

University of Louisville

ThinkIR: The University of Louisville's Institutional Repository

Electronic Theses and Dissertations

5-2009

Bridging the semantic gap in content-based image retrieval.

Joshua David Caudill
University of Louisville

Follow this and additional works at: <https://ir.library.louisville.edu/etd>

Recommended Citation

Caudill, Joshua David, "Bridging the semantic gap in content-based image retrieval." (2009). *Electronic Theses and Dissertations*. Paper 224.
<https://doi.org/10.18297/etd/224>

This Doctoral Dissertation is brought to you for free and open access by ThinkIR: The University of Louisville's Institutional Repository. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of ThinkIR: The University of Louisville's Institutional Repository. This title appears here courtesy of the author, who has retained all other copyrights. For more information, please contact thinkir@louisville.edu.

BRIDGING THE SEMANTIC GAP IN CONTENT-BASED IMAGE RETRIEVAL

By

Joshua David Caudill
B.S., CECS, University of Louisville, 2004
M.Eng., CECS, University of Louisville, 2005

A Dissertation
Submitted to the Faculty of the
Graduate School of the University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy

Department of Computer Engineering and Computer Science
University of Louisville
Louisville, Kentucky

May 2009

Copyright 2009 by Joshua David Caudill

All rights reserved

BRIDGING THE SEMANTIC GAP IN CONTENT-BASED IMAGE RETRIEVAL

By

Joshua David Caudill
B.S., CECS, University of Louisville, 2004
M.Eng., CECS, University of Louisville, 2005

A Dissertation Approved On

April 6, 2009

Date

by the following Dissertation Committee:

Dissertation Director

ACKNOWLEDGEMENTS

I would never have been able to finish my dissertation without the guidance of my committee members, help from friends, and support from my family.

I would like to express my deepest gratitude to my advisor, Dr. Hichem Frigui, for his excellent guidance, caring, patience, and providing me with an excellent atmosphere for doing research. I would like to thank the members of my committee, Dr. Adel Elmaghraby, Dr. James Graham, Dr. Olfa Nasraoui, and Dr. Patricia Cerrito, for their continual guidance and support. Many thanks to the Computer Engineering and Computer Science Department at the University of Louisville and the National Science Foundation for providing the financial support for this research.

I would like to thank Jason Meredith, who as a good friend, was always willing to help and give his best suggestions. It would have been a lonely lab without him. Many thanks to Oualid Missaoui, Aleksey Fadeev, Naouel Baili, Ouïem Bchir, Ahmed Chamseddine, Anis Hamdi, Lijun Zhang, Andrew D Karem, Hisham Zaghloul, Mohamed Ben Ismail and other researchers in the Multimedia Research Lab for helping me collect data and perform experiments. My research would not have been possible without their help.

I would like to thank Britney Broyles, whose friendship, support, and strength helped me maintain my focus and pushed me to complete my research. I would also like to thank Matthew Wright, who picked me up when I was down and made sure my spiritual path was as strong as my academic. Many thanks to Christy

Gearheart and Ashley Meredith for their countless hours revising this dissertation and other papers.

I would like to express my gratitude to my parents for their love, guidance, and understanding. I am also grateful to my many friends, family, and brother for their love and moral support. Finally, I would like to dedicate this dissertation to my grandmother, Dr. Maggie Miller, for all her years of proofing papers and being my inspiration to go farther in school.

ABSTRACT

BRIDGING THE SEMANTIC GAP IN CONTENT-BASED IMAGE RETRIEVAL

Joshua David Caudill

May 9, 2009

To manage large image databases, Content-Based Image Retrieval (CBIR) emerged as a new research subject. CBIR involves the development of automated methods to use visual features in searching and retrieving. Unfortunately, the performance of most CBIR systems is inherently constrained by the low-level visual features because they cannot adequately express the user's high-level concepts. This is known as the semantic gap problem.

This dissertation introduces a new approach to CBIR that attempts to bridge the semantic gap. Our approach includes four components. The first one learns a multi-modal thesaurus that associates low-level visual profiles with high-level keywords. This is accomplished through image segmentation, feature extraction, and clustering of image regions. The second component uses the thesaurus to annotate images in an unsupervised way. This is accomplished through fuzzy membership functions to label new regions based on their proximity to the profiles in the thesaurus. The third component consists of an efficient and effective method for fusing the retrieval results from the multi-modal features. Our method is based

on learning and adapting fuzzy membership functions to the distribution of the features' distances and assigning a degree of worthiness to each feature. The fourth component provides the user with the option to perform hybrid querying and query expansion. This allows the enrichment of a visual query with textual data extracted from the automatically labeled images in the database.

The four components are integrated into a complete CBIR system that can run in three different and complementary modes. The first mode allows the user to query using an example image. The second mode allows the user to specify positive and/or negative sample regions that should or should not be included in the retrieved images. The third mode uses a Graphical Text Interface to allow the user to browse the database interactively using a combination of low-level features and high-level concepts.

The proposed system and all of its components and modes are implemented and validated using a large data collection for accuracy, performance, and improvement over traditional CBIR techniques.

TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS	iii
ABSTRACT	v
LIST OF TABLES	xi
LIST OF FIGURES	xii
CHAPTER	
I INTRODUCTION	1
II RELATED WORK	10
A Feature Extraction	12
1 Color Features	12
2 Texture Features	13
3 Shape Features	13
B Distance Measures	14
C Segmentation and Indexing	15
1 The Fuzzy C-Means (FCM) Algorithm	16
2 The Competitive Agglomeration (CA) Algorithm	17
3 The Simultaneous Clustering and Attribute Discrimina- tion	18
4 Self-Organization of Oscillators Network (SOON) Algorithm	24
5 Self Organization and Visual Exploration of Large Data Sets	26
D User Interfaces for CBIR	27

1	Query-by-Visual-Example	27
2	Query-by-Visual-Region	28
E	Other Issues in CBIR	28
1	Fusion of Multiple Sets of Features	28
2	Bridging the Semantic Gap Problem	29
3	Image Annotation	30
III	CLUSTERING AND FEATURE DISCRIMINATION . .	40
A	Simultaneous Clustering and Attribute Discrimination (SCAD)	41
B	Case of Unknown Number of Clusters	44
C	Initialization & Distance Normalization for SCAD _c -CA	46
D	Annealing Schedule for Feature Discrimination	47
IV	UNSUPERVISED IMAGE REGION ANNOTATION . .	49
A	Multi-Modal Thesaurus Construction	50
1	Training Data Set	50
2	Feature Extraction and Vector Representation of Images	51
3	Learning Associations Between Visual Features and Key- words	56
4	Multi-Modal Thesaurus Construction	58
B	Unsupervised Image Annotation	59
1	Fuzzy Membership Generation	60
2	Keyword Weighting	61
C	Experimental Results and Validation	62
1	Comparative Analysis	67
D	Conclusions	70

V	FUSION OF MULTI-MODAL FEATURES FOR IMAGE	
	RETRIEVAL	74
A	Feature Descriptors	75
B	Multi-Modal Feature Fusion	79
1	Distance Mapping	79
2	Feature Relevance Weights	81
3	Feature Fusion	81
4	Hybrid Query and Query Expansion	83
C	Experimental Validation	84
1	Hybrid Query and Query Expansion	85
2	Fusion of Multiple Feature Sets	87
3	Subjective Evaluation	90
D	Conclusions	96
VI	REGION-BASED IMAGE RETRIEVAL	98
A	Hybrid Region Indexing	100
B	Retrieval by Boolean Composition	102
1	Query by Boolean Composition of Visual Prototypes	104
2	Query by Boolean Composition of Keywords	104
3	Query by Hybrid Boolean Composition	105
C	Ranking of Hybrid Boolean Composition	105
D	Experimental Results	106
E	Conclusions	109
VII	SEMANTIC VISUALIZATION AND NAVIGATION	111
A	Self Organization and Visual Exploration of Large Multi-Modal	
	Data Sets	112

B	Graphical Text Interface	114
C	Conclusions	118
VIII CONCLUSIONS AND FUTURE WORK		119
REFERENCES		125
CURRICULUM VITAE		136

LIST OF TABLES

TABLE		Page
1	List of words used to label the training images	64
2	Experimental Constant Values used in Application of SCAD _c -CA . .	65
3	Feature relevance weights for the sample clusters shown in Figure 13.	65
4	Accuracy of the four labeling methods averaged over all keywords. . .	70
5	Sample query images and the number of relevant images (among the top 50 retrieved) for each feature.	79
6	Feature relevance weights of the six individual feature sets determined through precision/recall used in fusion.	89

LIST OF FIGURES

FIGURE	Page
1	Illustration of the semantic gap problem. The image on the top represents the initial query. The other six images are retrieved by a typical CBIR system that uses visual features only. 3
2	Relationship of Core CBIR Topics 4
3	Overview of the proposed CBIR system. 5
4	Overview of the GUI to the proposed CBIR system. 7
5	Content-based image retrieval framework. 11
6	Illustration of the benefits of adding textual information to CBIR. (a) The query image. (b) The most similar 6 images using only low-level visual features. (c) Textual information is added to the query image. (d) The most similar images using visual and textual features. 32
7	Highlighted architecture of the proposed CBIR system component to perform unsupervised image annotation. 51
8	Examples of image-level annotations that refer to different and not segmented regions in the image. 52
9	Examples of segmented images using the CA; (a) Original Images, (b) Coarsely segmented images. 53
10	Representation of visual and textual features 56
11	Illustration of the image segmentation and annotation. (a) Image to be annotated. (b) The three regions of the image in (a). (c) Best three profiles matched to each region. 61

12	Segmentation and annotation of six images. The regions' boundaries are displayed as white lines, and the top word with its evidence is shown for each region.	62
13	Representative regions from 4 sample clusters. For each cluster few regions assigned to it are shown (The gray part of the image includes other regions not assigned to this cluster). The keywords above each image are those used to provide a global image annotation.	66
14	Visual profiles of the clusters in Figure 13. The six feature sets are shown with their representative regions. The $F^{H_{RGB}}$ is shown as a 64-bin histogram. The $F^{M_{HSV}}$ and $F^{M_{LUV}}$ moments are displayed as their mean color. The $F^{E_{HD}}$ is shown as a 5-bin bar plot representing the various angles. The $F^{W_{TD}}$ is shown with the mean and standard deviation of each frequency bank. The $F^{S_{HP}}$ feature lists the five values. Finally the dominant keywords in the cluster are shown as $F^{T_{XT}}$ (User Provided Keywords). To the right of each feature set, we show its relevance weight.	72
15	Comparison of the annotation accuracy using the proposed method and three other different annotation algorithms.	73
16	Architecture of the proposed CBIR system with the component that performs multi-modal querying and retrieval highlighted.	75
17	Precision/Recall curves of the six individual feature sets	78
18	Piecewise linear membership function, $\mu_i(d_i)$ used to map the distances of feature set i into membership values. A, B, and C correspond to the averages of the distances of the three closest images, the three images ranked at the middle, and the three furthest images respectively. . . .	81

19	Recall and precision of visual only features versus a hybrid query of visual and textual using query expansion. The results are averaged over the 1000 test images.	86
20	Sample query image where query expansion improves the results significantly. The first image is the query image. The others are the top 19 retrieved images, where X indicates that the image is from the same category, and thus is relevant. Most of the retrieved butterfly images do not share the same low-level features but they are labeled by the same set of keywords.	87
21	Sample query image where query expansion improves the results significantly. The first image is the query image. The others are the top 19 retrieved images, where X indicates that the image is from the same category, and thus is relevant. Labeling by the correct keywords allows the retrieval of images missed by the low-level features.	88
22	Precision/Recall of the proposed fusion method (Sum of weighted fuzzy memberships) and the standard sum of scaled distances. The performance of the best individual feature (F^{CSD}) is shown as a reference. .	90
23	Sample query image where the fusion based on the sum of weighted memberships outperforms the other methods. The first image is the query image. The others are the top 9 retrieved images, where X indicates that the image is from the same category, and thus is relevant.	91
24	Sample query image where the fusion based on the Choquet integral outperforms the other methods. The first image is the query image. The others are the top 9 retrieved images, where X indicates that the image is from the same category, and thus is relevant.	92
25	Screen shot of the Subjective Test Online Interface.	94

26	Average Subjective Test Results for Individual Queries.	95
27	Overall Average User Satisfaction for Algorithms, a) Using Top 10 Images to Query and b) Using Top 5 Images to Query.	95
28	Overall User Satisfaction Classified by Experience.	96
29	Overview of the proposed CBIR system component to perform region based image retrieval.	100
30	Snapshot of the Region Query Interface.	103
31	20 samples from the 200 Category Representatives	107
32	Positive and Negative query categories selected by the user to formulate a query. Here, the user is looking for images that contain horse/deer on grass but no flowers.	108
33	Comparison of the precision values for the query regions in Fig. 32 with and without textual features.	108
34	Comparison of the precision values when querying with and without tex- tual features, and when using the fuzzy neighborhood and ranked hybrid methods. The values are averaged over 30 queries.	109
35	Overview of the proposed CBIR system component to visualize high- dimensional visual and textual data.	113
36	Graphical Text Interface: Walk through from view of semantic database to selection of query image. (a) Initial semantic view of the system. (b) Zoomed in region. (c) Zoomed in region from image view. (d) Selecting an image after zooming.	116
37	Illustration of the filtering option to constrain the elements displayed based on selected keywords. (a) initial view for all semantic concepts. (b) view when only clusters signifying "grass" or "flower" are shown.	117

38	Reorganization of the axes to combine visual and textual features in the map. (a) view when both axes are based on textual features at different resolutions. (b) view when vertical axis is re-organized based on a color feature set.	117
----	---	-----

CHAPTER I

INTRODUCTION

Recent technological advances in the capture, storage, and transmission of large digital image and video collections, coupled with the steady growth of the internet, has created a need for intelligent methods to analyze this data effectively. As personal collections [1, 2] and online photo sharing communities [3, 4] grow, efficient storage techniques and digital libraries are being created. These large libraries have made it necessary to develop automated tools for storing, retrieving, organizing, and mining large multimedia databases to supplement traditional methods based on keyword indexing and retrieval. Image data offers unique advantages because it is relatively easy for humans to explore and interpret; However, for computer methods, it poses serious challenges.

To manage image databases, Content-Based Image Retrieval (CBIR) emerged as a new research subject [5, 6]. CBIR is the application of research techniques from various areas such as databases, pattern recognition, data mining, image processing, and multimedia to index and retrieve digital images in large databases. In particular, CBIR involves the development of automated methods that are able to recognize visual features of images - such as color, texture, and shape [7, 8]-, and to make use of this information to search, retrieve, and browse large image databases.

CBIR systems have been in existence for approximately two decades. While some applications have been built for commercial use [9, 10, 11], most exist within the university research domain [12, 13, 14] and in recent years have seen a surge in

prototype development (see [5, 15] for examples). Researchers have focused on various topics such as low-level image feature representation [16, 17, 18, 19], distance metrics [20, 21, 22, 23], visualization and navigation [24, 25], database categorization [26, 27, 28, 29, 30], and relevance feedback [31, 32]. Depending on the application domain, the size of the image collection, and the available a priori information, the above CBIR research topics exhibit varying degrees of difficulty.

As image databases continue to increase in size and become more complex in content, it is becoming increasingly more difficult to achieve high accuracy in retrieving visually similar images with a single feature set. As a result, diverse feature sets are being used and combined to provide a more accurate retrieval. Unfortunately the addition of low-level visual features has proved insufficient to improve the performance of CBIR systems. In particular, for large and generic image databases, the performance of most CBIR systems is inherently constrained by the low-level features because they cannot adequately express the user's high-level concepts. The gap in knowledge and understanding between low-level features and high-level concepts is known as the semantic gap [33, 34]. This problem is illustrated in Figure 1. The first image in this figure is the query image. Here, it is assumed that the user is searching for images that have flowers. A typical CBIR system would retrieve the images shown below the query image. As it can be seen, these images have similar color information. However, conceptually, they are very different.

In an attempt to bridge this gap, few approaches that integrate low-level visual features and textual keywords have been proposed [35, 36, 37, 38, 39, 40]. Unfortunately, manually labeling each image by a set of keywords is subjective and labor intensive. Moreover, region labeling (as opposed to global image labeling) may be needed, which makes manual labeling more tedious. To address this limitation,



Figure 1. Illustration of the semantic gap problem. The image on the top represents the initial query. The other six images are retrieved by a typical CBIR system that uses visual features only.

several algorithms that can annotate images/regions in an unsupervised (or semi-supervised) have been proposed in the past few years [38, 41, 39, 42, 43].

Another research issue in CBIR is how to combine the output of diverse multi-modal feature sets. This task has been overlooked by the CBIR community. Only methods that are based on distance scaling or normalization and simple list merging have been used [44]. In fact, different features can vary significantly with respect to the number of attributes, the dynamic ranges, and the adopted distance measures. Thus, fusion of these features is not trivial and can have a significant impact on the overall performance of the CBIR system.

The goal of this thesis is to develop algorithms that address the image annotation and the multi-modal feature fusion tasks. Our proposed approach combines topics from pattern recognition, data mining, image processing, and

multimedia to build an efficient CBIR system. The interactive relationship between these topics and their sub-topics is illustrated in Figure 2. The technical overview of our proposed system is illustrated in Figure 3. It has four main components that are highlighted below.

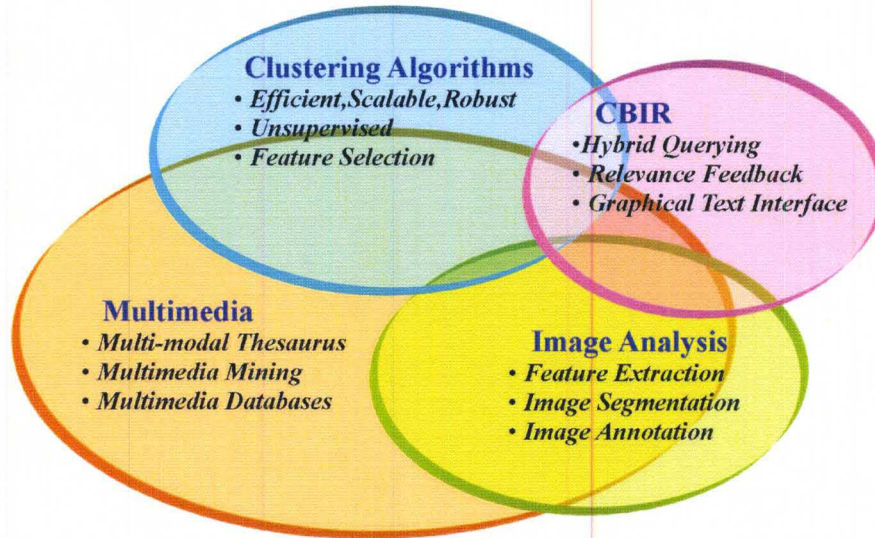


Figure 2. Relationship of Core CBIR Topics

1. **Learning Thesaurus:** This component uses a set of training images, that are annotated manually, to create a multi-modal thesaurus through clustering and feature weighting. The objective is to extract representative visual profiles corresponding to frequent homogeneous regions, and to associate these profiles with keywords. To accomplish this, the training images are segmented into homogeneous regions. Then, these regions are represented by visual descriptors combined with the image level annotations, and clustered into categories of regions that share common attributes. Clusters' representatives and their parameters are used to create profiles linking low-level image features and high-level concepts.
2. **Feature Extraction:** This component uses the developed multi-modal thesaurus to automatically annotate image regions. This is accomplished

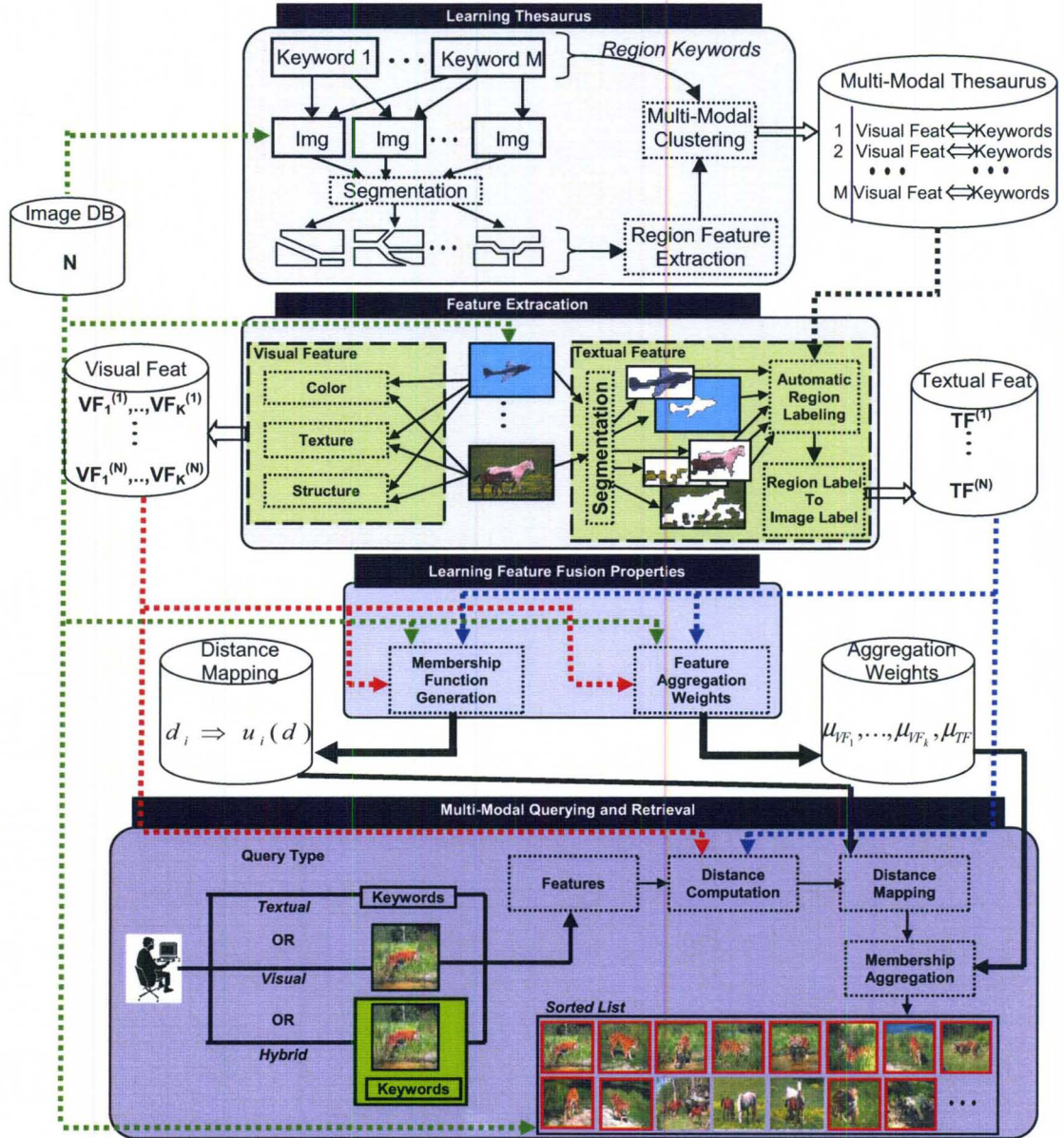


Figure 3. Overview of the proposed CBIR system.

through two steps. First, an un-annotated image is segmented into homogeneous regions. Then, fuzzy memberships are assigned to the regions that reflect their proximity to the thesaurus entries. These annotated regions can then facilitate textual region based searches, or be aggregated into image level annotations.

3. **Learning Feature Fusion Properties:** This component implements an efficient and effective method for fusing the retrieval results of the multi-modal features. Our method is based on learning and adapting fuzzy membership functions to the distribution of the features' distances. These memberships are then used to aggregate the results of the different features.
4. **Multi-Modal Querying and Retrieval:** This component uses the multi-modal thesaurus to perform hybrid querying and query expansion in the CBIR search process. Query expansion allows the enrichment of a visual query with textual data associated with the image. In particular, the images in the database, annotated using the second component, are made available to the hybrid queries to enrich the feature set and improve the relevancy of the retrieved images.

The above four components are integrated into a complete CBIR system comprised of three main query retrieval modes. The first mode is a classic CBIR retrieval with all four components integrated as shown in Figure 3. In the second mode, our system uses a novel region representation that allows the user to formulate a query by combining multiple regions of interest. This mode is useful when the user has a mental picture of what he/she is looking for but does not have an example image to initiate the query. The final mode uses a novel graphical text interface to perform semantic visualization and navigation. This mode allows for the initial navigation to be oriented around high-level concepts instead of randomly-selected images. The last two modes are illustrated in Figure 4.

This dissertation makes the following contributions to the area of Content-Based Image Retrieval:

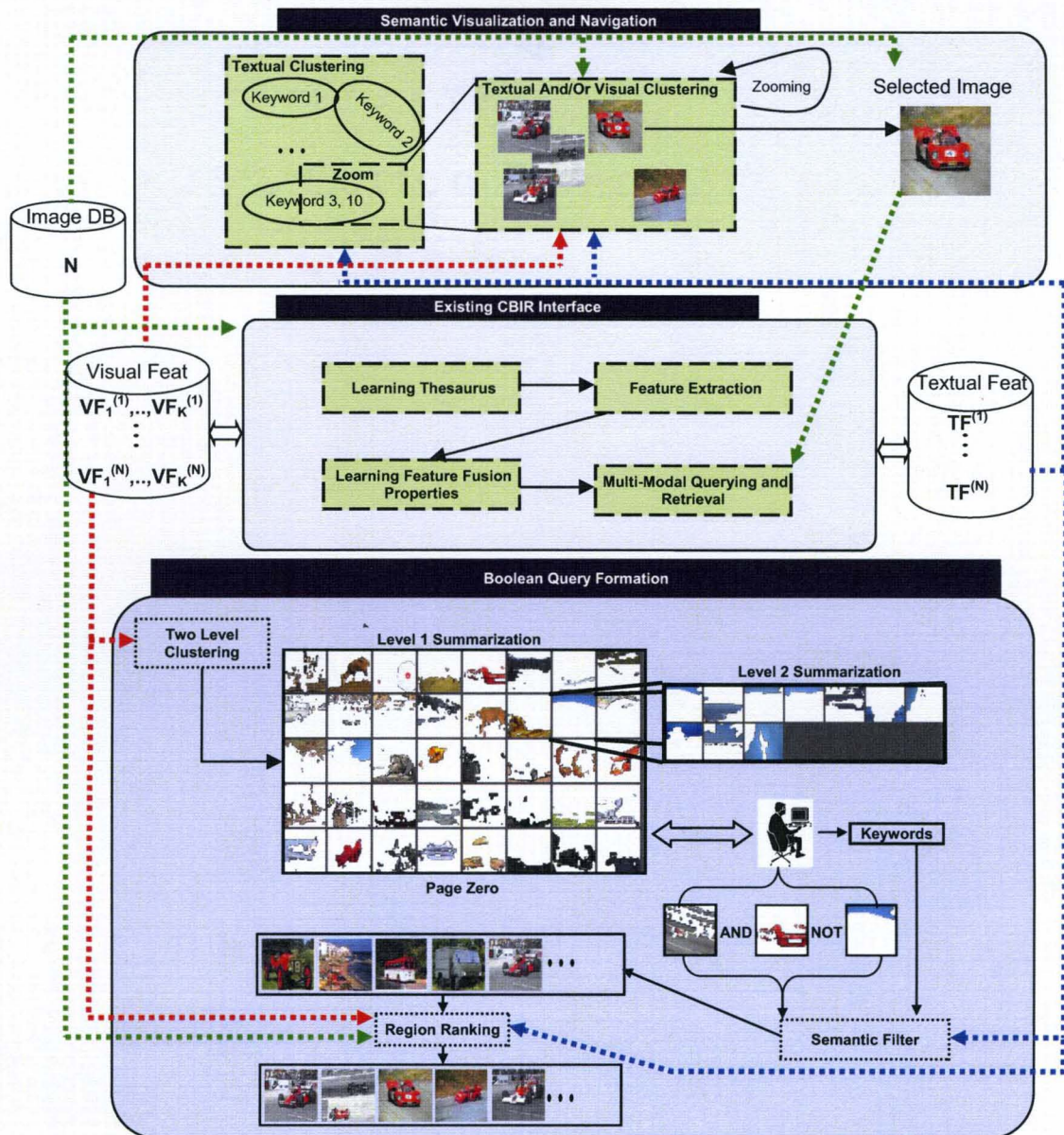


Figure 4. Overview of the GUI to the proposed CBIR system.

- Adaptation of an efficient clustering algorithm for image databases.
- Learning a multi-modal thesaurus to convert from one modality to another.
- Creating a novel approach to perform unsupervised region-based image annotation.

- Two efficient and effective methods for fusing the retrieval results of multi-modal features.
- Implementation of hybrid querying, query expansion, and concept refining.
- Efficient region-based image retrieval facilitated by boolean composition.
- A graphical text interface that visualizes high-dimensional multi-modal data for browsing and navigation in a two-dimensional platform.

The proposed CBIR system and its various components are validated using a large data set for accuracy, performance, and improvement over basic CBIR techniques. Our proposed image annotation algorithm outperforms three state-of-the-art approaches on average by 13% when labeling 10,000 images. Our efficient method for fusing the output of multi-modal features yields 6% higher precision on average than standard CBIR methods and 16% better retrieval performance than the best individual feature. Lastly, our region-based retrieval is 30% better than a similar state-of-the-art approach.

The CBIR system is implemented as a java framework built on a C# server. The server application maintains all data, clustering, and distance calculations in local memory. This currently places an upper limit on the size of our database equal to the amount of RAM on the machine. An image database of 55,000 with 6 multi-modal feature sets requires 1GB of memory. Using this implementation approach however, we average 0.83s on a query with the 55,000 image database using a 3.4Ghz Pentium IV with 4GB of RAM.

The organization of the rest of this dissertation is as follows: Chapter Two gives an overview of typical CBIR systems and their various components. Chapter Three contains our adaptations to an existing clustering algorithm for the CBIR application. Chapter Four describes our proposed approach to build a multi-modal

thesaurus in an unsupervised manner and its application to unsupervised image labeling. Chapter Five describes how to formulate hybrid queries and combine the output of the diverse feature sets. Chapter Six describes our proposed region representation that allows the user to formulate a query by combining multiple regions of interest. Chapter Seven describes the graphical interface of our CBIR system to perform semantic visualization and navigation. Finally, Chapter Eight gives our conclusions and highlights potential future work.

CHAPTER II

RELATED WORK

Content-Based Image Retrieval (CBIR) is the application of research techniques from various areas such as pattern recognition, data mining, image processing, and multimedia to index and retrieve digital images in large databases. In particular, CBIR involves the development of automated methods that are able to recognize visual features of images, and to make use of this information to index, search, retrieve, and browse large image databases. CBIR methods do not rely on human-inputted information such as captions or keywords, but more so on the content of the images themselves. Over the past few years, several CBIR prototypes have been developed (see [5, 15, 45, 46, 47, 30, 48, 49, 50, 51, 52] for examples). Most CBIR systems can be conceptually described by the framework depicted in Figure II.

CBIR systems make use of various types of user queries; most commonly, *query by sketch* and *query by example*. In query by sketch, a user draws a rough approximation and the system locates images matching the sketch. In query by example, the user selects an image that is representative of what he/she is looking for and the system retrieves the most similar images from the database. In almost all query approaches to CBIR, when an example image is given, its visual features are extracted and used to match against those in the database. Well defined distance measures are then used to compute the similarity between the query image and images in the database. The images are sorted according to their distance to

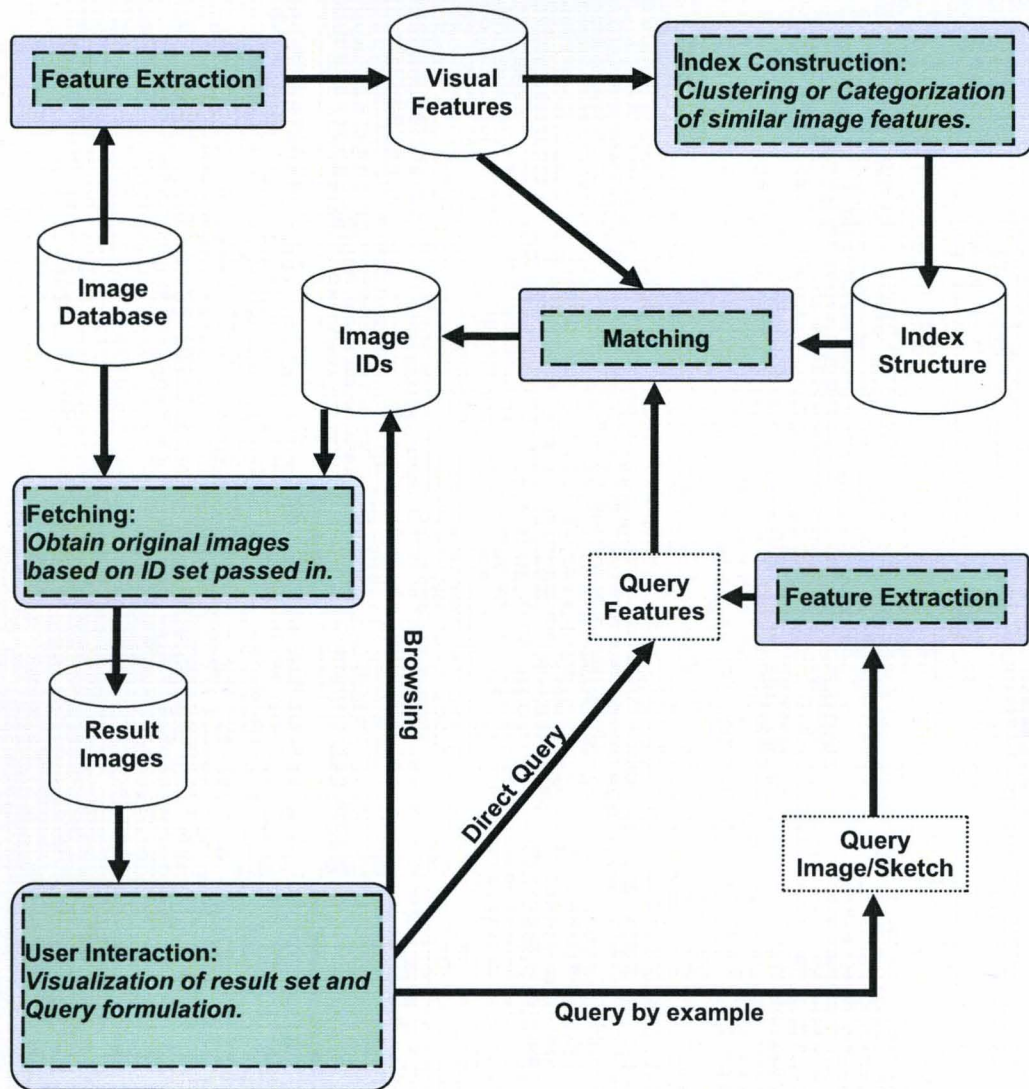


Figure 5. Content-based image retrieval framework.

the query, and the top k images are presented to the user.

In this chapter we outline the main steps involved in developing a complete CBIR system. While analyzing what it takes to make a CBIR system function, we will look at various approaches currently implemented.

A Feature Extraction

In a CBIR system, various visual features are extracted offline from each image for indexing purposes. Generic systems (e.g. [12, 13, 14]) make use of low-level features such as texture, color, and shape [7, 8]. Other specialized systems make use of higher-level features such as faces [53]. Searching for relevant images in a database is then converted to the problem of identifying images that have similar features. In the following, we outline common features that have been used extensively in the literatures. Most of these features have been adopted in the MPEG-7 standard [54].

1 Color Features

Since color is immediately perceived by humans when looking at an image, color features are the most widely used visual features. When using color features in CBIR, factors such as model selection, quantization and feature representation must be taken into consideration. The purpose of a color model is to facilitate the specification of colors in a standard way [55]. Common color models include Red-Green-Blue (RGB), Cyan-Magenta-Yellow (CMY), Cyan-Magenta-Yellow-Black (CMYK), Luma-Chrominance (YIQ), Hue-Saturation-Lightness (HLS), Hue-Saturation-Value (HSV), Luma-Blue Difference Chroma-Red Difference Chroma (YCbCr) and CIE-LUV.

Prior to describing an image by a color feature, the color spaces needs to be quantized. This step is used to reduce the possible available colors present to a smaller number. Color features are then extracted and represented by some color feature representation. There are many different schemes of varying complexity. For instance, the MPEG-7 standard [54] includes the Color Structure Descriptor, Scalable Color Descriptor, and the Dominant Color Descriptor [54]. The most

common and simplest approach used to represent an image's color feature is a color histogram, which represents the color distribution of an image.

2 Texture Features

The image texture feature describes properties such as smoothness, coarseness, and regularity. MPEG-7 standard includes various texture descriptors such as Edge Histogram Descriptor, Homogenous Texture Descriptor, and Perceptual Browsing Descriptor. Other commonly used texture features include the Gabor Wavelet Texture Feature [56], Tamura Texture Features [57, 58], and Wold Texture Features [59].

3 Shape Features

Shape information of objects in images is a very important visual feature. Few CBIR systems offer searching based on shape-based techniques. This is because these systems require image segmentation and describing each region by its shape feature. Segmentation is a very important step for the extraction of these features and includes procedures such as image smoothing, noise removal, and edge detection.

Shape-based techniques are categorized into boundary-based and region-based. Boundary-based shape representation uses the outer edges of objects while region-based uses the entire shape region. MPEG-7 shape features include Contour Shape Descriptor, Curvature Scale Space Descriptor, and the Angular Radial Transform. Other commonly used shape features include the Fourier Descriptor [60] and Moment Invariants [61].

B Distance Measures

Once all images in the database have their features extracted and stored, a method for comparing them needs to be established. This is usually accomplished using a distance measure. These measures will provide the means of ranking the images in the database to a query image and returning the most relevant ones. In the following, we outline some distance measures commonly used for this task.

The most common distance measure between two points x and y is the Euclidean distance (also referred to as the L2 distance), and is defined as

$$d^2(x, y) = \|x - y\|^2. \quad (1)$$

However, this distance can be unreliable and un-intuitive when the points in the feature space do not have a spherical distribution.

A more reliable distance measure is the Mahalanobis distance, defined as

$$d^2(x, y) = (x - y)^T \sum^{-1} (x - y),$$

where \sum^{-1} is the covariance matrix of all points (i.e. images) in the feature space. The Mahalanobis distance can represent data with non-spherical distribution as it takes into account the correlations among different features. The drawbacks of this distance include its high computational complexity, and the problems associated with computing and inverting the covariance matrix (\sum^{-1}) for high dimensional features spaces. A simple solution to this is to use a diagonal matrix instead of the full covariance matrix.

For some features the Mahalanobis and Euclidean distances may not be appropriate. Color histograms, for instance, need to consider the similarity between neighboring bins. The Quadratic distance [62], defined as

$$d^2(x, y) = (x - y)^T \mathbf{A} (x - y), \quad (2)$$

is commonly used for this task. In (2), \mathbf{A} is a matrix of weights denoting the similarity between bins in a histogram vector.

Other features, such as textual features, are high dimensional and tend to be binary. A typical distance used for this type of data is the Cosine distance, given by

$$d^2(x, y) = 1 - \frac{x \bullet y}{\|x\| \|y\|}. \quad (3)$$

This distance has been used extensively in text retrieval [63, 64].

C Segmentation and Indexing

Once the features of an image are extracted and the distance measures are defined, the next step in a typical CBIR system is to index the images for scalability and retrieval speed in large databases. Some systems index the images globally, others segment the images and index their regions. In both cases, clustering algorithms are the main tools used for segmentation and indexing. Clustering is the partitioning of a data set into subsets (clusters), so that the data in each subset share some common trait. A trait (feature) is defined as common with respect to a defined distance measure. The advantages of clustering are its unsupervised learning ability, and capability to support many distance measures. The most common clustering algorithms and techniques used in this thesis are outlined below.

Let $\mathcal{X} = \{\mathbf{x}_j \in \mathbb{R}^p | j=1, \dots, N\}$ be a set of N feature vectors in an p -dimensional feature space. Let $\mathbf{B} = (\beta_1, \dots, \beta_c)$ represent a C -tuple of prototypes each of which characterizes one of the C clusters. Each β_i consists of a set of parameters. Let u_{ij} represent the membership of \mathbf{x}_j in cluster β_i . The $C \times N$ fuzzy C -partition, $\mathbf{U} = [u_{ij}]$, satisfies [65]:

$$\begin{cases} u_{ij} \in [0, 1], & \forall i \\ 0 < \sum_{j=1}^N u_{ij} < N & \forall i, j \\ \sum_{i=1}^C u_{ij} = 1 & \forall j \end{cases} \quad (4)$$

1 The Fuzzy C-Means (FCM) Algorithm

The Fuzzy C-Means (FCM) algorithm [66] formulates the problem of fuzzily partitioning the N feature vectors into C clusters as minimization of the following objective function:

$$J(\mathbf{B}, \mathbf{U}; \mathcal{X}) = \sum_{i=1}^C \sum_{j=1}^N (u_{ij})^m d^2(\mathbf{x}_j, \beta_i). \quad (5)$$

In (5), $m \in (1, \infty)$ is a weighting exponent (fuzzifier) and $d(\mathbf{x}_j, \beta_i)$ is the distance from feature point \mathbf{x}_j to prototype β_i . Minimization of (4) with respect to \mathbf{U} , subject to the constraints in (4), gives us [66]

$$\left. \begin{aligned} u_{ij} &= \frac{1}{\sum_{k=1}^C \left(\frac{d^2(\mathbf{x}_j, \beta_i)}{d^2(\mathbf{x}_j, \beta_k)} \right)^{\frac{1}{m-1}}} && \text{if } I_j = 0 \\ u_{ij} &= 0 && \text{if } i \notin I_j \\ \sum_{i \in I_j} u_{ij} &= 1 && \text{if } i \in I_j \end{aligned} \right\} \text{if } I_j \neq 0 \quad (6)$$

where $I_j = \{i | 1 \leq i \leq C, d^2(\mathbf{x}_j, \beta_i) = 0\}$.

If the Euclidean distance

$$d^2(\mathbf{x}_j, \beta_i) = d_{ij}^2 = \|\mathbf{x}_j - \mathbf{c}_i\|^2, \quad (7)$$

where \mathbf{c}_i is the center of the i^{th} cluster is used in (5), then the FCM will seek spherical clusters. The update equation for the centroids is obtained by fixing the membership values and minimizing (5) with respect to \mathbf{c}_i . This minimization yields

$$\mathbf{c}_i = \frac{\sum_{j=1}^N (u_{ij})^m \mathbf{x}_j}{\sum_{j=1}^N (u_{ij})^m}. \quad (8)$$

The FCM algorithm is summarized below:

Fuzzy C-Means Algorithm

Fix the number of clusters C ;
 Fix m , $m \in (1, \infty)$;
Repeat
 Compute $d(\mathbf{x}_j, \beta_i)$ using (7);
 Update the partition matrix $U^{(k)}$ using (6);
 Update the centers using (8);
Until($\|\Delta \mathbf{U}\| < \epsilon$)

2 The Competitive Agglomeration (CA) Algorithm

The FCM requires specification of the expected number of clusters a priori. In many real applications, this may not be possible. In this case, several approaches to find the optimal C can be used [67, 68, 69, 70, 71]. In particular, the Competitive Agglomeration (CA) [71] is an efficient algorithm that determines the optimal C by minimizing the following objective function:

$$J(\mathbf{B}, \mathbf{U}, \mathcal{X}) = \sum_{i=1}^C \sum_{j=1}^N (u_{ij})^2 d^2(\mathbf{x}_j, \beta_i) - \alpha \sum_{i=1}^C \left[\sum_{j=1}^N u_{ij} \right]^2, \quad (9)$$

subject to the constraints in (4). In (9), $d^2(\mathbf{x}_j, \beta_i)$ represents the distance from feature vector \mathbf{x}_j to prototype β_i . The number of clusters C is dynamically updated in the CA.

Optimization of J with respect to \mathbf{U} yields [71]:

$$u_{st} = u_{st}^{\text{FCM}} + u_{st}^{\text{BIAS}}, \quad (10)$$

where

$$u_{st}^{\text{FCM}} = \frac{[1/d^2(\mathbf{x}_t, \beta_s)]}{\sum_{k=1}^C [1/d^2((\mathbf{x})_t, \beta_k)]}, \quad (11)$$

and

$$u_{st}^{\text{BIAS}} = \frac{\alpha}{d^2(\mathbf{x}_t, \beta_s)} (N_s - \bar{N}_t). \quad (12)$$

In (12),

$$N_s = \sum_{j=1}^N u_{sj} \quad (13)$$

is the fuzzy cardinality of clusters, and

$$\bar{N}_t = \frac{\sum_{k=1}^C [1/d^2(\mathbf{x}_t, \beta_k)] N_k}{\sum_{k=1}^C [1/d^2(\mathbf{x}_t, \beta_k)]}. \quad (14)$$

The choice of α in (9) reflects the importance of the second term relative to the first term. In [71], the authors recommend using

$$\alpha(k) = \eta_0 \exp(-k/\tau) \frac{\sum_{i=1}^C \sum_{j=1}^N (u_{ij})^2 d^2(\mathbf{x}_j, \beta_i)}{\sum_{i=1}^C [\sum_{j=1}^N u_{ij}]^2}, \quad (15)$$

where η_0 is the initial value, τ the time constant, and k is the iteration number. The CA algorithm is summarized below:

Competitive Agglomeration Algorithm

Fix the maximum number of clusters $C = C_{max}$;
Initialize iteration counter $k = 0$ and the fuzzy C partition $\mathbf{U}^{(0)}$;
Compute initial cardinalities N_i for $1 \leq i \leq C$ using (13);
Repeat
 Compute $d^2(\mathbf{x}_j, \beta_i)$ for $1 \leq i \leq C$ and $1 \leq j \leq N$;
 Update $\alpha(k)$ using (15);
 Update the partition matrix $\mathbf{U}^{(k)}$ using (10);
 Compute the cardinality N_i for $1 \leq i \leq C$ using (13);
 If $(N_i < \epsilon_1)$ discard cluster β_i ;
 Update the number of clusters C ;
 Update the prototype parameters;
 $k = k + 1$;
Until(prototype parameters stabilize)

3 The Simultaneous Clustering and Attribute Discrimination

The selection of feature subsets that best represents the given data is an issue concerning the design of a good learning algorithm. The performance of such algorithms suffers from the use of irrelevant features. To address this issue, several methods have been proposed to perform feature selection and weighting [72, 73, 74]. Feature selection completely removes irrelevant features, while feature weighting

extends on selection by assigning continuous weights to the features. Performance can degrade in both instances if feature weights are learned globally and do not take into account the fact that data can be made up of different groups. In this case, the data needs to be partitioned into groups and feature weights need to be learned for each group. One algorithm that can achieve this task is the Simultaneous Clustering and Attribute Discrimination (SCAD) algorithm [75, 76]. SCAD was designed to search for the optimal clusters' prototypes and the optimal relevance weights for each feature of each cluster. Two versions of SCAD were developed. The first one (SCAD₁) balances between two terms in a compound objective function. The second version (SCAD₂), minimizes a single term criterion that implements a discrimination exponent [76].

SCAD₁ Algorithm

SCAD₁ minimizes

$$J(\mathbf{C}, \mathbf{U}, \mathbf{V}; \mathcal{X}) = \sum_{i=1}^C \sum_{j=1}^N (u_{ij})^m \sum_{k=1}^n v_{ik} d_{ijk}^2 + \sum_{i=1}^C \delta_i \sum_{k=1}^n v_{ik}^2, \quad (16)$$

subject to (4) and

$$v_{ik} \in [0, 1] \quad \forall i, k; \quad \text{and} \quad \sum_{k=1}^n v_{ik} = 1, \quad \forall i. \quad (17)$$

In equation (16), v_{ik} represents the relevance weight of feature k in cluster i , and d_{ijk} is given by

$$d_{ijk} = |x_{jk} - c_{ik}|, \quad (18)$$

where x_{jk} is the k th feature value of data point \mathbf{x}_j , and c_{ik} is the k th component of the i th cluster center vector.

Optimization of J with respect to \mathbf{V} yields [75]:

$$v_{ik} = \frac{1}{n} + \frac{1}{2\delta_i} \sum_{j=1}^N (u_{ij})^m \left[\frac{\|\mathbf{x}_j - \mathbf{c}_i\|^2}{n} - d_{ijk}^2 \right], \quad (19)$$

where δ_i is computed in iteration t using

$$\delta_i^{(t)} = K \frac{\sum_{j=1}^N (u_{ij}^{(t-1)})^m \sum_{k=1}^n v_{ik}^{(t-1)} (d_{ijk}^{(t-1)})^2}{\sum_{k=1}^n (v_{ik}^{(t-1)})^2}. \quad (20)$$

The first term in (19), $(1/n)$, is the default value if all n features are treated equally, and no discrimination is performed. The second term is a bias that can be either positive or negative. It is positive for compact feature sets where the partial distance is, on average, less than the total distance (normalized by the number of features). If a feature set is compact, compared to the other features, for most of the points that belong to a given cluster (high u_{ij}), then it is very relevant for that cluster.

Minimization of J with respect to \mathbf{U} , subject to the constraints in (4), yields

$$u_{ij} = \frac{1}{\sum_{k=1}^C \left(\tilde{d}_{ij}^2 / \tilde{d}_{kj}^2 \right)^{\frac{1}{m-1}}}, \quad (21)$$

where $\tilde{d}_{ij}^2 = \sum_{k=1}^n v_{ik} d_{ijk}^2$ is a weighted Euclidean distance.

Minimization of J with respect to \mathbf{C} yields

$$c_{ik} = \begin{cases} 0 & \text{if } v_{ik} = 0, \\ \frac{\sum_{j=1}^N (u_{ij})^m x_{jk}}{\sum_{j=1}^N (u_{ij})^m} & \text{if } v_{ik} > 0. \end{cases} \quad (22)$$

SCAD₁ is an iterative algorithm that starts with an initial partition and alternates between the update equations of u_{ij} , v_{ik} , and c_i . The SCAD₁ algorithm is summarized below:

SCAD₁

Fix the number of clusters C ;
 Fix m , $m \in (1, \infty)$;
 Initialize the centers;
 Initialize the relevance weights to $1/n$;
 Initialize the fuzzy partition matrix \mathbf{U} ;
Repeat
 Compute d_{ijk}^2 for $1 \leq i \leq C$, $1 \leq j \leq N$, and $1 \leq k \leq n$;
 Update the relevance weights v_{ik} by using equation (19);
 Update the partition matrix $U^{(k)}$ by using equation (21);
 Update the centers by using equation (22);
 Update δ_i by using equation (20);
Until(centers stabilize)

SCAD₂ Algorithm

In [76], a new version of SCAD (SCAD₂), that minimizes a single term criterion instead of trying to balance between two terms as in SCAD₁, was proposed. SCAD₂ implements a discrimination exponent [76] to replace the second term in the objective function. It minimizes

$$J(\mathbf{B}, \mathbf{U}, \mathbf{V}; \mathcal{X}) = \sum_{i=1}^C \sum_{j=1}^N (u_{ij})^m \sum_{k=1}^n (v_{ik})^q d_{ijk}^2 \quad (23)$$

subject to (4) and (17). The exponent $q \in (1, \infty)$ is referred to as the discrimination exponent.

Minimization of J with respect to \mathbf{V} yields

$$v_{ik} = \frac{1}{\sum_{t=1}^n \left(\tilde{D}_{ik} / \tilde{D}_{it} \right)^{1/(q-1)}}, \quad (24)$$

where $\tilde{D}_{ik} = \sum_{j=1}^N (u_{ij})^m d_{ijk}^2$ is the measure of dispersion of the i^{th} cluster along the k^{th} dimension, and $\sum_{t=1}^n \tilde{D}_{it}$ is the total dispersion of the i^{th} cluster. In other words, the more compact the i^{th} cluster is along the k^{th} dimension (smaller \tilde{D}_{ik}), the higher the relevance weight, v_{ik} will be for the k^{th} feature.

Minimization of J with respect to \mathbf{U} subject to the constraints in (4) yields

$$u_{ij} = \frac{1}{\sum_{k=1}^C \left[\tilde{d}_{ij}^2 / \tilde{d}_{kj}^2 \right]^{\frac{1}{m-1}}}, \quad (25)$$

where

$$\tilde{d}_{ij}^2 = \sum_{k=1}^n (v_{ik})^q d_{ijk}^2. \quad (26)$$

While the discrimination exponent, q , is needed for finding feature weights v_{ik} , for the purpose of computing the fuzzy memberships, u_{ij} , it was recommended in [76] to set this exponent to one and use the distance measure in SCAD₁:

$$\tilde{d}_{ij}^2 = \sum_{k=1}^n v_{ik} d_{ijk}^2. \quad (27)$$

Minimization of J with respect to the centers \mathbf{C} yields the same equation update as in SCAD₁ (see equation (22)).

The SCAD₂ algorithm is summarized below:

SCAD₂

Fix the number of clusters C ;
 Fix m , $m \in (1, \infty)$;
 Fix the discrimination exponent q , $q \in (1, \infty)$;
 Initialize the centers and fuzzy partition matrix \mathbf{U} ;
 Initialize all the relevance weights to $1/n$;
Repeat
 Compute d_{ijk}^2 for $1 \leq i \leq C$, $1 \leq j \leq N$, and $1 \leq k \leq n$;
 Update the relevance weights matrix \mathbf{V} by using equation (24);
 Compute \tilde{d}_{ij}^2 by using equation (27);
 Update the partition matrix U by using equation (25);
 Update the centers by using equation (22);
Until(centers stabilize)

The Coarse Simultaneous Clustering and Attribute Discrimination Algorithm

Both versions of SCAD were designed to search for the optimal clusters' prototypes and the optimal relevance weight for each feature within each cluster.

However, for high dimensional data, learning a relevance weight for *each* feature may lead to overfitting. To avoid this situation, a coarse approach to feature weighting (called SCAD_c) was proposed in [77]. In SCAD_c, instead of learning a weight for each feature, the set of features is divided into logical subsets, and a weight is learned for each feature subset.

In [77], the authors assume that the p features have been partitioned into K subsets: FS^1, FS^2, \dots, FS^K , and that each subset, FS^s , includes k^s features. Let d_{ij}^s be the partial distance between \mathbf{x}_j and cluster i using the s^{th} feature subset. Let $\mathbf{V} = [v_{is}]$ be the relevance weight for FS^s with respect to cluster i . The total distance, D_{ij} , between \mathbf{x}_j and cluster i is then computed by aggregating the partial distances and their weights, i.e.,

$$D_{ij}^2 = \sum_{s=1}^K v_{is} (d_{ij}^s)^2. \quad (28)$$

SCAD_c minimizes

$$J(\mathbf{B}, \mathbf{U}, \mathbf{V}; \mathcal{X}) = \sum_{i=1}^C \sum_{j=1}^N u_{ij}^m \sum_{s=1}^K v_{is} (d_{ij}^s)^2 + \sum_{i=1}^C \delta_i \sum_{s=1}^K v_{is}^2, \quad (29)$$

subject to equation (4) and

$$v_{is} \in [0, 1] \quad \forall i, s; \quad \text{and} \quad \sum_{s=1}^K v_{is} = 1, \quad \forall i. \quad (30)$$

Optimization of J with respect to \mathbf{V} yields

$$v_{is} = \frac{1}{K} + \frac{1}{2\delta_i} \sum_{j=1}^N (u_{ij})^m \left[D_{ij}^2/K - (d_{ij}^s)^2 \right]. \quad (31)$$

Minimization of J with respect to \mathbf{U} , subject to the constraints in (4), yields

$$u_{ij} = \frac{1}{\sum_{k=1}^C (D_{ij}^2/D_{kj}^2)^{\frac{1}{m-1}}}. \quad (32)$$

Minimization of J with respect to the prototype parameters depends on the choice of d_{ij}^s . Since the partial distances are treated independent of each other (i.e., disjoint

feature subsets), and since the second term in (29) does not depend on the prototype parameters explicitly, the objective function in (29) can be decomposed into K independent problems:

$$J_s = \sum_{i=1}^C \sum_{j=1}^N u_{ij}^m v_{is} (d_{ij}^s)^2, \quad \text{for } s = 1, \dots, K. \quad (33)$$

Each J_s would be optimized with respect to a different set of prototype parameters.

For instance, if d_{ij}^s is an Euclidean distance, minimization of J_s would yield the following update equation for the centers of subset s ,

$$\mathbf{c}_i^s = \frac{\sum_{j=1}^N u_{ij}^m \mathbf{x}_j^s}{\sum_{j=1}^N u_{ij}^m}. \quad (34)$$

SCAD_c is summarized below:

Coarse SCAD

Fix the number of clusters C ;
 Fix m , $m \in (1, \infty)$;
 Initialize the centers and fuzzy partition matrix \mathbf{U} ;
 Initialize the relevance weights to $1/K$;
Repeat
 Compute $(d_{ij}^s)^2$ for $1 \leq i \leq C$, $1 \leq j \leq N$, and $1 \leq s \leq K$;
 Update the relevance weights v_{is} using (31);
 Compute D_{ij}^2 using (28);
 Update the partition matrix $U^{(k)}$ using (32);
 Update the centers using (34);
Until(centers stabilize)

4 Self-Organization of Oscillators Network (SOON) Algorithm

In [78], Frigui *et al.* introduced a clustering approach that combines concepts from clustering and synchronization of coupled oscillators. The algorithm is efficient, robust, unbiased to the cluster size, and can find an arbitrary number of clusters.

Let $\mathfrak{Y} = \{\mathbf{y}_j | j = 1, \dots, N; \mathbf{y}_j \in \mathbb{R}^p\}$ be a set of objects characterized by p attributes. Each object (\mathbf{y}_j) is represented by an IF oscillator (O_j) characterized by a phase variable ϕ_j , a state variable x_j , given by

$$x_i = f(\phi_i), i = 1, \dots, N, \quad (35)$$

where

$$f(\phi_j) = \frac{1}{b} [1 + (e^b - 1)\phi_j]. \quad (36)$$

In (36), b is a constant, usually fixed to 2, and determines the concavity of f . When the phase of O_i , x_i , reaches the threshold one, it fires, resets to zero, and the phases of the other oscillators $O_j (j \neq i)$ get updates using

$$x_j(t^+) = B(x_j(t) + \epsilon_i(\phi_j)). \quad (37)$$

In (37), $\epsilon_i(\phi_j)$ is a coupling function. It depends on the similarity between the firing oscillator (O_i) and the oscillator being updated (O_j). In [78], the authors use the following coupling function

$$\epsilon_i(\phi_j) = \begin{cases} C_E[1 - d_{ij}^2/\delta_0] & \text{if } d_{ij}^2 \leq \delta_0 \\ -C_I \left[\frac{d_{ij}^2 - \delta_0}{\delta_1 - \delta_0} \right] & \text{if } \delta_0 < d_{ij}^2 \leq \delta_1 \\ -C_I & \text{otherwise} \end{cases} \quad (38)$$

In (38), d_{ij} is the relative dissimilarity measure between O_i and O_j , C_E and C_I are the maximum excitatory and inhibitory coupling respectively. If two oscillators are similar (i.e. $d_{ij} < \delta_0$), then the coupling is excitatory. However, if $d_{ij} > \delta_0$, then the oscillators are dissimilar and coupling is inhibitory. Thus, the parameter δ_0 is related to the resolution of the desired partitions. A small δ_0 would partition the data into a large number of small clusters, while a large δ_0 would partition the data into a small number of large clusters.

The SOON algorithm is summarize below.

Self-Organization of Oscillators Network (SOON) Algorithm

```
Initialize phases  $\phi_i$  randomly for  $i = 1, \dots, N$ ;  
Repeat  
  Identify next oscillator to fire =  $\{O_i : \phi_i = \max_{j=1}^N \phi_j\}$ ;  
  Compute  $d_{ij}$ , for  $j = 1 \dots N, j \neq i$ ;  
  Bring  $\phi_i$  to threshold & adjust other phases:  
     $\phi_j = \phi_j + (1 - \phi_i)$  for  $j = 1, \dots, N$ ;  
  For all oscillators  $O_j (j \neq i)$  Do  
    Compute state variable  $x_j$  using (35) and (36);  
    Adjust state variables using (37);  
    Compute new phases:  $\phi_j(t^+) = f^{-1}(x_j(t^+))$ ;  
  Identify synchronized oscillators;  
  Update the parameters of the synchronized group;  
  Reset phases of synchronized oscillators;  
Until (Synchronized groups stabilize);
```

5 Self Organization and Visual Exploration of Large Data Sets

In addition to clustering the data, SOON could be used to map the high-dimensional data to a one-dimensional phase variable that reflects the pairwise similarity among the data points. Thus, this algorithm could also be used to visualize data. In [79], the authors proposed a visualization algorithm that explores this property of SOON. In particular, the phase values were used as a one-dimensional projection of the data. To obtain a n-dimensional phase space, SOON could be run n-times while varying resolution and/or distance measures.

For very large data, one-dimensional and two-dimensional phases of all data samples is too cluttered to reveal any useful information. Thus, the information must first be clustered into representative samples. This phase map of clustered points generates a general overview of the data, providing a global summary of the entire image database. The Self Organization and Visual Exploration (SOAVE) [79] algorithm visualizes high-dimensional data on a two-dimensional map as follows:

1. Summarization: The SOON algorithm is used to cluster data into a subset of

representative points. The phases of the resulting synchronized oscillators provides for the initial projection.

2. Mapping: The summarized data is allowed to self-organize for additional iterations at a different resolution or with a different distance measure to generate the second set of phases for two-dimensional space.
3. Visualization: A two-dimensional map is generated from the sets of phases and presents the user with the relative spatial distribution of the clusters' representatives.
4. Zooming: Zooming into a region of clusters requires a clustering algorithm (SOON_D) [79] that desynchronizes the phases. This approach explodes the elements within a cluster resulting in the re-clustering of the data as phases of dissimilar points diverge.

D User Interfaces for CBIR

Another main component of a CBIR system is the user interface. This component allows the user to initiate the query process and to visualize the retrieved images. In the following, we outline some of the common approaches.

1 Query-by-Visual-Example

Global query-by-visual-example (QBvE) is the most common interface mode in CBIR systems [9, 80, 81, 82, 83, 84]. In this mode, the user supplies a sample (e.g. an image) of what he/she is looking for and then the system retrieves items that are most similar to the submitted example. The main advantage of QBvE is that the user is not required to provide an explicit description of the image, which is instead computed by the system from the example image. A major drawback of this

approach is that it uses global features that cannot describe small objects within the image. Another problem in QBvE, known as the "*Page Zero*" problem [85], questions how does a user begin the search process without an example image. In other words, existing CBIR techniques are successful only if the user has a relevant starting point. Alternative visual browsing techniques (such as [86, 87]) help by providing an overview of the database but make sense for image search only if the goal is vague [5].

2 Query-by-Visual-Region

The QBvE paradigm may not provide reasonable results when the focus of the search is a specific object or part of an image. Query-by-visual-region (QBvR) allows for more specific queries by letting the user specify which part of the image is the target. A QBvR retrieval system segments the images into regions (objects) and retrieves based on the visual similarity between them. Existing QBvR systems [17, 18, 88] perform an exhaustive search among regions in the database from a single example region. The major drawback with the QBvR approach is that image segmentation is a difficult problem, inevitably makes mistakes, and can cause some degradation in performance. Moreover, QBvR approaches are computationally expensive, difficult to implement, tend to be domain specific, and have the same "*Page Zero*" problem as QBvE.

E Other Issues in CBIR

1 Fusion of Multiple Sets of Features

As image databases continue to increase in size and become more complex in content, it is becoming impossible to achieve high performance in retrieving visually similar images with a single feature set. As a result, diverse sets of features are

being used and combined to provide a more accurate retrieval. However, the task of effectively fusing the output of multiple descriptors has been overlooked by the CBIR community. Only methods that are based on distance scaling or normalization and simple list merging have been used [44]. In fact, the different features can vary significantly with respect to the number of attributes, the dynamic ranges, and the adopted distance measures. Thus, fusion of these features is not trivial and can have a significant impact on the overall performance of the CBIR system.

2 Bridging the Semantic Gap Problem

The performance of most CBIR systems is inherently constrained by the low-level features used to describe the content of the images; These low-level features cannot adequately express the user's high-level concepts, and as such give unsatisfactory results. This problem is referred to as the "Semantic Gap" [33, 34]. The semantic gap is the inability to reconcile high-level concepts as perceived by users and low-level features used to describe the content of the images. Through life experiences, humans gain knowledge that allows them to associate concepts with objects. Teaching a computer to make these connections, however, is a non-trivial problem.

Current solutions to bridge the semantic gap can be categorized into three main approaches. The first one is based on image database categorization [89, 90, 91, 92, 93, 94, 95]. In this approach, the goal is to partition the image database into clusters of similar images. During retrieval, only clusters that are similar to the query image are searched. Thus, if images within each cluster are semantically similar, the retrieved images will also be semantically similar. Thus, considerable effort has been made towards effective clustering algorithms for this task. These algorithms attempt to learn relevant features for each category, and use

partial supervision information to guide the clustering process [96, 97, 98, 99, 100, 94, 101, 102, 103].

The second type is based on relevance feedback [104, 98]. For example, a typical image retrieval process begins by using the QBvE approach, wherein the user provides the system with an example image [105]. The system then retrieves a set of images that are visually similar to the given query. Then, the user is asked to select which returned images are relevant to their interests. This is usually accomplished by allowing the user to flag any number of returned images as either positive (relevant) or negative (irrelevant) [25, 106, 107, 108]. Iteratively, the algorithm continues by using the information obtained from the features of the current query, and factors in the given feedback information to adjust or guide the query tract to adapt to the user’s perception. Examples of such guidance include using the relevance to shift the query in the feature space [25], learn or adjust feature relevance weights [106, 107], or attempt to predetermine relevance values for the remaining images in the database [108].

The third approach involves annotating the images and representing them by textual features to support text-based queries. While this could be done manually, it is tedious and not practical for large scale image databases. A viable alternative would involve automatic image labeling. Since our proposed approach falls into this category, in the following section, we outline few different methods that were proposed for this task.

3 Image Annotation

Image annotation has proven to be an important approach to bridge the semantic gap in CBIR systems. While no system can fully bridge the gap, experiments have shown that any step towards it can have a drastic improvement in

the accuracy of the CBIR.

An example of the benefit of adding textual information to a CBIR query is shown in Figure 6. Here, the query image is shown in Figure 6(a) and the results in Figure 6(b) are retrieved using only visual features. While these results are visually relevant, there is a noticeable semantic difference between the query and the results. In Figure 6(c), semantic information is added to the query. Adding these labels to the query can improve the results significantly as shown in Figure 6(d). The success of this approach depends on the accuracy and efficiency of the annotation process. For instance, if the database contains few thousand images, then manually labeling the images is tedious, but feasible. But in most real world applications with vast databases, labeling elements quickly becomes unrealistic. Automatic image annotation is thus the only feasible approach to accomplishing the task.

Automatic image annotation is not a trivial task. It cannot be accomplished with traditional pattern recognition techniques. When dealing with images, there are multiple problems in trying to use a classifier: (1) standard classification relies on labeled data for training and it is hard to collect large labeled data, (2) there are too many classes (typically in the 100's), (3) the image features are high-dimensional, (4) the values are continuous with no known range, and (5) the available training data may not be accurate.

Different approaches to image labeling have been proposed. Some of these algorithms annotate images at the region level, others are global and annotate at the image level. Annotation at the image level involves finding words/labels that best describe the entire image. Region level annotation requires segmenting the image into objects, regions, or blobs, and annotating each region. Region level annotation allows for direct object searching and in most cases, produces higher

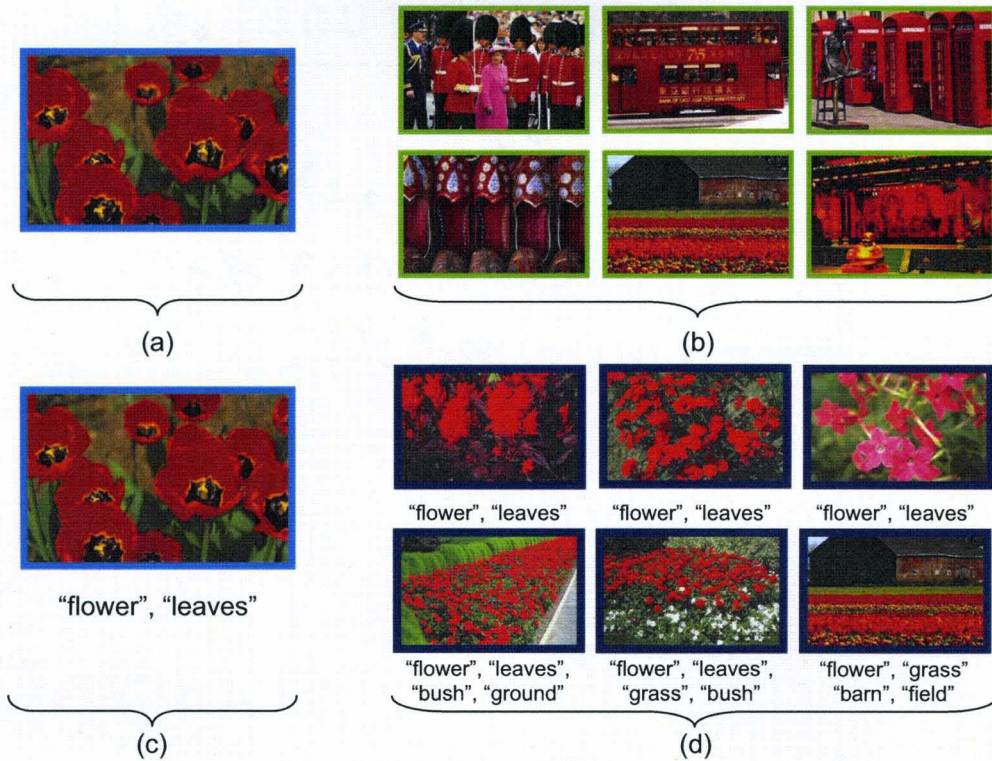


Figure 6. Illustration of the benefits of adding textual information to CBIR. (a) The query image. (b) The most similar 6 images using only low-level visual features. (c) Textual information is added to the query image. (d) The most similar images using visual and textual features.

retrieval accuracy since what is inside the image can be better represented. While there are further divisions in approaches into unsupervised and supervised, only the unsupervised methods will be described, as supervised approaches are semi-equivalent to manual labeling and can become tedious as well.

The main unsupervised approaches to image annotation are probabilistic, correlation based, latent semantic indexing, and data mining (clustering and association rule based). These approaches are outlined in the following sections. A few other approaches that do not fall into these are also discussed in the last subsection.

Probabilistic Approaches

Probabilistic approaches model the data to estimate the probability distributions and the local relevances of the features. The modeling of the visual features in conjunction with their textual features provides a model for translating the feature from one modality to another. In the following, few of these approaches are outlined.

In Duygulu *et al.* [38], the problem of image annotation is treated as a machine translation from one form (images regions) into another form (words). This model learns *lexicons* from an annotated training data set through a variant of the Expectation-Maximization (EM) [109] algorithm. These lexicons are then used to predict one representation (words) given another representation (image regions). In particular, a correspondence between the words assigned to an image and the regions representing the image can be learned. Annotation of a new test image consists of segmenting it into blobs, and choosing the word with the highest probability for each blob.

Another annotation approach is proposed by Barnard *et al.* [37]. Here, a number of models that can calculate the joint distribution and correspondence of image regions and words are learned. The annotation models are used to describe the distribution of regions and words, and a separate set of models are used to establish correspondence. Multiple models are integrated to reveal more information than any individual one. The annotation models considered in [37] include a Multi-Modal Hierarchical Aspect Model and a Multi-Modal Latent Dirichlet Allocation. In the first model, nodes generate image regions using a Gaussian distribution and words using a multinomial distribution. Each word is assumed to have come from a node in a hierarchical concept tree which is coherent with the model for nouns and verbs adopted by WordNet [110]. The closer a node is to the

root the more clusters share it. Root nodes give more general words (e.g. *sky*) than leaf nodes (e.g. *waves*). An image (*document*) is modeled by a sum over the clusters and weighted by the probability that the document is in the cluster.

In the mixture of multi-modal LDA model (MoM-LDA), each collection is modeled by a randomly generated mixture over latent factors. The outer plate is the repetition of I images, and each image has M blobs and N words. An EM algorithm [109] with variational E step calculates the maximum likelihood estimates of the Dirichlet, word multinomials, and Gaussian parameters. Co-occurrence of words and regions on a node can simulate correspondence between specific regions and words from the hierarchical clustering model. Lastly, weighted models provide for integrating correspondence and hierarchical clustering to strengthen the relationship between words and image regions.

The *translation model* in [38] was extended by Jin *et al.* [111] to eliminate uncorrelated words from those generated through usage of WordNet [110].

Uncorrelated words refers to those that label an image during annotation and are irrelevant to the image. This is done by using various semantic similarity measures between keywords and combining these to make a final decision using Dempster-Shafer evidence combination [112]. Some of the similarity measures used are the Resnik Measure (RIK) [113], Jiang and Conrath Measure (JNC) [114], Lin Measure (LIN) [115], Leacock and Chodorow Measure (LNC) [116], and Banerjee and Pedersen Measure (BNP) [117]. Each measure depicts different independent relationships (*evidence*) between words that is utilized and combined to create hypothesis' in Dempster's Rule. A threshold then removes keywords from the annotation list for that image.

Another probabilistic approach to image annotation is the Cross-Media Relevance Models (CMRM) [39]. This approach derives the probability of

generating a word given the blobs in an image in a simpler way. It does not assume a one-to-one correspondence between blobs and words in an image as the translation models. The CMRM assumes that a set of keywords $\{w_1 \dots w_n\}$ is related to the objects represented by blobs $\{b_1 \dots b_m\}$. A relevance model is the underlying probability distribution of an image I and contains all possible blobs and words that could appear in I . The probability of observing a word w from the relevance model for I is estimated by the conditional probability $P(w|I)$. Since the actual relevance model is unknown, Jeon *et al.* [39] estimate $P(w|I)$ by either sampling repeatedly from the distribution or by picking n words with the highest probability. An alternative approach that uses a Continuous-space Relevance Model (CRM) was also introduced in [118].

The Coherent language models (CLM) [119] are closely related to the cross-media relevance models. These models exploit word-to-word correlations in an image to strengthen annotation decisions. This approach benefits from three advantages over previous models: 1) it can determine the annotation length, 2) the number of annotated image examples can be reduced through active learning, and 3) it avoids the overfitting problem with the EM algorithm. Instead of predicting each annotated word independently, this model estimates coherent language models for a given image. In other words, the CLM treats the words annotating an image as a set, $\{w\}$, and attempts to maximize the conditional probability $P(\{w\}|I)$. To avoid dealing with an exponential number of word sets (with respect to vocabulary size), the CLM computes word probabilities, Θ_w , that determines the likelihood of each word to be used in the annotation.

The coherent language model with flexible length (CLMFL) is introduced and utilizes the fact that the estimated language model is based on Bernoulli distributions. This model provides for annotations of different lengths and the

ability to predict this length. Adding an active learning strategy to the CLMFL model can allow users to annotate images with the least averaged word probability. This reduces the uncertainty in determining the right model.

A different probabilistic image annotation approach has been proposed in [120]. In this approach, a two-dimensional multiresolution hidden Markov model (2D MHMM) is used to represent each concept. These models assign a likelihood value of the occurrence of an image based on the textual description of a concept. A high likelihood indicates a strong association and can be used to annotate new images.

Fan *et al.* [121, 122] proposed a multi-level approach to annotate image components with relevant semantic concepts. Salient objects are used for image content representation and feature extraction. Support Vector Machine (SVM) classifiers are honed to detect salient objects, and finite mixture models (FMM) are used for concept interpretation. The optimal model structure is determined through an adaptive EM algorithm that does not require careful initialization, can take advantage of negative samples, and can escape local extrema by reorganizing the distribution. Salient objects from an image are classified into the best matching semantic image concept with the maximum posterior probability.

In [123], a Latent Dirichlet Allocation (LDA) model is used for modeling associations between words and pictures. Three different annotation models were compared: the Gaussian-Multinomial Mixture Model, Gaussian-Multinomial LDA, and Correspondence LDA.

Correlation Based Approaches

Correlation approaches combine visual features with textual features and look for associations. These approaches tend to be fast and efficient, as modeling the data is not required as with probabilistic methods .

Wang *et al.* [124] and Pan *et al.* [125] use similar approaches to discover correlations between visual features and keywords through a correlation-based translation table. In [124], blob-tokens are created through a K-means clustering algorithm [126] that performs weighted feature selection, and in [127] blob-tokens are created through clustering with the G-means algorithm [127] which can adaptively learn the number of tokens. Both approaches proceed by creating a weighted matrix that combines visual features and keywords. This matrix is then used to create a translation table that measures the association between a term and a blob-token by the co-occurrence counts. The method in [125] is expanded further to use singular value decomposition (SVD) [128] for suppressing noise in the data before learning the association.

Another correlation based image annotation method uses Mixed Media Graph (MMG) to discover cross-modal correlations [129]. This approach represents all objects and attributes as nodes in a graph. Correlations are then discovered using a "random walk with restarts". This graph theory approach states that before a random walk chooses its next edge there is a probability it will go back to the beginning. These steady-state probabilities are then calculated to find the importance of node B with respect to node A.

Latent Semantic Approaches

Latent Semantic Indexing (LSI) [130] has been used mainly to index documents in text-retrieval. Cross-Language Latent Semantic Indexing (CL-LSI) is the technique of using LSI to retrieve queries in multiple languages where the queries themselves can be in different languages. The idea of CL-LSI was first presented by Landauer and Littman [131] with French and English documents. In [132, 33, 34], Hare *et al.* used a generalization of the CL-LSI to annotate images.

Their annotation however was not explicit. Unannotated images are simply placed in a semantic-space which can be queried by keyword.

In [133], Yu *et al.* extended the LSI to take into account the targeted values in the training set as well as the inputs in multi-label informed latent semantic indexing (MLSI). This approach not only preserves the information contained within inputs but learns correlations between multiple outputs. Unlike normal LSI which is purely unsupervised, MLSI is a supervised LSI and could be extended to image annotation as with the CL-LSI.

Another annotation model that has shown promising performance is based on Latent Semantic Analysis (LSA). Here, annotation is accomplished through finding the underlying semantic structures of words and image features in a linear Latent Space. For instance, in [42] Liu and Tang reveal these latent variables of words and visual features using Probabilistic LSA (PLSA). The authors have also extended this approach to use a Nonlinear Latent Space and capture the dependency of images and words using Image-Word Embedding (IWE). In [43], the authors compare PLSA to LSA, citing differences and benefits to both approaches.

Data Mining Based Approaches

Data mining techniques, and in particular Clustering and Association Rules, have been used to annotate images. In [134], Wang *et al.* used clustering to group similar visual tokens from images using a modified K-means algorithm. At each iteration, the algorithm determines which features are important to a given cluster and discards the remaining features. Then, using a method similar to the one presented in [125], keywords are linked to blob-tokens using a correlation table approach. Stan and Sethi [135] on the other hand, use multidimensional indexing to solve the high dimensionality and the non-Euclidean nature of the feature space.

Then, the primitive features and high-level concepts are mapped using IF-THEN rules using keyword rankings and cluster radius.

Another data mining technique that has been used for image annotation is association rule mining. Association rule mining has been used traditionally in applications such as market basket analysis [136]. It attempts to capture interesting relationships between attributes, thereby enhancing the understandability of the data. Association rule mining has also been applied to image data [137, 138, 139]. However, they have not been fully exploited for the case of multi-modal data to learn relationships among the different modalities.

Multiple instance learning (MIC) has also been applied to learn image annotation. For instance, in [140], Chen and Wang use his technique for region based image categorization. A collection of instance prototypes are learned that represent a class of instances more likely to appear in bags with specific labels. Every bag is then nonlinearly mapped to a point in the bag feature space and support vector machines are trained on this space.

CHAPTER III

CLUSTERING AND FEATURE DISCRIMINATION

As image databases continue to increase in size and become more complex in content, it is becoming impossible to achieve high performance in retrieving visually similar images with a simple feature set. As a result, diverse sets of features are being used and combined to provide a more accurate retrieval. However, this would impose additional requirements on several components of the CBIR system. For instance, in clustering, the different sets of features are not expected to be equally relevant in the different image categories. Consequently, irrelevant features can adversely affect cluster definitions. Thus, it is recommended to identify cluster-dependent feature-relevance weights. Unfortunately, most existing feature selection and weighting algorithms [141] are not suitable for unsupervised learning. Recently, an algorithm performing Simultaneous Clustering and Attribute Discrimination (SCAD) was proposed (refer to section §II.C.3). SCAD is designed to search for the optimal clusters prototypes and the associated optimal relevance weight for each feature within a cluster. Two versions of SCAD were developed: The first (SCAD₁, §II.C.3.3) balances between two terms in a compound objective function while the second (SCAD₂, §II.C.3.3) minimizes a single discrimination exponent term.

For high dimensional data, learning a relevance weight for *each* feature may result in overfitting. To avoid this, a coarse feature weighting approach, called SCAD_c [77] was proposed as an extension of SCAD₁ (see section §II.C.3.3). Instead

of learning a weight for each feature, the set of features is divided into logical subsets, and a single weight is learned for each of these subsets. For CBIR applications our initial experimentations have indicated that SCAD₂ is more stable than SCAD₁. Since a coarse version of SCAD₂ was not previously developed, we will first start by developing this algorithm and deriving the necessary conditions. To simplify notation, we will simply use SCAD to refer to SCAD version 2. Whenever we refer to the original version, we will explicitly mention that it is version 1.

In this chapter, we adopt the SCAD algorithm and enhance it to cluster image features into more meaningful groupings. First, we propose a coarse extension of SCAD that assigns relevance weights to feature subsets. Next, the algorithm is extended to partition the data into the optimal number of clusters by interpreting concepts from the Competitive Agglomeration (CA) described in section §II.C.2. Then, different techniques to deal with prototype initialization, distance normalization, and annealing schedule for feature discrimination are proposed.

A Simultaneous Clustering and Attribute Discrimination (SCAD)

We assume that the p features have been partitioned into K subsets: FS^1, FS^2, \dots, FS^K such that each subset FS^s includes k^s features. For CBIR applications, these subsets could include a set for color features, another for texture features, and yet another for textual features. Let d_{ij}^s be the partial distance between feature vector \mathbf{x}_j and cluster i using the s^{th} feature subset, and let $\mathbf{V} = [v_{is}]$ be the relevance weight for FS^s with respect to cluster i . SCAD_c minimizes

$$J(\mathbf{B}, \mathbf{U}, \mathbf{V}; \mathcal{X}) = \sum_{i=1}^C \sum_{j=1}^N (u_{ij})^m \sum_{s=1}^K (v_{is})^q (d_{ij}^s)^2, \quad (39)$$

subject to the constraints in (4) and (17).

To optimize J with respect to \mathbf{V} , we use the Lagrange multiplier technique

and minimize:

$$J(\Lambda, \mathbf{V}) = \sum_{i=1}^C \sum_{j=1}^N (u_{ij})^m \sum_{s=1}^K (v_{is})^q (d_{ij}^s)^2 - \sum_{i=1}^C \lambda_i \left(\sum_{s=1}^K v_{is} - 1 \right),$$

where $\Lambda = [\lambda_1, \dots, \lambda_C]^t$ are Lagrange multiplier constants. Since the rows of \mathbf{V} are independent, the optimization problem above can be reduced to the following C independent problems:

$$J_i(\lambda_i, \mathbf{V}_i) = \sum_{j=1}^N (u_{ij})^m \sum_{s=1}^K (v_{is})^q (d_{ij}^s)^2 - \lambda_i \left(\sum_{s=1}^K v_{is} - 1 \right), \text{ for } i = 1, \dots, C, \quad (40)$$

In (40), \mathbf{V}_i is the i^{th} row of \mathbf{V} . By setting the gradient of J_i to zero, one can solve the system of equations

$$\frac{\partial J_i(\lambda_i, \mathbf{V}_i)}{\partial \lambda_i} = \left(\sum_{s=1}^K v_{is} - 1 \right) = 0 \quad (41)$$

$$\frac{\partial J_i(\lambda_i, \mathbf{V}_i)}{\partial v_{is}} = q (v_{is})^{(q-1)} \sum_{j=1}^N (u_{ij})^m (d_{ij}^s)^2 - \lambda_i = 0 \quad (42)$$

for the relevance weights v_{is} and obtain

$$v_{is} = \frac{\left[1 / \sum_{j=1}^N (u_{ij})^m (d_{ij}^s)^2 \right]^{1/(q-1)}}{\sum_{s=1}^K \left[1 / \sum_{j=1}^N (u_{ij})^m (d_{ij}^s)^2 \right]^{1/(q-1)}}. \quad (43)$$

Simplifying (43), v_{is} reduces to

$$v_{is} = \frac{1}{\sum_{k=1}^K \left(\tilde{D}_{is} / \tilde{D}_{ik} \right)^{1/(q-1)}}, \quad (44)$$

where $\tilde{D}_{is} = \sum_{j=1}^N (u_{ij})^m (d_{ij}^s)^2$ is the measure of dispersion for the i^{th} cluster taking into account only the s^{th} feature set and $\sum_{k=1}^K \tilde{D}_{ik}$ is the cumulative dispersion of the i^{th} cluster. This relation implies that the more compact the i^{th} cluster is with respect to feature set s (smaller \tilde{D}_{is}), the higher the relevance weight, v_{is} will be for the s^{th} feature.

The discrimination exponent $q \in (1, \infty)$ controls the feature discrimination among the different subsets. For large values of q , there is little or no discrimination. For small values, there is greater discrimination. Minimization of J in (39) with respect to \mathbf{U} , subject to the constraints in (4), yields

$$u_{ij} = \frac{1}{\sum_{k=1}^C \left[\tilde{d}_{ij}^2 / \tilde{d}_{kj}^2 \right]^{\frac{1}{m-1}}}, \quad (45)$$

where

$$\tilde{d}_{ij}^2 = \sum_{s=1}^K (v_{is})^q (d_{ij}^s)^2 \quad (46)$$

is the total distance between \mathbf{x}_j and cluster i computed from aggregating the partial distances and their weights. As recommended in [76], we let $q = 1$ when computing the fuzzy memberships, i.e., we replace (46) with

$$\tilde{d}_{ij}^2 = \sum_{s=1}^K v_{is} (d_{ij}^s)^2. \quad (47)$$

Minimization of J with respect to the prototype parameters depends on the choice of d_{ij}^s . Since the partial distances are treated independent of each other (i.e. disjoint feature subsets), the objective function in (39) can be decomposed into K independent problems:

$$J_s = \sum_{i=1}^C \sum_{j=1}^N (u_{ij})^m (v_{is})^q (d_{ij}^s)^2, \quad \text{for } s = 1, \dots, K. \quad (48)$$

Each J_s would be optimized with respect to a different set of prototype parameters. For instance, if d_{ij}^s is an Euclidean distance, minimization of J_s would yield the following update equation for the centers of subset s

$$\mathbf{c}_i^s = \frac{\sum_{j=1}^N u_{ij}^m \mathbf{x}_j^s}{\sum_{j=1}^N u_{ij}^m}. \quad (49)$$

The SCAD_c algorithm is summarized below:

Coarse Simultaneous Clustering and Attribute Discrimination Algorithm

Fix the number of clusters C ;
 Fix m , $m \in (1, \infty)$;
 Fix the discrimination exponent q , $q \in (1, \infty)$;
 Initialize the centers and fuzzy partition matrix \mathbf{U} ;
 Initialize all the relevance weights to $1/K$;
Repeat
 Compute $(d_{ij}^s)^2$ for $1 \leq i \leq C$, $1 \leq j \leq N$, and $1 \leq s \leq K$;
 Update the relevance weights matrix \mathbf{V} using (44);
 Compute \tilde{d}_{ij}^2 using (47);
 Update the partition matrix U using (45);
 Update the centers using (49);
Until(centers stabilize)

B Case of Unknown Number of Clusters

The SCAD_c algorithm described in the previous section requires that the number of clusters be specified a priori. However, it is not always possible to estimate this value, and the final partition can be sensitive to this value. To address this issue, we integrate the objective function of the Competitive Agglomeration algorithm (explained in §II.C.2) into the objective function of SCAD_c . The resulting objective function would combine the advantages of the CA and the SCAD_c algorithms. This algorithm (called $\text{SCAD}_c\text{-CA}$) minimizes

$$J(\mathbf{B}, \mathbf{U}, \mathbf{V}; \mathcal{X}) = \sum_{i=1}^C \sum_{j=1}^N (u_{ij})^m \sum_{s=1}^K (v_{is})^q (d_{ij}^s)^2 - \alpha \sum_{i=1}^C \left[\sum_{j=1}^N u_{ij} \right]^2, \quad (50)$$

subject to the constraint in (4).

In (50), C is an upper bound of the expected number of clusters. Minimization of (50) with respect to \mathbf{V} yields the same equation of v_{is} as in SCAD_c (see (44)) since the new term does not depend on v_{is} . The same is true for determining the centers (see (49)). To optimize (50) with respect to \mathbf{U} , we use the Lagrange multipliers and

obtain

$$J(\mathbf{B}, \mathbf{U} | \mathcal{X}) = \sum_{i=1}^C \sum_{j=1}^N (u_{ij})^2 \sum_{s=1}^K (v_{is})^q (d_{ij}^s)^2 - \alpha \sum_{i=1}^C \left[\sum_{j=1}^N u_{ij} \right]^2 - \sum_{j=1}^N \lambda_j \left(\sum_{i=1}^C u_{ij} - 1 \right)$$

An updating equation for the memberships u_{ij} can be obtained by fixing \mathbf{B} and solving

$$\frac{\partial J}{\partial u_{ij}} = 2u_{ij} \sum_{s=1}^K (v_{is})^q (d_{ij}^s)^2 - 2\alpha \sum_{t=1}^N u_{it} - \lambda_j = 0, \text{ for } i \in \{1, \dots, C\}, j \in \{1, \dots, N\} \quad (51)$$

If we assume that the membership values do not change significantly from one iteration to the next, then (51) can be solved for u_{ij} , yielding

$$u_{ij} = \frac{2\alpha \times N_i + \lambda_j}{2 \sum_{s=1}^K (v_{is})^q (d_{ij}^s)^2}. \quad (52)$$

In (52),

$$N_i = \sum_{t=1}^N u_{it} \quad (53)$$

is the fuzzy cardinality of cluster i . Using (52), the constraint in (4), and solving for λ_j , one obtains

$$\lambda_j = \frac{1 - \alpha \sum_{k=1}^C [N_k / \sum_{s=1}^K (v_{kj})^q (d_{kj}^s)^2]}{\sum_{k=1}^C [1 / \sum_{s=1}^K (v_{kj})^q (d_{kj}^s)^2]} \quad (54)$$

Substituting equation (54) in equation (52), we obtain the update equation for the membership of feature point \mathbf{x}_j in cluster β_i :

$$u_{ij} = \alpha \frac{N_i}{\sum_{s=1}^K (v_{is})^q (d_{ij}^s)^2} + \frac{1 - \alpha \sum_{k=1}^C [N_k / \sum_{s=1}^K (v_{kj})^q (d_{kj}^s)^2]}{\sum_{k=1}^C [\sum_{s=1}^K (v_{is})^q (d_{ij}^s)^2 / \sum_{s=1}^K (v_{kj})^q (d_{kj}^s)^2]}$$

Rearranging the terms, we obtain

$$u_{ij} = u_{ij}^{\text{FCM}} + u_{ij}^{\text{BIAS}}, \quad (55)$$

where

$$u_{ij}^{\text{FCM}} = \frac{[1 / \sum_{s=1}^K (v_{is})^q (d_{ij}^s)^2]}{\sum_{k=1}^C [1 / \sum_{s=1}^K (v_{kj})^q (d_{kj}^s)^2]}, \quad (56)$$

and

$$u_{ij}^{\text{BIAS}} = \frac{\alpha}{\sum_{s=1}^K (v_{ij})^q (d_{ij}^s)^2} (N_i - \bar{N}_j). \quad (57)$$

In (57),

$$\bar{N}_j = \frac{\sum_{k=1}^C [1 / \sum_{s=1}^K (v_{kj})^q (d_{kj}^s)^2] N_k}{\sum_{k=1}^C [1 / \sum_{s=1}^K (v_{kj})^q (d_{kj}^s)^2]}. \quad (58)$$

The choice of α in (50) reflects the importance of the second term relative to the first term. We use the same scheme recommended by the authors in [71] to update α in every iteration to balance the two terms. That is, we use

$$\alpha(k) = \eta(k) \frac{\sum_{i=1}^C \sum_{j=1}^N (u_{ij})^2 \sum_{s=1}^K (v_{ij})^q (d_{ij}^s)^2}{\sum_{i=1}^C [\sum_{j=1}^N u_{ij}]^2} \quad (59)$$

where

$$\eta(k) = \eta_0 \exp(-k/\tau). \quad (60)$$

In (60), η_0 is the initial value, τ the time constant, and k is the iteration number.

C Initialization & Distance Normalization for SCAD_c-CA

One issue when combining distances from feature subsets of variable lengths is that their values can have very different dynamic ranges. Thus, it is critical that the distances be normalized within SCAD_c-CA to avoid any bias that may be caused by the dimensionality of the feature set. In our application, we use the fuzzy c means algorithm (FCM) (outlined in section §II.C1) to obtain an initial partition of the data set. This algorithm, treats all features equally important and assigns a fuzzy membership value u_{ij} to each point \mathbf{x}_j in cluster i .

Using these memberships, after few iterations of the FCM algorithm, we estimate the average distance for each feature subset s using

$$D_{avg}^s = \frac{\sum_{j=1}^N \sum_{i=1}^C u_{ij} \times d_{ij}^s}{\sum_{j=1}^N \sum_{i=1}^C u_{ij}}, \quad (61)$$

where d_{ij}^s is the partial distance of \mathbf{x}_j to \mathbf{c}_i with respect to the given feature subset s .

Using D_{avg}^s , equations (39)-(48),(50) can be modified such that the partial distances are replaced by the normalized partial distance, i.e.,

$$\tilde{d}_{ij}^s = \frac{d_{ij}^s}{D_{avg}^s} \quad (62)$$

in every iteration.

D Annealing Schedule for Feature Discrimination

In [76], the authors have argued that the discrimination exponent, q , can have a significant influence over the feature subset weights. In fact, selecting a value too high can result in all feature subsets being equally weighted, while a value too low emphasizes just one subset as important. Through experimentation, it was determined that it is better to initially begin with equally weighted subsets while the centroids are drastically moving, then have the weights discriminate certain subsets per cluster later in the computation. This leads to using annealing schedule for the discrimination exponent, q' . Let Q_{MAX} be the upper bound of q early in the algorithm, Q_{MIN} be the lower bound of q at the end of the algorithm, and Q_{BREAK} be the iteration that q' must reach Q_{MIN} . Then, q' is defined as

$$q' = \begin{cases} Q_{MAX} - (k) \frac{Q_{MAX} - Q_{MIN}}{Q_{BREAK}}, & \text{if } k < Q_{BREAK} \\ Q_{MIN} & \text{Otherwise} \end{cases} \quad (63)$$

where k is the current iteration. To translate this implementation into the existing equations, we simply let $q = q'$.

The resulting SCAD_c-CA algorithm is summarized below

SCAD_c-CA Algorithm

Fix the maximum number of clusters $C = C_{max}$;
Fix $m, m \in (1, \infty)$;
Fix FCM averaging counter l ;
Fix the discrimination exponent parameters Q_{MAX} and Q_{MIN} ,
 $Q_{MAX} \in (1, \infty)$, $Q_{MIN} \in (1, \infty)$, and $Q_{MAX} > Q_{MIN}$;
Fix the discrimination exponent counter Q_{BREAK} ,
 $Q_{BREAK} \in (l, \infty)$;
Initialize iteration counter $k = 0$;
Initialize the centers and fuzzy partition matrix \mathbf{U} ;
Initialize all the relevance weights to $1/K$;
Compute initial cardinalities N_i for $1 \leq i \leq C$ using (13);
Repeat
 Compute $d^2(\mathbf{x}_j, \beta_i) = (d_{ij}^s)^2$ using (1)-(3);
 Update the partition matrix $U^{(k)}$ using (6);
 Update the centers using (8);
 $k = k + 1$;
Until($k > l$)
Compute D_{avg}^s using (61);
Repeat
 Compute $q = q'$ using (63);
 Compute $(\tilde{d}_{ij}^s)^2 = (d_{ij}^s / D_{avg}^s)^2$ for $1 \leq i \leq C$,
 $1 \leq j \leq N$, and $1 \leq s \leq K$;
 Update $\alpha(k)$ using (59 and 60);
 Update the relevance weights v_{is} using (44);
 Compute \tilde{d}_{ij}^2 using (47);
 Update the partition matrix $U^{(s)}$ using (55);
 Update the centers using (49);
 Compute the cardinality N_i for $1 \leq i \leq C$ using (53);
 If ($N_i < \epsilon_1$) discard cluster β_i ;
 Update the number of clusters C ;
 Update the prototype parameters;
 $k = k + 1$;
Until(centers and prototype parameters stabilize)

CHAPTER IV

UNSUPERVISED IMAGE REGION ANNOTATION

The performance of most CBIR systems is inherently constrained by the used low-level features, and cannot give satisfactory results when the user's high level concepts cannot be expressed by low level features. In an attempt to bridge this semantic gap, few approaches that integrate low level visual features and user-defined textual keywords have been proposed [35, 36, 37]. Unfortunately, manually labeling each image by a set of keywords is subjective and labor intensive. Moreover, region labeling (as opposed to global image labeling) may be needed, which makes manual labeling more tedious. To address this issue, few algorithms that can annotate images/regions in an unsupervised (or semi-supervised) manner have been proposed in the past few years (refer to section §II.E.3).

In this chapter, we describe our proposed approach to annotate images at the region level to bridge the semantic gap in our CBIR system. Figure 7 displays the over all architecture of our proposed CBIR (shown initially in Figure 3) where the image annotation components are highlighted. Our approach, called Thesaurus Based Image Annotation (TBIA), is based on learning associations between low-level visual features and high-level textual keywords through multimedia data mining. These associations are then used to construct a multi-modal thesaurus that relates keywords to visual profiles through frequently co-occurring patterns. In particular, we adopt the SCAD_c-CA algorithm, that was developed in the previous chapter, to perform clustering and feature weighting simultaneously for the purpose

of learning inter-modality associations. This clustering algorithm is used to identify representative profiles that correspond to frequent homogeneous regions. The feature discrimination process, embedded in the clustering, would identify the relevant features in each profile. Then, representatives from each cluster and their relevant visual and textual features are used to build a multi-modal thesaurus. This thesaurus could be used to facilitate many tasks such as *auto-annotation*, *hybrid* searching and browsing, and *query expansion* as will be demonstrated throughout this thesis.

A Multi-Modal Thesaurus Construction

1 Training Data Set

We assume that a collection of images is available and that each image is annotated by few keywords. We do not assume that the annotation is complete or accurate. For instance, the image may contain many objects, and there is not a one to one correspondence between objects and words. This scenario is very common as images with annotations are readily available, but images where the regions themselves are labeled are rare and difficult to obtain.

Fig. 8 displays three images annotated at the image level from our training collection. Some keywords, such as "grass", can be clearly associated with color features. Others, such as, "house", may be associated with shape features. Other words may be associated with any combination of color, texture, and shape features. This information, if it could be learned, would improve the efficiency of image annotation and hybrid searching and browsing.

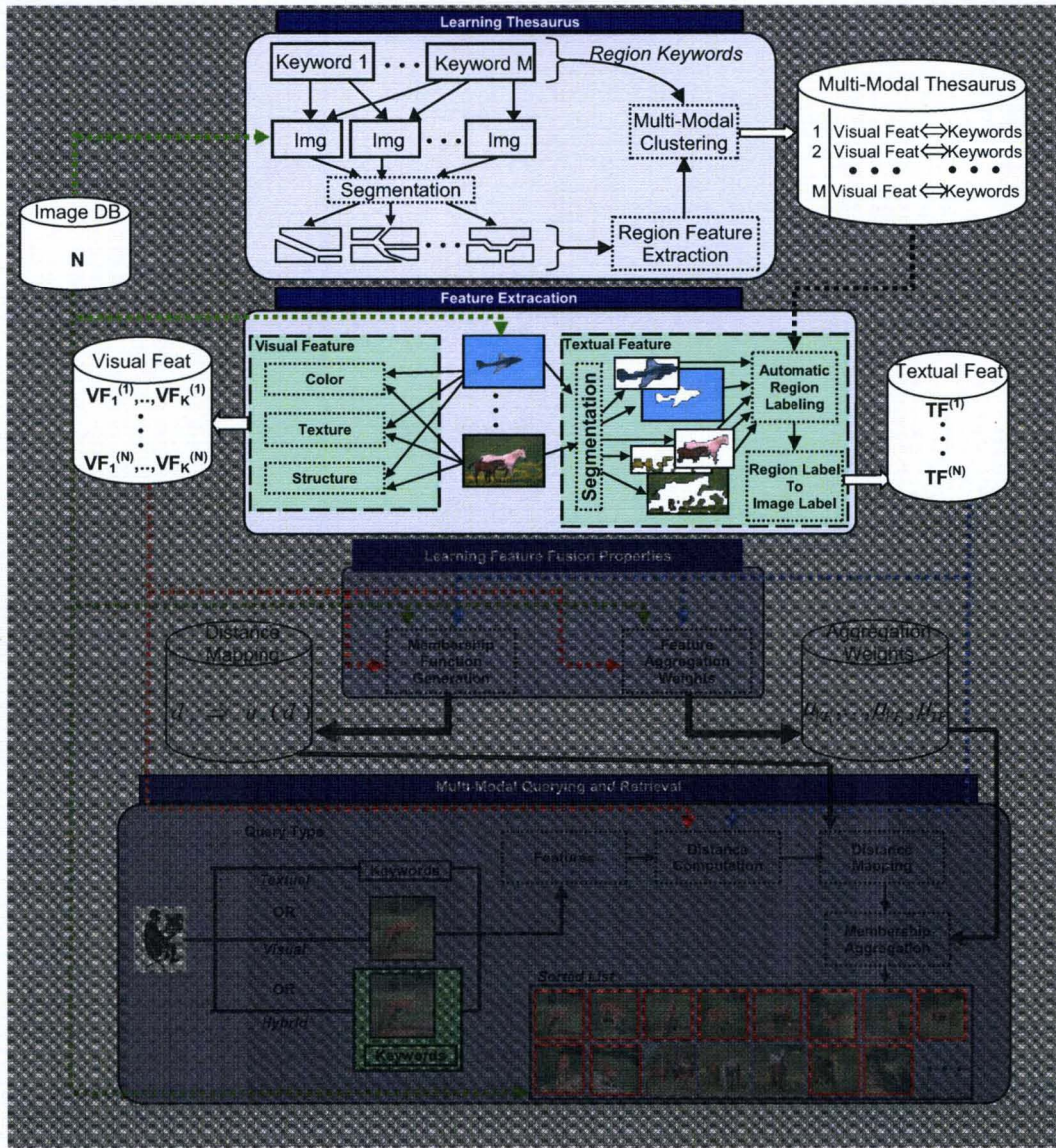


Figure 7. Highlighted architecture of the proposed CBIR system component to perform unsupervised image annotation.

2 Feature Extraction and Vector Representation of Images

First, each image in the training set is segmented into homogeneous regions based on color and/or texture features. While various segmentation algorithms could be used, in this work each image is coarsely segmented by clustering. The initial segmentation of all images in the database is carried out offline and the computation time may not be an issue. However, for image queries presented to the

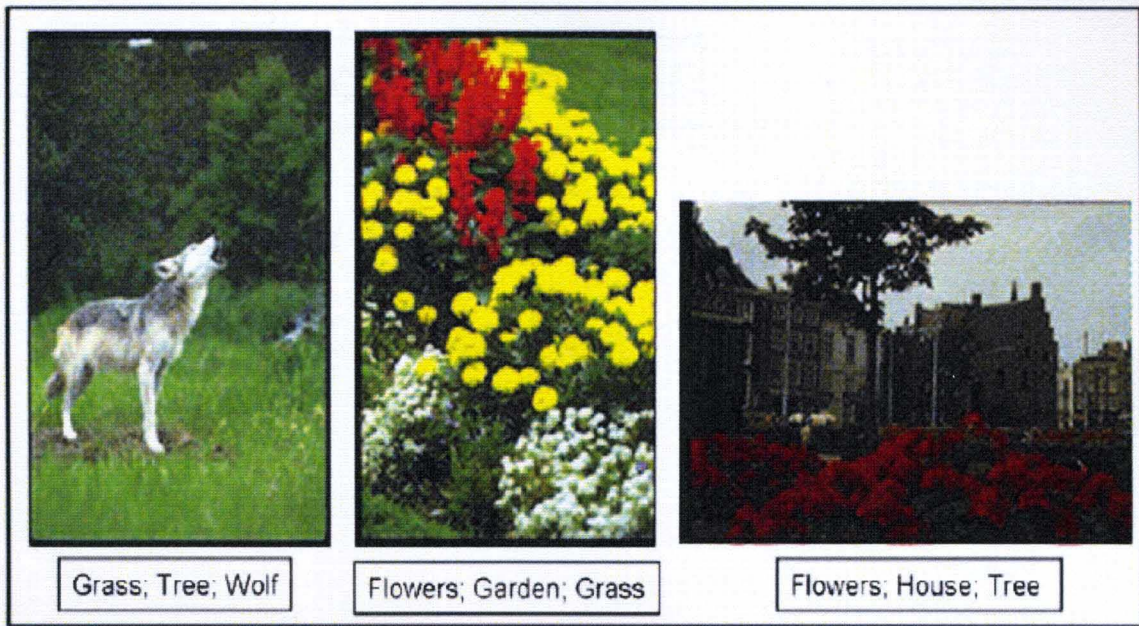


Figure 8. Examples of image-level annotations that refer to different and not segmented regions in the image.

CBIR system, segmentation must be carried out online at query time, requiring the segmentation algorithm to have a fast response. For this reason, instead of clustering every single pixel in the image, we extract one feature vector for a group of pixels in a fixed neighborhood. Moreover, we only use a simple feature that encodes the color histogram of the pixels in the neighborhood.

After the feature extraction, the Competitive Agglomeration (CA) algorithm [71] (outlined in section §II.C.2) is used to group the feature vectors into clusters. Our choice of this algorithm is based on its computational efficiency and its ability to identify the optimal number of clusters for each image. Figure 9 displays three sample images segmented using the above described approach. As it can be seen, our segmentation is coarse and we identify only the large regions.

After segmentation, each image region is described by standard visual features such as color, texture, shape, and a set of keywords. These features are



Figure 9. Examples of segmented images using the CA; (a) Original Images, (b) Coarsely segmented images.

briefly outlined below. For a more detailed description, we refer the reader to the references.

1. Wavelet Texture Descriptor: Each image is analyzed at different resolutions. The Haar filter bank is used to decompose the image into three scales [56]. This would result in a total of 10 components (approximation at scale three, and horizontal, vertical, and diagonal components at the three scales). Then,

the mean and standard deviation of the components are computed. This results in a 20-dimensional feature vector. We will refer to this feature set as F^{WTD} .

2. Edge Histogram Descriptor: The Edge Histogram Descriptor [54] encodes the structure of an image. First, simple edge detector operators are used to identify edges and group them into five categories: vertical, horizontal, 45° diagonal, 135° diagonal, and isotropic (non-edge). Then, a five bin histogram is used to represent the frequency of each edge category within each region. This results in a 5-dimensional feature vector. We will refer to this feature set as F^{EHD} .
3. RGB Color Histogram: The colors of all pixels within each region are uniformly quantized and represented by a histogram. We use a total of 64-bins and obtain a 64-dimensional feature vector. We will refer to this feature set as $F^{H_{RGB}}$.
4. HSV Color Moments: The Mean, standard deviation, and skewness of the distribution of the pixels within each region in the HSV color space are computed. This results in a 9-dimensional feature vector. We will refer to this feature set as $F^{M_{HSV}}$.
5. LUV Color Moments: The Mean, standard deviation, and skewness of the distribution of the pixels within each region in the LUV color space are computed. This results in a 9-dimensional feature vector. We will refer to this feature set as $F^{M_{LUV}}$.
6. Shape: For each region, the eccentricity, orientation, area, solidity, and extent are calculated. The eccentricity is calculated by first finding an ellipse with the same second-moments as the region and then computing the ratio of the

distance between the foci of the ellipse and its major axis length. The orientation is defined as the angle in degrees between the x-axis and the major axis of the ellipse containing the same second-moments as the region. The area is defined as the actual number of pixels within the region. The area is normalized by the total number of pixels in the image so that images of different sizes are comparable. The solidity is defined as the proportion of pixels in the convex hull that are also in the region. The extent is defined as the proportion of the pixels in the bounding box of the region that are also in the region. It is computed as the Area divided by the area of the bounding box. These features are represented by a 5-dimensional feature vector. We will refer to this feature set as F^{SHP} .

7. Use-Provided Keywords: We use the standard vector space model with term frequencies as features [142]. Let $\{w_1, w_2, \dots, w_p\}$ be the set of all keywords used to annotate the image database. Then, for each image region, we use a binary vector where the i^{th} element indicates the presence/absence of the i^{th} keyword in annotating the image. Since our data set is annotated by 97 keywords, this feature set is represented by a 97-dimensional feature vector. We will refer to this feature set as F^{TXT} .

Let $\{f_{j_1}^{(i)}, \dots, f_{j_{k_j}}^{(i)}\}$ be a k_j dimensional vector that encodes the j^{th} visual feature set of region R_i of a given image. An image that includes n regions (R_1, \dots, R_n) would be represented by n vectors of the form:

$$\underbrace{f_{11}^{(i)}, \dots, f_{1k_1}^{(i)}}_{\text{visual feat 1 of } R_i}, \dots, \underbrace{f_{C1}^{(i)}, \dots, f_{Ck_C}^{(i)}}_{\text{visual feat C of } R_i}, \underbrace{w_1, \dots, w_p}_{\text{Keywords}}, \quad i = 1 \dots n.$$

Figure 10 illustrates this image representation approach. It should be noted here that since the keywords are not specified per region, they are duplicated for each region representation. The assumption is that, if word w describes a given

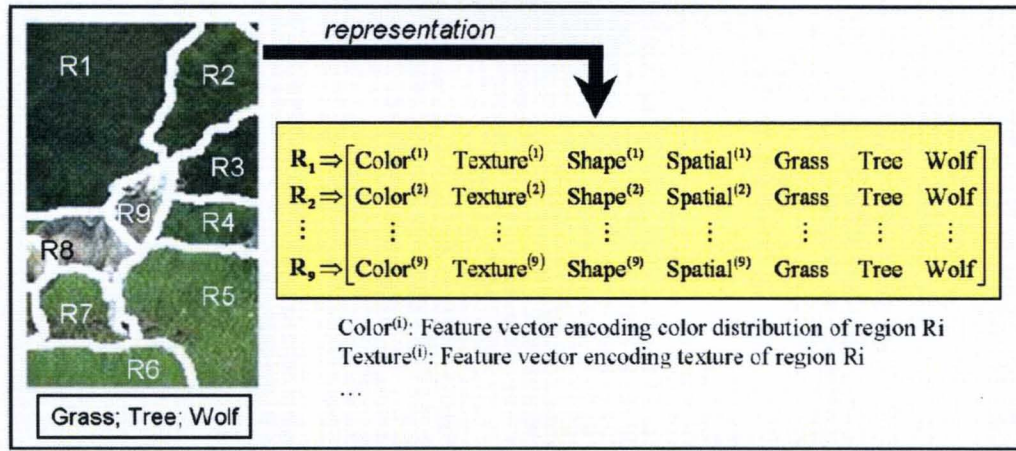


Figure 10. Representation of visual and textual features

region R_i , than a subset of its visual features would be present in many instances across the image database. Thus, an association rule among them could be mined. On the other hand, if none of the words describe R_i , then these instances would not be consistent and will not lead to strong associations.

3 Learning Associations Between Visual Features and Keywords

Developing a learning algorithm using the visual and textual features of the images is a challenging task. First, the training data is incomplete as the words are not specified for the different regions. Second, different types of features need to be extracted and combined. Third, the number of keywords is too large to treat the task as a standard classification problem where each word corresponds to one class. Last, different visual features are not equally important in characterizing different image regions. Highly relevant features for one group of regions may be completely irrelevant for another group. In this thesis, we propose using a data mining approach.

Association rule mining [136] has been used traditionally in applications such

as market basket analysis. It attempts to capture interesting relationships between attributes, thereby enhancing the understandability of the data. Association rule mining has also been applied to image data [137, 138]. However, they have not been fully exploited for the case of multi-modal data to learn relationships among the different modalities.

Using the image representation described in the previous section, a large collection of images could be mined to extract associations between the different feature sets. For instance, using a subset of images similar to those in Figure 8, we can extract association rules of the form:

*"If color is **green** and texture is **regular, fine, with dominant orientation at 90°** then keyword is **grass**." (shape and spatial location features are not relevant).*

*"if color is **yellow** and shape is **round** and location is **top of image**, then keyword is **sun**. (texture is not relevant)*

Using these type of rules, a multi-modal thesaurus can be built to:

- **Auto-annotate:** e.g., for a given region, if its color is "green" and has specific texture properties, one can automatically label it "grass".
- **Perform hybrid search and browsing:** The query can be expanded to include inter-modality associations and to include both visual and textual features. For example, if the user specifies "grass" as the keyword for searching, the search will also include the associated visual features.
- **Expand the query:** The query can also be expanded to include intra-modality associations. For instance a yellow and circular region (sun) may be associated with "reddish" regions located on the top (sunset view). Thus, a query image that has the picture of the sun (yellow) may retrieve sunset images that are not similar in color (red).

Several algorithms could be used to extract association rules from the proposed data representation. However, due to the uncertainties in the images/regions representation (duplicated words, incorrect segmentation, irrelevant features, ...), standard association rule extraction algorithms may not provide acceptable results. Our proposed approach overcomes these limitations by relying on the SCAD_c-CA algorithm described in the previous chapter. This algorithm is designed to search for the optimal clusters' prototypes and the optimal relevance weight for each feature of each cluster. The clustering component would be used to group similar regions and identify prototypical visual profiles. The feature weighting component would guide the clustering process to identify meaningful clusters with subsets of relevant features. We use the seven features (described in section §IV.A.2) as different feature subsets, and for each subset we use an appropriate distance measure. In particular, for F^{WTD} , F^{EHD} , F^{MHSV} , F^{MLUV} , and F^{SHP} we use the Euclidean distance (see eqn. (1)). For F^{HRGB} we use the Quadratic distance (see eqn. (2)) and for F^{TXT} we use the Cosine distance measure (see eqn. (3)).

Clustering all image regions in the database would result in clusters of regions that show common visual and textual attributes. Within each cluster, a correlation can be established between the visual and textual features. This correlation is the principle behind the proposed multi-modal thesaurus and its ability to bridge the semantic gap.

4 Multi-Modal Thesaurus Construction

For each cluster, we use its visual prototype (closest image to centroid), the visual features of its centroid, the relevance weights for each feature subset, and the dominant keywords from the textual feature set to form a visual profile, Q_i . In the following we let $\mathbf{FS}^1, \dots, \mathbf{FS}^K$ refer to F^{HRGB} , F^{MHSV} , F^{MLUV} , F^{WTD} , F^{EHD} , F^{SHP} ,

F^{TXT} respectively. Formally, let c_i^1, \dots, c_i^K represent the centers of the feature subsets $\mathbf{FS}^1, \dots, \mathbf{FS}^K$ of cluster i , and let v_i^1, \dots, v_i^K represent the feature relevance weights of these subsets. The feature subsets of profile Q_i are defined as

$$Q_i^{c_s} = c_i^s, s = 1, \dots, K, \quad (64)$$

and the relevance weights as

$$Q_i^{v_s} = v_i^s. \quad (65)$$

The visual prototype of profile Q_i , denoted Q_i^I , is then defined as the closest image, i.e.,

$$Q_i^I = \text{Image}(\arg \min_{k=1}^N D_i^k), \quad (66)$$

where $\text{Image}(k)$ is the image of region R_k , N is the number of images in the database, and

$$D_i^k = \sum_{s=1}^K \frac{Q_i^{v_s} \times \text{dist}(R_k^s, Q_i^{c_s})}{Q^{D_{avg}^s}}. \quad (67)$$

In (67), R_k^s is the s^{th} feature subset of region R_k . Finally, letting t represent the feature set that represents the keywords (i.e. F^{TXT}), the textual feature set for profile Q^i is defined by

$$Q_i^x = c_i^t. \quad (68)$$

The visual and textual profiles of all clusters constitute the multi-modal thesaurus.

B Unsupervised Image Annotation

The constructed multi-modal thesaurus could be used to annotate new unlabeled images. In the following, for clarity purposes, it is assumed that only one set of color and one set of texture features are used. Given a test image, it is first segmented into homogeneous regions using the CA algorithm following the same procedure used to segment the training images. Then, for each region, R_k , its color

feature , R_k^c , texture feature R_k^t , are extracted and compared to the profiles representatives using

$$D_{ik} = v_i^c \times \text{dist}(R_k^c, Q_i^c) + v_i^t \times \text{dist}(R_k^t, Q_i^t), i = 1, \dots, C. \quad (69)$$

In (69), Q_i^c and Q_i^t are the centers for the color and texture feature subsets of profile i , and v_i^c and v_i^t are their relevance weights. Based on the distances D_{ik} and the distribution of the clusters, several ways could be used to annotate R_k and assign an evidence value to each label. In our approach, we use a fuzzy labeling approach.

1 Fuzzy Membership Generation

Each region R_k would be assigned a fuzzy membership degree in all profiles using:

$$\mu_i(R_k) = \frac{1}{\sum_{p=1}^C (D_{ik}/D_{pk})^{2/(m-1)}}, \quad (70)$$

where m is a weighting exponent that controls the degree of fuzziness.

The process of assigning fuzzy memberships to different image regions in the different concepts is illustrated in Figure 11. Figure 11(a) displays a test image to be annotated. First, the image is segmented into homogenous regions. Figure 11(b) displays the three regions. Then, for each region, visual features are extracted (see section §IV.A.2), and a distance is computed to all profiles in the multi-modal thesaurus. After mapping the distances to memberships (using eqn. (70)) and sorting, the top three profile matches for each region are selected. Figure 11(c) displays the top three matching profiles for each image region. Below the representative region of each profile we display the membership value for the image region in that profile. Above each image profile, we display the top annotations (with their respective centroid feature values). For instance the first region can be assigned to one profile without ambiguity ($\mu = 0.999$). Using this profile one could annotate this region using the keyword "Flower". The third region on the other

hand, is not representative of any of the profiles and is assigned a low fuzzy membership to multiple clusters. Using the keyword feature subsets of the best matching profiles, we could label this region as "grass", "tree", "flower", "leaves", ... etc. with various degrees. In the next section, we outline our approach to combine the membership degrees in all profiles and the labels of these profiles to annotate image regions.

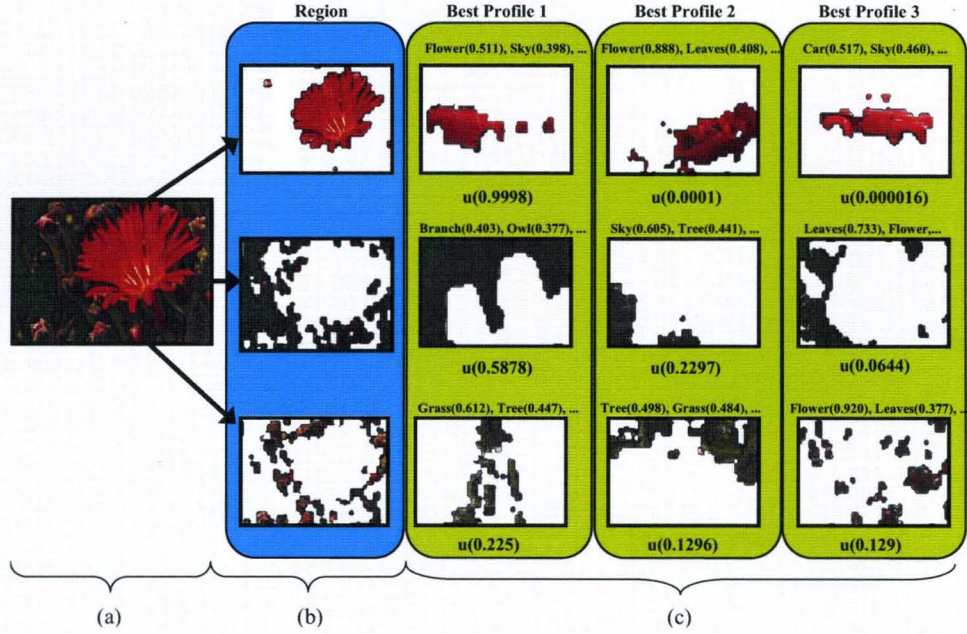


Figure 11. Illustration of the image segmentation and annotation. (a) Image to be annotated. (b) The three regions of the image in (a). (c) Best three profiles matched to each region.

2 Keyword Weighting

The keyword components of the prototypes, i.e. c_i^w , are biased by more frequent words. Frequent words tend to be present in more clusters, and their feature values may not reflect their actual relevance within the cluster. This is a well-known problem in text document classification and categorization. The standard approach to overcome this bias is to weigh the term frequencies by the inverse document frequencies (IDF) [142]. Using a similar approach, we first define

the inverse cluster frequency (*ICF*) of word j as

$$ICF(w_j) = \log(1 + \frac{C}{C_j}), \quad (71)$$

where C_j is the number of clusters that include the word w_j with a significant frequency.

Then, to reduce the bias of the more frequent words, the word frequencies in each cluster i , i.e. components of c_i^w , are scaled by the *ICF* using

$$\tilde{c}_i^w = ICF \times c_i^w. \quad (72)$$

Finally, the evidence value of assigning word w_j to region R_k is computed using

$$Evid(w_j^k) = \sum_{i=1}^C \mu_i(R_k) \times \tilde{c}_{ij}^w \quad (73)$$

In Figure 12, we illustrate our unsupervised annotation on six images. For each image, we show the boundaries of its region and we only show the top word and its evidence.

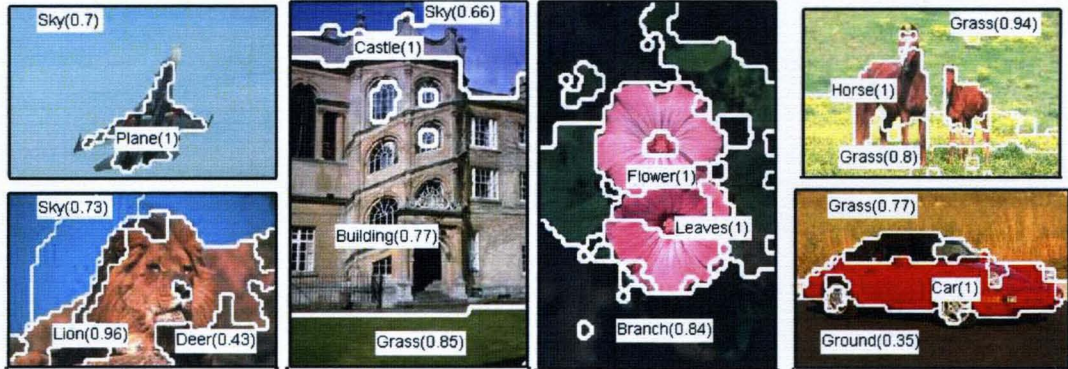


Figure 12. Segmentation and annotation of six images. The regions' boundaries are displayed as white lines, and the top word with its evidence is shown for each region.

C Experimental Results and Validation

In this section, we validate our proposed TBIA approach to learn image semantics and compare it to existing systems using a large image collection. The

data set used for this experiment consists of 9,264 labeled images from the Corel collection. Each image in the training set is manually labeled by 1 to 7 keywords. A total of 97 keywords were used which provide a global description of the images and are not explicitly associated with specific regions. Table 1 displays a list of all keywords. This data set is defined such that there are at least 50 images labeled by each keyword. Each image is coarsely segmented by clustering its color distribution as described in section §IV.A.2. Segmentation of all the training images resulted in a total of 40,051 regions, averaging to 4.32 regions per image. Each region is represented by the sets of visual and textual features outlined in section §IV.A.2.

The image regions, represented by the seven feature subsets, are clustered using the SCAD_c-CA algorithm. In this application, finding the optimum number of clusters (C) is not critical as long as it is large enough to avoid lumping different profiles into one cluster. Here, the results are reported when $C = 400$. Table 2 displays the parameters of the data set and other parameters used for clustering.

Sample results of this clustering step are illustrated in Figure 13. For each cluster, few regions (within each image, the other regions are masked and have a gray color) are displayed and for each image the keywords used to annotate it are shown. As it can be seen, SCAD has succeeded in identifying meaningful clusters of visually similar regions. Moreover, each cluster includes few consistent and dominant keywords that can provide a semantic description of the images assigned to it.

Table 3 displays the relevant feature weights for the four clusters displayed in Figure 13. The more dominant and consistent a feature across the assigned regions is, the higher the associated weights. For instance, the cluster "Sky" has a relatively high relevance weight for the color features and a low weight for the structure feature. Similarly, for the "Flower" cluster, a higher relevance weight is assigned to the texture feature. This is because the color of the different regions assigned to this

TABLE 1

List of words used to label the training images

antelope	cloud	helicopter	road
ape	column	hippo	rock
badger	cow	horse	sand
balloon	crocodile	leaves	sculpture
beach	deer	leopard	seal
bear	desert	lion	sheep
bird	dirt	lizard	skunk
bison	dog	llama	sky
boat	donkey	manatee	smoke
branch	elephant	mane	snake
bridge	fence	miscellaneous	snow
building	field	monkey	squirrel
bus	fire	mountain	stone
bush	fish	mushroom	sun
butterfly	flower	night	tiger
cactus	footballfield	opossum	train
car	forest	owl	tree
castle	fox	people	turtle
cat	frog	person	wall
cheetah	giraffe	pig	water
cherrytree	goat	plane	whale
chicken	grapes	porcupine	wolf
chipmunk	grass	rabbit	zebra
city	ground	raccoon	
cliff	groundhog	rhino	

cluster is not consistent.

In Figure 14 the extracted visual profiles of the clusters displayed in Figure 13 are displayed. For each profile, we show the image of the closest region (Q_i^I) along with a visualization of the features of its centroid. For the $F^{H_{RGB}}$ color features, a 64-bin histogram is displayed. For the $F^{M_{HSV}}$ and $F^{M_{LUV}}$ color moments, the mean is displayed as an image patch. For the $F^{E_{HD}}$ subset, the edge histogram of the five components indicating the proportion of horizontal, vertical, diagonal, anti-diagonal, and non-edge pixels in the region are shown. For the multi-resolution

TABLE 2

Experimental Constant Values used in Application of SCAD_c-CA

Category: Data Set

Definition	Constant Name	Constant Value
# Images	I	9,264
# Keywords	W	97
Max Keywords per Image	m	7
Avg. Image Height	r	236
Avg. Image Width	c	384

Category: Segmentation (CA)

Maximum # Clusters	C_{max}	10
Total Number Regions Created	N	40,051

Category: SCAD

# Feature Subsets	K	7
Maximum # Clusters	C	400
# Initial FCM Iterations	l	10
Discrimination Exponent Max	Q_{MAX}	10
Discrimination Exponent Min	Q_{MIN}	3
Discrimination Exponent Stabilization	Q_{BREAK}	30
Fuzzifier	m	1.1
CA Initial Value	η	0.01
CA Time Constant	τ	10

TABLE 3

Feature relevance weights for the sample clusters shown in Figure 13.

Cluster	$\mathbf{F}^{H_{RGB}}$	$\mathbf{F}^{M_{HSV}}$	$\mathbf{F}^{M_{LUV}}$	\mathbf{F}^{WTD}	\mathbf{F}^{EHD}	\mathbf{F}^{SHP}	\mathbf{F}^{TXT}
"Sky"	33%	12%	13%	14%	14%	7%	8%
"Flower"	6%	11%	10%	33%	22%	9%	9%
"Tiger"	12%	18%	16%	21%	14%	11%	8%
"Plane"	41%	15%	15%	9%	6%	6%	8%

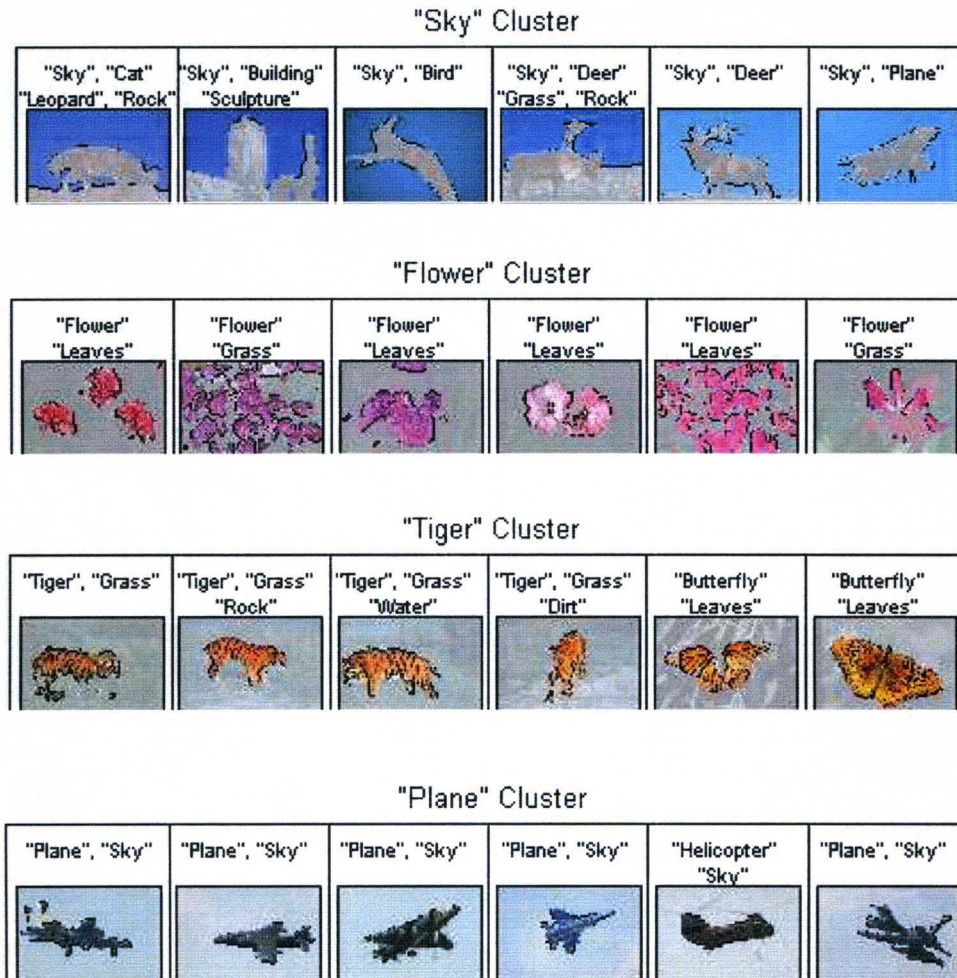


Figure 13. Representative regions from 4 sample clusters. For each cluster few regions assigned to it are shown (The gray part of the image includes other regions not assigned to this cluster). The keywords above each image are those used to provide a global image annotation.

wavelet features, the average values of the different wavelet components are displayed as bars, and on each bar, the variance of the corresponding components is indicated. For the shape features, the values of the five components are listed. For the textual features, the three dominant components are listed. Finally, the relevance weight of each feature set is shown.

The visual profiles correlate features from different modalities. For instance,

using the first profile in Figure 14, it can be deduced that: "if the color is *bluish* and the texture is *smooth*, then the label is *sky*". Similarly, using the second profile, one can deduce that: "if the color is *pinkish* and there are some *fine edges* (high resolution), then the label is *flower*".

In addition to the inter-modality correlation, the clusters identified by SCAD_c-CA can reveal intra-modality correlation. For example, the visual features of the "tiger" and "butterfly" in cluster three of Figure 13 are highly correlated. Also, the visual features of several "sky," "water," and "snow" clusters are correlated. Similarly, from the keywords assigned to the "Tiger" and "Plane" clusters in Figure 14, the words "sky" and "plane" and the words "grass" and "tiger" are highly correlated.

1 Comparative Analysis

To validate the proposed annotation method, two aspects need to be evaluated: how accurate is the algorithm in labeling images and how does it compare to existing methods.

A 4-fold cross validation on the data set is performed to determine how accurate the algorithm is, resulting in 6,948 training and 2,316 testing images per fold. The test images are automatically annotated by various state-of-the-art methods [40, 41, 39, 143]. Then the labeled images are used to determine the word accuracy and the overall accuracy of the labeling approach.

Let $N^U(w)$ be the set of images containing the true keyword (labeled by user) w . Let $N^A(w)$ be all the images containing the label w , generated by a labeling algorithm, within the top five global image labels. The $accuracy(w)$ is then defined

as

$$accuracy(w) = \frac{\| N^U(w) \cap N^A(w) \|}{\| N^U(w) \|}, \quad (74)$$

and the overall method *accuracy* as

$$accuracy = \frac{\sum_{w=1}^M accuracy(w)}{M}. \quad (75)$$

We compare the accuracy of our proposed approach (TBIA) with those generated by the following algorithms.

- **Image-to-Word Transformation (IWT):** This approach, proposed by Mori et al. [41] is a simple method that correlates image regions with keywords. First images are partitioned into block regions using a uniform grid. Then the visual features of the regions are clustered and the likelihood conditional probability is estimated by accumulating the word's frequencies in each cluster. Finally, the query image regions are compared to clusters and their likelihoods combined to find plausible words.
- **Cross-Media Relevance Model (CMRM):** In [39], the joint distribution of regions and words are learned. This distribution represents the cross-media relevance model and defines the underlying probability distribution of an image. The relevance model contains all possible regions and words that could appear. From these distributions, the regions corresponding to each test image are then used to generate words and associated probabilities. Each test image is therefore annotated by a vector of probabilities for all keywords.
- **Pair-wise Constrained Clustering Based Annotation (PCCBA):** In [143], Rui et al. presented a labeling approach that uses constraint based clustering. Their approach is motivated by the fact that image regions with different semantic words but similar appearance may be easily grouped during

clustering due to a sparse feature space. To overcome this problem, they impose some constraints on the process of clustering. This approach associates a cost with violating a constraint between pairs of regions. These constraints are derived by considering the language model underlying the annotations assigned to training images. Annotation is then performed through a greedy selection and joining algorithm that finds independent sub-sets of region clusters and employs a semi-naïve Bayesian model to compute the posterior probability of words given those independent sub-sets.

The *accuracy* of each word using the four considered algorithms is displayed in Figure 15 along with the frequency of that word in the data set. As it can be seen, the IWT and the PCCBA methods both perform well for frequent words. However, the TBIA and CMRM outperform on the lesser frequent terms and across the entire database on average. For clarity, words in Figure 15 that are not found by all methods are discarded. In Table 4 the overall *accuracy* of each system is shown.

The results for TBIA in Figure 15 vary with words based on their respective frequency. The frequent words are presented more in the clustering and, as such, have a higher correct percentage. Additionally, there are also some words that are simply un-predictable; they are either never used or always used in the wrong region.

Our TBIA approach has been shown to achieve a higher overall *accuracy* in image labeling and identify lesser frequent terms than other state-of-the-art approaches. The advantage of the TBIA is not only in its approach to association mining through the multi-modal thesaurus, but in utilizing a better clustering algorithm with feature relevance weights. Through assigning fuzzy memberships to different image regions, our approach does not require an accurate segmentation and allows multiple clusters to affect the outcome of the labeling. The addition of the inverse cluster frequency assists this approach in overcoming the bias problem of

frequent keywords appearing through clusters. This common problem in text document classification is not addressed in most image labeling systems.

TABLE 4

Accuracy of the four labeling methods averaged over all keywords.

Method	TBIA	IWT	CMRM	PCCBA
Accuracy	28%	13%	23%	12%

D Conclusions

In this chapter, we presented an unsupervised approach that extracts representative visual prototypes from large collections of images through a process of clustering and unsupervised feature selection. This approach creates visual profiles corresponding to frequent homogenous regions that are associated with keywords. To accomplish this, manually annotated images are segmented into homogeneous regions. Then, the regions are combined with the image level annotations and clustered into categories of regions that share common attributes. Clusters’ representatives and their parameters are used to create profiles linking low-level image features and high-level concepts.

The second component of our approach uses the multi-modal thesaurus to automatically annotate segmented regions. This part is accomplished through two steps. First, an un-annotated image is segmented into homogeneous regions. Then, fuzzy membership functions are used to label new regions based on their proximity to the thesaurus entries. These annotated regions can then facilitate textual region based searches, or be aggregated into image level annotations. We showed that our approach outperforms state-of-the-art methods on its ability to determine accurate annotations with infrequent annotations. Thus, our approach is more reliable when

the database is very large, and only few labeled samples are available.

In addition to summarizing the large number of regions by few visual prototypes, we showed that the identified clusters could be used to reveal inter- and intra-modality correlations. In particular, the inter-modality correlation could be used to extract associations between visual profiles and textual keywords. These associations, along with the cluster-dependent feature relevance weights, could be used to build a multi-modal thesaurus that could serve as a foundation for inter-modality translation, and for hybrid navigation and search in content-based image retrieval. For instance, a textual query using the terms "grass" could be expanded to include the associated visual features. Thus, allowing the user to use keywords to query unlabeled images. In the following chapters, we will show how these properties can be exploited to develop a CBIR that uses hybrid query and navigation.

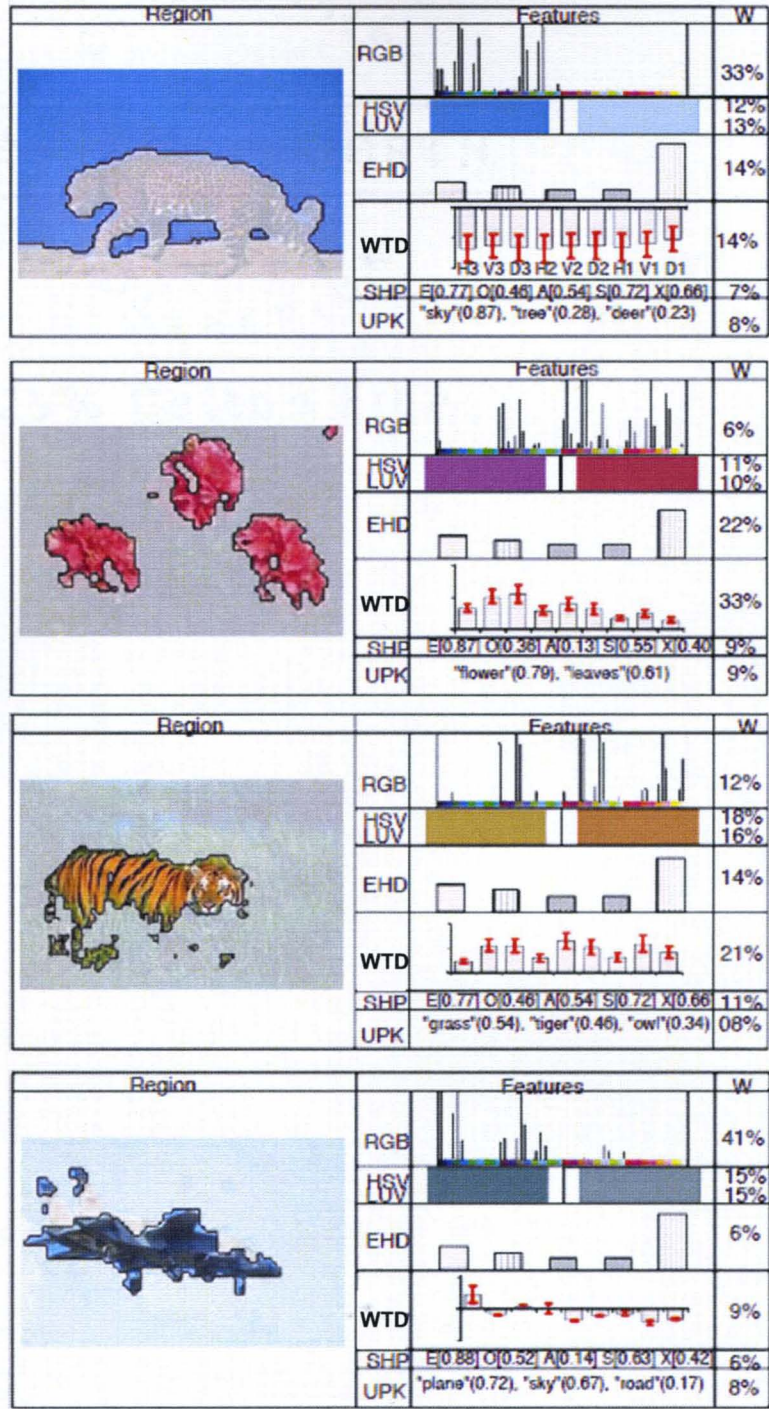


Figure 14. Visual profiles of the clusters in Figure 13. The six feature sets are shown with their representative regions. The F^{RGB} is shown as a 64-bin histogram. The F^{HSV} and F^{LUV} moments are displayed as their mean color. The F^{EHD} is shown as a 5-bin bar plot representing the various angles. The F^{WTD} is shown with the mean and standard deviation of each frequency bank. The F^{SHP} feature lists the five values. Finally the dominant keywords in the cluster are shown as F^{TXT} (User Provided Keywords). To the right of each feature set, we show its relevance weight.

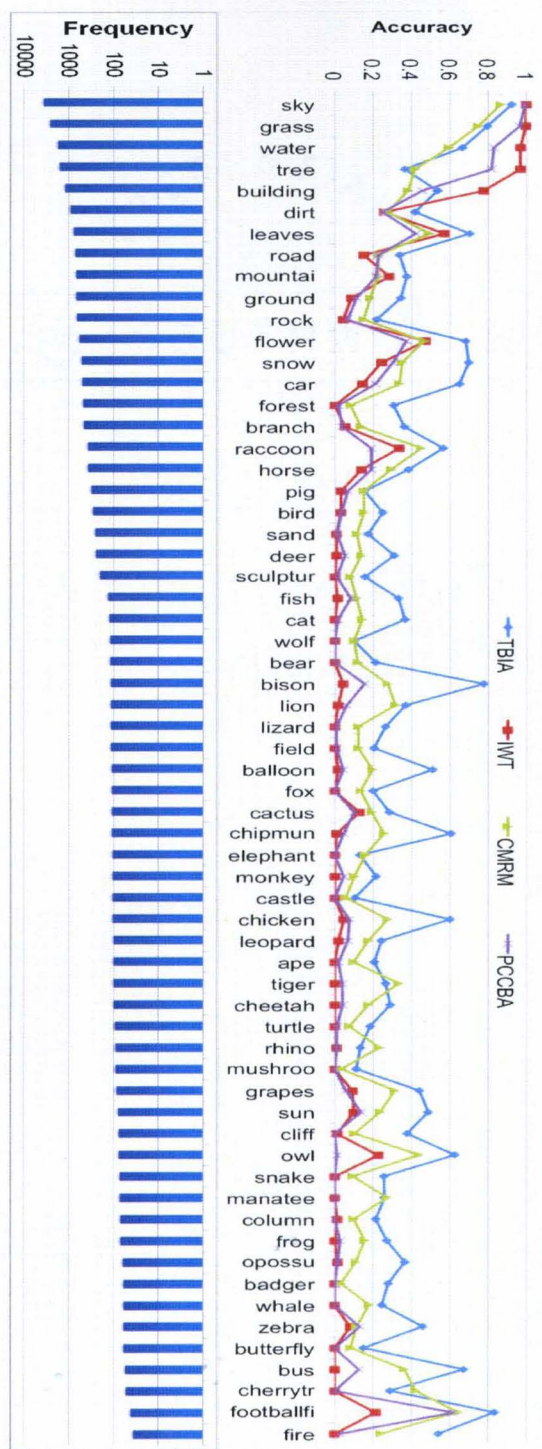


Figure 15. Comparison of the annotation accuracy using the proposed method and three other different annotation algorithms.

CHAPTER V

FUSION OF MULTI-MODAL FEATURES FOR IMAGE RETRIEVAL

As image databases continue to increase in size and become more complex in content, it is becoming impossible to achieve high performance in retrieving visually similar images with a single feature set. As a result, diverse sets of features are being used and combined to provide a more accurate retrieval. However, the task of effectively fusing the output of multiple descriptors has been overlooked by the CBIR community. Only methods that are based on distance scaling or normalization and simple list merging have been used [44]. In fact, the different features can vary significantly with respect to the number of attributes, the dynamic ranges, and the adopted distance measures. Thus, fusion of these features is not trivial and can have a significant impact on the overall performance of the CBIR system.

In this chapter, we present the component of our CBIR system that addresses the aforementioned issues. This component is highlighted in Figure 16. It includes two efficient and effective methods for fusing the retrieval results of the multi-modal features. The first method is based on learning and adapting fuzzy membership functions with the distribution of the features' distances. These memberships are then used to aggregate the results of the different features. The second technique is non-linear and is based on the discrete Choquet integral.

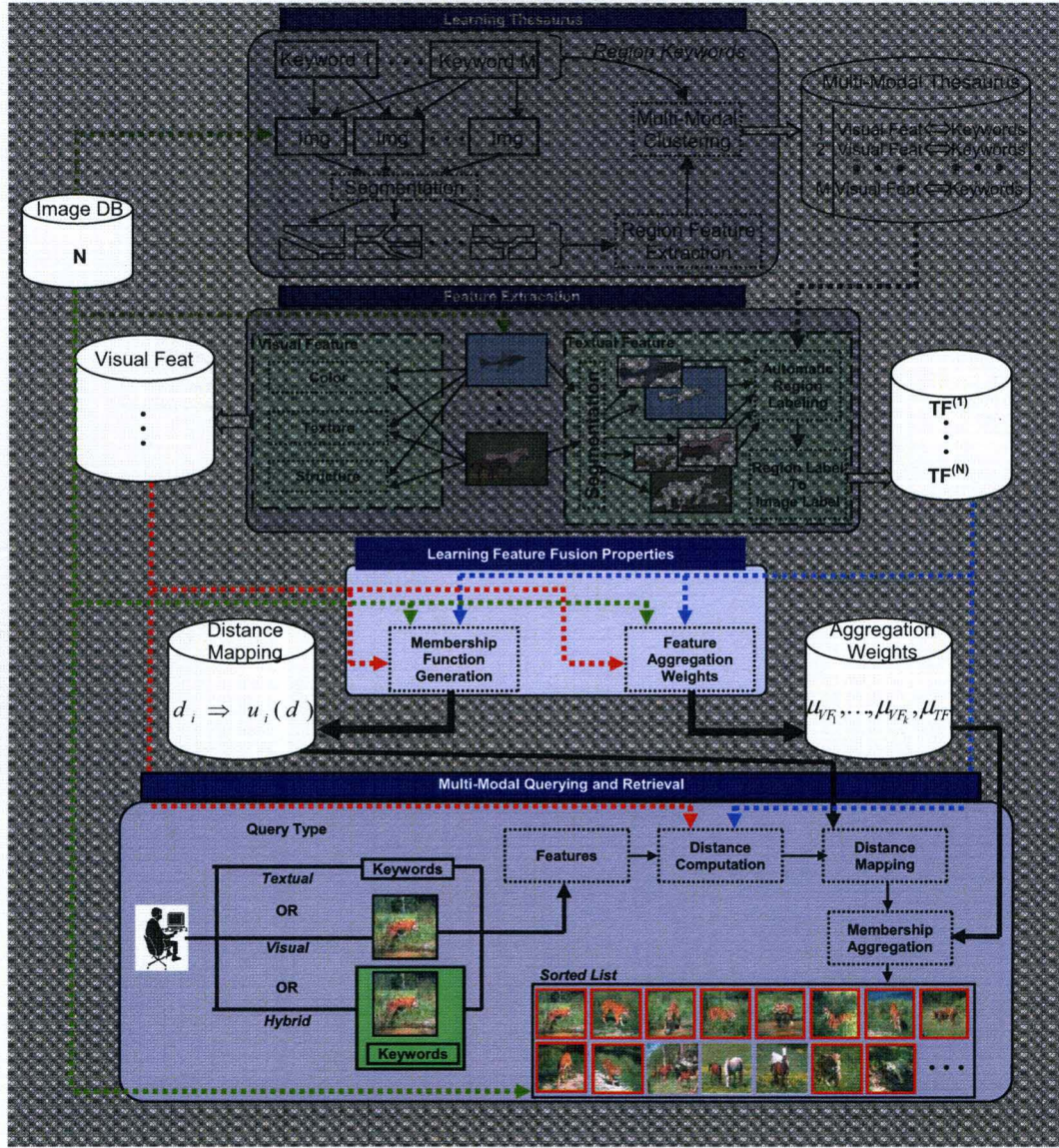


Figure 16. Architecture of the proposed CBIR system with the component that performs multi-modal querying and retrieval highlighted.

A Feature Descriptors

For global image retrieval purposes, features need to be extracted at the image level (not at the region level as in the previous chapter). In our CBIR system, each image is described by standard visual features that include color, texture, and a set of automatically labeled keywords. These features are outlined below. For a more detailed description, we refer the reader to the references.

structure of an image. First, simple edge detector operators are used to identify edges and group them into five categories: vertical, horizontal, 45° diagonal, 135° diagonal, and isotropic (non-edge). Then, the image is divided into 16 sub-images and local, global, and semi-local edge histograms are generated. This results in a 150-dimensional feature vector. We will refer to this feature set as F^{EHD} .

6. Textual Keywords: Using the multi-modal thesaurus, learned as described in the previous chapter, each image is segmented and its regions are labeled by the proposed TBIA algorithm. Since our dictionary includes 97 words, the TBIA Generated Keywords (TGK) feature would include 97 dimensions. Each component is the sum of evidence (computed using eq. (73)) of the corresponding word across all image regions. This results in a 97-dimensional feature vector. We will refer to this feature set as F^{TGK} .

To illustrate the need for fusion of different sets of features, the performance of the individual features is analyzed. Figure 17 displays the precision vs. recall curves for the six features defined above. These curves are generated using a database of 10,000 images from the Corel collection, partitioned into 100 pre-defined categories. The definition of these categories will be defined in section §V.C. For each category, we select 10 images randomly and use them as queries. For each query, we vary the number of retrieved images from 10 to 75 and compute the precision and recall values. Then, the results of all 1000 queries are averaged and displayed in Figure 17. As it can be seen, the performance of the different algorithms can vary significantly. Even though it is easy to rank these features based on their average performance, this does not mean that a given feature (e.g. F^{CSD}) is consistently better than all other features. For instance, in Table 5, one sample image is displayed for each case where one of the features has the largest

number of relevant images among the 50 retrieved images. It is interesting to note that there are few instances where even the F^{HTD} , which has a poor average performance for the used image collection, can retrieve the largest number of relevant images. The textual descriptor (F^{TGK}) also has a relatively low performance. This is because the annotation process is completely unsupervised and thus, is not very accurate. For instance, for the third and fifth images in Table 5, the F^{TGK} feature did not retrieve any relevant images in the top 50 images. This is because these query images were not annotated correctly. This usually occurs if the image segmentation is poor and/or the constructed multi-modal thesaurus does not include words that can describe the image. The above observations emphasize the need to effectively fuse the results of the different features to take advantages of their strengths without being affected by their weaknesses.

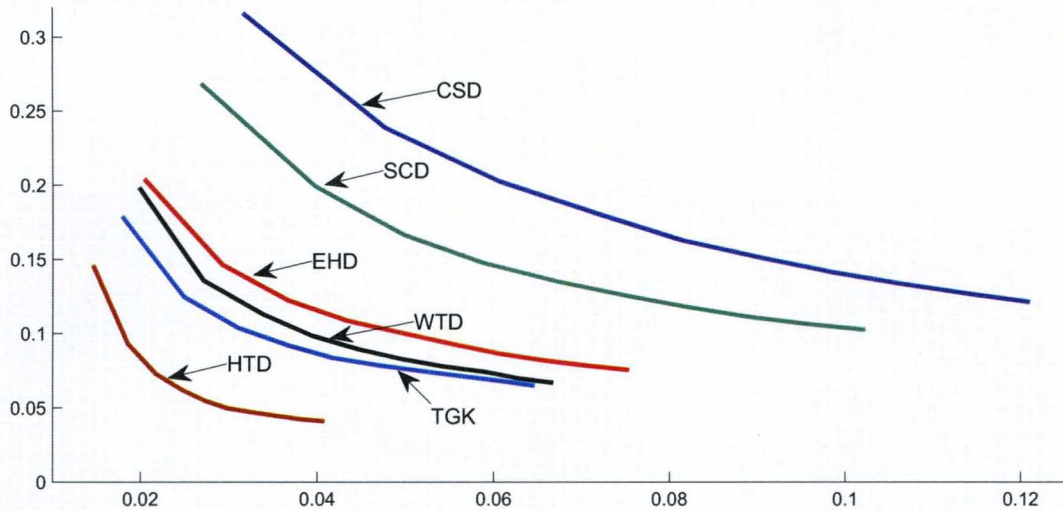







Figure 17. Precision/Recall curves of the six individual feature sets

TABLE 5

Sample query images and the number of relevant images (among the top 50 retrieved) for each feature.

Query	CSD	SCD	HTD	WTD	EHD	TGK
	37	26	2	7	18	21
	21	30	13	22	23	24
	3	3	11	6	4	0
	29	24	10	47	20	15
	8	2	7	17	31	0
	13	14	11	24	10	45

B Multi-Modal Feature Fusion

1 Distance Mapping

For each feature, we learn a fuzzy membership function that maps the distances to the $[0, 1]$ interval. The basic idea is to assign high membership values (close to 1) to distances that are relatively low and low membership values to relatively large distances. These membership functions could be designed based on the distribution of the distances within each feature using a small set of training images.

For simplicity, we use a piecewise linear function to model the memberships

functions as illustrated in Fig. 18. This function is characterized by three points: A, B, and C. These points are learned for each feature k using a set of training images based on the inter-category feature distance distributions. Let M_{kc} be the $n \times n$ pairwise distance matrix for all training images in category c , and let M_{kc}^s be the sorted distance matrix. Let

$$\alpha_k = \frac{\sum_{i=1}^C \frac{\sum_{j=1}^n \sum_{p=1}^3 M_{ki}^s[j][p]}{3 \times n}}{C}, \quad (76)$$

$$\beta_k = \frac{\sum_{i=1}^C \frac{\sum_{j=1}^n \sum_{p=n/2-1}^{n/2+1} M_{ki}^s[j][p]}{3 \times n}}{C}, \quad (77)$$

and

$$\gamma_k = \frac{\sum_{i=1}^C \frac{\sum_{j=1}^n \sum_{p=n-3}^n M_{ki}^s[j][p]}{3 \times n}}{C}. \quad (78)$$

In the above equations C is the number of image categories, and n is the number of training images in each category. In other words, A, B, and C correspond to the category averages of the distances of the three closest images, the three images ranked at the middle, and the three furthest images respectively. The membership function $\mu_k(d)$ of feature k , is then defined as

$$\mu_k(d) = \begin{cases} 1 & \text{if } d < \alpha_k \\ 1 + \frac{0.5}{\alpha_k - \beta_k}(d - \alpha_k) & \text{if } \alpha_k \leq d < \beta_k \\ 0.5 + \frac{0.5}{\beta_k - \gamma_k}(d - \beta_k) & \text{if } \beta_k \leq d < \gamma_k \\ 0 & \text{otherwise} \end{cases}$$

During retrieval, the function $\mu_k(d_k)$ would be used to map the partial distance (using feature set k) between the query image, q , and any image j in the database to a membership value in the $[0, 1]$ interval.

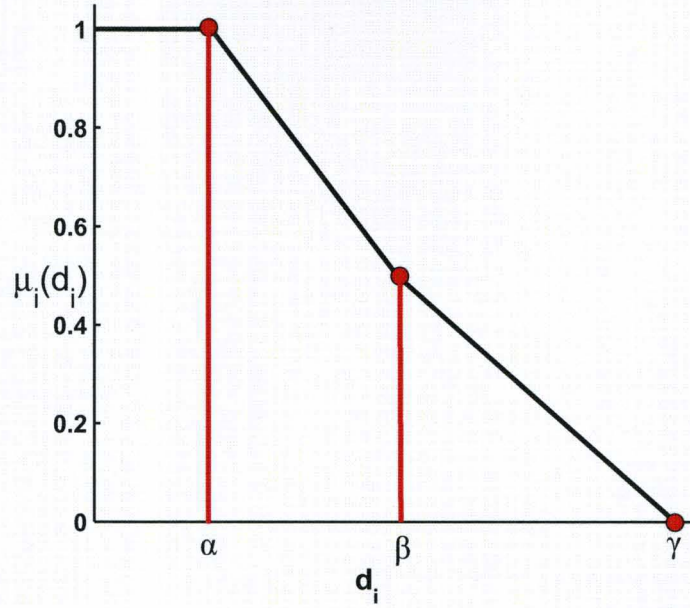


Figure 18. Piecewise linear membership function, $\mu_i(d_i)$ used to map the distances of feature set i into membership values. A, B, and C correspond to the averages of the distances of the three closest images, the three images ranked at the middle, and the three furthest images respectively.

2 Feature Relevance Weights

The different features are usually not equally important. Ideally, the importance of each feature depends on the location of the query image in the feature space and on the user's preferences. Thus, relevance weights for the features should ideally be updated dynamically using a relevance feedback mechanism. Since our proposed CBIR system does not involve feedback, only a global degree of worthiness is estimated for each feature. In particular, a weight, w_{fi} , is assigned to each feature, F_i , based on its relative performance. For instance, the area under the precision/recall curves could be used to estimate the feature relevance weights.

3 Feature Fusion

The features' memberships values and their relevance weights could be combined using several methods. In this chapter, we present two distinct

approaches. The first one is linear, and is based on a simple weighted combination. The second one is non-linear and is based on the discrete Choquet integral.

Sum of Weighted Fuzzy Memberships

The aggregated confidence assigned to each image j in the database that is at a distance $d_i(q, j)$ from the query image is computed using

$$\bar{\mu}(j) = \sum_{i=1}^K \mu_i(d(q, j)) \times w_i. \quad (79)$$

In (79), K is the total number of features to be combined. We will refer to the fusion using the above equation as the Sum of Weighted Fuzzy Memberships (SWFM).

Discrete Choquet Integral

Choquet integral based fusion [144, 145, 146] involves a nonlinear aggregation of algorithm confidence values. The aggregation operator is defined by the discrete Choquet integral with respect to a fuzzy measure. The Choquet integral aggregates confidence values by computing a weighted average of their sorted values. The weights are determined by a function of the fuzzy measure which depends on the ordering of the confidence values.

Using standard notation in this area, a fuzzy measure is defined as follows:

Definition 1 *Let $\mathcal{X} = \{x_1, \dots, x_n\}$ be an arbitrary set. A set function $g : 2^{\mathcal{X}} \rightarrow [0, 1]$ that satisfies the following requirements is called a fuzzy measure if*

1. $g(\emptyset) = 0, g(\mathcal{X}) = 1$
2. $A, B \subset \mathcal{X}$, and $A \subset B$, then $g(A) \leq g(B)$
3. if $\{A_i\}$ is an increasing subsequence of subsets of \mathcal{X} , then

$$\lim_{i \rightarrow \infty} g(A_i) = g\left(\bigcup_{i=1}^{\infty} A_i\right)$$

A fuzzy measure is a Sugeno measure (or a λ -fuzzy measure) if it satisfies the following additional condition for some $\lambda > -1$

$$4. \forall A, B \subset \mathcal{X} \text{ with } A \cap B = \emptyset$$

$$g(A \cup B) = g(A) + g(B) + \lambda g(A)g(B) \quad (80)$$

The value of λ can be uniquely determined for a finite set \mathcal{X} by solving

$$(\lambda + 1) = \prod_{i=1}^n (1 + \lambda g(\{x_i\})). \quad (81)$$

The value $g(\{x_i\})$ is called the density of the measure, and is interpreted as the importance of the single information source x_i .

Let \mathcal{X} be a set, g a fuzzy measure, and $h : \mathcal{X} \rightarrow [0, 1]$ be a function where $h(x)$ denotes the confidence value of x . The Choquet integral of h which respect to the fuzzy measure g can be defined as

$$C_g(h) = \int_x h(x) \circ g = \int_0^1 g(A_\alpha) d\alpha,$$

where $A_\alpha = \{x | h(x) \geq \alpha\}$. If \mathcal{X} is a discrete set, the Choquet integral can be computed as follows

$$C_g(h) = \sum_{i=1}^n [h(x_i) - h(x_{i-1})] g(A_i), \quad (82)$$

where $h(x_1) \leq h(x_2) \leq \dots \leq h(x_n)$, $h(x_0) = 0$, and $A_i = \{x_i, \dots, x_n\}$.

Using our CBIR context, the feature confidence function $\mu(x)$ will be used as the $h(x)$ function, and the feature relevance weights, w_i , will be used as the importance of the single source of information x_i , i.e., $g(x_i) = w_i$.

4 Hybrid Query and Query Expansion

The multi-modal thesaurus and the proposed feature fusion methods allow for hybrid querying, query expansion, and concept refining in a simple and natural way.

Hybrid querying allows the use of profiles that associate low-level features and concepts in the multi-modal thesaurus to more accurately represent the user's perception of the query. In creating a hybrid query, the user first selects a query that may include both images and keywords. The keywords are treated as an independent feature set and its retrieved results are fused with those obtained with the other visual features.

An extension to hybrid querying is query expansion (when concepts are unknown) and concept refining (when features are unknown). Query expansion matches the low-level features in the query to thesaurus profiles and retrieve the most relevant concepts. These concepts are converted to a textual feature vector and appended to the original query. Concept refining consists of taking the user's keywords query and expanding it by finding the best matching profile(s). Then a query of low-level features is created and is used to expand the original textual query. Concept refining can result in multiple queries being sent and their results being combined and ranked for one final set.

To prepare the image database for the use of hybrid querying and retrieval, all the images in the database are labeled offline using the unsupervised image annotation technique described in the previous chapter.

C Experimental Validation

For this experiment, we use the same 9,264 images used in section §IV.C for training. In particular, these images were used for learning the multi-modal thesaurus, the mapping of the different distances to fuzzy membership functions, and the aggregation weights of each feature set. An additional 10,000 images from the Corel collection were used for testing and evaluation. This set is partitioned into 100 categories with 100 images in each category, based on the Corel folders in which

they originated. For each category, we use all 100 images, even if they are not homogeneous. As a result, the ground truth of this collection is not accurate and the overall performance may be low, and not representative of the actual performance of the different features. Thus, only relative performance will be emphasized in the analysis of the results. To generate the queries, 10 images from each category were randomly selected and used exclusively for testing. The remaining images were used to populate the database.

1 Hybrid Query and Query Expansion

First, the performance of the additional high-level feature is evaluated when the textual feature set, F^{TGK} , is used to expand the query image. That is, we simulate a scenario where the user specifies a query image and retrieves similar images without being aware of the semantic labels assigned to the query image or to the other images in the database. That is, the initial query, which contains only low-level visual features of the query image, is expanded to include the automatically generated textual features. In other words, a query that includes low-level features only is transformed into a hybrid query that includes both visual and textual features. For now, the partial results of all individual features (five visual and one textual) are fused using a simple distance scaling method. For each query image, the distances generated by each feature are scaled within a fixed interval, (e.g. $[0,1]$), and the fusion is simply the sum of the scaled distance.

Figure 19 displays the precision vs. recall curve when only visual features are used and when the visual features are expanded to included the textual features. These curves were generated by varying the number of retrieved images from 10 to 75 and recording the number of correctly retrieved images. The results are averaged over all of the 1000 query images (10 images per category selected randomly). As it

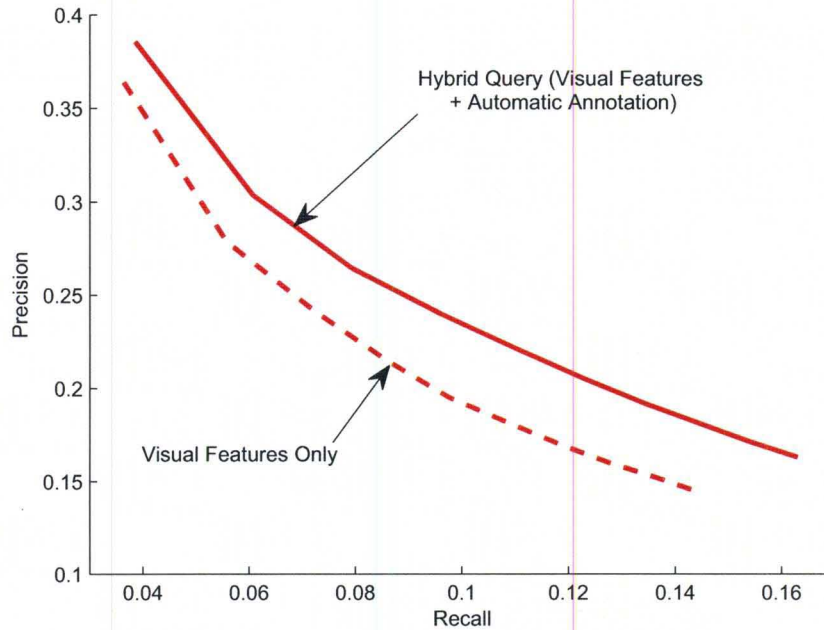


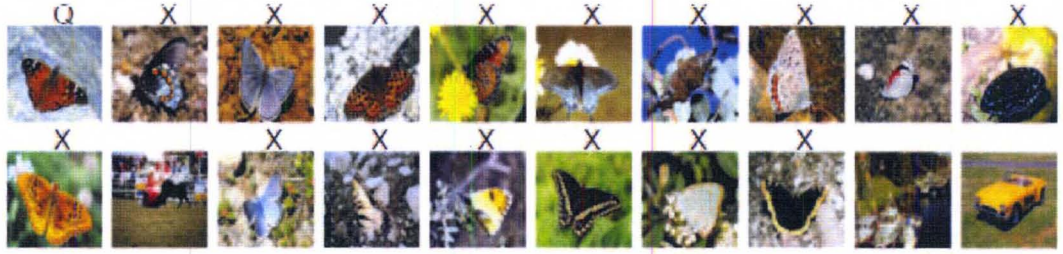
Figure 19. Recall and precision of visual only features versus a hybrid query of visual and textual using query expansion. The results are averaged over the 1000 test images.

can be seen, the additional textual features can improve the performance of the CBIR system significantly.

To illustrate the advantage of the textual features further, we select two query images and display the images retrieved by each method. The results are displayed in Figure 20 and Figure 21. In these figures, the first image is the query image, and the remaining 19 images are sorted in an increasing order of their total distances. For both figures, the hybrid query method retrieved more relevant images. In fact, most of the retrieved images (especially the butterfly image) do not share the same low-level features. They were retrieved because they were labeled correctly using words such as "butterfly", "leaves", "flowers", and "grass".



(a) Query Visual Features Only



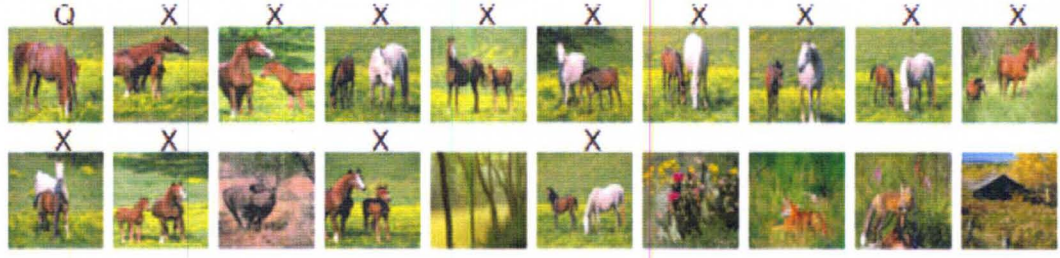
(b) Query Expansion (Visual Features + Automatic Annotation)

Figure 20. Sample query image where query expansion improves the results significantly. The first image is the query image. The others are the top 19 retrieved images, where X indicates that the image is from the same category, and thus is relevant. Most of the retrieved butterfly images do not share the same low-level features but they are labeled by the same set of keywords.

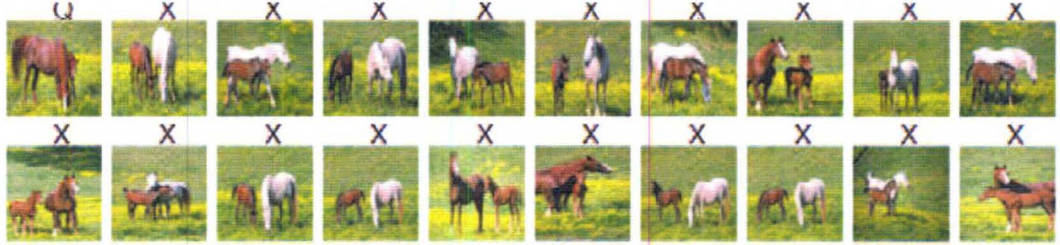
2 Fusion of Multiple Feature Sets

The results of the proposed fusion methods are compared with those obtained using two approaches commonly found in CBIR based on distance scaling or normalization and distance ranking. In the scaling method, for each query image, the distances generated by each feature are scaled within a fixed interval, (e.g. $[0,1]$). The fusion of the different feature is simply the sum of the scaled distance. In the ranking method, the distances generated by each feature are ranked in ascending order. The fusion is computed as the sum of the individual ranks. This method will be referred to as the sum of ranked distances.

To compute the fuzzy measures for the Choquet integral fusion, first the densities of the individual features are computed. For each feature, the area under



(a) Query Visual Features Only



(b) Query Expansion (Visual Features + Automatic Annotation)

Figure 21. Sample query image where query expansion improves the results significantly. The first image is the query image. The others are the top 19 retrieved images, where X indicates that the image is from the same category, and thus is relevant. Labeling by the correct keywords allows the retrieval of images missed by the low-level features.

the precision/recall curve to estimate these densities is used. A relatively more reliable feature would have a larger area, and thus, would be assigned a larger density value. The values of these densities, computed using eqn. (80) and the curves in Figure 17, are shown in Table 6.

For the linear fusion, the results are evaluated when the different features are equally weighted (i.e., $w_i=1$, for all i), and when the different features are weighted using the weights in Table 6 (i.e., $w_i=g_i$).

The precision/recall values, averaged over the 1000 test query images are displayed in Figure 22. In this figure, the performance of the best individual feature is shown as a reference curve. First, we note that all fusion methods can improve the results significantly. Second, of all fusion methods, the sum of the ranked

distances method has the worst performance. This is due mainly to the fact that this approach assigns integer ranks to the individual features and ignores the relative values of the distances. In other words, it does not take into account distances that are clustered or the possible large gaps in the sorted distances.

The other fusion methods have comparable results at high recall values. That is, when a large number of images are considered, these methods retrieve comparable number of relevant images. However, at a higher precision, when fewer images are considered, the Choquet-based and the sum of weighted fuzzy membership fusions have better performance. This means that these methods do a better job at ranking the relevant images. This is mainly due to the weights (or densities) assigned to the different features based on their average performance on a training set.

Even though some of the fusion methods have comparable average performances, their results on individual query images can vary significantly. In Figure 23 and Figure 24, the closest 9 images to two sample queries using three fusion methods are shown. In Figure 23, the fusion based on the sum of weighted memberships outperforms the other methods, while in Figure 24, the fusion based on the Choquet integral outperforms the other methods. The difference between these methods is more significant if more images are ranked and displayed.

TABLE 6

Feature relevance weights of the six individual feature sets determined through precision/recall used in fusion.

Feature	\mathbf{F}^{CSD}	\mathbf{F}^{SCD}	\mathbf{F}^{HTD}	\mathbf{F}^{WTD}	\mathbf{F}^{EHD}	\mathbf{F}^{TGK}
\mathbf{W}_f	0.238	0.220	0.118	0.167	0.164	0.165

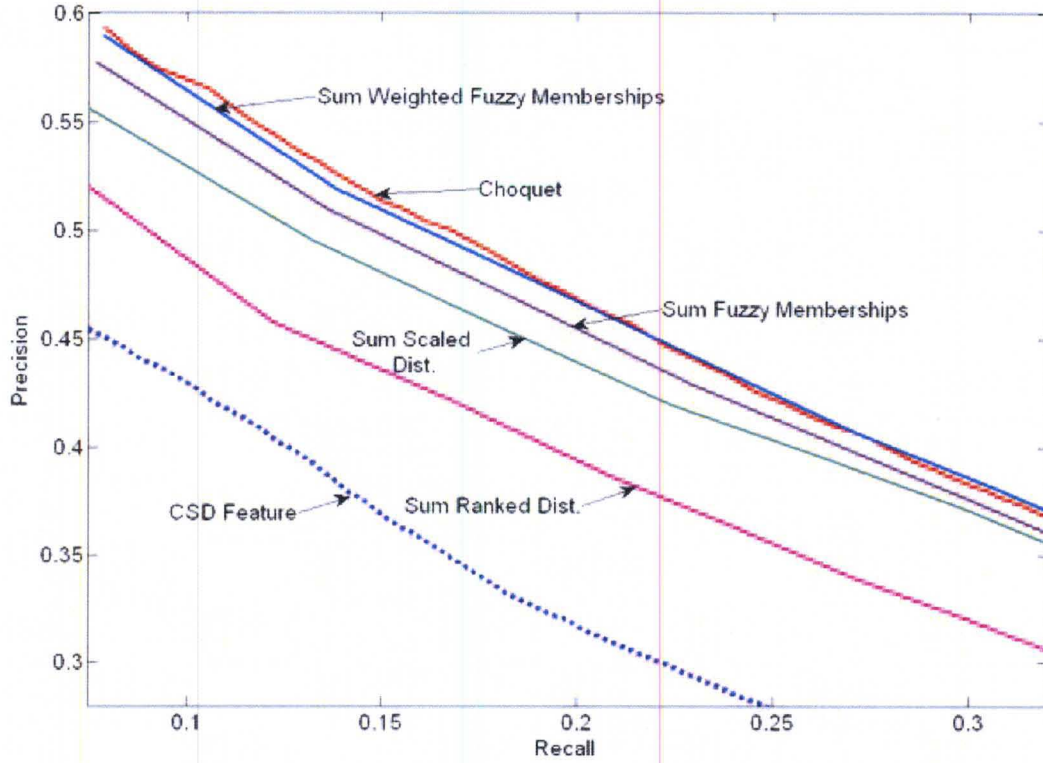
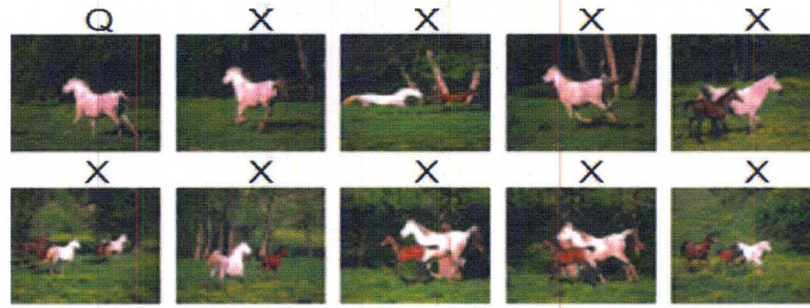


Figure 22. Precision/Recall of the proposed fusion method (Sum of weighted fuzzy memberships) and the standard sum of scaled distances. The performance of the best individual feature (F^{CSD}) is shown as a reference.

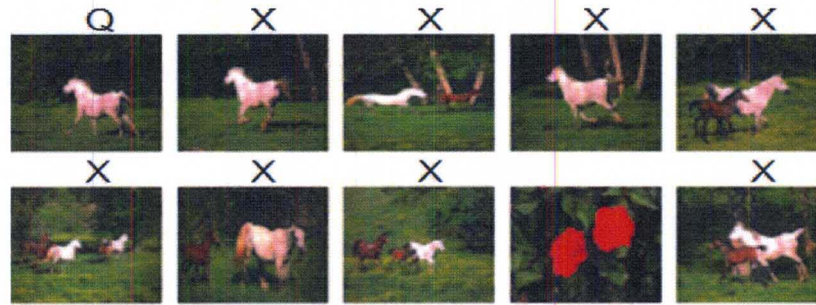
3 Subjective Evaluation

The precision/recall curves on fixed categories can provide limited information on how a retrieval method *should* perform. To truly assess the performance of a retrieval system some form of user satisfaction needs to be measured. In this section we describe our subjective experiment that was designed to provide a quantitative measure on the user's level of satisfaction for several query images.

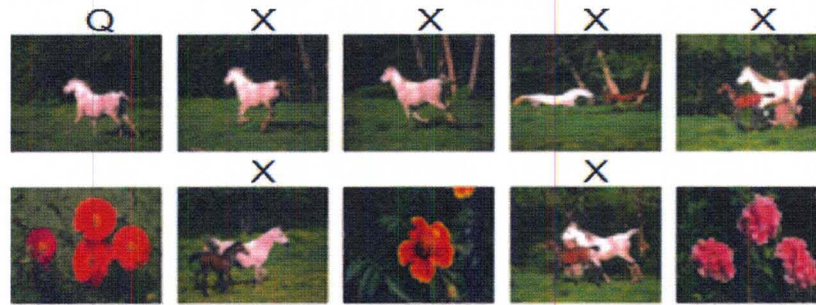
For this experiment, a more realistic data set is simulated containing 55,000 Corel images with many overlapping categories. From the 55,000 available images, 25 are randomly selected as query images. Four different retrieval approaches will be analyzed and compared in this experiment. The first one is a standard CBIR



(a) Fusion based on sum of weighted fuzzy memberships



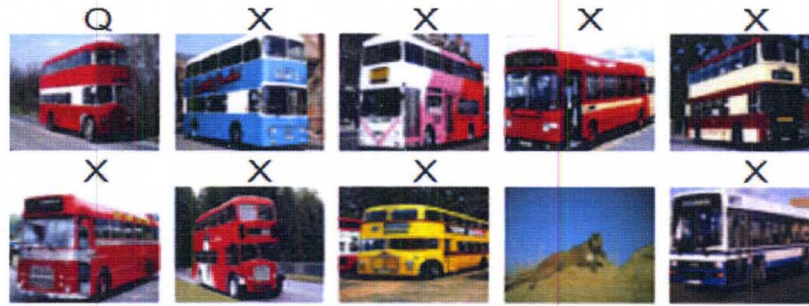
(b) Fusion based on the discrete Choquet integral



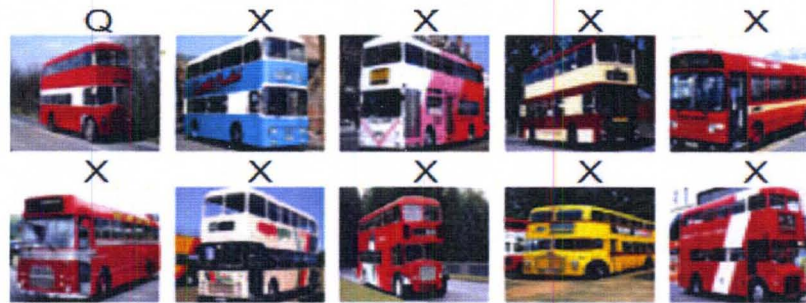
(c) Fusion based on the sum of scaled distances

Figure 23. Sample query image where the fusion based on the sum of weighted memberships outperforms the other methods. The first image is the query image. The others are the top 9 retrieved images, where X indicates that the image is from the same category, and thus is relevant.

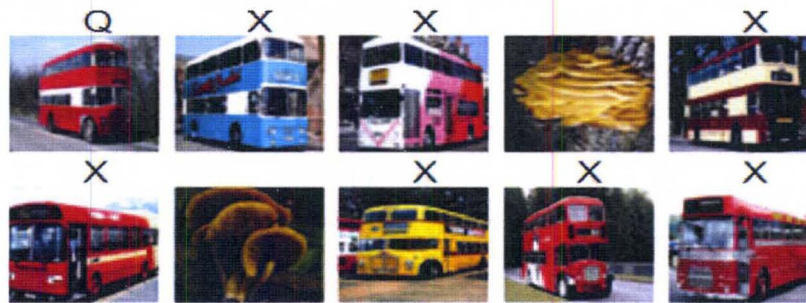
method that uses the sum of scaled distances (SSD). The second approach is the sum of weighted fuzzy memberships (SWFM). To compare the results of the SSD and fusion approach, all the features in the retrieval are used except the semantic information (TGK). To compare the effect of text, and in particular hybrid



(a) Fusion based on sum of weighted fuzzy memberships



(b) Fusion based on the discrete Choquet integral



(c) Fusion based on the sum of scaled distances

Figure 24. Sample query image where the fusion based on the Choquet integral outperforms the other methods. The first image is the query image. The others are the top 9 retrieved images, where X indicates that the image is from the same category, and thus is relevant.

querying, we manually labeled each of the 25 queries. These labels are compared to the automatically generated labels of the images in the database. This additional textual feature is added and all 6 feature sets are fused using the SWFM ("SWFM with text"). This approach permits all users of the experiment to have the same

textual feature and results. The final method of retrieval is trivial and examines what would happen if just random images are retrieved, thereby justifying the need for a CBIR system.

For each query, the top ten images from each of the four approaches are retrieved and stored. A user is shown the results of all algorithms at the same time for one query. Figure 25 displays a snapshot of our user interface. The query image is on the left of each row and the retrieved images per algorithm are beside it. To prevent any prior knowledge from influencing the user's preference, the algorithm results are randomly placed on different rows (i.e. Alg A for the last query may be Alg B in the next query). The interface also guarantees that if any images are present in multiple rows (approaches can return the same top images to the same query), they will receive the same rating. The available choices for the query-result pairs are: "Poor", "Minimal", "Average", "Reasonable", and "Good". No instructions are given to the user, so the definitions of these choices are purely opinionated based on the user's understanding of image retrieval systems.

Figure 26 shows the overall user satisfaction for each query per algorithm for the 67 users that participated in this evaluation. This value is calculated as a weighted query precision, where the weight is the *match* value assigned. Except for one of the 25 queries, both fusion methods outperform SSD, in most cases by 20%. In query 24 the SSD shows higher user satisfaction than SWFM, even though both still fall within the "minimal match" boundaries. For this query adding semantic information allows the fusion to jump to an "average match" range, above SSD. In general the fusion with text either matches or does better than fusion without text. In Figure 27, where the user satisfaction is averaged over all 25 query images, both fusion methods clearly outperform SSD and all methods are better than random which was expected. The overall algorithm averages are shown when only five



Figure 25. Screen shot of the Subjective Test Online Interface.

images are retrieved versus ten to determine the level of precision significance in the initial five. As Figure 27 shows there is almost a 10% increase in user satisfaction for SWFM and SWFM with text when only the first five results are examined, while

only about a 5% increase for the SSD. This phenomenon is attributed to the fusion methods returning images with better ranks, as Figure 22 suggested they should.

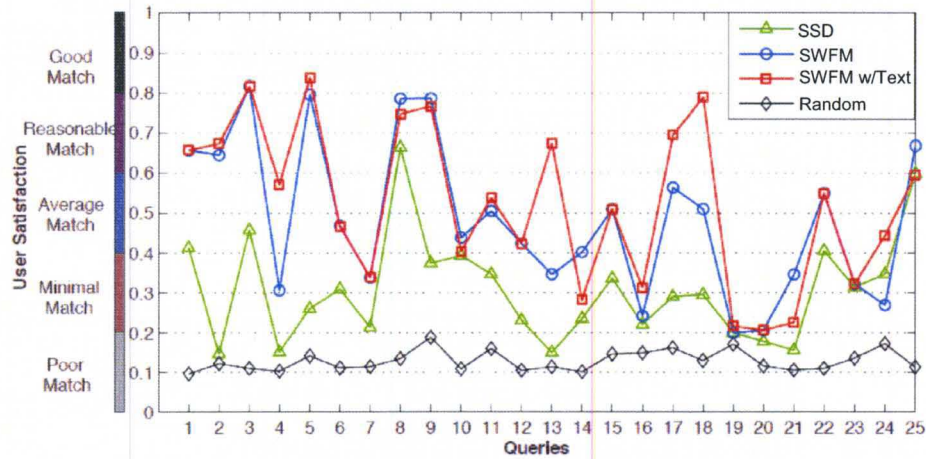


Figure 26. Average Subjective Test Results for Individual Queries.

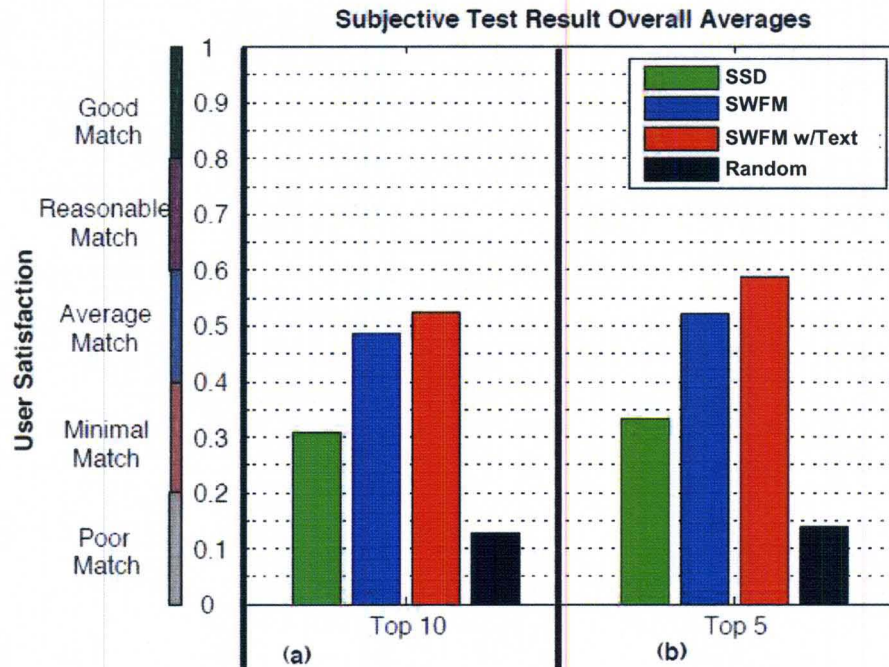


Figure 27. Overall Average User Satisfaction for Algorithms, a) Using Top 10 Images to Query and b) Using Top 5 Images to Query.

As part of the anonymous registration for the subjective test, users had to indicated their level of image retrieval experience (i.e. *Inexperienced*, *Average User*,

Knowledgeable, and Expert). Figure 28 shows the overall user satisfaction when the experienced users results are separated from the inexperienced. Here *experience* is defined as "knowledgeable" and "expert". Interestingly enough, the experienced users satisfaction is approximately 5% lower on all approaches then the inexperienced. We attribute this to the face that a knowledgeable user maybe more critical of the results and using the full range of possible satisfaction choices, whereas an inexperienced user may tend to use just "poor match" or "good match".

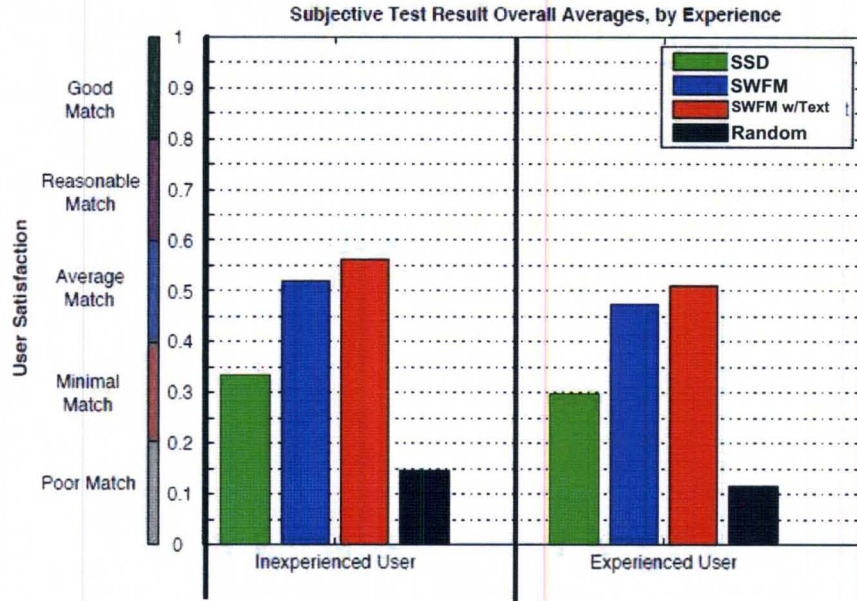


Figure 28. Overall User Satisfaction Classified by Experience.

These results confirm that even with a very large set of data (we are retrieving 0.018% of available information), the proposed fusion approaches and especially fusion with semantic information can considerable outperform the standard method of combining multiple features.

D Conclusions

In this chapter, we presented a generic approach to fuse the outputs of multiple features for CBIR. Our approach is based on mapping the distribution of

the distances of each feature to a fuzzy membership value, and assigning a degree of worthiness to each feature based on its average performance. The memberships and the feature weights are then aggregated to produce a confidence that could be used to rank the retrieved images. Two aggregation methods were presented and evaluated. The first one is linear and is based on a simple weighted combination. The second one is non-linear and is based on the discrete Choquet integral. Both approaches are computationally efficient and involve only simple multiplication and summation of the outputs of the individual features. The Choquet integral involves additional sorting of the individual outputs which is not a significant task if only few features are used. Thus, both methods could be used to fuse the results in a real-time mode.

The proposed CBIR system was validated and compared using a set of 10,000 pre-categorized images. Standard MPEG-7 features and a textual set of feature extracted automatically using our image annotation approach are used. It was shown that the system can improve the overall ranking of the retrieved images significantly and thus, provides a higher precision, especially at low recall values.

Our CBIR system was evaluated further using a larger set of 55,000 generic color images by analyzing the user's response to the retrieved images. The subjective test proved that our proposed CBIR system outperforms other systems.

Currently, our CBIR system is trained globally using simple membership functions and a set of training images. It is possible to integrate a relevance feedback component into the CBIR system to adapt the fusion parameters. In particular, the user's feedback could be used to adjust the parameters of the membership functions and to adjust the degree of worthiness assigned to each feature.

CHAPTER VI

REGION-BASED IMAGE RETRIEVAL

Most existing CBIR systems are based on global image features and have limited capabilities because they cannot capture local variations of the image properly. To overcome this deficiency, region-based image retrieval (QBvR, see section §II.D.2) has been proposed. Rather than deploying global features over the entire content, QBvR segments images into a number of homogeneous regions, which ideally should correspond to objects, and extract local features for each region. QBvR allows the user to search for images containing objects similar to those in a reference image. This object-level representation is intended to enhance the ability of capturing as well as representing the focus of the user's perception of image content. The main limitations of QBvR is that image segmentation is not a trivial task and image segments do not usually correspond to objects. Moreover, searching with multiple reference regions is less obvious to solve, and computationally is much more expensive.

Another issue in existing CBIR systems is known as the "*Page Zero*" problem [85]. This questions how can the user begin the search process without an example image. In other words, these systems assume that the user has a relevant starting point, which may not be always valid. In this case, alternative visual browsing techniques can help by providing an overview of the database. The "*Page Zero*" problem is more critical in QBvR as the query segments may come from different images.

To address the "*Page Zero*" and the computational efficiency problems in QBvR, Fauqueur et. al. [147, 148] proposed an approach that allows the user to perform mental image search by formulating a boolean composition of region categories. In this method, all images are segmented and their regions are categorized through clustering. One region (closest to cluster center) is selected to represent each category, and these regions are presented to the user to initiate the query process. The user can compose a boolean expression of region representatives. The search process is performed through the use of inverted tables of the region category labels, and thus, is very efficient. This QBvR system is unique in letting a user say "I want images that contain regions similar to these, and not similar to these". One limitation of the Fauqueur's approach is that a large number of images could be retrieved and these images are presented to the user without any sorting. Moreover, this system relies on visual features only. This is despite the fact that recent research has shown that the inclusion of automatically generated textual features, even if they are not perfect, can improve the results of CBIR systems significantly [35, 36]. This is because visual similarity is weakly correlated with the measures of similarity adopted for image comparison [149], and using text as an additional feature has the advantage of evaluating image similarity at a higher level of abstraction, providing better generalization.

In this chapter, we build on Fauqueur's mental image search application, enhancing it with recent advances in bridging the semantic gap for increased user satisfaction. This extension allows the user to formulate hybrid queries by selecting reference image regions and/or textual keywords that should (or should not) be included in the target images. This QBvR process is naturally facilitated by the boolean composition, and exhaustive search is not required. The highlighted section in Figure 29 shows the architecture of the region based image retrieval approach in

our CBIR system.

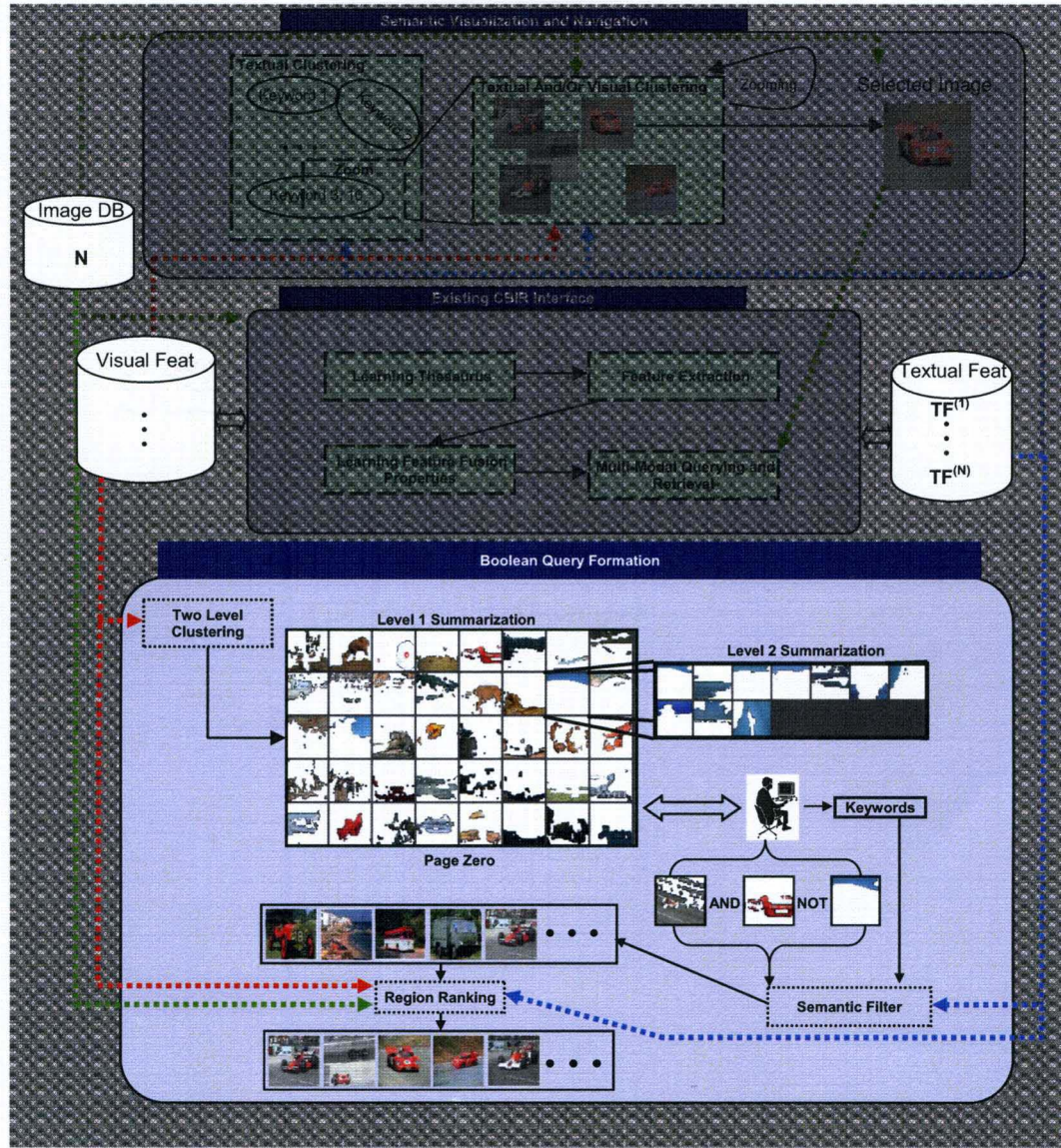


Figure 29. Overview of the proposed CBIR system component to perform region based image retrieval.

A Hybrid Region Indexing

We propose expanding Fauqueur’s mental image search [147, 148] by learning a multi-modal thesaurus (see section §IV.A) and integrating textual keywords in the indexing and retrieval processes. We use the Thesaurus Based Image Annotation

(TBIA) method [150] previously presented to learn image semantics in an unsupervised way. This approach has the advantages of being computationally simple, assigning labels with soft confidence values, and more importantly, assigning labels at the image region level.

The visual profiles of the identified region categories will be used to provide the user with a visual summary of the image database content. That is, they will be used to construct "*page zero*" of the QBVR system. The user can formulate a query by selecting regions that should and/or regions that should not be included in the retrieved images. To retrieve relevant images efficiently, we use an indexing scheme that combines techniques used in the classic text-based information retrieval with techniques used in [148] for visual information retrieval to facilitate hybrid query and retrieval.

First, we introduce few indexing tables to provide associations between images, region categories, and keywords. Let $CI(I)$ be a table that links an image I with all region categories that contain one of its regions. Similarly, let $IC(C)$ be the inverted table of $CI(I)$ which lists all the images that have at least one region assigned to category C . For the textual keywords, we first annotate the images using the learned thesaurus. Each region of each image is annotated based on the proximity of its visual feature to the prototypes in the thesaurus [151]. Then, we create indexing tables for the annotating keywords. Let $TC(W)$ be the set of images that contain regions annotated with word W , and let $Y(W)$ be a set of learned synonyms to W (including the word W itself).

Since the region clustering is not perfect, and since the user may have only a vague idea of what he/she is looking for, we expand the indexing scheme by implementing a range-query mechanism which includes the neighbors of a category. In [148], the authors define the neighbor category of category C_q as the set of

categories C_j such that

$$N^\gamma(C_q) = \{C_j, j = 1, \dots, N \mid \|p_q - p_j\| \leq \gamma\}. \quad (83)$$

In (83), N is the total number of region categories, p_q and p_j are the prototypes (feature vectors of the centroids) of categories C_q and C_j respectively, and γ is a fixed range radius threshold. When the user selects a region category, C_q , all categories within $N^\gamma(C_q)$ will also be considered.

The definition of $N^\gamma(C_q)$ in (83) makes sense only when the clusters are well defined, have spherical shapes, and the centroids are good representatives. However, this is rarely the case for generic image databases where images are represented by high-dimensional feature vectors, and where boundaries between image categories are fuzzy. To overcome this limitation, instead of using (83), we define $N^\gamma(C_q)$ using the distribution of all the regions within a category instead of a single representative point. In particular, we let

$$N^\gamma(C_q) = \{C_j, j = 1 \dots N \mid \frac{\sum_{i=1}^T u_{qi} - u_{ji}}{\sum_{i=1}^T u_{qi} + u_{ji}} \leq \gamma\}, \quad (84)$$

where T is the total number of image regions in the database, and $u_{qi} \in [0, 1]$ is the fuzzy membership of region i in category q . These membership values are generated by the fuzzy clustering algorithm used to categorize the regions. Using this fuzzy similarity definition, two categories are similar if most regions have similar membership degrees in the two categories. Using (84), we create a fifth indexing table, $N(C)$, to implement the range query. For each category C_q , this table includes a sorted list of all similar categories.

B Retrieval by Boolean Composition

The learned multi-modal thesaurus constitutes the query interface. This interface consists of iconic images of the regions' representatives and a list of

keywords. The user can easily formulate queries using the visual prototypes, the textual keywords, or a combination of both. Figure 30 displays a snapshot of our interface.

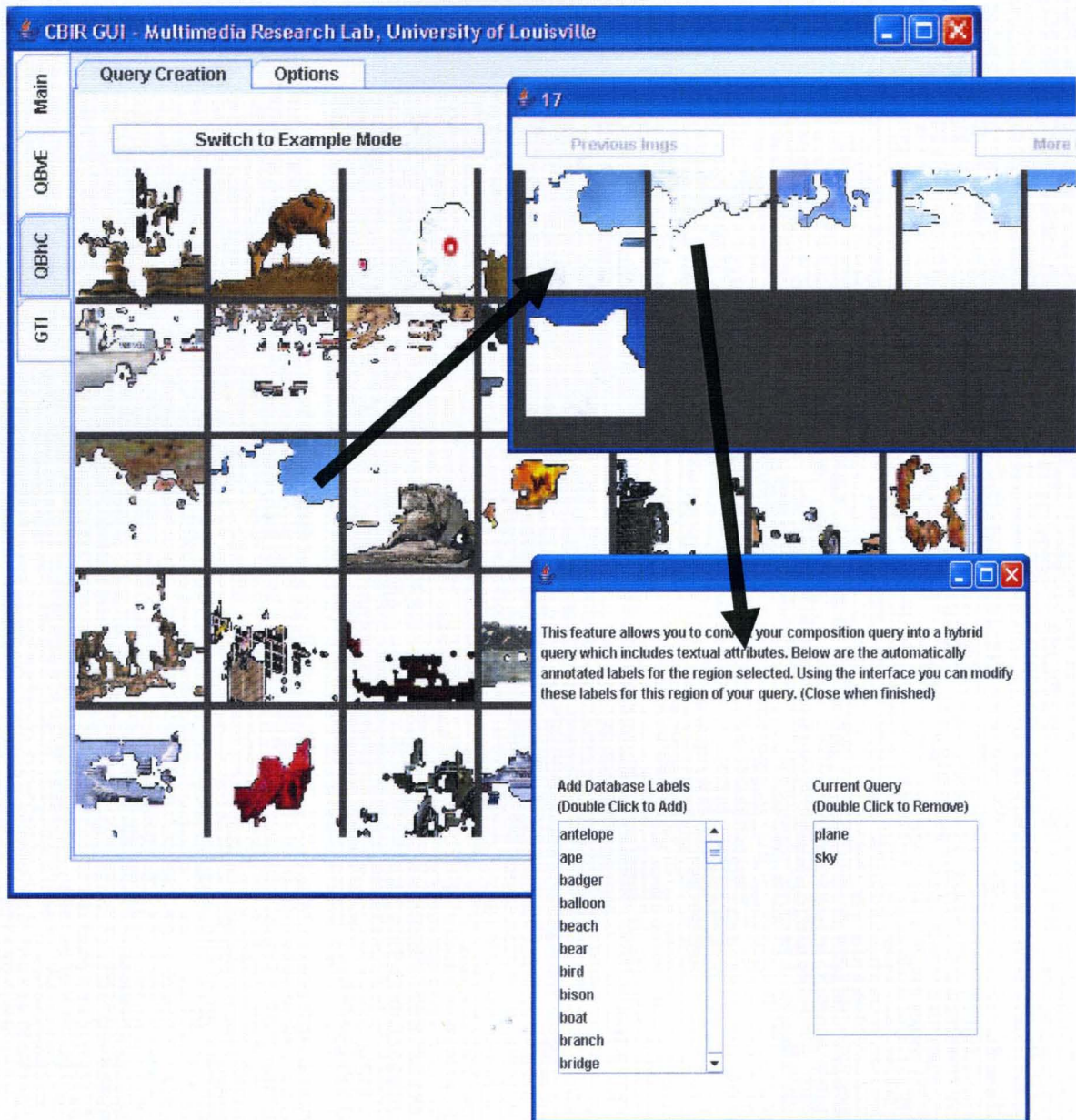


Figure 30. Snapshot of the Region Query Interface.

1 Query by Boolean Composition of Visual Prototypes

This mode allows the user to formulate queries such as: *"Find images that include regions similar to these ones but with no regions like these ones"*. Using the visual interface, the user selects positive and/or negative region categories. Let the Positive Query Categories, $PCQs = \{C_{pq_1}, \dots, C_{pq_M}\}$, represent the set of regions that the user has selected to be included in the target images. Similarly, let the Negative Query Categories, $NCQs = \{C_{nq_1}, \dots, C_{nq_R}\}$, represent the regions that the user has selected to be excluded. Each of the categories included in PCQ and NCQ would be expanded using its neighboring categories. Using the IC inverted tables, the system retrieves images that satisfy

$$S_{result} = S_Q \setminus S_{NQ}, \quad (85)$$

where

$$S_Q = \bigcap_{i=1}^M \left[\bigcup_{C \in N^\gamma(C_{pq_i})} IC(C) \right],$$

and

$$S_{NQ} = \bigcup_{i=1}^R \left[\bigcup_{C \in N^\gamma(C_{nq_i})} IC(C) \right].$$

2 Query by Boolean Composition of Keywords

In this mode, the user specifies a set of keywords that should and/or should not be included in the target images. These keywords could be specified from a list. They could also be extracted automatically from the labels assigned to the region categories selected by the user (as illustrated in Figure ??). In this thesis, we report the results using the latter method. Let $WC(C_q)$ be the set of words that annotate the query category C_q . Using the indexing tables $TC(W)$ and $Y(W)$, and the neighbor expansion of C_q , the system retrieves images that satisfy

$$T_Q = \bigcap_{i=1}^M \left[\bigcap_{w \in WC(N^\gamma(C_{pq_i}))} \bigcup_{z \in Y(w)} TC(z) \right]. \quad (86)$$

We should note that, for simplicity, (86) includes only the set of positive query categories. This equation could be easily expanded to include the negative set, NCQ , as well.

3 Query by Hybrid Boolean Composition

Equations (85) and (86) could be easily combined to retrieve images that satisfy both visual and textual conditions. One simple way to achieve this is to retrieve images that satisfy

$$S'_{result} = (T_Q \cap S_Q) \setminus S_{NQ} \quad (87)$$

C Ranking of Hybrid Boolean Composition

A drawback in the Fauqueur approach [147, 148] is the inability to sort the results when a large number of images are retrieved. We propose a method for sorting the returned images after the result set is defined. This is equivalent to having a query image and comparing it to a smaller list, S'_{result} , of filtered images using a normal CBIR approach. Since our image database is greatly decreased through the boolean composition approach, computational complexity is not dramatically increased. In this method, the query image is the union of the regions from the PCQs (as if they were from one image), and the query mode is a hybrid query utilizing the expanded labels available. For now we use the simple Sum Squared Distances (SSD) approach to determine the overall similarity. However, more effective methods could be implemented using feature membership, categorization feature weights, and feature fusion methods [152] as outlined in the previous chapter. Let

$$D_j^2 = \sum_{k=1}^M \left(\frac{\sum_{p \in R(k,j)} \sum_{s=1}^{K+1} v_{ks} (d_{pk}^s)^2}{|R(k,j)|} \right), \forall j \in S'_{result}. \quad (88)$$

where j is an image in S'_{result} , v_{ks} is the relevance weight of feature s in category k (identified during clustering with SCAD), M is the number of PCQs, and K is the number of visual features ($K + 1$ is if we include the textual feature). $R(k, j)$ is a function that returns the set of regions in j that are contained in $N^\gamma(C_{pqk})$.

In (88), D_j^2 will return distances representing how close an image in S'_{result} is to the original mental query. This ordering could be used to reduce the list of retrieved images, or to simply allow the user to focus on the first few images in the list.

D Experimental Results

For this experiment, the 9,264 images from section §IV.C were used to create the multi-modal thesaurus. An additional 4,000 Corel images were used for testing and evaluation. Each image is coarsely segmented by clustering its color distribution. The Competitive Agglomeration (CA)[71] was used to cluster each image into an optimum number of regions (as outlined in section §IV.A.2). Segmentation of all the test images resulted in 17,514 regions. Each region is then characterized by the a set of standard descriptors defined in section §IV.A.2. The 17,514 image regions were clustered using SCAD with $C=200$. Figure 31 displays a sample of 20 region representatives.

To illustrate the behavior of the proposed system, we assume that the user selects two positive and two negative query categories. These categories are shown in Figure 32. Here, it can be assumed that the user is interested in retrieving images that contain a horse or a deer on grass but not flowers. Using only the visual features of these categories and their neighbors (expanded using (84)), and using (85), the system retrieves 671 images. Using this setting, our system behaves similar to Fauqueur’s mental image search system [148]. Next, we use the labels assigned by

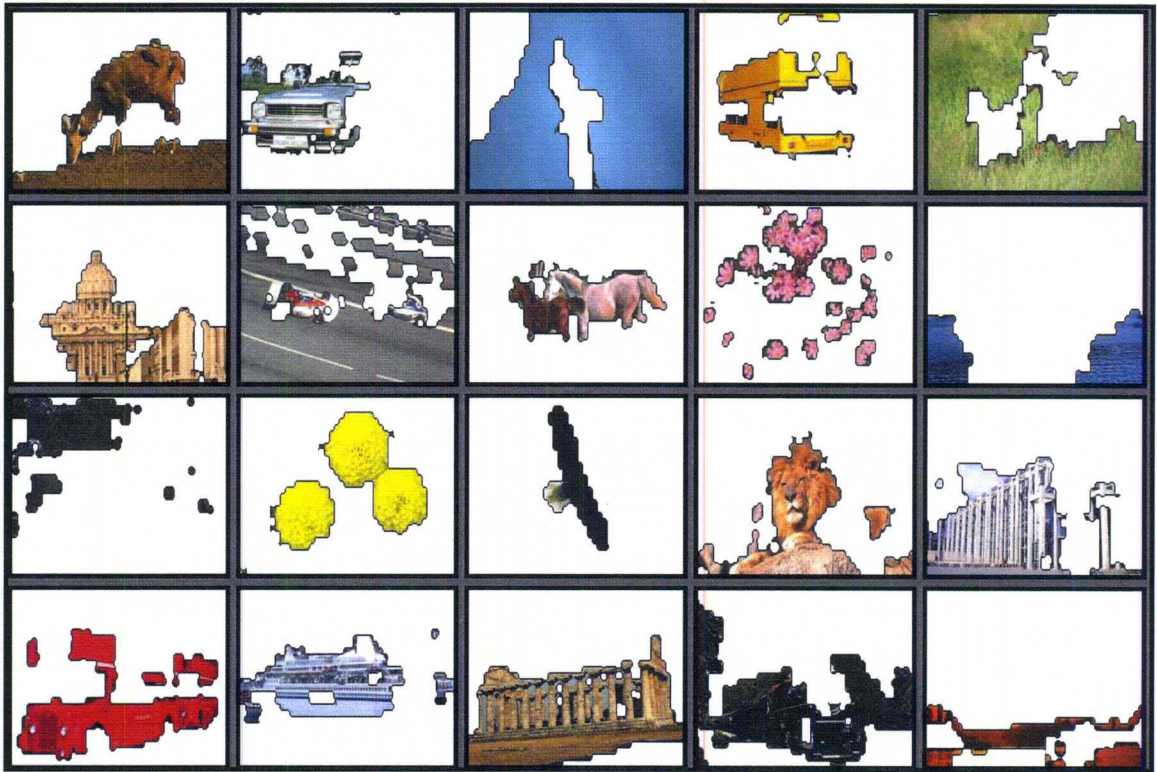
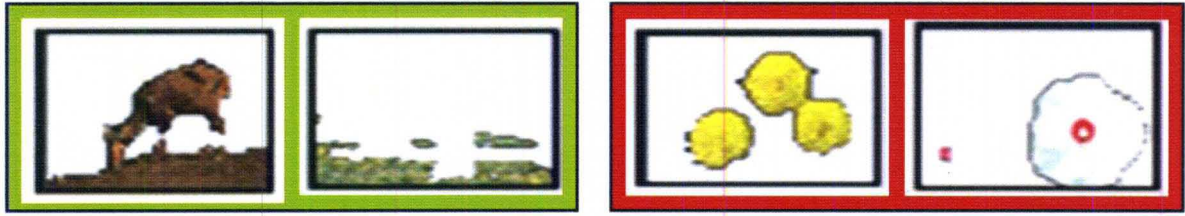


Figure 31. 20 samples from the 200 Category Representatives

the TBIA annotation algorithm (refer to chapter §IV), construct boolean compositions of keywords (using (86)), and retrieve images that satisfy the boolean expression in (87). In this case, the number of retrieved images reduces to 183, without losing many relevant images. In Figure 33, we compare the precision of querying with and without textual features for the most frequent categories among the retrieved images. As it can be seen, using the automatically generated region labels increases the precision of the relevant categories (e.g., cats, horses, deer, etc.) and reduces the precision of the irrelevant images (e.g., buildings, cars, foliage, etc.)

In a second experiment, we select 30 different queries and for each one we compute the precision when the number of retrieved images is varied from 10 to 70 by an increment of 10. For each query, we compare the results of four region-based image retrieval implementations; The original Fauqueur approach, visual features only using our fuzzy neighborhood, hybrid approach using visual and textual



(a) Positive Query Categories (b) Negative Query Categories

Figure 32. Positive and Negative query categories selected by the user to formulate a query. Here, the user is looking for images that contain horse/deer on grass but no flowers.

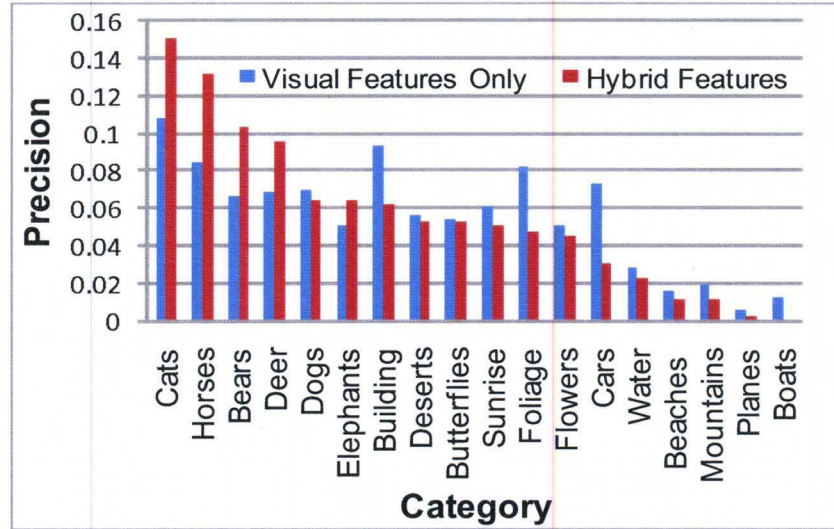


Figure 33. Comparison of the precision values for the query regions in Fig. 32 with and without textual features.

features, and the ranked hybrid using eqn. (88). The average precision values are shown in Figure 34. As it can be seen, each of our additions to the region-based CBIR improves the results significantly. The addition of the fuzzy neighborhood slightly improves the Fauqueur method, especially as more images are retrieved. This results from both methods not ranking their results, so the more images returned the greater the influence of the fuzzy association on the result set.

In the hybrid approach, as it can be seen, when the textual features are added to constrain the visual features, they do filter out many irrelevant images. The final addition of ranking the hybrid results improves the precision/recall dramatically, especially when fewer images are returned to the user.

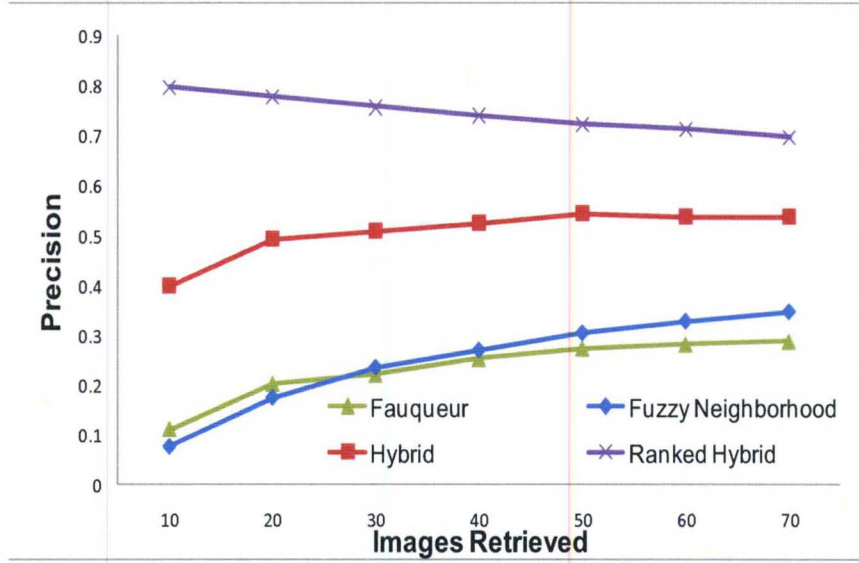


Figure 34. Comparison of the precision values when querying with and without textual features, and when using the fuzzy neighborhood and ranked hybrid methods. The values are averaged over 30 queries.

E Conclusions

In this chapter, we proposed an efficient region based image retrieval system that indexes and retrieves images using both visual and textual features. Our system segments all the images in the database and categorize their regions into groups of similar regions. To integrate high-level semantic features into the boolean composition of region categories, we use our thesaurus based image annotation algorithm to label image regions. The representative regions and their labels are then presented to the user who can formulate a query using a combination of positive and negative categories. This way, the user can formulate hybrid queries by selecting reference image regions and/or textual keywords that should (or should not) be included in the retrieved images. The keywords could also be implicitly selected as those used to label the reference regions. The search process is performed through the use of inverted tables of the region category labels, and thus, exhaustive search is not needed. The multi-modal features could be processed in parallel or

sequentially, where one modality could be used as a filter for the other modality.

Our approach builds on a previously developed system [148]. In the worst case scenario our region based image retrieval system will perform equal to this approach. The advantages of our additions are in removing the spherical assumption in determining neighborhoods, especially with high-dimensional data. Using the fuzzy membership approach allows the distribution of all regions within a category to affect the results, and allows multiple regions to belong to multiple categories.

The modification of the S_{NQ} allows for a more meaningful user query to be formed. In this instance, the user excludes regions similar to their NCQ's, however the Fauqueur constraint on the intersection of the NCQ's (effectively removing fewer images) is not present. Additionally textual information is added to support hybrid region based query. These results are then ranked in a novel way to display the closest images matching the user's desires without compromising system efficiency.

In this chapter, we have illustrated our approach when textual features are extracted from the labels assigned to the query regions and were used to constrain the visual features. We have shown that these additional features can filter out many irrelevant images and thus, improve the precision of the system. Other possible scenarios include allowing the user to specify keywords and using features from one modality to rank the images retrieved by another modality.

CHAPTER VII

SEMANTIC VISUALIZATION AND NAVIGATION

In the previous chapters we presented the to query-by-visual example at a global and region level for our CBIR system. However, there are cases where the user does not have a clear idea of what he or she is searching for. These users have no specific aim other than to find interesting things. For this reason, many CBIR systems incorporate a browsing and navigating interface. These systems organize and present the entire database of images to the user for browsing and navigation. Visual browsing techniques such as those in [86, 87] provide an overview of the database, but are practical only when the goal image is vague [5] and the domain of the image database is broad.

As image databases become increasingly large, it is not conceivable to browse the entire database at once; however, an overview of the entire database is desired. In this regard, the image library must create a set of images representative of the images located within. In order to summarize a database effectively, most visualization applications cluster solely based on the visual content of the images. This constrains the navigation, thereby introducing the semantic gap; "As long as the gap is there, use of the content-based retrieval for browsing will not be within the grasp of the general public as humans are accustomed to rely on the immediate semantic imprint the moment they see an image [5]."

Additionally, in order to effectively visualize high dimensional data, each point needs to be projected into a two or three-dimensional space. Techniques such

as Principal Component Analysis (PCA) [153], Singular Value Decomposition (SVD) [154], Kohonen Self Organizing Feature Maps (SOFM) [155], and Multi-Dimensional Scaling (MDS) [156] have been used for this purpose. However, these projection approaches are not easily scalable and require the entire data set to be loaded into memory.

In this chapter, we present the navigational component of our CBIR system that addresses the aforementioned issues. In addition to completing our CBIR system, we also address the existing browsing issues; Scalability and the semantic gap. We present the Graphical Text Interface (GTI) that visualizes high-dimensional multi-modal data for browsing and navigation in a two-dimensional platform. This approach provides a platform for dynamic updating that can account for both visual and semantic information. This component is highlighted in Figure 35. The clustering of the content is performed using the SOON algorithm [78] discussed below. The GTI can search online in realtime, actively adapt to different resolutions⁷, reorganize each axis independently, and perform centroid-free clustering to reduce the effect of the curse of dimensionality [157].

A Self Organization and Visual Exploration of Large Multi-Modal Data Sets

We have adopted the SOAVE algorithm (refer to section §II.C.5) to summarize and visualize our large multi-modal data collection. In particular, we have modified the following components:

1. Distance Measure: A common problem associated with most prototype-based clustering algorithms is that their performance degrades as the dimensionality of the data increases. For instance, using a centroid-based algorithm to cluster our data collection (in a n -dimensional space) may lead to poor

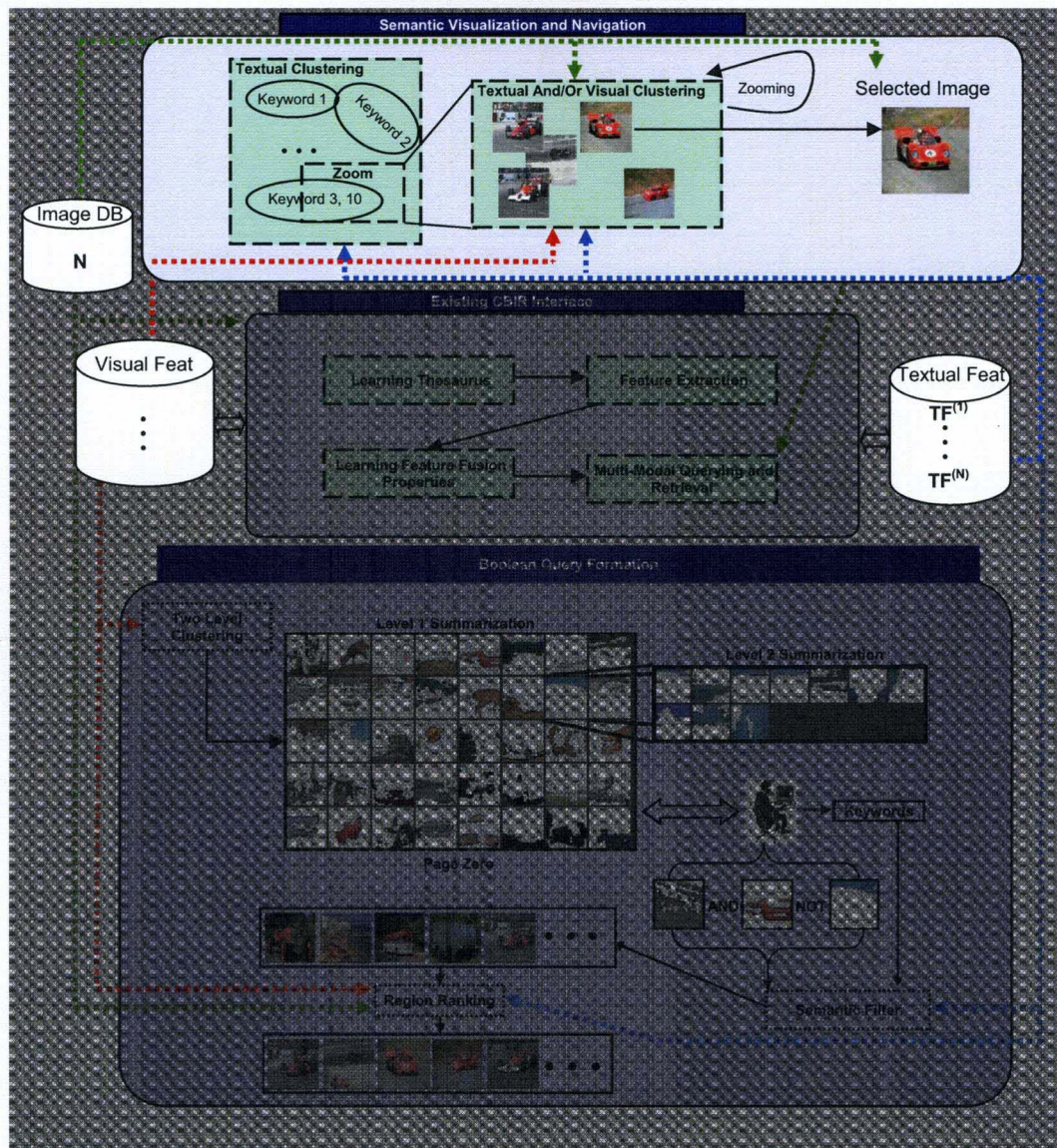


Figure 35. Overview of the proposed CBIR system component to visualize high-dimensional visual and textual data.

summarization and visualization. To overcome this limitation, we have modified SOAVE to use pairwise distances. In other words, when considering the assignment of a given point to a cluster, we compute its distance to all points within the cluster instead of the distance to the centroid of the cluster. This modification makes SOAVE more effective in clustering high dimensional data, but increases the computational and storage requirements. To maintain the scalability and efficiency of SOAVE, we have redesigned it to use linked

lists and various indexing schemes.

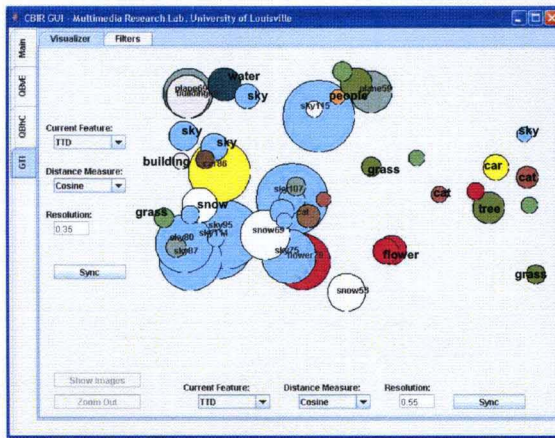
2. Mapping: To accommodate for the multi-modal features and semantic information in our data collection, we provide the user with options to map the data based on different distance measures, resolutions, or features. This is achieved by using SOON to cluster the data with the specified parameters and then using the phases to map the data to a one-dimensional space. This option allows each axis of the two-dimensional visualization space to function independently of the others, yielding semantic and visual comprehension of the image database. For instance, the horizontal axis may reflect the similarity with respect to the color feature while the vertical axis may reflect the semantic similarity using the keywords.
3. Zooming: To zoom into a specific region of clusters, the SOON algorithm is applied to each axis with a lower resolution while retaining the feature set and distance measures. This allows some of the clusters to desynchronize and break into smaller clusters. Thus, providing the user with a more detailed view.

B Graphical Text Interface

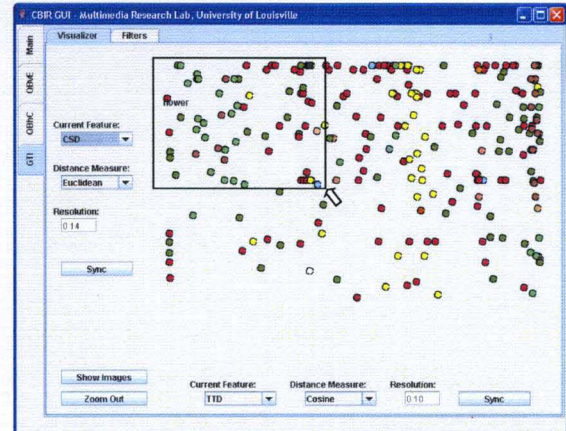
The Graphical Text Interface (GTI) is a two-dimensional map browser that visualizes high dimensional data using SOON at multiple resolutions. Using SOON allows each axis of the map to be dynamically clustered in real-time and computed independently of the other axis. This unique ability enables a user to keep one axis focused on textual information at a given resolution while continually redefining a different axis based on another feature set. Similar to SOAVE, the GTI provides the user with full control of each axis, giving the ability to select different features, distance measures, and clustering resolutions.

The GTI addresses the "*Page Zero*" issue (see chapter VI) in an intuitive way. This alternative visual browsing technique removes the visual requirement for formulating a query found in most CBIR systems. This is accomplished by presenting an overview of the database in a high-level conceptual form. At any point during browsing and navigation, the user may switch part or all of the interface to use visual features instead of textual features. This additional ability displays conceptually related images on a visual axis. Additionally, the system enables the user to view images based on the conceptual or visual representative of a cluster.

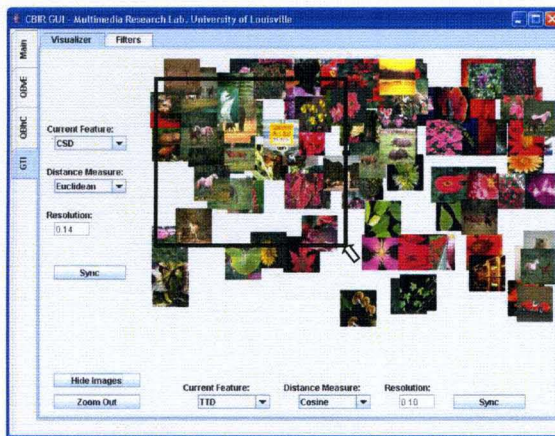
Figure 36 shows how the GTI facilitates browsing and navigation and how to select an image query if needed. In Figure 36(a), the initial interface is displayed providing an overview of the entire image collection. Each circle in the display represents one of the clusters, and these few clusters summarize the 10,000 images. The size of each circle reflects the relative size of the cluster and the color reflects the dominant color of the cluster representative image. The most frequent keyword is also displayed inside each circle to provide the user with semantic information. Within this view, the user can select a region of clusters to zoom into that area and visualize it at a lower resolution. In Figure 36(b), we display the map of the area selected in Figure 36(a). As it can be seen, this step has resulted in breaking the clusters into many smaller clusters. The dominant colors of all these tiny clusters are red, green, and yellow. Next we assume that the user zooms in further and selects the region highlighted in Figure 36(b). If the user is interested in only browsing the data collection, then he/she can stop after these steps, or he/she can select a different region to zoom into. If the user is interested in querying the database, then he/she can select the query example from the map as indicated in Figure 36(d). In this case, the system uses this example and switches to the query-by-visual-example mode.



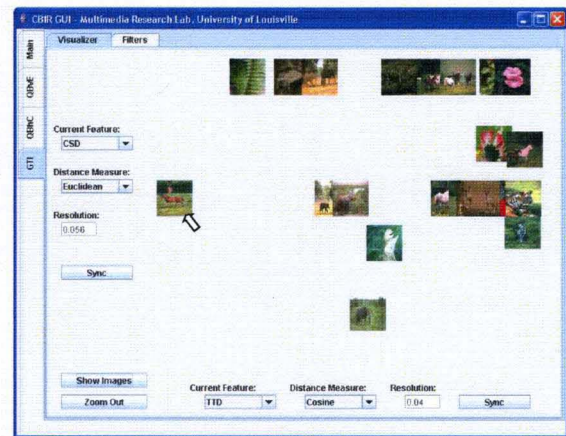
(a)



(b)



(c)



(d)

Figure 36. Graphical Text Interface: Walk through from view of semantic database to selection of query image. (a) Initial semantic view of the system. (b) Zoomed in region. (c) Zoomed in region from image view. (d) Selecting an image after zooming.

Another component of the GTI is the ability to filter out the results to be displayed based on semantic keywords. This browsing option is novel and is non-existent in current applications. It provides a unique platform for narrowing the semantic gap. Figure 37 shows how limiting the images to be visualized to those that include keywords "grass" and "flower" drastically reduces the number of clusters and images displayed. In Figure 37(a), the initial system is shown with all semantic concepts while in Figure 37(b), only clusters containing at least one of the filtered keywords within the top five cluster words is shown.

The final component of the GTI is the ability to re-organize each axis

C Conclusions

In this chapter, we presented the Graphical Text Interface (GTI) component of our CBIR system. This interface consists of a two-dimensional map browser for visualizing high-dimensional data. It is an adaptation of the SOAVE algorithm to perform summarization and projection of high-dimensional multi-modal data. It allows the user to visualize the data with respect to different feature sets, and zoom in and explore certain regions of the mapped space. The user could also filter the data to be visualized using a restricted set of keywords.

The GTI interface could be used to provide the user with an overview of the database. This approach provides a method of searching and navigating the database without requiring knowledge of the high-dimensional image content, while being able to dynamically adapt the visual feature space for individual needs. The GTI interface could also be used to overcome the "*page zero*" problem and guide the user in selecting an example image to initiate the query process. Once a query image is selected, our CBIR system switches to the query-by-visual-example mode.

CHAPTER VIII

CONCLUSIONS AND FUTURE WORK

We have presented various algorithms addressing the image annotation and multi-modal feature fusion tasks to narrow the semantic gap in Content-Based Image Retrieval. The developed algorithms combine topics from pattern recognition, data mining, image processing, and multimedia and were integrated in a complete CBIR system that has four main components. The first component uses a set of training images to learn a thesaurus. These images are manually annotated and used to create a multi-modal thesaurus through clustering and feature weighting. The objective is to extract representative visual profiles corresponding to frequent homogeneous regions and to associate these profiles with keywords. To accomplish this, the training images are segmented into homogeneous regions. Then, the regions are represented by visual descriptors, combined with the image level annotations, and clustered into categories of regions that share common attributes. Representatives of each cluster and its parameters are used to create profiles linking low-level image features and high-level concepts.

The second component of our CBIR system uses the developed multi-modal thesaurus to automatically annotate segmented regions. This was accomplished through two main steps. First, an unannotated image is segmented into homogeneous regions. Then, fuzzy membership functions are used to label new regions based on their proximity to the thesaurus entries. These annotated regions are then used to facilitate textual region based searches. We have showed that our

approach outperforms the state-of-the-art methods in its ability to determine accurate annotations especially for infrequent concepts. Thus, our approach is more reliable when the database is very large, and only few labeled samples are available.

The third component consists of an efficient and effective method for fusing the retrieval results of the multi-modal features. Our approach is based on mapping the distribution of distances for each feature to a fuzzy membership function and assigning a degree of worthiness to each feature based on its average performance. The memberships and the feature weights are then aggregated to produce a confidence value that is used to rank the retrieved images. Two aggregation methods were described and experimented with. The first is linear and is based on a simple weighted combination. The second one is non-linear and is based on the discrete Choquet integral. Both approaches are computationally efficient, requiring only simple multiplication and summation of the outputs of the individual features. The Choquet integral involves additional sorting of the individual outputs which is not significant if only few features are used. Thus, both methods could be used to fuse the results in a real-time mode. We have showed, using both subjective and objective experiments that our approach outperforms standard approaches that combine multiple features using distance scaling and ranking.

The fourth component of our CBIR uses the multi-modal thesaurus to perform hybrid querying and query expansion in the CBIR search process. In particular, the inter-modality correlation learned using the first component of the CBIR and represented in a multi-modal thesaurus is used to enrich and expand the visual query with textual data. We have showed that this query expansion can improve the accuracy of the retrieved images significantly. This is particularly true for images that are semantically similar but visually different.

The above four components were integrated and implemented into a complete

CBIR system that can run in three different modes. The first mode is a classic CBIR retrieval with all four components integrated. The user selects the query image from the database and can select any of the visual and/or textual features. The user has also the flexibility to modify the distances, the fusion methods, and to add semantic labels or use the labels assigned automatically by the second component of our system.

The second mode uses a novel region based approach that addresses the visual search when the user has a mental picture of what he is looking for but no sample image. This mode is an efficient region based image retrieval system. Our system first segments all the images in the database and categorizes their regions into groups of similar regions. Then, each region is labeled using our thesaurus based image annotation algorithm. The representative regions and their labels are then presented to the user who can formulate a query using a combination of positive and negative categories. Thus, the user can formulate hybrid queries by selecting reference image regions and/or textual keywords that should, or should not, be included in the retrieved images. Additionally the keywords could be implicitly selected as those used to label the reference regions. The search process is performed through the use of inverted tables of the region category labels; as such, exhaustive search is not necessary.

The final mode uses a novel Graphical Text Interface to perform semantic visualization and navigation, allowing for the initial navigation to be oriented around high-level concepts instead of randomly-selected images. This two-dimensional map browser visualizes high dimensional data using a clustering algorithm that can summarize the data at multiple resolutions. We use this algorithm to cluster and map the data dynamically using any of the features specified by the user. This unique ability lets the user keep one axis focused on

textual information at a given resolution, while continually refining a different axis using another feature set. In addition, the GTI provides the user with full control of each axis to modify the features, distance measures, clustering resolution, and various filtering options.

The presented CBIR system and its various components were validated using a large data set for accuracy, performance, and improvement over basic CBIR techniques. Our thesaurus based image annotation algorithm outperforms three state-of-the-art approaches on average by 13% when labeling 10,000 images. Our efficient method for fusing the output of multi-modal features yields 6% higher precision on average than standard CBIR methods and 16% better retrieval performance than the best individual feature. Lastly, our region-based retrieval is 30% better than a similar state-of-the-art approach.

The CBIR system is implemented as a java framework built on a C# server. The server application maintains all data, clustering, and distance calculations in local memory. Using this implementation approach, we average 0.83s on a query with a 55,000-image database using a 3.4Ghz Pentium IV with 4GB of RAM.

One current limitation of our CBIR system consists of the implementation method used to store the data. To achieve increased speed the data set and features are always in memory. This currently places an upper limit on the size of our database equal to the amount of RAM on the machine. A database of 55,000 images with 6 multi-modal feature sets requires 1GB of memory. As image features are added to provide more accurate retrieval the memory constraints increase. One solution to maintain time constraints of real-time querying would be to use multiple servers. Partitioning the database across platforms allows unlimited capacity while only slightly decreasing performance. For instance, one could have a color feature in one location and a texture feature in another. Sending the appropriate feature of a

query to a service running on that machine, the databases could return a subset of image results whose output could be fused. A similar method would allow splitting images of a database across servers, and not just their features.

Another limitation that faces our system is related to the scalability issue. First, we use a relatively small vocabulary size (100 words). In a more realistic scenario, a much larger vocabulary size may be needed. In this case, the vector space notation may not be appropriate, and thus, integrating the textual features into the clustering phase is not trivial. Second, the SCAD algorithm used to categorize the images and image regions is not scalable. That is it cannot handle a large data set that does not fit into memory. We are currently developing a scalable version of SCAD for large data that partitions the data, clusters the partitions, and then clusters the results. If this approach could produce similar output as SCAD, each partition could be clustered in parallel on separate machines or in separate threads and increase time performance.

Future research will expand on the contributions presented here, while investigating solutions to the possible limitations of the system. One such area is to examine exploiting the intra-modality correlations learned during the region clustering process. For instance, some colors such as the color of "planes" and the color of the "sky" may be correlated. These intra-modality correlations could be used to expand the query to include features not present in the query image but highly correlated.

Additionally, one can focus on enhancing the multi-modal thesaurus. The intra-modality correlations could not only be used to learn similar features, but to expose tightly coupled terms that should be investigated. For keywords that always appear together, no new knowledge is being gained and incorrect annotation/representation can take place. By adding images that contain each term

but not the other to the training set, the multi-modal thesaurus can begin distinguishing between these. We also will investigate adding more images for low frequency terms to not only allow them to be better distinguished, but decrease the penalty assigned to the higher frequency words through the ICF.

Finally, it is possible to integrate a relevance feedback component into our CBIR to further minimize the semantic gap. Relevance feedback has shown great results in focusing a users query, and if we could store/learn from this information we would strengthen our existing components. For instance, relevance feedback can be used to adapt the fusion parameters. Our current approach is trained globally using simple membership functions and a set of training images. Future enhancements could use the user's feedback to adjust the parameters of the membership functions and the degree of worthiness assigned to each feature. Similarly, the relevance feedback could be used to adjust the annotations of the images in the database. Storing the positive and negative information obtained from each query could not only strengthen existing keywords by modifying their values, but also highlight erroneous labels and add new terms to the database through hybrid querying. A modification to the fuzzy labeling process could then take this new knowledge into consideration as an expert equal to the profiles in the multi-modal thesaurus. Relevance feedback shows promise in expanding our annotation approach and providing a foundation for improving the accuracy of the system as a whole.

REFERENCES

- [1] Kodak, "Kodak easyshare gallery," <http://www.kodakgallery.com/>.
- [2] ImageShack Corp., "Imageshack," <http://www.imageshack.us/>.
- [3] Ludicorp, "flickr," <http://www.flickr.com/>.
- [4] Photobucket Inc., "Photobucket," <http://www.photobucket.com/>.
- [5] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Patt. Analysis Mach. Intell.*, vol. 22, no. 12, 2000.
- [6] A. D. Bimbo, *Visual Information Retrieval*, Morgan Kaufmann, 1999.
- [7] C. Chen, G. Gagaudakis, and P. Rosin, "Similarity-based image browsing," in *IEEE International Conference on Information Visualisation*, 2000.
- [8] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," in *Penn State University Technical Report CSE 06-009*, 2006.
- [9] W. Niblack, R. Barber, W. Equitz, M. Flickner, and al., "The qbic project: Querying images by content using color, texture, and shape," *Proc. SPIE (Storage and Retrieval for Image and Video Databases)*, vol. 1908, pp. 173–187, 1993.
- [10] J. R. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. C. Jain, and C.-F. Shu, "Virage image search engine: an open framework for image management," *Proceedings of SPIE*, vol. 2670, no. 76, 1996.
- [11] Seaborn, *Image Retrieval Systems*, M.A., 1997.
- [12] S. Marchand-Maillet and The Viper Team, "Viper," <http://viper.unige.ch/team/index.html>.
- [13] R. N. Cabral, "imgseek," <http://www.imgseek.net/>.
- [14] Q. Iqbal and J. K. Aggarwal, "Cires: A system for content-based retrieval in digital image libraries," in *Int. Conference on Control, Automation, Robotics and Vision*, 2002, pp. 205–210, CIRES CBIR System at <http://amazon.ece.utexas.edu/qasim/research.htm>.
- [15] Y. Rui, T. Huang, and S. Chang, "Image retrieval: current techniques, promising directions and open issues," *Journal of Visual Communication and Image Representation*, vol. 10, no. 4, 1999.

- [16] Amarnath Gupta and Ramesh Jain, "Visual information retrieval," *Communications of the ACM*, vol. 40, no. 5, pp. 70–79, 1997.
- [17] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: Image segmentation using expectationmaximization and its application to image querying," *IEEE Trans. Patt. Analysis Mach. Intell.*, vol. 24, no. 8, 2000.
- [18] Wei-Ying Ma and B. S. Manjunath, "Netra: A toolbox for navigating large image databases," *Multimedia Systems*, vol. 7, no. 3, pp. 184–198, 1999.
- [19] Rosalind W. Picard and Thomas P. Minka, "Vision texture for annotation," *Multimedia Systems*, vol. 3, no. 1, pp. 3–14, 1995.
- [20] Simone Santini and Ramesh Jain, "Similarity measures.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 9, 1999.
- [21] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vision*, vol. 40, no. 2, 2000.
- [22] D. P. Huttenlocher, G. A. Klanderman, and W. A. Rucklidge, "Comparing images using the hausdorff distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 9, 1993.
- [23] James Ze Wang, Jia Li, and Gio Wiederhold, "SIMPLIcity: Semantics-sensitive integrated matching for picture Libraries," *IEEE Trans. Patt. Analysis Mach. Intell.*, vol. 23, no. 9, pp. 947–963, 2001.
- [24] Baback Moghaddam, Qi Tian, Neal Lesh, Chia Shen, and Thomas S. Huang, "Visualization and user-modeling for browsing personal photo libraries," *Int. J. Comput. Vision*, vol. 56, no. 1-2, 2004.
- [25] S. Santini and R. Jain, "Integrated browsing and querying for image databases," *IEEE MultiMedia*, vol. 7, no. 3, pp. 26–39, 2000.
- [26] Bertrand Le Saux and Nozha Boujemaa, "Unsupervised categorization for image database overview," in *Visual Information and Information Systems*, 2002, pp. 163–174.
- [27] Yixin Chen, James Z. Wang, and Robert Krovetz, "Clue: Cluster-based retrieval of images by unsupervised learning," *IEEE Transactions on Image Processing*, vol. 14, no. 8, 2005.
- [28] J. Li and J. Z. Wang, "Automatic linguistic indexing of pictures by a statistical modeling approach," *IEEE Trans. Patt. Analysis Mach. Intell.*, vol. 25, no. 10, 2003.
- [29] Hichem Frigui, "Unsupervised learning of arbitrarily shaped clusters using ensembles of gaussian models," *Pattern Anal. Appl.*, vol. 8, no. 1-2, 2005.
- [30] G. Sheikholeslami, W. Chang, and A. Zhang, "Semquery: Semantic clustering and querying on heterogeneous features for visual data," *IEEE Trans. on Knowledge and Data Engineering*, vol. 14, no. 5, pp. 988–1002, 2002.

- [31] Yong Rui, Thomas S. Huang, and Sharad Mehrotra, "Relevance feedback: A power tool in interactive content-based image retrieval," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 8, no. 5, 1998.
- [32] Ye Lu, Chunhui Hu, Xingquan Zhu, HongJiang Zhang, and Qiang Yang, "A unified framework for semantics and feature based relevance feedback in image retrieval systems," in *ACM Multimedia*, 2000, pp. 31–37.
- [33] J. S. Hare, P. H. Lewis, P. G. B. Enser, and C. J. Sandom, "Mind the gap: Another look at the problem of the semantic gap in image retrieval," in *Proceedings of SPIE Multimedia Content Analysis, Management and Retrieval 2006*, 2006.
- [34] J. S. Hare, P. A. S. Sinclair, P. H. Lewis, K. Martinez, P. G. B. Enser, and C. J. Sandom, "Bridging the semantic gap in multimedia information retrieval: Top-down and bottom-up approaches," in *Proceedings of Mastering the Gap: From Information Extraction to Semantic Representation / 3rd European Semantic Web Conference*, 2006.
- [35] S. Sclaroff, M. Cascia, and S. Sethi, "Unifying textual and visual cues for content-based image retrieval on the world wide web," *Computer Vision and Image Understanding*, vol. 75, no. 1/2, pp. 86–98, 1999.
- [36] X. S. Zhou and T. S. Huang, "Unifying keywords and visual contents in image retrieval," *IEEE Multimedia*, vol. 9, no. 2, pp. 23–33, 2002.
- [37] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M. Jordan, "Matching words and pictures," *Journal of Machine Learning Research*, vol. 3, pp. 1107–1135, 2003.
- [38] P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth, "Object recognition as machine translation : Learning a lexicon for a fixed image vocabulary," in *Seventh International Conference on Computer Vision (ECCV)*, 2002, pp. 97–112.
- [39] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *ACM SIGIR*, 2003.
- [40] H. Frigui and J. Caudill, "Building a multi-modal thesaurus from annotated images," in *Proc. of the 18th Int. Conf. on Pattern Recognition (ICPR'06)*, 2006, vol. 4.
- [41] Y. Mori, H. Takahashi, and R. Oka, "Image to word transformation based on dividing and vector quantizing images with words," in *First Int. Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999.
- [42] W. Liu and X. Tang, "Learning an image-word embedding for image auto-annotation on the nonlinear latent space," in *Proceedings of ACM Multimedia*, 2005.
- [43] F. Monay and D. Gatica-Perez, "On image auto-annotation with latent space models," in *Proceedings of ACM Multimedia*, 2003.

- [44] Aksoy S. and Haralick R.M., "Feature normalization and likelihood-based similarity measures for image retrieval," *Pattern Recognition Letters*, vol. 22, pp. 563–582, 2001.
- [45] Y. Chen, J. Z. Wang, and R. Krovetz, "Clue: Cluster-based retrieval of images by unsupervised learning," *IEEE Trans. on Image Processing*, vol. 14, no. 8, pp. 1187–1201, 2005.
- [46] S. Santini, A. Gupta, and R. Jain, "Emergent semantics through interaction in image databases," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 3, pp. 337–351, 2001.
- [47] S. Santini and R. Jain, "Interfaces for emergent semantics in multimedia databases," in *Proceedings of SPIE Vol. 3656 Storage and Retrieval for Image and Video Databases VII*, 1999.
- [48] S.-F. Cheng, W. Chen, and H. Sundaram, "Semantic visual templates: Linking visual features to semantics," in *IEEE Int. Conference on Image Processing*, 1998, pp. 531–535.
- [49] R. Datta, W. Ge, J. Li, and J. Wang, "Toward bridging the annotation-retrieval gap in image search by a generative modeling approach," in *Proceedings of the ACM MM'06*, 2006.
- [50] J. Li, J. Z. Wang, and H. Wiederhold, "Irm: Integrated region matching for image retrieval," in *Proceedings of the Eighth ACM Int. Conference on Multimedia*, 2000, pp. 147–156.
- [51] J. Z. Wang and Y. Du, "Scalable integrated region-based image retrieval using irm and statistical clustering," in *Proceedings of the 1st ACM/IEEE-CS Joint Int. Conference on Digital Libraries*, 2001, pp. 268–277.
- [52] Y. Chen and J. Wang, "A region-based fuzzy feature matching approach to content-based image retrieval," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, pp. 1252–1267, 2002.
- [53] T. Hirayama, Y. Iwai, and M. Yachida, "Facial recognition system using efficient methods for facial scale variations," *IEEE SICE 2003 Annyal Conference*, vol. 3, no. 4–6, pp. 3236–3241, 2003.
- [54] B. S. Manjunath, P. Salembier, and T. Sikora, *Introduction to MPEG 7: Multimedia Content Description Language*, John Wiley, 2002.
- [55] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Addison Wesley Publishing Company, 1993.
- [56] R. Gonzalez, R. Woods, and S. Eddins, *Digital Image Processing*, Prentice Hall, 2004.
- [57] H. Tamura, S. Mori, and T. Yamawaki, "Texture features corresponding to visual perception," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-8, no. 6, pp. 460–473, 1978.

- [58] V. Castelli and L. D. Bergman (Eds.), *Image Databases: Search and Retrieval of Digital Imagery*, Wiley: New York, 2002.
- [59] F. Liu and R. W. Picard, "Periodicity, directionality, and randomness: Wold features for image modeling and retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 7, pp. 722–733, 1996.
- [60] Y. Rui, A. She, and T. Huang, "Modified fourier descriptors for shape representation - a practical approach," in *First International Workshop on Image Databases and Multi Media Search*, 1996, pp. 22–23.
- [61] L. Yang and F. Albregtsen, "Fast computation of invariant geometric moments: A new method giving correct results.," in *International Conference on Pattern Recognition (ICPR)*, 1994, pp. 201–204.
- [62] J. Hafner, H. Sawhney, W. Flickner, and W. Niblack, "Efficient color histogram indexing for quadratic form distance functions," *IEEE Trans. Patt. Analysis Mach. Intell.*, vol. 17, pp. 729–736, 1995.
- [63] G. Salton, *The SMART Retrieval System - Experiments in Automatic Document Processing*, Prentice-Hall, 1971.
- [64] K. Kukich, "A comparison of some novel and traditional lexical distance metrics for spelling correction," in *Proceedings of INNC-90-Paris*, 1990, pp. 309–313.
- [65] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York, 1981.
- [66] J. C. Bezdek, R. Ehrlich, and W. Full, "Fcm: The fuzzy c-means clustering algorithm," *Computers and Geosciences*, vol. 10, no. 2–3, pp. 191–203, 1984.
- [67] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, Prentice Hall, 1988.
- [68] R. Krishnapuram and C. P. Freg, "Fitting an unknown number of lines and planes image data through compatible cluster merging," *Pattern Recognition*, vol. 25, no. 4, pp. 385–400, 1992.
- [69] R. N. Davés and K. J. Patel, "Progressive fuzzy clustering algorithms characteristic shape recognition," in *North Amer. Fuzzy Information Processing Soc.*, 1990, pp. 121–124.
- [70] R. Krishnapuram, H. Frigui, and O. Nasraoui, "Fuzzy and possibilistic shell clustering algorithms and their application to boundary detection and surface approximation," *IEEE Trans. Fuzzy Systems*, vol. 3, no. 1, pp. 29–60, 1995.
- [71] H. Frigui and R. Krishnapuram, "Clustering by competitive agglomeration," *Pattern Recognition*, vol. 30, no. 7, pp. 1223–1232, 1997.
- [72] H. Almuallim and T. G. Dietterich, "Learning with many irrelevant features," in *Ninth National Conference on Artificial Intelligence*, 1991, pp. 547–552.

- [73] K. Kira and L. A. Rendell, "The feature selection problem: Traditional methods and a new algorithm," in *Tenth National Conference on Artificial Intelligence*, 1992, pp. 129–134.
- [74] L. A. Rendell and K. Kira, "A practical approach to feature selection," in *International Conference on Machine Learning*, 1992, pp. 249–256.
- [75] H. Frigui and O. Nasraoui, "Simultaneous clustering and attribute discrimination," in *Proceedings of the IEEE International Conference on Fuzzy Systems*, 2000, pp. 158–163.
- [76] H. Frigui and O. Nasraoui, "Unsupervised learning of prototypes and attribute weights," *Pattern Recognition Journal*, vol. 37, no. 3, pp. 567–581, 2004.
- [77] H. Frigui and S. Salem, "Fuzzy clustering and subset feature weighting," in *IEEE International Conference on Fuzzy Systems*, 2003.
- [78] M. B. H. Rhouma and H. Frigui, "Self-organization of pulse-coupled oscillators with application to clustering," *Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 180–195, 2001.
- [79] H. Frigui and K. Jampana, "Self organization and visual exploration of large high dimensional data sets," in *Self Organization and Visual Exploration of Large High Dimensional Data sets, held in conjunction with the SIAM Int. Conf. on Data Mining (SDM 2004)*, 2004, pp. 5–16.
- [80] J. R. Smith and S. F. Chang, "Quad-tree segmentation for texture-based image query," in *Proc. ACM Multimedia '94 Conf.*, 1994.
- [81] J. R. Smith and S.-F. Chang, "Visualeek: A fully automated content-based image query system," in *ACM Multimedia*, 1996, pp. 87–97.
- [82] P. Kelly and M. Cannon, "Query by image example: The candid approach," *Proc. SPIE*, vol. 2420, pp. 238–248, 1995.
- [83] S. F. Chang, J. R. Smith, H. J. Meng, J. Wang, and D. Zhong, "Finding images/video in large archives," in *D-Lib Mag.*, 1997.
- [84] S. Mehrotra, Y. Rui, M. Ortega, and T. S. Huang, "Supporting content-based queries over image in mars," in *Proc. 4th IEEE Int. Conf. Multimedia Computing and Systems*, 1997, pp. 623–633.
- [85] M. La Cascia, S. Sethi, and S. Sclaroff, "Combining textual and visual cues for content-based image retrieval on the world wide web," in *IEEE Workshop on Content-based Access of Image and Video Libraries (CBAIVL)*, 1998.
- [86] A. Hiroike, Y. Musha, A. Sugimoto, and Y. Mori, "Visualization of information spaces to retrieve and browse image data," in *Int. Conf. on Visual Information System (VIS)*, 1999.
- [87] J. Laaksonen, E. Oja, M. Koskela, and S. Brandt, "Analyzing low-level visual features using content-based image retrieval," in *Int. Conf. on Neural Information Processing (ICONIP)*, 2000.

- [88] N. Boujemaa, J. Fauqueur, M. Ferecatu, F. Fleuret, V. Gouet, B. Le Saux, and H. Sahbi, "Ikonfa: Interactive generic and specific image retrieval.," in *International workshop on Multimedia Content-Based Indexing and Retrieval*, 2001.
- [89] S. Guha, R. Rastogi, and K. Shim, "Cure: An efficient clustering algorithm for large databases," in *Proceedings of the ACM SIGMOD Conference*, 1998.
- [90] A. Hinneburg and D. Keim, "An efficient approach to clustering in large multimedia databases with noise.," in *Proc. of the 4th Int. Conf. on Knowledge Discovery and Data Mining (KDD'98)*, 1998, pp. 58–65.
- [91] S. Medasani and R. Krishnapuram, "Categorization of image databases for efficient retrieval using robust mixture decomposition," in *Proc. of the IEEE Workshop on Content-Bases Access of Image and Video Libraries*, 1998, pp. 50–54.
- [92] A.K.H. Tung, J. Han, L.V.S. Lakshmanan, and R.T. Ng, "Constraint-based clustering in large databases.," in *Proc. of the Int. Conf. on Database Theory (ICDT'01)*, 2001, pp. 405–419.
- [93] L. Saux, H. Bertand, and N. Boujemaa, "Image database clustering with svm-based class personalization," *Storage and Retrieval Methods and Applications for Multimedia, Proc. of the SPIE*, vol. 5307, pp. 9–19, 2003.
- [94] N. Grira, M. Crucinau, and N. Boujemaa, "Active semi-supervised fuzzy clustering for image database categorization," in *Proc. of the 7th ACM SIGMM Int. Workshop on Multimedia Information Retrieval*, 2005, pp. 9–16.
- [95] H. Frigui, C. Hwang, and F.C-H. Rhee, "Clustering and aggregation of relational data with application to image database categorization," *Pattern Recognition*, vol. 40, pp. 3053–3068, 2007.
- [96] K. Wagstaff and C. Cardie, "Clustering with instance-level constraints.," in *Proc. of the 17th Int. Conf. on Machine Learning*, 2000, pp. 1103–1110.
- [97] S. Basu, M. Bilenko, and R.J. Mooney, "Comparing and unifying search-based and similarity-based approaches to semi-supervised clustering.," in *Proc. of the Workshop on Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining Systems (ICML'03)*, 2003, pp. 42–49.
- [98] D. Cohn, R. Caruana, and A. McCallum, "Semi-supervised clustering with user feedback," 2003.
- [99] S. Basu, A. Banerjee, and R.J. Mooney, "Active semi-supervision for pairwise constrained clustering.," in *Proc. of the SIAM Int. Conf. on Data Mining (SDM-2004)*, 2004, pp. 333–344.
- [100] D. Klein, S.D. Kamvar, and C.D. Manning, "From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering.," in *Proc. of the 19th Int. Conf. on Machine Learning (ICML'02)*, 2002, pp. 307–314.

- [101] R. Ge, M. Ester, W. Jin, and I. Davidson, "Constraint-driven clustering,," in *Proc. of the 13th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining (KDD'07)*, 2007, pp. 320–329.
- [102] H. Frigui and J. Meredith, "Image database categorization under spatial constraints using adaptive constrained clustering," in *Proc. of the IEEE World Congress on Computational Intelligence*, 2008, pp. 2268–2276.
- [103] Y. Liu, R. Jin, and A.K. Jain, "Boostcluster: Boosting clustering by pairwise constraints," in *Proc. of the 13th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 2007, pp. 450–459.
- [104] H-J Zhang, Z. Chen an W-Y Liu, and M. Li, "Relevance feedback in content-based image retrieval," in *Invited Keynote, 12th Int. Conf. on New Information Technology*, 2002, pp. 1–8.
- [105] A. Kushki, P. Androutsos, K. N. Plantaniotis, and A. N. Venetsanopoulos, "Query feedback for interactive image retrieval," *IEEE Transactions On Circuits and Systems for Video Technology*, vol. 14, no. 5, pp. 644–655, 2004.
- [106] J. Urban, J. M. Jose, and C.J. van Rijsbergen, "An adaptive approach towards content-based image retrieval," in *Proc. of the Third International Workshop on Content-Based Multimedia Indexing*, 2003, pp. 119–126.
- [107] M.L. Kherfi and D. Ziou, "Relevance feedback for cbir: A new approach based on probabilistic feature weighting with positive and negative examples,," *IEEE Transactions On Image Processing*, vol. 15, no. 4, pp. 1017–1030, 2006.
- [108] Z.-H. Zhou, K.-J. Chen, and H.-B. Dai, "Enhancing relevance feedback in image retrieval using unlabeled data," *ACM Transactions on Information Systems (TOIS)*, vol. 24, no. 2, pp. 219–244, 2006.
- [109] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal Processing Magazine*, vol. 13, no. 6, pp. 47–60, 1996.
- [110] G. Miller, "Wordnet: A lexical database for english," *Comm. of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.
- [111] Y. Jin, L. Khan, L. Wang, and M. Awad, "Image annotations by combining multiple evidence & wordnet," in *ACM Multimedia*, 2005, pp. 706–715.
- [112] G. Shafer, *A Mathematical Theroy of Evidence*, Princeton University Press, 1976.
- [113] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1995.
- [114] J. Jiang and D. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," in *Proceedings of International Conference on Research in Computational Linguistics*, 1997.

- [115] D. Lin, “Using syntatic dependency as a local context to resolve word sense ambiguity,” in *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, 1997, pp. 64–71.
- [116] C. Leacock, “Combining local context and wordnet similarity for word sense identification,” in *WordNet: A Lexical Reference System and its Application*, 1998, pp. 265–283.
- [117] S. Banerjee and T. Pedersen, “Extended gloss overlaps as a measure of semantic relatedness,” in *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, 2003, pp. 805–810.
- [118] V. Lavrenko, R. Manmatha, and J. Jeon, “A model for learning the semantics of pictures,” in *Advances in Neural Information Processing Systems*, 2004.
- [119] R. Jin, J. Y. Chai, and L. Si, “Effective automatic image annotation via a coherent language model and active learning,” in *ACM Multimedia*, 2004, pp. 892–899.
- [120] J. Li and J. Z. Wang, “Automatic linguistic indexing of pictures by a statistical modeling approach,” *IEEE Trans. Patt. Analysis Mach. Intell.*, vol. 25, no. 9, pp. 1075–1088, 2003.
- [121] J. Fan, Y. Gao, and H. Luo, “Multi-level annotation of natural scenes using dominant image components and semantic concepts,” in *ACM Multimedia*, 2004, pp. 540–547.
- [122] J. Fan, Y. Gao, H. Luo, and G. Xu, “Automatic image annotation by using concept-sensitive salient objects for image content representation,” in *ACM SIGIR*, 2004, pp. 361–368.
- [123] D. M. Blei and M. I. Jordan, “Modeling annotated data,” in *Proceedings of ACM SIGIR*, 2003.
- [124] L. Wang, L. Khan, L. Liu, and W. Wu, “Automatic image annotation and retrieval using weighted feature selection,” in *ISMSE Conf.*, 2004.
- [125] J.-Y. Pan, H.-J. Yang, P. Duygulu, and C. Faloutsos, “Automatic image captioning,” in *KDD Conf.*, 2004.
- [126] J. A. Hartigan and M. A. Wong, “Algorithm as 136: A k-means clustering algorithm,” *Applied Statistics*, vol. 28, no. 1, pp. 100–108, 1979.
- [127] G. Hamerly and C. Elkan, “Learning the k in k-means,” in *Proceedings of the NIPS*, 2003.
- [128] G. H. Golub and C. Reinsch, “Singular value decomposition and least squares solutions,” *Numerische Mathematik*, vol. 14, no. 5, pp. 403–420, 1970.
- [129] J.-Y. Pan, H.-J. Yang, C. Faloutsos, and P. Duygulu, “Automatic multimedia cross-modal correlation discovery,” in *ICME Conf.*, 2004.

- [130] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society of Information Science*, vol. 41, pp. 391–407, 1990.
- [131] T. K. Landauer and M. L. Littman, "Fully automatic cross-language document retrieval using latent semantic indexing," in *Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research*, 1990, pp. 31–38.
- [132] J. S. Hare, P. H. Lewis, P. G. B. Enser, and C. J. Sandom, "A linear-algebraic technique with an application in semantic image retrieval," in *Proceedings of the Fifth Int. Conference on Image and Video Retrieval*, 2006, pp. 31–40.
- [133] K. Yu, S. Yu, and V. Tresp, "Multi-label informed latent semantic indexing," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2005, pp. 258–265.
- [134] L. Wang, L. Khan, L. Liu, and W. Wu, "Automatic image annotation and retrieval using weighted feature selection," in *Proceedings of the IEEE Sixth Int. Symposium on Multimedia Software Engineering (ISMSE'04)*, 2004.
- [135] D. Stan and I. K. Sethi, "Mapping low-level image features to semantic concepts," in *Proceedings of SPIE Conference: Storage and Retrieval for Media Databases*, 2001, pp. 172–179.
- [136] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proceedings of the VLDB*, Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, Eds., 1994, pp. 487–499.
- [137] C. Ordonez and E. Omiecinski, "Discovering association rules based on image content," in *Proceedings of the IEEE Advances in Digital Libraries*, Baltimore, 1999, pp. 38–49.
- [138] O. R. Zaïane, J. Han, Z. Li, S. H. Chee, and J. Chiang, "Multimediaminer: A system prototype for multimedia data mining," in *SIGMOD Conf.*, 1998, pp. 581–583.
- [139] H. Malik and J. Kender, "Clustering web images using association rules, interestingness measures, and hypergraph partitions," in *Proceedings of ACM ICWE'06*, 2006, pp. 48–55.
- [140] Y. Chen and J. Wang, "Image categorization by learning and reasoning with regions," in *Journal of Machine Learning Research* 5, 2004, pp. 913–939.
- [141] D. Wettschereck, D. W. Aha, and T. Mohri, "A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms," *Artificial Intelligence Review*, vol. 11, pp. 273–314, 1997.
- [142] G. Salton and M.J. McGill, *An Introduction to Modern Information Retrieval*, McGraw-Hill, 1983.

- [143] Shi Rui, Wanjun Jin, and Tat-Seng Chua, "A novel approach to auto image annotation based on pairwise constrained clustering and semi-naive bayesian model," in *11th International Multimedia Modelling Conference (MMM'05)*, 2003.
- [144] M. Grabisch, "Modelling data by the Choquet integral," in *Information Fusion in Data Mining*, V. Torra, Ed., pp. 135–148. Physica Verlag, Heidelberg, 2003.
- [145] P. D. Gader, B. Nelson, A. Hocaoglu, S. Auephanwiriyakul, and M. Khabou, "Neural versus heuristic development of Choquet fuzzy integral fusion algorithms for land mine detection," in *Neuro-fuzzy Pattern Recognition*, H. Bunke and A. Kandel, Eds., pp. 205–226. World Scientific Publ. Co., 2000.
- [146] H. Tahani and J.M. Keller, "Information fusion in computer vision uusing the fuzzy integral," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 20, no. 3, pp. 733–741, 1990.
- [147] J. Fauqueur and N. Boujemaa, "Mental image search by boolean composition of region categories," *Multimedia Tools and Applications*, vol. 31, no. 1, pp. 95–117, 2006.
- [148] J. Fauqueur and N. Boujemaa, "Logical query composition from local visual feature thesaurus," in *ACM Multimedia*, 2003.
- [149] N. Rasiwasia, P. J. Moreno, and N. Vasconcelos, "Bridging the gap: Query by semantic example," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 923–938, 2007.
- [150] H. Frigui and J. Caudill, "Region based image annotation," in *Int. Conference on Image Processing (ICIP'06)*, 2006.
- [151] H Frigui and J Caudill, "Mining visual and textual data for constructing a multi-modal thesaurus," in *Proc. of the SIAM Int. Conf. on Data Mining*, 2007.
- [152] H. Frigui, J. Caudill, and A. C. B. Abdallah, "Fusion of multi-modal features for efficient content-based image retrieval," in *WCCI2008*, 2008.
- [153] I.T. Jolliffe, *Principal Component Analysis*, Springer Verlag, 1986.
- [154] Gilbert Strang, *Linear Algebra and its application*, Academic Press, 1980.
- [155] T. Kohonen, *Self-Organizing Maps 3rd. ed.*, Springer, 2001.
- [156] J. B. Kruskal and M. Wish, *Multidimensional Scaling*, SAGE Publications, 1978.
- [157] A. Hinneburg and D. Keim, "Optimal grid-clustering: Towards breaking the curse of dimensionality," in *Proceedings of the Very Large Data Base Conference*, 1999.

CURRICULUM VITAE

- NAME: Joshua D. Caudill
- ADDRESS: Department of Computer Engineering and Computer Science
University of Louisville
Louisville, KY 40292
- EDUCATION: Pursuing Ph.D., Computer Science and Engineering,
May 2009
University of Louisville, Louisville, Kentucky
- M.Eng., Computer Engineering and Computer Science
with Highest Honors, May 2005
University of Louisville, Louisville, Kentucky
- B.S., Computer Engineering and Computer Science
with Highest Honors, December 2004
University of Louisville, Louisville, Kentucky
- PUBLICATIONS: 1. **J. Caudill** and H. Frigui, "Image Retrieval by Hybrid
Boolean Composition of Region Categories,"
International Conference on Image Processing (ICIP)
Submitted Jan 09.
2. **J. Caudill** and H. Frigui, "Automatic Image Annotation
and Fusion of Multi-Modal features for Semantic
Image Retrieval", Springer, Submitted Nov 08.
3. M. M. Ben Ismail, H. Frigui, and **J. Caudill**,
"Empirical Comparison of Automatic Image Annotation
Systems", International Workshops on Image
Processing Theory, Tools and Applications (IPTA 2008),
Sousse, Tunisia, Nov 2008.
4. H. Frigui, **J. Caudill**, and A. C. Ben Abdellah, "Fusion
of Multi-Modal Features for Efficient Content-Based
Image Retrieval", IEEE World Congress on
Computational Intelligence (WCCI 2008), Hong Kong,
June 2008.

5. H. Frigui and **J. Caudill**, "Mining Visual and Textual Data for Constructing a Multi-Modal Thesaurus", SIAM International conference on Data Mining, Minnesota, April 2007.
6. H. Frigui and **J. Caudill**, "Region Based Image Annotation", International Conference on Image Processing (ICIP 2006): 953-956, October 2006, Atlanta.
7. H. Frigui and **J. Caudill**, "Building a Multi-Modal Thesaurus from Annotated Images", International Conference on Pattern Recognition 2006 (ICPR 2006): 198-201, Hong Kong, August 2006.
8. H. Frigui and **J. Caudill**, "Unsupervised Image Segmentation and Annotation for Content-Based Image Retrieval", IEEE International Conference on Fuzzy Systems, Vancouver, BC, July 2006.

HONORS AND AWARDS:

- CECS Departmental Alumni Recipient (2005)
- Trustees Academic Scholarship (Full Tuition 2000-2005)
- Deans Scholar and Deans List (2000-2005)
- Recipient of KEES Scholarship Award (2000-2004)
- Cash Award USACE for Outstanding Performance (2004)
- Boy Scouts of America Eagle Scout (2000)
- Valedictorian of High School graduating class (2000)

- MEMBERSHIPS:
- Association for Computing Machinery (2006-present)
 - United States Fencing Association (2005-present)
 - Institute of Electrical and Electronics Engineers (2003-present)
 - Tau Beta Pi Engineering Honor Society (2002-present)
 - Golden Key International Honour Society (2001-present)
 - Honorable Order of Kentucky Colonels (2000-present)