MULTIPLE OBJECT RETRIEVAL IN IMAGE DATABASES USING
HIERARCHICAL SEGMENTATION TREE

by

WEI-BANG CHEN

CHENGCUI ZHANG, COMMITTEE CHAIR
OLIVIA AFFUSO
BARRETT R. BRYANT
ALAN P. SPRAGUE
CHIAO-WANG SUN

A DISSERTATION
Submitted to the graduate faculty of The University of Alabama at Birmingham,
in partial fulfillment of the requirement for the degree of
Doctor of Philosophy

BIRMINGHAM, AL

2012
MULTIPLE OBJECT RETRIEVAL IN IMAGE DATABASES USING HIERARCHICAL SEGMENTATION TREE

WEI-BANG CHEN

COMPUTER AND INFORMATION SCIENCES

ABSTRACT

The purpose of this research is to develop a new visual information analysis, representation, and retrieval framework for automatic discovery of salient objects of user’s interest in large-scale image databases. In particular, this dissertation describes a content-based image retrieval framework which supports multiple-object retrieval. The proposed research, unlike most existing works that are designed for single object retrieval or adopt heuristic multiple object matching scheme, aims at contributing to this field through the development of an image retrieval system that enables effective and efficient multiple-object retrieval and automatic discovery of the objects of interest through users’ relevance feedback. The key to achieving the above goal is a new systematic and hierarchical image representation, and a related learning and retrieval framework, the combination of which makes it possible for a machine to interpret an image in terms of its containing regions and their relationships. In this dissertation, an efficient and accurate hierarchical image segmentation algorithm based on multi-resolution analysis is developed to alleviate the over- and/or under-segmentation problems through the preservation of associative relationships between image regions in a hierarchical region-tree. This algorithm is designed in a way to simultaneously produce image segmentation results and hierarchical region-tree representations, which are typically obtained through two separate processes in existing approaches, so as to reduce the time complexity. With hierarchical region-tree representations, the relevance of an image to the query image is
thus measured according to the sub-tree comparison. As a full comparison of all sub-trees is unlikely to be feasible, an efficient strategy for selecting and comparing proper sub-trees is designed and developed. Another key contribution of this research is the seamless integration of users’ relevance feedback (RF) with the proposed multiple object retrieval system, which allows automatic discovery of the objects of users’ interest and is expected to improve the retrieval accuracy through feedback-retrieval loops. While there is a clear indication of needs for such interactive learning capabilities, we believe this is the first systematic attempt to formulate a comprehensive, intelligent, and interactive framework for multiple object retrieval in image databases that takes full advantage of a hierarchical region-tree representation.

Keywords: content-based image retrieval, multi-resolution image segmentation, hierarchical segmentation tree, multiple object retrieval, relevance feedback
DEDICATION

I dedicate this dissertation to my family
for their love, support and sacrifice.
ACKNOWLEDGEMENTS

It would not have been possible to write this dissertation without the help and support of the kind people around me, to only some of whom it is possible to give particular mention here.

First and foremost, I want to express my utmost gratitude to Dr. Chengcui Zhang for her guidance, inspiration, support and patience in the past seven years. Her mentorship was paramount in providing a well-rounded experience consistent with my long-term career goals. I would like to gratefully and sincerely thank my dissertation committee members, Dr. Olivia Affuso, Dr. Barrett R. Bryant, Dr. Alan P. Sprague and Dr. Chiao-Wang Sun, for their precious time, constructive comments, and sharing invaluable experience with me.

I am deeply indebted to my wife Wen-Lin Liu and my family for their unconditional love, unwavering support, and endless care throughout the years. Also, I would like to thank all my colleagues in the KDDM lab: Xin Chen, Song Gao, Ying Liu, John Osborne, Richa Tiwari, Chun Wei, Lin Yang, Ying Zhang and Liping Zhou, with whom I have enjoyed collaboration. I also appreciate many great help from Chu-Feng Chou, Chen-Chung Lin, Hui Ma, Meng Wong, and all my friends for your encouragement and the moments filled with joys. Finally, I owe my special thanks to my friendship partner Corky and Greta Clark for their kindness and warm-hearted assistants.

Thank you all for making my dream come true.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>iii</td>
</tr>
<tr>
<td>DEDICATION</td>
<td>v</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>x</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>xi</td>
</tr>
<tr>
<td>LIST OF ABBREVIATIONS</td>
<td>xiii</td>
</tr>
</tbody>
</table>

## CHAPTER

### 1 INTRODUCTION .........................................................................................................1

- A. Introduction to Image Retrieval ................................................................. 1
- B. Text-based Image Search Engine ................................................................. 3
- C. Content-based Image Retrieval ................................................................. 7
  - C.1 Full Image Search ................................................................................... 12
  - C.2 Object-based Image Retrieval ............................................................... 14
  - C.3 Relevance feedback ................................................................................ 17
  - C.4 Hierarchical Image Representation ....................................................... 19
- D. Summary of Image Retrieval ........................................................................ 21
- E. Semantic Image Retrieval and Its Challenges ............................................ 22
  - E.1 The Challenge of Semantic Gap ............................................................. 22
  - E.2 The Challenge of Image Segmentation .................................................... 24
  - E.3 The Challenge of Multiple-Object Retrieval .......................................... 29
- F. Research Objectives ..................................................................................... 34
- G. Structure of Dissertation ............................................................................ 35

### 2 PRELIMINARY WORKS ...........................................................................................36

- A. Content-based Image Retrieval .................................................................... 36
- B. Content-based Video Retrieval ..................................................................... 37
  - B.1 Abnormal Events Detection in Traffic Surveillance Videos .................. 38
  - B.2 Abnormal Events Detection in Indoor Surveillance Videos .................. 39
- C. Biomedical Visual Information Retrieval ................................................... 40
B.3  Image Scene Retrieval ............................................................................. 170
B.4  Extend MOIR Framework to Audio and Video Retrieval ..................... 171

7  LIST OF REFERENCES.................................................................................... 172
<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The number of images for each object of interest ................................. 76</td>
</tr>
<tr>
<td>2</td>
<td>Subjective segmentation quality assessment ............................................. 81</td>
</tr>
<tr>
<td>3</td>
<td>Performance comparison of DCD and CCH color features on full image search .... 85</td>
</tr>
<tr>
<td>4</td>
<td>The performance comparison details of five image matching algorithms .......... 90</td>
</tr>
<tr>
<td>5</td>
<td>The effectiveness comparison in single object retrieval with feedback .......... 94</td>
</tr>
<tr>
<td>6</td>
<td>The effectiveness comparison in multiple-object retrieval ........................ 122</td>
</tr>
<tr>
<td>7</td>
<td>Details of the 270 query images in fine art painting image retrieval ........... 159</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>13</td>
</tr>
<tr>
<td>7</td>
<td>18</td>
</tr>
<tr>
<td>8</td>
<td>26</td>
</tr>
<tr>
<td>9</td>
<td>28</td>
</tr>
<tr>
<td>10</td>
<td>44</td>
</tr>
<tr>
<td>11</td>
<td>46</td>
</tr>
<tr>
<td>12</td>
<td>49</td>
</tr>
<tr>
<td>13</td>
<td>55</td>
</tr>
<tr>
<td>14</td>
<td>57</td>
</tr>
<tr>
<td>15</td>
<td>60</td>
</tr>
<tr>
<td>16</td>
<td>63</td>
</tr>
<tr>
<td>17</td>
<td>65</td>
</tr>
<tr>
<td>18</td>
<td>69</td>
</tr>
</tbody>
</table>
### LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>AveP</td>
<td>average precision</td>
</tr>
<tr>
<td>CBIR</td>
<td>content-based image retrieval</td>
</tr>
<tr>
<td>CBVIR</td>
<td>content-based visual information retrieval</td>
</tr>
<tr>
<td>CCH</td>
<td>color-code histogram</td>
</tr>
<tr>
<td>CHMM</td>
<td>coupled hidden Markov models</td>
</tr>
<tr>
<td>DBMS</td>
<td>database management systems</td>
</tr>
<tr>
<td>DCD</td>
<td>dominant color descriptor</td>
</tr>
<tr>
<td>DRM</td>
<td>dynamic region matching</td>
</tr>
<tr>
<td>DWT</td>
<td>discrete wavelet transformation</td>
</tr>
<tr>
<td>EMD</td>
<td>earth mover’s distance</td>
</tr>
<tr>
<td>FIRM</td>
<td>feedback-based integrated region matching</td>
</tr>
<tr>
<td>FIS</td>
<td>full image search</td>
</tr>
<tr>
<td>HCK</td>
<td>hill-climbing based color k-means segmentation</td>
</tr>
<tr>
<td>IR</td>
<td>image retrieval</td>
</tr>
<tr>
<td>IRM</td>
<td>integrated region matching</td>
</tr>
<tr>
<td>MAP</td>
<td>mean average precision</td>
</tr>
<tr>
<td>MD5</td>
<td>message-digest algorithm 5</td>
</tr>
<tr>
<td>MGA</td>
<td>modified genetic algorithm</td>
</tr>
<tr>
<td>MIA</td>
<td>microarray image analysis</td>
</tr>
<tr>
<td>MIL</td>
<td>multiple-instance learning</td>
</tr>
</tbody>
</table>
MOIR  multiple-object image retrieval
MRIA  multi-resolution image analysis
PCA   principal component analysis
QBE   query-by-example
QBIC  query by image content
RBF   radial basis function
RF    relevance feedback
ROIs  regions of interest
STM   segmentation tree matching
SVM   support vector machine
TREC  text retrieval conference
UIMA  Unstructured Information Management Architecture
URL   uniform resource locator
CHAPTER 1
INTRODUCTION

A. Introduction to Image Retrieval

In the innovative process of cultural inheritance, precious knowledge is being preserved and accumulated across generations. Most knowledge was stored in various analogue forms in the past. For instances, historical records, legal provisions, and literature works are preserved as written or printed documents over centuries, which reveals the need to develop efficient filing systems to store and retrieve these archived documents. Soon after the invention of computers, it has changed the way how human knowledge is stored. With the development of science and technology, computing performance is rapidly increasing and the cost of storage is continuously decreasing. The evolution of digital technology promotes information storage migrating from analogue to digital form, and results in convenient information sharing and distribution (Li, Wang, & Wiederhold, 2000). The needs for efficient information retrieval emerged from the expeditious growth of digital information.

The history of the automatic information retrieval concept can be traced back to 1945, to a published article titled “As We May Think,” which was authored by Vannevar Bush and disclosed the idea of automatic search for relevant information from an enormous stock of knowledge (Bush, 1945). This groundbreaking idea promotes several key developments in the field of automated information retrieval in the 1950s and 1960s (Singhal, 2001). For instance, Hans Peter Luhn, a pioneer of information science, proposed a crucial statistical-based approach for automatic literary information indexing and retrieval in 1957 (Luhn,
Most conventional information retrieval systems are designed for text-based search because most unstructured content, e.g., article, speech, music, video, or still image, is digitized to the text form that is relatively easy to be indexed. For example, image metadata search, the most traditional and common approach of image retrieval, heavily relies on manually added metadata which includes text-based annotations such as captions, keywords, or image descriptions. Since 1980, the digital revolution has driven the explosion of digital devices on the market, which makes digital imaging emerge from its infancy in the past decade. With the growing number of digital camera and camera phone users, more and more digital images have become available. Therefore, in this dissertation we mainly focus on the topic of image retrieval.

As the adage suggests, “a picture is worth a thousand words.” Information embedded in an image usually provides a more clear and succinct way to present an idea than a substantial amount of text. The emerging needs in retrieving information from images brings researchers’ attention, and thus, image retrieval has been an extremely active research area in the past decade. Approaches in this interdisciplinary research domain require combining various kinds of techniques developed from database management, information systems, image processing, and computer vision (Gupta & Jain, 1997). Many efforts have been made to address this challenging issue. These efforts can be classified into two categories: (1) text-based image search engine, and (2) content-based image retrieval (CBIR).
B. Text-based Image Search Engine

In most conventional text-based image search systems, all images in the search scope, namely target images, must first be annotated in order to enable text-based image retrieval. The annotations such as file name, caption, keywords, tags, and other text-based descriptions, are stored in the associated metadata. Then, the text-based database management system (DBMS) retrieves images based on the annotations stored in the associated metadata (Luo, Wang, & Tang, 2003). For instance, Flickr, a well-known image and video hosting website created by Ludicorp in 2004 and acquired by Yahoo! in 2005, achieves text-based image/video search by introducing one of the Flickr’s key features - “Tags”. This key feature allows Flickr users to manually annotate a photo or a video with textual description, which enables Flickr search engine to index photos and videos in the databases based on the textual description, and helps users easily find stored content having something in common.

A text-based image search system is very efficient in indexing, but it also heavily relies on the text-based annotations in metadata to index the content stored in the databases. As a consequence, the quality of the annotations has a great impact on the quality of retrieval results. Creating high-quality image annotations usually requires extremely laborious manual works, and thus, it is not practical to perform manual annotation on large-scale datasets. In addition, the process of annotating database content is not only time consuming, but also very subjective. The reason is that with a limited set of words, it is very difficult to precisely describe all visual content in an image (Luo et al., 2003). Also, the end-user perception and interpretation of visual content varies from person to person.

For example, Figure 1 represents a beautiful scenic picture of “Stone Mountain”, a famous granite monolith located at east of Atlanta, Georgia, USA. When annotating this im-
age, some people may add tags such as “Stone Mountain”, “stone”, “lake”, “sky”, and “cloud”. Other people may add annotations such as “tree” and “reflection” to describe the same scenic photo. This exemplifies the subjectiveness of manual annotation.

Figure 1: An example image containing a scene from “Stone Mountain, Georgia (GA)” to demonstrate the subjectiveness of manual annotation

Although most state-of-the-art image search engines attempt to automate the image annotation process by analyzing the textual clues surrounded (e.g., images and their surrounding text on a web page), the image retrieval results produced from such search engines often appear poor or mismatched due to the fact that the surrounded text may be irrelevant to the content of an image.

For example, when we search images containing Stone Mountain, the watermelon image (as shown in Figure 2) may come up on top. The reason is that both the title and the caption on the original web page contain the search keywords “Stone Mountain” as outlined by red rectangles in Figure 2. This example demonstrates the problem that text-based image
retrieval heavily relies on image annotations or surrounded text rather than the visual content in the image.

Figure 2: An example that demonstrates the problem that text-based image retrieval heavily relies on image annotations or surrounded text rather than the visual content.

The last but not the least problem is that text-based image retrieval systems heavily rely on image annotations or surrounded text rather than semantics, and thus, cannot distinguish between homonyms. In Figure 3, we demonstrate using the keyword “stone” to perform a search with Google Images. From the top 20 retrieved results, 5 images are related to a famous American actress “Sharon Stone.” This mismatch is due to the fact that text-based image search engines cannot discriminate the keyword “stone” from the name of the actress.
“Sharon Stone”. In addition, several other images, such as stone wall, stone mouse and stone bathtub, also appear in the retrieval results. This is because the text descriptions of all these images contain the search keyword “stone”.

Figure 3: The top 20 images retrieved by Google Images using “stone” as the search keyword. From the top 20 retrieved results, 5 images are related to a famous American actress “Sharon Stone” and several other images contain stone mouse and stone bathtub, which demonstrates the problem that text-based image retrieval systems heavily rely on image annotations or surrounded text rather than semantics, and thus, cannot distinguish between homonyms.
In order to cope with these aforementioned issues that arise in text-based image retrieval systems, another research field, Content-Based Image Retrieval (CBIR), has been introduced to retrieve images based on the visual content of images.

C. Content-based Image Retrieval (CBIR)

In contrast to text-based image retrieval, content-based image retrieval (CBIR), also known as query by image content (QBIC) and content-based visual information retrieval (CBVIR), describes the automatic process of retrieving images from an image database according to the visual features extracted from the images. This concept was first introduced by Kato in 1992 (Kato, 1992). A comprehensive survey article by Lew et al. summarizes the recent achievements in the field of content-based image retrieval (Lew, Sebe, Djeraba, & Jain, 2006).

The content-based image retrieval (CBIR) framework suggests that three levels of abstraction can be retrieved as the content of an image. Level-3, the highest level of abstraction, comprises retrieval by abstract attributes which interpret the meaning, concept, or scenes depicted. Level-2, the middle level of abstraction, is called logical features, also known as derived features, which infer the identity of objects depicted in an image. Level-1, the lowest level of abstraction, is known as primitive visual features. Primitive visual features, such as color, texture, shape features and spatial locations, are some examples of fundamental information that can be directly extracted from images (Eakins & Graham, 1999). Figure 4 shows a sample image to exemplify the three levels of abstraction.
Figure 4: An example image for demonstration of the three levels of abstraction. The birthday scene represents the highest level of abstraction (level-3); the containing objects, such as the birthday cake, two glasses of wine, candles, ribbons, and gift box, represent the middle level of abstraction (level-2); the visual characters of object, such as the color of wine – dark red, the shape of cake – round, and the pattern of the gift box – striped, represent the lowest level of abstraction (level-1).

Figure 4 depicts a birthday scene which represents the highest level of abstraction. This scene is composed of several objects associated with the birthday subject, such as a birthday cake with some candles, a gift box, two glasses of red wine, and some decoration ribbons. These aforementioned objects belong to the middle level of abstraction. Each object can be described with a set of primitive visual features, such as color features, texture features, shape features, and spatial information, which are at the lowest level of abstraction.

For example, the color of the cake is white and pink; the texture of the cake is the decoration patterns; the shape of the cake is round; the spatial location of the cake is at the right part of the image.
In addition to the three levels of abstraction in the content-based image retrieval (CBIR), the architecture of a generic content-based image retrieval (CBIR) system usually employs a query-by-example (QBE) interface which allows users to provide a sample query image to the system (Kelly & Cannon, 1995). For instance, “Search by Image”, a novel feature introduced to Google Images in October 2011, is an example of such a QBE interface for content-based image retrieval systems. Users can access this novel feature by clicking on the camera icon which is located at the right side of the Google Images search bar. “Search by Image” allows users to search similar images on the Internet by submitting an arbitrary sample image to Goggle Images. As demonstrated in Figure 5, users can upload a sample image by either dragging & dropping one sample picture to the search bar (Figure 5a), or providing the uniform resource locator (URL) of a web image (Figure 5b).

Another example of the QBE interface is “Google Goggles”. Google Goggles, one of the existing commercial content-based image retrieval (CBIR) systems, demonstrates the power of QBE-based CBIR technology (Google, 2010). It supports photo-based searching on handheld devices by taking a sample picture with the built-in camera. As a query image, the sample picture is submitted to the search engine for retrieving related information. For example, if you take a photo or upload an image of the Eiffel Tower, Google Goggles would recognize it as landmark, and then, search for all information related to it, such as a map of Paris showing the location of the Eiffel Tower. Although Google Goggles has shown a promising performance on searching landmarks, book covers, artwork, places, wine, logos, and business cards, it has its limitation on searching natural scene objects such as food, cars, plants, or animals.
Figure 5a: Google Images – “Search by Image” allows users to search visually similar images by dragging and dropping an image into the search bar as query sample image.

Figure 5b: Google Images – “Search by Image” also allows users to provide the URL of a web image as the query image.

Typical content-based image retrieval (CBIR) systems view the query image and target images, i.e., all images in the database, as a collection of primitive visual features such as color, texture, shape and spatial location. On the basis of these primitive visual features, a content-based image retrieval system measures the similarity between a query image and each target image in the database, and then, it ranks and retrieves target images in the data-
base according to their similarity scores to the query image (Chen, Li, & Wang, 2004). From this image retrieval process, three fundamental steps can be summarized for content-based image retrieval framework, namely primitive visual feature extraction, multi-dimensional indexing, and retrieval system design (Rui, Huang, & Chang, 1997).

Methodology in Content-based image retrieval (CBIR) can be further divided into two major categories, including full image search frameworks and object-based image retrieval frameworks. The full image search framework, also known as non-region-based image retrieval, retrieves images based on the visual features extracted from the whole image (Wong & Po, 2004; Stone & Li, 1996; Kelly & Cannon, 1995; Pentland, Picard, & Sclaroff, 1994; Samadani, Han, & Katragadda, 1993). In general, full image search is relatively simple and efficient, but less human-centered. The reason is that humans find images on the basis of the high level concepts, such as objects (level-2 abstraction) or scenes (level-3 abstraction); however, global visual features used in the full image search only represent the level-1 abstraction, incapable of capturing the properties of those higher level concepts.

In contrast to full image search, another line of approaches is object-based image retrieval, also known as region-based image retrieval, which attempts to capture the high level concepts embedded in images such as objects. Object-based image retrieval framework retrieves images based on the visual features extracted from the regions that correspond to meaningful objects in images (Zhou, Zhang, Wan, Birch, & Chen, 2009; Ji, Yao, & Liang, 2008; Li, Zhang, Lu, & Ma, 2007; Zhang & Chen, 2005b; Carson, Belongie, Greenspan, & Malik, 2002).
C.1 Full Image Search

In content-based image retrieval (CBIR), the full image search, also known as non-region-based image retrieval, is the simplest baseline approach. A full image search framework has the characteristics of easy-to-implement and is relatively efficient in retrieval. A basic framework in this kind usually uses global features, i.e., the visual features extracted from the entire image, to describe the characteristic properties of that image. The similarity between the query image and each target image is then assessed according to the extracted global features. All target images are ranked in the decreasing order of their similarities to the query image. Then, the ranked target image list is returned to the user.

Owing to the compact feature vector, retrieving images based on global visual features is very efficient for matching, but it has some drawbacks. The first problem is that the local information and spatial configuration of image are missing (Alajlan, Kamel, & Freeman, 2006). Another problem is that basic full image search cannot accommodate the need to adapt to different focuses of attention of different users (Zhang, Chen, & Shyu, 2004). In Figure 6, we demonstrate the problems of basic full image search with an example.

In this example, our query image (as shown in the top-left corner in Figure 6) is a video frame captured from a drama video. We are interested in the “di yi”, the traditional pheasant patterned ceremonial robe of a queen, which appeared in the center of the query image. Our goal is to search visually similar images containing “di yi” using Google Images – “Search by Image”.

As demonstrated at the bottom of Figure 6, Google Images returns 16 visually “similar” images. It is obvious that the query image and all the 16 retrieved images have highly similar color composition. Specifically, all the retrieved top images contain three prominent
colors, i.e., blue, red, and brown. However, none of these retrieved images are relevant to our interest – “di yi”. The image retrieval results in this example faithfully exemplify the two aforementioned issues in basic full image search.

Figure 6: Search results for the query sample image “di yi” on Google Images – “Search by Image”. The query sample image containing the search target “di yi” is shown at the top of Figure 6. Sixteen top similar images are retrieved and displayed at the bottom of Figure 6. None of these visually similar images are related to our search target “di yi”. This example demonstrates the problems of basic full image search.
In general, full image search is relatively simple and efficient, but less capable of adapting to individual users’ needs. The reason is that humans find images based on the high level concepts, such as objects or scenes; however, global visual features used in the full image search cannot capture the properties of those high level concepts. In order to deal with these issues, object-based image retrieval framework and relevance feedback are then introduced.

C.2 Object-based Image Retrieval

In contrast to full image search, the object-based image retrieval framework (also known as region-based image retrieval framework) retrieves images based on the visual features extracted from the regions that correspond to meaningful objects in images. Several systems, such as Blobworld system (Carson, Belongie, Greenspan, & Malik, 2002; Carson, Thomas, Belongie, Hellerstein, & Malik, 1999), NeTra system (Ma & Manjunath, 1997a), and SIMPLIcity system (Wang, Li, & Wiederhold, 2001; Wang, Li, Chan, & Wiederhold, 1999), have been developed in the area of object-based image retrieval.

Blobworld, for instance, is one of the renowned object-based image retrieval systems. In order to enable object-based image search, this image retrieval system automatically splits images into segments ("blobs"), each of which is a homogenous region roughly corresponding to an object, with associated visual features, such as color and texture. Then, using the visual features extracted from each region, the system finds coherent image regions of users’ interests between a query image and all target images. Blobworld also introduces an efficient indexing structure to meet the needs of the large-scale retrieval.
Both Blobworld and NeTra systems adopt image segmentation to preserve local information, unlike the full image search which represents images with global features, and measure image similarity on the basis of local features extracted from individual regions. Besides, these systems also allow users to search target images based on a limited number of regions by merging single-region query results. However, these systems suffer from the imperfection of the segmentation algorithms, which can produce incorrect mapping between objects and segments, and thus, users often have difficulty to determine which regions should be used for retrieval. In addition, semantically-precise image segmentation is still an open problem in image processing (Li, Wang, Gray & Wiederhold, 1999; Li & Gray, 1998; Ma & Manjunath, 1997b; Malik, 1997; Zhu & Yuille, 1996), which makes region-based image matching scheme extremely difficult. Another problem of the Blobworld system and the NeTra system is that they still require users to explicitly specify the region of interest and are incapable of automatically discovering the regions of a user’s attention in the query image. The reason is that there is no learning mechanism involved in these region-based image retrieval systems to reveal the focuses of attention of users.

In order to address the aforementioned problems, Li et al. (2000) introduced the integrated region matching (IRM) scheme which measures the overall similarity between images according to the overall similarity between two sets of image regions. Another approach, dynamic region matching (DRM) was also introduced by Ji et al. (2008) to address the same issue. The idea behind DRM scheme is similar to that of the IRM scheme, but adopts a probabilistic fuzzy region matching algorithm to cope with the problem of inaccurate image segmentation. Further, DRM scheme enables the automatic discovery of the user’s region of interest (ROI) using “FeatureBoost” as a RF learning algorithm. However, both IRM and DRM
based approaches still suffer greatly from inaccurate segmentation especially over-segmentation, in addition to the other many problems they have.

These abovementioned approaches retrieve “objects” on the basis of a collection of independent segments/regions which may not individually correspond to semantic objects, without considering the associative relationships between image regions. On top of that, the adverse effect of inaccurate segmentation has become a major bottleneck that impedes the advancement of object-based image retrieval systems. For example, over-segmented regions that originate from different objects may be extremely similar, and thus, may aggravate the problem of false positives.

Another example, Li et al. (2000) assume in IRM scheme that larger segments are more important than smaller segments based on observers’ perception, and thus, introduce an “area percentage scheme” for assigning region significance in the IRM scheme. The area percentage scheme emphasizes on large regions and assigns a higher weight to them. However, this assumption is problematic in object-based image retrieval because the background region is often the largest region among all regions in many images. It is not reasonable to assign a very high significance value to background regions because doing so may overwhelm those middle or mid-small sized objects which may truly represent the objects of the user’s interest.

Another problem is that the integrated region matching (IRM) scheme is designed to measure the overall similarity between images according to the similarity between two sets of image regions. Like the Blobworld and the NeTra systems, the IRM scheme still does not involve any learning algorithm that enables the image retrieval system to progressively reveal the focuses of attention of users.
C.3 Relevance feedback

While those high level concepts of users’ interests come naturally to a human being, they pose a big challenge to computer systems due to the so-called semantic gap. This is because computer systems can only recognize those low level primitive visual features, but not necessarily the high level concepts. In order to bridge the semantic gap and automatically discover the focuses of attention of users, relevance feedback (RF) and multiple-instance learning (MIL) are two techniques introduced into object-based image retrieval systems.

Relevance feedback is an interactive process in which the user judges the quality of the retrieval results returned by the retrieval system (Su, Zhang, Li, & Ma, 2003; Rui, Huang, & Mehrotra, 1997b). The procured feedback from the user is then used to refine the original query. To learn from the user’s feedback information, multiple instance learning (MIL) is proposed to seamlessly integrates with relevance feedback for single object-based retrieval (Zhang et al., 2004; Huang, Chen, & Shyu, 2003; Zhang, Goldman, Yu, & Fritts, 2002). This supervised machine learning scheme has the assumption that the user’s concept can be represented as a single “best” object. However, the need for discovery of multiple objects of the user’s interest is also not uncommon, and it thus is more natural to have more than one significant object associated with a search target.

For instance, one user may want to search for those images containing a red bus running on the street as shown in Figure 7a; while another user may be more interested in a red bus parked on the grassland as shown in Figure 7b. This exemplifies the emerging need for multiple-object retrieval in image databases.
Figure 7: This example demonstrates the emerging need for multiple-object retrieval in image databases: a) a user is interested in a red bus running on the street while another user is interested in b) a red bus parked on the grassland.

As aforementioned, using IRM measure alone is not sufficient for dealing with the problem of multiple-object retrieval due to the fact that the IRM scheme only measures the overall similarity of two images in terms of their overall object similarity but cannot identify
the objects of users’ interest. In 2010, Zhang et al.’s work has demonstrated that combining integrated region matching (IRM) scheme with users’ relevance feedback (RF) in a multi-object based image retrieval framework makes the automatic discovery of user desired object(s) possible (Zhang, Zhou, Wan, Birch, & Chen, 2010). The success in automatic discovery of the user’s focuses of attention in Zhang et al.’s work indicates that the relevance feedback technique can be used in this research in order to identify the objects/regions of the users’ interests.

To this point, this research mainly target multiple-object retrieval, which focuses on the level-2 abstraction and the image retrieval based on it, but it will also be built on top of the feature representations from level-1 abstraction. In addition, the proposed framework needs to incorporate a learning mechanism with the relevance feedback technique in order to discover the focuses of attention of users. The proposed object-based image retrieval framework falls into the category that is a region-based CBIR with a learning mechanism. Several other works in the category include (Zhang, Zhou, Wan, Birch, & Chen, 2010; Zhang, Chen, Chen, Chen, & Shyu, 2005a; Zhang & Chen, 2005b; Zhang, Chen, & Shyu, 2004; Chen & Wang, 2004; Huang, Chen, & Shyu, 2003; Jing, Li, Zhang, Zhang, & Zhang, 2003a; Jing, Li, Zhang, Zhang, & Zhang, 2003b).

C.4 Hierarchical Image Representation

The goal of this research is to develop an effective and efficient multiple-object image retrieval framework which is more robust to the problem of over-segmentation. We believe that the key to relieving these aforementioned issues is a new systematic and hierar-
chical representation of visual information, and the corresponding analysis and retrieval framework that makes it possible for a machine to interpret an image in terms of its containing regions and their relationships. For this reason, it is essential to effectively preserve the spatial and neighboring relationships among regions in order to better model the image content. One possible solution is to adopt hierarchical image representation to preserve such relationships among regions.

To the best of our knowledge, most existing approaches construct a hierarchical image representation in two steps (Li, Guo, Song, & Xiao, 2011; Al-Qunaieer, Tizhoosh, & Rahnamayan, 2011; Vilaplana & Marques, 2007; Sumengen & Manjunath, 2005; Prewer & Kitchen, 2001; Xu, Duygulu, Saber, Tekalp, & Yarman-Vural, 2000). The first step is to perform segmentation at different image resolutions, and the second step is to construct the hierarchical representation of the image by associating regions from different resolutions. This two-step process has a low efficiency due to high time-complexity associated with the multi-scale image analysis.

Another performance issue in terms of efficiency emerges from using image hierarchy to preserve spatial and neighboring relationships. One critical issue is how to perform similarity comparison efficiently on the basis of hierarchical image representations. This is because there can be many sub-trees in one hierarchical region tree, not to mention when comparing all proper sub-tree pairs from a given pair of trees. For this reason, there is a demand for an efficient algorithm for matching two hierarchical region trees.
D. Summary of Image Retrieval

Content-based image retrieval (CBIR) differs from text-based image search engine in two aspects.

First, from the aspect of methodology, a content-based image retrieval framework retrieves desired images from image databases by means of analyzing the actual contents of images rather than searching through the text-based annotations in the metadata.

Second, from the aspect of data properties, a content-based image retrieval framework analyzes essentially unstructured data which are an array of pixel intensity values without any inherent meaning, whereas text-based image search engine searches essentially structured data which are a set of words with lexical semantics.

These characteristics make the essence of content-based image retrieval different from that of the conventional text-based image retrieval (Eakins & Graham, 1999). In addition, content-based image retrieval systems are desirable because they do not require humans to manually tag every image in the database with keywords, annotations, and/or descriptions.

As the image retrieval community suggests, the content-based image retrieval framework is proposed to complement the insufficiency of text-based image search engines, rather than replace them. Therefore, it is essential to integrate both frameworks as one in order to obtain a satisfactory retrieval performance (Rui et al., 1997a). Although both text-based and content-based image retrieval frameworks are equally important, we mainly focus on describing a framework that will improve the accuracy of object-based content-based image retrieval in this dissertation.

In brief, the goal of this research is to develop an effective and efficient multiple-object image retrieval framework which can alleviate the over-segmentation problem by in-
Introducing the hierarchical image representation, but does not suffer from the inefficiency during the construction of the image hierarchy or the comparison of hierarchical representations of images. In this dissertation, we introduce a multiple-object image retrieval system named (MOIR) in order to achieve the above goals. In the proposed MOIR framework, we develop an efficient algorithm named “Multi-Resolution Image Analysis” (MRIA) to perform image segmentation and construct the image hierarchy all in one run. This is achieved by designing a branch-and-bound-like algorithm that performs image segmentation and hierarchical tree construction concurrently, and the analysis progresses from low resolution to higher resolution and uses certain constraints to improve efficiency. In addition, we also design an efficient algorithm that is used to compare two image hierarchies representing two images.

In order to provide a better understanding of the proposed research, a review of the challenging issues in semantic image retrieval is presented in the next section.

E. Semantic Image Retrieval and Its Challenges

E.1 The Challenge of Semantic Gap

The semantic contents of images, what users generally look for, manifest in the top two levels (levels 2 & 3) of abstraction. Thus, CBIR researchers refer these two levels of abstraction as semantic image retrieval (Gudivada & Raghavan, 1995). In the process of semantic image retrieval, it is extremely difficult for computers to answer queries such as “find photos of birthday celebration scenes” or “find pictures with two glasses of wine” on the basis of their semantic meanings. This is because computers can only recognize low-level primitive visual features but cannot perceive high-level semantics contained in images as what
human beings can do. This raises one of the key issues in CBIR research which is how to bridge the “semantic gap” between low-level visual features and high-level user perceptions when modeling the content of image with primitive visual features (Eakins & Graham, 1999).

Many efforts have been made to reduce the semantic gap. These state-of-the-art approaches can be categorized into five classes according to the methodology including (1) describing high-level semantics with object ontology, (2) using machine learning techniques to associate primitive visual features with the query concepts, (3) introducing relevance feedback (RF) which identifies users’ perception in the iterative user intervention process, (4) supporting high-level image retrieval with semantic template, and (5) retrieving images on the Web based on both visual and textual information contained in Web pages (Liu, Zhang, Lu, & Ma, 2007). Among these five categories, relevance feedback (RF) technique is known to be an effective approach to reduce the semantic gap and has been an active research domain for the past decade. RF incorporates the user’s concept with the iterative learning process (Zhang, Chen, Chen, Chen, & Shyu, 2005a; Su, Zhang, Li, & Ma, 2003; Rui, Hunag, & Mehrotra, 1997). More specifically, relevance feedback (RF) technique is a supervised machine learning process that refines the query by means of either query modification or query reweighting (Porkaew & Chakrabarti, 1999). Query modification attempts to use feedback information to improve the estimation of query point so that the new query point can better represent the user’s desire. In query reweighting, it adjusts the importance of features by changing the corresponding weights. In this research, we propose to seamlessly integrate relevance feedback technique with the proposed image retrieval framework. Several challenging issues remain in incorporating the RF technique. The first issue is how to efficiently and effectively learn from a small set of feedback samples. The second issue is how to accumu-
late knowledge learned from the relevance feedback. The third issue is how to combine low-
level visual and high-level semantic features in a query, more specifically how to combine
keyword-based search with CBIR.

E.2 The Challenge of Image Segmentation

As aforementioned, content-based image retrieval (CBIR) system retrieves images
based on a collection of primitive visual features. These primitive visual features can be di-
extracted from either the entire image (as global features) or from regions (as local fea-
tures). Most current content-based image retrieval (CBIR) systems are designed as an object-
based image retrieval framework owing to the fact that users are usually more interested in
specific regions rather than the entire image. In addition, it would be a closer match to hu-
man’s perception system if images are represented at the region level (Jing et al., 2003).

In order to enable object-based search, it is essential to extract meaningful objects
embedded in images. An object-based image retrieval framework requires images partitioned
into regions of interest (ROIs). The process of partitioning an image into several constituent
components, each of which represents a meaningful object in the image, is called image seg-
mentation. Image segmentation is known to be one of the most challenging issues in the field
of image processing and computer vision. Many efforts have been made to improve the seg-
mentation accuracy. Most segmentation algorithms extract image segments on the basis of
either color (Rahimizadeh, Marhaban, Kamil, & Ismail, 2009; Lucchese & Mitra, 2001), tex-
ture (Xie & Mirmehdi, 2007), or edge features (Arbeláez, Maire, Fowlkes, & Malik, 2009;
Yu & Clausi, 2008). Only few segmentation algorithms combine multiple visual features
Another challenging issue in image segmentation is over-segmentation and under-segmentation. Due to the imperfection of segmentation algorithms, segmentation results obtained from most of the existing segmentation algorithms are often over-segmented and/or under-segmented. Over-segmented regions indicate that an object is divided into two or more smaller segments. On the other hand, under-segmented regions indicate that two or more objects are merged into a larger region. Figure 8 shows an image with both over-segmented and under-segmented regions produced from a state-of-the-art contour-based image segmentation algorithm (Arbeláez et al., 2009).

In Figure 8, we can observe from the segmented image that the white horse object is over-segmented since its two front legs appear in two different regions/segments than the main horse body region. On the contrary, the brown horse at the right of the image is under-segmented because some parts of horse are merged either with the tree or with the grass regions. In this example, the under-segmentation is due to the fact that the image segmentation algorithm partitions image into regions based on the object contours and there is no distinct edge between the brown horse and the tree shade regions.
Figure 8: An example image with both over-segmented and under-segmented regions. The segmentation results are produced from a state-of-the-art contour-based image segmentation algorithm. The original image (top) and the segmentation results (bottom)
Most recently, Zhao et al. proposed a supervised image segmentation framework – “CO3” (Zhao, Zhu, & Luo, 2010). This interactive image segmentation system requires users to scribble on the foreground and background areas in order to collect essential training data for building a region appearance model. The region appearance mode is then used to distinguish between the foreground and the background regions in the given image. In Figure 9, we demonstrate some segmentation results produced from the state-of-the-art CO3 framework. This approach can produce promising segmentation results. However, there are several drawbacks in this interactive image segmentation system.

First, the CO3 framework suffers from the scalability problem. The reason is that the CO3 framework requires users to be heavily involved in the process of training data collection, which makes it impractical for dealing with large-scale image databases. The second problem is that the discriminative model of the CO3 framework is actually a two-class classifier which can only distinguish between foreground and background regions and therefore, it cannot meet the needs of multiple-object retrieval.

While both over- and under-segmentation cause problems, under segmentation has a more negative impact on object-based image retrieval. The reason is that an under-segmented region represents several different objects with one region in the image, which is less useful in the object-based image retrieval. On the other hand, an over-segmented region could still represent part of an object from which the features extracted can still represent the object’s properties, though partially, but more truthfully than that extracted from an under-segmented region. Therefore, most existing segmentation algorithms tend to over-segment. Then, the main challenge is how to alleviate the problem of over segmentation in object-based image retrieval.
Figure 9a. Original image

Figure 9b. User scribbles – Foreground (Red) & Background (Green)

Figure 9c. Segmentation results produced from CO3
The Challenge of Multiple-Object Retrieval

In content-based image retrieval, most object-based CBIR systems consider each region, generated by segmentation, in an image as an object, and thus, these CBIR systems retrieve images on the basis of the similarity between two (or two sets of) regions, obtained from the query image and a target image in the database, respectively. However, when an image is over-segmented and/or under-segmented, the assumption of one-to-one mapping between object regions is no longer valid and becomes one-to-many and/or many-to-one mapping. The performance of region-based matching scheme thus suffers from both over-segmentation and under-segmentation, because the one-to-one mapping between segments and objects is no longer preserved.

The complex mapping of segment and object results in a less than robust similarity measure under the region-based matching scheme and therefore may negatively impact the image retrieval performance. Li et al. (2000) introduced the integrated region matching (IRM) scheme which is an overall image-to-image similarity measure based on the similarity of two sets of regions. However, the IRM scheme cannot identify users’ interest since it only matches similar regions from two images based on low-level visual features. Although IRM scheme uses the region size as the significance score of each region, it only emphasizes dominant regions but not the actual regions that interest users. For this reason, IRM scheme alone is not suitable for the object retrieval task. (Zhang, Zhou, Wan, Birch, & Chen, 2010; Zhou, Zhang, Wan, Birch, & Chen, 2009).

Another similar approach, dynamic region matching (DRM) was also introduced to address the same issue mentioned above (Ji et al., 2008). The idea behind the dynamic region matching (DRM) scheme is similar to that of the integrated region matching (IRM) scheme,
but adopts a probabilistic fuzzy region matching algorithm to deal with the problem of inaccurate image segmentation. In addition, the authors of DRM framework introduced a “FeatureBoost” learning algorithm in the relevance feedback (RF) process in order to automatically capture the user’s region of interest (ROI). However, both IRM and DRM schemes still suffer greatly from the inaccurate segmentation especially over-segmentation.

Most existing object-based image retrieval systems are based on single region matching, with the main limitation being that one individual image region can sometimes hardly represent the objects of the user’s desire especially when more than one region of interest is involved in the retrieval. Although the IRM scheme has been attempted to enable multiple region-based retrieval in image databases, it is not capable of discovering automatically user-preferred regions (Zhang, Zhou, Wan, Birch, & Chen, 2010; Zhou, Zhang, Wan, Birch, & Chen, 2009).

The first attempt to integrate relevance feedback (RF) into a multiple object-based image retrieval system was made in Zhang et al. (2004). In this framework, the Multiple Instance Learning (MIL) technique is used to learn user-preferred regions by examining the relevance feedback from users. More specifically, MIL considers a segmented image as a bag, and each region in a segmented image is an instance. When a user provides feedback, he/she labels each retrieved image as relevant (positive bag) or irrelevant (negative bag). A positive labeled bag implies that at least one region is relevant while a negative labeled bag indicates that none of the regions in that image are relevant. By iteratively analyzing users’ feedback with a multilayer feed-forward neural network and back-propagation, the MIL-based approach can progressively identify user-preferred regions. However, in this work, each user interested region needs to be discovered through one full run of MIL training process, which
is very time consuming, and the number of interested regions/objects must be known at the front. Impeded by the efficiency issue, only a handful of examples of two-object retrieval was presented in Zhang et al. (2004). Although Zhang (2005a) further enhanced the learning process with one-class support vector machine, the efficiency issue and the excessive burden for user input remain two challenging problems. In addition, these existing MIL-based image retrieval frameworks, although offer multi-object retrieval capability, all suffer greatly from over- and under-segmentation problems.

For these reasons, the needs in performing effective and efficient multiple objects search in images draw our attention and raise an important issue of how to model the content of an image based on its segmented regions. In order to accurately model the image content on the basis of segmented regions, it is essential to preserve relations among segmented regions. Most existing CBIR systems such as SIMPLicity (Wang et al., 2001) perform image retrieval without knowing the relations between segmented regions in images. In this regard, a hierarchical tree is known to be an efficient data structure that can be used to describe the relations among nodes and similarity between nodes in the tree. This motivates researchers to attempt a hierarchical representation of image regions in order to preserve the spatial and neighboring information for modeling the content of images (Lim, Arbeláez, Gu, & Malik, 2009; Du & Bui, 2006; Yu, Fritts, & Sun, 2002).

As aforementioned, image segmentation algorithms generally suffer from high time complexity. Only few image segmentation algorithms have the ability of providing real-time segmentation (Felzenszwalb & Huttenlocher, 2004). However, none of them can generate hierarchical representation of image segments. Therefore, it is impractical for an object-based CBIR system to perform hierarchical region-tree based image segmentation on the fly. There
is no exception even to the state-of-the-art image segmentation algorithm (Arbeláez et al., 2009). Built on the segmentation algorithm proposed in (Arbeláez et al., 2009), Lim et al. (2009) have proposed to build a region-tree to better model the content of images in term of its segmented regions. The entire process from image segmentation to region-tree construction is known to be a time-consuming process which takes more than 210 seconds to segment a 380×240 image, plus the time spent on the hierarchical region-tree construction which usually takes a comparable amount of time as that of the segmentation step. This motivates us to develop an effective and efficient image segmentation algorithm which uses combinatorial visual features for segmentation while at the same time preserves the relations as a hierarchical tree for the segmented regions.

In the communities of image processing, computer vision, and signal processing, pyramid representation describes an iterative process of smoothing and down-sampling, a process also known as multi-resolution analysis. This process generates a multi-scale signal representation of the original signal. Pyramid representation of an image is considered to be more efficient than the original image representation in terms of the amount of information preserved in pixel values because the later may contain a lot of redundant information (Burt & Adelson, 1983). Therefore, it is possibly a promising direction to use multi-resolution analysis as the core of the image segmentation algorithm. With the proposed approach, a hierarchical region-tree will naturally form during the process of image segmentation; the time complexity of segmentation can also be reduced at the same time since analyzing a lower resolution image is much more efficient than analyzing the original image and not all regions need to be segmented at all resolutions.
In existing multi-object retrieval frameworks, the object matching is achieved through the region matching. However, when the content of an image is represented as a hierarchical region-tree, the object matching becomes a sub-tree comparison. This is because an object in the hierarchical region-tree representation could be either a sub-tree if the object is over-segmented or a leaf node if the object is ideally segmented. Further, when under-segmentation occurs, more than two objects may exist in one leaf node. Therefore, using the hierarchical region-tree may increase the time complexity when matching objects through the sub-tree comparison. This poses another challenge to the proposed study where an efficient sub-tree comparison algorithm is required. In the proposed research, we design and implement an efficient sub-tree comparison strategy rather than a brute-force implementation of full sub-tree comparison.

Another challenge remaining in this research is how to automatically and effectively discover users’ interest given scarce and imbalanced feedback information as training data. In this research, we propose a hierarchical tree based relevance feedback scheme to identify users’ interest. The proposed scheme considers each proper subtree in the query image a merited candidate and adopts a machine learning mechanism to learn users’ interest. The machine learning mechanism in the proposed framework actually discovers candidates according to the result of proper sub-tree comparison and the user’s feedback so that it provides an opportunity to bridge the semantic gap between low-level primitive visual features and high-level users’ perceptions.
F. Research Objectives

Despite the advances in information retrieval and multimedia data analysis, there are still several challenging issues remaining in content-based image retrieval (CBIR) research. In this dissertation, the proposed research aims to provide an innovative image retrieval framework for multiple-object retrieval in image databases. This framework minimizes the need of user intervention, while maximizing the retrieval performance. The following objectives are expected to be achieved:

First, in this research, an efficient and accurate hierarchical image segmentation algorithm based on multi-resolution image analysis will be developed to alleviate the over-segmentation and under-segmentation problems through the preservation of associative relationships among image regions as a hierarchical region tree. The hierarchical representation of visual information reflects the image composition which evokes the interest of the viewer, and thus, enables a machine to interpret an image in terms of its containing regions and their relationships.

Second, with hierarchical region tree representations, the relevance of an image to the query image is thus measured according to the proper sub-tree comparison. As a full comparison of all sub-trees is unlikely to be feasible, we will design and develop an efficient strategy for comparing proper sub-trees.

Third, the proposed research will seamlessly integrate users’ relevance feedback (RF) into the proposed multiple-object retrieval system, which allows automatic discovery of the objects of users’ interest and is expected to improve the retrieval accuracy through feedback-retrieval loops. This user-centered design allows users to retrieve multiple objects in images
with little user intervention. We expect this CBIR research to provide a user-friendly environment for image retrieval that will enhance user experience in searching desired images.

G. Structure of Dissertation

The remainder of this dissertation is organized as follows. In Chapter 2, we present preliminary works related to the proposed research. In Chapter 3, we formally describe the details of the proposed multiple-object image retrieval framework. In Chapter 4, we introduce the evaluation metrics and conduct several experiments to demonstrate the efficiency and efficacy of the proposed multiple-object image retrieval framework. In Chapter 5, we extend the applications of the proposed multiple-object image retrieval framework to fine-art and painting image retrieval. In Chapter 6, we conclude this dissertation and discuss the possible future extensions to this research.
CHAPTER 2
PRELIMINARY WORKS

A. Content-based Image Retrieval

A feedback-based image clustering and retrieval framework is developed in this study. This framework adopts an innovative clustering algorithm, modified genetic algorithm (MGA), to group all image regions in the databases on the basis of their visual feature similarity. This clustering process effectively reduces the search scope in the subsequent retrieval step, and thus, makes this image retrieval framework more efficient. In addition, in this study, we design a feedback-based integrated region matching (FIRM) scheme which successfully integrates the integrated region matching (IRM) with relevance feedback (RF), and thus, improves the effectiveness of the object-based image retrieval. This paper also demonstrates the effectiveness of FIRM in object-based image retrieval by evaluating the system performance on a large-scale image database.

The contributions of this study include: (1) we develop an innovative clustering algorithm, modified genetic algorithm (MGA), and successfully apply this algorithm on image clustering in order to narrow down the search scope and reduce the time-complexity for retrieval; (2) we design a feedback-based integrated region matching (FIRM) scheme by integrating the integrated region matching (IRM) and relevance feedback (RF) into an image retrieval framework; (3) this robust and efficient framework significantly outperforms dynamic region matching (DRM) and support vector machines (SVM) in terms of retrieval accuracy (Zhang et al., 2010; Zhou et al., 2009).
This preliminary work in multiple-object retrieval is the foundation of the proposed CBIR research. Hierarchical representations of image content will be incorporated to alleviate the problem of over-segmentation that we encountered in our preliminary study. In addition, the IRM scheme used in our preliminary study will be replaced by an efficient proper sub-tree comparison scheme in order to provide multiple-object retrieval ability on the basis of hierarchical representations of images. Further, in the preliminary study, we only analyze the positive feedback with SVM. In the proposed research, we will develop a voting scheme that can potentially learn users’ interest from relevance feedback by considering positive feedback as well as negative feedback.

B. Content-based Video Retrieval

Videos are composed of moving images (frames). Therefore, content-based video retrieval can be considered as content-based image retrieval with extra temporal information. In our preliminary work, we further extend the concept of content-based image retrieval to video event retrieval in order to detect abnormal events such as illegal driving behavior and accident, in the traffic surveillance video and to identify abnormal events such as chasing, fighting, and robbery, in the indoor surveillance video.

The surveillance camera system has been widely used to ensure site security, which not only deters criminals but also provides solid evidence for settling disputes or investigating crimes. The interpretation of surveillance videos heavily relies on human, which requires dedicated personnel spending hours or days, to sequentially screen and review the entire video sequence. The event extraction from surveillance videos becomes even more difficult
when videos are recorded from multi-cameras since images from different cameras may be displayed concurrently in one screen or may be interleaved in a round-robin fashion. For this reason, it is essential to automate the process of surveillance so that the time and labor cost can be greatly saved. The goal of this research is to automate the monitoring process for assisting security personnel to identify abnormal events and to retrieve specific abnormal events from the video database.

B.1 Abnormal Events Detection in Traffic Surveillance Videos

One of our preliminary works in content-based video retrieval is that we successfully apply Principal Component Analysis (PCA) & Support Vector Machine (SVM) techniques with the relevance feedback (RF) and multiple-instance learning (MIL) for retrieving various abnormal driving behaviors and accidents in traffic surveillance videos.

The purpose of this research is to detect abnormal events such as illegal driving behavior and accident, in the traffic surveillance video and retrieve videos with similar events in the databases. In this research, two PCA-based algorithms, i.e., Eigenvehicle and PCA-SVM, are introduced to differentiate varieties of vehicle objects such as trucks, passenger cars, vans, and pick-ups. Further, an interactive Multiple Instance Learning (MIL) framework for semantic video mining and retrieval is used along with one-class SVM to find out the spatio-temporal patterns of user-interested moving vehicle behaviors from users’ feedback. The experimental results show the effectiveness of the proposed framework in detecting abnormal events such as illegal driving behaviors and accidents. By tightly integrating these key com-
ponents in a learning system, we ease the surveillance video retrieval problem (Chen, Zhang & Chen, 2007; Zhang, Chen & Chen, 2006)

B.2 Abnormal Events Detection in Indoor Surveillance Videos

In this work, we further extend the concept of content-based image retrieval to video event retrieval by successfully incorporating coupled hidden Markov models (CHMM), relevance feedback (RF), and multiple-instance learning (MIL) into a framework for retrieving various abnormal events, such as chasing, fighting, and robbery, in surveillance videos. These abnormal events involve not only the behavior of subjects but also the interactions between them.

In this research, we successfully incorporate CHMM with the content-based video retrieval framework to effectively model various abnormal events by representing the behavior of a person through one-chain and reflecting the interactions between two persons in the cross transitions between two chains. In addition, in order to provide a user-friendly semantic video retrieval system, the proposed framework adopts query-by-example (QBE) interface. Further, by incorporating the relevance feedback (RF) from the user, the learning system is able to guess the actual event of interest to some extent. To address this challenge, we incorporate Multiple Instance Learning (MIL) with relevance feedback in this framework.

The contribution of this work is that we successfully incorporate coupled hidden Markov models (CHMM) with relevance feedback (RF) and multiple-instance learning (MIL) for retrieving various abnormal events in surveillance videos, which, to the best of our
knowledge, is among the first effort to tackle the surveillance video retrieval problem this way (Zhang, Chen, Chen, Yang & Johnstone, 2010; Zhang, Chen, Zhou & Chen, 2009).

C. Biomedical Visual Information Retrieval

In this subsection, two previous directions related to image segmentation are discussed. In the microarray image research, we develop an innovative global-local thresholding scheme for microarray image segmentation. In the development of an automatic bacteria colony counter, Kolmogorov-Smirnov test, a statistical hypothesis test, is used in bacterial colony segmentation. We believe that, with proper extension, the thresholding scheme and the Kolmogorov-Smirnov test can be adapted and adopted in the proposed multi-resolution based image segmentation, as part of the future work.

C.1 Microarray Image Analysis

DNA microarray is an increasingly important technique that allows biologists to monitor expression levels of thousands of genes in parallel. This technique is widely used in biological research for studying genomic aberrations in cancer and other genetic diseases, and therefore, it has great potential for clinical diagnostics in the future. However, several critical problems remain unsolved with this technology. The first issue is related to automatic gridding and spot segmentation for microarray images. It has been reported that the quality of spot segmentation significantly influences data precision in the subsequent data analysis. However, nowadays microarray image analysis software still requires users’ fine tuning to
obtain acceptable results. Another important issue is how to automatically collect related information regarding all genes on the microarray slide for subsequent data analysis and data mining. To relieve researchers from manually correcting image processing results and manually collecting the related information for genes, we invented an automatic and robust method for microarray image analysis and the related information retrieval module which is integrated with our proposed database schema for managing microarray image data.

This study has far-reaching impacts: (1) we designed and implemented a Microarray Image Analysis (MIA) system, which provides users a convenient venue that enables automatic analysis of slide images and acquisition of accurate gene expression data from microarray slides; (2) the MIA system identifies and rectifies tilted image slides, discovers block boundaries, generates gridlines, recognizes spots, and finally extracts the accurate spot intensity values from the two image channels (red and green) in a microarray slide (Chen, Zhang, & Liu, 2007; Chen, Zhang, & Liu, 2006a; Chen, Zhang, & Liu, 2006b); (3) the MIA system, which is built on Unstructured Information Management Architecture (UIMA), provides a flexible, scalable, and extensible environment for users to perform various analysis tasks on microarray slide images (Chen & Zhang, 2007; Chen, Zhang, Liu, & Tiwari, 2007).

C.2 Automatic Bacteria Colony Counter

Bacterial colony enumeration has applications in many different assays such as antibiotic screening, toxicology testing, and genotoxicity measuring. The counting of bacterial colony is usually performed by well-trained technicians manually. Thus, this manual enu-
meration process has a very low throughput, and is time consuming and labor intensive in practice.

Our contributions in this study include: (1) we introduced a fully automatic yet cost-effective bacterial colony counter which can not only count but also classify colonies through image segmentation; (2) this automatic bacteria colony counter can significantly reduce the manual labor by automatically detecting the dish/plate region and extracting and counting colonies, both through image segmentation; (3) our proposed method can recognize chromatic and achromatic images and thus can deal with both colored and clear medium; (4) our automatic bacteria colony counter is software centered and can accept general digital camera images as its input, and therefore, the total cost is kept low comparing to other ad hoc automatic counters; (5) in order to differentiate colonies of different species, the proposed counter adopts one-class Support Vector Machine (SVM) with Radial Basis Function (RBF) as the classifier (Chen & Zhang, 2009a; Chen & Zhang, 2008); (6) our proposed counter demonstrates a promising performance in terms of both precision and recall, and is robust and efficient in terms of labor- and time-savings (Chen & Zhang, 2009b; Zhang, Chen, Liu, & Chen, 2008; Chen & Liu, 2008; Zhang & Chen, 2007).
CHAPTER 3
MULTIPLE-OBJECT IMAGE RETRIEVAL FRAMEWORK

A. Framework Overview

Our goals in this research include: (1) introducing an object-based content-based image retrieval (CBIR) system with a query-by-example (QBE) interface, (2) reducing the negative impact of inaccurate segmentation by developing novel algorithms which can efficiently represent images with hierarchical region trees and perform image similarity comparison through comparing the two corresponding hierarchical region trees, and (3) integrating machine learning with relevance feedback (RF) technique to automatically discover the objects/regions of the users’ interests, thus, enabling multiple-object retrieval in image databases. To achieve the abovementioned goals, we design a multiple-object image retrieval system (MOIR) in this research, and the high level architecture of the proposed multiple-object image retrieval (MOIR) framework is illustrated in Figure 10.

In brief, the proposed multiple-object image retrieval process starts with the user’s submission of a query sample image. Then, the proposed multi-resolution image analysis (MRIA) is performed on the query image to build a hierarchical region tree. In the next step, the proposed multiple-object image retrieval (MOIR) framework measures the similarity between the query image and each target image in the database, which is achieved by the comparison of two hierarchical region trees, representing the query image and target image, respectively. Then, target images are ranked in the decreasing order of their similarities to the
query image. According to the top 20 images in the ranked list, the user provides feedback to the retrieval system with the purpose to refine the retrieval results.

**Figure 10:** The high level architecture of the proposed multiple-object image retrieval (MOIR) framework which includes (1) user submission of a query sample image, (2) construction of hierarchical region tree with the proposed multi-resolution image analysis (MRIA), (3) measuring image similarity with the proposed hierarchical region tree comparison algorithm, and (4) using relevance feedback (RF) technique to refine the image retrieval results.

**B. Region Property Extraction**

In content-based image retrieval, image similarity is calculated on the basis of primitive visual features. The commonly used primitive visual features include color features, texture features, and shape features. In this research, we use color features because color is the most basic constituent element of visual content. Further, color features are relatively stable and robust to various image transformations, compared to other visual features.
Many early color-based image retrieval systems describe the entire image or a segmented region with mean color (Smith & Chang, 1996). Further, these systems use various distance measures such as Euclidean distance to describe the similarity between two regions, represented by two mean colors (Smith et al., 1996). However, these approaches might not be sufficient under some circumstances. For example, many image retrieval systems extract color features from the HSV color space because it is more pertinent to human visual perception. HSV color space represents a color with three components, i.e., hue, saturation, and value (luminance or brightness). As a directional distribution, the hue component organizes the visible light spectrum in a circular form as a color wheel, which is demonstrated in Figure 11.

With its current setup, the color wheel starts with red at the 0 degrees, followed by yellow at the 60 degrees, green at the 120 degrees, cyan at the 180 degrees, blue at the 240 degrees, and violet at the 300 degrees. In this color space, the hue component is measured on the range of $[0, 1)$, corresponding to 0 degree to 360 degrees on the color wheel. In a directional distribution, 0 degree and 360 degrees both represent the same hue, i.e., red, on the color wheel. It is thus not suitable to use mean hue values to represent color features.

For instance, assume that a segmented region consists of two colors, red and cyan, where the hue value of red is 0 and that of cyan is 0.5. By taking the average of the hue values, we obtain 0.25 which corresponds to a totally different color - green. This exemplifies that the mean hue value is not appropriate in representing the color feature of a region.

Another example is given below to demonstrate the inappropriateness of using Euclidean distance in the HSV color space. Assume that we measure color similarity between red and one of the following two colors: violet-red, and orange. The hue values for red, violet-red, and orange are 0, 0.9, and 0.1, respectively. The Euclidian distance between red and
violet-red is 0.9, while the distance between red and orange is 0.1. However, the actual distance between red and violet-red should also be 0.1 on the color wheel. This indicates that it is not suitable to directly use a conventional distance measure on a directional distribution such as hue.

Figure 11: The color wheel demonstrates the non-interval quantization of hue used to extract dominant color descriptor from images.
In order to avoid the aforementioned problems in directional distribution, one of the possible solutions is to use histogram to represent the color property of regions. In this research, we have tested and compared the effectiveness of two histogram-based approaches commonly used in the object-based image retrieval. These two approaches include: color code histogram and the MPEG-7 dominant color descriptor.

B.1 Color Code Histogram

Color histogram describes the distribution of the various colors in an image, which is a very effective representation of that image for the purpose of image retrieval or object recognition (Shapiro & Stockman, 2001). An image color histogram is composed of tabular frequencies erected over discrete intervals, also known as bins, each of which represents a specific color range. Using color histogram to represent an image is relatively stable since color histogram is invariant to various transformations such as translation, rotation about the imaging axis, small off-axis rotations, scale changes, and partial occlusion (Shapiro et al., 2001).

6-bit color code histogram, also known as 64-bin color code histogram, is one of the widely used color histogram techniques. The 6-bit color code histogram is originally proposed by Swain & Ballard in 1991 for use in object recognition (Swain & Ballard, 1991). To construct a 6-bit color code histogram, we first convert each pixel in a color image into a 6-bit color-code by taking the 2 most significant bits of each of R, G, and B color components. This process replaces similar colors within a range by a single value, and transforms a RGB
image to an index image with $2^6 = 64$ color codes. Then, a 64-bin histogram can be created on the basis of the normalized frequency of each color code.

Assume that two 6-bit color code histograms are denoted as $A = \{a_0, a_2, ..., a_{63}\}$ and $B = \{b_0, b_2, ..., b_{63}\}$, respectively. The distance function ($D$) between the two 6-bit color code histograms representing two image regions is defined as:

$$D(A, B) = \sqrt{\sum_{i=0}^{63} (b_i - a_i)^2}$$

The similarity score can be further calculated by normalizing the distance value between the two 6-bit color code histograms with the maximum distance, i.e., $\sqrt{2}$, in the feature space.

$$\text{similarity}(A, B) = 1 - \left( \frac{D(A, B)}{\sqrt{2}} \right)$$

Figure 12 demonstrates an original color image on the left, and the 6-bit color code histogram of the original color image is shown on the right.
B.2 Dominant Color Descriptor

Another commonly used histogram-based approach is the MPEG-7 dominant color descriptor (Shao, H., Wu, Y., Cui, W., & Zhang, J., 2008). MPEG-7 dominant color descriptor is efficient and effective in describing the entire or a portion of an image with representative color distributions. In this research, we adopt Shao et al.’s implementation to extract dominant color descriptor features from images (Shao et al., 2008). The feature extraction process is briefly described as follows.

1. Converting all color values to HSV color space.
2. Reduce the total number of colors to enhance the efficiency by encoding the true color space into 72 distinct quantized colors $C$ with the following non-interval quantization algorithm.

$$C = 9 \times H + 3 \times S + V$$
where $c \in [0, 71]$, $h \in [0, 360^\circ]$, $s \in [0, 1]$ and $v \in [0, 1]$. The non-interval quantization of the hue component is illustrated in Figure 11.

3. Build a 72-bin histogram based on coded colors. Each bin in the histogram represents the frequency of a coded color. Then, histogram normalization is performed, which produces the percentage for each coded color.

4. Rank coded colors in descending order of their percentage values, and the top $K$ coded colors are selected as the dominant colors, while the non-dominant colors are no longer considered. In this paper, we set $K=8$ according to Shao et al.’s implementation.

5. Normalize the percentage values of the $K$ dominant colors by the sum of their values, and then, set the corresponding percentage value to zeros for all non-dominant colors,
which results in a 72-dimensional feature vector as the dominant color descriptor.

The similarity measure between two dominant color descriptors, such as \( A = \{a_0, a_1, \ldots, a_{71}\} \) and \( B = \{b_0, b_1, \ldots, b_{71}\} \), can be calculated with the following formula.

\[
\text{similarity}(A, B) = \sum_{l=0}^{71} \min(a_{l}, b_{l})
\]

We conducted a preliminary study on the basis of the full image search in order to compare image retrieval performance of the above two color features in terms of their mean average precision (MAP) measure which is defined and detailed in Section D, Chapter 4. Based on 201 multiple-object queries, the MAP value of the MPEG-7 dominant color descriptor is 10.78\%, and that of the 6-bit color code histogram is 8.57\%. The results of the experiment which is detailed in Section E.1, Chapter 4, show that the MPEG-7 dominant color descriptor significantly outperforms 6-bit color code histogram in the full image retrieval. For this reason, the proposed multiple-object image retrieval (MOIR) framework adopts MPEG-7 dominant color descriptor as the primitive visual features to avoid the aforementioned problems in directional distribution.

C. Hierarchical Image Representation

In object-based image retrieval, it is crucial to extract meaningful objects from images. This object extraction process is known as image segmentation. In an ideally segmented
image, each segmented region represents a meaningful object. Unfortunately, there is no per-
fekt segmentation algorithm, and image segmentation is recognized as one of the most com-
plicated tasks in image processing and still remains unsolved. An inaccurately segmented
image suffers from over- and/or under-segmentation problems. Over-segmentation refers to a
segmentation algorithm that partitions a meaningful object into two or more regions. Contra-
rily, under-segmentation refers to a segmentation that lacks the ability to differentiate be-
tween two or more meaningful objects, resulting in a segmented region that contains two or
more objects.

Both over-segmentation and under-segmentation have negative impacts on object-
based image retrieval. In general, object-based search is achieved through matching two sets
of segmented regions, representing a query image and an arbitrary target image, respectively.
This process may suffer from the incorrect region matching due to objects in images being
over-segmented and/or under-segmented. Under-segmented objects usually cause a bigger
problem than over-segmented objects because in the case of over-segmentation, search algo-
rithms may still be able to find the object of interest through matching part(s) of the object,
i.e., regions corresponding to an over-segmented object. An under-segmented object, howev-
er, can be rarely matched in this way. Thus, most existing image segmentation algorithms
tend to over-segment an image.

Since image segmentation quality has a direct impact on the retrieval accuracy in ob-
ject-based image retrieval, in order to reduce its impact, researchers introduce hierarchical
image representation to preserve the relationship among segments (Arbeláez, Maire, Fowlkes,
& Malik, 2009; Ahuja & Todorovic, 2008; Burt, Hong, & Rosenfeld, 1981). The hierarchical image
representation is a flexible and convenient way to mirror the multi-scale processing in the human visual system.

The conventional approach to hierarchical image representation is a two-step process, including image segmentation followed by region tree construction. The first step is to perform segmentation on images presented in different resolutions from the highest (the original image) to the lowest, producing a segmentation mask for each resolution. The second step is to construct the hierarchical representation of the image, i.e., a region tree, by associating segments from different resolutions. However, these multi-level analysis approaches suffer from a high computational complexity. First, performing a full-scale segmentation at each different image resolution is itself complicated enough, let alone the need of one extra run through all resolutions to associate segmented regions. One of our goals in this research is to design a novel hierarchical image segmentation algorithm that possesses the following characteristics:

1. Preserving the spatial relationships among segmented regions in a hierarchical region tree that represents an image;
2. Performing image segmentation and hierarchical region tree construction in a concurrent manner to reduce the computational complexity;
3. Including a branch-and-bound-like algorithm that performs image analysis from low resolution to higher resolution in order to mitigate the inefficiency problem during the multi-level analysis.
In this dissertation, a multi-resolution image analysis (MRIA) algorithm is proposed that performs hierarchical image segmentation with the above desired characteristics. The proposed multi-resolution image analysis (MRIA) algorithm is inspired by the human visual system.

Imagine you are standing on an open field and a red sports car is moving toward you at a very far distance. Initially, your eyes can only see a tiny red object without any detail due to the visual acuity of the visual system. When the tiny red object is moving closer, your visual system is able to recognize the object as a red sports car but still cannot capture fine details of the car. Later, when the car approaches close enough, your eyes can distinguish fine details of the car such as the vehicle brand logo and the textures of wheels.

The above observation indicates that our visual system has limited resolving power and our brain only recognizes an object when our visual system provides enough details, the combinatorial of various primitive visual features, about the object being observed. This phenomenon also implies that when an object is located at a far distance, our visual system can only perceive down-sampled signals from the object. In other words, the human visual system cannot capture enough details about that object until the sampling rate climbs up to a certain level. The entire process reflects that the human brain actually performs a multi-resolution analysis through our visual system, which motivates us to adopt a similar multi-level analysis process into the proposed multi-resolution image analysis (MRIA) algorithm.

In signal processing, down-sampling is known to be a process that removes bandwidth in high-frequency and preserves bandwidth in low-frequency in data. Therefore, the most prominent regions in an image can be obtained even with a low sampling rate, while the detailed information can be revealed at higher sampling rates.
Figure 13 exemplifies a series of multi-resolution images where (a) is the original image, (b) is 1/2 down-sampled image of the original image in each dimension, (c) is 1/4 down-sampled image of the original image in each dimension, (d) is 1/8 down-sampled image of the original image in each dimension, and (e) is 1/16 down-sampled image of the original image in each dimension. In Figure 13 (a), the high-frequency signals such as the black and white stripes on the zebra are distinct. After a series of down-samplings, the black and white stripes on the zebra become blurred in (c), and totally vanish in (d) & (e). This indicates that the high-frequency signals are removed from the image in the process of down-sampling.

Figure 13: A series of multi-resolution images where (a) is the original image, (b) is 1/2 downsampling image of the original image at each dimension, (c) is 1/4 downsampling image of the original image at each dimension, (d) is 1/8 downsampling image of the original image at each dimension, and (e) is 1/16 downsampling image of the original image at each dimension. This figure also exemplifies that the high-frequency signals, such as the black and white stripes on the zebra, are removed from the image in the down-sampling process.
In general, the most prominent regions in images usually indicate either backgrounds or a target object in close-up shot. If we progressively increase the sampling rate, more and more details will be become evident for each prominent region discovered previously. The process of gradually increasing the sampling rate naturally forms a region-based hierarchical tree with different levels of details. Moreover, as an added benefit, performing analysis on low resolution images is much more efficient than that of the high resolution images.

The flowchart of the proposed multi-resolution image analysis (MRIA) is depicted in Figure 14. As aforementioned, one of our goals in this research is to mitigate the inefficiency of the multi-level analysis. To achieve this goal, our approach is to design a branch-and-bound-like algorithm that performs image analysis from the lowest resolution and progresses to higher resolutions only if necessary.
Figure 14: The proposed multi-resolution image analysis (MRIA) framework for hierarchical image representation.
Apparently, the analysis of a low resolution image is much faster than that of a high resolution image. Therefore, the first step in the proposed algorithm is to obtain down-sampled images. Discrete wavelet transformation (DWT) is known to be an efficient method to transform the original image into a series of down-sampled images. In this research, Haar wavelet transform is used to produce a series of down-sampled images by reducing image size by half in each dimension (Haar, 1910) each time it gets down-sampled. In addition, a minimal image size of 8-by-8 pixels is preset as a constraint because any image smaller than this preset size will be too coarse to differentiate meaningful objects.

The second step starts with creating a root node to represent the entire image at the lowest resolution and extracting primitive visual features from the lowest resolution image. By considering the entire image as one region, the first level image segmentation is performed on the region, producing a segmentation mask. It is worth mentioning that any image segmentation algorithm can be used and seamlessly integrated into the proposed multi-resolution image analysis (MRIA) algorithm. In this research, we adopt an image segmentation algorithm which is implemented on the basis of the renowned seeded region growing algorithm (Shih, F. T. & Cheng, S., 2005; Adams, R. & Bischof, L., 1994) because of its efficiency and efficacy. The details of the proposed region-growing based image segmentation are presented in the next section (Section D, Chapter 4).

In the next step, we increase the image resolution and up-scale the segmentation mask in order to obtain more details about each segmented region. For each segmented region, a child node is created to represent that region, and primitive visual features are extracted from the region, followed by the second level image segmentation on each region produced by the first level segmentation. With more details revealed for a region at each higher resolution, the
subsequent higher-level segmentation plays an important role in determining the homogeneity of the region. More specifically, a region is considered homogeneous if no new segment can be segregated from that region, indicating that there is no need to further process this region in the subsequent analysis. On the other hand, for a region that is not sufficiently homogeneous will likely be further segmented into smaller segments at a higher image resolution, and a new segmentation mask will be produced for that region. This process will continue until either of the following criteria is satisfied. The two stopping criteria are: (1) no region can be further segmented, or (2) the highest image resolution is reached.

However, about 4 percent of images in our dataset cannot be properly processed using the aforementioned multi-resolution image analysis (MRIA) algorithm. The reason is that those images usually contain highly similar foreground and background such that the multi-resolution image analysis stopped at the lowest image resolution due to the fact that no region can be further segmented.

Figure 15, for instance, depicts a town with mud houses and a few trees. The original image (size: 384×256) is displayed in (a), and a series of images at different resolutions is demonstrated in (b) to (f). In the multi-resolution images, (b)-(e) are intermediate resolution images, and (f) is an image at the lowest solution (12×8). We can observe from these images that image details below certain resolution are almost completely removed and the differences between foreground and background are invisible, in particular (e) and (f). This example suggests that in some images, objects such as the foreground and the background cannot be well differentiated at low image resolutions, and thus, it is not reasonable to immediately stop the multi-resolution image analysis process if segmentation at the current resolution exposes no object at all.
In order to cope with this issue, we slightly modify the process of multi-resolution image analysis by relaxing the stopping criteria for processing images. Specifically, the modified algorithm will not apply the criterion of “no region can be further segmented” until the total number of segmented regions in the image is ≥ 2. This provides a chance for objects invisible to a low resolution analysis to be extracted at a higher resolution where more details become available.

Figure 15: A sample image that contains highly similar foreground and background. This example shows the same image at multiple-resolutions that contains the scene of a town with mud houses and a few trees. (a) is the original image (384×256); (b)-(e) are intermediate resolution images, and (f) is image at the lowest solution (12×8). In (e) and (f), images details are almost completely removed and the differences between foreground and background are invisible.

D. Seeded Region Growing Segmentation

Seeded region growing, a well-known pixel-based image segmentation algorithm, expands a region from a set of seed pixels by comparing a seed pixel with its immediate neighbors and gradually adding similar neighbor pixels into the current growing region. A region
is stopped growing when no spatially connected pixel can be further added into the region. The image segmentation process is completed when each pixel in the image has been assigned to one of the regions.

The seeded region growing approach requires a set of seed pixels provided by the users or procured from an automatic seed selection algorithm (Shih et al., 2005). In addition, the algorithm needs to determine the cutoff value in terms of pixel similarity on the basis of the connectivity or pixel adjacency information. Further, a minimum region restriction may be adopted in the region growing algorithm. Applying seeded region growing approaches to segment images often produces homogeneous regions which correspond well to the observed edges.

In this research, we propose a modified seeded region growing algorithm which enables the segmentation of an entire image (rectangular shaped) or a sub-image with an arbitrary shape. Given an arbitrarily shaped sub-image \( R_0 \), the goal is to divide \( R_0 \) into a set of smaller homogeneous regions if there is any. Initially, all pixels in \( R_0 \) are labeled as zeros indicating that these pixels do not belong to any region. In order to grow a new region \( R_i \) \((i \in \mathbb{Z}^+)\), an arbitrary zero-labeled pixel is selected as the seed \( S \). For efficiency and simplicity, the proposed algorithm chooses the first zero-labeled pixel in \( R_0 \) according to the column-major indexing of pixels. A queue \( Q \) is created for storing seeds, started with the insertion of \( S \). The first seed \( s \) in \( Q \) is retrieved for expanding the current region \( R_i \). The pixel corresponding to \( s \) is labeled with a special label ‘\( L \)’, indicating that the pixel is assigned to the new region \( R_i \). Subsequently, we compute the similarity value \( v \) between \( s \) and each of its immediate neighbor pixels. If a similarity value \((v)\) is greater than or equal to the pre-defined cutoff value \((\text{cutoff})\), the algorithm assigns the corresponding immediate neighbor pixel to the new re-
region $R_i$ by labeling the pixel with ‘$L$’. Further, the pixel is also considered as a new seed and added in $Q$. In this research, we empirically determine a cutoff value ($cutoff$) for each different image resolution as follows: $cutoff = 80\%$ for Level 1 image resolution (the lowest image resolution); $cutoff = 90\%$ for Level 2 image resolution; $cutoff = 97\%$ for Level 3 image resolution or higher. Again, a new seed $s$ is retrieved from $Q$ for further expanding the current region $R_i$. This process continues until $Q$ is empty. Then, all pixels labeled as $L$ are returned as the region $R_i$.

Sometime, the size of a new region is too small to likely represent a meaningful object. For this reason, the returned region $R_i$ is considered valid only if its size exceeds a predefined threshold value. The threshold value of size is empirically determined with the following equation.

\[
\text{threshold value} = \left(0.07 \times \sqrt{Iw^2 + Ih^2}\right)^2
\]

where $Iw$ and $Ih$ represent the image width and image height at the current image resolution. For a valid region $R_i$, the corresponding pixels are then labeled with a positive integer $i$, representing the newly formed region. Otherwise, the corresponding pixels are labeled with negative integer $-i$.

The above process continues until all pixels in $R_0$ are non-zero labeled. To this point, the sub-image $R_0$ has been divided into positive labeled and negative labeled region. Recall that some regions may be labeled as “$-i$” because of their small size. In order to handle these negatively labeled regions, we measure the similarity between a negatively labeled region and each positively labeled region. Then, the negatively labeled region is assigned to the pos-
itively labeled region to which it has the highest similarity in terms of the dominant color descriptor. The flowchart of the proposed region growing algorithm is illustrated in Figure 16.

Figure 16: The flowchart of the proposed modified seeded region growing algorithm.
In summary, the proposed MRIA algorithm first segments an image at the lowest resolution, and performs subsequent segmentation for each previously generated region only when necessary, i.e., when that region is not sufficiently homogeneous. During the same process, a hierarchical tree representation is constructed (in a top-down manner) along with the multi-resolution segmentation results. The key point in this process is to avoid unnecessary image segmentation at any higher image resolution – if a sub-tree/branch, which represents a region in the image, is considered homogenous, it will be removed from any subsequent segmentation (pruning of the analysis space). In this way, the computational cost can be dramatically reduced.

Although we can use the image hierarchy to preserve the associative relations among regions, the negative impact of over-segmentation still remains unsolved for object-based image retrieval until we make use of the hierarchical tree matching in the image retrieval process. In the next step, we utilize the preserved associative relations to alleviate the over-segmentation problem in retrieval by introducing the hierarchical region tree matching.

**E. Hierarchical Region Tree Matching**

With the proposed multi-resolution image analysis (MRIA) algorithm, the query image and all target images in the database are segmented into regions at multiple resolutions. For each image, the relations among those segmented regions are concurrently preserved in the form of a hierarchical region tree. As aforementioned, an image hierarchy reflects that image’s visual composition, and thus, provides a way to model not only visually but spatially the visual content of that image. Figure 17 provides some examples of hierarchical region
trees. A typical hierarchical tree consists of three types of nodes including one root node ($R$), leaf nodes ($L$), and inner nodes ($I$). The root node represents the entire image as a single region. A leaf node represents a region with consistent visual features and cannot be further partitioned into sub-regions in that feature space. An inner node represents a region that consists of at least two sub-regions. In this dissertation, we refer to a sub-tree of a tree $T$ as a tree consisting of a node and all of its descendants in $T$. Thus, the sub-tree corresponding to the root node is the entire tree; the sub-tree corresponding to any inner node ($I$) in $T$ is defined as a proper sub-tree ($P$). For each proper sub-tree ($P$) or leaf ($L$) in a hierarchical image representation, it can represent multiple objects, a single object, or part of an object.

Figure 17: Four examples of hierarchical region trees. The hierarchical region trees from left to right represent a query image, an over-segmented image, an ideally segmented image, and an under-segmented image, respectively.

Figure 17 demonstrates four hierarchical region trees $T_1, T_2, T_3$ and $T_4$ which model the content of a query image ($T_1$) and three target images ($T_2, T_3$ and $T_4$) in the database, respectively. As shown in Figure 17, symbols $R$, $L$, and $I$ represent the root node, a leaf node,
and an inner node, respectively. Numbers in the subscripts indicate the ordinal value of a specific type of node ($L$ or $I$), at that level. The numbering of ordinal values restarts from 1 at each new level. The corresponding regions from different hierarchical region trees, i.e., from different images, have the same color.

Traditional CBIR frameworks, such as SIMPLIcity, measure object relevance through the comparison of two sets of objects without considering the relationships among regions in an image (Wang et al., 2001; Wang et al., 1999). Unlike the conventional CBIR frameworks, using hierarchical region tree in the proposed object-based CBIR system provides additional information on the relationships among the regions in an image and is expected to reduce the negative impact of inaccurate segmentation, especially over-segmentation. Taking into account the relationships along with individual image regions allows the proposed CBIR framework to better measure the similarity between two regions (not restricted to leaf nodes only) from two images. This idea transforms the object comparison problem into proper sub-tree comparison.

As aforementioned, image segmentation is an extremely difficult problem. Although an object may be ideally-segmented, quite often an object suffers from over-segmentation and/or under-segmentation problems. An ideally-segmented object corresponds to a leaf node in a tree, but a leaf node may represent an under-segmented region which contains two or more objects. An over-segmented object corresponds to an inner node in a tree. Figure 17 depicts an over-segmented object in $T_2$, an ideally-segmented object in $T_3$, and an under-segmented region in $T_4$, as elaborated below.

For convenience, we will use a shorthand representation to refer to a node in a tree throughout the rest of this paper. The shorthand representation is defined as:
The root node is at Level 1. For example, when we refer to the inner node \((I_1)\) located at Level 3 in tree \(T_2\), the shorthand representation of this node is \((2, 3, I.1)\).

In Figure 17, as indicated by the same color, \((2, 3, I.1)\) in \(T_2\) corresponds to the same object ideally segmented in \(T_1\) \((1, 3, L.2)\) and \(T_3\) \((3, 4, L.1)\), but is further partitioned into \((2, 4, L.1)\) and \((2, 4, L.2)\) in \(T_2\). This indicates that this object in \(T_2\) is over-segmented. \((1, 3, L.1)\) and \((1, 3, L.2)\) represent two ideally segmented objects in tree \(T_1\). However, the corresponding nodes do not exist in \(T_4\). This is because that the node \((4, 3, L.1)\) in \(T_4\), which should correspond to the node \((1, 2, I.1)\) in \(T_1\), is under-segmented. In other words, two objects \((1, 3, L.1)\) and \((1, 3, L.2)\) are both included in one region \((4, 3, L.1)\) in \(T_4\) but they cannot be separated by segmentation on that image. For illustration purposes, we depict these two nodes from \(T_1\) in \(T_4\) with red dotted circles as \((4, 4, L.1)\) and \((4, 4, L.2)\), though they do not exist in \(T_4\). Although the nodes \((1, 3, L.1)\) and \((1, 3, L.2)\) probably cannot be matched with any node in \(T_4\), their parent node \((1, 2, I.1)\) can still be matched to \((4, 3, L.1)\).

From the above examples, three types of comparison can be concluded, including leaf to leaf (L-L) comparison, leaf to sub-tree or sub-tree to leaf (L-P/P-L) comparison, and sub-tree to sub-tree (P-P) comparison. The above three types of comparisons are actually performed through measuring the similarity between their primitive visual features. The L-L comparison measures the similarity between two segments which correspond to two leaf nodes. The L-P/P-L comparison simply measures the similarity between a segment that corresponds to a leaf node and a region consisting of a set of segments that correspond to a sub-
tree. The P-P comparison calculates the similarity between two sets of segments that correspond to two sub-trees, respectively.

We expect that the similarity measure derived from the above three types of comparisons can reduce the negative impact of over-segmentation. This is because when matching two objects that either or both are over-segmented, the optimal object matching can still be achieved through a L-P/P-L or P-P comparison. However, we are not very optimistic about using hierarchical region trees to alleviate the problem of under-segmentation. Our take on this is that most existing image segmentation algorithms, especially those used in object-based image retrieval systems, tend to over-segment an image so that the retrieval performance is largely affected by over-segmentation (Carson et al., 2002). Thus, we argue that by alleviating the problem of over-segmentation, the state-of-the-art of multiple-object image retrieval can be advanced. In this research, we make sure that the proposed hierarchical image segmentation algorithm tends to over-segment an image but is bounded by an acceptable rate of such.

A performance issue in terms of efficiency also emerges from the aforementioned comparisons. This is because there could be many sub-trees in one hierarchical region tree, not to mention when comparing all proper sub-tree pairs from a given pair of trees. For this reason, an efficient algorithm for matching two hierarchical region trees is required. In this research, an efficient approach, namely hierarchical segmentation tree matching (STM) algorithm, for matching two proper sub-trees is developed. In order to avoid excessive computational cost in proper sub-tree comparison, our idea is to calculate the sub-tree similarity based on previously calculated similarity values during subsub-tree comparison, similar to the idea
of dynamic programming. We use the following example to explain the proposed proper sub-tree comparison algorithm.

![Tree A (Query Image)](image1) ![Tree B (Target Image)](image2)

Figure 18: Matching two hierarchical region trees – A and B, representing a query image (A) and a target image (B), respectively.

Figure 18 exemplifies two hierarchical region trees – A and B, representing a query image (A) and a target image (B), respectively. In matching two hierarchical region trees, our goal is to find, for each node in A, its best matching node in B. Recall that when building a region tree, all nodes are inserted in the order of top-to-down and left-to-right. In order to reuse the previously calculated similarity values, the tree comparison is performed in the reverse order.

The comparison starts from matching the rightmost node in A, which is the leaf node (A7), with each node in B. In this round of matching, the 3 L-L comparisons, i.e., A7-B5, A7-B4, and A7-B3, come up first and execute in that order. After that, there are 2 L-P compari-
sons, i.e., A7-B2 and A7-B1 (in that order). When performing a L-P comparison, the similarity between a leaf node and a sub-tree is defined as the highest similarity between the leaf node and a node in the sub-tree (including the root node of that sub-tree). However, there is no need to rematch the leaf node in the query image with every child node in that sub-tree of B. In fact, according to our reverse order of comparison, the similarity between that leaf node in A and every child node in the sub-tree of B has been previously calculated already.

Following the same process, the comparison continues and at a later time reaches the matching of an inner node (A3) with each node in B. In this round of matching, there are 3 P-L comparisons, i.e., A3-B5, A3-B4, and A3-B3. Following that, there are 2 P-P comparisons, i.e., A3-B2 and A3-B1. In each P-L comparison involved in this step, since the similarity between each child node of A3 and each leaf node of B has been calculated already during previous steps, there is no need to calculate them again, and the only additional computation incurred is the calculation of similarity between A3 itself and that leaf node in B.

When comparing two proper sub-trees such as A3-B2, we first measure the inner node similarity (INS) which is defined as the similarity between the two root nodes of the two sub-trees. If the inner node similarity exceeds a predefined threshold value (≥ 90% similarity in our case), we further measure the highest child node similarity (CNS) between the two sub-trees. It is worth noting that the CNS can be directly derived from the child node similarity scores calculated in previous steps. The proper sub-tree similarity (PSS) is defined as the maximum of INS and CNS as formalized in the following equation.

\[ PSS = \max\{INS, CNS\} \]
Assume there are $m$ target images in the database. The similarity value between the query image and each target image can be efficiently measured using the proposed hierarchical region tree comparison algorithm, resulting in a vector of length $n$, where $n$ is the number of nodes in the query image.

The collection of the aforementioned vectors forms a matrix of size $m$-by-$n$, namely the node similarity matrix. Each row in the matrix represents a target image, and each column heading in the matrix corresponds to a node in the query image. An entry $[m_i, n_j]$ records the highest similarity value between the $n_j^{th}$ node in the query image and all the nodes in the hierarchical tree of the $m_i^{th}$ image. According to the similarity scores stored in the node similarity matrix, the proposed multiple-object image retrieval framework can obtain the overall similarity by calculating the row sum, returning a ranked list of images to the user as the initial retrieval results. In addition, the node similarity matrix is used in the subsequent user relevance feedback process which progressively discovers the object(s) of the user’s interests.

F. Relevance Feedback

In addition to the development of an efficient sub-tree similarity measure, another challenge remaining in the domain is how to discover the objects of the user’s interest given the user’s scarce and imbalanced feedback information as training data. We also want to avoid the proper sub-tree comparison during feedback iterations due to the expensive computational cost of sub-tree matching. For these reasons, our idea is to build a classifier that makes the maximum use of users’ relevance feedback, learns user-desired object(s) from the node similarity matrix and user feedback, and refines the retrieval results.
To achieve this goal, the first step is to collect the user’s feedback on the retrieval results. As aforementioned, the proposed multiple-object image retrieval (MOIR) framework calculates the row sum from the node similarity matrix which represents the overall similarity between the query image and each target image. The multiple-object image retrieval (MOIR) system ranks the target images in the descending order of their similarity to the query image, and returns the top 20 images (as the initial results) to the user for feedback. The user then provides feedback on those 20 images by assigning each either a positive or a negative label. A positive label is given if and only if the image contains all the objects of the user’s interest. Otherwise, a negative label is provided. The user’s feedback is then used by the retrieval system to learn his object(s) of interest. Since only 20 images are returned to the user for feedback, the amount of feedback information is scarce in nature and can be extremely imbalanced (e.g., only 2~3 images are positive among the top 20). However, returning more images for user feedback could bring a big burden to the user.

The second step is to associate the user’s feedback with the node similarity matrix. Recall that in the node similarity matrix, each column heading represents a node in the query image, and each row represents a target image. Since the user-desired object(s) must exist in the query image, one or more columns in the node similarity matrix represent the object(s) of the user’s interest.

However, it is not a trivial task to directly identify the relevant column(s), i.e., relevant object(s), in the node similarity matrix due to scarce feedback information. Instead of directly identifying the relevant column(s), we propose to adopt one-class support vector machine (SVM) (Schölkopf, Platt, Shawe-Taylor, Smola, & Williamson, 1999) to build a classifier and let the classifier determine the importance of each column/object in the query image.
The idea is that we consider each row in the node similarity matrix as a feature vector used for SVM training, representing the similarity between the query image and a target image in terms of object similarity. Further, we use the user’s feedback on each returned top target image as a class label. All positive samples belong to one class which represents relevant images while all negative samples belong to another class which represents irrelevant images. Then, a set of distinct target images with the user’s feedback are cumulatively collected as training samples through each feedback iteration.

The training samples are fed to the one-class SVM to train the classifier. This trained classifier is then used to test the relevance of all target images in the database and rank them according to their decision values generated from the SVM classifier. In this way, we can progressively refine the retrieval results by maximizing the usage of all of the user’s feedbacks collected through multiple iterations without sacrificing the efficiency because there is no need to recalculate the node similarity matrix.

In summary, by addressing and attempting to solve the above challenging issues, we expect that the development of the multi-resolution image analysis (MRIA) will provide us with an efficient tool to simultaneously produce image segmentation results and hierarchical region-tree representations, which are typically obtained through two separate processes with existing approaches. The hierarchical image representation is expected to alleviate the object matching problem due to the negative impact of over-segmentation. Further, with such a hierarchical representation, the relevance of a target image to the query image, in terms of their object similarity, is thus measured according to their proper sub-tree similarity. In the proposed framework, we also design and develop an efficient strategy to compare proper sub-trees. An innovative relevance feedback scheme is also proposed to bridge the semantic gap,
providing the capability for the system to learn the object(s) of the user’s interest with a very small and unbalanced training set, which maximizes the usage of the user’s feedback in query refinement and avoids the expensive proper sub-tree comparison. By means of the seamless integration of users’ relevance feedback (RF) with the proposed multiple-object image retrieval (MOIR) system, it allows automatic discovery of the objects of the user’s interest and improves the retrieval accuracy through feedback-retrieval loops, as demonstrated in next chapter.
CHAPTER 4

EVALUATION METRICS AND EXPERIMENTAL RESULTS

A. Dataset Description

The Corel Photo CDs are an image database that has been commonly and widely adopted by CBIR researchers for image retrieval evaluation. For examples, many recent papers published in the top-tier journals and conferences use Corel image database for image retrieval evaluation (Wu, Jin, & Jain, 2012; Wang, Hoiem, & Forsyth, 2012; Liu, Li, Zhang, & Xu, 2011; Zhang, Xiang, Zhou, Ye, & Mu, 2011; Liu, Zhang, Hou, Li, & Yang, 2010). The Corel image dataset contains more than 800 theme categories each of which is composed of 100 images. The image collection in Corel image dataset is mostly natural scene images. The experiments in this dissertation are performed on a subset of images collected from Corel Image Database.

The experimental dataset consists of 10,000 images originated from 100 theme categories in Corel Photo CDs. It is worth mentioning that we use Corel set in a different way from conventional methods. Unlike the traditional way where Corel theme category labels are used as ground-truth, we procure our own ground-truth for evaluation. Specifically, we define 50 objects and manually annotate images containing these objects. Many of these objects (e.g., blue sky, red car, snow, and roadway) occur in multiple Corel theme categories. Our ground-truth labels are these manually annotated objects instead of Corel theme category labels. Table 1 demonstrates the number of images for each object of interest used in the experiments.
Table 1: The number of images for each object of interest

<table>
<thead>
<tr>
<th>Object</th>
<th># of Image</th>
<th>Object</th>
<th># of Image</th>
<th>Object</th>
<th># of Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>penguin</td>
<td>66</td>
<td>Tiger</td>
<td>102</td>
<td>dinosaur</td>
<td>101</td>
</tr>
<tr>
<td>pyramid</td>
<td>24</td>
<td>blue sky</td>
<td>1,810</td>
<td>snow</td>
<td>584</td>
</tr>
<tr>
<td>bonsai</td>
<td>100</td>
<td>red bus</td>
<td>53</td>
<td>white rabbit</td>
<td>16</td>
</tr>
<tr>
<td>red car</td>
<td>109</td>
<td>Roadway</td>
<td>319</td>
<td>yellow car</td>
<td>34</td>
</tr>
<tr>
<td>yellow bus</td>
<td>23</td>
<td>Shoji</td>
<td>18</td>
<td>red airplane</td>
<td>28</td>
</tr>
<tr>
<td>bullet</td>
<td>23</td>
<td>Gun</td>
<td>76</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

B. Complexity Analysis

B.1 Multi-Resolution Image Analysis (MRIA) Algorithm

For an original image of size $m \times n$, assume that the number of down-sampling levels is $k$. The size of the down-sampled image at the lowest resolution (level-1) is $a \times b$ where $a = \frac{m}{2^k}$ and $b = \frac{n}{2^k}$.

For the proposed multi-resolution image analysis algorithm, the worst case is that we need to segment the entire image for all of the down-sampled images. In the worst case ($k \to \infty$), the total number of pixels needed for processing is

$$a \times b + 2^1 a \times 2^1 b + 2^2 a \times 2^2 b + \cdots + 2^k a \times 2^k b$$
Introducing $a = \frac{m}{2^k}$ and $b = \frac{n}{2^k}$ into the above equation, we can obtain the following equation.

$$\frac{m}{2^k} \times \frac{n}{2^k} + \frac{m}{2^{(k-1)}} \times \frac{n}{2^{(k-1)}} + \frac{m}{2^{(k-2)}} \times \frac{n}{2^{(k-2)}} + \cdots + \frac{m}{2^0} \times \frac{n}{2^0} \approx O(1.3mn)$$

The above equation indicates that the total number of pixels needed for processing in the worst case is $O(1.3mn)$.

Based on our experiment, the maximum value of $k$ is 6, and the proposed multi-resolution image analysis (MRIA) algorithm stops by level 3 for more than 57% images and by level 4 for more than 98% images. This indicates that the proposed multi-resolution image analysis (MRIA) algorithm is very efficient in performing image segmentation and tree construction since in most cases the proposed multi-resolution image analysis (MRIA) algorithm only needs to process about 2% (level 3) to 8% (level 4) of the original image size. We also experimentally analyze the complexity of the proposed multi-resolution image analysis (MRIA) algorithm in Section C.1, Chapter 4.

B.2 Hierarchical Region Tree Comparison Algorithm

As mentioned in Section E, Chapter 3, there are three types of tree node comparisons, including: leaf to leaf comparison (L-L), sub-tree to leaf comparison (P-L), and sub-tree to sub-tree comparison (P-P). The proposed hierarchical region tree comparison algorithm performs comparison not only on L-L but also on P-L and P-P whose complexity is determined
by the number of inner nodes. Given the same number of leaf nodes, a binary tree has the
most inner nodes, thus the worst case is the comparison of two binary trees.

Assume the numbers of leaf nodes in binary trees $A$ and $B$ are $m$ and $n$, respectively.
In a binary tree, the number of inner nodes is the number of leaf nodes minus 1, and thus, the
numbers of inner nodes in $A$ and $B$ are $m - 1$ and $n - 1$, respectively.

The time complexity of each type of tree node comparisons is calculated as follows.

L-L: $m \times n$

P-L: $(m - 1) \times n$ /* without loss of generality, assume we compare all the
sub-trees in $A$ with all the leaf nodes in $B$. */
P-P: $(m - 1) \times (n - 1)$

Therefore, the overall complexity is the sum of the three types of tree node comparisons,
which can be obtained from the following calculation.

$m \times n + (m - 1) \times n + (m - 1) \times (n - 1)$
$= mn + mn - n + mn - m - n + 1$
$= 3mn - 2n - m + 1$
$\approx O(mn)$

From the above equation, the worst case overall time complexity for the proposed hi-
erarchical region tree comparison is $O(mn)$. 
Based on our experiment, the average number of inner nodes in our dataset is about 3.2. Further, while an inner node has a branching factor of at least 2 in our case, many inner nodes have more than two children, suggesting that the actual total number of P-L and P-P comparisons is even lower.

C. Multi-Resolution Image Analysis (MRIA)

The performance of the proposed multi-resolution image analysis (MRIA) is evaluated through two experiments, including the efficiency analysis and the efficacy analysis.

C.1 *MRIA Efficiency Analysis*

In this experiment, we objectively assessed the performance of the proposed multi-resolution image analysis (MRIA) algorithm in terms of segmentation efficiency. The segmentation efficiency quantifies the efficiency of a segmentation algorithm on the basis of the total number of pixels analyzed in the algorithm. The segmentation efficiency of an algorithm $A$ applied to an image $I$ is defined as 1 minus the ratio of the total number of pixels analyzed to the number of pixels in the original image, which is formalized in the following equation.

\[
\text{Segmentation Efficiency}(A, I) = 1 - \frac{\sum_{i=1}^{n_l} \# \text{ of pixels analyzed at level } i}{\# \text{ of pixels in the original image}}
\]
where $i$ represents the level in the multi-resolution image pyramid, and $i = 1$ indicates the lowest image resolution in the image pyramid; $n_i$ is the level of the highest image resolution processed for image $I$.

Based on our experiment on the 10,000 images, the average segmentation efficiency of the proposed multi-resolution image analysis (MRJA) is 98.26%. This indicates that the proposed MRJA approach is very efficient in segmenting an image. In addition, during multi-resolution segmentation, a hierarchical image representation is naturally formed, without the need for any additional run through the data.

C.2 **MRJA Efficacy Analysis**

In this research, we introduce a subjective quality assessment experiment to evaluate the efficacy of the proposed multi-resolution image analysis (MRJA) algorithm in terms of image segmentation quality. This experiment compares the segmentation results generated from the proposed MRJA algorithm and a hill-climbing based color $k$-means segmentation algorithm (HCK) (Achanta, Estrada, Wils, & Süssstrunk, 2008; Ohashi, Aghbari, & Makinouchi, 2003). To ensure the integrity of subjective evaluation, 9 evaluators performed a blind review through a web interface as shown in Figure 19. The evaluators vote the best segmented image from the two displayed segmented images produced by our algorithm and HCK, respectively. The evaluation system also provides a neutral option, if both segmented images are comparable. The results of the assessment are presented in Table 2.
Figure 19: The blind review web interface for evaluating the efficacy of the proposed MRIA algorithm. In this web interface, the original image is displayed in the middle of the interface, and the two segmentation masks produced from the proposed multi-resolution image analysis (MRIA) and the hill climbing based color k-means (HCK) are randomly displayed at the left and right of the original image. This web evaluation system requests evaluators to select the best segmentation results from the two segmentation masks displayed or choose the original image if the two segmentation masks displayed are comparable (equally good or equally bad). In the efficacy evaluation, 200 randomly selected images are displayed to 9 evaluators and the evaluators’ feedbacks are then collected for the subsequent analysis.

<table>
<thead>
<tr>
<th></th>
<th>Comparable</th>
<th>MRIA is better</th>
<th>HCK is better</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRIA vs. HCK8</td>
<td>20%</td>
<td>67%</td>
<td>13%</td>
</tr>
<tr>
<td>MRIA vs. HCK10</td>
<td>19%</td>
<td>73%</td>
<td>8%</td>
</tr>
</tbody>
</table>

Table 2: Subjective segmentation quality assessment
In Table 2, the numbers ‘8’ and ‘10’ represent the different number of bins used in the HCK algorithm. Table 2 demonstrates that the image segmentation quality of the proposed MRIA algorithm significantly outperforms the HCK algorithm.

D. Evaluation Metrics for Multiple-Object Image Retrieval

Average precision (AveP) and mean average precision (MAP) are two commonly used standard evaluation metrics in the text retrieval conference (TREC) community, which are also used in the following experiments in order to fairly compare the proposed multiple-object image retrieval (MOIR) framework to other existing frameworks.

We choose these two measures because we need a measure that takes into account not only the relevance of retrieved images but also their ranks in the returned list. Average precision (AveP) and mean average precision (MAP) are known to be two such measures that simultaneously take into account precision, recall, and rank. In addition, these two evaluation metrics have been shown to have exceptionally good discrimination power and robustness.

D.1 Average Precision (AveP)

The definition of the average precision (AveP) measure for a query $q$ is formalized in the following equation.

$$AveP(q) = \frac{\sum_{k=1}^{n}(P(k) \times rel(k))}{\text{number of retrieved relevant documents}}$$
where \( k \) is the rank in the sequence of the retrieved list, and \( n \) is the number of retrieved images. \( P(k) \) indicates the precision at cut-off \( k \) in the list, and \( rel(k) \) represents an indicator function equaling 1 if the item (image) at the rank \( k \) is relevant, zero otherwise.

D.2  \textit{Mean Average Precision (MAP)}

The mean average precision (MAP) is the arithmetic mean of the average precision values obtained from a set of queries, which is defined in the equation below.

\[
\text{MAP} = \frac{\sum_{q=1}^{Q} \text{AveP}(q)}{Q}
\]

where \( Q \) is the number of queries.

The above two measures, i.e., average prevision (AveP) and mean average precision (MAP), provide an objective and comprehensive view when comparing the performance of the proposed framework to other existing approaches.

E.  MOIR Framework Performance Assessment

Using MAP measure, in the following experiments we compare the proposed multiple-object image retrieval (MOIR) framework to three state-of-the-art object-based image retrieval frameworks, including integrated region matching (IRM) (Li et al., 2000), feedback-based image clustering and retrieval framework (FIRM) (Zhang et al., 2010), and dynamic
region matching (DRM) (Ji et al., 2008). In order to make this experiment a fair comparison, SVM is integrated into IRM for learning the user’s feedback since IRM itself is designed for matching two sets of segments but without the ability to incorporate the user’s relevance feedback.

E.1 Comparison of Color Feature Representations

In this research, two commonly and widely used histogram-based color feature representations, i.e., color code histogram (CCH) and MPEG-7 dominant color descriptor (DCD), are introduced to elude the problems of directional color distribution aforementioned in Section B, Chapter 3. In order to identify which color feature representation is more effective in image retrieval, we conduct the following experiment to compare these histogram-based color feature representations.

In this experiment, we evaluate the effectiveness of the two color feature representations aforementioned on the basis of the full image search framework which compares two images by measuring their similarity according to the global visual features extracted from each image. Specifically, the 6-bit color code histogram (CCH) features and the MPEG-7 dominant color descriptor (DCD) features are extracted from the entire image for each image in our dataset. The 6-bit color code histogram features (CCH) represent an image as 64-dimension feature vector, while the MPEG-7 dominant color descriptor (DCD) represents an image as 72-dimension feature vector.

As shown in Table 3, 201 multiple-object query images are selected from the dataset, which includes 9 different query object combinations, including: bonsai + shoji (18), blue
sky + red bus (34), pyramid + blue sky (21), white rabbit + snow (11), gun + bullet (19), red airplane + blue sky (18), red car + roadway (45), yellow bus + roadway (19), yellow car + roadway (16) where the number in the parentheses represents the number of query images in that combination.

Table 3: Performance comparison of MPEG-7 dominant color descriptor (DCD) and 6-bit color code histogram (CCH) color features on full image search

<table>
<thead>
<tr>
<th>Number of Query Images</th>
<th>Query Objects</th>
<th>CCH</th>
<th>DCD</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>bonsai + shoji</td>
<td>18.73%</td>
<td>19.09%</td>
</tr>
<tr>
<td>34</td>
<td>blue sky + red bus</td>
<td>12.46%</td>
<td>18.29%</td>
</tr>
<tr>
<td>21</td>
<td>pyramid + blue bus</td>
<td>7.49%</td>
<td>9.48%</td>
</tr>
<tr>
<td>11</td>
<td>white rabbit + snow</td>
<td>12.70%</td>
<td>14.79%</td>
</tr>
<tr>
<td>19</td>
<td>gun + bullet</td>
<td>7.97%</td>
<td>12.13%</td>
</tr>
<tr>
<td>18</td>
<td>red airplane + blue sky</td>
<td>7.59%</td>
<td>9.85%</td>
</tr>
<tr>
<td>45</td>
<td>red car + roadway</td>
<td>3.25%</td>
<td>4.09%</td>
</tr>
<tr>
<td>19</td>
<td>yellow bus + roadway</td>
<td>6.51%</td>
<td>6.93%</td>
</tr>
<tr>
<td>16</td>
<td>yellow car + roadway</td>
<td>6.71%</td>
<td>7.22%</td>
</tr>
<tr>
<td>201</td>
<td>Overall Average</td>
<td>8.57%</td>
<td>10.78%</td>
</tr>
</tbody>
</table>
We assess the image similarity between a query image and each target image in the dataset according to the similarity measure described in Section B1 and Section B2, Chapter 3. Then, all target images are ranked according to their similarities to the query image. For each query image, the average precision (AveP) value is calculated on the basis of the ranked target image list. For each query image category, the mean average precision (MAP) value is subsequently calculated as the mean of the AveP values of its member images.

In Table 3 each row represents a query image category. The first column shows the number of query images in the corresponding query image category. The second column indicates the user-desired query objects. The third and the fourth columns are the mean average precision values for the 6-bit color code histogram feature representation and the MPEG-7 dominant color descriptor feature representation, respectively.

From the experimental results shown in Table 3, the mean average precision (MAP) values of the 6-bit color code histogram (CCH) feature representation for the 9 query image categories are: “bonsai + shoji” (18.73%), “blue sky + red bus” (12.46%), “pyramid + blue bus” (7.49%), “white rabbit + snow” (12.70%), “gun + bullet” (7.97%), “red airplane + blue sky” (7.59%) + “red car + roadway” (3.25%), “yellow bus + roadway” (6.51%), and “yellow car + roadway” (6.71%).

In addition, the mean average precision (MAP) values of the MPEG-7 dominant color descriptor (DCD) feature representation for the 9 query image categories are: “bonsai + shoji” (19.09%), “blue sky + red bus” (18.29%), “pyramid + blue bus” (9.48%), “white rabbit + snow” (14.79%), “gun + bullet” (12.13%), “red airplane + blue sky” (9.85%) + “red car + roadway” (4.09%), “yellow bus + roadway” (6.93%), and “yellow car + roadway” (7.22%).
Further, the overall mean average precision (MAP) values of the 6-bit color code histogram (CCH) feature representation and the MPEG-7 dominant color descriptor (DCD) feature representation are 8.57% and 10.78%, respectively.

The experimental results indicate that the MPEG-7 dominant color descriptor (DCD) feature representation significantly outperforms that of the 6-bit color code histogram feature representation. This suggests that the MPEG-7 dominant color descriptor (DCD) feature representation can better characterize images. Therefore, based on the experimental results, we adopt the MPEG-7 dominant color descriptor (DCD) feature representation in this research in order to not only better represent images but also avoid the problems in directional distribution.

E.2 Comparison of Image Matching Algorithms

In Section C, Chapter 4 we have demonstrated the efficacy and efficiency of the proposed multi-resolution image analysis (MRIA) algorithm for image segmentation and hierarchical region tree construction. However, without an efficient and effective hierarchical region tree comparison algorithm, using hierarchical image representation alone cannot relieve the problem of imperfect image segmentation. In this research, we develop an efficient hierarchical region tree comparison algorithm, namely segmentation tree matching (STM), to work together with hierarchical image representation in order to ease the problem of over-segmentation in object-based image retrieval.

To assess the effectiveness of the proposed segmentation tree matching (STM) algorithm, we compare the proposed segmentation tree matching (STM) algorithm to several
state-of-the-art approaches. These state-of-the-art approaches we use for comparison include
the earth mover’s distance (EMD) (Pele & Werman, 2009), the integrated region matching
(IRM, the core algorithm of SIMPLIcity, Stanford Univ.) (Li, et al., 2000), feedback-based
image clustering and retrieval framework (FIRM) (Zhang, et al., 2010), and dynamic region
matching (DRM) (Ji, et al., 2008).

The assessment is based on 201 query images containing multiple user-desired ob-
jects as shown in Table 4. These query images consist of 9 different query object combina-
tions, including: bonsai + shoji (18), blue sky + red bus (34), pyramid + blue sky (21), white
rabbit + snow (11), gun + bullet (19), red airplane + blue sky (18), red car + roadway (45),
yellow bus + roadway (19), yellow car + roadway (16) where the number in the parentheses
represents the number of query images in that combination.

In this experiment, we adopt the mean average precision (MAP) value as the evalua-
tion metric and compare the performance of STM to that of the state-of-the-art algorithms.
The detailed experimental results are presented in Table 4. In Table 4, each row represents a
query image category. The first column shows the number of query images in the corre-
sponding query image category. The second column indicates the user-desired query objects.
The columns 3-7 are mean average precision values of the STM (the proposed segmentation
tree matching), EMD (the earth mover’s distance), IRM (integrated region matching), FIRM
(feedback-based image clustering and retrieval framework), and DRM (dynamic region
matching), respectively.

From the experimental results shown in Table 4, the mean average precision (MAP)
values of the proposed segmentation tree matching (STM) algorithm of the 9 query image
categories are: “bonsai + shoji” (21.78%), “blue sky + red bus” (26.28%), “pyramid + blue
bus” (6.80%), “white rabbit + snow” (14.08%), “gun + bullet” (10.75%), “red airplane + blue sky” (8.24%) + “red car + roadway” (5.79%), “yellow bus + roadway” (8.90%), and “yellow car + roadway” (7.56%). Among the results produced by five different algorithms, STM has the best performance in 5 of the 9 image categories and a close-to-best performance in the “white rabbit + snow” category, demonstrating the highest overall MAP value (12.37%), followed by EMD (11.82%), IRM (11.55%), and FIRM (11.55%). DRM has a significantly lower overall average MAP value (5.99%) than the others (as shown in Figure 20).

![Comparison of Image Matching Algorithms](image)

**Figure 20:** Comparison of image matching algorithms: STM, EMD, IRM, FIRM, and DRM based on 201 query images containing multiple user-desired objects. Among the 5 image matching algorithms, the proposed hierarchical region tree comparison algorithm (STM) has the best performance (12.37%) in terms of the mean average precision value.
Table 4: The performance comparison details of five image matching algorithms

<table>
<thead>
<tr>
<th>Number of Query Images</th>
<th>Query Objects</th>
<th>STM</th>
<th>EMD</th>
<th>IRM</th>
<th>FIRM</th>
<th>DRM</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>bonsai + shoji</td>
<td>21.78%</td>
<td>21.25%</td>
<td>20.43%</td>
<td>20.43%</td>
<td>8.57%</td>
</tr>
<tr>
<td>34</td>
<td>blue sky + red bus</td>
<td>26.28%</td>
<td>22.57%</td>
<td>21.60%</td>
<td>21.60%</td>
<td>6.05%</td>
</tr>
<tr>
<td>21</td>
<td>pyramid + blue bus</td>
<td>6.80%</td>
<td>8.32%</td>
<td>8.19%</td>
<td>8.19%</td>
<td>5.74%</td>
</tr>
<tr>
<td>11</td>
<td>white rabbit + snow</td>
<td>14.08%</td>
<td>14.18%</td>
<td>14.23%</td>
<td>14.23%</td>
<td>9.61%</td>
</tr>
<tr>
<td>19</td>
<td>gun + bullet</td>
<td>10.75%</td>
<td>11.67%</td>
<td>11.57%</td>
<td>11.57%</td>
<td>6.39%</td>
</tr>
<tr>
<td>18</td>
<td>red airplane + blue sky</td>
<td>8.24%</td>
<td>11.53%</td>
<td>11.30%</td>
<td>11.30%</td>
<td>6.82%</td>
</tr>
<tr>
<td>45</td>
<td>red car + roadway</td>
<td>5.79%</td>
<td>4.36%</td>
<td>4.34%</td>
<td>4.34%</td>
<td>3.67%</td>
</tr>
<tr>
<td>19</td>
<td>yellow bus + roadway</td>
<td>8.90%</td>
<td>8.07%</td>
<td>8.13%</td>
<td>8.13%</td>
<td>5.48%</td>
</tr>
<tr>
<td>16</td>
<td>yellow car + roadway</td>
<td>7.56%</td>
<td>7.22%</td>
<td>7.40%</td>
<td>7.40%</td>
<td>6.46%</td>
</tr>
<tr>
<td>201</td>
<td>Overall Average</td>
<td>12.37%</td>
<td>11.82%</td>
<td>11.55%</td>
<td>11.55%</td>
<td>5.99%</td>
</tr>
</tbody>
</table>

In summary, based on 201 queries containing multiple-object of users’ interests, the experimental results show that the proposed STM algorithm outperforms EMD and the other existing approaches used in our comparison even at the initial retrieval. This indicates that the proposed STM is more effective than other state-of-the-art image matching algorithms even without using relevance feedback. In addition, the experimental results suggest that the
proposed STM algorithm can relieve the over-segmentation problem since all the image matching algorithms use the same color feature and compare images based on the same segmented results. The difference in the retrieval accuracy can only be caused by the use of different matching mechanisms, and STM is a clear winner in this comparison.

E.3 Single-Object Retrieval

In previous section, we have demonstrated that the proposed segmentation tree matching (STM) algorithm outperforms other state-of-the-art image matching algorithms in terms of whole image similarity captured through the similarity between two sets of objects. However, it is quite often to see query images not only contain user-desired objects but also irrelevant objects. Thus, it is essential for an image retrieval system to learn users’ query intention in order to improve the image retrieval accuracy.

To learn users’ query intention, it would be much easier if an image retrieval system requests a user to explicitly specify the object(s) of his/her interest. However, it would be not only troublesome but also impractical for a user to put so many efforts in the user interaction cycle. For example, some existing systems request users to select the regions that correspond to desired objects, but it would be very difficult and confusing for users to complete this requirement when the image is over-segmented or under-segmented., let alone asking the user to select multiple objects from such an image.

In the proposed multiple-object image retrieval (MOIR) framework, our goal is to minimize the amount of user effort, from which essential information for discovery of user-desired objects can be procured by the proposed approach. For this purpose, we introduce the
relevance feedback (RF) technique, a supervised machine learning mechanism, to analyze users’ feedback.

In object-based image retrieval, retrieving a single object is the most fundamental function. In order to demonstrate the effectiveness of the proposed multiple-object image retrieval (MOIR) framework in single object retrieval, we conduct the following experiment which evaluates the proposed MOIR framework based on 560 query images from 11 query object categories including dinosaur (101), red bus (53), pyramid (24), white rabbit (16), bullet (23), yellow car (34), yellow bus (23), bonsai (100), tiger (102), penguin (66), and shoji (18) where the number in the parentheses represents the number of query images in that category.

The proposed segmentation tree matching (STM) algorithm is designed to compare two segmentation trees representing two images. The STM algorithm does not have the ability to discover user-desired objects by itself. To enable this, we integrate with STM a one-class support vector machine (SVM) with RBF kernel, which seamlessly cooperates with the STM algorithm to gradually discover user-desired object(s) from the user’s feedback.

Among the four state-of-the-art image matching algorithms, i.e., EMD, IRM, FIRM, and DRM, neither EMD nor IRM involve a relevance feedback mechanism. In order to conduct a fair comparison between the proposed MOIR framework (STM+SVM) and these state-of-the-art approaches, in this experiment, SVM is incorporated into IRM for learning from the user’s feedback since IRM itself is designed for matching two sets of segments/objects without considering the user’s query intention. We do not compare the proposed MOIR framework (STM+SVM) to EMD because no suitable existing relevance feedback mechanism can be readily integrated with EMD. Only FIRM and DRM have a built-in
relevance feedback mechanism so that they can be directly compared with the proposed MOIR framework (STM+SVM) without any framework modification.

It is worth noticing that sometimes there is no relevant image in the returned top 20 images at the initial retrieval, which may fail some relevance feedback learning algorithms, such as FIRM. The reason is that these algorithms only take into account the relevant images in the user’s feedback. In order to make the experiment a fair comparison, five randomly selected relevant images are automatically inserted into the top 20 returned images as the top 5 images at the initial retrieval, serving as the starting point for all the algorithms used in this comparison.

Similarly, we adopt the mean average precision (MAP) value in this experiment as the performance metric and compare the performance of the MOIR framework (STM+SVM), to that of the state-of-the-art approaches. The detailed experimental results are presented in Tables 5.1-5.4, in which each row represents a query image category. The first column shows the number of query images in the corresponding query image category. The second column indicates the user-desired query object. The 3rd to 6th columns are mean average precision values of the 1st – 4th feedback iterations, respectively. The percentage improvement of each subsequent feedback iteration over the 1st feedback iteration is shown in the parentheses.
Table 5.1: The effectiveness of MOIR (STM+SVM) in single object retrieval with feedback

<table>
<thead>
<tr>
<th>Number of Query Images</th>
<th>Query Object</th>
<th>Feedback Iteration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>101</td>
<td>dinosaur</td>
<td>21.40%</td>
</tr>
<tr>
<td>53</td>
<td>red bus</td>
<td>23.00%</td>
</tr>
<tr>
<td>24</td>
<td>pyramid</td>
<td>4.95%</td>
</tr>
<tr>
<td>16</td>
<td>white rabbit</td>
<td>8.73%</td>
</tr>
<tr>
<td>23</td>
<td>bullet</td>
<td>10.26%</td>
</tr>
<tr>
<td>34</td>
<td>yellow car</td>
<td>4.58%</td>
</tr>
<tr>
<td>23</td>
<td>yellow bus</td>
<td>9.12%</td>
</tr>
<tr>
<td>100</td>
<td>bonsai</td>
<td>4.98%</td>
</tr>
<tr>
<td>102</td>
<td>tiger</td>
<td>8.32%</td>
</tr>
<tr>
<td>66</td>
<td>penguin</td>
<td>4.09%</td>
</tr>
<tr>
<td>18</td>
<td>shoji</td>
<td>16.83%</td>
</tr>
<tr>
<td>560</td>
<td>Overall Average</td>
<td>11.00%</td>
</tr>
</tbody>
</table>
Table 5.2: The effectiveness of IRM+SVM in single object retrieval with feedback

<table>
<thead>
<tr>
<th>Number of Query Images</th>
<th>Query Object</th>
<th>Feedback Iteration</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>101</td>
<td>dinosaur</td>
<td>18.07%</td>
<td>26.15% (+8.08)</td>
<td>28.45% (+10.38)</td>
<td>29.19% (+11.12)</td>
<td></td>
</tr>
<tr>
<td>53</td>
<td>red bus</td>
<td>18.16%</td>
<td>26.52% (+8.36)</td>
<td>33.82% (+15.66)</td>
<td>38.54% (+20.38)</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>pyramid</td>
<td>5.28%</td>
<td>3.07% (-2.21)</td>
<td>3.66% (-1.62)</td>
<td>4.16% (-1.12)</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>white rabbit</td>
<td>7.66%</td>
<td>4.68% (-2.98)</td>
<td>5.00% (-2.66)</td>
<td>5.09% (-2.57)</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>bullet</td>
<td>8.40%</td>
<td>8.31% (-0.09)</td>
<td>10.06% (+1.66)</td>
<td>9.21% (+0.81)</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>yellow car</td>
<td>4.02%</td>
<td>3.42% (-0.60)</td>
<td>3.34% (-0.68)</td>
<td>3.61% (-0.41)</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>yellow bus</td>
<td>7.23%</td>
<td>7.74% (+0.51)</td>
<td>8.61% (+1.38)</td>
<td>8.66% (+1.43)</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>bonsai</td>
<td>5.26%</td>
<td>8.04% (+2.78)</td>
<td>8.56% (+3.30)</td>
<td>9.17% (+3.91)</td>
<td></td>
</tr>
<tr>
<td>102</td>
<td>tiger</td>
<td>7.21%</td>
<td>5.75% (-1.46)</td>
<td>6.35% (-0.86)</td>
<td>7.21% (+0.00)</td>
<td></td>
</tr>
<tr>
<td>66</td>
<td>penguin</td>
<td>4.36%</td>
<td>4.63% (+0.27)</td>
<td>4.70% (+0.34)</td>
<td>5.14% (+0.78)</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>shoji</td>
<td>15.79%</td>
<td>19.12% (+3.33)</td>
<td>24.69% (+8.90)</td>
<td>23.99% (+8.20)</td>
<td></td>
</tr>
<tr>
<td>560</td>
<td>Overall Average</td>
<td>9.58% (8.08)</td>
<td>12.00% (+2.42)</td>
<td>13.63% (+4.05)</td>
<td>14.52% (+4.94)</td>
<td></td>
</tr>
</tbody>
</table>
Table 5.3: The effectiveness of FIRM in single object retrieval with feedback

<table>
<thead>
<tr>
<th>Number of Query Images</th>
<th>Query Object</th>
<th>Feedback Iteration</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>dinosaur</td>
<td>26.97%</td>
<td>26.87%</td>
<td>(-0.10)</td>
<td>26.77%</td>
<td>(-0.20)</td>
</tr>
<tr>
<td>53</td>
<td>red bus</td>
<td>19.99%</td>
<td>20.00%</td>
<td>(+0.01)</td>
<td>20.01%</td>
<td>(+0.02)</td>
</tr>
<tr>
<td>24</td>
<td>pyramid</td>
<td>6.89%</td>
<td>6.90%</td>
<td>(+0.01)</td>
<td>6.92%</td>
<td>(+0.03)</td>
</tr>
<tr>
<td>16</td>
<td>white rabbit</td>
<td>8.94%</td>
<td>8.94%</td>
<td>(+0.00)</td>
<td>8.94%</td>
<td>(+0.00)</td>
</tr>
<tr>
<td>23</td>
<td>bullet</td>
<td>9.87%</td>
<td>9.86%</td>
<td>(-0.01)</td>
<td>9.86%</td>
<td>(-0.01)</td>
</tr>
<tr>
<td>34</td>
<td>yellow car</td>
<td>4.09%</td>
<td>4.09%</td>
<td>(+0.00)</td>
<td>4.09%</td>
<td>(+0.00)</td>
</tr>
<tr>
<td>23</td>
<td>yellow bus</td>
<td>7.43%</td>
<td>7.44%</td>
<td>(+0.01)</td>
<td>7.44%</td>
<td>(+0.01)</td>
</tr>
<tr>
<td>100</td>
<td>bonsai</td>
<td>7.53%</td>
<td>7.52%</td>
<td>(-0.01)</td>
<td>7.52%</td>
<td>(-0.01)</td>
</tr>
<tr>
<td>102</td>
<td>tiger</td>
<td>8.38%</td>
<td>8.38%</td>
<td>(+0.00)</td>
<td>8.38%</td>
<td>(+0.00)</td>
</tr>
<tr>
<td>66</td>
<td>penguin</td>
<td>5.17%</td>
<td>5.16%</td>
<td>(-0.01)</td>
<td>5.17%</td>
<td>(+0.00)</td>
</tr>
<tr>
<td>18</td>
<td>shoji</td>
<td>20.48%</td>
<td>20.47%</td>
<td>(-0.01)</td>
<td>20.47%</td>
<td>(-0.01)</td>
</tr>
<tr>
<td>560</td>
<td>Overall Average</td>
<td>12.40%</td>
<td>12.39%</td>
<td>(-0.01)</td>
<td>12.37%</td>
<td>(-0.03)</td>
</tr>
<tr>
<td>Number of Query Images</td>
<td>Query Object</td>
<td>Feedback Iteration</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>-----------------------</td>
<td>--------------</td>
<td>--------------------</td>
<td>------</td>
<td>------------</td>
<td>------------</td>
<td>------------</td>
</tr>
<tr>
<td>101</td>
<td>dinosaur</td>
<td>10.87%</td>
<td>12.94%</td>
<td>(+2.07)</td>
<td>14.63%</td>
<td>(+3.76)</td>
</tr>
<tr>
<td>53</td>
<td>red bus</td>
<td>5.60%</td>
<td>6.74%</td>
<td>(+1.14)</td>
<td>8.43%</td>
<td>(+2.83)</td>
</tr>
<tr>
<td>24</td>
<td>pyramid</td>
<td>4.94%</td>
<td>5.71%</td>
<td>(+0.77)</td>
<td>6.74%</td>
<td>(+1.80)</td>
</tr>
<tr>
<td>16</td>
<td>white rabbit</td>
<td>6.70%</td>
<td>9.50%</td>
<td>(+2.80)</td>
<td>10.56%</td>
<td>(+3.86)</td>
</tr>
<tr>
<td>23</td>
<td>bullet</td>
<td>6.04%</td>
<td>6.97%</td>
<td>(+0.93)</td>
<td>7.72%</td>
<td>(+1.68)</td>
</tr>
<tr>
<td>34</td>
<td>yellow car</td>
<td>3.68%</td>
<td>4.03%</td>
<td>(+0.35)</td>
<td>4.62%</td>
<td>(+0.94)</td>
</tr>
<tr>
<td>23</td>
<td>yellow bus</td>
<td>4.84%</td>
<td>5.09%</td>
<td>(+0.25)</td>
<td>5.66%</td>
<td>(+0.82)</td>
</tr>
<tr>
<td>100</td>
<td>bonsai</td>
<td>3.87%</td>
<td>4.93%</td>
<td>(+1.06)</td>
<td>5.70%</td>
<td>(+1.83)</td>
</tr>
<tr>
<td>102</td>
<td>tiger</td>
<td>4.46%</td>
<td>5.42%</td>
<td>(+0.96)</td>
<td>6.12%</td>
<td>(+1.66)</td>
</tr>
<tr>
<td>66</td>
<td>penguin</td>
<td>4.32%</td>
<td>6.13%</td>
<td>(+1.81)</td>
<td>6.86%</td>
<td>(+2.54)</td>
</tr>
<tr>
<td>18</td>
<td>shoji</td>
<td>8.58%</td>
<td>11.76%</td>
<td>(+3.18)</td>
<td>14.89%</td>
<td>(+6.31)</td>
</tr>
<tr>
<td>560</td>
<td>Overall Average</td>
<td>5.85%</td>
<td>7.20%</td>
<td>(+1.35)</td>
<td>8.28%</td>
<td>(+2.43)</td>
</tr>
</tbody>
</table>
Figure 21 summarizes the retrieval performance of each framework in terms of MAP. The red bar represents the proposed multiple-object image retrieval (MOIR) framework; the green bar represents the IRM with SVM; the blue bar represents the FIRM; and the yellow bar represents the DRM. From Figure 21, it can be observed that after 4 feedback iterations, the MAP value of the proposed MOIR framework reaches 15.52%, which is 1%, 3.17%, and 6.1% higher than that of the IRM+SVM, FIRM, and DRM, respectively. The MOIR framework significantly outperforms other frameworks since the number of queries (560 queries) and the size of retrieval scope (10,000 target images) are large enough for us to claim that even a 1% increase in MAP value is significant. Further, the MAP value of MOIR steadily increases through the feedback iterations, which indicates the robustness and effectiveness of the relevance feedback.

![Image](image_url)  
**Figure 21:** Single object retrieval results (560 queries)

<table>
<thead>
<tr>
<th>Iteration</th>
<th>MOIR</th>
<th>IRM + SVM</th>
<th>FIRM</th>
<th>DRM</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>11.00%</td>
<td>9.58%</td>
<td>12.40%</td>
<td>5.85%</td>
</tr>
<tr>
<td>3</td>
<td>11.95%</td>
<td>12.00%</td>
<td>12.39%</td>
<td>7.20%</td>
</tr>
<tr>
<td>4</td>
<td>13.81%</td>
<td>13.63%</td>
<td>12.37%</td>
<td>8.28%</td>
</tr>
<tr>
<td>5</td>
<td>15.52%</td>
<td>14.52%</td>
<td>12.35%</td>
<td>9.42%</td>
</tr>
</tbody>
</table>
We demonstrate the performance of the proposed MOIR framework (STM+SVM) and other state-of-the-art frameworks, i.e., IRM+SVM, FIRM, and DRM, in retrieving single object by using a sample query image containing a “red bus” as the user-desired object. Figures 22-25 exemplify the top 20 retrieved images returned from the MOIR (STM+SVM), IRM+SVM, FIRM, and DRM, respectively. Each figure includes the retrieval results in the initial retrieval and at the subsequent 4 feedback iterations.

In the image retrieval user interface, the query image is shown at the top row. In addition, the user-desired query object, the number of relevant images in the top 20 returned images, and the mean average precision (MAP) value are displayed in the text box located to the right of the query image. The top 20 images in the ranked target image list are returned to the user for result display. The display of the top 20 images is spatially arranged in a way that their similarity (produced by a certain matching algorithm) to the query image decreases from top-to-bottom and left-to-right. On the top of each image, the value in the square brackets is either 0 (irrelevant) or 1 (relevant), indicating the user’s feedback for that image. A relevant image must contain all objects of interest, or irrelevant otherwise.

As aforementioned, 5 relevant images are added as the top 5 images in the initial retrieval in order to conduct a fair comparison. This can be observed in the initial retrieval results, i.e., Figure 22(1), Figure 23(1), Figure 24(1), and Figure 25(1). Further, we can observe that there are many irrelevant images, much more than there are positive ones, returned in the initial retrieval because the image retrieval system has no prior knowledge about a user’s query intention and retrieves target images purely according to the overall similarity. After the 1st feedback iteration, we can observe that the proposed MOIR framework effectively learns user-desired object(s) from the feedbacks and gradually refines the results. This sug-
gests that the integration of the proposed MOIR framework with RF technique is effective in discovery of user-desired object.
Figure 22(1): Example of MOIR image retrieval results – initial retrieval.
Figure 22(2): Example of MOIR image retrieval results – 1st feedback.
Figure 22(3): Example of MOIR image retrieval results – 2nd feedback.
Figure 22(4): Example of MOIR image retrieval results – 3rd feedback.
Figure 22(5): Example of MOIR image retrieval results – 4th feedback.
Figure 23(1): Example of IRM+SVM image retrieval results – initial retrieval.
Figure 23(2): Example of IRM+SVM image retrieval results – 1st feedback.
Figure 23(3): Example of IRM+SVM image retrieval results – 2nd feedback.
Figure 23(4): Example of IRM+SVM image retrieval results – 3rd feedback.

Query Object: "red bus" - [4] - AveP : 0.079409
Figure 23(5): Example of IRM+SVM image retrieval results – 4th feedback.

Query Object: "red bus" - [3] - AveP : 0.051446
Figure 24(1): Example of FIRM image retrieval results – initial retrieval.
Figure 24(2): Example of FIRM image retrieval results – 1st feedback.
Figure 24(3): Example of FIRM image retrieval results – 2nd feedback.
Figure 24(4): Example of FIRM image retrieval results – 3rd feedback.
Figure 24(5): Example of FIRM image retrieval results – 4th feedback.
Figure 25(1): Example of DRM image retrieval results – initial retrieval.
Figure 25(2): Example of DRM image retrieval results – 1st feedback.
Figure 25(3): Example of DRM image retrieval results – 2nd feedback.
Figure 25(4): Example of DRM image retrieval results – 3rd feedback.
581095.jpg

Query Object: "red bus" - [1] - AveP : 0.037394

Figure 25(5): Example of DRM image retrieval results – 4th feedback.
E.4  \textit{Multiple-Object Retrieval}

Similar to the single object image retrieval, in this experiment we demonstrate the effectiveness of the MOIR framework in multiple-object retrieval based on 201 query images from 9 different query object combinations, including: bonsai + shoji (18), blue sky + red bus (34), pyramid + blue sky (21), white rabbit + snow (11), gun + bullet (19), red airplane + blue sky (18), red car + roadway (45), yellow bus + roadway (19), yellow car + roadway (16) where the number in the parentheses represents the number of query images in that combination.

In this experiment, we again adopt the mean average precision (MAP) value as the evaluation metric and compare the performance of the MOIR framework (STM+SVM), to that of the state-of-the-art approaches. The detailed experimental results are presented in Tables 6.1-6.4, in which each row represents a query image category. The first column shows the number of query images in the corresponding query image category. The second column indicates the user-desired query object. The columns 3-6 are mean average precision values of the 1\textsuperscript{st} – 4\textsuperscript{th} feedback iterations, respectively. The percentage improvement of each subsequent feedback iteration over the 1\textsuperscript{st} feedback iteration is shown in the parentheses.
Table 6.1: The effectiveness of MOIR (STM+SVM) in multiple-object retrieval.

<table>
<thead>
<tr>
<th>Number of Query Images</th>
<th>Query Objects</th>
<th>Feedback Iteration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>bonsai + shoji</td>
<td>16.83%</td>
</tr>
<tr>
<td>34</td>
<td>blue sky + red bus</td>
<td>26.91%</td>
</tr>
<tr>
<td>21</td>
<td>pyramid + blue sky</td>
<td>5.84%</td>
</tr>
<tr>
<td>11</td>
<td>white rabbit + snow</td>
<td>14.74%</td>
</tr>
<tr>
<td>19</td>
<td>gun + bullet</td>
<td>11.06%</td>
</tr>
<tr>
<td>18</td>
<td>red airplane + blue sky</td>
<td>9.24%</td>
</tr>
<tr>
<td>45</td>
<td>red car + roadway</td>
<td>5.86%</td>
</tr>
<tr>
<td>19</td>
<td>yellow bus + roadway</td>
<td>9.68%</td>
</tr>
<tr>
<td>16</td>
<td>yellow car + roadway</td>
<td>7.38%</td>
</tr>
<tr>
<td>201</td>
<td>Overall Average</td>
<td>12.16%</td>
</tr>
</tbody>
</table>
Table 6.2: The effectiveness of IRM+SVM in multiple-object retrieval.

<table>
<thead>
<tr>
<th>Number of Query Images</th>
<th>Query Objects</th>
<th>Feedback Iteration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>bonsai + shoji</td>
<td>15.79%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(19.12%)(+8.90)</td>
</tr>
<tr>
<td>34</td>
<td>blue sky + red bus</td>
<td>18.65%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(24.69%)(+15.65)</td>
</tr>
<tr>
<td>21</td>
<td>pyramid + blue sky</td>
<td>6.30%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(5.68%)(-0.62)</td>
</tr>
<tr>
<td>11</td>
<td>white rabbit + snow</td>
<td>13.20%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(6.73%)(-6.47)</td>
</tr>
<tr>
<td>19</td>
<td>gun + bullet</td>
<td>8.63%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(8.98%)(+0.31)</td>
</tr>
<tr>
<td>18</td>
<td>red airplane + blue sky</td>
<td>10.68%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(10.68%)(-0.43)</td>
</tr>
<tr>
<td>45</td>
<td>red car + roadway</td>
<td>4.20%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(5.80%)(+1.60)</td>
</tr>
<tr>
<td>19</td>
<td>yellow bus + roadway</td>
<td>7.96%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(8.90%)(+1.26)</td>
</tr>
<tr>
<td>16</td>
<td>yellow car + roadway</td>
<td>7.09%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(5.20%)(-1.47)</td>
</tr>
<tr>
<td>201</td>
<td>Overall Average</td>
<td>9.98%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(11.00%)(+3.38)</td>
</tr>
</tbody>
</table>
Table 6.3: The effectiveness of FIRM in multiple-object retrieval.

<table>
<thead>
<tr>
<th>Number of Query Images</th>
<th>Query Objects</th>
<th>Feedback Iteration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20.48%</td>
</tr>
<tr>
<td>18</td>
<td>bonsai + shoji</td>
<td>21.62%</td>
</tr>
<tr>
<td>34</td>
<td>blue sky + red bus</td>
<td>8.14%</td>
</tr>
<tr>
<td>21</td>
<td>pyramid + blue sky</td>
<td>14.26%</td>
</tr>
<tr>
<td>11</td>
<td>white rabbit + snow</td>
<td>11.57%</td>
</tr>
<tr>
<td>19</td>
<td>gun + bullet</td>
<td>11.27%</td>
</tr>
<tr>
<td>18</td>
<td>red airplane + blue sky</td>
<td>4.35%</td>
</tr>
<tr>
<td>45</td>
<td>red car + roadway</td>
<td>8.13%</td>
</tr>
<tr>
<td>19</td>
<td>yellow bus + roadway</td>
<td>8.13%</td>
</tr>
<tr>
<td>16</td>
<td>yellow car + roadway</td>
<td>7.41%</td>
</tr>
<tr>
<td>201</td>
<td>Overall Average</td>
<td>11.56%</td>
</tr>
</tbody>
</table>
Table 6.4: The effectiveness of DRM in multiple-object retrieval.

<table>
<thead>
<tr>
<th>Number of Query Images</th>
<th>Query Objects</th>
<th>Feedback Iteration</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>bonsai + shoji</td>
<td>8.58%</td>
<td>11.76%</td>
<td>(+3.18)</td>
<td>14.75%</td>
<td>(+6.17)</td>
</tr>
<tr>
<td>34</td>
<td>blue sky + red bus</td>
<td>6.06%</td>
<td>7.27%</td>
<td>(+1.21)</td>
<td>9.25%</td>
<td>(+3.19)</td>
</tr>
<tr>
<td>21</td>
<td>pyramid + blue sky</td>
<td>5.77%</td>
<td>6.55%</td>
<td>(+0.78)</td>
<td>7.72%</td>
<td>(+1.95)</td>
</tr>
<tr>
<td>11</td>
<td>white rabbit + snow</td>
<td>9.62%</td>
<td>13.59%</td>
<td>(+3.97)</td>
<td>15.00%</td>
<td>(+5.38)</td>
</tr>
<tr>
<td>19</td>
<td>gun + bullet</td>
<td>6.39%</td>
<td>6.93%</td>
<td>(+0.54)</td>
<td>7.64%</td>
<td>(+1.25)</td>
</tr>
<tr>
<td>18</td>
<td>red airplane + blue sky</td>
<td>6.82%</td>
<td>7.75%</td>
<td>(+0.93)</td>
<td>8.38%</td>
<td>(+1.56)</td>
</tr>
<tr>
<td>45</td>
<td>red car + roadway</td>
<td>3.68%</td>
<td>4.36%</td>
<td>(+0.68)</td>
<td>4.95%</td>
<td>(+1.27)</td>
</tr>
<tr>
<td>19</td>
<td>yellow bus + roadway</td>
<td>5.48%</td>
<td>5.80%</td>
<td>(+0.32)</td>
<td>6.42%</td>
<td>(+0.94)</td>
</tr>
<tr>
<td>16</td>
<td>yellow car + roadway</td>
<td>6.47%</td>
<td>6.72%</td>
<td>(+0.25)</td>
<td>6.94%</td>
<td>(+0.47)</td>
</tr>
<tr>
<td>201</td>
<td>Overall Average</td>
<td>5.99%</td>
<td>7.12%</td>
<td>(+1.13)</td>
<td>8.25%</td>
<td>(+2.26)</td>
</tr>
</tbody>
</table>
The performance for each framework is summarized in Figure 26 which shows that after four feedback iterations, the proposed multiple-object image retrieval (MOIR) framework significantly outperforms IRM+SVM, FIRM, and DRM by 3.25%, 6.02%, and 8.09%, respectively. Similarly, the MAP value of our algorithm also increases through the feedback iterations, again indicating the learning effectiveness and robustness from users’ relevance feedback. Figures 27(2) and 27(5) exemplify the top 20 images retrieved by MOIR after the 1st feedback and the 4th feedback iterations for a sample query image and their corresponding AveP values, respectively. The query image is displayed on the top left corner.

![Figure 26: Multiple-object retrieval results (201 queries)](image-url)
We demonstrate the performance of the proposed MOIR framework (STM+SVM) and other state-of-the-art frameworks, i.e., IRM+SVM, FIRM, and DRM, in retrieving multiple objects by using an image containing a “red bus” and “blue sky” as two user-desired objects. Figures 27-30 exemplify the top 20 retrieved images returned from the MOIR (STM+SVM), IRM+SVM, FIRM, and DRM, respectively. Each figure includes the retrieval results in the initial retrieval and at the subsequent 4 feedback iterations.

In the initial retrieval, the proposed MOIR framework retrieves many bus, but not necessarily red bus, images as shown in Figure 27(1). As aforementioned, since the top 5 images in the initial retrieval are always those automatically added relevant images, they are irrelevant to the quality of the initial retrieval. Therefore, we can ignore the top 5 images in the initial retrieval when evaluating the retrieval result. In the remaining images in Figure 27(1), we can observe that only 1 image (rank: 6th) contains both red bus and blue sky, 2 images (rank: 16th and 19th) contain neither red bus nor blue sky, and the rest of the images contain either red bus or blue sky.

In the subsequent feedback iterations, the learning algorithm was able to gradually learn from the user’s feedback that he/she is interested in red object and blue object. As shown in Figure 27(3), for example, the retrieval system returns images such as a red hot air balloon with blue sky (rank: 4th), a red airplane with blue sky (rank: 7th), a skier in red skiing suit with blue sky (rank: 10th), a red car with blue sky (rank: 12th), and a man wearing red yashmak and belt with blue sky (rank: 16th). As more feedbacks collected in the user interaction cycles, our proposed learning mechanism enables MOIR system to capture user’s query intention. For instance, Figure 27(5) shows that 15 out of 20 returned images contain both red bus and blue sky.
Figure 28 exemplifies that IRM+SVM framework can also benefit from the relevance feedback mechanism. However, it is obvious that the proposed MOIR framework has a better learning curve owing to the contribution of the hierarchical image representation and the segmentation tree comparison. Figures 29 and 30 show that for this query, the MAP values of FIRM and DRM do not gradually increase through the feedback iterations in retrieving multiple objects although the experimental results in Section E.3 Chapter 4 have shown that the relevance feedback mechanisms of these frameworks are effective in single object retrieval. This observation indicates that the learning algorithms used in FIRM and DRM are not robust for multiple-object retrieval.

In summary, in this chapter, Section C shows that the proposed multiple-resolution image analysis (MRIA) is effective and efficient. Section E presents a series of experiments that demonstrate the effectiveness of the proposed multiple-object image retrieval (MOIR) framework in both single and multiple-object retrieval.
Figure 27(1): Example of MOIR image retrieval results –initial retrieval.
Query Object: "blue sky", "red bus" - [4] - AveP : 0.1866

Figure 27(2): Example of MOIR image retrieval results – 1st feedback.
Figure 27(3): Example of MOIR image retrieval results – 2nd feedback.
Figure 27(4): Example of MOIR image retrieval results – 3rd feedback.
Figure 27(5): Example of MOIR image retrieval results – 4th feedback.
Figure 28(1): Example of IRM+SVM image retrieval results – initial retrieval.
Figure 28(2): Example of IRM+SVM image retrieval results – 1st feedback.
Figure 28(3): Example of IRM+SVM image retrieval results – 2nd feedback.
Figure 28(4): Example of IRM+SVM image retrieval results – 3rd feedback.
Figure 28(5): Example of IRM+SVM image retrieval results – 4th feedback.
Figure 29(1): Example of FIRM image retrieval results – initial retrieval.
Figure 29(2): Example of FIRM image retrieval results – 1st feedback.
Figure 29(3): Example of FIRM image retrieval results – 2nd feedback.
Figure 29(4): Example of FIRM image retrieval results – 3\textsuperscript{rd} feedback.
Figure 29(5): Example of FIRM image retrieval results – 4th feedback.

Query Object: "blue sky", "red bus" - [1] - AveP : 0.04804
Figure 30(1): Example of DRM image retrieval results –initial retrieval.
<table>
<thead>
<tr>
<th>Image ID</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>581004</td>
<td>-</td>
</tr>
<tr>
<td>103032</td>
<td>-</td>
</tr>
<tr>
<td>196055</td>
<td>-</td>
</tr>
<tr>
<td>204061</td>
<td>-</td>
</tr>
<tr>
<td>222033</td>
<td>-</td>
</tr>
<tr>
<td>213016</td>
<td>-</td>
</tr>
<tr>
<td>415022</td>
<td>-</td>
</tr>
<tr>
<td>173084</td>
<td>-</td>
</tr>
<tr>
<td>103076</td>
<td>-</td>
</tr>
<tr>
<td>123049</td>
<td>-</td>
</tr>
<tr>
<td>106085</td>
<td>-</td>
</tr>
<tr>
<td>701037</td>
<td>-</td>
</tr>
<tr>
<td>213083</td>
<td>-</td>
</tr>
<tr>
<td>110034</td>
<td>-</td>
</tr>
<tr>
<td>106044</td>
<td>-</td>
</tr>
<tr>
<td>521020</td>
<td>-</td>
</tr>
<tr>
<td>106019</td>
<td>-</td>
</tr>
<tr>
<td>181099</td>
<td>-</td>
</tr>
<tr>
<td>387075</td>
<td>-</td>
</tr>
<tr>
<td>212009</td>
<td>-</td>
</tr>
</tbody>
</table>

Figure 30(2): Example of DRM image retrieval results – 1st feedback.
Figure 30(3): Example of DRM image retrieval results – 2\textsuperscript{nd} feedback.
Figure 30(4): Example of DRM image retrieval results – 3rd feedback.
Figure 30(5): Example of DRM image retrieval results – 4\textsuperscript{th} feedback.
CHAPTER 5
AN APPLICATION FOR FINE ART PAINTING IMAGE RETRIEVAL

A. Introduction

The need for novel techniques to store and retrieve digital fine art images has emerged from the growing popularity of the Internet (Lombardi, Cha, & Tappert, 2004). It is not unusual that users are interested in finding a particular piece of artwork using image search engines such as Google Images, Yahoo Images, and Bing Images. The emerging need often involves a process of finding database images that encompass all the objects and their spatial arrangements in the query image. The abovementioned process is also known as the sub-image retrieval problem (Chiang & Huang, 2008).

It is worth noticing that unlike searching dedicated fine art image databases containing high quality painting images, web images containing artworks are originated from various sources around the world and thus, the image quality varies. For example, some pieces of artworks are printed on T-shirts, mugs, and other commercial products; some photos taken in a gallery or a museum may contain paintings created by famous artists; some fine art publications provide close-up views of the artworks in order to explain the details of the pieces.

Most conventional sub-image retrieval frameworks partition query and target images into a fixed number of blocks as grids in order to preserve the spatial information, and then, they attempt to match blocks between the query image and each target image in order to find whether the query target exists as a sub-image in a target image. These approaches assume that the query target appeared in the query and target images have similar size. This assump-
tion may cause a problem when search images with various sizes and different resolutions such as web images. Unfortunately, not many sub-image retrieval frameworks can deal with arbitrary-size sub-image queries on the basis of color and spatial similarity (Sebe, Lew, & Huijsmans, 1999). Only a few works perform sub-image retrieval based on multi-scale. Using multi-scale analysis on sub-image retrieval can ease the problem of matching query target on images with various sizes and different resolutions. However, these approaches construct the hierarchical image representation using a fixed image partition on multi-resolution images (Luo & Nascimento, 2004; Luo & Nascimento, 2003; Sebe et al., 1999). The use of hierarchical image representation and a fixed image partition brings two problems: First, the comparison of hierarchical image representation created from a fixed image partition on multi-resolution images suffers from high computational complexity since the algorithm needs to find pairs of similar blocks across different image resolutions. Second, these approaches may suffer from noises due to blocks cannot precisely map to the query target. For this reason, due to the variety of web images, it is not a trivial task to search a certain piece of artwork from databases containing images collected from the Internet.

In this dissertation, we have proposed a multiple-object image retrieval (MOIR) framework for natural scene image retrieval. In contrast to natural scene images, fine art (oil/water color) painting images are man-made artificial images with an entirely different set of color, texture, lightness, and edge characteristics. In order to demonstrate the effectiveness and robustness of the proposed multiple-object image retrieval (MOIR) framework, in this chapter we show the result of further applying MOIR framework to the problem of sub-image retrieval on fine art painting images collected from the Internet.
The two questions we want to answer with this study are: 1) Can MOIR be effectively applied to sub-image retrieval, without changing any core technique or framework architecture? 2) When applied to a different image data set that has totally different characteristics than the image set based on which we have developed and tested the proposed framework, how effective/robust will MOIR be, without any parameter tuning or adaptation? The first question is related to the analysis of the extent of applicability of MOIR; and the second question is related to the robustness testing of MOIR.

B. Experiments
B.1 Dataset Description

Unlike the Corel image database that contains mostly natural scene images, the fine art painting images are drawing pieces created by artists, and thus, they have totally different properties than that of natural scene images. In addition, the growing popularity of image search on the Internet has brought researchers’ attention to retrieving and indexing web images. Unlike the Corel image database that comprises high quality Photo CD images, the quality of web images varies widely because these images are created by heterogeneous sources on the Internet. In order to evaluate the robustness of the proposed multiple-object image retrieval (MOIR) framework in retrieving web images, we further collected a set of fine art painting images from the Internet and performed more evaluations.

To procure our fine art image dataset, we used Google Image Search to find and download more than 2,000 fine art images from the Internet. Subsequently, we removed redundant images in our fine art image collection using Message-Digest Algorithm 5 (MD5).
After filtering, our dataset contains 1978 non-redundant (oil/water color) painting images from the web. Each image in the dataset is manually verified and annotated with the artist and the title of the piece. The artist and the title details of this fine art image collection are as listed below.

**Claude Monet (5 pieces with 129 images)**

- Camille Monet in Japanese Costume (16)
- Canal in Amsterdam (19)
- Jar of Peaches (25)
- The Grand Canal (48)
- The Green Wave (21)

**Edouard Manet (17 pieces with 547 images)**

- Berthe Morisot (40)
- Boating (28)
- Bunch of Asparagus (12)
- Execution of Emperor Maximilian of Mexico (31)
- Flowers in a Crystal Vase (26)
- House in Rueil (35)
- Mlle Victorine in the Costume of a Matador (12)
- Music in the Tuileries (47)
- Nana (35)
- Olympia (40)
The Dead Christ with Angels (17)
The Luncheon on the Grass (61)
The Old Musician (31)
The Railway (49)
The Rue Mosnier with Flags (24)
The Spanish Singer (31)
Young Flautist (28)

**Pablo Picasso (12 pieces with 365 images)**

Card Player (12)
Dora Maar au Chat (57)
La Lecture (22)
Les Demoiselles d'Avignon (83)
Les Noces de Pierrette (24)
Massacre in Korea (35)
Nude under a Pine Tree (27)
Nude with Joined Hands (4)
Reading Woman (9)
Seated Bather (29)
Studio with Plaster Head (31)
Violin and Grapes (32)

**Pierre-Auguste Renoir (18 pieces with 695 images)**
A Girl with a Watering Can (43)
By the Water (12)
Claude Monet Painting in His Garden at Argenteuil (22)
Dance at Bougival (62)
Dance at Le Moulin de la Galette (52)
Dance in the City (57)
Dance in the Country (37)
Girl Braiding Her Hair (30)
Girl with a Hoop (23)
Girls at the Piano (61)
Jeanne Durand-Ruel (16)
Julie Manet with Cat (27)
Luncheon of the Boating Party (73)
Mme. Charpentier and Her Children (37)
On the Terrace (57)
Portrait of Ambroise Vollard (18)
The Swing (44)
The Theater Box (24)

**Van Gogh (8 pieces with 242 images)**
Field with Poppies (19)
Irises 1 (45)
Irises 2 (5)
Iris 3 (10)

Starry Night (81)

Tree Roots and Trunks (28)

Undergrowth with Two Figures (36)

Wild Roses (18)

In summary, this dataset consists of 1978 images, covering 60 distinct pieces of artworks from 5 artists. Figure 31 demonstrates some sample images of one of the pieces in our fine art image dataset.
It is worth noticing that images in our fine art image dataset are procured from various sources on the Internet, and thus, the quality of these web images varies. In addition, we can observe that some images are photos containing the piece taken from different angles; some images may have a slight color difference, and some images are close-up views of the
original piece. All of these can add complexity to this sub-image retrieval problem, causing potential false positives and false negatives during retrieval.

B.2 Performance Comparison

To demonstrate the effectiveness of the proposed multiple-object image retrieval (MOIR) framework in retrieving fine art images, we compare MOIR framework to other commonly used approaches, including integrated region matching (IRM) and full image search (FIS). For FIS framework, we also extract different color features, including 6-bit color code histogram (CCH) and MPEG-7 dominant color descriptor (DCD) to represent images. Further, in order to make the experiment a fair comparison, we use one-class SVM as the relevance feedback learning mechanism for all the frameworks in comparison. Moreover, as mentioned in Section E.3, Chapter 4, sometimes there is no relevant image in the returned top 20 images at the initial retrieval, which may fail some relevance feedback learning algorithms, such as FIRM. The reason is that these algorithms only take into account the relevant images in the user’s feedback. Therefore, five randomly selected relevant images are automatically inserted into the top 20 returned images as the top 5 images at the initial retrieval, serving as the starting point for all the algorithms used in this comparison.

It is worth noticing that in this experiment we do not automatically add 5 relevant images in the initial retrieval since one-class SVM adopted can learn users’ intention from relevant and irrelevant image in users’ feedback.
All frameworks are evaluated with the mean average precision (MAP) metric on the basis of 270 query images, including 9 pieces from 4 artists, selected from the fine art image collection as shown in Table 7.
Table 7: Details of the 270 query images in fine art painting image retrieval

<table>
<thead>
<tr>
<th>Artist</th>
<th>Title</th>
<th>Number of Query Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Claude Monet</td>
<td>The Green Wave</td>
<td>21</td>
</tr>
<tr>
<td>Edouard Manet</td>
<td>House in Rueil</td>
<td>35</td>
</tr>
<tr>
<td>Edouard Manet</td>
<td>The Railway</td>
<td>49</td>
</tr>
<tr>
<td>Pablo Picasso</td>
<td>Reading Woman</td>
<td>9</td>
</tr>
<tr>
<td>Pablo Picasso</td>
<td>Nude under a Pine Tree</td>
<td>27</td>
</tr>
<tr>
<td>Pablo Picasso</td>
<td>La Lecture</td>
<td>22</td>
</tr>
<tr>
<td>Pierre-Auguste Renoir</td>
<td>On the Terrace</td>
<td>57</td>
</tr>
<tr>
<td>Pierre-Auguste Renoir</td>
<td>Girl with a Hoop</td>
<td>23</td>
</tr>
<tr>
<td>Pierre-Auguste Renoir</td>
<td>Julie Manet with Cat</td>
<td>27</td>
</tr>
</tbody>
</table>

270

Figure 32 presents the performance comparison results of fine art image retrieval. Based on 270 queries, MOIR is comparable to IRM+SVM (those two are the best among all) in the initial retrieval and significantly outperforms all the other methods, i.e., IRM+SVM, FIS+SVM (CCH) and FIS+SVM (DCD), from the 2nd feedback iteration onward. At the 4th feedback iteration, the proposed MOIR framework significantly outperforms IRM+SVM,
FIS+SVM (CCH), and FIS+SVM (DCD) by 3.92%, 11.58%, and 12.98%, respectively. The MAP value of our algorithm also increases monotonically through the feedback iterations, again indicating the effectiveness and robustness of the relevance feedback.

In addition, Figures 33, 34, and 35 exemplify the top 20 retrieved images at the initial retrieval, after the 1st feedback, and after the 4th feedback iterations for a sample query image and their corresponding AveP values, respectively. It is worth mentioning that the proposed MOIR framework can not only find images with different view angles and cropping as respectively shown in the 12th and 16th ranked images in Figure 33, but also retrieve even photos containing the user-desired piece of artwork as shown in the 12th ranked image in Figure 35.
Figure 32: Fine art painting image retrieval results (270 queries)
Figure 33: An example of MOIR fine art image retrieval results – initial retrieval.
Figure 34: An example of MOIR fine art image retrieval results – 1st feedback.
Figure 35: An example of MOIR fine art image retrieval results – 4th feedback.
C. Summary

In summary, based on experimenting with the 1978 available web fine art images we collected and manually annotated, the proposed multiple-object image retrieval (MOIR) framework has been shown to be directly applicable and an effective approach in sub-image retrieval. Not only can our framework search high quality natural scene image databases such as the Corel image collection effectively, it can also be effectively applied to retrieving artificial images with different visual characteristics and various qualities such as the fine art painting images collected from the Internet.
CHAPTER 6
CONCLUSIONS AND FUTURE WORK

In this dissertation, an image retrieval framework (MOIR) based on hierarchical image representation is developed for multiple-object retrieval. We evaluate the performance of the proposed image retrieval framework in terms of efficacy and efficiency. Finally, we demonstrate an application of the proposed framework to fine art painting image search. In this chapter, we summarize the contributions of this dissertation, and explore possibilities for future work.

A. Summary of Contribution

In this dissertation, the novelties of the proposed research are manifold as listed below.

1) We propose a multi-resolution, hierarchical segmentation algorithm (MRIA) that performs image segmentation and hierarchical region tree construction simultaneously across multiple resolutions. MRIA is different from existing hierarchical image segmentation. Existing methods need to perform full segmentation at each resolution, while MRIA segments an image starting at a very low resolution, and proceeds to segmentation at a higher resolution only when needed. Therefore, MRIA is more efficient than existing methods.
2) We design an efficient hierarchical tree comparison algorithm which greatly reduces the computational cost involved in similarity comparison.

3) The proposed MOIR system enables customized (user-guided) multiple-object retrieval, i.e., to retrieve images where all the user-desired objects appear. User feedback is necessary to discover user interest but asking the user to manually outline all the objects of his/her interest is a tedious (labor-intensive and user-unfriendly) process. Instead, we ask a user to click on only a small number of relevant images that contain all of his/her desired objects. This is semi-supervised learning with extremely incomplete user input and a very small training set, with two critical pieces of information missing: the number of objects desired and what they are, making the problem space two magnitudes larger than the whole-image retrieval; thus it is very hard to design a learning algorithm that can learn effectively and also has reasonably low complexity. To our best knowledge we are the first to explore the combined problem of hierarchical image segmentation, representation, and customized object-based image retrieval and to have made non-trivial progress.

In summary, this research provides a human-centered image retrieval framework which allows users to perform multiple-object retrieval in image databases. To enhance the retrieval performance, this innovative framework seamlessly integrates a multi-resolution hierarchical segmentation algorithm which produces image segmentation results and a region-based hierarchical tree concurrently in an efficient and effective way. In addition, introducing the region-based hierarchical tree can preserve the relations among segmented regions, which also eases the over-segmentation issue in the subsequent object matching process. Fur-
ther, the proposed sub-tree comparison approach provides an efficient way of performing object matching and multi-object retrieval. Moreover, we maximize the usage of the user’s feedback in query refinement and avoid the expensive proper sub-tree comparison in the feedback process. By means of the seamless integration of the user’s relevance feedback (RF) with the proposed multiple-object image retrieval (MOIR) system, it allows automatic discovery of the object(s) of the user’s interest and improves the retrieval accuracy through feedback-retrieval loops. Finally, we have demonstrated that the proposed MOIR framework is a generalized approach for image retrieval, which can deal with not only the natural scene images but also the artificial images such as fine art painting images.

B. Future Work

The work presented in this dissertation is just the first step to explore the field of multiple-object image retrieval. In this section, we outline several possible research directions that will be investigated as future work.

B.1 Image Segmentation

The proposed multiple-object image retrieval (MOIR) framework adopts hierarchical image representation and hierarchical region tree matching to cope with the problem of over-segmentation. Although the problem of under-segmentation is temporarily relieved by applying the proposed approach on over-segmented images, the under-segmentation problem still
remains unresolved. To further enhance the effectiveness of MOIR framework, it is essential to improve the image segmentation algorithm as one of our future works.

As we demonstrated in Chapter 1 that supervised image segmentation framework, such as CO3, can produce promising segmentation results by learning from users’ elaborate scribbling process. This indicates the inevitable trade-off between image segmentation quality and user involvement. Therefore, it is possible to get a reasonable segmentation quality with minimum input from the users. We believe that an image segmentation algorithm with semi-supervised learning mechanism may be a possible direction to resolve the problem of inaccurate image segmentation.

B.2 Feature Representation

In this dissertation work, we extract color features for representing the properties of segmented regions since color features are relatively stable and robust to image transformation. In fact, any visual features that are suitable for segmentation and robust during down-sampling can be readily plugged into our system, though our focus is not to explore the best features for segmentation or image retrieval. We will be working on integrating more visual features in the future.

For example, for fine art image retrieval, it is possible to introduce scale-invariant feature transform (SIFT) feature to represent the characteristics of regions. This is because SIFT feature is relatively stable and robust to image transformation and it has been shown an effective feature representation for exact object matching. Therefore, we plan to improve the
performance of the fine art image retrieval system by introducing the SIFT feature representation as one of our future works.

In addition, it is quite often to see the need to retrieve intensity images in the real-world. For example, most medical images such as x-ray image belong to intensity images. However, the proposed MOIR framework is design for retrieving color images, and thus, the current approach is not robust to gray-level images. This may limit the applicability of the proposed multiple-object retrieval framework. In order to apply the proposed approach to other fields, we need to cope with the abovementioned issue by introducing or designing suitable feature representation for intensity images as another future work.

B.3  Image Scene Retrieval

As aforementioned that users are usually interested in searching images on the basis of the level 2 (object) or level 3 (scene) abstractions. Scene-based image retrieval is more challenge since it involves not only the relationships among objects but also the psychological interpretation of the images. Although this dissertation mainly targets object-based image retrieval, which focuses on level 2 abstraction, the proposed framework can be further extended to scene-based image retrieval, which focuses on level 3 abstraction, by utilizing the spatial and neighbor relationships preserved in the hierarchical image representation and using relevance feedback technique to procure psychological interpretation of the images from the users.
B.4  \textit{Extend MOIR Framework to Audio and Video Retrieval}

In addition to image retrieval, searching audio and video data are two other lines of research in content-based document retrieval. Similar to image segmentation, both audio and video data can be partitioned into segments based on temporal or spatio-temporal information (Zhang, Whalley, & Brooks, 2009; Grundmann, Kwatra, Han, & Essa, 2010). This implies that audio and video data can also be hierarchically represented. Therefore, it is highly possible that the proposed MOIR framework can be applied to the content-based audio and video retrieval. We plan to further extend the proposed MOIR framework to content-based audio and video retrieval as one of our future works.
LIST OF REFERENCES


Bayes/MDL for multiband image segmentation, *IEEE Transactions on Pattern Analysis and Ma-
chine Intelligence*, 18(9), 884-900.