieu and Viola, 2000], we would like to investigate the role of query example further with two key observations.

First, in query by example (QBE), the probability of relevance depends on the similarity between query q and image x. Next, we note that the set of relevant images R does not exist until a query has been specified. However we can construct

prior categories of images $C_k, k = 1, \cdots, M$ as some prototypical instances of R and compute the memberships of q and x to these prior categories for contextual similarity.

We believe that both local (intra-image) and global (inter-class) similarities play complementary roles in image matching and ranking. Using a Bayesian formulation, we have

$$P(R|q,x) = \frac{P(q,x|R) \cdot P(R)}{P(q,x)} \tag{6.11}$$

We observe that $P(q,x)$ tends to be small if q and x are similar (i.e. less likely to find similar images than dissimilar pair in a large database). On the other hand, $P(q,x|R)$ tends to be large if q and x are similar with respect to R (i.e. q and x are more likely to co-occur in R if they belong to R). And $P(R)$ is constant for a given query session.

Hence $P(R|q,x)$ is proportional to the similarity between q and x given R (denoted as $\mu(q,x)$) and the similarity between q and x in terms of their contents (denoted as $\lambda(q,x)$) i.e.

$$P(R|q,x) \propto \mu(q,x) \star \lambda(q,x). \tag{6.12}$$

where $\star$ is some confluence operator to combine the similarities.

For the purpose of retrieval, Equation (6.12) provides us a principled way to rank images x by their probabilities of relevance to the user's information need as represented by the query example q. Indeed when the similarities $\mu(q,x)$ and $\lambda(q,x)$ are expressed in the form of probabilistic distance (i.e. inverse of similarity) such as the KL distance (or cross-entropy) (Equation (6.4)), ordering images from the smallest distance to the largest distance is the manifestation of the minimum cross-entropy principle ([Kapur and Kesava, 1992], pp. 13).

To recapitulate, the principle requires us to choose the *a posterior* probability distribution x such that it satisfies all given constraints, and it has the minimum cross-entropy relative to the specified *a priori* distribution. In the case of QBE, q plays the role of *a priori* distribution. Hence Equation (6.12) echoes the *Probability Ranking Principle* in text information retrieval [Robertson and Sparck Jones, 1976]. We quote the principle below for ease of reference:

The Probability Ranking Principle

If a reference retrieval system's response to each request is a ranking of the documents in the collection in order of decreasing probability of relevance to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data have been made available to the system for this purpose, the overall effectiveness of the system to its user will be the best that is obtainable on the basis of those data.

Figure 6.1 depicts our proposed similarity integration framework in a schematic block diagram, using similarity matching based on SSR and SSC indexes for $\lambda(q, x)$ and $\mu(q, x)$ respectively. At the topmost row, an image to be indexed is subjected to multi-scale view-based detection against SSRs learned from selected image regions (second row, left). The soft detection decisions are reconciled and aggregated spatially into compact indexes. Indexes of selected classes are used to learn the SSCs for the image collection (second row, right). Given a query and an image in the database (bottom), the matching is being done at both the content and context levels as well as combined and ranked under probabilistic principles.

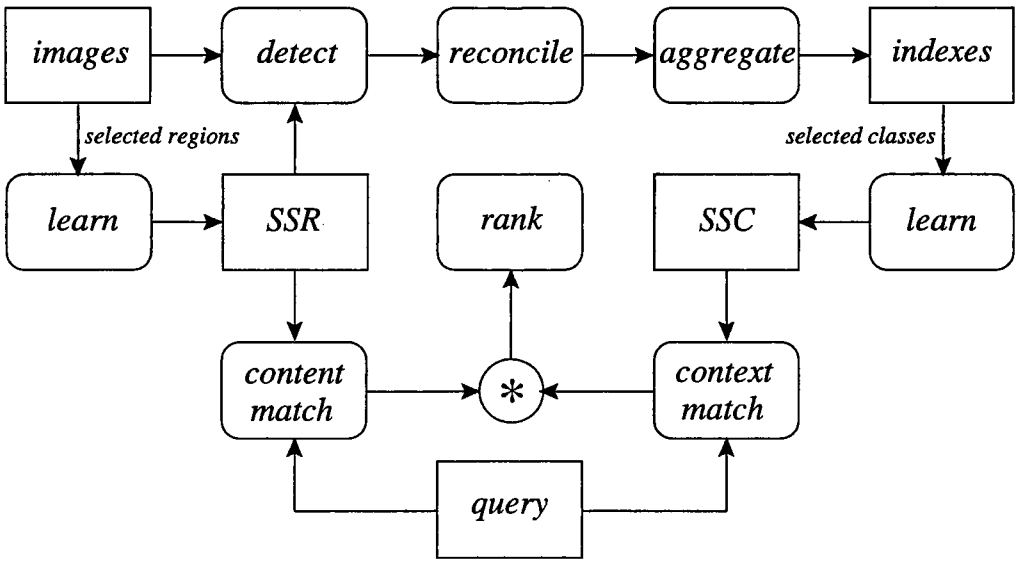

Figure 6.1: System flow of indexing and retrieval with similarity integration

Since only ranking matters for practical image retrieval and the actual relation

between $\mu(q,x)$ and $\lambda(q,x)$ in Equation (6.12) is unknown, we consider two simple schemes in our empirical evaluation in the next chapter. A multiplicative scheme is defined as

$$P(R|q,x) = \mu(q,x) \times \lambda(q,x), \tag{6.13}$$

and a linear combination ($\omega \in [0,1]$) scheme is

$$P(R|q,x) = \omega\mu(q,x) + (1-\omega)\lambda(q,x). \tag{6.14}$$

Besides substituting SSR and SSC similarity matching for $\lambda(q,x)$ and $\mu(q,x)$ respectively, we have also experimented with similarities based on DSR and LCP indexes respectively.

6.4 Dual Frameworks

Based on the integrated similarity matching scheme described above, both the local (i.e. SSR, DSR) and global (SSC, LCP) image semantics are unified in the matching and ranking of images for similarity-based retrieval. In particular, following different computational pathways, as depicted in Figure 6.2, result in dual cascading learning frameworks that combine both intra-image and inter-class semantics for image indexing and retrieval.

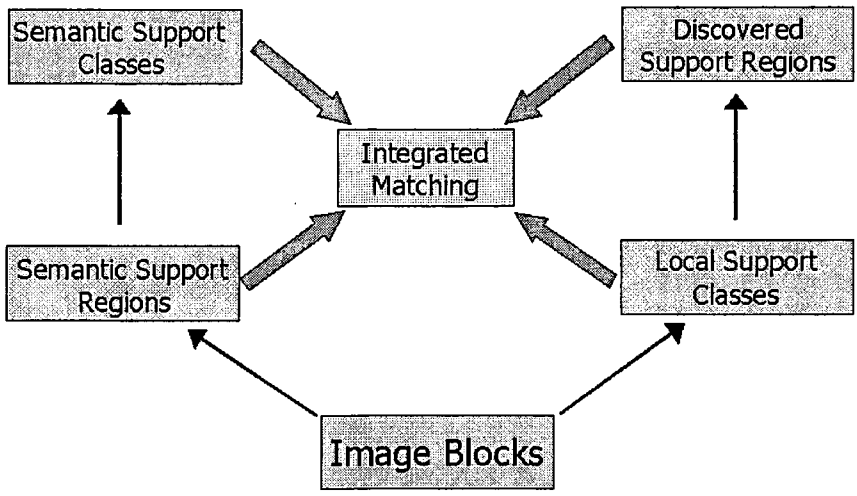

Figure 6.2: Dual cascading image indexing and matching frameworks

The computational pathway on the left-hand-side of Figure 6.2 corresponds to the semantics design framework based on supervised learning. In this design framework, support vector detectors for SSRs that are determined for an application domain are trained. The reconciled and aggregated detection-based indexes then serve as input patterns for support vector learning of image classifiers to generate inter-class (SSC) image indexes. During retrieval, similarities based on both indexes are combined to rank images. Hence the path is based on construction of local semantics in order to generate global semantics before their respective image indexes can be integrated in similarity matching.

On the other hand, the semantics discovery framework based on hybrid supervised and unsupervised learning provides an alternative flow for indexing and retrieval, shown as the computational pathway on the right-hand-side of Figure 6.2. In this discovery framework, support vector image classifiers are first trained on local image blocks from a small number of class-labeled images. Then local semantic patterns are discovered from clustering the image blocks with high classification output. Training samples are induced from cluster memberships for support vector learning to form local semantic pattern detectors. In the similar manner, the similarities based on LCP-based indexes and DSR-based indexes are combined to rank images during retrieval. Thus, in contrast to the left computational pathway, the flow starts with global semantics to induce local semantics before their respective image indexes are used in combined similarity matching.

Chapter 7

Query and Retrieval

Why think? Why not try the experiment?
John Hunter

Performance is your reality. Forget everything else
Harold Geneen

7.1 Overview

In this chapter, to address the semantic interpretation problem mentioned in Section 1.1.3, we present three query formulation and associated query processing methods, namely Qyery by Class/Event (QBCE), Query by Spatial Icons (QBSI), and Query by Multiple Examples (QBME), as means to reduce the ambiguity and subjectivity in query interpretation.

While QBCE supports queries at the high-level semantics using predefined image class or event labels, QBSI allows visual query formulation based on spatial arrangement of visual icons, representing predefined local visual semantics. These two query methods are easily supported by the global and local indexing schemes presented in previous chapters respectively. While they allow explicit specification of visual semantics, the QBME method defines the information need in implicit manner based on multiple image examples, as a simple extension of the conventional

query-by-example method.

The three query methods are evaluated using different sets of queries on 2400 real consumer images. The data set is first described in the next section. In each of the subsequent sections devoted to the three query methods separately, the details of query processing, queries and ground truth, and experimental results are described respectively.

7.2 Test Collection

In this thesis, we have decided to evaluate our proposed indexing schemes and query methods on broad domain consumer images. As explained in Chapter 1, consumer images exhibit very high content variations with very few annotations. They are very challenging for content-based image retrieval evaluation.

We are fortunate to have access to a collection of 2400 consumer images from a single family (Mr. Jean-Luc Lebrun) for the experiments in this thesis. These genuine consumer images are taken over 5 years in several countries with both indoor and outdoor settings. The images are those of the smallest resolution (i.e. 256×384) from Kodak PhotoCDs, in both portrait and landscape layouts. After removing possibly noisy marginal pixels, the images are of size 240×360. The indexing process automatically detects the layout and applies the corresponding tessellation template for portrait or landscape layout. On one hand, the small resolution of the images allows for more efficient processing. On the other hand, they pose greater challenge for feature extraction and visual concept detection.

To have a feel of the content diversity in our 2400 collection, we show 72 (3%) of them in Figure 7.1. For outdoor images, the content varies from natural landscape (beach, lakeside, river, pond, park, forrest, garden, mountain, rocky area etc) to city scenes (urban area, rural area, crowded street, market, road with vehicles, swimming pool, temple, mosque, castle etc) from different countries and cultures (Singapore, France, China, Cambodia, Malaysia, Indonesia etc). The indoor images are taken with different focuses (portrait of single person or a few people, groups of different sizes, people having meal, cultural performance, wedding ceremony, interior layout, display of objects like painting, toys, antique collection etc). In both outdoor and

indoor images, the subject of focus could be people (or faces in photo frame), statues, animals, flowers, buildings (or their miniature in theme park) etc and their mixture with occlusion, taken with different posture, during day or night, from different viewpoints, and at different distances. Figure 7.2 illustrates some of the photos of bad quality (e.g. faded, over-exposed, blurred, dark etc). We did not remove these bad quality photos from our test collection in order to reflect the complexity of the original data. There are 188 (approx. 7.8%) such kind of noisy and ambiguous photos in our 2400 test collection.

7.3 Query by Class/Event (QBCE)

7.3.1 Query Processing

Query by Class/Event (QBCE) refers to query based on predefined image class or event labels. The event-based retrieval framework proposed in Section 5.2 can be used to support this kind of queries. Given a selected class or event label, the query processing algorithm of QBCE needs to decide the set of relevant images in the database for the selected class or event and return them to the user. In addition, the returned images should be ranked in descending order of relevance for efficient browsing of the thumbnail images.

Recall that in the winner-take-all approach for event-based image retrieval, the relevance measure $R(M_i, Z)$ of any image with index Z can be computed against the learned Event (or class) model M_i in three simple steps. First the output of an Event model M_i, denoted as $S(M_i, Z)$, can be computed as

$$S(M_i, Z) = \frac{\exp^{\mathcal{M}_i(Z)}}{\sum_j \exp^{\mathcal{M}_j(Z)}}, \tag{7.1}$$

where $\mathcal{M}_i(Z)$ denotes a SVM classifier output.

Next the winner k is decided by

$$k = argmax_i S(M_i, Z). \tag{7.2}$$

Figure 7.1: Sample consumer photos in the 2400 test collection

Figure 7.2: Some consumer images of bad quality

Lastly the relevance measure of Z to Event model M_i is calculated as

$$R(M_i, Z) = \begin{cases} S(M_k, Z), & \text{if } i = k \\ 0, & \text{otherwise} \end{cases} \tag{7.3}$$

7.3.2 Queries and Ground Truth

We have performed two sets of QBCE experiments. The first experiment for event-based retrieval includes 5 common events from the proposed event taxonomy as shown in Figure 5.1. They are events `meal`, `wedd`, `park`, `wtsd`, `pool` as listed in Table 7.1. For each event, the list of photos considered relevant to the event is constructed from the 2400 photos by 3 human subjects. The sizes of the ground truth lists are shown in Table 7.1. Figure 7.3 shows 3 sample photos of each event.

Table 7.1: Events and ground truth (G.T.) sizes

Label	Event	G.T.
`meal`	Having meal	76
`wedd`	Wedding	241
park	At the park	304
wtsd	Along waterside	114
pool	At swimming pool	52

Figure 7.3: Sample photos of each (column) event in Table 7.1

The second experiment for QBCE involves 11 image categories (i.e. all nodes except the root node and the `miscellaneous` node) in the proposed consumer image category hierarchy shown in previous chapters and repeated in Figure 7.4 for ease of

reference. Again for each of the 11 categories considered, the list of relevant photos is constructed from the 2400 photos by 3 human subjects. The sizes of the ground truth lists are shown in Table 7.2. Figure 7.5 shows some sample photos of these categories. Note that the `wtsd` category is different from the `wtsd` event in Table 7.1 as the latter does not include images taken at the beaches.

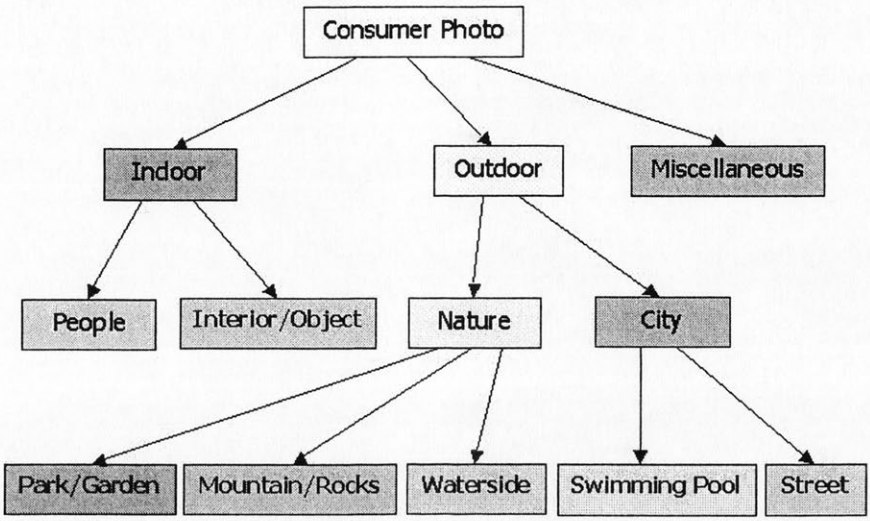

Figure 7.4: A hierarchy of consumer image categories

Table 7.2: Categories and ground truth (G.T.) sizes

Label	Category	G.T.
`indr`	indoor	994
`outd`	outdoor	1218
`misc`	miscellaneous	188
`inpp`	people indoor	860
`inob`	interior/object	134
`city`	city	697
`natr`	nature	521
`pool`	swimming pool	52
`strt`	street	645
`wtsd`	waterside/beach	150
`park`	park/garden etc	304
`mtrk`	mountain/rocks	67

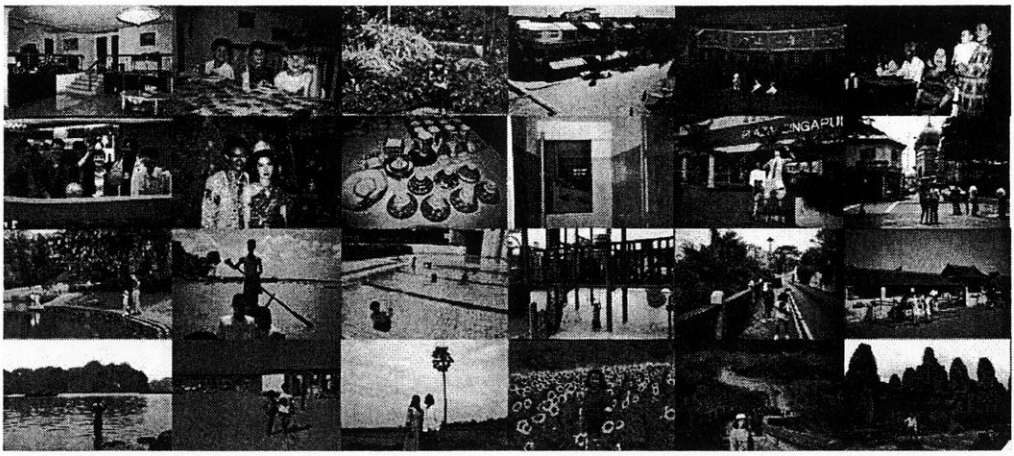

Figure 7.5: Two sample photos for each category listed in Table 7.2

7.3.3 Experimental Results

For the event-based retrieval experiment (Table 7.1), the learning of each of the 5 events is based on a training set of only 10 labeled photos to simulate practical situations (i.e. a user only needs to label 10 photos for each event). To ensure unbiased training samples, 10 different training sets are generated from the ground truth list for each event based on a uniform random distribution. The learning and retrieval of each event are thus performed 10 times and the results are averages over these 10 runs. Note that for each of these runs, the photos used for training are removed from the ground truth when computing the precision and recall values.

To retrieve photos of an Event, a user simply selects one of the 5 event labels. Based on the Event models, the relevance measures were computed as given above. We compare our event modeling and retrieval approach (denoted as "EBR") with a baseline method, color histogram of 11 key colors (red, green, blue, black, grey, white, orange, yellow, brown, pink, purple) in the HSV color space, as adopted by the original PicHunter system [Cox et al., 2000] (denoted as"HSV-11").

Table 7.3 lists the average precisions (over 10 runs) of retrieval for each event among the top 20 and 30 retrieved photos for the two methods and the percentages of improvement by the EBR method over the HSV-11 method.

From Table 7.3, we see that our method performs extremely well for the `park` and `wedd` events and fairly well for the `pool` event. The `meal` and `wtsd` events remain as challenging problems to be addressed, likely due to the lack of consistency in their

Table 7.3: Average precisions at top numbers of photos

Avg. Prec.	HSV-11	EBR	Improved
meal, top 20	0.08	0.16	100%
meal, top 30	0.08	0.17	113%
wedd, top 20	0.56	0.91	63%
wedd, top 30	0.58	0.90	55%
park, top 20	0.48	0.96	100%
park, top 30	0.43	0.96	123%
wtsd, top 20	0.20	0.36	80%
wtsd, top 30	0.17	0.31	82%
pool, top 20	0.16	0.61	281%
pool, top 30	0.13	0.51	292%

visual contents. Nevertheless, in all cases, our method significantly outperforms feature-based approach such as color histogram.

In the second experiment for the 11 image categories shown in Table 7.2, 20% of the 2400 photos is designated as training samples to learn a category. Similarly, to ensure unbiased training samples, 10 different sets of positive training samples are generated from the ground truth list for each category based on uniform random distribution. The negative training samples of a given category are positive training samples from other categories that do not overlap with the category.

The evaluation of retrieval precision is carried out hierarchically with respect to the category tree in Figure 7.4. The test data for the category of a child node in a run is the ground truth list of its parent node minus the training samples used for learning the category of the child node in the run. For example, to evaluate the retrieval performance of nature (natr) photos, the ground truth list of outdoor less the training samples used for the nature category is taken as the test data. As another example, the test data for indoor is all the 2400 photos minus the training samples for learning the indoor category. The learning and retrieval of each category were performed 10 times and the results are averages over these 10 runs. Table 7.4 lists the average precisions (over 10 runs) of retrieval among the top 20, 30, and 50 retrieved photos for the 11 categories.

From Table 7.4, we observe that, on average and up to first 50 thumbnails, a user is guaranteed to get almost all relevant photos of the respective categories except less

Table 7.4: Average precisions at top numbers of photos

Avg.Prec.	Top 20	Top 30	Top 50
indr	0.94	0.96	0.96
outd	1.00	1.00	1.00
inpp	0.99	0.99	0.99
inob	0.84	0.75	0.56
natr	0.96	0.96	0.95
city	0.95	0.94	0.93
park	1.00	0.99	0.98
mtrk	0.41	0.27	0.16
wtsd	0.92	0.89	0.66
pool	0.47	0.32	0.21
strt	0.99	0.99	0.99

so for the categories interior/object (inob) and waterside (wtsd), and even less so for the categories mountain/rocks (mtrk) and swimming pool (pool). The reasons for poorer performance in these 4 categories are two-fold. First these categories have much fewer positive training samples (i.e. 27, 30, 13, 10) for statistical learning. Moreover, they also comprise images of varied contents (Figure 7.5: interior versus object(s), mountain versus rocks, river-side and lakeside versus beach (no water visible), pool water dominant versus pool with other structure). We believe that with more training samples, their performance would be raised.

We believe that high precision values at top number of retrieved photos is important. In practice, this implies that the user can easily locate relevant photos in one or two pages of photo thumbnails displayed on a computer screen. If the client device is a mobile device such as PDA and cellphone with limited display area (say 4 to 6 thumbnails per screen), our approach can sustain a high precision value that shows many relevant photos in the first few pages before the user loses his or her patience.

7.4 Query by Spatial Icons (QBSI)

7.4.1 Query Processing

Query by Spatial Icons (QBSI) is a new query formulation method that allows explicit specification of visual semantics in terms of "what" and "where". A QBSI query is composed as a spatial arrangement of visual icons, hence the name. Query processing for QBSI involves both pattern-based and logic-based computation. A Visual Query Term (VQT) q specifies a region R where a local semantic region (such as SSR) S_i should appear. A query formula then chains these VQTs together via appropriate logical operators. The truth value $\lambda(q, x)$ of a VQT q for any image x is simply defined as

$$\lambda(q, x) = T_i(R) \tag{7.4}$$

where $T_i(R)$ is an aggregated detection-based index as defined in Equation (3.29) of Section 3.6.

In our current implementation, we support a two-level Is-A hierarchy of SSRs (Figure 3.5) though it can be extended to deeper or other forms of hierarchies (e.g. Part-Whole hierarchy) (c.f. Section 3.7). A VQT can involve a more specific visual semantics (e.g. swimming pool water, denoted as `Water:Pool`) or a more abstract semantics (e.g. water, denoted as `Water`). On the other hand, the spatial constraint R defines the location and size of the specified visual semantics as drawn on a canvas.

As described in Section 3.7, the truth value $D_k(R)$ of a VQT that specifies a more abstract visual semantics C_k (i.e. `People`, `Sky`, `Ground`, `Water`, `Foliage`, `Mountain`, `Building`, and `Interior`) can be computed in terms of the truth values of more specific visual semantic classes S_i that belong to C_k,

$$D_k(R) = \max_i T_i(R). \tag{7.5}$$

A QBSI query Q can be specified as a disjunctive normal form of VQT (with or without negation),

$$Q = (q_{11} \wedge q_{12} \wedge \cdots) \vee \cdots \vee (q_{c1} \wedge q_{c2} \wedge \cdots). \tag{7.6}$$

Then the query processing of query Q for any image x is to compute the truth value $\lambda(Q, x)$ using appropriate logical operators. As uncertainty values are involved in SSR detection and indexing, we adopt fuzzy operations [Klir and Folger, 1992] as follows:

$$\lambda(\bar{q}, x) = 1 - \lambda(q, x), \tag{7.7}$$

$$\lambda(q_i \wedge q_j, x) = \min(\lambda(q_i, x), \lambda(q_j, x)), \tag{7.8}$$

$$\lambda(q_i \vee q_j, x) = \max(\lambda(q_i, x), \lambda(q_j, x)). \tag{7.9}$$

In short, the query processing algorithm of QBSI deals with the certainties $T_i(R)$ and $D_k(R)$ of word labels S_i and C_k (e.g. `Water:Pool`, `Water`) extracted from image region R respectively. These are abstraction learned upon low-level features such as color and texture. The indexes do not store the feature values anymore and hence the matching does not involve low-level features.

Nevertheless, the vocabulary for QBSI is limited by the semantics that can be learned and detected in image content. For instance, abstract concepts such as 'happiness' and 'Africa' would require presence of additional text annotation associated with the images which are not always available in certain application domains (e.g. consumer photos). In this thesis, we focus on semantics that can be extracted from the image content alone.

In our existing web-based prototype, an intuitive graphical interface is provided for a user to specify a QBSI query. To specify a VQT, the user first selects a SSR (specific or abstract) from a palette of icons associated with the SSR. Then a spatial image region based on the selected icon can be drawn by clicking and dragging a rectangular box in a canvas. If the user wishes to apply a negation operator, he or she can click on the "NOT" button followed by the drawn region. A yellow cross will be superimposed on the selected region. The user can continue to specify more VQT in a conjunct by repeating the above steps. The user can also start a new conjunct in the disjunctive normal form (Equation (7.6)) by clicking on the "OR" button to bring up a new window with canvas and icons. A reset button is provided to clear all the icons drawn for a conjunct in a given window. A typical screen shot is given in Figure 7.6 (note that only a subset of the visual icons are displayed in

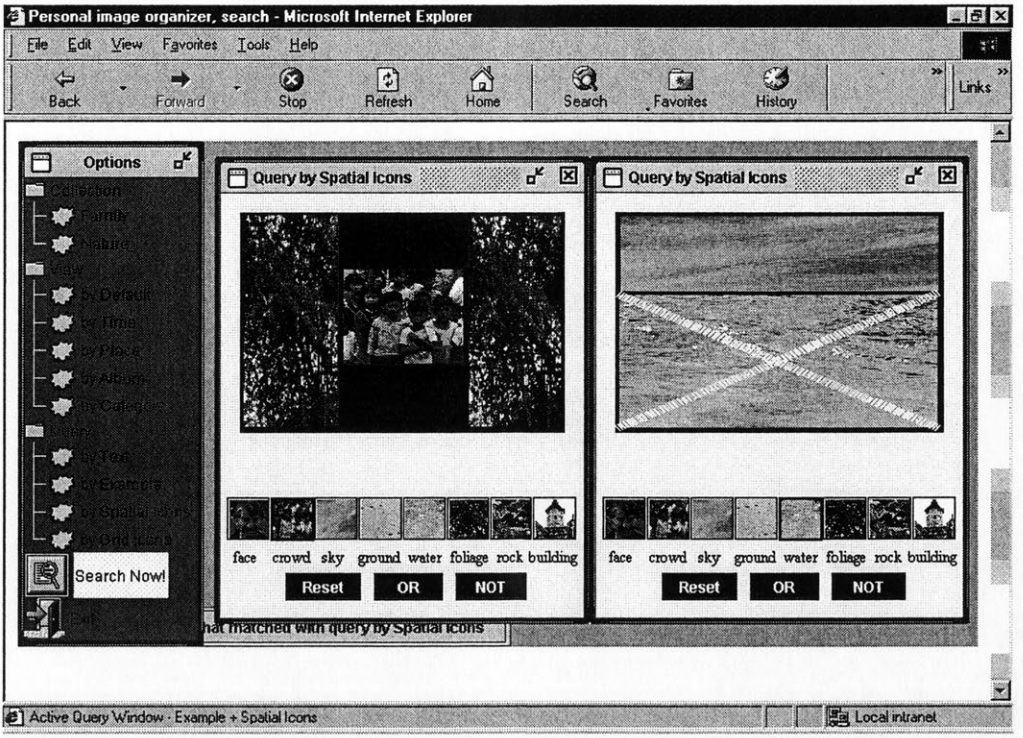

Figure 7.6: A screen shot for QBSI interface

this prototype). In particular, the figure illustrates a disjunct of two conjuncts, one with 3 visual query terms (left) and the other with 2 visual query terms (right), one of which is a negation on water.

As the region specified by a VQT is arbitrary, the precise computation of $T_i(R)$ using Equation (3.29) on reconciled small regions z_k is not cost effective in terms of speed and storage. Hence as a trade-off in our implementation, we pre-indexed the images using a uniform 3×3 spatial tessellation with the 26 SSRs defined in Figure 3.5 based on Equations (3.24) and (3.29). The truth value of a VQT q with region R and SSR S_i is approximated as,

$$\lambda(q, x) = \frac{\sum_{Z_j \in Z} T_i(Z_j)}{|Z|} \tag{7.10}$$

where Z consists of any of the 3×3 blocks that has more than half of its area covered by region R.

Another QBSI interface that corresponds to the 3×3 indexing grid is also supported. That is, the user can click on an icon associated with a SSR and fill any

block in the 3×3 grid canvas with the selected icon. In a similar way, a negation operator ("NOT" button) can be applied to a block (which will be crossed in yellow) and a new window with grid and icons can be invoked ("OR" button) to start a new conjunct.

The ImageScape system [Lew, 2000] also allows placement of icons (face, sky, water, tree/grass, and sand/stone) on a canvas to create a query. However, unlike our QBSI approach, the spatial extent of the placed icons is not emphasized. Moreover, it is not clear in [Lew, 2000] that how a query of semantic icons is processed. Last but not least, no proper evaluation has been reported.

As an enhancement to Query by Example (QBE) method, the Query By Multiple Regions (QBMR) approach [Moghaddam et al., 2001] allows a composition of query from multiple regions from example images with or without spatial layout. Our QBSI approach can complement the QBMR method in two aspects. It is useful when the user is not looking for specific visual similarity but rather more abstract visual concepts. The QBSI interface can also be used to obtain an initial set of relevant images for QBMR as the latter still suffers from the boostrapping problem. Furthermore, the QBSI approach does not need the computation of best matching region and best spatial configuration as required by QBMR [Moghaddam et al., 2001]. The query processing of QBSI, which is based on principled fuzzy operations, is simple and efficient.

Another novel feature in our approach not available in the above works is hierarchy of visual concepts. That is, SSRs can be structured into Is-A or Part-Whole hierarchy for detection, indexing, and query. For example, a sky SSR class can further be divided into subclasses of clear, cloudy, and blue skies with associated specific detectors. A QBSI query can then involve a specific type of sky or a generic sky concept. We will demonstrate this kind of queries and the underlying query processing in our experiments below. Another interesting structural mechanism is to detect and index a SSR in terms of its parts (e.g. [Mohan et al., 2001]).

Note that the free-form QBSI interface as shown in Figure 7.6 and grid-based QBSI interface mentioned above are two of the query functions provided in our operational prototype which is implemented in C and Java with Microsoft Access. Our web-based system also allows query by examples, query by text, query by mixture of

query modes, browsing along different dimensions (time, place, people, categories), data management (e.g. addition, deletion, copying of photos and albums), text annotation, SMIL-based [W3C, 2001] slideshow authoring and presentation with music. Last but not least, separate tools are also provided for uploading of images to the web, visual queries (QBE and QBSI) and slideshow presentation on PocketPC.

7.4.2 Queries and Ground Truth

To evaluate the effectiveness of QBSI using SSR-based image indexes for the 2400 consumer images, we have designed 15 QBSI queries. They are illustrated in Figures 7.7 to 7.11.

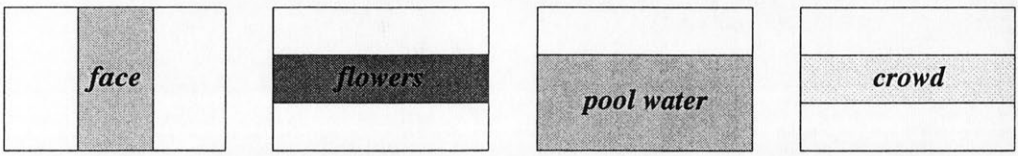

Figure 7.7: QBSI queries Q01 to Q04

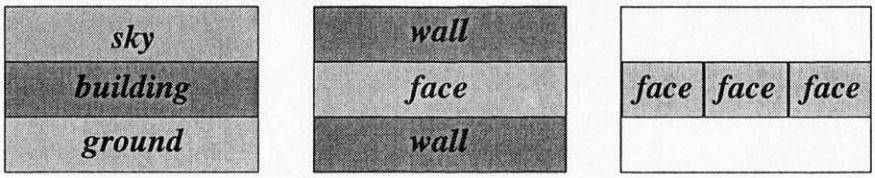

Figure 7.8: QBSI queries Q05 to Q07

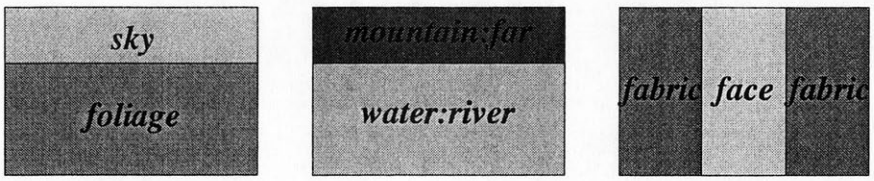

Figure 7.9: QBSI queries Q08 to Q10

While queries Q01 to Q04 focus on single VQT, queries Q05 to Q15 demonstrate multiple VQTs. In particular, query Q06 is composed to look for indoor images

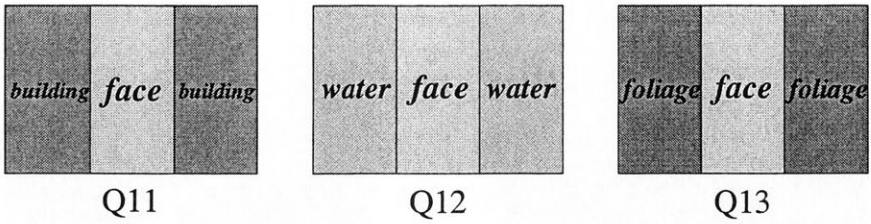

Figure 7.10: QBSI queries Q11 and Q13

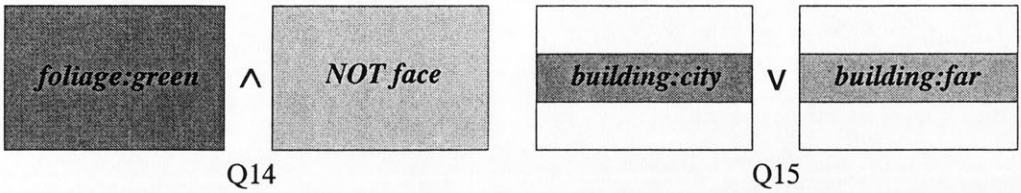

Figure 7.11: QBSI queries Q14 and Q15

with close-up of people. Query Q07 specifies faces in 3 different regions to enforce "small group of people". Query Q10 intends to retrieve images related to wedding events whereby auspicious fabric can be seen. Query Q14 shows the use of the negation operator. Last but not least, query Q15 illustrates the usefulness of disjunct operator. All the queries except Q05 and Q08 involve specific SSRs. Queries Q05 and Q08 are based on superclass of SSRs. Queries Q11 to Q13 illustrates the flexibility of mixing SSR (face) and the superclasses (building, water, and foliage). Our SSR indexing framework supports query with different levels of visual semantics and their mixture.

7.4.3 Experimental Results

The image indexes to support QBSI are computed based on Equations (3.24) and (3.29) with face detection enhancement [Rowley et al., 1998]. With our modular framework, the replacement of object detection decisions is simple as described in Section 3.10.

Table 7.5 lists the number of relevant images among the top 20 and 30 retrieved images as well as the size of the ground truth (G.T.) for each of the queries tested. As shown in the table, the average precisions for the top 20 and 30 retrieved images are 0.79 and 0.70 respectively, which we consider effective for practical applications. Interestingly, queries Q02 and Q09 demand small number of specific images (i.e.

Table 7.5: Precisions at top retrieved images for QBSI experiment

Query	Top 20	Top 30	G.T.
Q01	14	24	590
Q02	18	23	26
Q03	14	16	44
Q04	16	19	78
Q05	19	26	281
Q06	14	20	302
Q07	20	20	380
Q08	18	25	83
Q09	12	16	19
Q10	14	17	112
Q11	16	25	523
Q12	11	16	61
Q13	18	25	259
Q14	18	25	107
Q15	15	20	234
Avg	*15.8*	*21.1*	

less than 30; around 1%) to be found among 2400 images. The recall among top 30 retrieved images is high with recall values 0.88 (23/26) and 0.84 (16/19) respectively (i.e. almost all the relevant images are found among the top 30 retrieved images).

Next we show the top retrieved images for 3 of the 15 queries, namely queries Q02, Q05, and Q07, in Figures 7.12, 7.13, and 7.14 respectively. In the figures, the top 18 images retrieved are shown in top-down, left-to-right order of decreasing relevance.

Figure 7.12: Top 18 retrieved images for QBSI query Q02

For query Q02 (Figure 7.7), the intention was to look for images with flowers

Figure 7.13: Top 18 retrieved images for QBSI query Q05

Figure 7.14: Top 18 retrieved images for QBSI query Q07

(c.f. `Foliage:Floral` in Figure 3.5) at the center. Among the top 18 images shown in Figure 7.12, only image 15 is irrelevant as the flower regions is considered too small.

With query Q05 (in Figure 7.8), we look for images with a spatial layout of sky, building, and ground (c.f. Figure 3.5). Only the last image in Figure 7.13 is a false positive where the greyish water was incorrectly detected as ground.

In the case of query Q07 (in Figure 7.8) that looks for small group of people appearing at the center of an image (c.f. `People:Face` in Figure 3.5), the top 18 images shown in Figure 7.14 are all found in the ground truth list for the query.

Compared to existing query formulation methods, our QBSI approach allows explicit specification of visual semantics as illustrated by the 15 queries in Figures 7.7 to 7.11. Consider the case of Query By Canvas (QBC) reviewed in Section 2.8. How would a user express visual concepts such as flowers, faces, and buildings using color and texture or their combination? Query by Sketches (QBS) (Section 2.8) is not very useful either as the shapes of flowers, faces, sky, water etc are ill-defined. Compared to the ImageScape system [Lew, 2000] that also allows placement of visual icons as query, our QBSI approach has richer expressive power as we support spatial constraints (Q01 to Q15), negation (Q14), disjunction (Q15), and concept hierarchy (Q05, Q08, Q11-13).

7.5 Query by Multiple Examples (QBME)

7.5.1 Query Processing

Query by Example (QBE) suffers from the bootstrapping problem of finding a suitable query image to start with. Nevertheless, QBE is still an intuitive and useful query method for similarity-based retrieval because it is simple to perform and is unique to image retrieval (QBE has less appeal for retrieval of other media such as text, music, and video). It is an attractive option for query formulation on a mobile device whereby more elaborate typing and drawing are usually inconvenient to perform.

Query by Multiple Examples (QBME) is a natural extension of QBE. Multiple examples are in general helpful in describing the information need involving higher

level of semantics. For instance, to describe images taken near a river or beach, different visual examples of a riverside or beach scene are necessary. As another example, image examples of buildings, street, roadside etc will be useful to convey to an image retrieval system that images of a city or urban scene are sought after. In fact, groups of query images have been considered useful contextual hints for query formulation ([Smeulders et al., 2000], p.1369).

Query processing for QBE and QBME mainly involves similarity matching. Hence all the indexing schemes proposed in this thesis that support similarity (or dissimilarity) matching (SSR, DSR, SSC, LCP) (Section 6.2) as well as the integrated similarity matching framework (Section 6.3) are applicable.

In practice, what matters is the ranking of the images returned for a QBE query, whether they are sorted in ascending order of distance measure or descending order of similarity values, based on the distance and similarity functions defined in Sections 6.2 and 6.3. However, as the indexes proposed in this thesis can be viewed as histograms, we wish to point out some resemblance to histogram intersection [Swain and Ballard, 1991]. That is, we can define content-based similarity λ between a query q (with m local blocks Z_j) and an image x (with m local blocks X_j) based on L_1 distance measure (city block distance) as

$$\lambda(q,x) = 1 - \frac{1}{2m}\sum_j \sum_i |T_i(Z_j) - T_i(X_j)| \tag{7.11}$$

This is indeed equivalent to histogram intersection with further averaging over the number of local histograms, m, except that the bins have semantic interpretation such as SSRs, DSRs, and LCPs.

There is a trade-off between content symmetry and spatial specificity. If we want images of similar semantics with different spatial arrangement (e.g. mirror images) to be treated as similar, we can have larger tessellated block in SAM (i.e. the extreme case will be a single block that covers the entire image, similar to the effect of a global histogram). However in applications where spatial locations are considered differentiating, local histograms will provide good sensitivity to spatial specificity. Furthermore, we can attach different weights to the blocks (i.e. Z_j, X_j)

to emphasize the focus of attention (e.g. center) as follows

$$\lambda(q,x) = \frac{\sum_j \omega_j \cdot \lambda(Z_j, X_j)}{\sum_k \omega_k}, \tag{7.12}$$

where ω_j are weights, and $\lambda(Z_j, X_j)$ is the similarity between two image blocks defined as

$$\lambda(Z_j, X_j) = 1 - \frac{1}{2}\sum_i |T_i(Z_j) - T_i(X_j)|. \tag{7.13}$$

Now, when a query Q has multiple examples (i.e. QBME), $Q = \{q_1, q_2, \cdots, q_K\}$, the similarity is computed as

$$\lambda(Q,x) = \max_i \lambda(q_i, x). \tag{7.14}$$

Note that if distance measure is adopted instead, then the max operator is replaced by the min operator.

7.5.2 Queries and Ground Truth

To evaluate QBME, we have designed 16 semantic queries and their ground truth (G.T.) among the 2400 test collection based on the consensus of 3 human subjects. These queries and their sizes of ground truth are listed in Table 7.6. That is, for each query, every human subject has to look through the entire collection to build the list of relevant images. Note that queries Q01-Q02, Q04-Q12 correspond to the semantic categories shown in Table 7.2.

Fig. 7.15 and 7.16 show, in top-down left-to-right order, 3 relevant images for queries Q01-Q08 and Q09-Q16 respectively. As we can see from these sample images, the relevant images for any query considered here exhibit highly varied and complex visual appearance. Hence to represent each query, the 3 human subjects selected 3 (i.e. $K = 3$ in Equation (7.14)) relevant photos as query examples for our experiments because a single query image is far from satisfactory to capture the semantic of any query. Indeed single query images have resulted in poor precisions and recalls in our initial experiments. The precisions and recalls were computed without the query images themselves in the lists of retrieved images.

Table 7.6: The 16 semantic queries used in QBME experiments

Query	Description	G.T.
Q01	indoor	994
Q02	outdoor	1218
Q03	people close up	277
Q04	people indoor	860
Q05	interior or object	134
Q06	city scene	697
Q07	nature scene	521
Q08	at a swimming pool	52
Q09	street or roadside	645
Q10	along waterside or beach	150
Q11	in a park or garden	304
Q12	at mountain area	67
Q13	buildings	239
Q14	close up, indoor	73
Q15	small group, indoor	491
Q16	large group, indoor	45

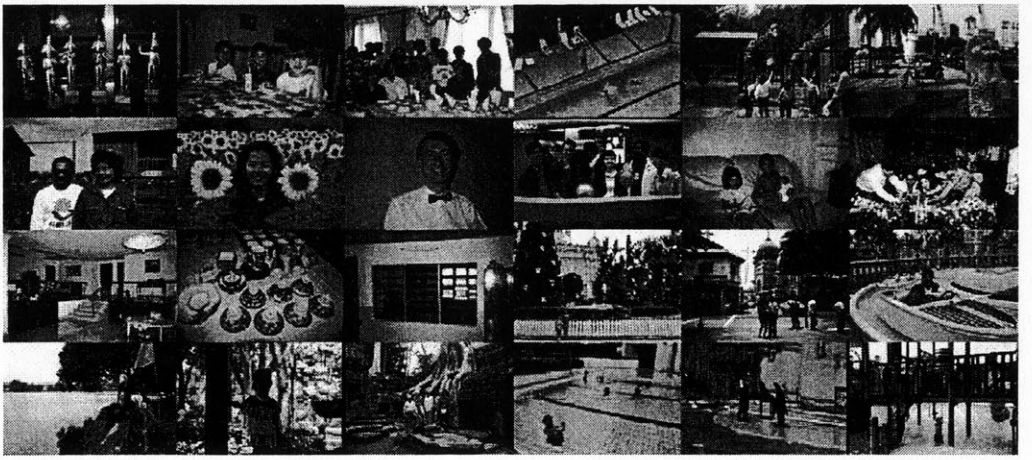

Figure 7.15: Sample consumer photos associated with queries Q01 to Q08

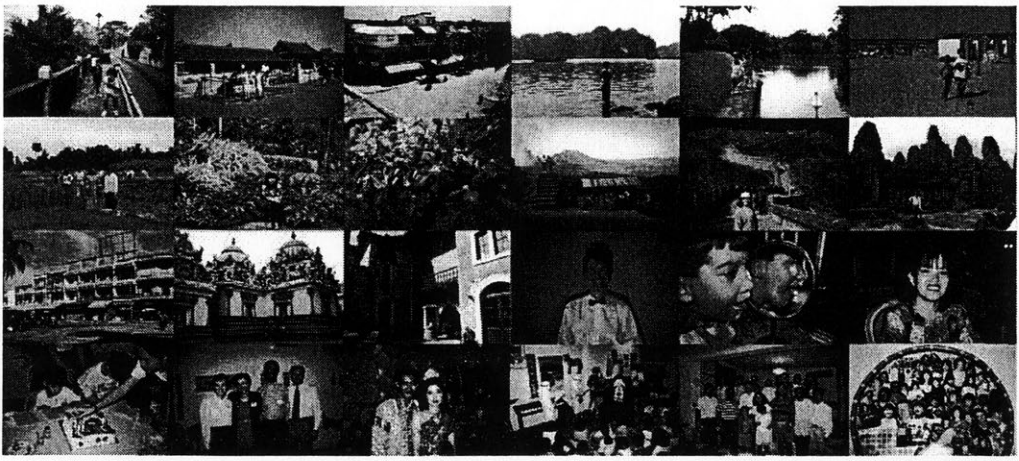

Figure 7.16: Sample consumer photos associated with queries Q09 to Q16

7.5.3 Scope of Comparison

In the QBME experiments, we compare indexing schemes and similarity integration schemes proposed in this thesis with a feature fusion approach that combines color and texture in a linearly optimal way (denoted as "CTO") in both quantitative and qualitative aspects. For each approach, we have conducted experiments with various system parameters in order to select their best performances for final comparison. Overall average precisions (denoted as P_{avg}) and average precisions at top 30 retrieved images (denoted as P_{30}) over 16 queries are used as the selection criteria.

In the next subsection, we first compare and select the best feature fusion configuration for the "CTO" approach. Next, we will look at the effects of different parameters on SSR-based indexing and retrieval. The optimal system parameters will be adopted for other indexing schemes (DSR, SSC, LCP) when applicable. In the subsection that follows, the best weighting coefficients used in similarity integration for the semantics design (denoted as "Dsgn") and semantics discovery (denoted as "Dscv") approaches will be determined. Lastly, we compare the indexing schemes (SSR, DSR, SSC, LCP) and similarity integration schemes (Dsgn and Dscv) with feature fusion approach (CTO) quantitatively and qualitatively in the last two subsections respectively.

7.5.4 Indexing based on Fusion of Color and Texture

For the similarity-based retrieval experiments in this thesis, we have decided to compare our proposed semantic indexing and integrated matching solutions with the fusion of color and texture features instead of other methods such as region-based matching for two reasons. First our initial attempt with region segmentation using 500 outdoor images [Wu et al., 2000a] does not scale up on the 2400 collection. Indeed, very high dimensions are required for the CTO approach to produce reasonable performance as we shall see below. Next we have adopted similar color and texture features for the CTO approach to demonstrate the advantage gained from the abstraction layer (mid-level) introduced by our indexing schemes.

For the color-based signature, both global and local (4×4 grid) color histograms of b^3 ($b = 4, 5, \cdots, 17$) number of bins in the RGB color space as well as the HSV-11 histogram [Cox et al., 2000] described in Section 7.3.3 are computed on an image. Their retrieval performances in terms P_{avg} (P_{30}) over 16 queries are listed in Table 7.7). In the case of global color histograms, the performance saturated at 4096 ($b = 16$) and 4913 ($b = 17$) bins with $P_{avg} = 0.363$ and $P_{30} = 0.577$. Hence the one that used fewer number of bins is preferred. Among the local color histograms attempted, the one with 2197 bins ($b = 13$) gives the best average precisions with $P_{avg} = 0.381$ and $P_{30} = 0.598$. Histogram intersection [Swain and Ballard, 1991] is used to compare two color histograms.

For the texture-based signature, we have adopted the mean and standard deviation of Gabor coeffients and the associated distance measure (Equation (3.12)) [Manjunath and Ma, 1996]. The Gabor coefficients are computed with 5 scales and 6 orientations. Convolution windows of $20 \times 20, 30 \times 30, \cdots, 60 \times 60$ are attempted. Similarly, both global and local (4×4 grid) signatures were experimented. Table 7.8 shows their average precisions P_{avg} (P_{30}) over 16 queries. The best results are obtained when 20×20 windows are used. We obtained $P_{avg} = 0.251$ and $P_{30} = 0.300$ for global signatures and $P_{avg} = 0.235$ and $P_{30} = 0.379$ for local signatures. These inferior results when compared to those of color histograms lead us to conclude that simple statistical texture descriptor is less effective than color histogram for heterogeneous consumer image contents.

The color and texture signatures are combined for image matching and retrieval

Table 7.7: Average precisions by global and local color histograms

b	Label	Global	Local
	HSV-11	0.290 (0.425)	0.353 (0.508)
4	RGB-64	0.305 (0.490)	0.345 (0.548)
5	RGB-125	0.323 (0.467)	0.357 (0.538)
6	RGB-216	0.332 (0.498)	0.367 (0.573)
7	RGB-343	0.340 (0.527)	0.367 (0.575)
8	RGB-512	0.350 (0.527)	0.372 (0.567)
9	RGB-729	0.353 (0.540)	0.376 (0.585)
10	RGB-1000	0.356 (0.548)	0.377 (0.579)
11	RGB-1331	0.357 (0.563)	0.377 (0.588)
12	RGB-1728	0.357 (0.548)	0.379 (0.596)
13	RGB-2197	0.359 (0.556)	0.381 (0.598)
14	RGB-2744	0.359 (0.560)	0.379 (0.592)
15	RGB-3375	0.361 (0.569)	0.380 (0.590)
16	RGB-4096	0.363 (0.577)	0.380 (0.596)
17	RGB-4913	0.363 (0.577)	0.381 (0.590)

Table 7.8: Average precisions by global and local texture histograms

Window Size	Label	Global	Local
20×20	Gabor-20	0.251 (0.300)	0.235 (0.379)
30×30	Gabor-30	0.250 (0.296)	0.234 (0.367)
40×40	Gabor-40	0.248 (0.294)	0.234 (0.358)
50×50	Gabor-50	0.239 (0.319)	0.234 (0.348)
60×60	Gabor-60	0.247 (0.263)	0.233 (0.342)

as follows. The distance measures between a query and an image for the color and texture methods are first normalized within $[0, 1]$ and then combined linearly, similar to that shown in Equation (6.14) with color and texture matching replacing the roles of μ and λ respectively. Among the relative weights ω attempted at regular 0.1 intervals (Table 7.9), the best fusion is obtained at $P_{avg} = 0.38$ and $P_{30} = 0.61$ with equal color influence and texture influence for global signatures. In the case of local signatures, the fusion peaked when the local color histograms are given a dominant influence of 0.9 resulting in $P_{avg} = 0.38$ and $P_{30} = 0.59$. As shown in the last row of the table, almost identical performance values are obtained when a multiplicative fusion operator (Equation (6.13)) is used. Hence for performance comparison below, CTO approach will be represented by the fusion of global signatures RGB-4096 and Gabor-20 (0.38 (0.61)).

Table 7.9: Average precisions by fusion of global/local color/texture similarities

ω	Global (RGB-4096, Gabor-20)	Local (RGB-2197, Gabor-20)
0.1	0.304 (0.496)	0.275 (0.483)
0.2	0.340 (0.558)	0.304 (0.554)
0.3	0.364 (0.577)	0.328 (0.565)
0.4	0.378 (0.594)	0.347 (0.581)
0.5	0.384 (0.606)	0.362 (0.583)
0.6	0.384 (0.600)	0.373 (0.594)
0.7	0.381 (0.592)	0.379 (0.585)
0.8	0.376 (0.581)	0.382 (0.588)
0.9	0.370 (0.583)	0.383 (0.594)
(*)	0.384 (0.606)	0.383 (0.592)

7.5.5 Indexing based on SSRs

To begin with, we compare different matching functions and spatial aggregation templates in Table 7.10. The matching functions compared are: L_1-norm city block distance ("CBD"), L_2-norm Euclidean distance ("EUD"), Kullback-Leibler distance ("KLD"), symmetric Kullback-Leibler distance ("KLS"), and cosine similarity ("COS"). For the spatial aggregation templates, "Center-Focus" denotes a weighted tessellation (Equation (7.12)) that focuses at the center of image as illus-

trated in Figure 7.17. The weights ω_j are shown below each block in the figure.

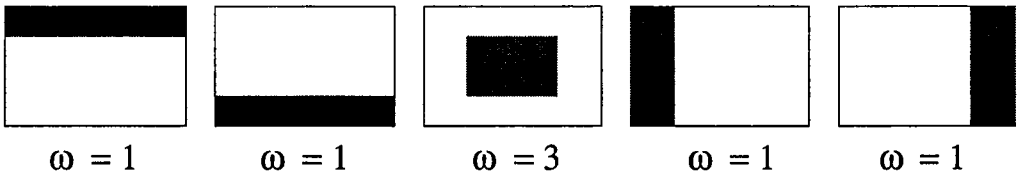

Figure 7.17: A spatial aggregation template that focuses at image center

An evenly weighted "4 × 4 Grid" template is also attempted. The SVM learning of the 26 SSRs (Figure 3.5, Table 3.2) is based on polynomial kernel ($Poly_2$) with modified similarity function (Equation 3.27) on composite feature vector z (z^c and z^t) (c.f. Sections 3.3.3 and 3.3.4). The softmax function (Equation (3.24)) was adopted for the normalization of SVM classification outputs.

Table 7.10: Average precisions by different matching functions and spatial aggregation templates

	Center-Focus	4 × 4 Grid
CBD	0.415 (0.656)	0.440 (0.679)
EUD	0.369 (0.588)	0.381 (0.640)
KLD	0.371 (0.575)	0.358 (0.573)
KLS	0.417 (0.638)	0.406 (0.650)
COS	0.401 (0.558)	0.399 (0.579)

From Table 7.10, we conclude that city block distance (CBD) is most effective in terms of average precisions P_{avg} and P_{30}. Furthermore, on average, spatial aggregation template based on a 4 × 4 grid is preferred (especially so for the case of CBD). Hence for the rest of our experiments, a 4 × 4 grid with CBD is adopted as the spatial aggregation template and matching function respectively.

Next we compare the two different classification normalization schemes proposed in Section 3.3.2, namely the softmax scheme (Equation (3.24)) and the hybrid winner-take-all and softmax scheme. As shown in Table 7.10, the performance indicators P_{avg} and P_{30} for the softmax scheme are 0.44 and 0.68 respectively. In the case of the hybrid scheme, better values of 0.45 and 0.70 have been achieved for P_{avg} and P_{30} respectively. Hence the hybrid scheme is used for the rest of the experiments.

Last but not least, we compare retrieval performance of SSR-based indexes using different classifiers as shown in Table 7.11. "SICA-167" and "SICA-226" refer to the alternative incremental learning algorithm described in Section 3.8 trained on 375 and 554 region samples respectively. The numbers 167 and 226 are the number of prototypes created for the hidden layer after learning. "SVM-C" and "SVM-T" are SVM classifiers (polynomial kernel of degree 2) trained on only color (z^c) and texture (z^t) feature vectors respectively. Their classification outputs are combined via voting to form "SVM-M". The last three rows refer to the SVM classifiers based on different kernel functions as described in Section 3.3.4.

Table 7.11: Average precisions by different classifiers

	P_{avg}	P_{30}
SICA-167	0.392	0.631
SICA-226	0.395	0.623
SVM-C	0.412	0.619
SVM-T	0.218	0.298
SVM-M	0.417	0.623
$Poly_2$	0.454	0.696
$Poly_6$	0.441	0.648
RBF_1	0.422	0.592

From Table 7.11, we observe that the discriminative power of the SICA classifiers are slightly worse off than the SVM classifiers. Color feature is more effective for retrieval than the texture feature for the heterogeneous consumer images in our test collection (i.e. SVM-C versus SVM-T). Post-classification fusion by SVM-M does not improve the performance further. Last but not least, though $Poly_6$ and RBF_1 classifiers have better generalization performance on test set for region learning (c.f. Section 3.3.4), the computationally simpler $Poly_2$ has turned out to attain the best retrieval performance for our queries and data set.

In summary, the best SSR-based indexing scheme for similarity-based retrieval using the 16 queries on 2400 consumer images uses SVM classifiers with degree 2 polynomial kernel to learn 26 SSR classes from 375 training samples. The SSRs are detected and normalized using a hybrid winner-take-all and softmax scheme, reconciled and aggregated according to a 4×4 grid template, and similarity matching

is performed based on city block distance. The same parameters are adopted for the indexes based on DSRs, SSCs, and LCPs whenever applicable for consistency in the experiments and performance comparison below.

7.5.6 Similarity Integration

The retrieval performances for indexing schemes (SSR, DSR, SSC, and LCP) as well as similarity integation schemes (Dsgn and Dscv) will be shown and compared to CTO in the next subsection. In this subsection, we show how the best performances of Dsgn and Dscv are selected based on different parameter values in similarity integation.

Table 7.12 lists P_{avg} and P_{30} for the Dsgn approach that is based on the integration of SSR and SSC similarities. The results of linear combination (Equation (6.14)) are obtained with ω computed at 0.1 intervals. The last row shows the result of multiplicative fusion (Equation (6.13)). Note that μ refers to the SSC similarity and λ is the SSR similarity. The best performance is obtained with $P_{avg} = 0.59$ and $P_{30} = 0.78$ using the linear combination with $\omega = 0.5$, suggesting equal importance of both intra-content and inter-class similarities.

Table 7.12: Average precisions by integration of SSR and SSC similarities

ω	P_{avg}	P_{30}
0.1	0.516	0.746
0.2	0.557	0.765
0.3	0.579	0.781
0.4	0.589	0.783
0.5	0.590	0.777
0.6	0.585	0.785
0.7	0.575	0.767
0.8	0.561	0.742
0.9	0.546	0.719
(*)	0.568	0.769

Similarly, Table 7.13 lists P_{avg} and P_{30} for the Dscv approach that is based on the integration of DSR and LCP similarities. The results of linear combination with ω computed at 0.1 intervals are listed followed by that of multiplicative fusion. Note

that LCP similarity plays the role of μ and DSR similarity is used for λ. Once again, equal contribution of both intra-content and inter-class similarities in linear combination has resulted in the best performance with $P_{avg} = 0.52$ and $P_{30} = 0.76$.

Table 7.13: Average precisions by integration of DSR and LCP similarities

ω	P_{avg}	P_{30}
0.1	0.494	0.700
0.2	0.508	0.708
0.3	0.517	0.731
0.4	0.522	0.744
0.5	0.523	0.760
0.6	0.521	0.756
0.7	0.515	0.754
0.8	0.505	0.748
0.9	0.492	0.719
(*)	0.522	0.752

7.5.7 Quantitative Comparison

We compare the best QBME performances of the indexing schemes (SSR, DSR, SSC, and LCP) and the similarity integation schemes (Dsgn and Dscv) against that of the CTO method in this subsection.

First, we compare CTO with Dsgn (semantics design approach). The Precision/Recall curves (averaged over 16 queries) for CTO, SSR, SSC, and Dsgn in Figure 7.18 illustrate the improvement at various recall values of the Dsgn methods over the CTO method. Table 7.14 shows the average precisions among the top 20, 30, 50, and 100 retrieved images as well as the overall average precisions for the methods compared including individual SSR and SSC indexing. The relative improvements (in percentage) of Dsgn over CTO are also shown in the last column.

In a similar manner, Figure 7.19 shows the Precision/Recall curves (averaged over 16 queries) for CTO, DSR, LCP, and Dscv and Table 7.15 lists the average precisions among the top 20, 30, 50, and 100 retrieved images as well as the overall average precisions for these methods. The relative improvements (in percentage) of Dscv over CTO are shown in the last column.

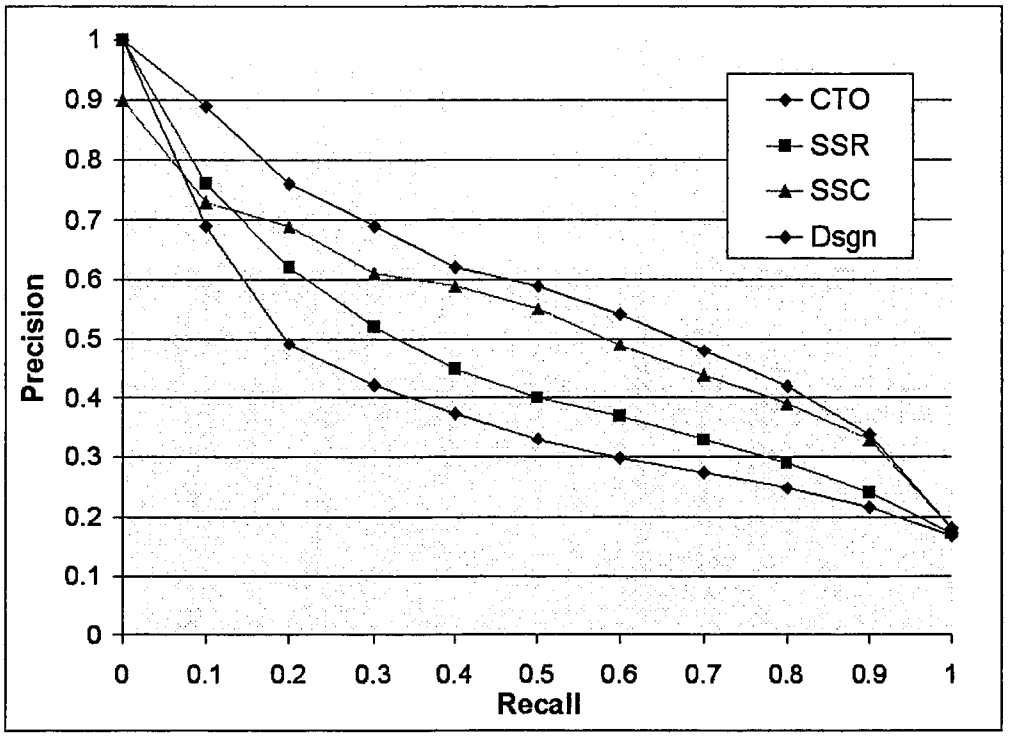

Figure 7.18: Precision/Recall curves for CTO, SSR, SSC, and Dsgn

Table 7.14: Average precisions at top retrieved images (CTO, SSR, SSC, Dsgn)

Avg.Prec.	CTO	SSR	SSC	Dsgn	%
At 20	0.65	0.76	0.71	0.84	29
At 30	0.61	0.70	0.68	0.78	28
At 50	0.55	0.62	0.64	0.72	31
At 100	0.49	0.54	0.58	0.65	33
overall	*0.38*	*0.45*	*0.53*	*0.59*	*55*

Table 7.15: Average precisions at top retrieved images (CTO, DSR, LCP, Dscv)

Avg.Prec.	CTO	DSR	LCP	Dscv	%
At 20	0.65	0.71	0.70	0.80	23
At 30	0.61	0.68	0.69	0.76	25
At 50	0.55	0.63	0.63	0.70	27
At 100	0.49	0.57	0.58	0.62	27
overall	*0.38*	*0.48*	*0.48*	*0.52*	*37*

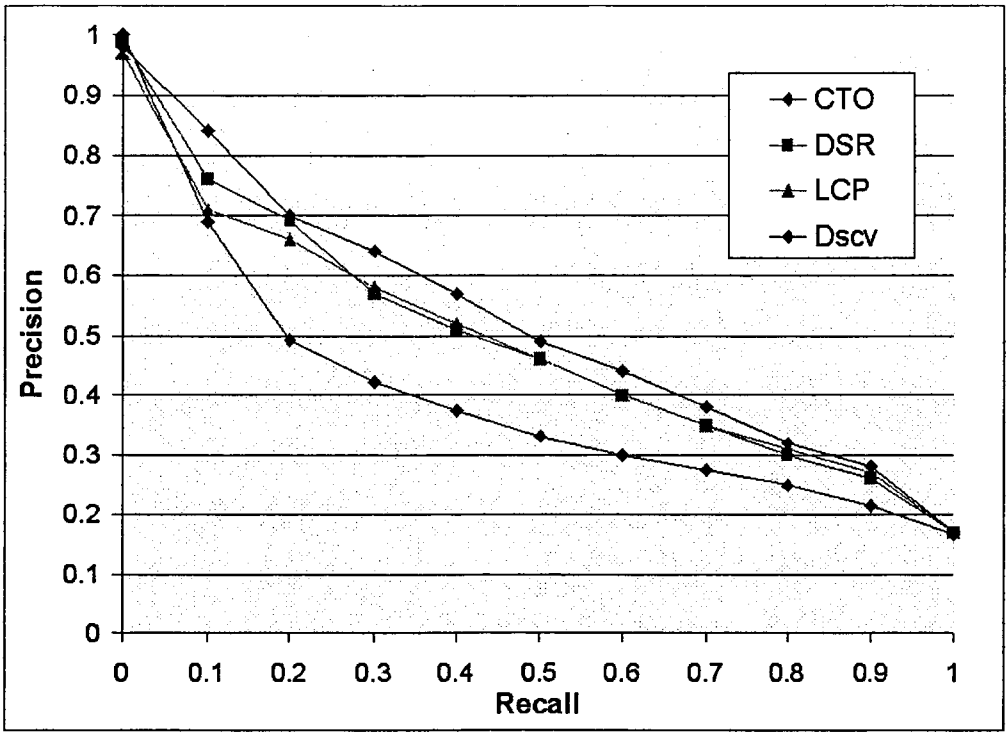

Figure 7.19: Precision/Recall curves for CTO, DSR, LCP, and Dscv

From Figures 7.18 and 7.19 as well as Tables 7.14 and 7.15, we can see that the proposed indexing methods (SSR, SSC, DSR, LCP) and similarity integration schemes (Dsgn, Dscv) outperform the feature fusion approach CTO significantly. The integration of intra-content (SSR, DSR) and inter-class (SSC, LCP) similarities also achieves better retrieval performance than individual indexing methods.

To better contrast the performance differences between the Dsgn and Dscv approaches with the CTO approach, their Precision/Recall curves (averaged over 16 queries) are plotted as shown in Figure 7.20 and their average precisions for top retrieved images are consolidated in Table 7.16.

Figure 7.21 compares the average precisions of each of the 16 queries for CTO, Dsgn, Dscv methods. The random retrieval method (i.e. $G.T./2400$) (denoted as "RND") is also included as a baseline comparison. The curves are plotted based on the descending precision values of the RND method, indicating the increasing difficulty of the queries. The actual precision values are listed in Table 7.17.

In a nutshell, our proposed approaches Dsgn and Dscv achieved high average precisions of 0.59 and 0.52 respectively, which are significant improvements of 55%

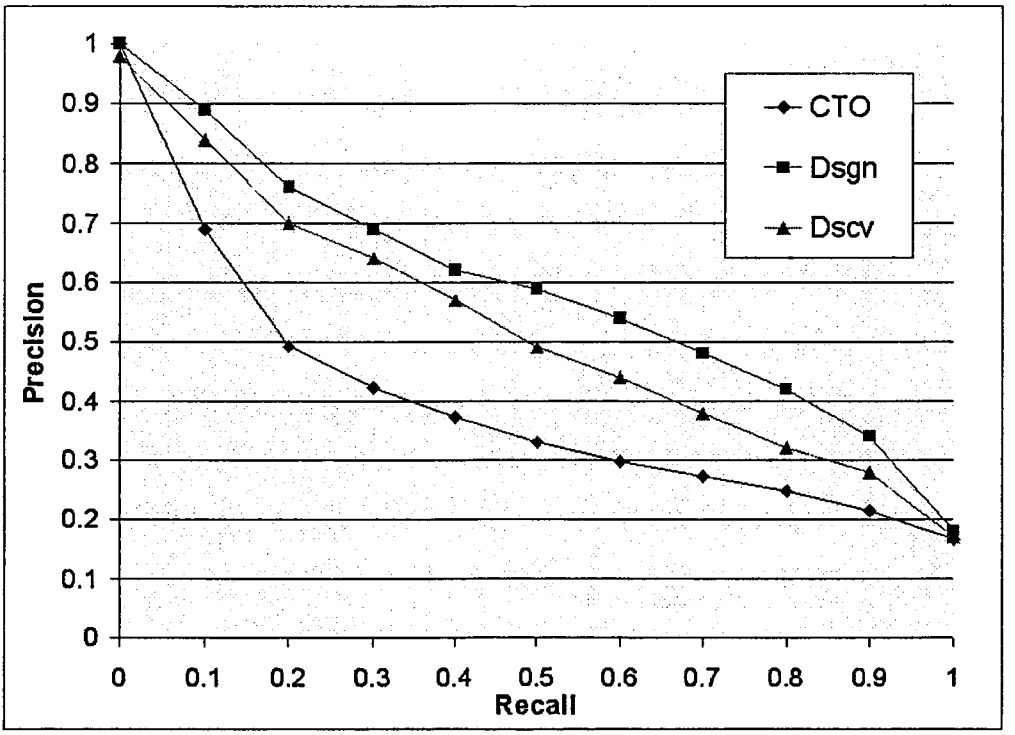

Figure 7.20: Precision/Recall curves for CTO, Dsgn, and Dscv

Table 7.16: Average precisions at top retrieved images (CTO, Dsgn, Dscv)

Avg.Prec.	CTO	Dsgn	Dscv
At 20	0.65	0.84	0.80
At 30	0.61	0.78	0.76
At 50	0.55	0.72	0.70
At 100	0.49	0.65	0.62
overall	*0.38*	*0.59*	*52*

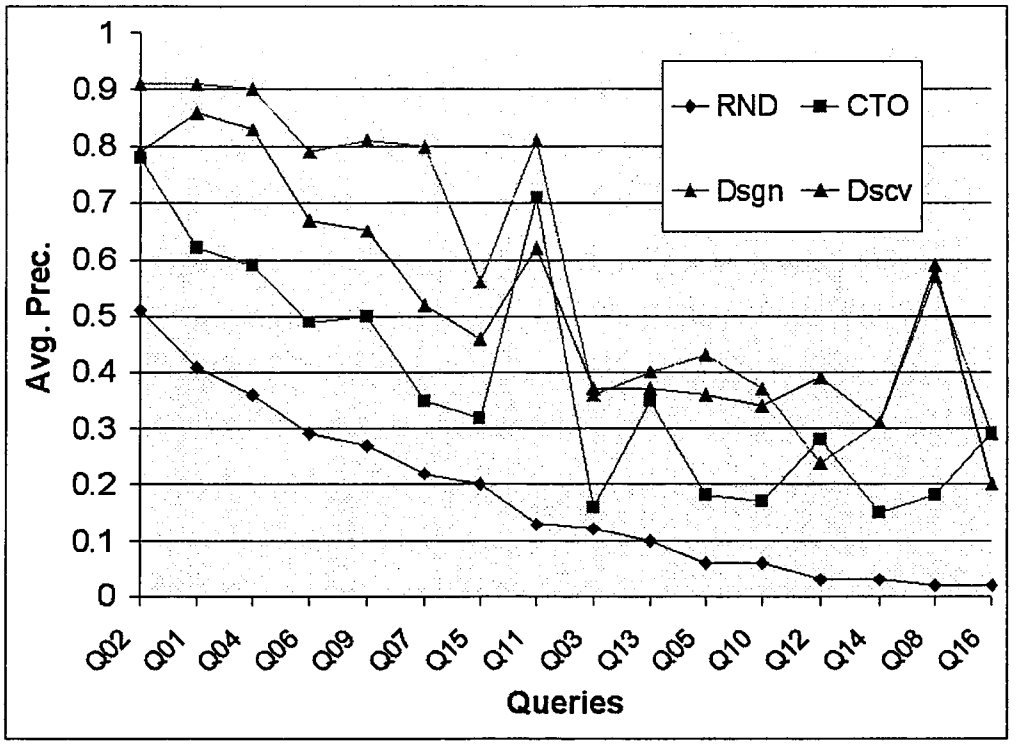

Figure 7.21: Average precisions of each query for RND, CTO, Dsgn, and Dscv

Table 7.17: Average precisions for each of the 16 queries

Query	Description	RND	CTO	Dsgn	Dscv
Q01	indoor	0.41	0.62	0.91	0.86
Q02	outdoor	0.51	0.78	0.91	0.79
Q03	people close-up	0.12	0.16	0.36	0.37
Q04	people indoor	0.36	0.59	0.90	0.83
Q05	interior or object	0.06	0.18	0.43	0.36
Q06	city scene	0.29	0.49	0.79	0.67
Q07	nature scene	0.22	0.35	0.80	0.52
Q08	at a swimming pool	0.02	0.18	0.57	0.59
Q09	street or roadside	0.27	0.50	0.81	0.65
Q10	waterside or beach	0.06	0.17	0.37	0.34
Q11	in a park or garden	0.13	0.71	0.81	0.62
Q12	at mountain area	0.03	0.28	0.24	0.39
Q13	building close-up	0.10	0.35	0.40	0.37
Q14	portrait, indoor	0.03	0.15	0.31	0.31
Q15	small group, indoor	0.20	0.32	0.56	0.46
Q16	large group, indoor	0.02	0.29	0.29	0.20

and 37% over that of the CTO method (last row of Table 7.16). The performance gap is consistently evident across various recall values in the Precision/Recall curves as shown above. Indeed both Dsgn and Dscv outperformed CTO in all except two queries (Q12 and Q16 (tie) for Dsgn; Q11 and Q16 for Dscv) in average precisions as seen in Table 7.17.

On the other hand, the small footprint of the proposed image indexes also has an added advantage in storage space and retrieval efficiency. Suppose a 4-byte floating point number is required for each soft classification output used in the Dsgn and Dscv methods. Then a Dsgn or Dscv image index requires less than 2 kilobytes of storage and simple operations on small number of vectors. This would have great advantage over the need to represent and process very high dimension of color and texture features and yet not achieving the same level of retrieval performance.

In summary, the proposed image indexes (SSR, SSC, DSR, and LCP) and their similarity integration realize semantic abstraction via prior learning and detection of visual classes when compared to direct indexing based on low-level features. The compact representation that accommodates imperfection and uncertainty in detection also resulted in much better performance than the optimal fusion of very high dimension of color and texture features in our QBME experiments using 16 semantic queries on 2400 unconstrained consumer photos. Hence we feel that the computational resources devoted to prior learning of visual classes and their detection during indexing are good trade-off for concise semantic representation as well as effective and efficient retrieval performance.

7.5.8 Qualitative Comparison

In this subsection, using the top retrieved images for 3 of the 16 QBME queries, we provide a qualitative comparison between the CTO approach and our proposed approach. For the latter, as Dsgn (semantics design approach) has the most favorable performance, it is chosen as the candidate for comparison. In the following, the top 18 images retrieved by each method (excluding the queries images) are shown in top-down, left-to-right order of decreasing relevance as computed by each method. In particular, we have selected queries Q08, Q10 and Q14 to illustrate the comparison as they refer to semantic categories of images have not been considered in previous

study [Vailaya et al., 2001].

For Q08 (Figure 7.22), the intention was to look for images taken at pool areas. The query images include a swimming pool, an sea lion show at a zoo, and a water playground with wooden structure.

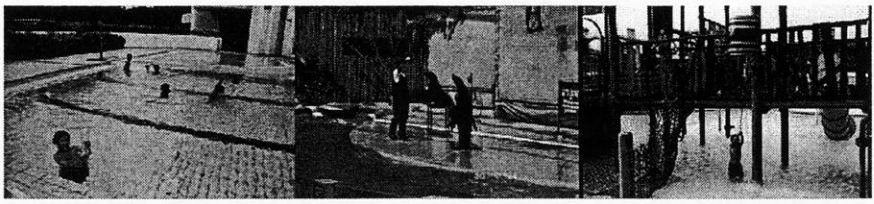

Figure 7.22: Query Q08 "at a swimming pool"

Figure 7.23: Top 18 retrieved images for query Q08 by CTO

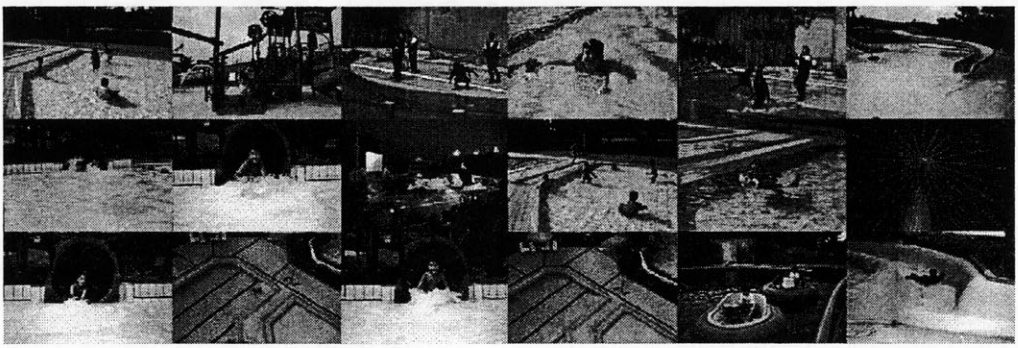

Figure 7.24: Top 18 retrieved images for query Q08 by Dsgn

In Figure 7.23, only 8 images $(1-3, 5, 10-11, 14, 18)$ are part of the grouth truth list of Q08. Apparently without semantic interpretation, CTO had assigned high similarity values to images that share similar color and texture profiles with the second and third query images, hence resulting in many top irrelevant images.

In particular, images 4, 6, 8, 13, 16 are indoor images. In the case of Dsgn approach, all the top 18 images shown in Figure 7.24 except image 12 are relevant. The good result can be attributed to the detection of the SSR `Water:Pool` (c.f. Table 3.2) though the blue tower in image 12 has resulted in false detection.

In the case of Q10 (Figure 7.25), relevant images should be those taken at the river or lake (first query image), pond (second query image), beach (last query image).

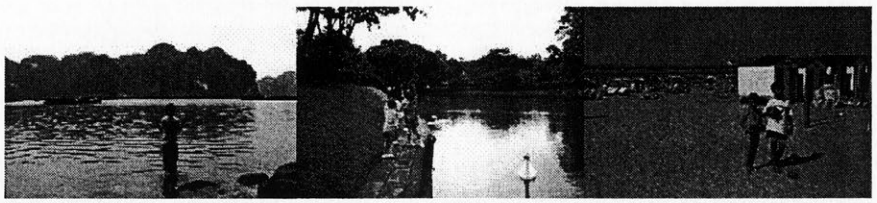

Figure 7.25: Query Q10 "along waterside"

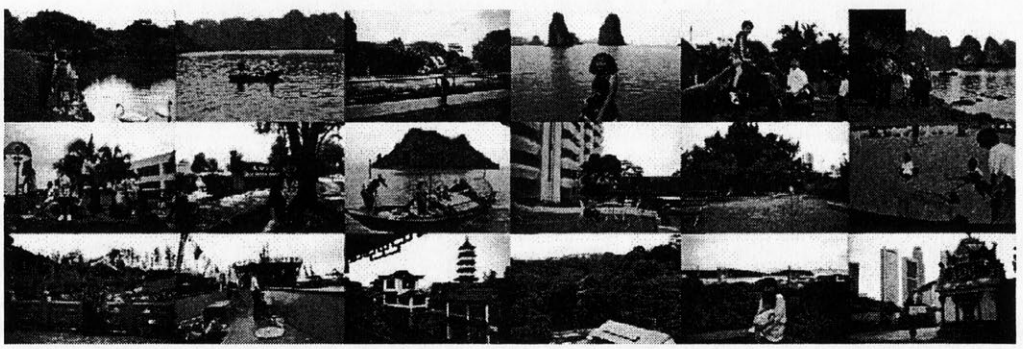

Figure 7.26: Top 18 retrieved images for query Q10 by CTO

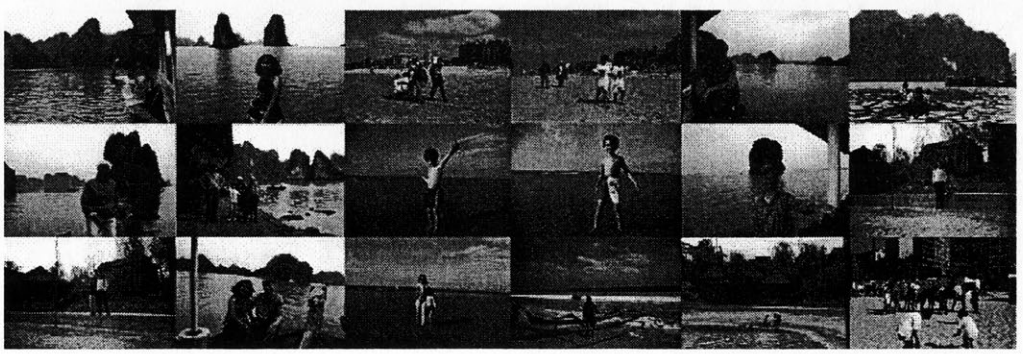

Figure 7.27: Top 18 retrieved images for query Q10 by Dsgn

From Figure 7.26, we observe that the CTO approach only retrieved 6 relevant

images ($1-2, 4, 6, 9, 12$). In this case, the colors green and grey with slight texture had resulted in the false positives (images $3, 5, 7-8, 10-11, 13-18$). On the other hand, the Dsgn method produced all relevant images at the top 18 images as shown in Figure 7.27. These images are the response to the first and last query images. The response to second query image follows these images but they are outside the top 18 list.

For query Q14 (Figure 7.28), the three query images depict typical images of people close-up indoor: woman with dark background, man in front of wall, two men in front of bright background.

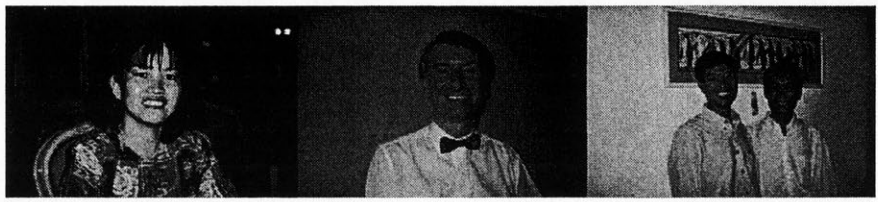

Figure 7.28: Query Q14 "people close-up indoor"

Figure 7.29: Top 18 retrieved images for query Q14 by CTO

The CTO approach retrieved 10 relevant images ($1-6, 8-10, 16$) for query Q14 (Figure 7.29) while the Dsgn approach only returned 2 irrelevant images ($14-15$) among the top 18 images (Figure 7.30). With a modular framework that can incorporate object detectors such as face detector seemlessly, the performance of the Dsgn approach can be further improved whenever more accurate object detectors become available from the computer vision research community.

Figure 7.30: Top 18 retrieved images for query Q14 by Dsgn

Chapter 8

Conclusion

The best way to predict the future is to invent it
Alan Kay

8.1 Contributions

As described in Chapter 1, the research conducted and presented in this thesis has been motivated from three axes:

- Broad Consumer Images (Section 1.1.1)
  The proliferation of digital cameras and camera phones calls for solutions to address the genuine problem of organizing and accessing voluminous consumer images. User studies have shown that manual annotation, either typed or spoken, is neither desirable nor comprehensive, hence ineffective for semantic retrieval. Thus content-based indexing and search is necessary on its own, especially useful for finding images sharing some common visual attributes, and for complementing browsing and facilitate annotation. In addition, unconstrained consumer images pose great challenges for content-based retrieval research due to their content variations and visual complexities;

- "Keywords" in Visual Data (Section 1.1.2)
  While keywords are simple, relatively effective, and practical for indexing and

retrieval of text documents, the equivalent counterpart for visual data is neither well-studied nor well-understood. This is because pixel-based visual data are ill-defined and in general underconstrained for visual perception and understanding. Based on common sense description of a scenery image, both visual type and locality information are considered important for visual data description. A computational scheme is sought after formalizing and automating spatial semantic indexing for effective retrieval;

- Semantic Gap (Section 1.1.3)
  While low-level features such as color, texture, and shapes can be computed from images, extraction of high-level semantic information needed for effective query interpretation remains a challenge for computer vision. This semantic gap is the manifestation of the semantics extraction and interpretation problems. Extraction of complete semantics from image data is hard as robust segmentation and general object recognition for broad domain images are unsolved problems. On the other hand, user intrepretation of queries and images is usually complex, subjective and ambiguous due to differences in tasks, gender, culture, education background, etc.

In Chapter 2, after clarifying the difference between pattern classification and image retrieval in Section 2.1, the key developments in text-based retrieval, feature-based retrieval, region-based retrieval, object-based retrieval, and probabilistic retrieval are reviewed in Sections 2.2 to Section 2.6 respectively. Significant works in other related research areas such as image classification, query formulation, feature fusion, and automatic annotation are also covered in Sections 2.7 to 2.10 respectively.

As pointed out in Section 1.2, before embarking on the research described in this thesis, we have pursued content-based image retrieval research along the direction of unsupervised learning [Lim, 1999a] [Lim, 1999b] [Lim, 1999c] [Lim, 2000c] [Lim, 2000d] as well as handcrafted construction of Visual Keywords [Lim, 2000a] [Wu et al., 2000a] [Lim, 2000b] [Lim, 2001b] [Lim, 2001a]. The process of creating image signature based on the Visual Keywords methodology has also been patented [Lim, 2003]. With the encouraging results obtained from these efforts, we are motivated to extend and deepen the framework substantially.

In this thesis, through the descriptions from Chapter 3 to Chapter 7, we have proposed a suite of technical solutions to address the research challenges for content-based image retrieval as listed in Section 1.1.4. The research results, including previous work (1999 to 2001) and current extensions, are published (and submitted) as 1 patent, 4 book chapters, 5 journal articles, and 27 conference papers.

In a nutshell, we have conceptualized and presented dual cascaded learning frameworks [Lim and Jin, 2003c] [Lim and Jin, 2004f] that integrate both local and global semantics, namely a semantics design approach and a semantics discovery approach. More specifically, our original research contributions are listed as follows:

- **Semantics Design**

  The semantics design framework [Lim and Jin, 2004a] [Lim and Jin, 2004b] [Lim and Jin, 2004i] provides a structured methodology to design, learn, and detect image semantics for building content-based image indexing and retrieval systems. Within the framework, two complementary indexing schemes have been proposed:

  - *Semantic Support Regions* (SSRs)

    In this local indexing scheme (Chapter 3), hierarchy of visual concepts called Semantic Support Regions (SSRs) are designed and constructed using statistical learning algorithms such as support vector machines [Lim et al., 2003b] [Lim and Jin, 2003b] [Lim and Jin, 2004h] and supervised incremental clustering algorithms (SICA) [Lim, 1993] [Lim, 1996] [Lim and Jin, 2002a] [Lim and Jin, 2002b] from labeled image blocks. During indexing, the learned SSRs are detected from tessellated image blocks of multiple resolutions. The soft detection decisions are reconciled and aggregated spatially as semantic histograms for image matching and retrieval.

  - *Semantic Support Classes* (SSCs)

    Instead of conventional image classification, the class memberships of a given image with respect to pre-defined image categories known as Semantic Support Classes (SSCs) are computed as a form of global image index and used in similarity matching (Section 5.3) [Lim and Jin, 2004d].

The SSCs can be viewed as prototypical categories for a content domain to anchor the context of a query and a database image for similarity-based retrieval (Chapter 6) [Lim and Jin, 2004i].

- **Semantics Discovery**

  The semantics discovery framework proposed in Chapter 4 [Lim and Jin, 2004c] [Lim and Jin, 2004d] [Lim and Jin, 2004g] is a new attempt to minimize human annotation effort in the construction of semantic image indexing and retrieval systems. The framework uses class-labeled only images to bootstrap recurrent intra-class and discriminative inter-class semantic regions. Within the framework, two complementary indexing schemes have been developed:

  - *Discoverd Semantic Regions* (DSRs)

    The Discoverd Semantic Regions (DSRs) induced from the semantic discovery framework are modeled using discriminative statistical learning algorithm (SVM) to form local visual detectors (Chapter 4). Playing the role of SSRs, equivalent multi-scale DSR detection, reconciliation, and aggregation steps on an image are used to form semantic index for matching and retrieval [Lim and Jin, 2004d] [Lim and Jin, 2004g].

  - *Local Class Patterns* (LCPs)

    The discriminative classifiers learned from class-labeled images in the semantic discovery framework are applied to local image block classification to form Local Class Patterns (LCPs) (Section 5.4). Similar to SSR-based and DSR-based indexing, LCP detection vectors on an image are reconciled and aggregated as semantic index for similarity matching [Lim and Jin, 2004c].

- **Learning and Integration**

  A common theme that runs across the proposed solutions in this thesis is learning and integration. All the proposed indexing schemes based on SSRs (Chapter 3), DSRs (Chapter 4), SSCs (Section 5.3), and LCPs (Section 5.4) as well as event modeling for event-based retrieval (Section 5.2) are all built upon statistical learning in a modular manner. In addition, information integration

approach motivated from Bayesian probability theory is proposed to combine similarities resulting from both intra-image and inter-class index matching (Chapter 6) [Lim and Jin, 2004a] [Lim and Jin, 2004b] [Lim and Jin, 2004i]. The integrated similarity matching scheme has been shown to be more superior than individual image index in similarity-based retrieval experiments (Section 7.5).

- **Representing and Detecting Visual Semantics for Indexing**
  A framework to represent and detect mid-level visual semantics for image indexing has been proposed in the thesis [Lim and Jin, 2002b] [Lim et al., 2003b] [Lim and Jin, 2004e] [Lim and Jin, 2004h]. It has the following innovative aspects:

  - *Just-In-Time Feature Fusion*
    Instead of Early or Late Feature Fusion approaches, we have proposed and implemented the fusion of color and texture features in the SVM kernel functions (Section 3.3.3). The method is more effective than other Early Fusion methods compared in the SSR learning experiments without the drawbacks of the Late Fusion methods as pointed out.

  - *Multi-Scale Segmentation-Free Indexing*
    A novel image indexing framework based on learning and detection of visual concepts from tessellated image blocks without region segmentation has been proposed (Sections 3.4, 3.5, 3.6) [Lim and Jin, 2004h]. The processing steps are inherently parallel. They allow detection outcomes from multiple resolutions to be reconciled and aggregated according to flexible spatial configuration.

  - *Abstraction Hierarchy*
    The representation of visual concepts for detection and indexing can be extended with IS-A and Part-Whole hierarchies to include more abstract visual concepts and more complex visual objects respectively (Section 3.7) [Lim et al., 2003c]. A two-level visual concept hierarchy (e.g. Sky:Clear/Cloudy/Blue) has been demonstrated in the thesis.

- *Post-Detection Segmentation*

  As a by-product of detection-based indexing, we have proposed an unconventional region segmentation algorithm (Section 3.9). Instead of the popular segmentation-recognition processing flow, an incremental clustering algorithm operating in the space of detection vectors from tessellated image blocks to form coherent coarse-grained object regions has been developed and tested [Lim, 2001b].

- **Query Formulation and Processing**

  As an attempt to reduce the ambiguity and subjectivity in query interpretation, three query formulation and associated query processing methods have been proposed in the thesis.

  - *Query by Class/Event*

    Query by Class/Event (QBCE) supports query at high-level semantics using predefined image class or event labels (Section 7.3.1). An event-based retrieval framework proposed in Section 5.2 is used to support this kind of query [Lim and Jin, 2002c] [Lim and Jin, 2003a] [Lim and Jin, 2003b]. More specifically, Events, such as those shown in the event taxonomy described in Section 5.2, are modeled based on statistical learning and a winner-take-all approach is used to compute the relevance score of an image for a query event.

  - *Query by Spatial Icons*

    Query by Spatial Icons (QBSI) allows visual query formulation based on spatial arrangement of visual icons, representing predefined local visual semantics (Section 7.4.1) [Lim and Jin, 2004e] [Lim and Jin, 2004h]. Supported by the semantic indexing schemes such as SSRs, query processing for QBSI involves both pattern-based and logic-based computation. Query with mixed levels of visual semantics can also be supported.

  - *Query by Multiple Examples*

    Query by Multiple Examples (QBME) is a natural extension of conventional QBE (Section 7.5.1). Multiple query images are useful contextual

hints for query formulation. Query processing for QBME involves similarity matching, hence it is supported by all the local and global indexing schemes and similarity integration schemes proposed in this thesis [Lim and Jin, 2004a] [Lim and Jin, 2004b] [Lim and Jin, 2004c].

- **Empirical Evaluation on 2400 Heterogeneous Consumer Images**
  A comprehensive empirical evaluation has been carried out using 2400 heterogenous consumer images to illustrate the usefulness of the proposed indexing schemes, similarity integration schemes, and query methods. Two sets of experiments, using 5 events and 11 categories, are conducted for QBCE with very promising results (Section 7.3.3). The QBSI query method is evaluated using 15 visual queries, achieving average precisions of 0.79 and 0.70 for the top 20 and 30 retrieved images respectively (Section 7.4.3). Sample retrieval results are also shown in Section 7.4.3.

  The proposed indexing schemes and similarity integration schemes are evaluated against a feature fusion approach in the QBME experiments using 16 semantic queries (Section 7.5.3). The proposed indexing methods (SSR, SSC, DSR, LCP) and similarity integration schemes (Dsgn, Dscv) outperform the feature fusion approach significantly (Section 7.5.7). In particular, the proposed approaches Dsgn and Dscv achieved high average precisions of 0.59 and 0.52 respectively, which are significant improvements of 55% and 37% over that of the feature fusion method (0.38). The integration of intra-content (SSR, DSR) and inter-class (SSC, LCP) similarities also achieves better retrieval performance than individual indexing methods, thus justifying the integration scheme motivated from Bayesian principles. Sample retrieval results are illustrated in Section 7.5.8.

In conclusion, as suggested by the title of the thesis, innovations in statistical learning, multi-scale segmentation-free indexing, and similarity integration as described above form a common theme for bridging the semantic gap in content-based image retrieval (Figure 8.1). More specifically, with reference to Figure 1.3, the semantic extraction problem is alleviated with the semantics design ($SSR + SSC$) and semantics discovery ($DSR+LCP$) frameworks. On the other hand, event-based

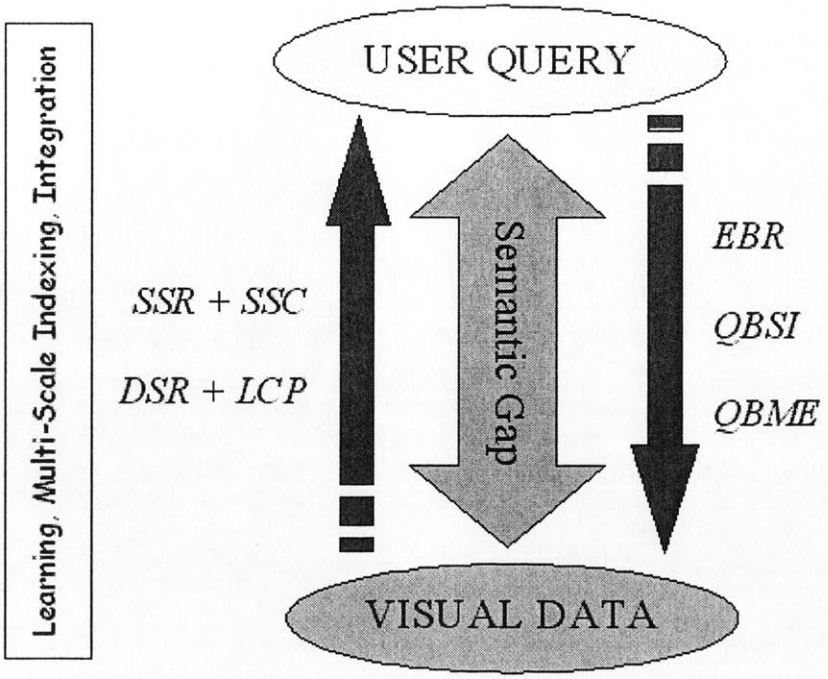

Figure 8.1: Proposed solutions for bridging the semantic gap

retrieval (EBR), QBSI, and QBME have been proposed to deal with the semantic interpretation problem.

8.2 Related Collaborations and Extensions

In the course of research of this thesis, the author has also started investigating several extensions related to the research presented in this thesis with other collaborators. They are described in this section as follows.

8.2.1 Fusion with Conceptual Graph Image Representation

Conceptual graphs, a knowledge representation formalism that handles concepts and hierarchies of concepts as well as relations and hierarchies of relations easily [Sowa, 1984] [Sowa, 2000], has been extended for image content representation [Mechkour, 1995] [Ounis and Pasa, 1998]. Figure 8.2 depicts an example of conceptual graph representation of image. Although conceptual graphs allow abstract description and indexing of image content (left of Figure 8.3) [Mulhem and Lim, 2002],

manual effort has been the only means to construct this kind of highly descriptive indexes [Mechkour, 1995] [Ounis and Pasa, 1998]. The local semantic indexing schemes such as SSRs proposed in this thesis fills the gap between computable low-level features and high-level conceptual graph representation (left of Figure 8.3).

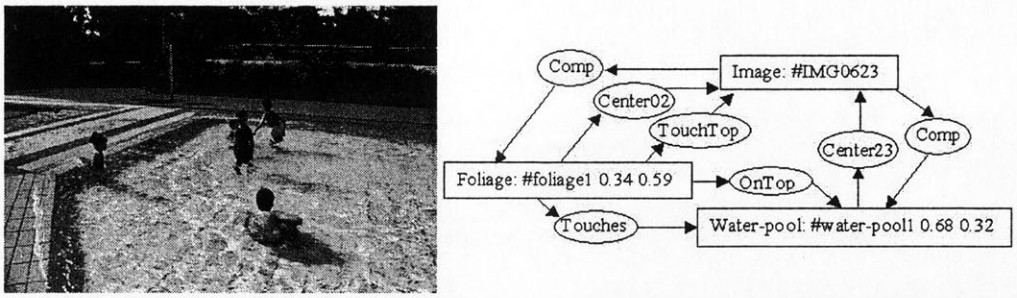

Figure 8.2: An example of conceptual graph representation of image (img0623)

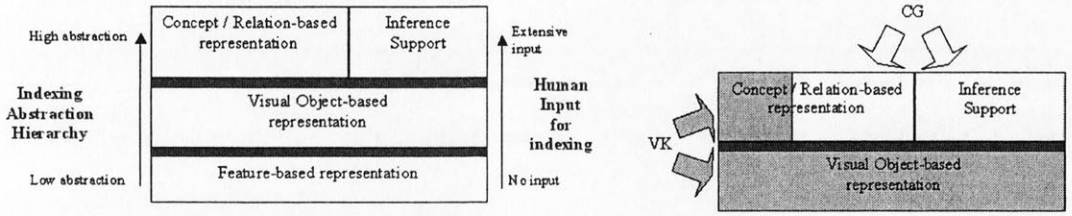

Figure 8.3: Abstraction levels of indexing

In particular, the SSR-based indexing approach (denoted as 'VK' on the right of Figure 8.3) realizes a mid-level index representation and enables automatic construction of conceptual graph (CG) index for an image that involves simple concept hierarchy and relations [Mulhem and Lim, 2002] [Lim et al., 2003b] [Lim et al., 2003c] [Mulhem et al., 2003] [Mulhem and Lim, 2003]. More specifically, the fusion of SSR-based and CG-based indexes has been developed and experimented in similarity-based retrieval [Mulhem and Lim, 2002] [Mulhem et al., 2003] and event-based retrieval [Lim et al., 2003b] [Lim et al., 2003c]. Last but not least, the integration with time information has also been explored [Mulhem and Lim, 2003]. These works were carried out as an international collaboration project with senior scientist Dr. Philippe Mulhem from CNRS, France.

8.2.2 Mapping Mid-Level Representation to Video Events

The notion of mid-level representation such as SSRs and DSRs for image indexing has been extended for semantic video indexing. Figure 8.4 depicts such a mid-level framework for video event detection. Recurrent spatio-temporal patterns with semantic meanings, known as audio-visual keywords, are characterized based on appropriate audio and visual features. Examples of audio-visual keywords include a shot or subshot with view of goal post during formation of an attack in soccer video, a candle-blowing moment of a birthday party in a home video, rocket launching in a documentary video etc. After designing and constructing the audio-visual keywords via supervised learning from examples, the mapping between audio-visual keywords and video events is learned via probabilistic models such as Hidden Markov Models (HMM).

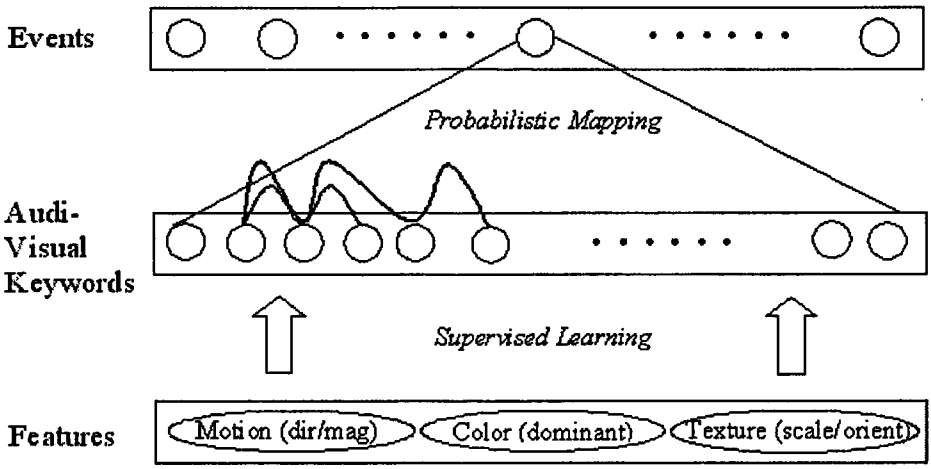

Figure 8.4: A mid-level mapping approach to video event detection

Figure 8.5 shows the flow of video event detection based on detection of audio and visual keywords executed in parallel with post-processing. Two M.Sc. students with School of Computing, National University of Singapore, under the formal supervision of the author, have implemented part of the event detection framework for soccer videos. While Mr. Haiping Sun focused on the extraction of the visual keywords for soccer video [Sun et al., 2003], Mr. Yulin Kang extended the visual keyword set [Kang et al., 2004b] and studied event detection with both grammatical rules [Kang et al., 2003] and HMM models [Kang et al., 2004a].

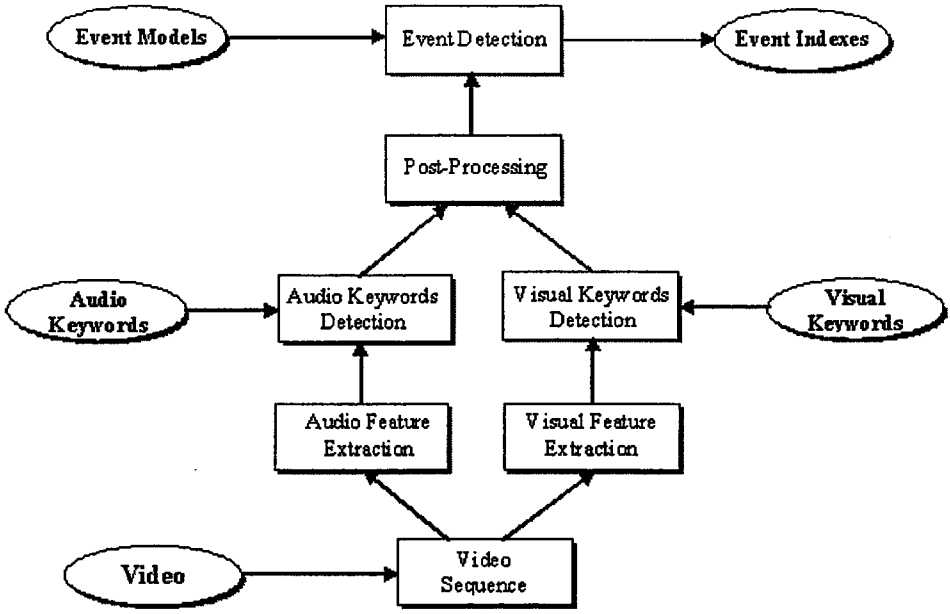

Figure 8.5: Flow of video event detection via audio-visual keywords

8.2.3 Photo Summarization for Visual Communication

Motivated by the wide spread of mobile camera phones and multimedia messaging service (MMS), a framework for automatic organization of personal image libraries based on the analysis of image creation time stamps and image contents to facilitate browsing and summarization of images has been proposed [Lim et al., 2003a] [Li et al., 2003b]. The proposed photo summarization framework has two main phases, namely photo sequence partitioning and key photo selection, as illustrated in Figure 8.6.

Both photo creation time and image content are exploited to efficiently partition image sequences ordered by time stamps. In each partition, key photo(s) is selected based on different criteria such as presence of object(s) such as clear face, monument, image quality such as contrast, sharpness etc. In practice, a mobile user can simply click on a "summarize" button on his/her phone and a summary of the photos stored on the phone will be sent as a MMS message to the recipient(s) who can enjoy it as a slideshow. The current prototype based on Nokia 7650 only implements simple content analysis method such as color histograms. We would adopt more advanced indexing methods presented in the thesis for content analysis and key photo selection.

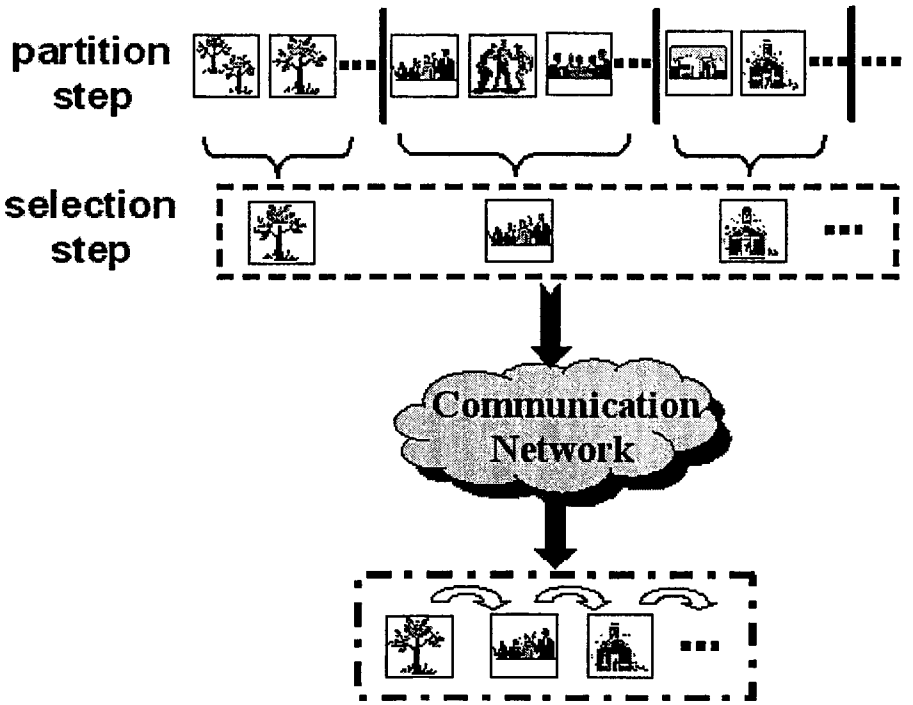

Figure 8.6: A photo summarization framework for visual communication

8.2.4 Snap2Tell: Mobile Scene-based Information Retrieval

A novel mobile scene-based information retrieval framework called Snap2Tell has been conceptualized and prototyped for tourism and education applications (Figure 8.7). The system consists of a mobile client device with camera and an application server. Based on the image of a scene (e.g. monument, nature landscape etc) or an object (painting, statue, flower, animal etc) captured from the camera plus possibly other information such as keywords and location information from positioning device, a user can query relevant information about the subject in the image using the mobile device. Image recognition and multi-modal information integration will be performed at the server to select, customize, and return relevant information to the user.

A key research challenge lies in the area of invariant object and scene recognition. The image captured of an object and a scene has to be recognized as one of the image-based models stored on the server, regardless of its viewpoint and scale under different lighting conditions. Location information obtained from GPS or GSM network can be used to reduce the search space at the database server. The current

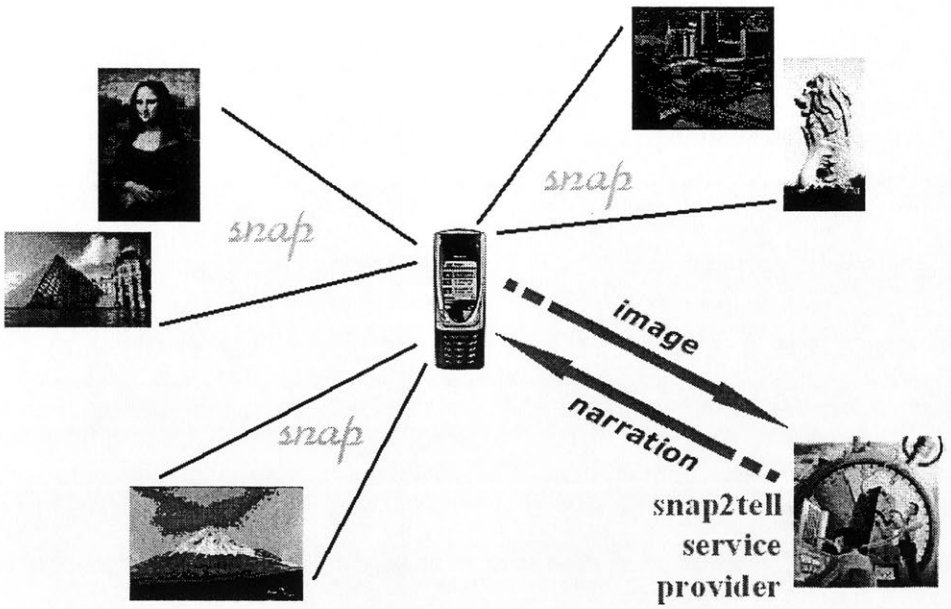

Figure 8.7: A Snap2Tell framework for scene-based information retrieval

Snap2Tell implementation utilizes color histograms for image indexing and matching. We believe that more advanced indexing schemes such as DSR will provide more discriminative power in scene recognition.

8.2.5 Roadmap

Figure 8.8 summarizes the past and current research efforts described above. The research started off with unsupervised approach ("Clustering & SVD"). As the resulting clusters have weak semantic interpretation, the opposite extreme of handcrafted strong semantics ("Handcrafted Visual Keywords") was attempted. With some success of the "Handcrafted Visual Keywords", the supervised approach was extended to the "Semantic Support Regions" framework, upon which, a new query method "Query by Spatial Icons" was designed and implemented in a comprehensive prototype. Through the development of "Event-Based Retrieval" (EBR), "Semantic Support Classes" was conceptualized to complement SSRs in integrated similarity matching. Both SSRs and EBR were also used to build and complement conceptual graph representation for both similarity-based retrieval and relational graph-based retrieval with French collaborator.

Meanwhile, semantics discovery framework based on both "Local Class Patterns"

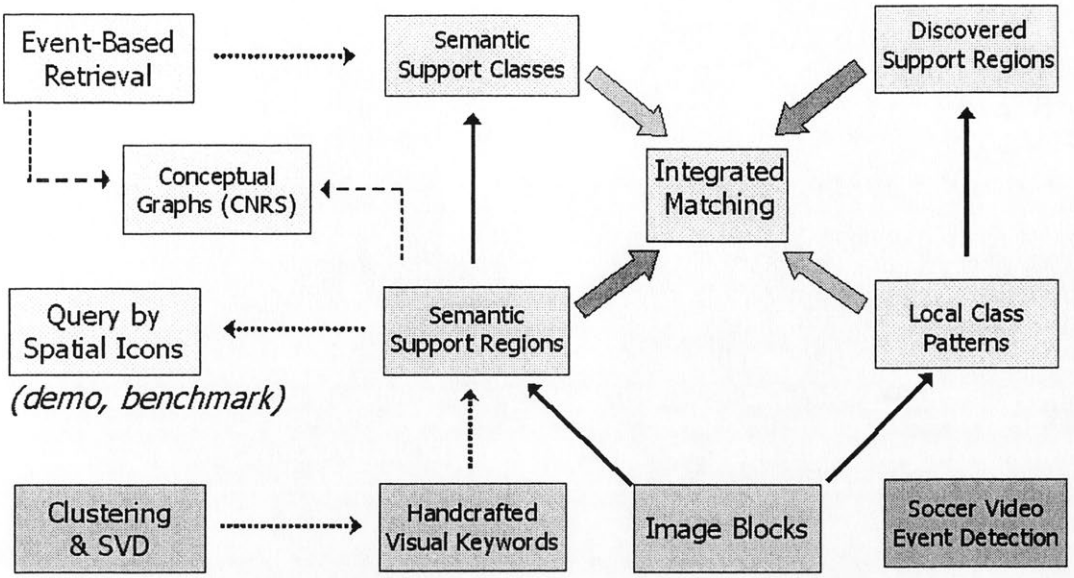

Figure 8.8: A summary of past and current research efforts

and "Discovered Semantic Regions" with similarity integration was developed as a dual parallel alternative to the semantics design path. Also, two M.Sc. students were supervised to extend the mid-level representation to the video domain. In particular, the extraction of mid-level visual representation for soccer event detection has been explored. Last but not least, two application frameworks and prototypes (not shown in Figure 8.8), namely photo summarization and Snap2Tell, have also been developed with the help of two polytechnic students and a software engineer.

8.3 Future Directions

The thesis represents a snapshot of the research effort to bridge the semantic gap in content-based image retrieval. The consolidated results presented in the thesis are not cast in stone. In the author's opinion, the following directions are worth pursuing:

- *Semantics Design Framework*

  As discussed before (Section 3.10), if two-class SVMs are replaced by one-class SVMs [Manevitz and Yousef, 2001] that only require positive examples for learning, we only need to train a new detector or classifier (SSRs, DSRs,

SSCs, LCPs) using new positive examples without re-training all the existing detectors or classifiers. Then efficient re-indexing of images with only the new detectors or classifiers can be carried out as suggested in Section 3.10. We shall adopt one-class SVMs when the parameter sensitivity problem has been resolved [Manevitz and Yousef, 2001].

At the same time, more sophiscated ways for the estimation of posterior probability from SVM outputs (e.g. [Platt, 1999b]) may be experimented. This will be useful in furnishing a more complete probabilistic treatment for the integrated similarity matching scheme. Besides broad consumer images, we would also like to apply the semantics design framework to other domains such as art images, medical images etc.

- *Semantics Discovery Framework*

  A more immediate interest to enhance the semantics discovery framework is to overcome the cluster validity problem as discussed in Section 4.6. The mean shift clustering algorithm [Fukunaga and Hostetler, 1975] [Chang, 1995], a simple iterative procedure that shifts each data point to the average of data points in its neighborhood, is a good candidate to approach the problem.

  An equally important research topic is the investigation of the trade-off between the coverage of sampling based on tessellated image blocks and the computational cost. It is even more important to explore new ways to exercise more control over the region semantics to be discovered, perhaps with additional constraints such as image selection for each class, image focus areas for sampling, etc imposed when domain knowledge is available.

  Problems and applications such as scene recognition [Torralba and Oliva, 2003] for the Snap2Tell application, robotic navigation [Torralba and Sinha, 2001] would be researched and experimented in the near future. In particular, in contrary to the approach that relies on pre-segmentation low-level features [Torralba and Sinha, 2001] [Torralba and Oliva, 2003] for scene categorization, we would like to explore what and how local semantics can be discovered and represented for scene recognition.

- *Multi-Modal Indexing for Consumer Images*

With the proliferation of camera phones, more consumer images will be taken using the camera phones than digital cameras. For effective indexing and accessing of personal multimedia diaries (e.g. such as those created using the forthcoming LifeBlog software at www.nokia.com/lifeblog), integration of indexes based on multiple modalities such as time stamps, location, image content, audio etc is mandatory. Based on our current work on automatic photo summarization [Li et al., 2003b] [Lim et al., 2003a], a more complete personal media management system utilizing multiple indexing cues would be developed.

- *Semantic Video Indexing*

  In the current implementation of soccer event detection [Kang et al., 2004b] [Kang et al., 2004a], the detection of audio-visual keywords does not incorporate certainty information. Hence the mid-level representation is brittle. Errors committed at this level would affect the performance of event detection at the next level. Thus a soft mid-level output representation scheme should be explored. Another interesting extension is to replace the mid-level SVM classifiers by HMMs. The objective is to explore the possibility of achieving audio-visual keyword segmentation and recognition in an integrated manner, instead of the current shot/subshot segmentation before SVM recognition. Last but not least, after the HMMs have learned the temporal mapping between audio-visual keywords and video events, we would like to extract and interpret the rules from the HMM states and transition probabilities.

- *Image-Text Association*

  Learning the association between images and text is a promising research trend [Benitez and Chang, 2003a] [Barnard et al., 2003b] [Li and Wang, 2003]. Existing works in this area either rely on segmented regions [Barnard et al., 2003b] or utilize low-level features for unsupervised learning [Benitez and Chang, 2003a] and generative model learning [Li and Wang, 2003]. Our segmentation-free semantic regions learned or discovered based on discriminative learning may provide an interesting and promising alternative representation for image-text association.

- *Image Code*

  Vision researchers are interested in how images are represented by the neurons in primary visual cortex. In particular, Barlow [Barlow, 1989] proposed that the neurons represent input data with independent components, a factorial code that performs reducdancy reduction. Image code theories were later tested empirically by Olshausen and Field [Olshausen and Field, 1996] [Olshausen and Field, 1997] using natural images. In essence, each image patch is represented as a linear combination of 'basis' patches, such that the mixing coefficients are as sparse as possible. Typically, 2D Gabor functions are chosen as the basis functions and independent components analysis is used to model the data [Lewicki and Olshausen, 1999]. In recent years, the image code framework has been extended to color and stereo images [Hoyer and Hyvarinen, 2000], image sequences [Hyvarinen et al., 2003], and higher-order structures [Karklin and Lewicki, 2003]. Figure 8.9 shows an illustration of representing an image patch in terms of basis images and the 160 basis images computed for color images in [Hoyer and Hyvarinen, 2000].

Each image patch is represented as a linear combination of basis patches

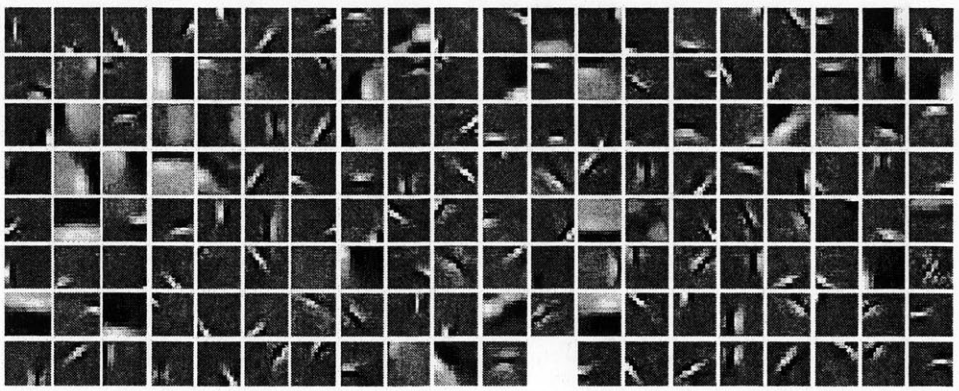

Figure 8.9: An illustration of image code and basis images for color images

In our proposed image indexing framework, an image patch can also be viewed as a linear combination (i.e. histogram) of local image models (e.g. SSRs,

DSRs) built from statistical learning. It would be interesting to develop the link between our detection-based semantic image representation and image code, though the development of corresponding neural evidence may not be viable. As researchers from the data compression community begin to see that signal compression and statistical classification share many goals and properties, both in theory and in practice (e.g. [Ozonat and Gray, 2004]), our motivation is to explore the use of higher-order image code for next generation image compression systems.

Appendix A

List of Publications

People who have accomplished work worthwhile have had a very high sense of the way to do things. They have not been content with mediocrity. They have not confined themselves to the beaten tracks; they have never been satisfied to do things just as others so them, but always a little better. They always pushed things that came to their hands a little higher up, this little farther on, that counts in the quality of life's work. It is constant effort to be first-class in everything one attempts that conquers the heights of excellence.

Orison Swett Marden (1850 - 1924)

In this Appendix, we list the publications that are related to the research presented in the thesis. They are divided into book chapters, journal articles, and conference papers. There is also a granted patent related to image indexing.

The list of publications includes those that are published or accepted during the Ph.D. candidature (March 2002 to July 2004) and those that are published before the candidature. For all papers (except one) listed here, the author of the thesis is either the first or second author. For papers whose first author is someone else, they are joint works with collaborators and students as reported in Section 8.2. For publications before 2002 and the granted patent, the author is the sole author who originated and performed the research.

A.1 Book Chapters

1. Lim, J.H. and Jin, J.S. (2004). From classification to retrieval: exploiting pattern classifiers in semantic indexing and retrieval. In U. Srinivasan & S. Nepal (eds.), *Managing Multimedia Semantics.* Idea Group Publishing (to appear).

2. Mulhem, P., Lim, J.H., Leow, W.K., and Kankanhalli, M.S. (2003). Advances in digital home photo albums. In S. Deb (ed.), *Multimedia Systems and Content-based Image Retrieval.* Idea Group Publishing (ISBN 1-59140-265-4), pp. 201–226.

3. Lim, J.H. (2000). Visual keywords: From text IR to multimedia IR. In F. Crestani & G. Pasi (eds.), *Soft Computing in Information Retrieval: Techniques and Applications*, Physica-Verlag, Springer Verlag, Germany (ISBN 3-7908-1299-4), pp. 77–101.

4. Lim, J.H. (2000). Visual Keywords. In J.K. Wu, M.S. Kankanhalli, J.H. Lim, & D.J. Hong, *Perspectives on Content-Based Multimedia Systems.* Kluwer Academic Publishers (ISBN 0-7923-7944-6), pp. 209–238.

A.2 Journal Papers

1. Lim, J.H. and Jin, J.S. (2004). A structured learning framework for content-based image indexing and visual query. *Multimedia Systems Journal* (to appear).

2. Lim, J.H. and Jin, J.S. (2004). Combining intra-image and inter-class semantics for consumer image retrieval. *Pattern Recognition* (to appear).

3. Lim, J.H., Tian, Q., and Mulhem, P. (2003). Home photo content modeling for personalized event-based retrival. *IEEE Multimedia*, **10**(4): 28–37.

4. Lim, J.H. (2001). Building visual vocabulary for image indexation and query formulation. *Pattern Analysis and Applications*, **4**(2/3): 125–139.

5. Lim, J.H. (2000). Photographs retrieval and classification by visual keywords and thesaurus. *New Generation Computing*, **18**(2): 147–156.

A.3 Conference Papers

1. Kang, Y.L., Lim, J.H., Kankanhalli, M.S., Xu, C.S., and Tian, Q. (2004). Goal detection in soccer video using audio/visual keywords. In *Proc. of IEEE International Conference on Image Processing 2004*, pp. 1629–1632.

2. Lim, J.H. and Jin, J.S. (2004). Combining local class patterns and discovered semantics for image retrieval. In *Proc. of IEEE International Conference on Image Processing 2004*, pp. 401–404.

3. Lim, J.H. and Jin, J.S. (2004). Unifying local and global content-based similarities for home photo retrieval. In *Proc. of IEEE International Conference on Image Processing 2004* (invited paper for special session on Content Understanding for Home Photo Management), pp. 2371–2374.

4. Tang, Q., Sun, H.P., Lim, J.H., Jin, J.S., and Tian, Q. (2004). A generic mid-level representation for semantic video analysis. In *Proc. of IEEE International Conference on Image Processing 2004*, pp. 629–632.

5. Kang, Y.L., Lim, J.H., Tian, Q., Kankanhalli, M.S., and Xu. C.S. (2004). Visual keywords labelling in soccer video. In *Proc. of IEEE International Conference on Pattern Recognition 2004*.

6. Lim, J.H. and Jin, J.S. (2004). Cascading classifiers for consumer image indexing. In *Proc. of International Conference on Pattern Recognition 2004*, pp. 546.1–4.

7. Lim, J.H. and Jin, J.S. (2004). Learning and integrating semantics for image indexing. In *Proc. of Pacific-Rim International Conference on Artificial Intelligence 2004*, pp. 823–832.

8. Lim, J.H. and Jin, J.S. (2004). Image retrieval using spatial icons. In *Proc. of IEEE International Conference on Multimedia & Exposition 2004*.

9. Lim, J.H. and Jin, J.S. (2004). Semantics discovery for image indexing. In Tomas Pajdla & Jiri Matas (eds.), *Proc. of European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004*. Springer-Verlag, Germany, LNCS 3021, pp. 270–281.

10. Kang, Y.L., Lim, J.H., Tian, Q., and Kankanhalli, M.S. (2003). Soccer video event detection with visual keywords. *IEEE Pacific-Rim Conference on Multimedia 2003, Singapore, Dec. 15-18, 2003.*

11. Li, J., Lim, J.H., and Tian Q. (2003). Automatic summarization for personal digital photos. *IEEE Pacific-Rim Conference on Multimedia 2003, Singapore, Dec. 15-18, 2003.*

12. Sun, H.P., Lim, J.H., Tian, Q., and Kankanhalli, M.S. (2003). Semantic labelling of soccer video. *IEEE Pacific-Rim Conference on Multimedia 2003, Singapore, Dec. 15-18, 2003.*

13. Lim, J.H. and Jin, J.S. (2003). Using dual cascading learning frameworks for image indexing. *Visual Information Processing, Sydney, Australia, Dec. 2003.*

14. Lim, J.H. and Jin, J.S. (2003). Support regions and images for photo event retrieval. In *Proc. of IEEE International Conference on Image Processing, Barcelona, Spain, Sep. 14-17, 2003*, pp. II 515–518.

15. Lim, J.H. and Jin, J.S. (2003). Learning consumer photo categories for semantic retrieval. In *Proc. of International Joint Conference on Artificial Intelligence, Acapulco, Mexico, Aug. 9 - 15, 2003*, pp. 1413–1414.

16. Mulhem, P. and Lim, J.H. (2003). Home photo retrieval: time matters. In *Proc. of International Conference on Image and Video Retrieval (CIVR), Urbana, IL, USA, July 24-25, 2003*, pp. 321–330.

17. Lim, J.H., Mulhem, P., and Tian, Q. (2003). Event-based home photo retrieval. In *Proc. of IEEE International Conference on Multimedia & Exposition, Baltimore, USA, July 6-9, 2003*, pp. II. 33–36.

18. Lim, J.H., Li, J., Mulhem, P., and Tian, Q. (2003). Content-Based summarization for personal image library. In *Proc. of ACM/IEEE Joint Conference on Digital Libraries, Houston, USA, May 27-31, 2003*, pp. 393.

19. Lim, J.H. and Jin, J.S. (2002). Semantic indexing and retrieval of home photos. In *Proc. of International Conference on Automation, Robotics, & Computer Vision, Singapore, Dec. 2-5, 2002* (invited paper), pp. 186–191.

20. Lim, J.H. and Jin, J.S. (2002). Home photo indexing using learned visual keywords. In *Proc. of Visual Information Processing, Sydney, Australia, Dec. 2002*, pp. 69–74.

21. Mulhem, P. and Lim, J.H. (2002). Symbolic photograph content-based retrieval. In *Proc. of ACM Conference on Information & Knowledge Management, McLean, VA, USA, Nov. 4-9, 2002*, pp. 94–101.

22. Lim, J.H. and Jin, J.S. (2002). Image indexing and retrieval using visual keyword histograms. In *Proc. of IEEE International Conference on Multimedia & Exposition, Lausanne, Switzerland, Aug. 26-29, 2002*, pp. 213–216.

23. Lim, J.H. (2001). Fuzzy object patterns for visual indexing and segmentation. In *Proc. of FUZZ-IEEE, Melbourne, Australia, Dec. 2-5, 2001*, pp. 77–80.

24. Lim, J.H. (2000). Explicit query formulation with visual keywords. In *Proc. of ACM Multimedia, Los Angeles, CA, USA, Oct 30-Nov 3, 2000*, pp. 407–409.

25. Lim, J.H. (1999). Learnable visual keywords for image classification. In *Proc. of ACM Digital Libraries, Berkeley, CA, USA, Aug. 11-14, 1999*, pp. 139–145.

26. Lim, J.H. (1999). Learning visual keywords for content-based retrieval. In *Proc. of IEEE International Conference on Multimedia & Computing Systems, Florence, Italy, Jun. 7-11, 1999*, pp. 169–173.

27. Lim, J.H. (1999). Categorizing visual contents by matching visual keywords. In N. Huijsmans & A.W.M. Smeulders (eds.), *Visual Information and Visual Information Systems, Proc. of Visual Information Systems 3, Amsterdam,*

The Netherlands, Jun. 2-4, 1999. Springer Verlag, Berlin, LNCS 1614, pp. 367–374.

28. Lim, J.H. (1996). Incremental neural classifier with prototype reduction. In *Proc. of International Conference on Automation, Robotics, & Computer Vision, Singapore, Dec. 3-6, 1996*, pp. 933–936.

29. Lim, J.H. (1993). Incremental case-based pattern classifier. *International Conference on Artificial Neural Networks, Amsterdam, The Netherlands, Sep. 13-16, 1993.*

A.4 Patent

1. Lim, J.H. (2003). Method and Apparatus for Indexing and Retrieving Images using Visual Keyword. *US Patent No 6,574,378* (3/6/2003), *GB Patent No. 2362078* (22/1/2003), *SG Patent No. 82279* (30/6/2003).

Bibliography

[Adams et al., 2003] Adams, W. et al. (2003). Semantic indexing of multimedia content using visual, audio, and text cues. *Eurasip Journal on Applied Signal Processing*, 2003(2).

[Aksoy and Haralick, 2002] Aksoy, S. and Haralick, R. (2002). A classification framework for content-based image retrieval. In *Proc. of ICPR 2002*, pages 503–506.

[Amir et al., 2003] Amir, A. et al. (2003). The ibm semantic concept detection framework. Slides presented at TRECVID Workshop 2003. http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html.

[Armitage and Enser, 1997] Armitage, L. and Enser, P. (1997). Analysis of user need in image archives. *Journal of Information Science*, 23(4):287–299.

[Bach et al., 1996] Bach, J. et al. (1996). Virage image search engine: an open framework for image management. In *Proc. of the SPIE Conference on Storage and Retrieval for Image and Video Databases IV, Vol. 2670*, pages 76–87.

[Barlow, 1989] Barlow, H. (1989). Unsupervised learning. *Neural Computation*, 1:295–311.

[Barnard et al., 2003a] Barnard, K. et al. (2003a). The effects of segmentation of feature choices in a translation model of object recognition. In *Proc. of CVPR 2003.*

[Barnard et al., 2003b] Barnard, K. et al. (2003b). Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135.

[Barnard and Forsyth, 2001] Barnard, K. and Forsyth, D. (2001). Learning the semantics of words and pictures. In *Proc. of ICCV 2001*, pages 408–415.

[Benitez et al., 1998] Benitez, A., Beigi, M., and Chang, S. (1998). Using relevance feedback in content-based image metasearch. *IEEE Internet Computing*, 2(4):59–69.

[Benitez and Chang, 2002a] Benitez, A. and Chang, S.-F. (2002a). Perceptual knowledge construction from annotated image collections. In *Proc. of IEEE ICME 2002.*

[Benitez and Chang, 2002b] Benitez, A. and Chang, S.-F. (2002b). Semantic knowledge construction from annotated image collections. In *Proc. of IEEE ICME 2002.*

[Benitez and Chang, 2003a] Benitez, A. and Chang, S.-F. (2003a). Automatic multimedia knowledge discovery, summarization and evaluation. *IEEE Trans. on Multimedia.* (submitted).

[Benitez and Chang, 2003b] Benitez, A. and Chang, S.-F. (2003b). Image classification using multimedia knowledge networks. In *Proc. of IEEE ICIP 2003.*

[Benitez et al., 2000] Benitez, A., Smith, J., and Chang, S.-F. (2000). Medianet: a multimedia information network for knowledge representation. In *Proc. of the SPIE Conference on Internet Multimedia Management Systems (IS&T/SPIE-2000), Vol. 4210.*

[Bezdek, 1981] Bezdek, J. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms.* Plenum, New York.

[Bishop, 1995] Bishop, C. (1995). *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford.

[Blanz et al., 1996] Blanz, V. et al. (1996). Comparison of view-based object recognition algorithms using realistic 3d models. In von der Malsburg, C., von Seelen, W., Vorbruggen, J., and Sendhoff, B., editors, *Artificial Neural Networks - ICANN 1996 Springer Lecture Notes in Computer Science, Vol. 1112*, pages 251–256. Springer-Verlag.

[Bolle et al., 1998] Bolle, R., Yeo, B., and Yeung, M. (1998). Video query: research directions. *IBM Journal of Research and Development*, 42(2):233–252.

[Bradshaw, 2000] Bradshaw, B. (2000). Semantic based image retrieval: a probabilistic approach. In *Proc. of ACM Multimedia 2000*, pages 167–176.

[Bridle, 1990] Bridle, J. (1990). Probabilistic interpretation of feedforward classification network ouptuts, with relationships to statistical pattern recognition. In Soulie, F. F. and Herault, J., editors, *Neurocomputing: Algorithms, Architectures and Applications*, pages 227–236. Springer-Verlag, New York.

[Brodley et al., 1999] Brodley, C. et al. (1999). Content-based retrieval from medical image databases: A synergy of human interaction, machine learning and computer vision. In *Proc. of AAAI*, pages 760–767.

[Burges, 1998] Burges, C. (1998). A tutorial on support vector machines for pattern recognition. *Knowledge Discovery and Data Mining*, 2(2):121–167.

[Burges and Scholkopf, 1997] Burges, C. and Scholkopf, B. (1997). Improving the accuracy and speed of support vector learning machines. In Mozer, M., Jordan, M., and Petsche, T., editors, *Advances in Neural Information Processing Systems 9*, pages 375–381. MIT Press, Cambridge, MA.

[Carson et al., 1997] Carson, C., Belongie, S., Greenspan, H., and Malik, J. (1997). Region-based image querying. In *Proc. of CVPR 1997 Workshop on Content-Based Access of Image and Video Libraries.*

[Carson et al., 2002] Carson, C., Belongie, S., Greenspan, H., and Malik, J. (2002). Color- and texture-based image segmentation using em and its application to image query and classification. *IEEE Tran. PAMI*, 24(8):1026–1038.

[Chang, 2002] Chang, S.-F. (2002). The holy grail of content-based media analysis. *IEEE Multimedia*, 9(2):6–10.

[Chang and Chen, 2001] Chang, S.-F. and Chen, W. (2001). vismap: an interactive image/video retrieval system using visualization and concept maps. In *Proc. of IEEE ICIP 2001.*

[Chang et al., 1998a] Chang, S.-F., Chen, W., and Sundaram, H. (1998a). Semantic visual templates: linking visual features to semantics. In *Proc. of IEEE ICIP 1998.*

[Chang et al., 1997a] Chang, S.-F. et al. (1997a). Videoq: an automatic content-based video search system using visual cues. In *Proc. of ACM Multimedia 1997.*

[Chang et al., 1998b] Chang, S.-F. et al. (1998b). A fully automated content-based video search engine supporting spatio-temporal queries. *IEEE Trans. on Circuits and Systems for Video Technology*, 8(5).

[Chang et al., 1997b] Chang, S.-F., Smith, J., Beigi, M., and Benitez, A. (1997b). Visual information retrieval from large distributed on-line repositories. *Communications of the ACM*, 40(12):63–71.

[Chang, 1995] Chang, Y. (1995). Mean shift, mode seeking, and clustering. *IEEE Trans. on PAMI*, 17(8):790–799.

[Chen and Tan, 2003] Chen, J. and Tan, T. (2003). Improved method for image retrieval using speech annotation. In *Proc. of Intl. Conf. on Multi-Media Modeling*, pages 15–32.

[Chen et al., 2001] Chen, J., Tan, T., and Mulhem, P. (2001). Method for photograph indexing using speech annotation. In *Proc. of IEEE PCM 2001*, pages 867–872.

[Chen and Wang, 2002] Chen, Y. and Wang, J. (2002). A region-based fuzzy feature matching approach to content-based image retrieval. *IEEE Trans. on PAMI*, 24(9):1252–1267.

[Chen et al., 2004] Chen, Y., Wang, J., and Krovetz, R. (2004). Clue: cluster-based retrieval of images by unsupervised learning. *IEEE Trans. on Image Processing*, page (to appear).

[Cinque et al., 2000] Cinque, L. et al. (2000). Retrieval of images using rich-region descriptions. *Journal of Visual Languages and Computing*, 11:303–321.

[Cooper et al., 2003] Cooper, M. et al. (2003). Temporal event clustering for digital photo collections. In *Proc. of ACM Multimedia 2003*, pages 364–373.

[Cox et al., 2000] Cox, I. et al. (2000). The bayesian image retrieval system, pichunter: theory, implementation and psychophysical experiments. *IEEE Trans. on Image Processing*, 9(1):20–37.

[Cristianni and Shawe-Taylor, 2000] Cristianni, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines.* Cambridge University Press.

[Daoudi and Matusiak, 2000] Daoudi, M. and Matusiak, S. (2000). Visual image retrieval by multiscale description of user sketches. *Journal of Visual Languages and Computing*, 11:287–301.

[Daugman, 1980] Daugman, J. (1980). Two-dimensional spectral analysis of cortical receptive field profiles. *Vision Research*, 20:195–203.

[Daugman, 1988] Daugman, J. (1988). Complete discrete 2d gabor transforms by neural networks for image analysis and compression. *IEEE Trans. ASSP*, 36:1169–1179.

[Deerwester et al., 1990] Deerwester, S. et al. (1990). Indexing by latent semantic analysis. *J. of the Am. Soc. for Information Science*, 41:391–407.

[Del Bimbo and Pala, 1997] Del Bimbo, A. and Pala, P. (1997). Visual image retrieval by elastic matching of user sketches. *IEEE Trans. on PAMI*, 19:121–132.

[Deng and Manjunath, 1999] Deng, Y. and Manjunath, B. (1999). An efficient low-dimensional color indexing scheme for region based image retrieval. In *Proc. of IEEE ICASSP 1999.*

[Duda and Hart, 1973] Duda, R. and Hart, P. (1973). *Pattern Classification and Scene Analysis.* John Wiley & Sons.

[Duygulu et al., 2002] Duygulu, P. et al. (2002). Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. In *Proc. of ECCV 2002*, pages IV. 97–112.

[Ebadollahi et al., 2002] Ebadollahi, S., Chang, S.-F., and Wu, H. (2002). Echocardiogram videos: summarization, temporal segmentation, and browsing. In *Proc. of IEEE ICIP 2002.*

[Ennesser and Medioni, 1995] Ennesser, F. and Medioni, G. (1995). Finding waldo, or focus of attention using local color information. *IEEE Trans. PAMI*, 17(8):805–809.

[Enser, 1993] Enser, P. (1993). Query analysis in a visual information retrieval context. *Journal of Document and Text Management*, 1(1):25–52.

[Enser, 1995] Enser, P. (1995). Pictorial information retrieval. *Journal of Documentation*, 51(2):126–170.

[Enser, 2000] Enser, P. (2000). Visual image retrieval: seeking the alliance of concept based and content based paradigms. *Journal of Information Science*, 26(4):199–210.

[Enser and Sandom, 2003] Enser, P. and Sandom, C. (2003). Towards a comprehensive survey of the semantic gap in visual image retrieval. In Bakker, E. et al., editors, *Proc. of CIVR 2003 LNCS 2728*, pages 291–299.

[Fergus et al., 2003] Fergus, R., Perona, P., and Zisserman, A. (2003). Object class recognition by unsupervised scale-invariant learning. In *Proc. of IEEE CVPR 2003.*

[Flickner et al., 1995] Flickner, M. et al. (1995). Query by image and video content: the qbic system. *IEEE Computer*, 28(9):23–30.

[Forsyth, 2001] Forsyth, D. (2001). Benchmarks for storage and retrieval in multimedia databases. In *Proc. of SPIE vol. 4676*, pages 240–247.

[Frost et al., 2000] Frost, C. et al. (2000). Browse and search patterns in a digital image database. *Information Retrieval*, 1:287–313.

[Fukunaga and Hostetler, 1975] Fukunaga, K. and Hostetler, L. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. on Information Theory*, 21:32–40.

[Gevers and Smeulders, 1997] Gevers, T. and Smeulders, A. (1997). Pictoseek: a content-based image search system for the world wide web. In *Proc. Visual 1997*, pages 93–100.

[Gonzalez and Woods, 1992] Gonzalez, R. and Woods, R. (1992). *Digital Image Processing.* Addison-Wesley Publishing Company.

[Graham et al., 2002] Graham, A., Garcia-Molina, H., Paepcke, A., and Winograd, T. (2002). Time as essence for photo browsing through personal digital libraries. In *Proc. of JCDL 2002*, pages 326–335.

[Greenspan et al., 2001] Greenspan, H., Goldberger, J., and Ridel, L. (2001). A continuous probabilistic framework for image matching. *Journal of Computer Vision and Image Understanding*, pages 1–23.

[Haering et al., 2000] Haering, N., Qian, R., and Sezan, M. (2000). A semantic event-detection approach and its application to detecting hunts in wildlife video. *IEEE Trans. CSVT*, 10(6):857–867.

[Hoyer and Hyvarinen, 2000] Hoyer, P. and Hyvarinen, A. (2000). Independent component analysis applied to feature extraction from colour and stereo images. *Network: Computation in Neural Systems*, 11(3):191–210.

[Hsu and Chang, 2004] Hsu, W. and Chang, S. (2004). Generative, discriminative, and ensemble learning on multi-modal perceptual fusion toward news video story segmentation. In *Proc. of IEEE ICME 2004.* (to appear).

[Huang et al., 1998] Huang, J., Kumar, S., and Zabih, R. (1998). An automatic hierarchical image classification scheme. In *Proc. of ACM Multimedia 1998*, pages 219–228.

[Hyvarinen et al., 2003] Hyvarinen, A., Hurri, J., and Vayrynen, J. (2003). Bubbles: a unifying framework for low-level statistical properties of natural image sequences. *Journal of the Optical Society of America A: Optics, Image Science, and Vision*, 20(7):1237–1252.

[Iyengar et al., 2003] Iyengar, G., Nock, H., and Neti, C. (2003). Discriminative model fusion for semantic concept detection and annotation in video. In *Proc. of ACM Multimedia 2003, Nov. 2-8, 2003, Berkeley, California, U.S.A.*, pages 255–258.

[Jaimes, 2003] Jaimes, A. (2003). *Conceptual Structures and Computational Methods for Indexing and Organization of Visual Information.* PhD thesis, Graduate School of Arts and Sciences, Columbia University.

[Jaimes and Chang, 2001] Jaimes, A. and Chang, S.-F. (2001). Learning structured visual detectors from user input at multiple levels. *International Journal of Image and Graphics.*

[Jain and Healey, 1998] Jain, A. and Healey, G. (1998). A multiscale representation including opponent color features for texture recognition. *IEEE Trans. on Image Processing*, 7(1):124–128.

[JEITA, 2002] JEITA (2002). Exif 2.2. http://www.exif.org/specifications.html.

[Joachims, 1998] Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. In *Proc. of the European Conference on Machine Learning*, pages 137–142.

[Joachims, 1999] Joachims, T. (1999). Making large-scale svm learning practical. In Scholkopf, B., Burges, C., and Smola, A., editors, *Advances in Kernel Methods - Support Vector Learning*, pages 169–184. MIT-Press.

[Johnson and Wichern, 1988] Johnson, R. and Wichern, D. (1988). *Applied Multivariate Statistical Analysis.* Prentice Hall International, Inc.

[Kang et al., 2004a] Kang, Y., Lim, J., Kankanhalli, M., Xu, C., , and Tian, Q. (2004a). Goal detection in soccer video using audio-visual keywords. In *Proc. of IEEE ICIP 2004*, pages 1629–1632.

[Kang et al., 2004b] Kang, Y., Lim, J., Tian, Q., Kankanhalli, M., , and Xu, C. (2004b). Visual keywords labelling in soccer video. In *Proc. of ICPR 2004.*

[Kang et al., 2003] Kang, Y., Lim, J., Tian, Q., and Kankanhalli, M. (2003). Soccer video event detection with visual keywords. In *Proc. of IEEE PCM 2003, Singapore, Dec. 15-18, 2003.*

[Kapur and Kesava, 1992] Kapur, J. and Kesava, H. (1992). *Entropy Optimization Principles with Applications.* Academic Press.

[Karklin and Lewicki, 2003] Karklin, Y. and Lewicki, M. (2003). Learning higher-order structures in natural images. *Network: Computation in Neural Systems*, 14:483–499.

[Keister, 1994] Keister, L. (1994). User types and queries: impact on image access systems. In *Challenges in Indexing Electronic Text and Images.* Learned Information.

[Kelly et al., 1996] Kelly, P., Cannon, M., and Barros, J. (1996). Efficiency issues related to probability density function comparison. In *Proc. of SPIE vol. 2670, Storage and Retrieval for Still Images and Video Databases IV*, pages 42–49.

[Klir and Folger, 1992] Klir, G. and Folger, T. (1992). *Fuzzy Sets, Uncertainty, and Information.* Prentice Hall.

[Kohonen, 1997] Kohonen, T. (1997). *Self-Organizing Maps (Second Edition).* Springer.

[Kumar et al., 2002] Kumar, S., Loui, A., and Hebert, M. (2002). Probabilistic classification of image regions using an observation-constrained generative approach. In *Proc. of First Intl. Workshop on Generative-Model-Based Vision.*

[Kutics et al., 2003] Kutics, A. et al. (2003). Linking images and keywords for semantics-based image retrieval. In *Proc. of IEEE ICME 2003*, pages 777–780.

[Landauer et al., 1998] Landauer, T., Laham, D., and Foltz, P. (1998). Learning human-like knowledge by singular value decomposition: a progress report. In M.I. Jordan, M. K. and Solla, S., editors, *Advances in Neural Information Processing Systems 10*, pages 45–51. MIT Press, Cambridge.

[Larkey and Croft, 1996] Larkey, L. and Croft, W. (1996). Combining classifiers in text categorization. In *Proc. of ACM SIGIR 1996*, pages 289–297.

[LeCun et al., 1995] LeCun, Y. et al. (1995). Learning algorithms for classification: a comparison on handwritten digit recognition. *Neural Networks*, pages 261–276.

[Lew, 2000] Lew, M. (2000). Next-generation web searches for visual content. *IEEE Computer*, 33(11):46–52.

[Lewicki and Olshausen, 1999] Lewicki, M. and Olshausen, B. (1999). A probabilistic framework for the adaptation and comparison of image codes. *Journal of the Optical Society of America A: Optics, Image Science, and Vision*, 16(7):1587–1601.

[Lewis and Ringuette, 1994] Lewis, D. and Ringuette, M. (1994). A comparison of two learning algorithms for text categorization. In *Proc. of ACM SIGIR 1994*, pages 81–93.

[Li et al., 2003a] Li, B., Goh, K., and Chang, E. (2003a). Confidence-based dynamic ensemble for image annotation and semantics discovery. In *Proc. of ACM Multimedia 2003*, pages 195–206.

[Li et al., 2003b] Li, J., Lim, J., and Tian, Q. (2003b). Automatic summarization for personal digital photos. In *Proc. of IEEE PCM 2003, Singapore, Dec. 15-18, 2003.*

[Li and Wang, 2003] Li, J. and Wang, J. (2003). Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. on PAMI*, 25(10):1–14.

[Li et al., 2000] Li, J., Wang, J., and Wiederhold, G. (2000). Integrated region matching for image retrieval. In *Proc. of ACM Multimedia 2000*, pages 147–156.

[Lim, 1993] Lim, J. (1993). Incremental case-based pattern classifier. In *Proc. of the Intl. Conf. on Artificial Neural Networks, Amsterdam, The Netherlands, Sep. 13-16, 1993.*

[Lim, 1996] Lim, J. (1996). Incremental neural classifier with prototype reduction. In *Proc. of ICARCV 1996, Singapore, Dec. 3-6, 1996*, pages 933–936.

[Lim, 1999a] Lim, J. (1999a). Categorizing visual contents by matching visual keyword. In Huijsmans, D. and Smeulders, A., editors, *Visual Information and Information Systems, Proc. of Visual 1999, Amsterdam, The Netherlands, Jun. 2-4, 1999*, pages 367–374.

[Lim, 1999b] Lim, J. (1999b). Learnable visual keywords for image classification. In *Proc. of ACM Digital Libraries 1999, Berkeley, CA, USA, Aug. 11-14, 1999*, pages 139–145.

[Lim, 1999c] Lim, J. (1999c). Learning visual keywords for content-based retrieval. In *Proc. of IEEE ICMCS 1999, Florence, Italy, Jun. 7-11, 1999*, pages 169–173.

[Lim, 2000a] Lim, J. (2000a). Explicit query formulation with visual keywords. In *Proc. of ACM Multimedia 2000, Los Angeles, CA, USA, Oct 30-Nov 3, 2000*, pages 407–409.

[Lim, 2000b] Lim, J. (2000b). Photographs retrieval and classification by visual keywords and thesaurus. *New Generation Computing*, 18(2):147–156.

[Lim, 2000c] Lim, J. (2000c). Visual keywords. In Wu, J. et al., editors, *Perspectives on Content-Based Multimedia Systems*, pages 209–238. Kluwer Academic Publishers.

[Lim, 2000d] Lim, J. (2000d). Visual keywords: from text ir to multimedia ir. In F.Crestani and G.Pasi, editors, *Soft Computing in Information Retrieval: Techniques and Applications*, pages 77–101. Physica-Verlag, Springer Verlag, Germany.

[Lim, 2001a] Lim, J. (2001a). Building visual vocabulary for image indexation and query formulation. *Pattern Analysis and Applications*, 4(2):125–139.

[Lim, 2001b] Lim, J. (2001b). Fuzzy object patterns for visual indexing and segmentation. In *Proc. of FUZZ-IEEE 2001, Melbourne, Australia, Dec. 2-5, 2001*, pages 77–80.

[Lim, 2003] Lim, J. (2003). Method and apparatus for indexing and retrieving images using visual keyword. US Patent No 6,574,378, GB 2362078, SG Patent No. 82279.

[Lim and Jin, 2002a] Lim, J. and Jin, J. (2002a). Home photo indexing using learned visual keywords. In *Proc. of VIP 2002, Sydney, Australia, Dec. 2002*, pages 69–74.

[Lim and Jin, 2002b] Lim, J. and Jin, J. (2002b). Image indexing and retrieval using visual keyword histograms. In *Proc. of IEEE ICME 2002, Lausanne, Switzerland, Aug. 26-29, 2002*, pages 213–216.

[Lim and Jin, 2002c] Lim, J. and Jin, J. (2002c). Semantic indexing and retrieval of home photos. In *Proc. of ICARCV 2002, Singapore, Dec. 2-5, 2002 (invited paper)*, pages 186–191.

[Lim and Jin, 2003a] Lim, J. and Jin, J. (2003a). Learning consumer photo categories for semantic retrieval. In *Proc. of IJCAI 2003, Acapulco, Mexico, Aug. 9 - 15, 2003*, pages 1413–1414.

[Lim and Jin, 2003b] Lim, J. and Jin, J. (2003b). Support regions and images for photo event retrieval. In *Proc. of IEEE ICIP 2003, Barcelona, Spain, Sep. 14-17, 2003*, pages II.515–518.

[Lim and Jin, 2003c] Lim, J. and Jin, J. (2003c). Using dual cascading learning frameworks for image indexing. In *Proc. of VIP 2003, Sydney, Australia, Dec. 2003.*

[Lim and Jin, 2004a] Lim, J. and Jin, J. (2004a). Cascading classifiers for consumer image indexing. In *Proc. of ICPR 2004*, pages 546.1–4.

[Lim and Jin, 2004b] Lim, J. and Jin, J. (2004b). Combining intra-image and inter-class semantics for consumer image retrieval. *Pattern Recognition.* (to appear).

[Lim and Jin, 2004c] Lim, J. and Jin, J. (2004c). Combining local class patterns and discovered semantics for image retrieval. In *Proc. of IEEE ICIP 2004*, pages 401–404.

[Lim and Jin, 2004d] Lim, J. and Jin, J. (2004d). From classification to retrieval: exploiting pattern classifiers in semantic indexing and retrieval. In Srinivasan, U.

and Nepal, S., editors, *Managing Multimedia Semantics*. Idea Group Publishing. (to appear).

[Lim and Jin, 2004e] Lim, J. and Jin, J. (2004e). Image retrieval using spatial icons. In *Proc. of IEEE ICME 2004*.

[Lim and Jin, 2004f] Lim, J. and Jin, J. (2004f). Learning and integrating semantics for image indexing. In *Proc. of PRICAI 2004*, pages 823–832.

[Lim and Jin, 2004g] Lim, J. and Jin, J. (2004g). Semantics discovery for image indexing. In *Proc. of ECCV 2004, Prague, Czech Republic, May 10-14, 2004*, pages 270–281.

[Lim and Jin, 2004h] Lim, J. and Jin, J. (2004h). A structured learning framework for content-based image indexing and visual query. *Multimedia Systems Journal*. (to appear).

[Lim and Jin, 2004i] Lim, J. and Jin, J. (2004i). Unifying local and global content-based similarities for home photo retrieval. In *Proc. of IEEE ICIP 2004 (invited special session on Content Understanding for Home Photo Management)*, pages 2371–2374.

[Lim et al., 2003a] Lim, J., Li, J., Mulhem, P., and Tian, Q. (2003a). Content-based summarization for personal image library. In *Proc. of ACM/IEEE JCDL 2003, Houston, USA, May 27-31, 2003*, page 393.

[Lim et al., 2003b] Lim, J., Mulhem, P., and Tian, Q. (2003b). Event-based home photo retrieval. In *Proc. of IEEE ICME 2003, Baltimore, USA, July 6-9, 2003*, pages II.33–36.

[Lim et al., 2003c] Lim, J., Tian, Q., and Mulhem, P. (2003c). Home photo content modeling for personalized event-based retrival. *IEEE Multimedia*, 10(4):28–37.

[Lin et al., 2003] Lin, C., Tseng, B., and Smith, J. (2003). Videoannex: Ibm mpeg-7 annotation tool for multimedia indexing and concept learning. In *Proc. of IEEE ICME 2003*.

[Lin and Hauptmann, 2002] Lin, W. and Hauptmann, A. (2002). News video classification using svm-based multimodal classifiers and combination strategies. In *Proc. of ACM Multimedia 2002, Dec. 1-6, 2002, Juan-les-Pins, France.*

[Lipson et al., 1997] Lipson, P., Grimson, E., and Sinha, P. (1997). Configuration based scene classification and image indexing. In *Proc. of IEEE CVPR 1997*, pages 1007–1013.

[Liu et al., 2000] Liu, W., Sun, Y., and Zhang, H. (2000). Mialbum - a system for home photo management using the semi-automatic image annotation approach. In *Proc. of ACM Multimedia 2000*, pages 479–480.

[Lu et al., 2000] Lu, Y. et al. (2000). A unified framework for semantics and feature based relevance feedback in image retrieval systems. In *Proc. of ACM Multimedia 2000*, pages 31–37.

[Ma and Manjunath, 1997a] Ma, W. and Manjunath, B. (1997a). Edge flow: a framework for boundary detection and image segmentation. In *Proc. of IEEE CVPR 1997.*

[Ma and Manjunath, 1997b] Ma, W. and Manjunath, B. (1997b). Netra: a toolbox for navigating large image databases. In *Proc. of IEEE ICIP 1997*, pages I. 568–571.

[Ma and Manjunath, 1999] Ma, W. and Manjunath, B. (1999). Netra: a toolbox for navigating large image databases. *Multimedia Systems*, 7(3):184–198.

[Maenpaa et al., 2002] Maenpaa, T., Pietikainen, M., and Viertola, J. (2002). Separating color and pattern information for color texture discrimination. In *Proc. of ICPR 2002, Aug. 11-15, Quebec City, Canada*, pages 668–671.

[Manevitz and Yousef, 2001] Manevitz, L. and Yousef, M. (2001). One-class svms for document classification. *Journal of Machine Learning Research*, 2:139–154.

[Manjunath and Ma, 1996] Manjunath, B. and Ma, W. (1996). Texture features for browsing and retrieval of image data. *IEEE Trans. on PAMI*, 18(8):837–842.

[Markkula and Sormunen, 2000] Markkula, M. and Sormunen, E. (2000). End-user searching challenges indexing practices in the digital newspaper photo archive. *Information Retrieval*, 1:259–285.

[Martinez and Serra, 2000] Martinez, A. and Serra, J. (2000). A new approach to object-related image retrieval. *Journal of Visual Languages and Computing*, 11:345–363.

[Mechkour, 1995] Mechkour, M. (1995). An extended model for image representation and retrieval. In *Proc. of the Intl. Conf. on Database and Expert System Applications*, pages 395–404.

[Mills et al., 2000] Mills, T. et al. (2000). Shoebox: a digital photo management system. Technical Report 2000.10, AT&T Laboratories Cambridge.

[Minka and Picard, 1997] Minka, T. and Picard, R. (1997). Interactive learning using a society of models. *Pattern Recognition*, 30(4).

[Mitchell, 1997] Mitchell, T. (1997). *Machine Learning*. McGraw-Hill.

[Moghaddam et al., 2001] Moghaddam, B., Biermann, H., and Margaritis, D. (2001). Regions-of-interest and spatial layout for content-based image retrieval. *Multimedia Tools and Application*, 14:201–210.

[Moghaddam et al., 1998] Moghaddam, B., Wahid, W., and Pentland, A. (1998). Beyond eigenfaces: probabilistic matching for face recognition. In *Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pages 30–35.

[Mohan et al., 2001] Mohan, A., Papageorgiou, C., and Poggio, T. (2001). Example-based object detection in images by components. *IEEE Trans. on PAMI*, 23(4):349–361.

[Mojsilovic and Gomes, 2002] Mojsilovic, A. and Gomes, J. (2002). Semantic based categorization, browsing and retrieval in medical image databases. In *Proc. of IEEE ICIP*.

[Mulhem and Lim, 2002] Mulhem, P. and Lim, J. (2002). Symbolic photograph content-based retrieval. In *Proc. of ACM CIKM 2002, McLean, VA, USA, Nov. 4-9, 2002*, pages 94–101.

[Mulhem and Lim, 2003] Mulhem, P. and Lim, J. (2003). Home photo retrieval: time matters. In *Proc. of Intl. Conf. on Image and Video Retrieval (CIVR), Urbana, IL, USA, July 24-25, 2003*, pages 321–330.

[Mulhem et al., 2003] Mulhem, P., Lim, J., Leow, W., and Kankanhalli, M. (2003). Advances in digital home photo albums. In Deb, S., editor, *Multimedia Systems and Content-based Image Retrieval*, pages 201–226. Idea Group Publishing.

[Muller et al., 1997] Muller, K.-R. et al. (1997). Predicting time series with support vector machines. In Gerstner, W., Germond, A., Hasler, M., and Nicoud, J.-D., editors, *Artificial Neural Networks - ICANN 1997, Springer Lecture Notes in Computer Science, Vol. 1327*, pages 999–1004. Springer-Verlag.

[Muller et al., 2001] Muller, K.-R. et al. (2001). An introduction to kernel-based learning algorithms. *IEEE Trans. on Neural Networks*, 12(2):181–202.

[Naphade et al., 2003] Naphade, M. et al. (2003). A framework for moderate vocabulary semantic visual concept detection. In *Proc. IEEE ICME 2003*, pages 437–440.

[Naphade and Huang, 2001] Naphade, M. and Huang, T. (2001). A probabilistic framework for semantic video indexing, filtering, and retrieval. *IEEE Trans. on Multimedia*, 3(1):141–151.

[Naphade et al., 2002] Naphade, M., Kozintsev, I., and Huang, T. (2002). A factor graph framework for semantic video indexing. *IEEE Trans. on Circuits and Systems for Video Technology*, 12(1):40–52.

[Naphade and Smith, 2003] Naphade, M. and Smith, J. (2003). Learning visual models of semantic concepts. In *Proc. IEEE ICIP 2003*.

[Nefian and Hayes III, 1999] Nefian, A. and Hayes III, M. (1999). Face recognition using an embedded hmm. In *Proc. of the IEEE Conf. on Audio and Video-based Biometric Person Authentication*, pages 19–24.

[Olshausen and Field, 1996] Olshausen, B. and Field, D. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609.

[Olshausen and Field, 1997] Olshausen, B. and Field, D. (1997). Sparse coding with an overcomplete basis set: a strategy employed by v1? *Vision Research*, 37:3311–3325.

[Ornager, 1996] Ornager, S. (1996). View a picture: theoretical image analysis and empirical user studies on indexing and retrieval. *Swedis Library Research*, 2:31–41.

[Ortega et al., 1997] Ortega, M. et al. (1997). Supporting similarity queries in mars. In *Proc. of ACM Multimedia 1997*, pages 403–413.

[Osuna et al., 1997] Osuna, E., Freund, R., and Girosi, F. (1997). An improved training algorithm for support vector machines. In Principe, J., Giles, L., Morgan, N., and Wilson, E., editors, *Proc. of IEEE Workshop on Neural Networks for Signal Processing*, pages 276–285.

[Ounis and Pasa, 1998] Ounis, I. and Pasa, M. (1998). Relief: Combining expressiveness and rapidity into a single system. In *Proc. of ACM SIGIR 1998*, pages 266–274.

[Ozonat and Gray, 2004] Ozonat, K. and Gray, R. (2004). Image classification using adaptive boosting and tree-structured discriminant vector quantization. In *Proc. of the Data Compression Conference, Snowbird, UT, USA*, page (to appear).

[Papageorgiou et al., 1998] Papageorgiou, P., Oren, M., and Poggio, T. (1998). A general framework for object detection. In *Proc. of ICCV*, pages 555–562.

[Paschos, 2001] Paschos, G. (2001). Perceptually uniform color spaces for color texture analysis: an empirical evaluation. *IEEE Trans. on Image Processing*, 10(6):932–937.

[Pass and Zabih, 1999] Pass, G. and Zabih, R. (1999). Comparing images using joint histograms. *Multimedia Systems*, 7:234–240.

[Pentland et al., 1995] Pentland, A., Picard, R., and Sclaroff, S. (1995). Photobook: content-based manipulation of image databases. *Intl. J. of Computer Vision*, 18(3):233–254.

[Picard, 1995] Picard, R. (1995). Toward a visual thesaurus. In *Proc. of Springer-Verlag Workshops in Computing, MIRO'95, Glasgow, Sep. 1995.*

[Pietikainen et al., 2002] Pietikainen, M., Maenpaa, T., and Viertola, J. (2002). Color texture classification with color histograms and local binary patterns. In *Proc. of Second Intl. Workshop on Texture Analysis and Synthesis, June 1, Copenhagen, Denmark*, pages 109–112.

[Platt, 1999a] Platt, J. (1999a). Fast training of support vector machines using sequential minimal optimization. In Scholkopf, B., Burges, C., and Smola, A., editors, *Advances in Kernel Methods - Support Vector Learning*, pages 185–208. MIT Press.

[Platt, 1999b] Platt, J. (1999b). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Smola, A. et al., editors, *Advances in Large Margin Classifiers.* MIT Press.

[Platt, 2000] Platt, J. (2000). Autoalbum: clustering digital photographs using probabilistic model merging. In *Proc. of IEEE Workshop on Content-Based Access of Image and Video Libraries.*

[Platt et al., 2003] Platt, J., Czerwinski, M., and Field, B. (2003). Phototoc: automatic clustering for browsing personal photographs. In *Proc. of IEEE PCM 2003.*

[Poirson and Wandell, 1996] Poirson, B. and Wandell, B. (1996). Pattern-color separable pathways predict sensitivity to simple colored patterns. *Vision Research*, 36(4):515–526.

[Puzicha et al., 1999] Puzicha, J., Buhmann, J., Rubner, Y., and Tomasi, C. (1999). Empirical evaluation of dissimilarity measures for color and texture. In *Proc. of ICCV*, pages 1165–1172.

[Ratan and Grimson, 1997] Ratan, A. and Grimson, W. (1997). Training templates for scene classification using a few examples. In *Proc. IEEE Workshop on Content-Based Analysis of Images and Video Libraries*, pages 90–97.

[Robertson and Sparck Jones, 1976] Robertson, S. and Sparck Jones, K. (1976). Relevance weighting of search terms. *J. of the Am. Soc. for Info. Sc.*, 27:129–146.

[Rodden, 1999] Rodden, K. (1999). How do people organize their photographs? In *Proc. of the BCS IRSG Colloquium, Electronic Workshops in Computing*, pages 142–152.

[Rodden and Wood, 2003] Rodden, K. and Wood, K. (2003). How people manage their digital photos? In *Proc. of ACM CHI 2003.*

[Rowe and Eads, 1994] Rowe, L.A. Boreczky, J. and Eads, C. (1994). Indices for user access to large video database. In *Proc. of SPIE 2185, Storage and Retrieval for Image and Video Databases II*, pages 150–161.

[Rowley et al., 1998] Rowley, H., Baluja, S., and Kanade, T. (1998). Neural network-based face detection. *IEEE Trans. on PAMI*, 20(1):23–38.

[Rui et al., 1997] Rui, Y., Huang, T., and Mehrotra, S. (1997). Content-based image retrieval with relevance feedback in mars. In *Proc. of IEEE ICIP*, pages 815–818.

[Salton, 1971] Salton, G., editor (1971). *The SMART System - Experiments in Automatic Document Processing.* Englewood Cliffs, NJ: Prentice Hall.

[Santini et al., 2001] Santini, S., Gupta, A., and Jain, R. (2001). Emergent semantics through interaction in image databases. *IEEE Trans. on KDE*, 13(3):337–351.

[Schmid, 1996] Schmid, C. (1996). Identifying speaker with support vector networks. In *Proc. of Interface 1996, Sydney.*

[Schmid, 2001] Schmid, C. (2001). Constructing models for content-based image retrieval. In *Proc. of CVPR*, pages 39–45.

[Scholkopf, 2000] Scholkopf, B. (2000). Statistical learning and kernel methods. Technical Report MSR-TR-2000-23, Microsoft Research Technical Report.

[Selinger and Nelson, 2001] Selinger, A. and Nelson, R. (2001). Minimally supervised acquisition of 3d recognition models from cluttered images. In *Proc. of CVPR*, pages 213–220.

[Smeulders et al., 2000] Smeulders, A. et al. (2000). Content-based image retrieval at the end of the early years. *IEEE Trans. on PAMI*, 22(12):1349–1380.

[Smith and Chang, 1996] Smith, J. and Chang, S.-F. (1996). Visualseek: a fully automated content-based image query system. In *Proc. of ACM Multimedia 1996*, pages 87–98.

[Smith and Chang, 1997] Smith, J. and Chang, S.-F. (1997). Visually searching the web for content. *IEEE Multimedia*, 4(3):12–20.

[Smith and Chang, 1999] Smith, J. and Chang, S.-F. (1999). Integrated spatial and feature image query. *Multimedia Systems*, 7(2):129–140.

[Smith et al., 2001] Smith, J. et al. (2001). Integrating features, models, and semantics for trec video. In *NIST Special Publication 500-200: The Tenth Text REtrieval Conference (TREC 2001)*, pages 240–249.

[Snoek and Worring, 2002] Snoek, C. and Worring, M. (2002). Multimodal video indexing: a review of the state-of-the-art. *Multimedia Tools and Applications*, page (to appear).

[Snoek et al., 2004] Snoek, C., Worring, M., and Hauptmann, A. (2004). Detection of tv news monologues by style reconstruction. In *Proc. of IEEE ICME 2004*, page (to appear).

[Sowa, 1984] Sowa, J., editor (1984). *Conceptual Structures: Information Processing in Mind and Machine.* Addison-Wesley Publisher.

[Sowa, 2000] Sowa, J., editor (2000). *Knowledge Representation: Logical, Philosophical and Computational Fundations.* Brooks/Cole Publisher.

[Sparck Jones and Willett, 1997] Sparck Jones, K. and Willett, P., editors (1997). *Readings in Information Retrieval.* Morgan Kaufmann Publishers, Inc.

[Sun et al., 2003] Sun, H., Lim, J., Tian, Q., and Kankanhalli, M. (2003). Semantic labelling of soccer video. In *Proc. of IEEE PCM 2003, Singapore, Dec. 15-18, 2003.*

[Sundaram and Chang, 2000] Sundaram, H. and Chang, S.-F. (2000). Determining computable scenes in films and their structures using audio visual memory models. In *Proc. of ACM Multimedia 2000.*

[Sung and Poggio, 1998] Sung, K. and Poggio, T. (1998). Example-based learning for view-based human face detection. *IEEE Trans. on PAMI*, 20(1):39–51.

[Swain and Ballard, 1991] Swain, M. and Ballard, D. (1991). Color indexing. *Intl. J. Computer Vision*, 7(1):11–32.

[Szummer and Picard, 1998] Szummer, M. and Picard, R. (1998). Indoor-outdoor image classification. In *Proc. of IEEE Int. Work. on Content-based Access of Image and Video Databases*, pages 42–51.

[Tao and Grosky, 2000] Tao, Y. and Grosky, W. (2000). Image indexing and retrieval using object-based point feature maps. *Journal of Visual Languages and Computing*, 11:323–343.

[Taycher et al., 1997] Taycher, L., Cascia, M., and Sclaroff, S. (1997). Image digestion and relevance feedback in the imagerover www search engine. In *Proc. of Visual 1997*, pages 85–91.

[Tieu and Viola, 2000] Tieu, K. and Viola, P. (2000). Boosting image retrieval. In *Proc. of IEEE CVPR 2000*, pages 1228–1235.

[Torralba and Oliva, 2003] Torralba, A. and Oliva, A. (2003). Statistics of natural image categories. *Network: Computation in Neural Systems*, 14:391–412.

[Torralba and Sinha, 2001] Torralba, A. and Sinha, P. (2001). Indoor scene recognition. Technical Report AI Memo 2001-015, CBCL Memo 202, MIT AI Lab.

[Town and Sinclair, 2000] Town, C. and Sinclair, D. (2000). Content-based image retrieval using semantic visual categories. Technical Report 2000.14, AT&T Laboratories Cambridge.

[Tseng et al., 2003] Tseng, B. et al. (2003). Normalized classifier fusion for semantic visual concept detection. In *Proc. IEEE ICIP 2003.*

[Tuceryan and Jain, 1998] Tuceryan, M. and Jain, A. (1998). Texture analysis. In C.H. Chen, L.F. Pau, P. W., editor, *The Handbook of Pattern Recognition and Computer Vision (2nd Edition)*, pages 207–248. World Scientific Publishing Co.

[Vailaya et al., 2001] Vailaya, A. et al. (2001). Bayesian framework for hierarchical semantic classification of vacation images. *IEEE Trans. on Image Processing*, 10(1):117–130.

[Vapnik, 1979] Vapnik, V. (1979). *Estimation of Dependences Based on Empirical Data [in Russian].* Nauka, Moscow (English translation: Springer Verlag, New York, 1982).

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

[Vapnik, 1998] Vapnik, V. (1998). *Statistical Learning Theory.* Wiley, New York.

[Vasconcelos and Lippman, 2000] Vasconcelos, N. and Lippman, A. (2000). A probabilistic architecture for content-based image retrieval. In *Proc. of IEEE CVPR 2000*, pages 1216–1221.

[Vendrig and Worring, 2003] Vendrig, J. and Worring, M. (2003). Interactive adaptive movie annotation. *IEEE Multimedia*, 10(3):30–37.

[W3C, 2001] W3C (2001). Synchronized multimedia integration language (smil 2.0). http://www.w3.org/TR/smil20/.

[Wang et al., 2001] Wang, J., Li, J., and Wiederhold, G. (2001). Simplicity: semantics-sensitive integrated matching for picture libraries. *IEEE Trans. on PAMI*, 23(9):947–963.

[Wang et al., 2003] Wang, L., Chan, K., and Zhang, Z. (2003). Bootstrapping svm active learning by incorporating unlabelled images for image retrieval. In *Proc. of IEEE CVPR 2003.*

[Weber et al., 2000] Weber, M., Welling, M., and Perona, P. (2000). Unsupervised learning of models for recognition. In *Proc. of ECCV 2000*, pages 18–32.

[Wu et al., 2000a] Wu, J., Lim, J., and Hong, D. (2000a). Toward semantics level indexing and retrieval of images and video. In *Proc. of Real World Computing Symposium, Tokyo, Japan, Jan. 17-19, 2000*, pages 159–164.

[Wu et al., 2000b] Wu, Y., Tian, Q., and Huang, T. (2000b). Discriminant-em algorithm with application to image retrieval. In *Proc. of IEEE CVPR 2000*, pages 1222–1227.

[Xie et al., 2004] Xie, L., Xu, P., and Chang, S.-F. (2004). Structure analysis of soccer video with domain knowledge and hidden markov models. *Pattern Recognition Letters*, page (to appear).

[Zhao et al., 2000] Zhao, W., Chellappa, R., Rosenfeld, A., and Phillips, P. (2000). Face recognition: a literature survey. Technical Report UMD CAR-TR-948, University of Maryland.

[Zhu et al., 2002] Zhu, L., Rao, A., and Zhang, A. (2002). Theory of keyblock-based image retrieval. *ACM Trans. on Information Systems*, 20:224–257.