

Visual Intelligence for Online Communities

Commonsense Image Retrieval by Query Expansion

James Jian Dai

B.S. Computer Science
University of British Columbia
June 2002

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
in partial fulfillment for the degree of
Masters of Science in Media Arts and Sciences
at the Massachusetts Institute of Technology, June 2004.

Copyright Massachusetts Institute of Technology 2004. All rights reserved.

Author

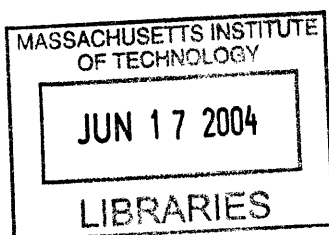
James Jian Dai
Program in Media Arts and Sciences

Certified by

John Maeda
Associate Professor of Design and Computation
Thesis Supervisor

Accepted by

Andrew B. Lippman
Chair, Departmental Committee on Graduate Studies
Program in Media Arts and Sciences



ROTCH

Visual Intelligence for Online Communities

Commonsense Image Retrieval by Query Expansion

James Jian Dai

B.S. Computer Science, University of British Columbia, June 2002

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning, in partial fulfillment for the degree of
Masters of Science in Media Arts and Sciences at the
Massachusetts Institute of Technology, June 2004

Abstract

This thesis explores three weaknesses of keyword-based image retrieval through the design and implementation of an actual image retrieval system. The first weakness is the requirement of heavy manual annotation of keywords for images. We investigate this weakness by aggregating the annotations of an entire community of users to alleviate the annotation requirements on the individual user. The second weakness is the hit-or-miss nature of exact keyword matching used in many existing image retrieval systems. We explore this weakness by using linguistics tools (WordNet and the OpenMind Commonsense database) to locate image keywords in a semantic network of interrelated concepts so that retrieval by keywords is automatically expanded semantically to avoid the hit-or-miss problem. Such semantic query expansion further alleviates the requirement for exhaustive manual annotation. The third weakness of keyword-based image retrieval systems is the lack of support for retrieval by subjective content. We investigate this weakness by creating a mechanism to allow users to annotate images by their subjective emotional content and subsequently to retrieve images by these emotions.

This thesis is primarily an exploration of different keyword-based image retrieval techniques in a real image retrieval system. The design of the system is grounded in past research that sheds light onto how people actually encounter the task of describing images with words for future retrieval. The image retrieval system's front-end and back-end are fully integrated with the Treehouse Global Studio online community – an online environment with a suite of media design tools and database storage of media files and metadata. The focus of the thesis is on exploring new user scenarios for keyword-based image retrieval rather than quantitative assessment of retrieval effectiveness. Traditional information retrieval evaluation metrics are discussed but not pursued. The user scenarios for our image retrieval system are analyzed qualitatively in terms of system design and how they facilitate the overall retrieval experience.

Thesis Advisor: John Maeda

Title: Associate Professor of Design and Computation

Visual Intelligence for Online Communities

Commonsense Image Retrieval by Query Expansion

James Jian Dai

B.S. Computer Science
University of British Columbia
June 2002

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
in partial fulfillment for the degree of
Masters of Science in Media Arts and Sciences
at the Massachusetts Institute of Technology, June 2004.

Thesis Reader
Walter Bender
Executive Director, MIT Media Lab
Senior Research Scientist, Program in Media Arts and Sciences

Thesis Reader
Whitman Richards
Professor of Cognitive Sciences
Computer Science and Artificial Intelligence Laboratory

To my parents, without you I would not be here.

Acknowledgements

First of all I would like to thank my thesis readers. Thank you Walter Bender for bringing the entirety of your expertise and experience to help shed light on the process and to steer me in the general direction of here. Thank you Professor Richards for being utterly honest with me about the bad and the good and for giving my own passions relevance. If I come back to MIT in the future, it will be because this place was good enough to keep a professor as respectable, genuine, professional and caring as you.

Secondly I would like to thank the people who gave me a nurturing place to explore my passions. Thank you members of the Theater Arts faculty who I will be forever grateful to have worked with: Alan Brody, Michael Ouellette, Tommy DeFrantz, Janet Sonenberg and Jay Scheib. Thank you for giving this rather assuming graduate student time, place and mind to be nosey. Janet, thank you for helping me in my time of need and for caring about me since the first e-mail through the ether. Jay, thank you for your trust and friendship. Welcome to MIT, please push in all directions. Thank you to the Rinaldi collective for your smiles and enthusiasm. Diane Brainerd, Bill Fregosi, Leslie Held, Mike Katz, Karen Perlow and Elizabeth Jochum, you are the true puppet masters of theater at MIT. Michele Oshima, thank you for kicking my butt in the right direction and for being yourself.

Thirdly I would like to thank my friends. It is my desire and curiosity to understand the invisible strings that bind us that makes me truly believe there is something worthwhile to be captured on stage. You know who you are. Thank you suitees (Grace, Roger, Aditi, Bing) for giving me a home. Thank you Aaron for the fun. Thank you Parmesh for reaching out and for being your beautiful self. Thank you Austin, Jinger, Jacob, Gustavo, Hugo, Helen, Bushra, Melissa, Hiroko, Basel, Ashlie for being my friends. Thank you Karen, Anna and Sarah for the memories of me fumbling towards myself. Thank you to the PLW team (Carlos, Noah, Pat, Marc, James, Ben, Allen) for the inspiration and companionship. Thank you the Grad Ring team (Alvar, Justin, Lucy, Akshay, Peg) for making a tradition with me. Thank you Mike Wu and Maria Klawe for setting me on the course to graduate school in the first place. Thank you Hilary for being there for me at all hours and for putting up with me. And thank you Stella for giving my life bite.

Finally, I would like to thank the person who adopted me last summer. But first, thank you Linda Peterson, Pat Solakoff, Mitch Resnick, Hiroshi Ishii, Walter Bender, Glorianna Davenport, James Seo, Anindita Basu, Andrew Sempere, Kimiko Ryokai and Janet Sonenberg for helping me find my advisor, John Maeda. John, words fail to describe how fortunate I feel to have a mentor who is truly the wind beneath my wings. Thank you for your honesty, trust and care...

If I follow my heart and do something I am proud of in the world, it is because I met you.

Table of Contents

1 Introduction.....	10
1.1 The Problem.....	10
1.2 The Need.....	11
1.3 The Approach.....	12
2 Extended Example	13
2.1 Scenario – Image retrieval by exact keyword matching.....	13
2.2 Scenario – Query expansion by dictionary definition	13
2.3 Scenario – Query expansion by conceptual relationships	14
2.4 Scenario – Retrieval by emotional content	15
3 Theoretical Background	16
3.1 Overview of Image Retrieval Research.....	16
3.2 Theories of Perception.....	16
3.2.1 Structuralism	17
3.2.2 Gestalt Theory	18
3.2.3 Marr’s Computational Theory of Perception	19
3.2.4 The Importance of Social Context and Subjective Emotions	20
3.3 Approaches to Image Retrieval.....	21
3.3.1 Content-Based Image Retrieval.....	22
3.3.2 Keyword-Based Retrieval	24
3.3.3 Integrated Approaches to Image Retrieval.....	29
3.4 Focus of Thesis	29
4 System Design and Implementation.....	31
4.1 The Treehouse Studio Framework	31
4.2 The Image Retrieval System.....	33
4.2.1 Image Annotation Application	33
4.2.2 Image Browse and Retrieval Application	35
4.3 User Interaction in the Image Retrieval Application.....	50
4.4 Image Retrieval System Design	51
4.4.1 Server Client Architecture.....	51
4.4.2 Image Database Architecture	53
4.5 User Scenarios Retraced	54
4.6 Image Retrieval System Summary	55
5 Evaluation.....	57
5.1 Functional comparisons with other image retrieval systems.....	57
5.2 Traditional Information Retrieval Evaluation Methods	59
5.3 User Feedback.....	60
5.3.1 Keyword annotation and knowledge aggregation feedback	60
5.3.2 Query expansion feedback	61
5.3.3 Emotional content retrieval feedback	62
5.4 Future extensions.....	62
6 Conclusions.....	64
7 References	65

List of Figures

Figure 1 – September 11, 2001	10
Figure 2 – How many words is this picture worth?	12
Figure 3 – Exact keyword retrieval example	13
Figure 4 – Dictionary definition example	14
Figure 5 – Conceptual relationship retrieval example	15
Figure 6 – Emotion retrieval example	15
Figure 7 – Numbers “4” and “5” are hidden	17
Figure 8 – Numbers “4” and “5” are visible here	18
Figure 9 – Grouping by proximity example	19
Figure 10 – Another grouping by proximity example	19
Figure 11 – Grouping by similarity example	19
Figure 12 – Which direction do the triangles point?	19
Figure 13 – Taxonomy of image representation methods	21
Figure 14 – Image query in a hypothetical feature space of RGB colors	22
Figure 15 – Example of keywords of an image in the PictureQuest database	24
Figure 16 – Framework for classifying visual descriptors	27
Figure 17 – Treehouse login page	31
Figure 18 – Treehouse user home page	32
Figure 19 – The vector-based Draw application	32
Figure 20 – The pixel-based Photo application	33
Figure 21 – The Annotation Application	34
Figure 22 – The Image Retrieval Application	35
Figure 23 – Keyword tab – direct retrieval	36
Figure 24 – Exact keyword matching	37
Figure 25 – Image retrieval by query expansion	37
Figure 26 – WordNet senses for “dog”	38
Figure 27 – Hyponyms of two definitions of “dog”	39
Figure 28 – A sample concept tree	39
Figure 29 – Insertion of new terms into the DSH	41
Figure 30 – The Nouns tab – query expansion by WordNet	42
Figure 31 – A second definition of “apple”	43
Figure 32 – Another example of WordNet query expansion	43
Figure 33 – Using OpenMind to expand queries by conceptual relationships	45
Figure 34 – A theory of basic emotional groups	46
Figure 35 – Emotional response annotation with the Annotation Application	47
Figure 36 – Retrieval by emotional response	48
Figure 37 – Retrieval of similar emotions	48
Figure 38 – Retrieval of opposite emotion	49
Figure 39 – Aggregate user feedback for emotional retrieval	49
Figure 40 – Nouns tab links	51
Figure 41 – Context tab links	51
Figure 42 – Image Retrieval System Client-Server architecture	52
Figure 43 – PictureQuest image database and keyword-based retrieval system	58
Figure 44 – Google Image Search for “kiwi”	58

List of Tables

Table 1 – Distribution of Classes by average percentage across three describing tasks...	25
Table 2 – Class frequency by percentage for describing tasks and Sorting task.....	26
Table 3 – Emotional keyword synonyms	47
Table 4 – The dictionary table in the Treehouse image database.....	53
Table 5 – Treehouse database table matching images to keywords	53
Table 6 – Treehouse database table for emotion keywords	54
Table 7 – Treehouse database table for emotional keyword aggregate	54
Table 8 – Information retrieval evaluation metrics.....	59

1 Introduction

1.1 The Problem



Figure 1 – September 11, 2001

The above picture (Figure 1) is an image of the second hijacked airplane about to slam into the World Trade Center on September 11, 2001. How would you describe this to a friend who has never seen it? Would you describe the contents of the image (twin towers, airplane, smoke), or would you describe the historical context around the event itself (the worst terrorist attack on U.S. territory in history), or would you describe the horrific emotional landscape that gripped the entire nation that will be forever associated with this image?

How would you describe this image so that you can retrieve it sometime in the future?

Herein lies the complexity of image retrieval. The task could be approached from many different angles that involve very different disciplines. We could try to describe the picture as a collection of low-level pictorial features such as color, line and texture. This approach taps into theories of perception from *Psychology* (Doorn and Vries 2000) and

Cognitive Science (Pinker 1985; Wade and Swanston 1991; Hoffman 1998). The *Computer Vision* (Marr 1982) and *Image Processing* (Veltkamp, Burkhardt et al. 2001) communities have explored ways of automating the low-level feature extraction process. We could also describe the picture in terms of its composition of pictorial parts, taking into account the layout and balance of the visual elements. This approach requires knowledge of how the human visual system works and how the eye travels in the image. We would be starting from the fields of *Psychology* and *Neurobiology* as well as *Graphic Design* (Dondis 1973). Alternatively, we could try to come up with meaningful categories by which we could distinguish this picture from all other pictures. This approach could lead us to note the social context around the event depicted in the picture as well as the biography of the photographer. We could also categorize the image by its medium of photography as opposed to a painting or drawing. Describing an image by its medium and context would go beyond the basic pictorial elements in the image itself. There is a long history of coming up with categorical structures for organizing images in *Library Systems* (Jorgensen 2003) and *Information Sciences* (Salton and McGill 1983) that often make use of ideas from the study of *Art History* (Berger 1973). There is a rich history of *Information Retrieval* (Jing and Croft 1994; Lim, Seung et al. 1997) research that have focused primarily on the retrieval of text-based documents with techniques that may be applicable to *Image Retrieval* (Jorgensen 2003).

The needs for visual information retrieval differ greatly based on the task domain and the numerous approaches originating from vastly different disciplines illustrate this fact. For some domains, such as sorting medical scans, automated methods of object recognition may be enough. For others, for example, a theater designer trying to find an image of a flag that conveys a sense of tension to be used as part of a scenic set, an approach that just takes into account color and texture is simply not enough. The image's meaning and emotional evocativeness must also be considered.

1.2 The Need

As our personal and professional collections of digital visual data grow at a rapid rate, there is an increasing need to address digital image browsing and searching. With the billions of images available on the Internet there is an urgent need for systems that organize them and help us quickly find what we need. Libraries have very extensive categorical systems that help us find a few books among millions. What would be a solution that helps us organize images? An image maybe worth a proverbial thousand words but how many words does it take to capture the essence of an image so that they can be used to later retrieve that image accurately? Should we try to use words at all? Is there a way to automate the categorization process? Can computers help? Please take a look at the picture in Figure 2 below.

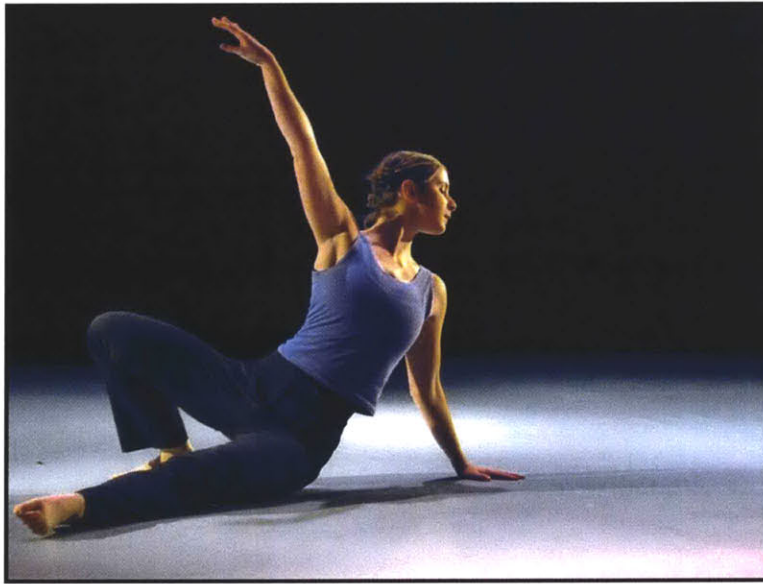


Figure 2 – How many words is this picture worth?

Would this picture be adequately described as “a young woman dancing”? What about the mood of the picture and the emotions it elicits? Just like art appreciation, our visual experience is highly visceral and subjective (Finke 1989). How can we build image retrieval systems that respect the human aspects of visual perception since humans are predominantly visual creatures (Sekuler and Blake 2001)?

1.3 The Approach

This thesis examines image retrieval by exploring ways to facilitate image retrieval in terms of how people see images and what people see in images. The task domain is focused primarily on personal digital images but the methods can be easily extended to other types of images and media. We examine how real people actually describe images and use this data to inform the system design and development.

Our approach is to explore image retrieval by keywords that users have associated with them. In Chapter Three, we discuss the theoretical foundation and background to this approach and in Chapter Four we present the design and implementation of an image retrieval system that explores three traditional shortcomings of keyword-based image retrieval:

1. The requirement of heavy manual annotation of keywords for images.
2. Overcoming the hit-or-miss nature of exact keyword matching.
3. Categorizing and retrieving images by their emotional content.

In Chapter Two, we provide an extended example of the power and accessibility of this approach. In Chapter Five we examine the effectiveness and weakness of our image retrieval system and discuss ways to extend it. Finally, we present a summary of our exploration in Chapter Six.

2 Extended Example

2.1 Scenario – Image retrieval by exact keyword matching

Jordan is a junior at college who has been playing with the Treehouse online community for a couple of weeks. He has browsed the image database and added some keywords to a few images that he thought was interesting. For a humanities class Jordan was asked to write a paper on the topic of “The Fantastical Inner Life of Poets”. He wants to use an evocative image to go with the title page of his paper. Jordan turns to the Google Image Search (Google 2004) and tries a few searches with the search words of “life” and “poets”. The results are either mundane images from personal websites or graphic icons from websites (such as “poets.gif”). None of them strike him as fitting for what he had in mind. He suddenly remembers that he had previously annotated a Treehouse image with the word “fantastical” because he had thought it was interesting. If his memory serves him right, that image might just work. He logs into the Treehouse image retrieval system and types in the keyword “fantastical” and lo and behold the image he remembers appears (Figure 3).

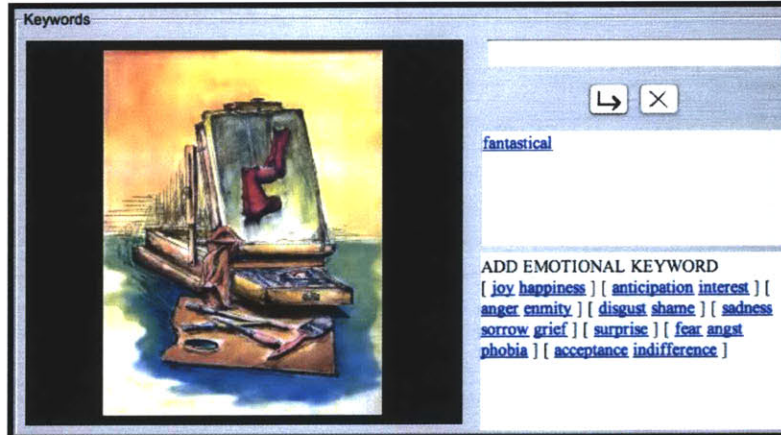


Figure 3 – Exact keyword retrieval example

This image will do just fine, Jordan muses. He retrieves it and inserts it into his paper.

2.2 Scenario – Query expansion by dictionary definition

Michael is a twelve-year-old student in seventh grade doing his homework. The assignment is to create a collage of pictures of fruits. He logs onto the Treehouse online community and begins searching for pictures of fruits he can use. Using the direct image search, he types in “apple” and gets some good looking pictures of apples. He does the same for some “peach” but that query returned no results. Frustrated, Michael clicks on

the results tab labeled *Nouns*. To his surprise this page is filled with images of other types of fruits. There is an image of a banana and orange among other things. Somehow the system picked up the fact that “apple” is a fruit and has returned other images that were annotated with the names of other types of fruits from the database (Figure 4).

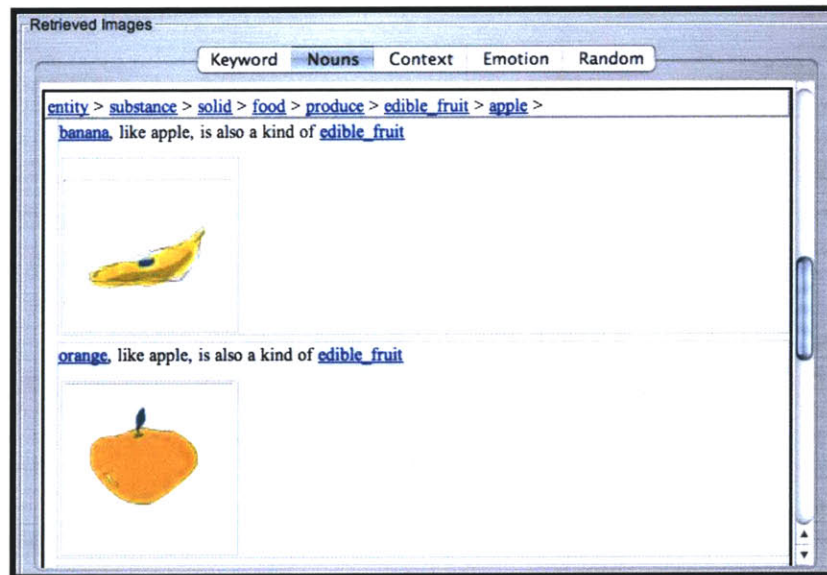


Figure 4 – Dictionary definition example

Michael picks the images he likes and adds them to his project.

2.3 Scenario – *Query expansion by conceptual relationships*

Joanna is a 67-year-old grandmother who wants to send out invitations to her grandchildren to visit her in sunny California. She decides to make a personal invitation card with pretty pictures for the children. She logs onto the Treehouse online community and searches for pictures of the beach since she lives right next to a beautiful sandy stretch. She puts “beach” as the search word and sends in the query. No images that were annotated with the word “beach” were found. However, upon examination of the *Context* results tab, she sees that the system has returned images that were annotated with words that are somehow conceptually related to the word “beach”. For example, the system returned pictures tagged with the words “ocean” and “sand” (Figure 5).

She decides that some of these could be useful and downloads them.

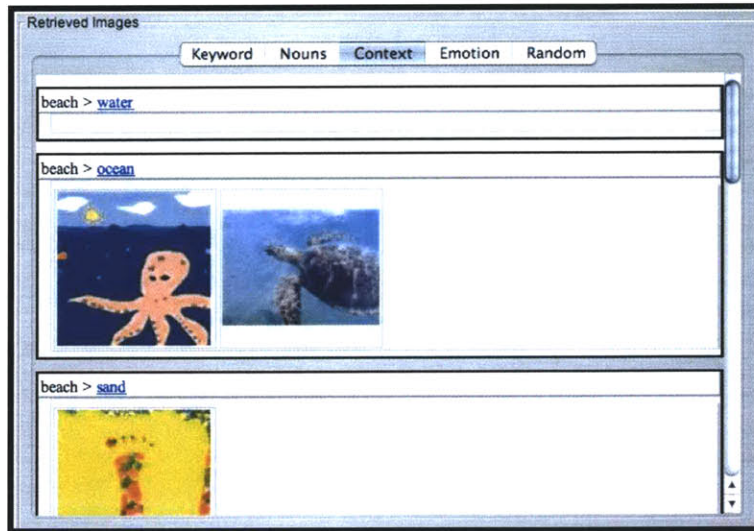


Figure 5 – Conceptual relationship retrieval example

2.4 Scenario – Retrieval by emotional content

Peter is a teenager in high school who wants to become a visual artist. His friend Sarah has been hospitalized for a week from a skiing accident. Peter loves pictures and wants to send Sarah a picture to cheer up her day. He logs onto the Treehouse image retrieval system. He clicks on the *Emotions* tab and initiates a request for pictures that members of the community have annotated to elicit the emotion of “happiness”. He chooses a funny sketch of a frog and downloads the image (Figure 6).

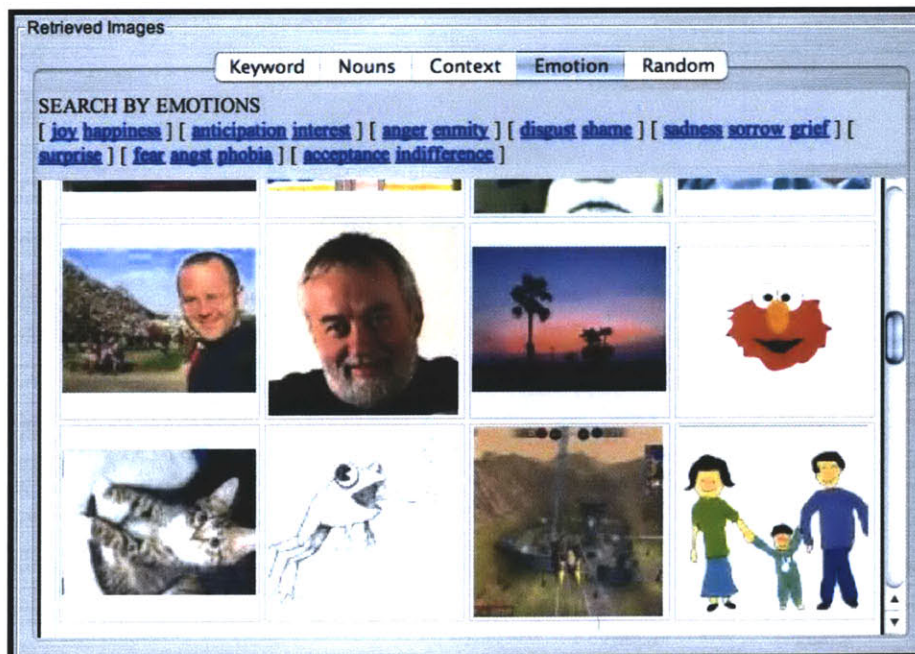


Figure 6 – Emotion retrieval example

Peter prints out the image and makes it into a card for Sarah.

3 Theoretical Background

3.1 Overview of Image Retrieval Research

The term *image retrieval* refers to the task of finding one or more images from a set of many (Jorgensen 2003). This includes both *image searching* where a user is hunting for a specific image that they have in mind as well as *image browsing* where the user has only a rough idea of the kind of image they are after (Kuchinsky, Pering et al. 1999). Image browsing often takes place by the user navigating through some sort of structure by which all images in the system are organized (Yee, Swearingen et al. 2003). The system's structure allows the user to examine sub-collections of similar images. A user could approach the task of *image retrieval* simply by examining the entire collection one by one. This would be analogous to looking for a book in the library by looking at each book one by one – not a very efficient route. Libraries have extensive systems for organizing books logically to help users find books. This organizational data may include title, author, date, genre, type and call number. Image retrieval systems can also help users by associating descriptive data with images (Yee, Swearingen et al. 2003). This descriptive data (called *meta-data* because it is data about data) can help organize images logically to facilitate their retrieval. Retrieving images would involve the user specifying a search query in terms of the descriptive data the system uses to organize the images. This search query represents the image or type of image the user would like to find.

Existing image retrieval systems use a variety of methods to organize and describe images. Research in this field can be roughly divided into two ways of representing images. How an image retrieval system represents an image is very important as the representational scheme directly affects how users engage with the system (since the search query needs to be expressed in terms of how the image is stored). One representational scheme, here termed Content-Based Image Retrieval, describes images by visual features that can be automatically detected by image processing. The other scheme, here termed Keyword-Based Retrieval, uses words as the descriptive data. The following sections go over some general theories of perception that image retrieval research stands upon. We then discuss Content-Based Image Retrieval and Keyword-Based Retrieval in terms of these theories of perception and expound the strengths and weaknesses of their representational schemes and retrieval mechanisms. This chapter ends describing the focus of research presented in this thesis and how it relates to past image retrieval research.

3.2 Theories of Perception

Humans are predominantly visual creatures, how people see images is directly related to how people see the world around us; thus image retrieval research is strongly rooted in

theories of perception (Doorn and Vries 2000). In this section we will describe some ideas that have been particularly influential to image retrieval research. Perception happens when stimuli from the world around us reach our senses and are cognitively processed for meaning. For example, although there are many sounds in our environment, we recognize some as words that carry linguistic meaning. Visual perception happens when light from our environment hits our retinas. The properties of the light are processed to help us recognize a visual scene; for example, the light frequencies are deduced as color (it is interesting to note that color does not exist in the world around us except in our heads). Researchers have long debated about how this visual processing takes place and what the fundamental building blocks for visual cognition are (Spoehr and Lehmkuhle 1984).

3.2.1 Structuralism

The early *Structuralism* view of perception situates the senses as independent of each other. *Structuralists* propose that for each detectable elementary physical event there is a specialized neural feature detector responding to the event and that a corresponding elementary observation (or sensation) would be experienced (Hochberg 1964). Their procedure was to discover the underlying fundamental sensations and their physiological bases, and the laws by which these elements combine. All other qualities for which we can find no such receptors (such as distance, solidity, social attributes, facial expressions, and so forth) were to be built up out of these units that were to combine by simple addition. The basic failures of this theory lay in the addition hypothesis. Many building blocks of visual perception such as color, position, shape and size, do not predict the observations made when those supposed building blocks are combined. For example, Figure 7 and Figure 8 below contain all parts of the letters “4” and “5”.

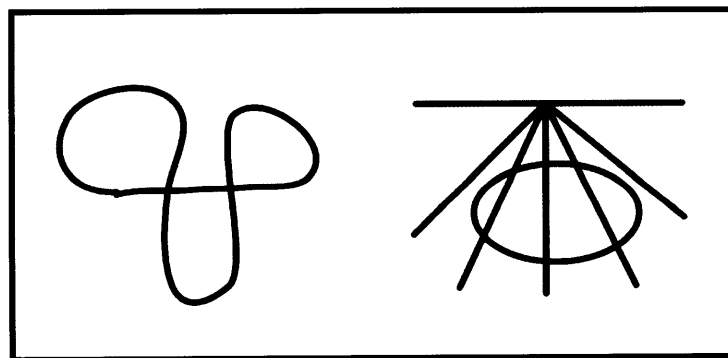


Figure 7 – Numbers “4” and “5” are hidden

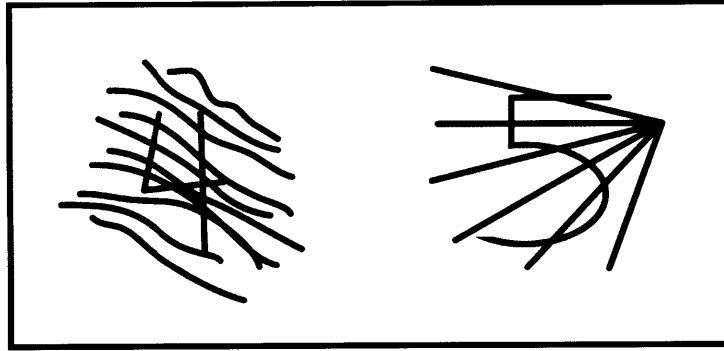


Figure 8 – Numbers “4” and “5” are visible here

In the above illustrations, all the point stimuli of the number “4” and “5” are clearly visible. The additional lines obscure the numbers in Figure 7 but not in Figure 8. This suggests that perception may be more than an aggregate of point stimuli, as Structuralism proposes, and that there are evidently laws of organization at work – factors that depend on the relationships between the parts of the patterns of stimulation (Hochberg 1964).

3.2.2 Gestalt Theory

Gestalt theory was the first serious attempt to deal with perception as other than an assemblage of independent point sensations with the goal of finding natural units of analysis of perception with which to replace the artificial sensations and to explain these new units in terms of a totally revised picture of how the nervous system works. Gestalt theorists criticized structuralism, believing that a percept is *not* composed of sensations:

If we take a new look at the world of perception, unbiased by any structuralist assumptions, what do we find as the most natural units of analysis? In the world of sight - not meaningless tiny patches of light and color, but whole shaped regions, set off or bounded by their contours, which appear the same whether they fall on one particular set of cells on the retina or on another: as you shift your gaze even slightly to one side of the number at the bottom of the page, a totally new set of cones is stimulated, yet the shape you see remains the same... before we undertake detailed psychophysical measurement, before we seek to understand the underlying physiological mechanisms, we must discover the rules that govern the appearance of shapes and forms (Hochberg 1964).

Gestalt theorists derived numerous laws of organization. Most of which can be summarized under the “law” of simplicity (“we see what is simplest to see”). Whether by early perceptual learning or by inborn arrangement, our nervous systems seem to choose those ways of seeing the world that keep perceived surfaces and objects as simple and constant as possible.



Figure 9 – Grouping by proximity example

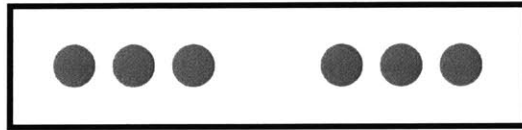


Figure 10 – Another grouping by proximity example

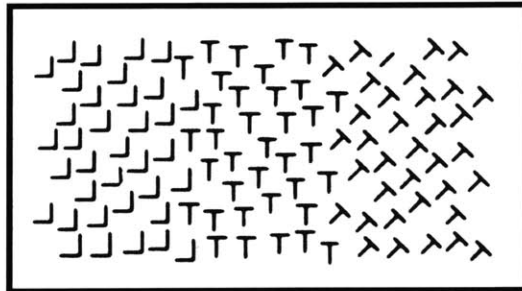


Figure 11 – Grouping by similarity example

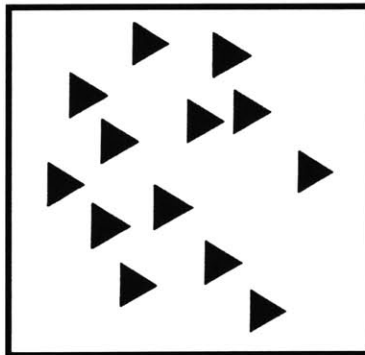


Figure 12 – Which direction do the triangles point?

For example, there are many common examples of grouping by similarity in everyday life. If like objects are located near each other in the physical world, we tend to group them together perceptually (Solso 1994). In Figure 9 we see three groups of vertical lines and in Figure 10 we see two groups of circles. In Figure 11, the T's in the center are perceived as belonging to a different class of objects. This grouping by similarity can be the most noticeable when there are conflicting possibilities. Take a look at Figure 12, in which direction do the triangles point? Does this direction change?

3.2.3 Marr's Computational Theory of Perception

David Marr's computational theory of perception begins with point stimuli of Structuralism and suggests a path to perception that incorporates the recognition of Gestalt shapes and surfaces. He believed that the processes involved in low-level vision produce a series of representations providing increasingly detailed information about the

visual environment. Three kinds of representations are identified. First of all is a primal sketch that provides a two-dimensional description of the main light-intensity changes in the visual input including info about edges, contours and blobs. Secondly a two-and-a-half dimensional sketch incorporates a description of depth and orientation of visual surfaces that makes use of info provided by color, texture and motion. Finally, a three dimensional representation describes the shapes of objects in space that is independent of the observer's viewpoint (Marr 1982). Perception involves matching the representation constructed from the visual stimulus against a catalogue of previously learned three-dimensional models stored in memory. Research in cognitive neurophysiology has found evidence to support this perspective (Doorn and Vries 2000).

3.2.4 The Importance of Social Context and Subjective Emotions

The above theories of perception focus primarily on the neurological processing of visual stimulus, they do not attempt to consider the social context and subjectivity of the perceptual experience. As suggested by the figure of the terrorist attack of September 11, 2001 in Chapter One, sometimes the social context and emotional content of an image are inseparable from the perception of an image. In considering major problems with theories of visual perception, Gordon (Gordon 1989) highlights the need for taking into account context and subjectivity: "there is a fundamental flaw in the idea that the eventual explanation of perception will be physiological: namely that neurophysiology remains 'inside' the organism, whilst perception involves the external world. Neural events may be isolated entities, but stimuli arise from within a context, a context that shapes our conscious experience. A general theory of visual will have to respect this fact, and this means that the language of such a theory is more likely to be psychological than physiological".

In research it is possible to adopt a tough-minded, behaviorist approach to perception with only controlled, measured responses. However, as human beings we are intimately aware of our own consciousness. It would be impossible for purely objective research into color vision to discover that certain color combinations are very unpleasant, or that some colors appear warm, others cold (Gordon 1989). Modern phenomenologists argue for the inclusion of subjective experience in accounts of perception. They insist that our perception of say, a house, transcends any limited vantage point: we 'see' the volume of the house, its solidity, even when the only visible aspect is the front. Our phenomenal experience includes the knowledge that we are 'inside' our bodies. We know what things would look like from alternative vantage points. It is important for image retrieval to take into account the subjective nature of human perception.

Emotions are relevant to image retrieval as they play a large part in the subjective perception of images (Solomon 2003). They affect not only visual processing but also retrieval. A person's perception and subsequent recall of an event or image is strongly dependent on their emotional state during both perception and recall (Evans 2001). It has been shown that in recalling events, people make use of emotional keywords associated with those events. Documenting and categorizing images based on affective response has not yet been explored in depth as has been done with color (Itten 1973). One obstacle is

exactly the subjective and changeable nature of affective response – that it differs person to person and time by time even for the same person. However, research has shown that most affective response to images are “either somewhat neural or high-arousal pleasant or high-arousal unpleasant” (Lang, Bradley et al. 1997). Similar reduction was found in user perception of facial expressions. The six “basic” emotion were found to reduce to a pleasantness-unpleasantness dimension (Scherer and Ekman 1984). On the pleasant-unpleasant dimension, happiness and surprise are at one end while disgust, anger and sadness are at the other. This research suggest that despite the subjective nature of emotions and perception, it may be possible to use broad categories to describe images by users’ affective responses to them (Matravers 1998).

3.3 Approaches to Image Retrieval

We stated above that research in image retrieval could be roughly divided into two approaches based on the type of descriptive data the systems use to document images. This taxonomy of representations in image retrieval is shown in Figure 13.

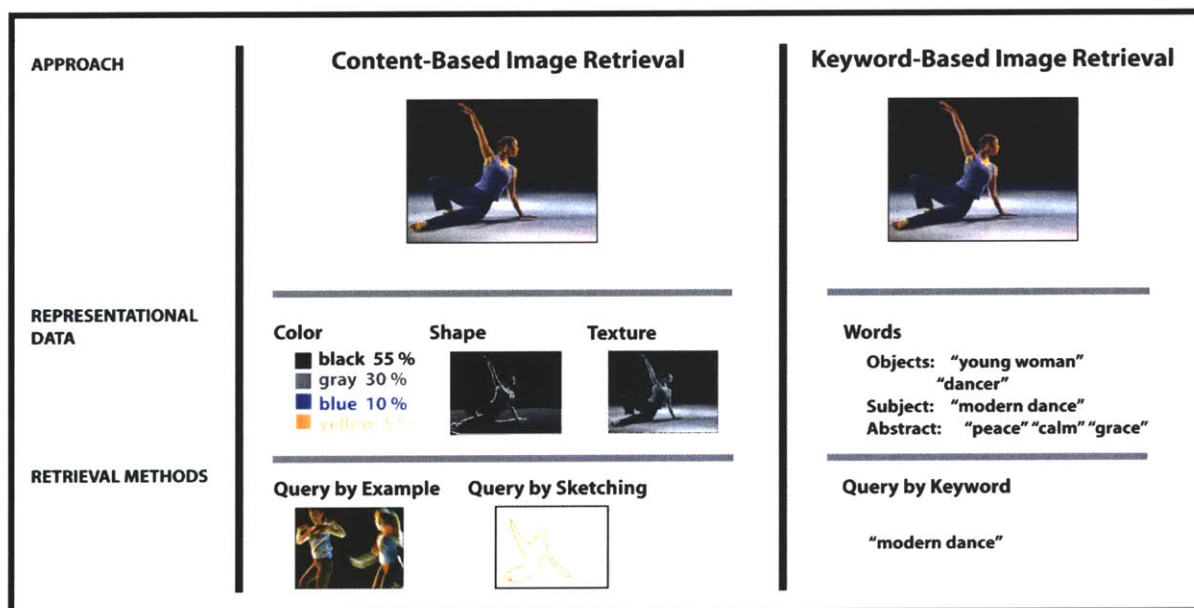


Figure 13 – Taxonomy of image representation methods

Content-Based Image Retrieval describes images by visual features that can be automatically detected by image processing while Keyword-Based Retrieval uses words as the descriptive data. We will now go over the advantages and disadvantages of both of these approaches to image representation and retrieval. It is important to note that the two approaches are not dichotomous approaches to image retrieval. They are two ways of representing images that can be used in the same system in an integrated fashion. We will describe some such integrated systems towards the end of this chapter. But first, the two approaches to image representation are described separately for clarity’s sake.

3.3.1 Content-Based Image Retrieval

Content-Based Image Retrieval (CBIR) is discussed here as one scheme for representing images used in image retrieval systems (Chang, Smith et al. 1997; Gupta and Jain 1997). The goal of CBIR is to describe a set of images based on their visual properties and to facilitate image retrieval by using the visual properties. CBIR is rooted in research from Computer Vision and Image Processing. In this approach, images are automatically analyzed for low-level features such as color, shape, texture and light intensity (McDonald and Tait 2003). The features extracted from images are used to construct a multidimensional feature space. All the images are then placed in this space. A well-known theory postulates that human similarity perception is based on the measurement of an appropriate distance in a metric psychological space (Shepard 1962; Torgeson 1965; Carroll and Arabie 1980). In this theory, it is assumed that a set of features models the stimulus' properties so that it can be represented as a point in a suitable feature space (Bimbo 1999). Similarity between images in CBIR systems is calculated by distance functions between the locations of images in the system's feature space. In order to retrieve images, the user can either Query by Example (QBE) by specifying an example image and asking the system to look for similar ones or Query by Sketching (QBS) where they can sketch out representations of images they are looking for. The query is located as a point in the feature space and image relevance is returned by the similarity calculation between the query point and all images in the set. For example, in Figure 14, the rectangles represent images of the database placed in a 3-axis color space. A query by the example image of the red apple is located in the space according to its color features (notice it is near the axis for the color red). The darkened images are those in the database that are close to the query image in the representational space, they become the results of the query and are returned by order of their proximity to the apple image. In this representational scheme, the vector distance between images in the space is used to calculate result relevance. This means queries have to be formulated as a point in this space.

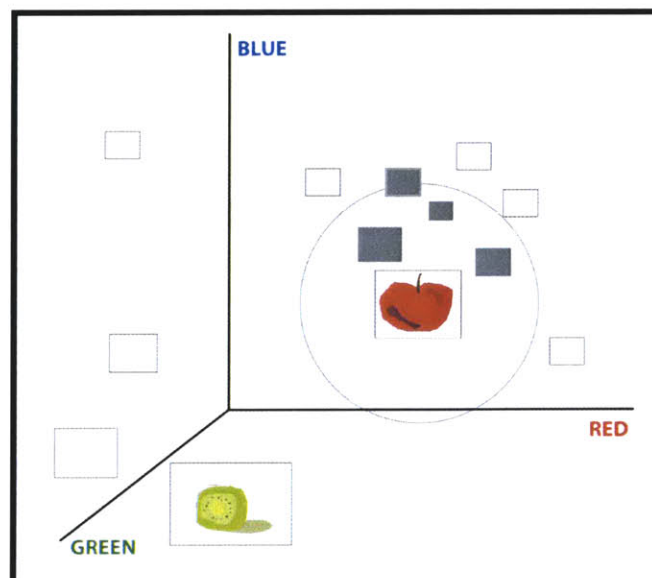


Figure 14 – Image query in a hypothetical feature space of RGB colors

CBIR systems follow theories of perception to create system representations that best simulate human cognitive representation (Salton and McGill 1983). As discussed in the section on theories of perception, visual cognition is concerned with the translation of point stimuli such as light intensity and color into higher-level representations of objects and shapes. Thus CBIR systems face the tough challenge of object recognition to simulate the functions of human cognitive processing. Whereas humans have little difficulty in picking out salient objects in images even in occlusion, this is a difficult task for machine recognition with success varying greatly among different task domains. In order to imbue the recognized regions features with meaning, CBIR systems often have a training mechanism whereby a human user labeled a sample region (e.g. “sky”), and the system goes ahead and tries to automatically find all instances of “sky” in all image sets (Pentland, Picard et al. 1993). They use image statistics to try to recover the semantic context of a scene automatically (Torralba and Oliva 2002). Because this is not a perfect process, CBIR systems typically have *relevance feedback* so that users engaged retrieving images could tell the system which regions have been correctly or incorrectly recognized (Gong 1998). Some exemplary CBIR systems include the QBIC system (Flickner, Sawhney et al. 1995) for feature space construction, the Photobook system for region modeling and relevance feedback (Pentland, Picard et al. 1993) and others that have examined different feature spaces such as color histograms (Vasile and Bender 2001), texture and others (Jaimes and Chang 2000).

3.3.1.1 Advantages

The advantage of the CBIR approach is that the extraction of descriptive data can be automated or partially automated. For large image databases, this means the size of the collection could be scaled up easily. Another advantage of the CBIR approach is the adaptive and learning nature of the system. As more users use the system and provide *relevance feedback* and train more models (e.g. “grass”, “pavement” etc.) the more intelligent and accurate the overall system becomes (at least in theory).

3.3.1.2 Disadvantages

Although CBIR systems tap in to theories of perception to construct feature spaces that reflect the human visual processing system, they have an extremely long way to go towards simulating actual human perception. In a way, the human brain still computes at an astonishing rate as compared to computational systems; it is somehow able to go through Marr’s steps for constructing a three-dimensional representational scene from perceived light intensities, in *real time*. Even if computational power were aligned with human processing, the process of visual perception is still not completely known. The primary disadvantage of the CBIR approach, as indicated by many studies in psychology, is the inadequacy of the feature vector model and metric distances as a model of human perception (Bimbo 1999). For example, a color feature space may judge two images to be similar because they contain similar portions of a particular red. However, a human being might see drastically different images of a beautiful sunset and a can of ketchup. This problem is due largely due to the subjectivity of human image perception and is commonly referred to as the Semantic Gap. Meaning that is gathered from an image is more than can be aggregated from low-level features such as color, shape and texture. A converse retrieval problem is such that except in very specific domains with expert users,

it is difficult for users to specify what they are looking for in terms of low-level features which is what needs to happen for a query to be formulated (Bimbo 1999). This implies serious limitations to Query By Sketching and Query by Example (Doorn and Vries 2000). The Semantic Gap problem in CBIR has been tackled from many perspectives including the use of combinations of multiple similarity models and iterative user relevance feedback (Schettini 1994). These attempts have had varied success but face the underlying challenge of adapting users to the way a machine represents images with low-level features. The task of trying to deconstruct an image into components of a representational scheme based on low-level features can be at odds with the subjective nature of how images are perceived and described by people.


3.3.2 Keyword-Based Retrieval

Keyword-Based Retrieval is the second general approach to image representation that uses words to describe images. Keyword-Based Retrieval is based on giving captions to images and retrieving images by querying with text and finding matches between query words and caption words. Although most systems work by having a list of keywords associated with each image, this text data can also be structured. For example, the online image database PictureQuest (PictureQuest 2004) has a collection of about 500,000 images that are annotated with a “description” field (for example, “a teen with a dog using a laptop computer outdoors”) as well as a “suggests” field (for example, “day, pet, lifestyle, technology, youth”) (Figure 15).

PictureQuest
Info

1-800-764-7427

This watermarked comping image may be used for preview purposes only. To save this image to your hard drive, click and hold (Mac), or right-click (PC), then select "Save image". The non-watermarked version of this image is available through PictureQuest.com.



© S. Wanke/PhotoLink/ Photodisc/ PictureQuest

Image #: 125609

Description:	Boy jumping a fence.
Suggest:	family and lifestyles, v15, color, horizontal, exterior, center, lifestyle, jumping, boy, fence, energy, person, child, silhouette, blue, black, pink, 15290, LS005631
Credit:	© S. Wanke/PhotoLink/ Photodisc/ PictureQuest
Copyright:	© S. Wanke/PhotoLink, 2000
Collection:	Photodisc
Artist:	S. Wanke/PhotoLink

Figure 15 – Example of keywords of an image in the PictureQuest database

Jorgensen (Jorgensen 2003) has pursued extensive research into how people approach the task of describing visual images. In one experiment, people were asked to describe images for three different tasks (Viewing, Search and Memory). For the Viewing Task, participants were asked to produce a simple spontaneous description, writing down words or phrases that “pop into their heads” until they could think of no more, or the time for

the task ended. The descriptive Search Task context was such that participants were told to imagine that these were images they wished to find within some type of image storage and retrieval system. The participants were told to visualize their ideal “intermediary,” to whom they would be describing the image. In the descriptive Memory Task, participants from the Viewing task were later asked to write descriptions of the images from memory. The attributes described by participants were grouped into broad categories (or “classes”). These attribute classes are shown in Table 1 below. Two broad types of Classes emerged, those of “Perceptual” attributes, related to the physical content of the image, and “Interpretive” attributes, which are stimulated perceptually but require additional internal interpretive and intellectual processes in order to name the attribute. Perceptual Classes include OBJECTS, PEOPLE, COLOR, SPATIAL LOCATION and VISUAL ELEMENTS. Interpretive Classes include STORY, ART HISTORICAL INFORMATION, PEOPLE RELATIONS, EXTERNAL RELATION, ABSTRACT and DESCRIPTION. VIEWER RESPONSE included personal reactions to the images.

CLASS	Viewing	Search	Memory	Average
OBJECTS	34.3%	27.4%	26.2%	29.3%
PEOPLE	8.7%	10.3%	11.1%	10.0%
COLOR	9.2%	9.7%	9.0%	9.3%
STORY	7.4%	10.8%	9.4%	9.2%
SPATIAL LOCATION	8.3%	10.7%	7.7%	8.9%
DESCRIPTION	6.0%	9.0%	8.8%	8.0%
VISUAL ELEMENTS	7.2%	5.4%	9.2%	7.2%
ART HISTORICAL INFO.	3.8%	5.7%	7.6%	5.7%
PEOPLE RELATED	5.2%	3.9%	2.6%	3.9%
EXTERNAL RELATION	3.3%	3.8%	4.0%	3.7%
VIEWER RESPONSE	3.7%	1.9%	3.1%	2.9%
ABSTRACT	3.0%	1.5%	1.3%	2.0%

Table 1 – Distribution of Classes by average percentage across three describing tasks

Jorgensen noted that the given task context affected how people described images. However, the Perceptual Classes were consistently named across the tasks. Of these, OBJECTS were the most cited. From the Interpretive Classes, the occurrence of attributes in the STORY Class suggest that in addition to subject categorization, specific human activities are of interest as well as other “story” elements. For example, an image of a whale jumping out of the water was described as “whale *performing* by jumping out of the water”. Attributes which formed part of the story of an image were those which described the “Who, What, When, and Where” aspects of the image representation. Specific attributes included activities, events, settings, and time aspects of the story as well as affective aspects such as emotions and relationships among people:

*This man seems to be paying close attention to what he sees
The woman appears to be overly concerned with appearance – she appears to be
putting on lipstick and looking at some kind of colorful material
[she is] worn out from too many years of stress*

Attributes representing emotions, abstract ideas, themes, and atmosphere also appeared in the descriptions:

Apparent female control of beast
Picture gives a feeling of triumphance [sic]
The kid is having a great time
Angry looking – ready to fight
Sinister, violence, threatening
Seems to have a religious connotation

All together, attributes representing story elements, emotions and relationships, themes, and abstract ideas such as the “atmosphere” of an image accounted for about 15% of the attributes named in the image describing tasks.

Jorgensen then presented users with a Sorting Task in order to elicit additional attributes that may shed light on how people perceive similarity among images. As described above, similarity is a very important notion in image retrieval as the similarity between a query and images in the database can be used to generate a ranking of results for a query. The participants were asked to sort seventy-seven color and black-and-white images into groups, such that if they wished to find an image at a later time, this would be the group they would look. The results of the Sorting Task are presented with the results of the Describing Task for comparison in Table 2 below.

CLASS	Describing tasks Average	Sorting task
OBJECTS	29.3%	8.7%
PEOPLE	10.0%	8.6%
COLOR	9.3%	2.6%
STORY	9.2%	8.5%
SPATIAL LOCATION	8.9%	0.8%
DESCRIPTION	8.0%	2.7%
VISUAL ELEMENTS	7.2%	3.8%
ART HISTORICAL INFO.	5.7%	23.8%
PEOPLE RELATED	3.9%	4.0%
EXTERNAL RELATION	3.7%	8.7%
VIEWER RESPONSE	2.0%	13.8%
ABSTRACT	2.9%	14.1%

Table 2 – Class frequency by percentage for describing tasks and Sorting task

Jorgensen notes several interesting aspects to this set of data when compared to data from describing tasks. The first is the much lower occurrence of the OBJECTS Class when compared to the describing tasks, in which it was the most frequently occurring class. The second is the higher occurrence of ABSTRACT attributes, which describe such things as abstract concepts, themes, and symbolic interpretations. The third is the Sorting Task also showed a strong effect for the presence of humans or human activity. The fourth is the low frequency of the perceptual attributes Classes of COLOR, SPATIAL LOCATION, and VISUAL ELEMENTS. Therefore, the Sorting Task participants appear

to rely more heavily on what are termed “interpretive” attributes rather than perceptual attributes, at least in terms of what they choose to communicate about these images in their process of sorting and placing the images in groups. This suggests that thematic and abstract ideas are more important in browsing than in search.

The two broad types of attributes of the Perceptual and Interpretive lend support the PictureQuest system’s annotation scheme, which has an entry for descriptive keywords (perceptual) and one for “suggests” keywords (more interpretive). Jorgensen proposes a structure called the Pyramid that provides even more levels of separation for attributes ranging from the specific to the abstract (Figure 16).

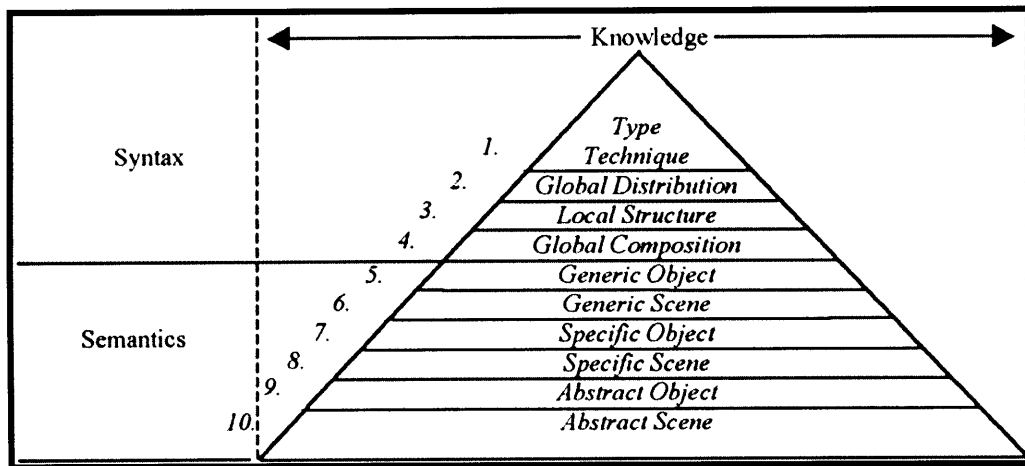


Figure 16 – Framework for classifying visual descriptors

In considering future directions of Keyword-Based Retrieval, Jorgensen points out that although studies have demonstrated that many descriptions of image content include a “story” connected with the image, the more affective and abstract aspects are generally considered too subjective to be addressed within formal systems. There is increasing recognition that these attributes should also be addressed, especially given that the more interpretive attributes appear to be very important in browsing (Jorgensen 2003).

3.3.2.1 Advantages

The primary advantage of the Keyword-Based Retrieval approach over CBIR is the lack of the Semantic Gap problem. As mentioned above, CBIR systems represent images in feature spaces constructed with low-level features whose similarity calculations may not resonate with human perception. A Query by Example with an image of a red apple may return a red sunset and a red table as well as other red apples simply because of their proximity in the low-level feature representation even though the images may not seem similar as judged by human perception that takes into account a user’s subjective reaction to an image. The keyword approach does not suffer from this problem because words *are* how people describe images to each other and therefore have hope of more fully capturing what people want to express about images.

As noted above, when asked to describe images, people cited the Perceptual Classes (attributes of visual elements from the images) consistently. However, they also cited the Interpretive Classes, especially when asked to group pictures by similarity. While some Perceptual Attributes can be detected by the CBIR approach because they represent what is visually *in* the picture, the Interpretive Attributes (such as theme, emotion etc.) are lost to the Semantic Gap. Words, on the other hand, allow users to describe both Perceptual and Interpretive Attributes.

3.3.2.2 Disadvantages

There are two major disadvantages of the Keyword-Based Retrieval approach as compared to CBIR. Firstly, there is the requirement of manual annotation. Secondly, many different words can be used to describe the same concept; this can lead to mismatches between words used by the annotator in labeling and image and later by a user when trying to retrieve an image.

The most obvious disadvantage of the Keyword-Based Retrieval approach is the requirement of manual labeling of images either with keywords or complete sentences. We may be able to provide a very extensive structure such as the Pyramid (Figure 16) for having categories of keywords at various levels of abstraction; however, we still need people to fill in those fields for each and every image for Keyword-Based Retrieval to be able to retrieve them. One possible approach to lessen the load on the user is to open up the indexing task to a wider audience such as an online community (Jorgensen 2003). This is the approach taken here and will be further discussed later.

A second disadvantage to Keyword-Based Retrieval is the nature of words. As highlighted by in “The Vocabulary Problem” (Furna, Landauer et al. 1987), in the English language the same word could have different meanings while different words could be used to describe the same thing, and people rarely agree on which word to use. Not only that, the subjective nature of perception is such that depending on our emotional moods and the context, the *same* person could describe an image differently at different times (Evans 2001). Herein lies the hit-or-miss problem of keyword retrieval. For any concept a user may want to describe, there is no single term that perfectly captures the concept. Yet most keyword-based retrieval systems function by finding exact matches between query terms and caption keywords. In these systems, an image can only be retrieved if a search term perfectly matches one of the keywords associated with the image. For a query to be successful, a user has to guess the word that was chosen to describe a concept when it was annotated. The hit-or-miss problem makes image retrieval by exact keyword matching like a shot in the dark; the onus is on the user to guess between many words that may all describe the same concept. This is likely to be more of a problem for Interpretive Attributes than Perceptual ones as people may be more likely to agree on the physical objects of a scene than the themes it evokes. However the problem also exists with Perceptual Attributes. For example, a picture of an apple could be described as a “fruit”, “apple” or a “Golden Delicious”. The problem here is not that words are ambiguous but the opposite. There are simply too many ways of describing what is found in an image. One solution is to get the user to provide keywords at different

levels of classification as guided by a structure such as the Pyramid (Figure 16). This way, the words “fruit”, “apple” and “Golden Delicious” can all be entered at different levels of abstraction. However, this puts a large amount of onus on the user to exhaustively annotate all images. In this thesis we explore using linguistic packages to automatically expand keywords semantically. With semantic query expansion, the user only has to enter the keyword “apple” and the image retrieval system would understand that an “apple” is a “fruit” and that a “Golden Delicious” is a type of “apple”. It would also understand that “apples” are often found at “picnics” and have “seeds”. A query for “apple” would then retrieve images that were annotated with the keyword “apple” as well as images annotated with keywords that are conceptually related to what an “apple” is. Query expansion helps alleviate the hit-or-miss problem by getting the machine (rather than the user) to examine the many different ways a concept could be expressed.

3.3.3 Integrated Approaches to Image Retrieval

Content-Based Image Retrieval (CBIR) and Keyword-Based Image Retrieval have been described above as employing different representational schemes for images and having different query formulation. Although we have described the two approaches to image retrieval separately they are not dichotomous and can co-exist in the same system. For example, CBIR techniques could be used to recognize and group similar textures. These textures could then be given a name (such as “grass”) and the system could then automatically look for other images with that texture and automatically classify them (with “grass”). This way the system would take advantage of image processing to automate part of the annotate process. A pioneering system for this type of integration was Photobook (Pentland, Picard et al. 1993). Recent systems have further integrated low-level features and semantic keywords with relevance feedback that affects both representations (Lu, Zhang et al. 2003). Future consumer image retrieval systems will likely use some combination of feature space representation and semantic keyword representation, the balance of which would be grounded in theories of perception (Vasile and Bender 2001).

3.4 Focus of Thesis

This thesis explores three weaknesses of keyword-based image retrieval. They are:

1. The requirement of heavy manual annotation of keywords.
2. The hit-or-miss nature of exact keyword matching.
3. The lack of support for retrieval by subjective content.

We explore methods of overcoming these shortcomings through the design and implementation of an actual image retrieval system. The methods we investigate to address the issues above are:

1. We explore the use of community-based annotation to alleviate the onus of manual annotation on individual users. Every user can take advantage of *all* annotations done by any member of the community. For example, a user can

- query for “apple” and retrieve images that were annotated with the word “apple” even if they didn’t create that image-keyword association.
2. We explore using linguistic tools (WordNet and the OpenMind Commonsense database) to perform automatic query expansion to overcome the hit-or-miss problem of keyword-based image retrieval. For example, queries for “youth” could also retrieve images annotated with “teenager” as both terms are conceptually related.
 3. We explore creating a mechanism for users to annotate images by their subjective emotional content and to subsequently query for images by emotional content. Since the emotional response of users to images are likely to be drastically different due to the subjective nature of emotions, we explore methods of aggregating user response to provide more accurate retrieval.

The emphasis of the thesis on creating intuitive and accessible methods of keyword-based image retrieval and how they can be practically realized in a larger multimedia framework. We discuss how traditional image retrieval systems are formally evaluated but do not apply a rigorous evaluation to our system. Instead we focus on creating different scenarios for keyword-based image retrieval and discuss new ways of facilitating the user’s retrieval task. The goal of the exploration is to come up with ways to alleviate the onus on the user and whenever possible to transfer the labor to the retrieval system.

The image retrieval system we create is integrated with the Treehouse Global Studio online community. Treehouse Studio is an online community that has a set of media design tools. It also has a database backend for storing the created media files. The image retrieval system augments the Treehouse Studio backend to store image metadata that facilitate the retrieval of visual media files such as pictures and drawings. We have also built a set of graphic user interfaces that allow users to annotate images with keywords and perform queries and view retrieval results.

In this chapter we have laid out the theoretical framework and background of our research. In the next Chapter we will present the design and implementation of our image retrieval system and situate it in the Treehouse Studio environment. In Chapter Five we discuss the strengths and weaknesses of our image retrieval system and ways of extending it to better facilitate the user experience. We present a summary of our exploration in Chapter Six.

4 System Design and Implementation

4.1 The Treehouse Studio Framework

The framework for our image retrieval system is the Treehouse Global Studio online community. This section provides a brief overview of the Treehouse Studio environment. The rest of the Chapter describes the design and implementation of the image retrieval system and how it integrates with the Treehouse Studio infrastructure.

Treehouse Studio (Maeda 2004) is a long-term research initiative of the Physical Language Workshop at the MIT Media Laboratory. It is a vision for a global online community for the creation, sharing and commerce of digital arts. Treehouse Studio aims to bring new digital arts creation tools to a broad Internet population by packaging them as Java programs that run inside of a web browser. These tools empower a broad audience to participate in new forms of digital media creation. Treehouse serves as a framework for conducting research into digital media forms and collaborative processes.

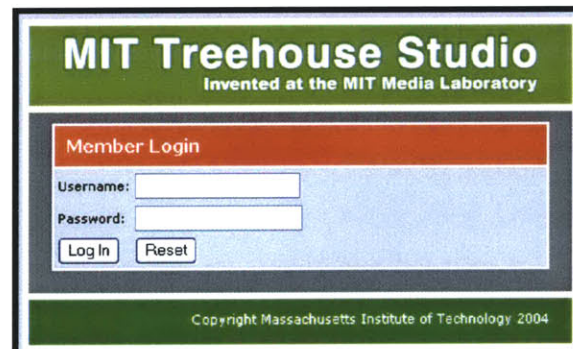


Figure 17 – Treehouse login page

Once a user logs into the system through the login page (Figure 17) they see their own home page (Figure 18). Please note that Treehouse Studio is still under development and all screenshots are from April 2004.

From their home page, users have access to a set of tools with which they can create different types of digital media (from the “Toolbox” panel of Figure 18). The toolset includes a vector-based drawing tool (Figure 19), a pixel-based photo-editing tool (Figure 20), a sound editing tool, a video-editing tool, a 3D sculpting tool and a slideshow presentation tool as well as others. Although many of these types of media creation tools already exist as commercial applications, the goal of redeveloping them as part of Treehouse is to make them accessible both in terms of user interface design and deployment. Since all the tools are built with Java, they only require an updated web browser to run and can be accessed anywhere in the world.

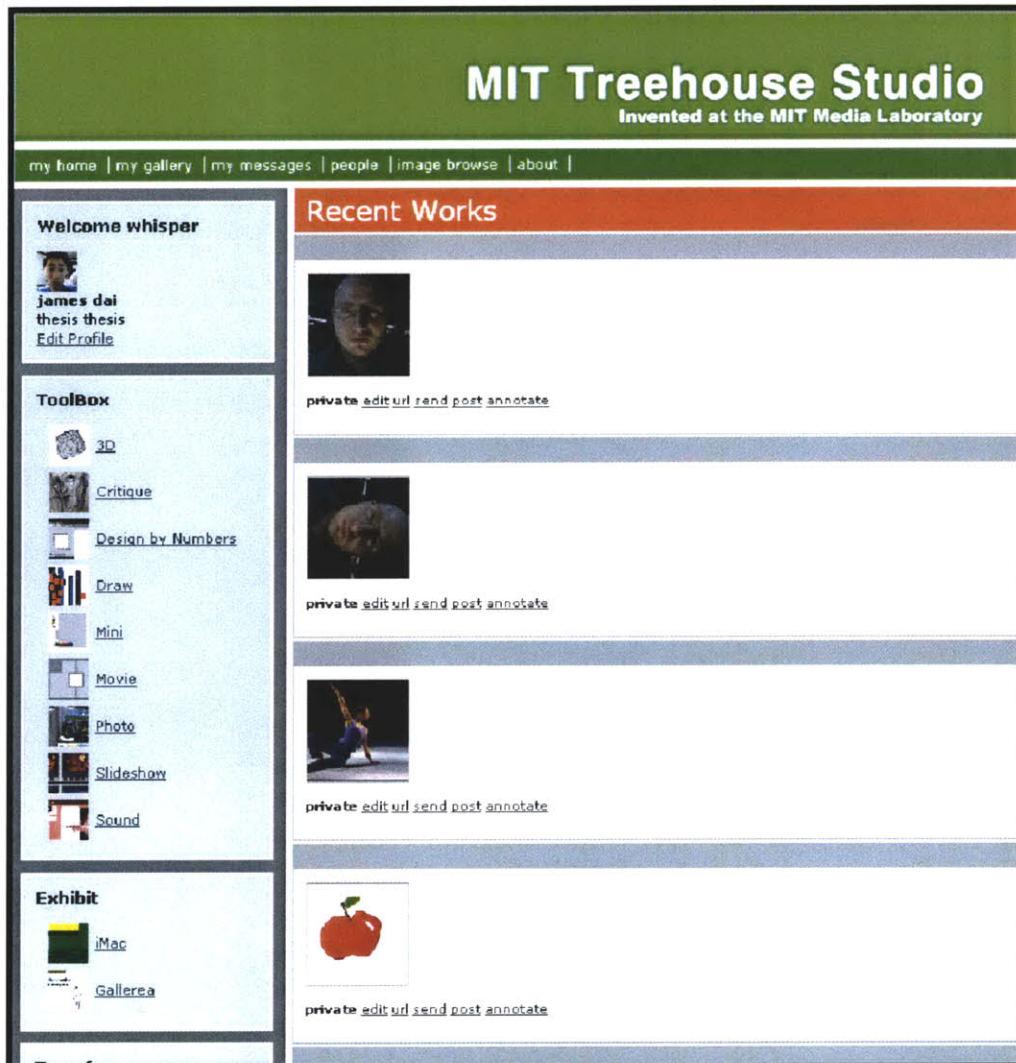


Figure 18 – Treehouse user home page

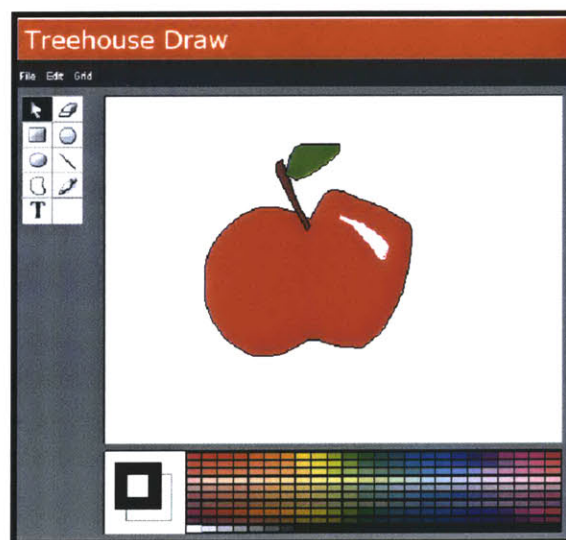


Figure 19 – The vector-based Draw application

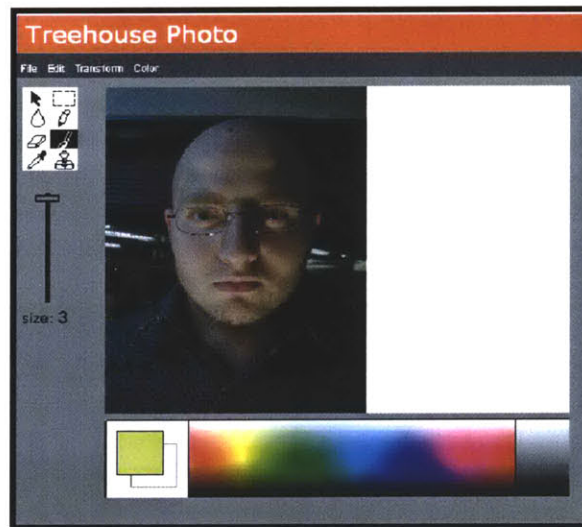


Figure 20 – The pixel-based Photo application

When a user creates a new work with one of these tools, the result is stored on the Treehouse server. A link to the new work then appears in the user's Recent Works section on their home page (Figure 18). Treehouse also allows users to import digital images to their account either by direct upload from their computer or via e-mail attachments sent to the Treehouse e-mail server. These images then become part of the user's collection and can be edited with the Draw and Photo applications.

Thus the Treehouse Studio community has both a web-based graphic user interface to a set of media design tools as well as a backend database server to store media files created and edited with the Treehouse tools. In terms of image retrieval, the drawings and photos created with the Draw and Photo applications are the most relevant since they are visual images.

We now move on to describe the image retrieval system we have created within the Treehouse Studio framework to allow Treehouse users to annotate, browse and retrieve images stored in the community. Everything described below has been designed and implemented as part of this thesis work.

4.2 The Image Retrieval System

4.2.1 Image Annotation Application

We have created a tool with which images can be annotated with keywords. Each piece of work in a user's collection can be annotated with the Annotation Application (Figure 21). This tool is accessed through the *annotate* link beside each media item in the Recent Works section of a user's Treehouse home page (Figure 18). When loaded, the image appears on the left side of the Application (Figure 21).

The Annotation Application allows users to enter new keywords for an image by typing into the text box at the top right and pushing the enter key. For example, in Figure 21, the keywords “dancer”, “hand” and “arm” have been added to the image of the dancer. These keywords can also be spatially pinpointed in the picture. In Figure 21, the keyword “hand” has been pinpointed to the dancer’s raised hand. The circular pinpoint region can also be resized; this pinpoint information is saved but not currently used for image retrieval. An extension of the system may use them for image processing for object recognition or spatial relationship analysis between objects.

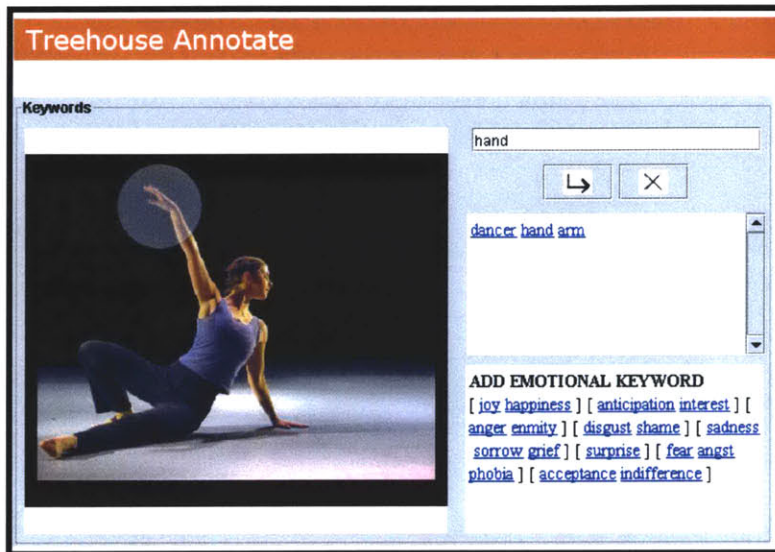


Figure 21 – The Annotation Application

When a new keyword is entered into the textbox on the top right of the Annotation Application, it is compared against a list of common words (called a stop list). The stop list contains words such as pronouns (“he”, “she”, “it”) and other common words that do not help distinguish images (such as “when” and “if”). If the word entered is not a stop word, then it is accepted and saved in the Treehouse server database in association with the current image. Currently, the system only accepts single keywords as unit of annotation and does not accept sentences.

What is surprising about using the Annotation Application is the speed with which a user can add keywords and pinpoint them. The keywords that have been associated with an image are immediately and automatically saved to the database and appear as hyperlinks. When a keyword is selected with a click, it can be deleted by clicking on the image button with an X on it. The keyword can also be pinpointed by moving the mouse over the image and clicking. Once set, dragging with the left mouse button will relocate the pinpoint and dragging with the right mouse button will resize it. The location and resize are all limit-bound to relevant values and the highlight circle is translucent against the image for clarity.

The bottom right section of the Annotation Application has an area for adding emotional keywords. This is discussed in detail later.

4.2.2 Image Browse and Retrieval Application

There is an “image browse” link from the user home page (Figure 18) that launches the application for browsing and retrieving images (called the Image Retrieval Application for short). A screenshot of the Image Retrieval Application is shown in Figure 22.

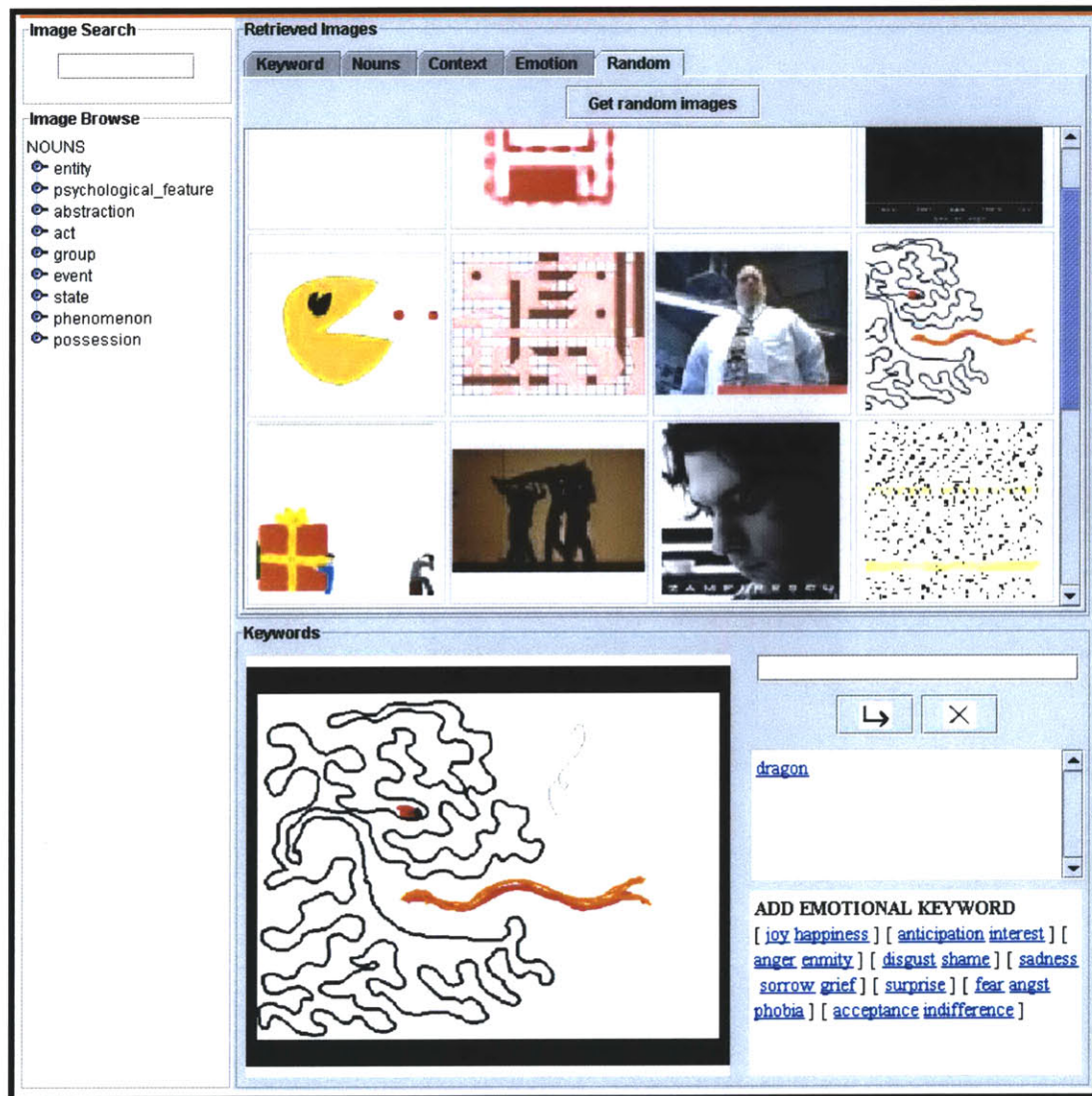


Figure 22 – The Image Retrieval Application

The Image Retrieval Application operates on *all* images that exist in Treehouse Studio database. Its interface contains numerous parts that work together to allow the user to search and retrieve images from the Treehouse database by a number of different methods. The top left “Image Search” section has a textbox for initiating new queries. The bottom left “Image Browse” section will be discussed later. The top right *Retrieved Images* region of the interface has five tabs: *Keyword*, *Nouns*, *Context*, *Emotion* and *Random*. Figure 22 shows the *Random* tab. This tab allows the user to click on the “Get random images” button in order to retrieve 20 random images from the database. These images are displayed in the scrollable region below the tab. This is the region for

displaying retrieved images. When a retrieved image is selected by a mouse click, it shows up enlarged in the bottom right and can be annotated. The bottom right *Keywords* section of the interface is exactly the same as the Annotation Application described above (Figure 21), except that it is embedded into the Image Retrieval Application.

A user initiates new queries by typing the query term into the *Image Search* section's textbox. Currently only single word query terms are examined. When a new query is launched, the *Retrieved Images* section is updated to show the results. Below, we will go through the search processes in detail.

4.2.2.1 Keyword Tab – Direct Retrieval

When a new query is launched from the *Image Search* textbox, The *Keyword* tab of the *Retrieved Images* section returns images that were annotated with that exact search word. For example, Figure 23 shows the *Keyword* results of a query for “apple”. All three of the returned images have the exact word “apple” as a keyword. As described earlier, this form of direct retrieval has the hit-or-miss weakness of exact keyword matching.

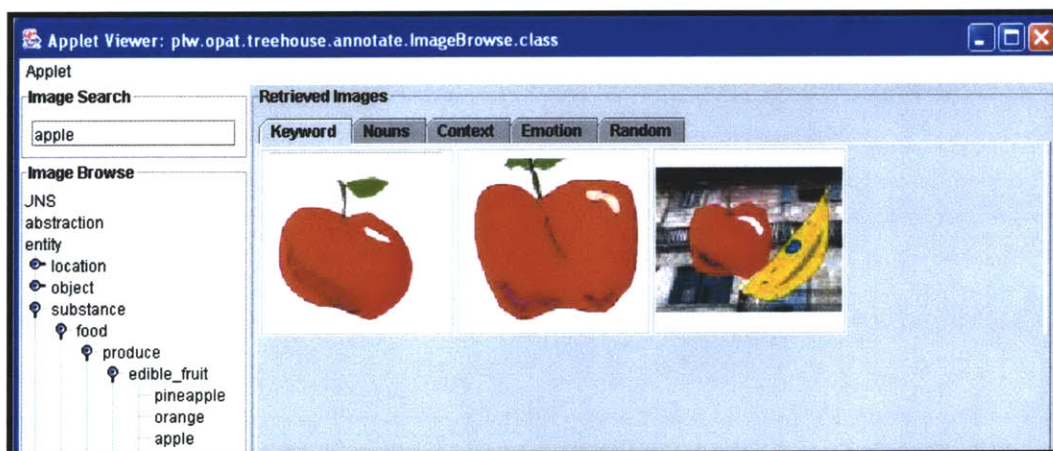


Figure 23 – Keyword tab – direct retrieval

When new keywords are added to images via the Annotation Application, they are saved to the Treehouse backend database. The application does not save which user has created the annotation. When a user initiates a new query, for example, for “apple” as above, the Image Retrieval Application looks up all images in the database that were annotated with the word “apple”, regardless of which user created that annotation. This fulfills the community-based annotation scheme that is one of the foci of the thesis. Individual users do not have to annotate the entire Treehouse image database; they tap into the collective aggregate annotations of the community. A search query initiated by an individual user makes use of all annotations created by any member of the community.

The *Keywords* tab of the Image Retrieval Application shows results by performing exact keyword matching between the query term and keywords of all images in the database. With exact keyword matching, images and their associated keywords exist in isolation. A query has to be compared with each image-keyword association separately and in a

binary fashion. For example, Figure 24 shows a query for “peach”. It just so happens that none of the images in the database have a keyword of “peach”, so no results are returned.

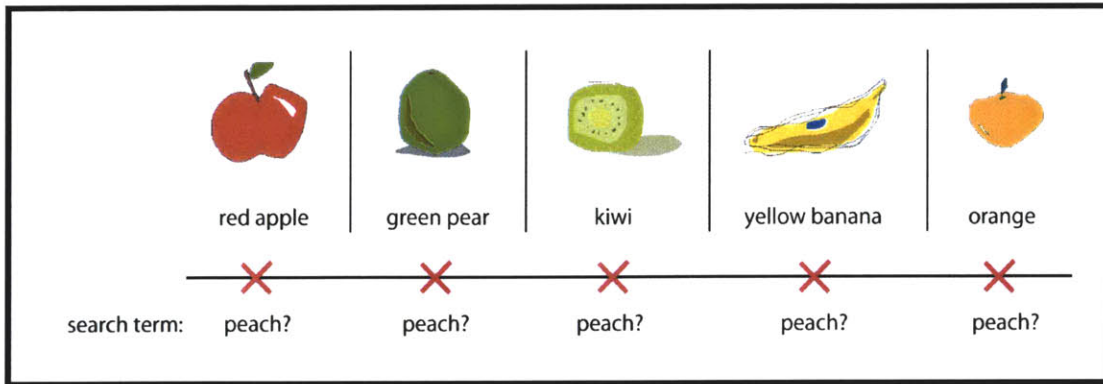


Figure 24 – Exact keyword matching

With exact keyword matching, the image either contains the query term exactly as entered or it doesn't, the query is either a hit with the image or a miss. The nature of the English language is such that there is many different words can describe the same concept. Thus a user may have to try a few different queries before finding the proper query word that exactly matches the keyword of images they want to retrieve.

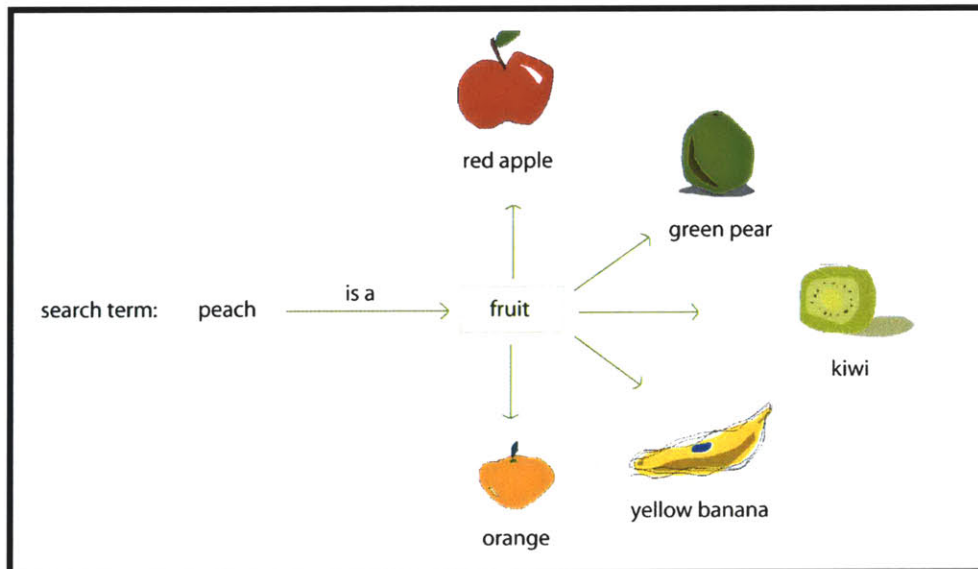


Figure 25 – Image retrieval by query expansion

What we would like to do is get the retrieval system to do more work on the query and to understand the meaning of the query word and to return images in the database that have keywords that are conceptually related to the query term. For example, in the example of the “peach” query above, the system finds that there are no images directly annotated with “peach”, but it should understand something about what a “peach” is, namely that it is a “fruit” and that an “orange” is also a “fruit”. The retrieval system should present these other “fruit” images to the user as potentially relevant results in the absence of

“peaches” (Figure 25). A query in this scheme of things is only a starting point that spreads out and returns a collection of images tagged with conceptually related keywords. This way, the system presents the user with images that do exist in the database so the user can examine them and decide how to refine their query to find images they are interested in.

Query expansion is the second focus of this thesis. Expanding a search query has been previously explored with text-based document retrieval by the *Information Retrieval* community (Akrivas, Wallace et al. 2002) (Gonzalo, Verdejo et al. 1998) (Deerwester, Dumais et al. 1990) (Takagi and Tajima 2001) (Lee 2001). Here, we explore the potential benefits of query expansion for the task of image retrieval. If exact keyword matching is like the user casting a fishing pole into exact spots in a pool of images with no guarantee of hooking onto an image, then query expansion gives users a way to dive into the image database and swim around groups of relevant images until they find some that are useful.

While the *Keywords* tab of the Image Retrieval Application shows results of exact keyword matching between query words and keywords of images in the database, the *Nouns* and *Context* tabs perform query expansion on the query term. These processes are described below.

4.2.2.2 Nouns Tab – Query Expansion by WordNet

WordNet is a linguistic dataset that organizes English nouns, verbs, adjectives and adverbs into synonym sets, each representing one underlying lexical concept (Fellbaum 1998). It can be thought of as a dictionary that can be computationally accessed. Each word in the system can have multiple definitions (what WordNet calls “senses”, as shown in Figure 26 below).

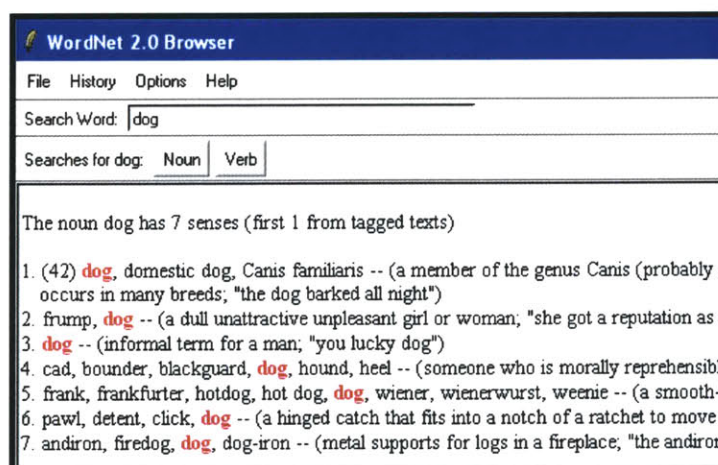


Figure 26 – WordNet senses for “dog”

Hierarchies of hyponyms (“is a kind of” relations) provide distinction between different definitions of the same word. For example, Figure 27 shows the hyponym tree for the first two definitions of “dog”.

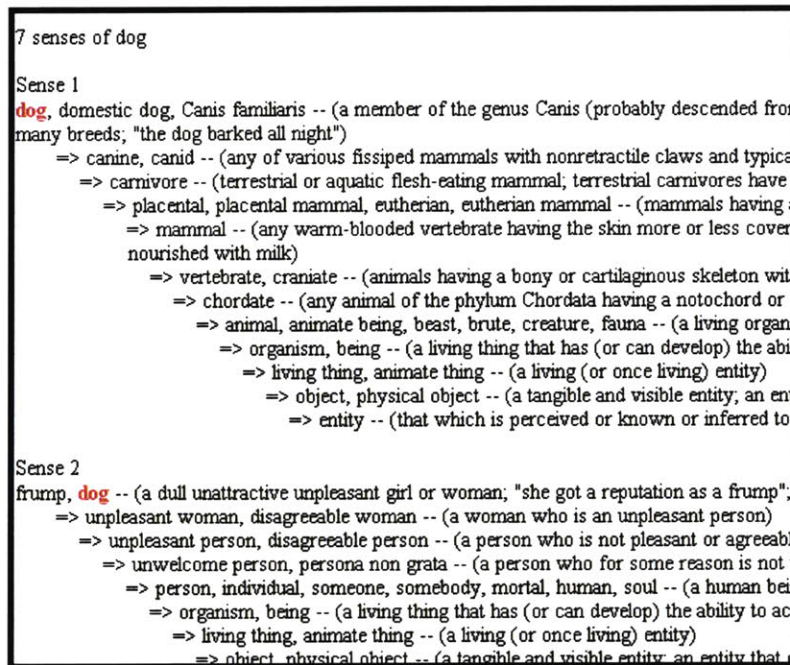


Figure 27 – Hyponyms of two definitions of “dog”

Thus WordNet is a series of hierarchies of concepts with each path from a root node to a leaf node providing a distinct definition for that leaf node. For example, the concept “orange” includes these two definitions in the WordNet hierarchies:

- Definition of “orange” as a fruit: orange > fruit_tree > tree > vascular_plant > organism > object > entity.
- Definition of “orange” as a color: orange > chromatic_color > color > visual_property > property > attribute > abstraction.

WordNet could be used to programmatically expand a search query in the following manner. Suppose a part of WordNet looks like the following:

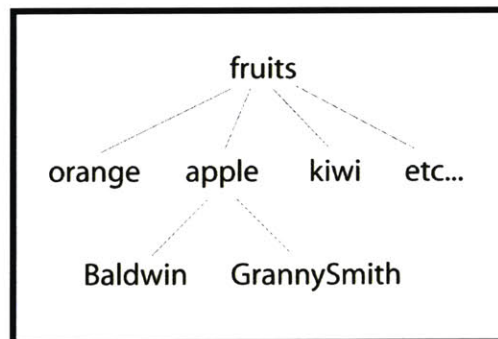


Figure 28 – A sample concept tree

Suppose a user initiates a query for “apple”, the image retrieval system would first locate the node of the query term on this tree. It would then expand the query semantically by

searching for the parent, children and sibling nodes of the concept in the Treehouse image database. Thus a search for apple could return images that were not only annotated with the word “apple” but also with words such as “fruits”, “orange”, “kiwi” or “GrannySmith”. In this way, WordNet could be used to expand a query based on its dictionary meaning to return hypernym concepts (apple is a kind of fruit) and hyponym concepts (GrannySmith is a kind of apple) as well as sibling concepts (other kinds of fruits).

Above, we mentioned that aggregating community annotations would lessen the onus of manual annotation on the individual user. Expanding a query by making use of the knowledge of WordNet would also lessen the manual annotation requirements of users. They no longer have to exhaustively annotate at multiple levels of abstraction (such as “fruit”, “apple” and “GrannySmith”). An image annotated with “apple” will also be reachable with queries of “fruit” and “GrannySmith” when the system performs query expansion with WordNet.

A major obstacle to taking advantage of the WordNet hierarchies in query expansion is the huge size of WordNet and the numerous obscure concepts that exist in its definition hierarchies. For example, the hyponym structure of one definition of “dog” is dog > canine > carnivore > placental > mammal > vertebrate > chordate > animal > organism > living_thing > object > entity. While the concept of “chordate” maybe useful for grouping WordNet concepts, it is a very uncommon word and users are unlikely to use it to annotate an image. So “chordate” and many other nodes in the WordNet hierarchies may have few or no images associated with them. These “empty” nodes will lessen the effectiveness of WordNet-based query expansion.

This problem can be overcome by extracting from WordNet a sub-tree consisting of node terms that actually have images associated with them in the image database. One way to do this is to “grow” a new tree by adding all the annotation keywords from Treehouse Studio, one keyword at a time. A parent node with no images associated with it would only be added when it is needed to organize two or more leaf nodes. This type of structure has been called the Dynamic Semantic Hierarchy because it is constantly updated to reflect a highly populated tree of relevant nodes (Yang, Wenyin et al. 2001).

We have implemented a Dynamic Semantic Hierarchy (DSH) in our image retrieval system to facilitate WordNet-based query expansion. We will now describe the algorithm used to construct our DSH and how it expands on the original DSH algorithm proposed by Yang et al. (Yang, Wenyin et al. 2001).

When our image retrieval system is loaded, it begins with an empty tree as the Dynamic Semantic Hierarchy (DSH). Each keyword (w) in the database is added to the DSH one by one following these steps:

1. Find the definitions of w in WordNet. These are nodes $[N_1, N_2, \dots, N_K]$ in the WordNet hierarchies, where K is the number of definitions.
2. For each node N_j , perform the following:

3. Starting from N_j , trace bottom-up along the links in WordNet hierarchy, until the first ancestor node already exists in the Dynamic Semantic Hierarchy (DSH). This ancestor node is denoted N_a .
4. Let $[C_1, C_2, \dots, C_M]$ be the direct children of N_a in the DSH, where M is the number of children.

For $i=1$ to M

Find the lowest common ancestor node N_{co_a} of both C_i and N_j in the WordNet hierarchy.

If $N_{co_a} \neq N_a$, goto step 6.

5. Insert N_j into DSH as a direct child of N_a .
Exit the algorithm.
6. Insert N_{co_a} into DSH as a child of N_a .
Remove C_i as a child of N_a and insert it as a child of N_{co_a} .
If $N_j \neq N_{co_a}$, insert N_j into DSH as a child of N_{co_a} .
Exit the algorithm.

For example, Figure 29 illustrates the insertion operations in our DSH. Section (b) shows the DSH after one keyword (“parrot”) has been inserted. Section (c) shows the subsequent insertion of the keyword “horse” and section (d) shows what the DSH looks like after “hen” has been added. Part (a) of the figure demonstrates how the DSH overcomes the distribution weakness of WordNet hierarchies – it shows that all leaf nodes of the DSH are terms that actually have images associated with them (“horse”, “elephant”, “parrot” and “hen”). Terms such as “chordate” and “vertebrate”, which exist in WordNet but don’t have images associated, are left out of the DSH. The algorithm only adds parent nodes that have no images associated with them (such as “mammal”) when they are needed to help organize the DSH.

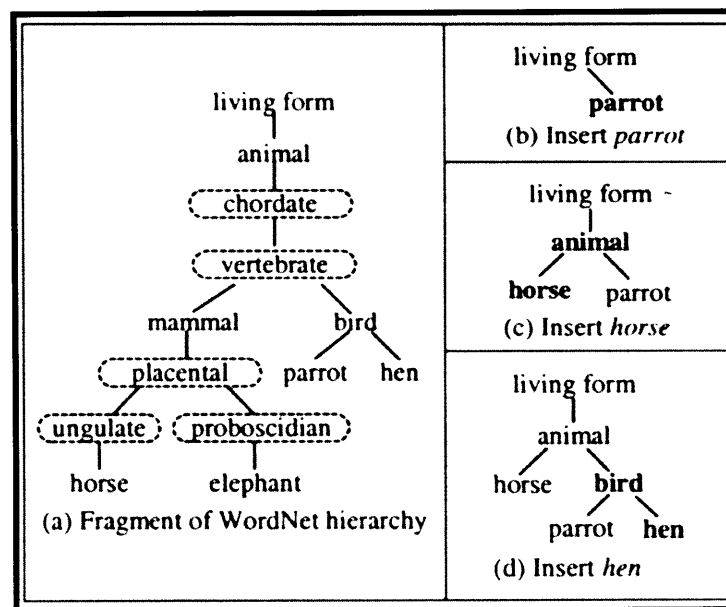


Figure 29 – Insertion of new terms into the DSH

It is important to note that Figure 29 doesn't completely capture the insertion algorithm. For each term (such as "parrot"), WordNet has multiple definitions. Figure 29 only shows the insertion of one possible definition. Yang et al (Yang, Wenyin et al. 2001) have demonstrated the construction of DSH for these single definition insertions. Our algorithm above extends their algorithm to take into account multiple possible definitions. So in actuality multiple copies of "parrot" with different definitions would be inserted at different parts of the tree (this is not illustrated but is what indeed happens with our system).

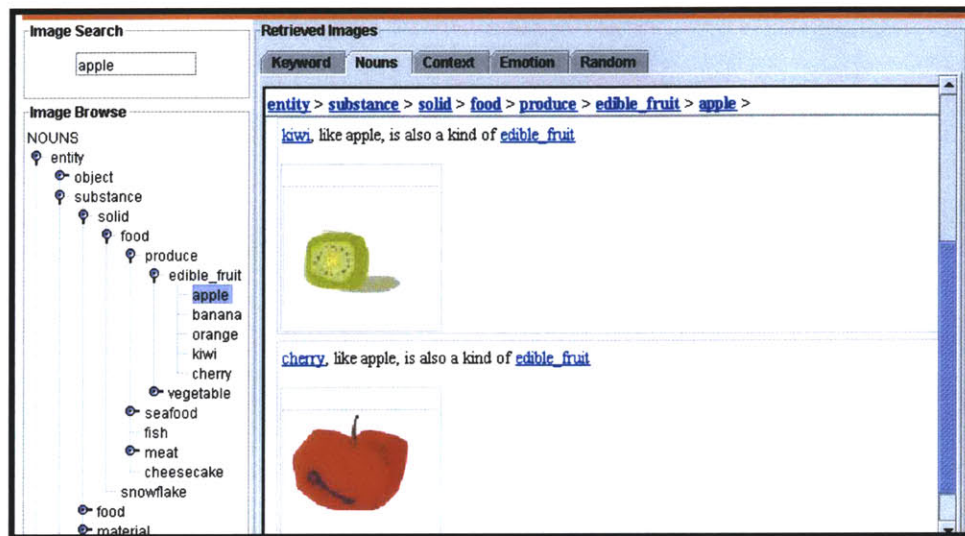


Figure 30 – The Nouns tab – query expansion by WordNet

Figure 30 shows the *Nouns* tab of the Image Retrieval Application. This tab shows the query expansion that has been performed on the search term "apple" with WordNet. The left side of the figure shows the Dynamic Semantic Hierarchy that has been constructed by inserting *all* the keywords in the database. The *Nouns* tab shows the result of the WordNet query expansion for the query term "apple". White title bars show the definitions of "apple", the one visible in Figure 30 is "apple" as an "edible_fruit". On the left side we can locate "apple" on the DSH. It shows that other "edible_fruit"s include "banana", "orange", "kiwi" and "cherry". The *Nouns* tab shows that indeed, the search has returned some of these items from the database ("kiwi" and "cherry"). The query for the concept "apple" has been expanded to return the parent, siblings and children of the "apple" node in the DSH if they have images associated with them. Although the scroll pane of the *Nouns* tab only shows the single "edible fruit" definition of "apple", other definitions are also similarly retrieved. For example, the query also returned "apple" as a "fruit tree" (Figure 31) and the query returned other types of fruit trees ("cherry" and "pear").

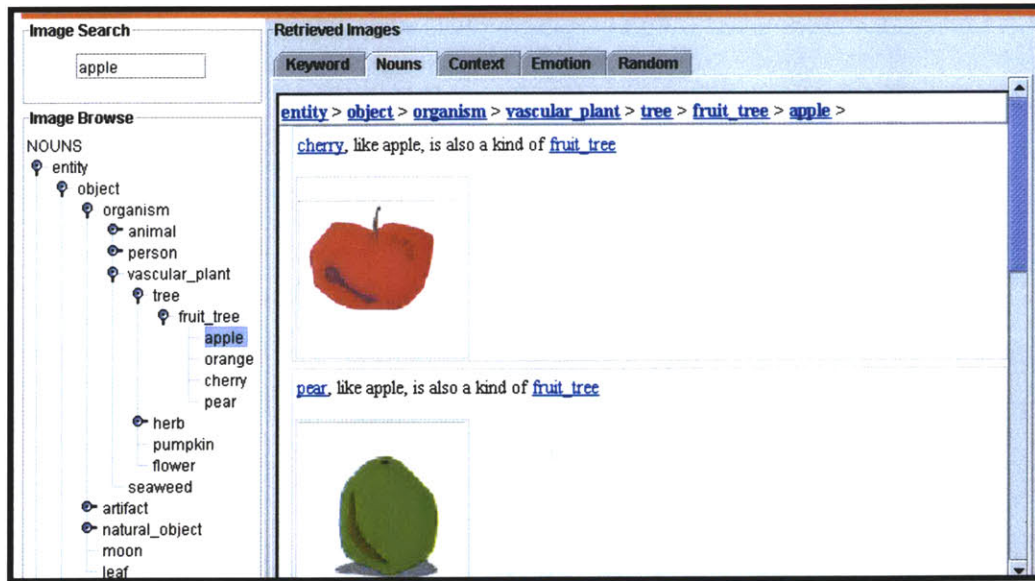


Figure 31 – A second definition of “apple”

Figure 32 below shows another example of WordNet-based query expansion. Here a search for “reptile” has returned two types of “reptiles”, “turtle” and “dragon” (which are children of the “reptile” node). The search has also returned a sibling of “reptile”, a “frog”.

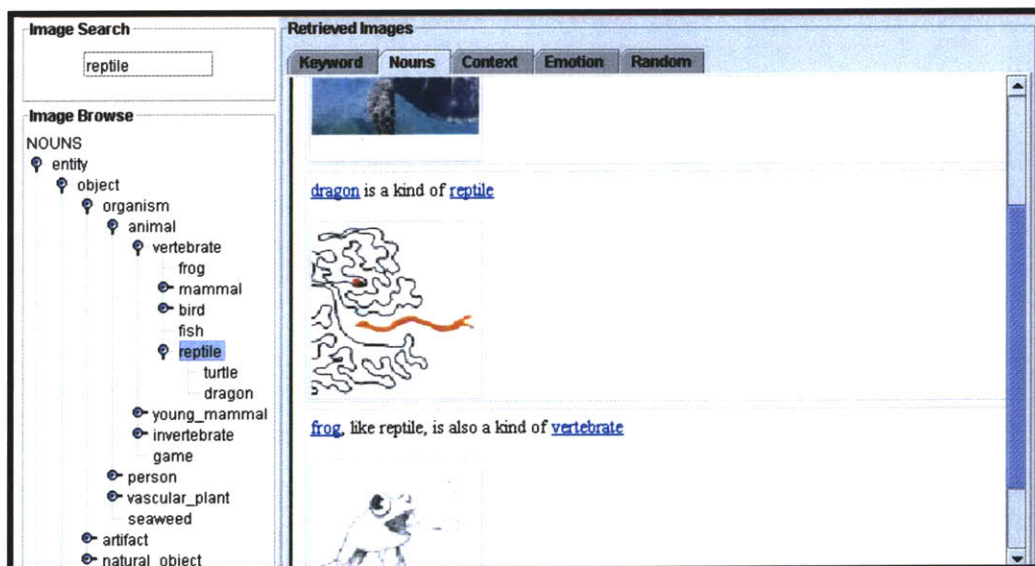


Figure 32 – Another example of WordNet query expansion

Each time the Annotation Tool is used to add a new keyword to an image, the Dynamic Semantic Hierarchy is updated to reflect the change. The algorithm described above runs and new leaf nodes as well as common parent nodes are automatically inserted in the proper place. The user can also explore the DSH tree by using their mouse; a click on any node launches a new query for that concept. This is shown on the left side of Figure 32.

4.2.2.3 Context Tab – Query Expansion by conceptual connections

Using WordNet as a tool for query expansion works for expanding search terms by their dictionary-based meanings. However, this type of query expansion only tackles part of the problem. In describing the hit-or-miss problem above, we mentioned that for any given concept there may be many words that describe it. Using WordNet allows us to expand a concept to find general or specific types of the concept. But it does not use other conceptual connections to the query term. For example, WordNet helps the system know that an orange is related to an apple because both are fruits, however, it would not help the system know that apples often have stems, are most likely red or green or are likely to be found in fruit baskets. For these other conceptual connections, we make use the data from the OpenMind Commonsense (OMCS) database (Liu and Singh 2003). OpenMind, like WordNet, is a semantic network of concepts and their relationships to each other. Whereas WordNet primarily has *is-a* type relationships (the hyponyms and hypernyms discussed above), the OMCS relations are much more varied. The goal of OpenMind is to capture “commonsense” relationships between concepts so computers can be more intelligent in understanding what a human takes for granted. For example, knowledge such as “cats have fur” or “the sky is normally blue” are common knowledge to a human being but seem so simple that they are left out of formal linguistic systems such as WordNet. OpenMind contains millions of conceptual relationships that capture such everyday knowledge such as “soap is used to wash hair”, “weddings have brides and grooms” and “apples are often red”. Other semantic networks like OpenMind exist, for example, the Cyc project also capture commonsense data (Lenat, Guha et al. 1990). However, OpenMind was chosen because it is open source and freely available. The knowledge in the OpenMind semantic network was collected piece meal online from users around the world. Treehouse Studio espouses a similar model of collective knowledge aggregation. Since both projects also originated from the MIT Media Lab, there is also future potential for crosspollination and data sharing.

The Image Retrieval Application overcomes the shortcomings of WordNet query expansion by using the OpenMind Commonsense data to get the most common contexts for the query term. By “context” here we mean other types of conceptual relationships beyond the “is-a” relationships of WordNet. For example, the contexts for “groom” may include “wedding” and “marriage” (“weddings” normally have a “groom” and “marriage” is between a “groom” and “bride”). Our image retrieval system uses OpenMind to find concepts that are highly related to a query term, then it looks in the Treehouse database to see whether there are any images that are annotated with these context words. If so, they are returned on the *Context* tab.

For example, Figure 33 shows a query for “family”. The OMCS data returned ten conceptual contexts for “family”, the top three, which are shown in Figure 33, are “man”, “woman” and “parent”. The images from the Treehouse database that are tagged with these words are returned.

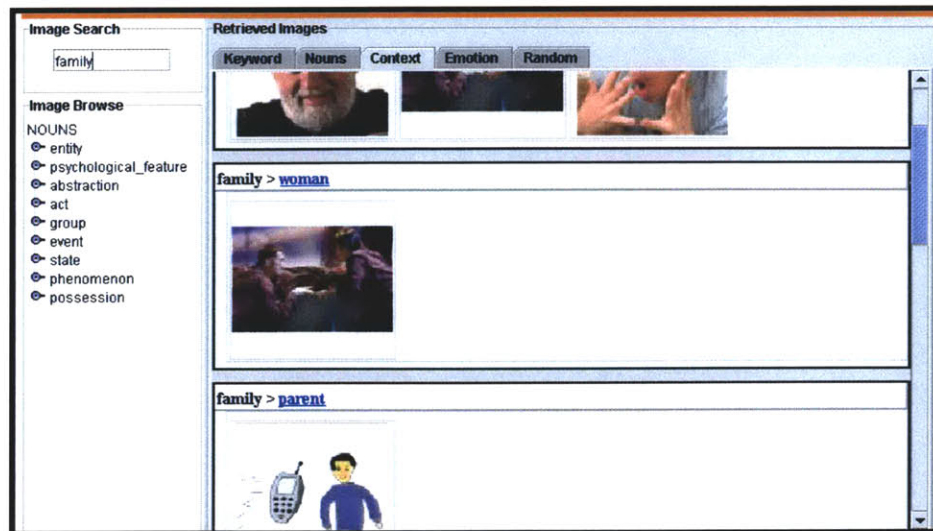


Figure 33 – Using OpenMind to expand queries by conceptual relationships

Together, WordNet and the OpenMind Commonsense database expand a single query term to other terms that are conceptually related. A query is no longer a shot in the dark that demands an exact match between query keywords and image caption keywords. By making use of the knowledge from WordNet and OpenMind, the keywords of images in the Treehouse database are connected with each other in a semantic network. The retrieval system locates a query at a certain point in this network. The retrieval system, instead of the user, analyzes the query term for meaning and presents what it thinks are other relevant images depending on semantic connections in the network of keywords. Query expansion by WordNet and OpenMind alleviate the hit-or-miss weakness of exact keyword matching retrieval. It is good to note that the *Keywords* tab also returns exact keyword matched results. The query expansion results of the *Nouns* and *Context* tabs can be used when exact keyword matching fails.

4.2.2.4 Emotion Tab – Search by emotional response

Above we have described how our image retrieval system uses WordNet and OpenMind data to explore the second focus of this thesis, that of query expansion. In this section we will examine the third and final focus of the thesis, facilitating the annotation and retrieval of images by their subjective emotional content.

As described in Chapter Three, humans often cite the emotions when asked to describe images, especially for the task of sorting images into categories to facilitate retrieval. Most past image retrieval research on keyword-based retrieval have focused on what is observable *in* the images, on the Perceptual Attributes rather than the Interpretive Attributes such as emotions (Jorgensen 2003). For example, Yang et al. (Yang, Wenying et al. 2001) focus solely on nouns visible in images which are more or less agreeable between different people. This avoids the thorny issues of quality control and consistency that come with the subjective nature of the Interpretive Attributes. However, as stated earlier, the visual experience is an inherently subjective one. So the retrieval of images based on their subjective content needs to be explored. To address this gap in research,

we have implemented a mechanism for users to provide feedback about their emotional response to an image. As part of our investigation we aggregate user feedback in order to bypass the consistency issue of emotional response.

For our image retrieval system we limit the number of emotional keywords used in annotation and retrieval. The emotional keywords available are from an affective theory (Scherer and Ekman 1984) that groups emotions into eight groups in a wheel formation as shown in Figure 34. Each emotion is similar in nature to its neighbors and roughly opposite in nature to those across from the wheel. This strict structure is supported by evidence of the basic emotions common to all humans, despite differences in societal upbringing (Stein and Oatley 1992). A color palette was chosen for the emotion wheel to help illustrate it.

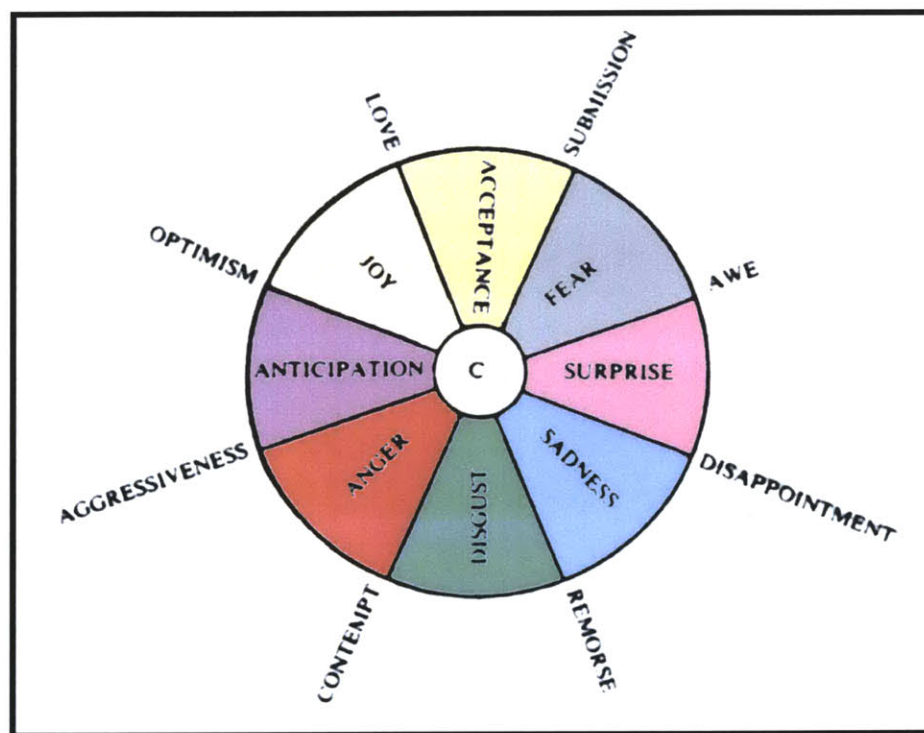


Figure 34 – A theory of basic emotional groups

In order to address the issue that people may use different words to describe similar emotions, synonyms are made available for each of the eight groups. The synonyms for the eight groups can be found in Table 3 below. These are chosen rather arbitrarily as the idea is to convey a sense of a category of emotion rather than trying exhaustively to capture all words that are related to a category.

Emotion	Synonyms
Joy	happiness
Anticipation	interest
Anger	enmity
Disgust	shame
Sadness	sorrow, grief
Surprise	
Fear	angst, phobia
Acceptance	indifference

Table 3 – Emotional keyword synonyms

As shown in Figure 35 below, the Annotation Application contains a section in the bottom-right for adding any of these emotional keywords to an image by clicking on it with the mouse.

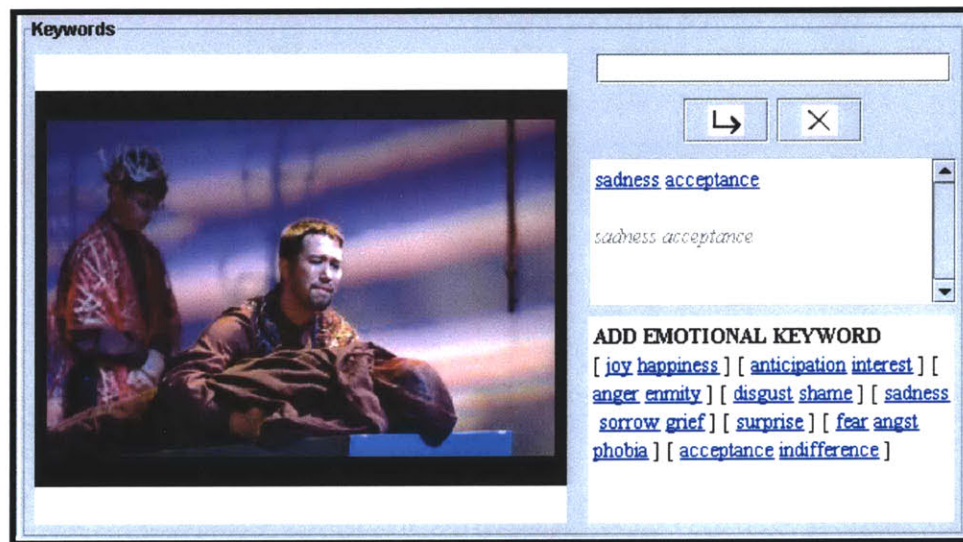


Figure 35 – Emotional response annotation with the Annotation Application

Once images are tagged with emotional response keywords, the Image Retrieval Application allows retrieval by these emotions through the *Emotion* tab. A search is initiated when the user clicks on one of the emotion keywords at the top of the *Emotion* tab. For example, Figure 36 shows the result of a query initiated by a click on “joy”. The wheel affective theory is used to expand an emotional query by retrieving similar emotions as well as the opposite emotion. For example, a query for the “joy” emotional category will return images tagged directly with that emotional group (Figure 36) along with images from the similar groups of “anticipation” and “acceptance” (Figure 37) as well as images from the opposite group of “sadness” (Figure 38).

It is not clear whether or not this query expansion for emotional keywords would be useful for users. It was added to the system as a possible analogy to the query expansion for regular keywords.

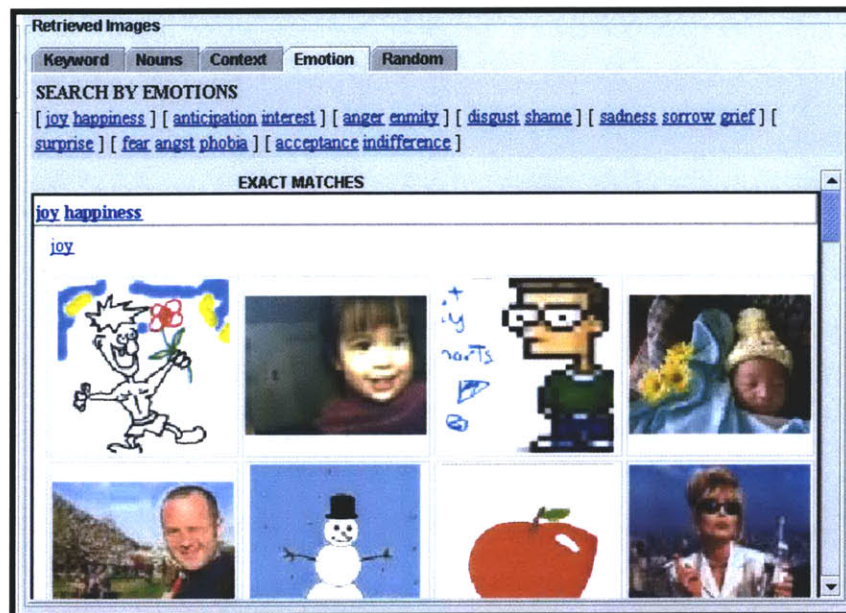


Figure 36 – Retrieval by emotional response

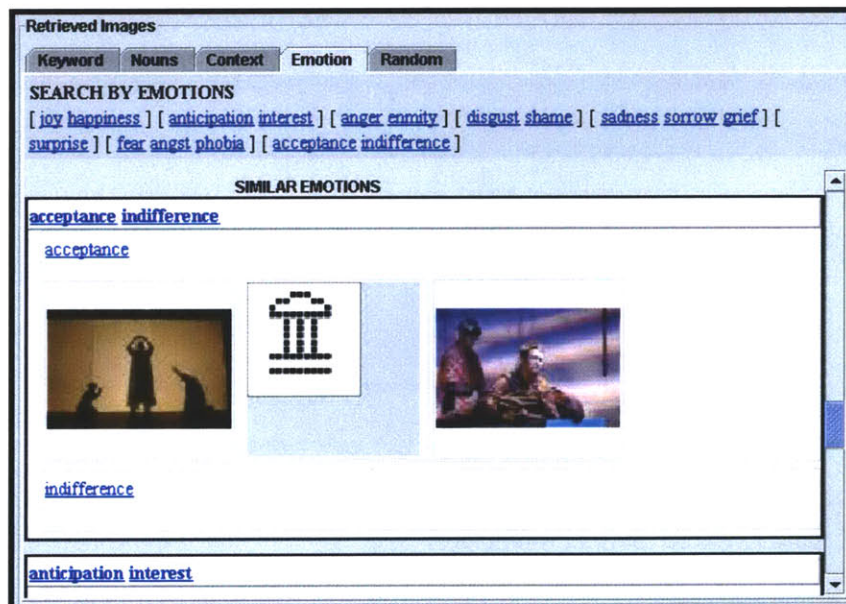


Figure 37 – Retrieval of similar emotions



Figure 38 – Retrieval of opposite emotion

As mentioned above, the subjective nature of emotions poses serious challenges for the retrieval process. Emotional response to images could differ greatly by user and even by the same user experiencing the image at different times. The Image Retrieval Application deals with this challenge by keeping track of the total number of users who label an image with a particular emotional word and uses this count to rank results. A search for images that have the emotion “surprise” returns images tagged with that keyword ranked by the number of users who annotated the image with “surprise” as shown in Figure 39.

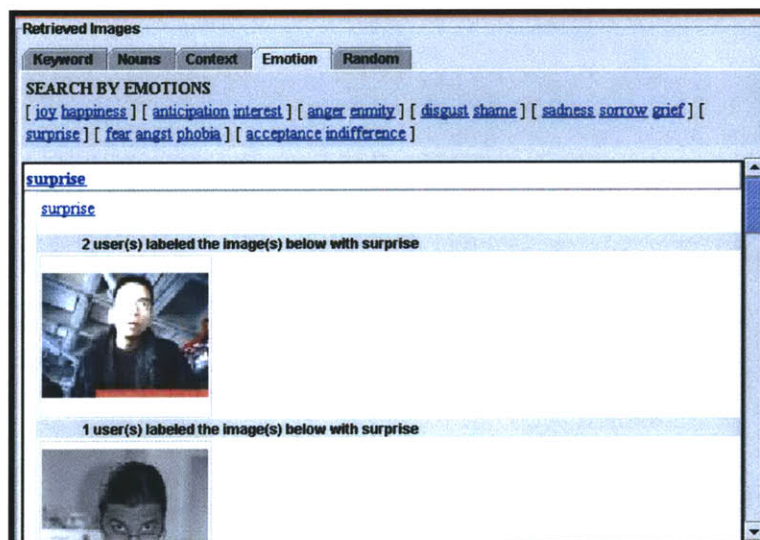


Figure 39 – Aggregate user feedback for emotional retrieval

This replaces a binary keyword-matching representation (that asks “is this image sad?”) with a probabilistic one (“how many people think this image is sad?”) An image could be

labeled as “sad” by a certain number of users while simultaneously being labeled as “anger” by a different number of users. This algorithm simply uses the total aggregate number of users who consider an image to be “happy” in order to rank results of a query for “happy” images. It does not consider that a number of users may also have considered this to be “sad”. A more advanced algorithm would take into account the distribution of the aggregate user response for calculating the retrieval ranking. Finding the optimal ranking algorithm is beyond the scope of this thesis. The focus here is on aggregating user annotations to deal with the subjectivity of emotional annotation.

4.3 User Interaction in the Image Retrieval Application

When a user enters a search term into the *Image Search* area at the top left region of the Image Retrieval Application, a query is initiated. This search modifies the *Keyword*, *Nouns*, and *Context* tabs simultaneously. As described in detail above, the *Keyword* tab returns results that directly contain the query term, the *Nouns* tab returns results after having expanded the query with WordNet and the *Context* tab returns results after having expanded the query with the OpenMind Commonsense data. The Dynamic Semantic Hierarchy tree structure at the left of the Image Retrieval Application reflects the subset of WordNet data that are in use by the images in the database. This tree structure is automatically updated when new keywords are added. The user can also explore the tree by expanding its branches. They can click on any node to launch a new query for that node. The *Emotions* tab is operated differently in that queries by the emotional groups are launched directly from this tab when a user click on one of the emotion words in the tab (Figure 36 above).

The images returned can be clicked on to open them in the Annotation Application at the bottom right of the Image Retrieval Application. This enables on-the-fly annotation during the image retrieval process.

The query expansion fundamentally changes how a user goes about the task of retrieving and browsing images. Previously with exact keyword matching, the user had to spend effort to come up with the perfect query term. This is an error-prone task that will likely require the user to reiterate multiple times without promise of results. Query expansion loosens the constraints and unloads much of the search effort onto the computer. The query term the user supplies provides a semantic starting point for the search. Because all the images in the database are connected to each other through the semantic web of knowledge that threads through their keywords, a query term can navigate along the threads of this web to seek out images that are labeled with terms that are conceptually related to the query term. When these results are returned, the related words are shown along with any associated images. Figure 40 shows the expanded terms in the *Nouns* tab and Figure 41 shows the expanded terms in the *Context* tab.

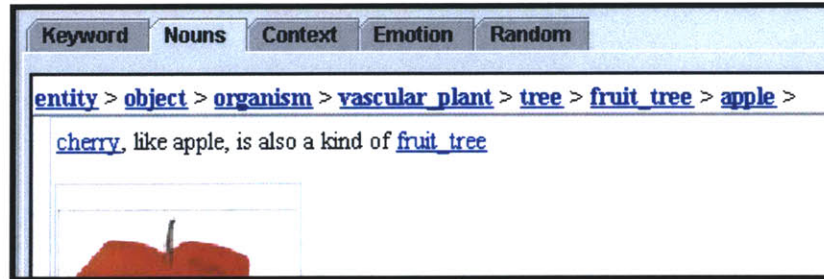


Figure 40 – Nouns tab links

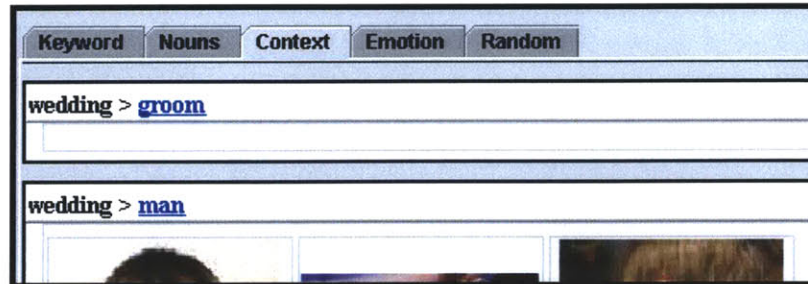


Figure 41 – Context tab links

These words are in fact hyperlinks that can be clicked on to launch a new search on that term. In this way, the user can continue to navigate the image database through the semantic network that connects all of the images' keywords by meaning. The initial query is the only time the user has to enter a search term; the computer then presents the user with possible paths to refine the search. For example, Figure 40 shows the results of a query for “wedding”. In the figure we see that “groom” and “man” are both related to “wedding”. At this point the user can click on the word “groom” to launch a new query for “groom”. The user can continuously navigate through the image database in this manner, moving closer to the concepts and images they seek.

4.4 Image Retrieval System Design

4.4.1 Server Client Architecture

The Image Retrieval Application, like all other Treehouse applications, is built as a Java applet and runs inside a web browser. The image retrieval processes require the use of the large WordNet and OpenMind data libraries. Loading them in the client applet is impractical (WordNet is about 20 MB worth of text data and OpenMind is of similar size). In order to overcome this problem, the Image Retrieval Application uses a client-server model with an extremely lightweight client and a heavyweight server. The Image Retrieval Application only handles the visual representation of the current state of image retrieval. It communicates with Treehouse server processes running on server machines through Java Servlet technology. The server processes is where the actual heavy data manipulation takes place.

Upon server initialization, WordNet and the OpenMind Commonsense database are loaded into memory. The Dynamic Semantic Hierarchy is constructed and populated with

all the keywords that are in the Treehouse image database. This is a process that could take minutes to complete but is acceptable because the server is rarely restarted. Once the server process is initialized along with the rest of Treehouse Studio, it is ready to do the data manipulation required by the Image Retrieval Application running on the client machines of users. Once the Dynamic Semantic Hierarchy has been completely constructed, the server constantly updates it with new keywords that users add with the Annotation Application. Whenever an Image Retrieval Application is loaded at a client machine, the server sends it a copy of the Dynamic Semantic Hierarchy to display.

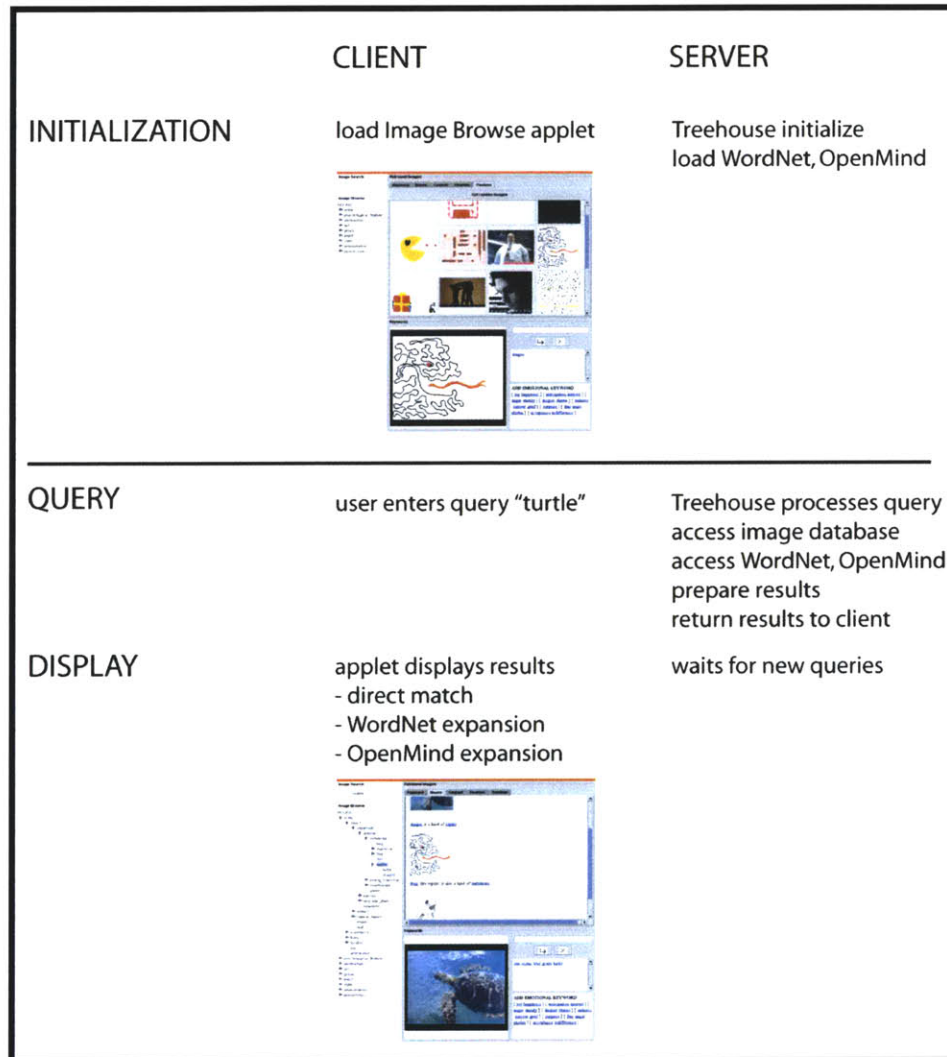


Figure 42 – Image Retrieval System Client-Server architecture

The server handles the search queries initiated by users. It analyzes the query term for WordNet query expansion and returns to the client applet all relevant WordNet definitions along with related terms and images as described above. The server also processes the query for OpenMind query expansion. It looks up the top ten contexts of the concept using the OpenMind data and again returns any images associated with the context words. The client applet receives the results from the server and displays them in

the Nouns and Context tabs. The client-server communications for our image retrieval system is illustrated in Figure 42.

The image retrieval system is integrated with the rest of the Treehouse Studio suite of tools. Treehouse uses unique integer numbers to represent all data in the database. So instead of passing bulky image data to the Image Retrieval applet, the server process simply passes the identification number. The Image Retrieval applet then uses the Treehouse Studio image loading mechanism to display the images.

4.4.2 Image Database Architecture

As mentioned above, Treehouse manages the database storage of image data and associates each image with a unique identification number. The Image Retrieval Application uses these numbers to manage the image-keyword association and the keyword-based retrieval of images. There are four database tables used facilitate the image retrieval process. The first (Table 4) is a dictionary. It contains all the keywords that exist in the system. Each entry represents a unique keyword. The data of the table include the unique id that is automatically assigned to each word, the word itself and the stem of the word. The stem is the word stripped to its lemmatized root form; it is useful so that queries for different versions of the same word can be matched with each other (for example, queries for “walked” can return images tagged with “walking” because they have the same stem “walk”).

id	word	stem
3	Apple	apple
4	Apples	apple

Table 4 – The dictionary table in the Treehouse image database

imageID	keywordID	pinX	pinY	pinRadius
23	3	34	50	10
23	4	0	20	20

Table 5 – Treehouse database table matching images to keywords

The second table maintains the associations between images and their keywords (Table 5). The imageID is the Treehouse identification number of the image and the keywordID is the identification number of the keyword from the dictionary table. The pinX, pinY and pinRadius entries capture the pinpoint information for this keyword.

When a query takes place, it is first located on the dictionary table and then its keywordID is used in conjunction with the second table to retrieve the ids of all images that contain that word. The image ids are then used to display the results in the Image Retrieval Application.

Two more tables are used in the image database to facilitate emotional retrieval because we need to keep track of the aggregate of individual user annotations in order to enable the ranking scheme used for emotional retrieval described above. The first is a table for

keeping track of the association between image and emotional word as well as which user created this association (Table 6).

username	imageID	emotionWord
Joe	1332	Happiness
Michele	1332	Joy
Mike	1332	Happiness

Table 6 – Treehouse database table for emotion keywords

imageID	emotionWord	count
1332	Happiness	2
1332	Joy	1

Table 7 – Treehouse database table for emotional keyword aggregate

The second table keeps track of the cumulative results of individual annotations (Table 7). When a user annotates an image with a particular emotional word, this association is stored in the database and the count for the total number of users who have made this association is updated. During retrieval, these tables are used to return images ranked by the cumulative aggregate of associations as described above.

4.5 User Scenarios Retraced

We will now describe how the system works in the background to retrieve images in the four scenarios described in Chapter Two.

In the first scenario, Jordan had previously added the keyword “fantastical” to an image (Figure 3). The keyword “fantastical” was added as a new keyword in the database in Table 4 and the keyword-image association was stored in the database in Table 5. When Jordan initiates a query for “fantastical”, the Image Retrieval Application sends the query to the server processes running on the Treehouse server machines. The server processes perform database table lookups on Table 5 to find the image that has been annotated with “fantastical”. The image Jordan had annotated is found and returned to the Image Retrieval Application and is shown to Jordan.

In the second scenario, Michael’s query for “apple” has been expanded by its dictionary definition. The query is passed from the Image Retrieval Application to the processes running on the Treehouse server just as in the first scenario. The server processes have in memory an updated version of the Dynamic Semantic Hierarchy (DSH) that contains all the keywords that have been entered by all users. The word “apple” is located at multiple places on the DSH. It is found under the parent node of “fruits” as well as “fruit tree”. The server uses the database tables as in the above scenario to find images that have been annotated with the parent, sibling and children of the “apple” nodes. For example, it looks up “fruits” (a parent of “apple” on the DSH) and “orange” (a sibling of “apple” under the common parent of “fruits”) as well as specific apples (children of “apple”) if they exist. All the expanded queries that contain results are packaged up and returned to the Image Retrieval Application and shown to the user (Figure 4). This way, Michael’s

query for “apple” returns direct matches for “apple” as well as other fruits. These other results are useful for him because he is looking for pictures of fruits to make a collage of images. It would have taken him much longer to query each type of fruit independently because he doesn’t know which fruits exist in the Treehouse image database.

In the third scenario, Joanna’s query of “beach” returned no direct results because no image in the Treehouse image database has been annotated with the word “beach”. However, the server processes have expanded the query to look up conceptually related words to “beach”, such as “ocean” and “sand”. These words do happen to have images associated with them and are returned to Joanna in the *Context* tab. If the query has not been expanded by the sever using commonsense conceptual connections between words, Joanna’s search would have returned no images at all.

In the fourth and final scenario, Peter uses the *Emotions* tab to search for images that elicit “happiness”. Previously, Treehouse users have annotated various images with different emotional keywords by using the Annotation Application. The Treehouse database tables keep track of these emotional keyword associations with the images. Table 6 keeps track of which user annotated which image with which emotional keyword and Table 7 keeps track of the total number of users who have annotated an image with a particular emotional keyword. When Peter initiates a query by the emotional keyword “happiness”, the server processes look up these tables to find images annotated with “happiness”. These images are returned to Peter in order of the total number of users who have described an image with “happiness”. This emotional retrieval mechanism allows Peter a quick way of getting at the emotional content of the images that exist in the Treehouse image database. The system automatically aggregates the community’s emotional annotations to deal with the fact that different users may perceive images differently.

4.6 Image Retrieval System Summary

In Chapter we have discussed how the design and implementation of our image retrieval system explores three weaknesses of keyword-based image retrieval. These three foci of the thesis are:

1. The requirement of heavy manual annotation of keywords.
2. The hit-or-miss nature of exact keyword matching.
3. The lack of support for retrieval by subjective content.

We have investigated these three foci in our image retrieval system by the following means:

1. We explored the use of community-based annotation to alleviate the onus of manual annotation on individual users. Every user can take advantage of *all* annotations done by any member of the community. Also, automatic query expansion further reduces the requirement of exhaustive manual annotation

since a query for a concept will be expanded to retrieve hypernyms and hyponyms of that concept as well as its siblings.

2. We explored using linguistic tools (WordNet and the OpenMind Commonsense database) to perform automatic query expansion to overcome the hit-or-miss problem of keyword-based image retrieval. WordNet is used to expand query terms via *is-a* type relationships where as OpenMind is used to expand query terms along other conceptual relationships.
3. We explored the creation of a mechanism for users to annotate images by their subjective emotional content and to subsequently query for images by emotional content. Emotional response is highly subjective and there are issues of consistency. We faced this challenge by aggregating user feedback for ranking retrieved images.

The next Chapter goes over strengths and weaknesses of our system in terms of how well they address the shortcomings of keyword-based image retrieval.

5 Evaluation

This Chapter presents a preliminary examination of our image retrieval system in terms of its goals. We discuss traditional metrics for quantitative evaluation of information retrieval systems but do not apply them to our system due to lack of time as well as a fundamental mismatch between query expansion and these evaluation parameters.

Instead we focus on a brief comparative analysis of our image retrieval system and existing image retrieval systems. We also gather a first round of feedback from Treehouse users in order to distill potential areas for improvement and future directions.


5.1 Functional comparisons with other image retrieval systems

There exist many online image databases that can be queried by a variety of methods. Those that implement a keyword-based image retrieval mechanism primarily use exact keyword matching. The PictureQuest (PictureQuest 2004) system described earlier (Figure 43) is one such system. PictureQuest uses exact keyword matching but presents an annotation structure to facilitate categorizing different types of keywords. There is a “Description” field for objects depicted in the scene and a “Suggest” field for more subjective and abstract ideas the image is related to. Having a “Suggest” field can prompt annotators of the system to capture some more subjective aspects of an image (Jorgensen 2003). However, because exact keyword matching is used, the annotators of the images in the system must aggressively try to exhaustively capture all ideas an image may be related to. Thus the keyword lists of PictureQuest images are very long and require a lot of manual annotation by a small army of annotators. This solution is not feasible for systems that manage personal images because exhaustive annotation is too burdensome for casual users.

The Treehouse image retrieval system does not present the user with an annotation structure to help them come up with different types of keywords. This is because the Treehouse system does not require the user to exhaustively annotate an image with all possible keywords. Instead the user is invited to enter a few keywords and to allow the system to perform conceptual query expansion on these keywords during the query process to retrieve semantically related images. This lessens the burden of annotation on the user. Whether or not this affects retrieval efficiency will depend highly on the type and size of the image database.

PictureQuest
1-800-764-7427
Info

This watermarked comping image may be used for preview purposes only. To save this image to your hard drive, click and hold (Mac), or right-click (PC), then select "Save image". The non-watermarked version of this image is available through PictureQuest.com.




© S. Wanke/PhotoLink/ Photodisc/ PictureQuest

Image #: 125609

Description:	Boy jumping a fence.
Suggest:	family and lifestyles, v15, color, horizontal, exterior, center, lifestyle, jumping, boy, fence, energy, person, child, silhouette, blue, black, pink, 15290, LS005631
Credit:	© S. Wanke/PhotoLink/ Photodisc/ PictureQuest
Copyright:	© S. Wanke/PhotoLink, 2000
Collection:	Photodisc
Artist:	S. Wanke/PhotoLink

Figure 43 – PictureQuest image database and keyword-based retrieval system

Another popular image search system is Google Image Search (Google 2004). The Google Image Search examines *all* images mined from the entire Internet archive that has been collected by the Google search engine. Again, exact keyword matching is used. For example, a query for “kiwi” will return images with the “kiwi” in its filename (Figure 44).


[Web](#) [Images](#) [Groups](#) [News](#) [Froogle](#) [New!](#) [more »](#)

[Advanced Image Search](#)
[Preferences](#)


Moderate SafeSearch is on

Images
Results 1 - 20 of about 87,900 for kiwi [definition]. (0.19 seconds)


Show: [All sizes](#) - [Large](#) - [Medium](#) - [Small](#)




kiwi.jpg
325 x 236 pixels - 8k
www.wegmans.com/.../produce/fruit/images/kiwi.jpg



kiwi.jpg
277 x 313 pixels - 15k
www.kidzone.ws/animals/birds4.htm



kiwi.gif
412 x 434 pixels - 94k
www.laturi.org/pic/kiwi.gif



kiwi.jpg
292 x 433 pixels - 28k
blizzardguy.tripod.com/bemie/

Figure 44 – Google Image Search for “kiwi”

Google Image Search has some intelligence about using the words in the immediate context of the image on a webpage to go slightly beyond exact keyword search. But no active query expansion is pursued. Query expansion for Google Image Search is infeasible because of the huge amount of images that are archived (in the billions). So exact keyword matching will always return many results. The problem lies in the usefulness of the retrieved images. Unlike the Treehouse Studio system where the user

can search their own private collection of images as well as that of the community, Google Image Search retrieves images that come from *anywhere* on the Internet. It is clearly serving a different purpose than the system described here. But it is interesting to note its effort in reaching into the textual context of an image to go beyond exact keyword retrieval. This has similarities to our use of linguistics packages to expand a query term in the retrieval process.

5.2 Traditional Information Retrieval Evaluation Methods

Traditional methods for evaluating information retrieval systems use quantitative measures of *precision* and *recall* (Jorgensen 2003). These quantifiers are calculated according to Table 8.

	Judgment by evaluator	
	Relevant	Not relevant
Retrieved	A (correctly retrieved)	B (falsely retrieved)
Not retrieved	C (missed)	D (correctly retrieved)

Table 8 – Information retrieval evaluation metrics

They two evaluation metrics are calculated as follows:

- **Precision** = $A / (A + C)$ = relevant correctly retrieved / all **relevant**
- **Recall** = $A / (A + B)$ = relevant correctly retrieved / all **retrieved**

These metrics for evaluating the effectiveness of information retrieval systems can be applied to many different types of data (such as text documents, websites, images, video or sound) as long as there is a gold standard by which a returned result can be judged to be relevant or irrelevant to a given query. This gold standard likely requires a human judge, especially for multimedia retrieval systems where relevance is more difficult to quantify and its detection is difficult to automate.

Although the system described in this thesis is one for visual information retrieval, query expansion cannot be easily assessed in strict *precision* and *recall* terms because it essentially retrieves many images that are not directly associated with the query but are rather are conceptually related to it. The goal of query expansion is to return many *potentially* relevant images in order to guide the user along semantic paths towards useful images. It becomes difficult to assess query-expanded results in terms of *precision* and *recall* because it is difficult to determine which images can be deemed as *relevant* for a query. Another reason for abandoning the traditional approach is that our exploration of retrieval by emotional response is inherently rooted in the subjective experience that defies the adaptation of a gold standard developed by a single judge.

The focus of the research has been on finding new ways to facilitate the user annotation, search and browse experience. Instead of the quantitative evaluation metrics described above, we will examine our system qualitatively for strengths and weaknesses. We will examine how well the system supports the user's image retrieval experience.

5.3 User Feedback

User feedback of the system was gathered informally from users of the Treehouse online community. Other observations were made as the system was demonstrated at two half-day Media Lab Open House events in March 2004 at which the system was used by dozens of people. The system is fully integrated with the functioning Treehouse online community that is open to a selected group of about 50 users. These users include MIT students, researchers and research sponsors of the MIT Media Lab. We acknowledge that this is not a naïve user group, but will focus on getting a preliminary first round of feedback to distill possible extensions to the image retrieval system.

We present user feedback below in three sections that reflect the three foci of the thesis.

5.3.1 Keyword annotation and knowledge aggregation feedback

The automatic saving ability of the Annotation Application was found to be easy to use and intuitive. In order to annotate an image, the user simply had to type a new keyword in the text field of the Annotation Application and press enter. The new keyword immediately showed up in the list of keywords for the image and the association is saved to the database behind the scenes. The next time *any* user loads this image, they would see the newest keyword in association. The user interface to pinpointing a keyword and resizing the pinpoint area was also found to be intuitive once the user is introduced to it. However, the introduction could be made clearer with a status bar for the application. Actually, the graphical user interface of the whole image retrieval system lacks direction-providing status bars.

The knowledge aggregation of image keyword data across all users of the community was found to be extremely useful for beginning users to quickly see some images in the database in a meaningful way. Since all keywords become immediately connected to each other through conceptual links provided by WordNet and OpenMind data, a user is able to immediately begin issuing queries to search and browse images of the database without having to annotate images themselves first. The Dynamic Semantic Hierarchy was found to be useful as a guide for concepts that do not yet have images associated with them. For example, a user can easily notice that a parent concept such as “fruits” only has “apple” and “banana” as its children. This means that the database contains no images labeled with other types of fruit, which can prompt a user to create such images using the Treehouse application toolkit (such as the Draw application). This is a great way for individual users to actively fill in the gaps of the community aggregate knowledge and is something that is missing from the OpenMind Commonsense data collection process (Singh 2002).

The absence of any image processing in the system was found to be lacking. If the Treehouse database had ten images of apples, then a user or group of users would have to separately annotate all ten images with “apple”. A valuable extension to the system would be the incorporation of some mechanism for a user to teach the system what an apple looks like so that the system can go ahead and try to automatically annotate other

such concepts. The pinpoint mechanism of the Annotation Application could be utilized to help this object recognition since it allows users to locate keywords in sub-regions of images. This would require some form of *relevance feedback* mechanism by which users can judge and check the results of automated annotation. The co-existence of such Content-Based Image Retrieval (CBIR) techniques along with Keyword-Based Image Retrieval is beginning to appear in some other systems (Lu, Zhang et al. 2003) and seems to be a good way to take advantage of the benefits of both approaches. Keywords will be able to provide semantic depth to address the Semantic Gap weakness of CBIR systems while CBIR techniques can help overcome the shortcomings of the requirement for manual annotation for keywords. This is an exciting direction for future image retrieval systems to explore (Jorgensen 2003).

5.3.2 Query expansion feedback

Although the WordNet and OpenMind query expansion results that show up in the *Nouns* and *Context* tabs were found to be useful, users still used the *Keywords* tab which returned exact keyword matching results. Exact retrieval was useful for the situations where users knew their intended target. This was often the case with specific images and keywords a user had them added to the system, such as in the first scenario presented in Chapter Two. Exact keyword matching was still the fastest way to get at an image when the user knows exactly which words to search for to get the specific image they want. On the other hand, the query expansion was much more useful for serendipitous browsing (Kuchinsky, Pering et al. 1999), when the user did not have a particular target image in mind. One scenario is when a user is searching for *potential images* to use in an image collage with a particular theme. For example if the theme is fruits, then using the query expansion to get all fruit images in the database at once is much more efficient than separate searches for specific types of fruit. This is true especially since the user may not know beforehand all the fruits they are interested in and which of these terms have images associated with them in the Treehouse database.

The word links in the results pages of the query expanded tabs (*Nouns* and *Context*) worked as designed and allowed a user to continue to refine their search just by clicking on related terms. However, this image browsing experience did not support a “back” button to return to a prior search item, nor did the system have a “history” list of the most recent searches.

The system can be extended by keeping track of a user’s history and to use this knowledge to further confine the context of relevance in query expansion. Potential mechanisms for history tracking have been demonstrated in other systems (Shneiderman and Kang 2000). But this system affords the unique opportunity to use the OpenMind data to find more and more specific conceptually related terms. For example, a first query of dog could return contexts such as “beach”, “leash” and “bark”. However, after the user traverses down a link (lets say “beach”), then the OpenMind data can look up contexts for both concepts simultaneously (“dog” and “beach”) and more specific conceptually related terms may arise (such as “Frisbee” or “catch”). Of course there must also be ways of disabling the feature so users can have a fresh start when desired.

History tracking could also help rank retrieved images. Currently, images are returned in random order on the *Nouns* and *Context* tabs. One image tagged with “apple” is just as likely to appear near the top of the returned results list as another image tagged with “apple”. The tracking of the user’s history of queries can help make it easier to distinguish between which image is more relevant by examining the relevance of the other keywords to the history of queries.

5.3.3 Emotional content retrieval feedback

The single-click method for adding emotional keywords in the Annotation application was well received. Currently, users can retrieve images by emotional content by clicking on an emotion word at the top of the *Emotions* tab. This caused some confusion because it is different from how the *Nouns* and *Context* tabs are searched (via the *Image Search* text field). Other than this inconsistency in the search experience, the Image Retrieval Application was found to be compact, intuitive and informative. The system could be made better by a redesign of the search front-end for a consistent experience across all the results tabs.

Currently, the emotional content retrieval ranks results by the number of users who have labeled an image with a particular emotion. The absolute number of users is used in the ranking algorithm. This could be problematic because different images may receive different viewer ship and some could have more chances at being annotated than others. For example, for a query of “happy”, an image tagged by four people as “happy” and four people as “sad” will be ranked higher than an image that only has three people who have tagged it as “happy”. A more advanced ranking algorithm will take into account the dispersion of aggregate user annotations for emotional content. This way the image could have a normalized score for “happiness” or “sadness” that takes into account the spread of the types of emotional groups chosen by different users. So a query for “happy” images would return in order of decreasing amounts of “happiness” as calculated by the normalization algorithm.

Another weakness of the system is that although it is possible to see the results of WordNet expansion, OpenMind expansion and emotional retrieval on separate tabbed windows, there is currently no way to integrate the methods for a more sophisticated query. For example, a user may want to find all images that have semantic connections to “ocean” ranked by their “happiness” level. This type of integrated retrieval would be an invaluable addition to the system.

5.4 Future extensions

In summary of the feedback from previous sections, the limitations of our image retrieval system include:

- The lack of image processing abilities to augment the semantic power for a hybrid CBIR and Keyword-based image search that exploits the advantages of both approaches.
- The lack of tracking of the history of user queries.
- The lack of using history information to automatically refine search.
- The lack of integrated search capabilities between semantic contexts and emotions.
- The lack of a ranking system for query expansion retrieval.
- The lack of a ranking system for emotional content retrieval that normalizes the aggregation of user annotations across different emotional groups.

Some of these limitations have been explored else in other image retrieval systems (Salton and McGill 1983). Other limitations, such as those dependent on an emotional retrieval mechanism, are entirely novel to this system. The issue of scalability has also not been yet adequately explored. This is a vital limitation as image retrieval behavior depends highly on how populated the image database is.

Some obvious extensions to the system that would be beneficial would be remedies for the limitations discussed above. Other potential extensions include:

- Go beyond single word annotations to take into account sentences.
- Add more subjective ways for categorizing images. For example, *smell* and the *tactile* nature of what is depicted in the image can be added in a similar fashion as the *Emotions* tab. These data points can also be integrated with query expansion. A possible combination query may then be for an image that is conceptually related to “ocean”, which is “happy” and smells “sweet”.
- Currently, individual keywords are associated with certain parts of an image or the entire image. There is no mechanism for keeping track of the relationships between keywords. This might be used to visually teach the system, for example, that all cars have wheels and an engine. This type of visual relationship information could then be fed back to the OpenMind Commonsense database to enrich the OpenMind dataset.
- Opening up the annotation process to the world outside of the Treehouse online community. We noted that users had fun with the *Random* tab to randomly retrieve images and to annotate ones they thought were interesting. Currently, the image retrieval system is integrated with Treehouse Studio completely and is only accessible to users of the community. It would be beneficial to open up the annotation process to the rest of the web community (ESPGame 2003).

These extensions would continue the trend of offloading the processing requirements of the user onto the system and bring us closer to an intuitive and accessible image retrieval system for the casual user.

6 Conclusions

In this thesis we have explored three weaknesses of keyword-based image retrieval through the design and implementation of an actual image retrieval system. The first weakness is the requirement of heavy manual annotation of keywords for images. We investigated this weakness by aggregating the annotations of an entire community of users to alleviate the annotation requirements on the individual user. The second weakness is the hit-or-miss nature of exact keyword matching used in many existing image retrieval systems. We explored this weakness by using linguistics tools (WordNet and the OpenMind Commonsense database) to locate image keywords in a semantic network of interrelated concepts so that retrieval by keywords is automatically expanded semantically to avoid the hit-or-miss problem. Such semantic query expansion further alleviates the requirement for exhaustive manual annotation. The third weakness of keyword-based image retrieval systems is the lack of support for retrieval by subjective content. We investigated this weakness by creating a mechanism to allow users to annotate images by their subjective emotional content and subsequently to retrieve images by these emotions.

We have based the design and implementation of our image retrieval system on past research that sheds light into how users actually describe images in order to explore the three foci of thesis. Our image retrieval system has successfully created some new ways of getting at images with words. Our system tries to do as much of the query analysis for the user as possible and presents the user with a wide array of potentially relevant results to enable them to further refine their search.

The system was built with a light-client, heavy-server model such that all the semantic processing happens on the server. This means the mechanisms for query expansion described here can be easily added to existing image database systems (such as PictureQuest) without drastic redesign of the front-end user interface. Existing commercial systems can easily enhance their systems with the technology described here without drastic changes to their backend infrastructure.

7 References

- Akrivas, G., M. Wallace, et al. (2002). "Context-Sensitive Semantic Query Expansion." Proceedings of IEEE International Conference on Artificial Intelligence Systems.
- Berger, J. (1973). Ways of Seeing. London, BBC and Penguin Books.
- Bimbo, A. D. (1999). Visual Information Retrieval. San Francisco, Morgan Kaufmann Publishers.
- Carrol, J. D. and P. Arabie (1980). "Multidimensional scaling." Annual Review of Psychology 31: 607-649.
- Chang, S. F., J. R. Smith, et al. (1997). "Visual Information Retrieval From Large Distributed Online Repositories." Communications of the ACM 40(12): 63-71.
- Deerwester, S., S. Dumais, et al. (1990). "Indexing by Latent Semantic Analysis." Journal of the American Society for Information Science 41(6): 391-407.
- Dondis, D. A. (1973). A Primer of Visual Literacy. Cambridge, MA, MIT Press.
- Doorn, M. v. and A. P. d. Vries (2000). "The Psychology of Multimedia Databases." Proceedings of ACM Conference on Digital Libraries: 1-9.
- ESPGame (2003). ESPGame.
- Evans, D. (2001). Emotion. Oxford, Oxford University Press.
- Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. Cambridge, MA, MIT Press.
- Finke, R. A. (1989). Principles of Mental Imagery. Cambridge, MA, MIT Press.
- Flickner, M., H. Sawhney, et al. (1995). "Query by image and video content: the QBIC system." IEEE Computer 28(9): 23-32.
- Furna, G. W., T. K. Landauer, et al. (1987). "The Vocabulary Problem in Human-System Communication." Communications of the ACM 30(11): 964-971.
- Gong, Y. (1998). Intelligent Image Databases: Towards Advanced Image Retrieval. Boston, Kluwer Academic Publishers.
- Gonzalo, J., F. Verdejo, et al. (1998). "Indexing with WordNet synsets can improve text retrieval." Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems.
- Google (2004). Google Image Search.
- Gordon, I. E. (1989). Theories of Visual Perception. Chichester, John Wiley & Sons.
- Gupta, A. and R. Jain (1997). "Visual Information Retrieval." Communications of the ACM 40(5): 71-79.
- Hochberg, J. E. (1964). Perception. Englewood Cliffs, N.J., Prentice Hall.
- Hoffman, D. D. (1998). Visual Intelligence: How we create what we see. New York, W. W. Norton & Company.
- Itten, J. (1973). The Art of Color: The Subjective Experience and Objective Rationale of Color. New York, Van Nostrand Reinhold Company.
- Jaimes, A. and S.-F. Chang (2000). "A Conceptual Framework for Indexing Visual Information at Multiple Levels." Proceedings of IS&T/SPIE Internet Imaging 3964.

- Jing, Y. and W. B. Croft (1994). "An Association Thesaurus for Information Retrieval." Proceedings of RIAO: 146-160.
- Jorgensen, C. (2003). Image Retrieval: Theory and Research. Oxford, UK, Rowman & Littlefield.
- Kuchinsky, A., C. Pering, et al. (1999). "FotoFile: A Consumer Multimedia Organization and Retrieval System." Proceedings of CHI The Conference on Human Factors in Computing Systems: 496-503.
- Lang, P. J., M. M. Bradley, et al. (1997). Motivated Attention: Affect, Activation, and Action. Attention and Orienting: Sensory and Motivational Processes. P. J. Lang, R. F. Simons and M. T. Balaban. Mahwah, N.J., Lawrence Erlbaum Associates: 97-135.
- Lee, H.-M. (2001). "Interactive Query Expansion Based on Fuzzy Association Thesaurus for Web Information Retrieval." Proceedings of IEEE International Fuzzy Systems Conference: 724-727.
- Lenat, D. B., R. V. Guha, et al. (1990). "Cyc: toward programs with common sense." Communications of the ACM 33(8): 30-49.
- Lim, J.-H., H.-W. Seung, et al. (1997). "Query Expansion for Intelligent Information Retrieval on Internet." Proceedings of Parallel and Distributed Systems: 656-662.
- Liu, H. and P. Singh (2003). OMCSNet: A Commonsense Inference Toolkit.
- Lu, Y., H. Zhang, et al. (2003). "Joint Semantics and Feature Based Image Retrieval Using Relevance Feedback." IEEE Transactions on Multimedia 5(3): 339-347.
- Maeda, J. (2004). Treehouse Studio.
- Marr, D. (1982). Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. San Francisco, W. H. Freeman and Company.
- Matravers, D. (1998). Art and Emotion. Oxford, Clarendon Press.
- McDonald, S. and J. Tait (2003). "Search Strategies in Content-Based Image Retrieval." Proceedings of Conference on Information Retrieval: 80-87.
- Pentland, A., R. W. Picard, et al. (1993). "Photobook: Content-Based Manipulation of Image Databases." MIT Media Laboratory Perceptual Computing Technical Report 255.
- PictureQuest (2004). PictureQuest: Royalty Free and Rights Protected Images Online.
- Pinker, S. (1985). Visual Cognition. Cambridge, MA, MIT Press.
- Salton, G. and M. J. McGill (1983). Introduction to Modern Information Retrieval. New York, McGraw-Hill.
- Scherer, K. R. and P. Ekman (1984). Approaches to Emotion. Hillsdale, N. J., Lawrence Erlbaum Associates.
- Schettini, R. (1994). "Multicoloured object recognition and location." Pattern Recognition Letters 15: 1089-1097.
- Sekuler, R. and R. Blake (2001). Perception. Boston, McGraw Hill.
- Shepard, R. N. (1962). "The analysis of proximities: multidimensional scaling with unknown distance function Part I and II." Psychometrika 27: 127-140; 219-246.
- Shneiderman, B. and H. Kang (2000). "Direct Annotation: A Drag-and-Drop Strategy for Labeling Photos." Proceedings of Conference on Information Visualization: 88-95.
- Singh, P. (2002). The Open Mind Common Sense project, KurzweilAI.net.

- Solomon, R. C. (2003). What is an Emotion? Oxford, Oxford University Press.
- Solso, R. L. (1994). Cognition and the Visual Arts. Cambridge, MA, MIT Press.
- Spoehr, K. T. and S. W. Lehmkuhle (1984). Visual Information Processing. San Francisco, W. H. Freeman and Company.
- Stein, N. L. and K. Oatley (1992). Basic Emotions. Hillsdale, Lawrence Erlbaum Associates.
- Takagi, T. and M. Tajima (2001). "Query Expansion Using Conceptual Fuzzy Sets For Search Engines." Proceedings of IEEE International Fuzzy Systems Conference: 1303-1308.
- Torgeson, W. S. (1965). "Multidimensional scaling of similarity." Psychometrika 30: 379-393.
- Torrallba, A. and A. Oliva (2002). "Depth Estimation from Image Structure." IEEE Transactions on Pattern Analysis and Machine Intelligence 24(9): 1226-1238.
- Vasile, A. and W. Bender (2001). "Image Query Based on Color Harmony." Proceedings of SPIE: 4299-56.
- Veltkamp, R. C., H. Burkhardt, et al. (2001). State-of-the-Art in Content-Based Image and Video Retrieval. Dordrecht, Kluwer Academic Publishers.
- Wade, N. J. and M. Swanston (1991). Visual Perception. London, Routledge.
- Yang, J., L. Wenyin, et al. (2001). "Thesaurus-Aided Approach For Image Browsing and Retrieval." Proceedings of IEEE International Conference on Multimedia and Exposition: 313-316.
- Yee, K.-P., K. Swearingen, et al. (2003). "Faceted Metadata for Image Search and Browsing." Proceedings of CHI The Conference on Human Factors in Computing Systems: 401-408.