PERCEIVED FEATURES AND SIMILARITY OF IMAGES: AN INVESTIGATION INTO THEIR RELATIONSHIPS AND A TEST OF TVERSKY'S CONTRAST MODEL

Abebe Rorissa, B.S., M.S.

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

May 2005

APPROVED:

Samantha K. Hastings, Major Professor and
Interim Dean of the School of Library and
Information Sciences
Brian C. O'Connor, Committee Member and
Associate Director of the Interdisciplinary
Information Science Doctoral Program
Linda Schamber, Committee Member
Gerald Knezek, Committee Member
Sandra L. Terrell, Dean of the Robert B. Toulouse
School of Graduate Studies

Rorissa, Abebe, *Perceived features and similarity of images: An investigation into their relationships and a test of Tversky's contrast model.* Doctor of Philosophy (Information Science), May 2005, 125 pp., 15 tables, 15 figures, references, 115 titles.

The creation, storage, manipulation, and transmission of images have become less costly and more efficient. Consequently, the numbers of images and their users are growing rapidly. This poses challenges to those who organize and provide access to them. One of these challenges is similarity matching. Most current content-based image retrieval (CBIR) systems which can extract only low-level visual features such as color, shape, and texture, use similarity measures based on geometric models of similarity. However, most human similarity judgment data violate the metric axioms of these models.

Tversky's (1977) contrast model, which defines similarity as a feature contrast task and equates the degree of similarity of two stimuli to a linear combination of their common and distinctive features, explains human similarity judgments much better than the geometric models. This study tested the contrast model as a conceptual framework to investigate the nature of the relationships between features and similarity of images as perceived by human judges. Data were collected from 150 participants who performed two tasks: an image description and a similarity judgment task. Qualitative methods (content analysis) and quantitative (correlational) methods were used to seek answers to four research questions related to the relationships between common and distinctive features and similarity judgments of images as well as measures of their common and distinctive features.

Structural equation modeling, correlation analysis, and regression analysis confirmed the relationships between perceived features and similarity of objects hypothesized by Tversky (1977). Tversky's (1977) contrast model based upon a combination of two methods for measuring common and distinctive features, and two methods for measuring similarity produced statistically significant structural coefficients between the independent latent variables (common and distinctive features) and the dependent latent variable (similarity). This model fit the data well for a sample of 30 (435 pairs of) images and 150 participants (χ^2 =16.97, *df*=10, *p* = .07508, RMSEA= .040, SRMR= .0205, GFI= .990, AGFI= .965). The goodness of fit indices showed the model did not significantly deviate from the actual sample data.

This study is the first to test the contrast model in the context of information representation and retrieval. Results of the study are hoped to provide the foundations for future research that will attempt to further test the contrast model and assist designers of image organization and retrieval systems by pointing toward alternative document representations and similarity measures that more closely match human similarity judgments.

ACKNOWLEDGMENTS

In the writing of this dissertation, I have been guided, assisted, and supported by several people. It would be impossible to list all of them here. My special thanks go to my mentor and chair of my dissertation committee, Dr. Samantha Hastings, for not only opening new windows of opportunity for me, but also for being a consistent source of academic, social, and economic support. I would also like to thank my dissertation committee members, Dr. Linda Schamber, Dr. Brian O'Connor, and Dr. Gerald Knezek. They have all introduced me to new ideas, tools, methods, and techniques and provided me with invaluable input and feedback.

I wish to express my gratitude to the many people at Addis Ababa University, National University of Lesotho, University of Namibia, and University of North Texas, without whose assistance and advice I wouldn't be where I am today. To all participants of this study, I owe a great deal of gratitude.

Finally, my sincere thanks go to Daniel Gelaw, for being a friend in need, and to my wife and son, Mekdes and Markos, for their generosity, love, and unfaltering support throughout the entire period of my studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

INTRODUCTION

General Background

People from all walks of life as well as organizations use images for their daily activities including business, medical treatment, education, and entertainment. Images are "becoming an integral part of human communication" (Chang, Smith, Beigi, & Benitez, 1997, p. 63). The creation, storage, manipulation, and transmission of images have become less costly and more efficient. People can even capture, store, transmit and print images using their mobile telephones. Consequently, the numbers of both images and their users are growing rapidly. For instance, one of the major suppliers of stock images, Getty Images, Inc. (http://www.gettyimages.com) has over 70 million still images in its collection. This presents a tremendous challenge for those who design and implement image storage and retrieval systems. Two of the main challenges relate to indexing or representation of images and to the measurement of their similarity for organization and retrieval purposes as well as visualization of both stored and retrieved sets of images. Researchers and practitioners disagree on the significant attributes that should be considered for indexing images and this has been reflected in results of investigations that looked into the appropriateness of some of the traditional image indexing tools (Jörgensen, 2003). Furthermore, there is lack of research into the exact nature of image perception by users and the criteria they use to make similarity judgments.

Similarity plays an important role in human perception (Melara, 1992; Tversky, 1977; Tversky & Gati, 1978) and information organization and retrieval (Qin, 2000; Santini & Jain, 1999; Zhang & Korfhage, 1999a, 1999b; Zhang & Rasmussen, 2001). For instance, in libraries and other types of information systems, information objects are categorized according to their similarity (or proximity) based on their physical form and/or intellectual content. A major component of any information retrieval system is similarity matching to determine inter-document similarity, which is the degree of similarity between documents (or their representations) in the system, and/or the degree of similarity between representation of a user's query and documents (or their representations) in the system.

Even though geometric/spatial models of similarity are widely utilized to determine inter-document similarity and in visualizations of sets of documents stored in information retrieval systems and/or sets of retrieved documents, some weaknesses in their metric assumptions/axioms have been identified by Tversky (1977). Consequently, Tversky (1977) formulated an alternative set-theoretical model known as the contrast model. This study used the contrast model as the conceptual framework to investigate the nature of the relationships between perceived features and similarity of images.

Statement of the Problem

The amount of research on text representation and retrieval of text documents dwarfs that on representation and retrieval of image documents (Jörgensen, 1995; Lynch, 1991; Shatford, 1986). Even within the image indexing and retrieval literature, content-based (automatic/machine-based) image indexing and retrieval literature outnumbers the concept-based (manual, text-based) literature (Chu, 2001), even though there is continued dependence on concept-based image indexing and retrieval (Enser, 2000). Lacking from both sets of literature are detailed investigations of the nature of human similarity judgments of images, specifically the relationships between human similarity judgments and the common and distinctive features of images.

Image attributes are the key to their description, indexing, and representation (Rasmussen, 1997). However, only a handful of researchers (Greisdorf & O'Connor, 2002a, 2002b; Hastings, 1995; Jörgensen, 1995, 1996, 1998, 2003; Lee, 2001) have conducted studies to identify attributes of images generally perceived by users and current concept-based image indexing mechanisms hardly rely upon any sort of theoretical foundation (Jörgensen, 2003). One of these researchers who undertook such a study found that there is "a very incomplete match between the attributes addressed by major textual image indexing systems and thesauri in use and the attributes described by participants in empirical research" (Jörgensen, 2003, p.243).

Moreover, lack of research regarding the nature of visual similarity has already been identified (Jörgensen, 1995; Santini & Jain, 1999). Despite the major role similarity plays in human perception (Melara, 1992; Tversky, 1977; Tversky & Gati, 1978) and information organization and retrieval (Qin, 2000; Santini & Jain, 1999; Zhang & Korfhage, 1999a, 1999b; Zhang & Rasmussen, 2001), the various psychological models of similarity such as Tversky's (1977) contrast model have not been used enough to investigate problems in library and information science. Even though the geometric/spatial models of similarity (especially multidimensional scaling) are widely applied in information retrieval in general and image retrieval in particular to determine the extent of inter-document similarity and for visualizations of sets of stored or retrieved text and image documents (e.g. Rubner, 1999), some weaknesses in their metric axioms (the fact that human similarity judgment data violate these axioms) have been identified by Tversky (1977). Therefore, there is a need for research that uses the contrast model as a framework not only to bridge the gap in the relevant literature, but also to explore alternative methods of image indexing and retrieval. This study attempts to achieve that by investigating the nature of the relationships between perceived features and similarity of images.

Research Questions

This study used Tversky's (1977) contrast model as a general framework and attempted to answer the following research questions:

RQ1: Which methods measure the common and unique/distinctive features of images well?

RQ2: To what extent does the contrast model fit human similarity judgments and features/attributes data for a sample of images?

RQ3: What is the relationship between perceived similarity of images, as judged by humans, and their features/attributes identified and described/listed by humans?

RQ4: What are the relative weights given to common and unique/distinctive features in human similarity judgments of images?

Purpose of the Study

The study uses both qualitative (content analysis) and quantitative (correlational) methods to investigate the above research questions pertinent to perceived features and similarity of images. Its main purposes are:

- To determine, for a sample of images, methods that measure the common and unique/distinctive features of images well through testing the three measurement models for common and unique/distinctive features using structural equation modeling (SEM).
- To investigate the extent of the fit of Tversky's (1977) contrast model to human similarity ratings and common and unique/distinctive features data for the same sample of images using structural equation modeling (SEM);
- To determine the relationship between perceived similarity of images, as judged by humans, and their common and unique/distinctive features/attributes, identified and described/listed by humans; or specifically, to see whether higher common features and lesser unique/distinctive features result in higher average similarity ratings; and
- To determine, for the same sample of images, the relative weights given to common and unique/distinctive features in human similarity judgments through structural equation modeling (SEM) and regression analysis.

Significance of the Study

Given the importance of attributes and similarity matching and measurement in text and image indexing and retrieval, this study is among the first known to investigate perceived features and similarity of images, in the context of their representation and retrieval, using the contrast model as a framework. Results of this study will serve as a foundation for future research that will attempt to further test the model. Results may inform image indexers about important attributes. Furthermore, results obtained from the test of the contrast model will inform designers of retrieval systems about similarity measures that match perceived similarity and could be used to formulate alternative similarity measures for image retrieval systems that mimic perceived similarity. This research will begin to fill a gap in the literature.

The study is also significant in that its results challenge the widely accepted notion that documents (both text and image) can be represented as points in a continuous multidimensional space and their similarity computed using distance functions that satisfy the three metric axioms of minimality ($\delta(x,y) \geq \delta(x,x)=0$), symmetry ($\delta(x,y)=\delta(y,x)$), and the triangle inequality ($\delta(x,z) \leq \delta(x,y) + \delta(y,z)$), where x, y, and z are three points representing three objects in the multidimensional space and δ is a metric distance function. Instead, the results support the idea that these documents can be represented as sets of discrete features and that their similarity can be determined by contrasting their common and unique/distinctive features with varying weights for common and unique/distinctive features. Results of the tests of the measurement models will inform future research as to which measures of the common and unique/distinctive features of images measure the constructs well and provide reliability and validity coefficients to serve as references.

Basic Assumptions

It is safe to assume that human participants in the study are able to perceive and identify possible features/attributes of images and are also able to describe/list them. Furthermore, verbal and written descriptions as well as any data derived thereof can be used for further analysis to address the research questions (Ericsson & Simon, 1980).

Perceived similarity, like any perceptual continuum, is assumed to be scalable (measurable). Participants are assumed to be able to produce at least interval level data when asked to judge the degree of similarity of pairs of stimuli.

Despite recent developments in content-based automatic feature extraction and indexing of images using low-level features such as color, shape, and texture (Faloutsos et al., 1994; Flickner et al., 1995; Holt et al., 1997; Rui, Ortega, Huang, & Mehrotra, 1999; Smith & Chang, 1996), textual descriptions of image attributes remain popular methods for indexing (Enser, 2000; Jörgensen, 1995, 1996, 1998, 2003). Quite often, users formulate their queries or describe images using words (or natural language) as well (O'Connor, O'Connor & Abbas, 1999).

Definitions of Terms

Image

An image is "a visual representation of an object or scene or person produced on a surface" (Hyperdictionary, 2003). In the context of this study, an image is a digital color picture/photo of various types of objects with both indoor and outdoor surroundings stored in the Joint Photographic Experts Group (JPEG) format and displayed on a standard personal computer monitor.

Perception

Human perception is the process of receiving information about a stimulus/object through one's senses and organizing it as well as interpreting that information (Hyperdictionary, 2003). In this study, perceived features and similarity of images are considered to be results of more than just "feeling" them through one's senses. They are results not only of visual experience but are also results of interpretations of visual and other attributes based on participants' past experience and knowledge.

Attribute/Feature

In this study, the two terms, attribute and feature, are used interchangeably. An attribute of a stimulus (or object) is a characteristic of contents of the stimulus and an image attribute is a characteristic of both its visual and nonvisual contents. Layne (1994, p. 586) defines an image attribute as "what is depicted or represented in the image" while Jörgensen (2003, p. 3) expands on what is depicted in an image by stating that it is "not limited to purely visual characteristics, but includes other cognitive, affective, or interpretive responses to the image such as those describing spatial, semantic, or emotional characteristics."

Representation

Representation is a mechanism where "one thing stands for another" (O'Connor, 1996, p. 11). It is a concept that encompasses other concepts such as indexing and abstracting. O'Connor (1996) emphasizes the significance of attributes of entities in their representation. Representation of a document (both text and image) is, among others things, the creation of a surrogate for the document and can take the form of a record in a database, a textual/verbal description, a smaller version of the document (especially for image documents), a single or a set of parts/sections of the document (e.g. a frame(s) extracted from a moving image document), or an assigned set of terms that stand for contents of the document being represented.

Similarity Measure

In the context of image and text document retrieval, a similarity measure is a metric used to determine either the proximity/similarity of the representation of a document (or representations of a collection of documents) to the representation of a query or the proximity of representations of documents (or inter-document similarity) in a collection. The most widely used similarity measures are the distance-based (such as the Minkowski metric) and angle (cosine)-based similarity measures. In content-based image retrieval (CBIR), the distance-based measures are more popular (Gupta, Santini & Jain, 1997).

Limitations and Delimitations of the Study

One major limitation of the study stems from the fact that there are no standard test sets of images similar to the Text Retrieval Conference (TREC) document sets. Consequently, because the sample of images chosen for this study comes from the collection of images (color photographs taken by a number of photographers) published on a CD enclosed with the book by O'Connor & Wyatt (2004), results may not be generalizable to any other population of images (such as a collection of art images, a stock photo collection, etc.). Due to the specific data collection technique chosen (paired comparison, which involves $n(n-1)/2$ pairs of images for a sample of size n) to solicit similarity judgments of images by participants, the size of the sample of images for the study (30) may not be large enough to generalize results of the study to other stock photo or image collections. However, it is still large enough for the results to be valid for collections similar to the population of images.

The test of Tversky's (1977) contrast model is restricted only to similarity ratings of the form "how similar are images *a* and *b*" rather than subject/referent similarity ratings of the form such as "how similar is image a to image b." Because dissimilarity ratings were not collected, relationships between similarity and dissimilarity of images were not determined.

Summary

This chapter provides the general and theoretical background for the study and the delineation of the problem under study. The contrast model is established as the framework for the study and research questions concerning perceived features and similarity of images are presented. The research questions address the nature of the relationship between similarity and common and distinctive features of images including their relative weights and best measures, as well as the extent to which the contrast model fits the data for a sample of images. The main purposes, significance, and basic assumptions of the study are outlined. Limitations and delimitations of the study, in terms of the sample of images selected as well as data collection techniques used, are also identified. The next chapter presents a critical review of the literature that is relevant to this study.

CHAPTER 2

REVIEW OF THE LITERATURE

Introduction

This chapter presents a critical review of the literature relevant to the topic of the dissertation, the nature of the relationships between perceived features and similarity of images. More specifically, it presents a historical review of the main constructs of the study such as similarity and features/attributes, and a review of the theoretical framework for the study, such as psychological models of similarity in general and Tversky's (1977) contrast model in particular. A detailed review of the literature on similarity measurement in image retrieval, image features/attributes, representation, indexing and organization completes the chapter.

Psychological Models of Similarity

Similarity is one of the most important and well-researched constructs (Goldstone, 1999; Tversky, 1977). According to Melara (1992), the concept of similarity is central to the field of psychology and researchers' understanding of similarity as a construct must be anchored in the concept of perception. Others believe that similarity, as a construct, is "fundamental to theories of perception, learning, and judgment" (Tversky & Gati, 1978, p. 79) and that the "ability to assess similarity lies close to the core of cognition" (Goldstone, 1999, p. 757). Similarity "refers to the outcome of a comparison among entities, usually a comparison based on many of the entities' properties. Objects are similar to the degree that they have features in common and do not have distinctive features" (Sloman & Rips, 1998, p. 4). According to Tversky (1977, p. 327), similarity "serves *[as]* an organizing principle by which individuals classify objects, form concepts, and make generalizations." Unzicker, Jüttner & Rentschler (1998, p. 2289) argue that "classes or categories consist of collections of objects that are grouped by similarity." Most similarity analyses involve collection of data on ratings, by humans, of the degree of similarity of pairs of stimuli or the sorting of stimuli into groups based on their similarity and the use of geometric models that equate observed dissimilarities between stimuli to their metric distances between the points on a coordinate space (Tversky, 1977).

From around 1850, psychophysicists such as Fechner started to study the nature of human similarity judgments through investigations of relationships between physical and psychological changes. Fechner created a scale to measure the psychological change and called it the "just noticeable difference" or jnd (Melara, 1992). By creating this scale, Fechner laid the foundations for scaling or measurement of similarity relations as well as scaling or measurement in psychophysics, a field of psychology that studies the relationship between the physical world and its psychological representation/experience.

By introducing the concept of just noticeable difference, Fechner was also laying the foundations for the earliest model of similarity, based on the idea that jnd is a fixed entity. This assumption was challenged by Louis Leon Thurstone, who says a human observer gives "different comparative judgments on successive occasions about the same pair of stimuli" (Thurstone, 1927, p. 274). This led to the formulation of his law of comparative judgment, another classic model of similarity, which defines similarity relations as probabilistic.

Even though the two classic models of similarity pioneered by Fechner and Thurstone did not at first have a big impact in explaining similarity relations, they definitely influenced the scaling of similarity of stimuli as well as the development of multidimensional scaling (Melara, 1992). Multidimensional scaling is a geometric/spatial model of similarity where stimuli are represented as points on an n-dimensional ($n \geq 2$) space, usually Euclidean (Dunn-Rankin, Knezek, Wallace, & Zhang, 2004), and the distance between any two points determines the degree of psychological similarity of the two stimuli represented by the two points. The smaller this distance, the more similar the two stimuli are and vice versa.

While Fechner's just noticeable difference and Thurstone's Law of Comparative Judgment "examined the psychological properties of a single dimension of experience," multidimensional scaling "allowed an investigator to determine how many psychological dimensions subjects used when judging similarity" (Melara, 1992, p. 316). The geometric/spatial models of similarity, more specifically multidimensional scaling, are based on the assumption that the multidimensional space, on which the stimuli are represented as points in the space, is metric. Given any three points x, y and z in a multidimensional space, a metric space is one that satisfies the three metric axioms/conditions of positivity or minimality (distance between x and y is zero if they are identical and positive if they are distinct); symmetry (the distance from x to y is equal to the distance from y to x); and triangle inequality (the distance between x and z is less than or equal to the sum of the distance between x and y and the distance between y and z) (Figure 1 shows the three points in a two-dimensional space).

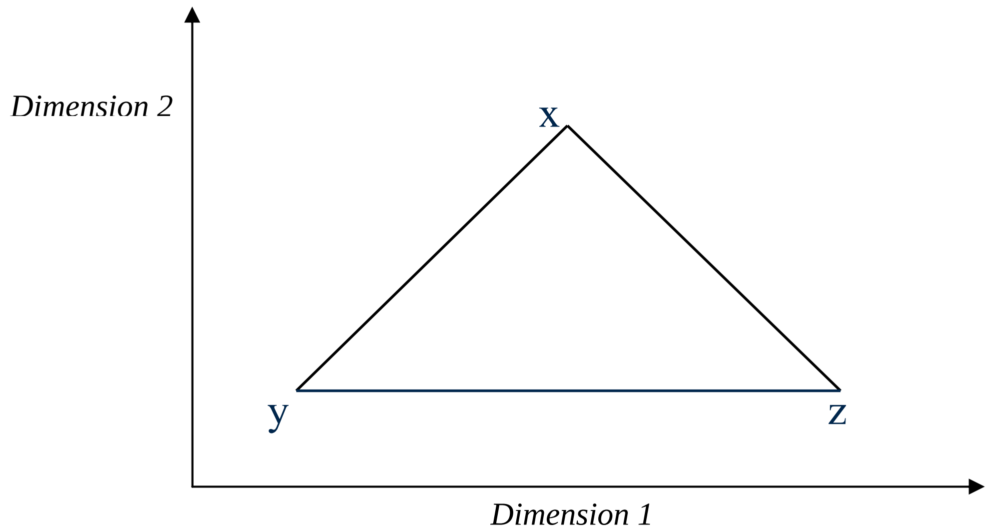
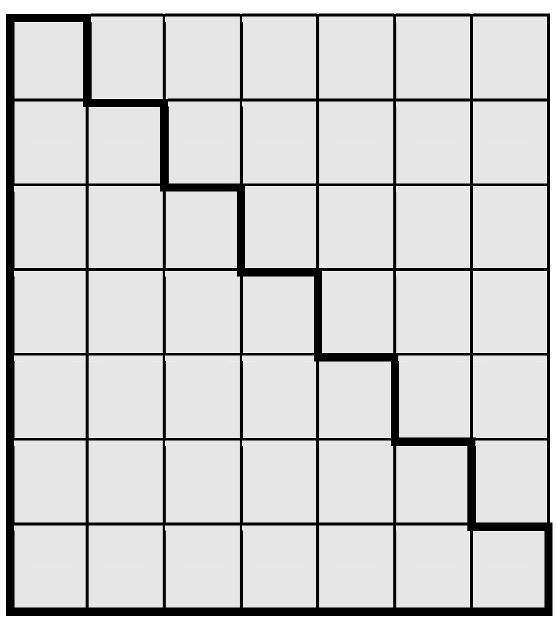


Figure 1. Three points in a two-dimensional space forming a triangle.

Even though the idea of representing stimuli as points in a multidimensional space is attractive, it requires a mechanism by which distances between the points are determined. The first researcher who attempted to tackle this issue and thus became a catalyst for the development and use of geometric/spatial models of similarity was Attneave (Melara, 1992). While it is easier to know what the dimensions of simple perceptive stimuli are and to determine, through experiments, how these dimensions combine to estimate overall similarity of stimuli (Attneave, 1950), this is not the case with more complex stimuli. Attneave (1950) limited his search for metrics or rules for determining the distances between the points in the multidimensional space, and hence the overall similarity of the stimuli represented by those points, to the Minkowski family of metrics. The two metrics chosen by Attneave (1950) were the Euclidean and the city-block or Manhattan metrics, two metrics which later became among the most popular in similarity measurement for text and image retrieval. Given two points, X and Y, in a pdimensional space ($p \geq 1$) with ($X_1, X_2, X_3, ..., X_p$) and ($Y_1, Y_2, Y_3, ..., Y_p$) as their respective coordinates, the distance between the two points, in terms of the Minkowski metric, is given by: $d(X,Y)=\sqrt[r]{\sum_{i=1}^{p}|X_i - Y_i|^r}$. When r=1, $d(X,Y)=\sum_{i=1}^{p}|X_i - Y_i|$, it is called the city-block or the Manhattan distance. When r=2, $d(X,Y)=\sqrt{\sum_{i=1}^{p}(X_i - Y_i)^2}$, it is the familiar Euclidean distance.

Figure 2 is an illustration of the city-blockdistance between two points, A and B. The rectangles in Figure 2 represent city blocks while the lines represent streets. In order to travel from one point to the other, one has to travel the same number of blocks regardless of the route taken.

A



B

Figure 2. "City-Block" distance between two points.

In Figure 3, the Euclidean distance between the two points x(x_1,x_2) and y(y_1,y_2) in a two-dimensional space is the hypotenuse of the right-angled triangle, which, according to the well-known Pythagoras theorem, is the square root of the sum of the squares of the lengths of the two legs, or $d(x, y) = \sqrt{\sum_{i=1}^{2} (x_i - y_i)^2}$.

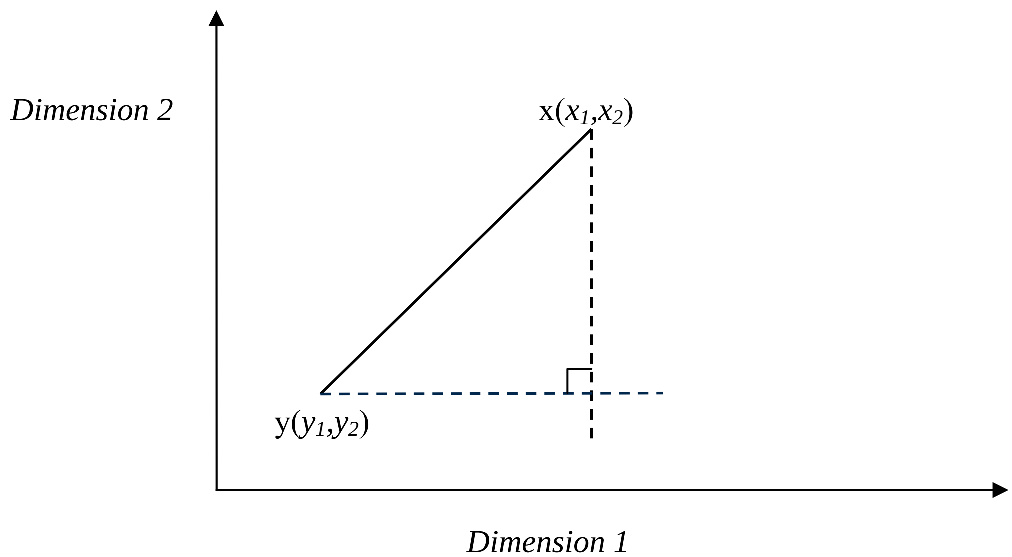


Figure 3. Euclidean distance between two points in a two-dimensional space.

Attneave, in 1950, is one of the early researchers to notice some of the weaknesses of the geometric/spatial models of similarity, especially the fact that human similarity ratings data do not satisfy the metric axioms. Others have gone even further and have not only formulated alternative models of similarity (Tversky, 1977), but also tested the alternative models under several conditions using a variety of stimuli (Gati & Tversky, 1984; Tversky & Gati, 1978; Tversky & Gati, 1982). Even with the introduction of non metric multidimensional scaling by Shepard (1962a, 1962b), which made possible the analysis of original similarity judgment data using a single technique through monotonic transformation, geometric/spatial models of similarity did not escape from criticism due to the fact that they, again, violate the metric axioms as well as a property of lines in the Minkowski family of geometries called segmental additivity (Melara, 1992). A line satisfies the property of segmental additivity if for three points A, B, and C on the line, in the same order, the sum of the distances from A to B and B to C equals the distance from A to C.

In line with Attneave's (1950) argument that metric axioms are not satisfied by similarity judgment data, hence the apparent weakness of geometric/spatial models of similarity, Tversky (1977) formulated an alternative set theory-based model called the contrast model. Tversky's (1977) contrast model, unlike geometric/spatial models of similarity, does not represent stimuli as points in a multidimensional space. Rather, it defines stimuli as sets of features and the similarity of any two stimuli as a linear function of a measure of their common and unique/distinctive features. According to the contrast model, two stimuli are more similar if they have more common features than unique distinctive features and human judges "place more weight on common features when judging similarity, and more weight on distinctive features when judging dissimilarity" (Melara, 1992, p. 346). Another assertion that sets the contrast model apart from the geometric/spatial models of similarity is that, while the geometric/spatial models consider dissimilarity to be the opposite of similarity, this is not always a given in the contrast model, even though Tversky & Gati (1978) report a near-perfect (r = -0.98) negative correlation between similarity and difference ratings of 21 pairs of countries. Perhaps the clearest distinction between the geometric/spatial models of similarity and the feature-based contrast model emanates from the fact that while dimensions have mutually exclusive levels (e.g. color has levels of red, green, blue, etc.), features are dichotomies (a stimulus either has or does not have a particular feature) (Rosch & Lloyd, 1978). This study looks only at the nature of human similarity judgments, rather than dissimilarity judgments, using the contrast model as a framework.

Most of the geometric/spatial models assume that humans, in their judgments of similarity and/or dissimilarity of stimuli, pay equal attention to the various dimensions. Shepard (1964) noticed that humans attach unequal weights to different dimensions and suggested that any analysis of similarity judgment data take into account individual differences among humans. This gave rise to a scaling procedure called individual differences scaling (or INDSCAL) (Melara, 1992).

Despite several critics pointing out the weaknesses of geometric/spatial models of similarity, these models still remain the dominant and most used (Tversky, 1977; Tversky & Gati, 1978). Apart from the geometric/spatial and Tversky's (1977) contrast models, the other two models of similarity of note are the transformational and alignment-based models. When it comes to well-structured stimuli, both geometric/spatial and feature-based models of similarity may not explain their similarity relations well. The transformational and alignment-based models are better suited to explaining the nature of similarity of such stimuli. Transformational models are based on the notion that two stimuli are more similar if few numbers of operations are required to make the two stimuli identical, by transforming one of the two (Goldstone, 1999). According to Goldstone (1999, p. 758), in alignment-based models "comparing things involves not simply matching features, but determining which elements correspond to or align with one another." In addition, in alignment-based models, features that align with one another may also need to have similar functions.

The four major models of similarity reviewed above, namely the geometric/spatial (e.g. MDS), feature-based (e.g. the contrast model), transformational, and alignment-based models, may not be well suited to explaining similarity relations of every possible set of stimuli due to weaknesses peculiar to each one of them. The strength of a similarity model should be measured, in part, on the basis of how well it explains the nature of human similarity judgments. An attempt was made in this study to test how well the contrast model does explain human similarity judgments of images.

The Contrast Model

People organize, group and categorize things based on their degree of similarity and separate them based on their degree of difference or dissimilarity. What makes two things similar has been a focus of several investigations in psychology, cognitive and behavioral sciences, and related fields for over 100 years (Melara, 1992). Through these investigations, a number of theories and models have been formulated and tested to explain perceived similarity (Attneave, 1950; Shepard, 1962a, 1962b; Thurstone, 1927; Tversky, 1977). One of these models is Tversky's (1977) contrast model. Tversky challenged the basic assumptions/axioms of the geometric/spatial models of similarity.

In a seminal paper, Tversky (1977) not only showed that metric axioms/assumptions of geometric/spatial models of similarity are violated by human similarity judgments data, but he also formulated and tested an alternative set-theoretical model of similarity called the contrast model. According to the contrast model, similarity judgment is a feature contrast task and the degree of similarity of two objects (or stimuli) is a linear combination of their common and unique/distinctive features. Furthermore, the model posits that two stimuli are more similar if they have more common features and fewer unique/distinctive features. They are less similar if they have more unique/distinctive features and fewer common features.

Figure 4 illustrates the relations between two sets. Given two stimuli *a* and *b* and their respective feature sets *A* and *B*, the perceived similarity of *a* and *b*, denoted by *s(a,b)*, is expressed as a linear function of the measures of their common and unique/distinctive features (Tversky, 1977; Tversky & Gati, 1978), and is given by:

$S(a,b) = \theta f(A \cap B) - \alpha f(A-B) - \beta f(B-A)$, where:

- $A \cap B$ represents the common features of a and b,
- A-B represents features of a that b does not have (distinctive features of a),
- B-A represents features of b that a does not have (distinctive features of b),
- θ, α, and β reflect weights given to the common and unique/distinctive features and are non-negative ($\theta, \alpha, \beta \geq 0$),
- S is an interval scale such that S(a,b) > S(c,d) if and only if s(a,b) > s(c,d), that is, a and b are more similar than c and d are,
- f is an additive function (that is, $f(A \cup B) = f(A) + f(B)$), whenever A and B are disjoint ($A \cap B = \emptyset$).

Another form of the contrast model, called the *ratio model*, is given by:

$S(a,b) = \dfrac{f(A \cap B)}{f(A \cap B) + \alpha f(A-B) + \beta f(B-A)}$, where $\alpha, \beta \geq 0$. The *ratio model* defines a normalized value of similarity such that $0 \leq S \leq 1$. It is a generalized form of set-theoretical models of similarity, including the contrast model (Tversky, 1977).

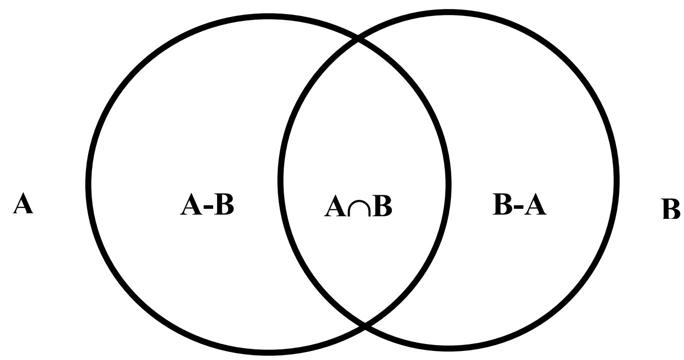


Figure 4. Graphical representation of the relation between two sets.

In the above equations, the function f, apart from being a measure of the common and distinctive features of the stimuli, is also an indicator of the salience or prominence of the stimuli. The information content, intensity, frequency, familiarity, and good form of the stimuli are among factors that contribute to their salience (Tversky & Gati, 1978).

Results of several tests of the contrast model agree that humans focus their attention more on common features when judging similarity than when judging differences of stimuli (Gati & Tversky, 1984; Johnson, 1981; Tversky, 1977; Tversky & Gati, 1978). On the effects of common and distinctive features of stimuli on their similarity and differences, Tversky & Gati (1978) add, “it is reasonable to assume that enlarging the measure of the common features increases similarity and decreases difference, whereas enlarging the measure of the distinctive features decreases similarity and increases difference” (p. 81). The above statement is perhaps another good reason for the use of correlational research methods in this study to investigate the relationships between similarity and the measures of common and distinctive features of images (see RQ3 & RQ4). Furthermore, when asked to rate the similarity of stimulus a to stimulus b (a being the subject and b the referent), humans tend to focus more on the subject (stimulus a) than the referent (stimulus b) and among different types of features, humans "attend primarily to features that have classificatory significance" (Tversky & Gati, 1978. p. 81).

The main reason for Tversky (1977) to formulate and test the contrast model, as indicated earlier, is the violation of the metric axioms, which are the basis of geometric/spatial models of similarity, by similarity rating data. Given three points x, y, and z in a multidimensional space and a metric distance function δ, the three metric axioms can be expressed as (Tversky, 1977):

Minimality: $\delta(x,y) \geq \delta(x,x)=0$

Symmetry: $\delta(x,y) = \delta(y,x)$

Triangle Inequality: $\delta(x,z) \leq \delta(x,y) + \delta(y,z)$

In support of his argument of the violation of the minimality axiom by similarity ratings, Tversky (1977) cites the fact that an object may not be recognized as itself all the time. In other words, two identical stimuli may not always be judged to be the same. Tversky (1977) provided evidence that the symmetry axiom, even though it is one of the basic assumptions of similarity theories, does not always hold. In his study of 21 pairs of countries, he noted that participants judged the similarity of North Korea to Red China to be greater than the similarity of Red China to North Korea. He attributes this asymmetry to the fact that "the variant [*North Korea*] is more similar to the prototype [*Red China*] than vice versa" (Tversky, 1977, p. 328), a notion supported by Rosch (1975). A t-test confirmed the asymmetry as well for both rated similarity and confusion data (Tversky, 1977; Tversky & Gati, 1978). The third metric axiom that drew criticism from Tversky

(1977) is the triangle inequality. To illustrate how this axiom is violated by similarity relations of three stimuli, he notes that even though Jamaica is similar to Cuba (geographic location) and Cuba is similar to Russia (ideology/politics), that does not make Jamaica and Russia similar.

As discussed in the preceding section, the geometric/spatial model and Tversky's (1977) contrast model of similarity are set apart by methods used to represent stimuli and determine their similarity. Geometric/spatial models represent stimuli as points in a multidimensional metric space and their similarity is determined by a distance function that is assumed to satisfy the three metric axioms. On the other hand, Tversky's (1977) contrast model represents stimuli as sets of features and the similarity of two stimuli is a function of the measures of their common and distinctive features, based on different weights for common and distinctive features. Tversky (1977) defines a feature or attribute of a stimulus as its components, whether concrete properties or abstract attributes. This definition was adopted for this study.

The first attempts to test the contrast model were made by Tversky (1977) and Tversky & Gati (1978) using both perceptual/visual stimuli (e.g. figures, letters of the alphabet, schematic faces) and semantic stimuli (countries, vehicles). Tversky (1977), in addition to showing the violation, by similarity ratings, of the three metric axioms, found significant correlations between average similarity ratings by 48 judges and two measures of common features (r=0.68 & 0.84) and distinctive features (r= -0.36 & -0.64) of 12 vehicles (bus, car, truck, motorcycle, train, airplane, bicycle, boat, elevator, cart, raft, sled). He also showed that the linear combination of measures of common and distinctive features account for close to 76% (multiple correlation coefficient, R=0.87) of the variance in similarity, a result that clearly supports the contrast model and, again, attests to the significance of this research for understanding the nature of human similarity judgments of images using the contrast model as a framework.

Subsequent to its formulation and testing by Tversky (1977), several researchers have either tested or used the contrast model as a framework for their studies in marketing and advertising (Johnson, 1981, 1986; Johnson & Horne, 1988), psychology (Ben-Shakhar & Gati, 1992; Dopkins & Ngo, 2001; Gati & Tversky, 1984), and consumer research (Ulhaque & Bahn, 1992). Johnson (1981) first tested Tversky's (1977) contrast model using consumer products such as soft drinks, frozen desserts, appliances, and fruits. Results of his study show the asymmetry property of similarity judgments data and they "support the generality of Tversky's theory in a consumer products context" (Johnson, 1981, p. 115).

Johnson (1986) also conducted a direct test of the contrast model through similarity, dissimilarity, and subject/referent similarity judgment tasks. He devised an indirect method for operationalizing the function f in the contrast model [$S(a,b)= \theta f(A \cap B) - \alpha f(A-B) - \beta f(B-A)$], which is a measure of the common and distinctive features of stimuli. The method involves dividing the total number of common (or distinctive) features attributed to the stimuli by each participant by the total number of participants who listed at least a single common (or distinctive) feature. It is one of the methods adopted for this study (*method 2*) and it is described in detail in chapter 3.

To determine the effects of common and distinctive features on similarity, dissimilarity, and subject/referent similarity judgments of various products, thereby testing the contrast model, Johnson (1986) used linear regression, one of the data analysis methods used in this study as well. Johnson (1986) divided a total of 87 participants into four groups to take part in the similarity/dissimilarity judgment tasks while all 87 participants completed the stimulus description task with a two minutes time constraint on each judgment (or pair of stimuli) to ensure validity. In order to conduct the study with several participants simultaneously, Johnson (1986) used a written format of the stimulus description task, an approach partially followed in this study.

Results of his correlation and regression analyses of the similarity/dissimilarity judgments data provided not only a direct support for the contrast model, they also proved Tversky's (1977) assertion that "common features add to similarity and detract from dissimilarity whereas distinctive features have the opposite effect" (Johnson, 1986, p. 55). Johnson (1986) came to the conclusion that "proximity judgments are well predicted by a simple linear combination or contrast of the average number of common and distinctive features associated with the *[stimuli]* being judged" (p. 56). This, once again, confirms the contrast model and offers support to the choice, for this study, of correlational (linear) methods such as linear regression, correlation, and structural equation modeling to test Tversky's (1977) contrast model using images as stimuli.

The main focus of Gati and Tversky's (1984) 15 studies involving verbal descriptions of people, meals, farms, sea scenes, faces, medical symptoms, schematic faces, profiles, landscapes, as well as drawings of sea scenes, images of faces, and pictures of sea scenes was the estimation of the relative weight, using the contrast model as a framework, of common to distinctive features on similarity judgments. They, once again, validated findings of earlier researchers and concluded that "the relative weight of common to distinctive features was higher in judgments of similarity" (Gati & Tversky, 1984, p. 341). This statement is relevant particularly to one of the research questions in this study (RQ4). A similar study by Ben-Shakhar and Gati (1992), through manipulation of components of stimuli (by removing or adding components), supports Gati and Tversky's (1984) assertions regarding relative weights of common and distinctive features on similarity judgments. However, Ben-Shakhar and Gati (1992) recommend further research to corroborate their results.

A few authors (Dopkins & Ngo, 2001; Medin, Goldstone, & Gentner, 1993; Tversky, 1977) suggested a possible effect of category relationships of stimuli, in addition to their common and distinctive features, on similarity judgments. Dopkins and Ngo (2001) specifically sought to investigate the contribution of category relationships to similarity judgments. Through their two experiments, one involving names of 15 vegetables and 15 non-vegetables paired with the higher-level semantic category "vegetable" (i.e. 30 pairs of stimuli) and another experiment involving 28 pairs of names of vegetables and 63 pairs formed from seven vegetable and nine non-vegetable names, Dopkins and Ngo (2001) found, using regression analysis, that category relationships had an effect on similarity judgments. They conclude that "A pair of concepts is perceived to be more similar if they are bound by a category inclusion relation than if they are not" (p. 251), that is, vegetable/vegetable pairs were judged to be more similar than vegetable/non-vegetable pairs. Perhaps this is true for conceptual/semantic stimuli such as names of vegetables with category relationships. However, it may not have any impact on this study because images, which are perceptual/visual stimuli, were used to test the contrast model.

In summary, while researchers in other fields such as psychology, advertising and marketing, and consumer research have realized the usefulness of Tversky's (1977) contrast model to understand and explain the nature of human similarity judgments of perceptual/visual, conceptual/semantic, and verbal stimuli, with results supporting the contrast model, no attempt has been made by library and information science researchers to test and use the model. This is despite the fact that the contrast model explains human similarity judgments much more than the other models of similarity and the fact that similarity measurement is a major component/function of most text and image document representation and retrieval systems that use the vector-space model of document representation. Hence, this study is significant in that it is the first investigation to look into the nature of human similarity judgments in the context of image retrieval using the contrast model as a theoretical framework. Results of the study will inform designers of image retrieval systems about alternative models of similarity measurement and assist the formulation of similarity measures that produce similar results to human judges.

Similarity Measurement in Image Retrieval

Similarity matching is the key task in information seeking, storage, and retrieval by both people and machines (Qin, 2000; Santini & Jain, 1999; Zhang & Korfhage, 1999a, 1999b; Zhang & Rasmussen, 2001). The majority of computerized systems use the vector-space model (VSM) of document representation (Zachary, 2000), in which similarity matching is achieved through similarity measures. Rubner (1999, p. 7), in explaining the function and importance of similarity measures in content-based image retrieval (CBIR) systems, states that: "In order for an image retrieval system to find images that are visually similar to the given query, it should have both a proper representation of the images' visual features and a measure that can determine how similar or dissimilar the different images are from the query."

Similarity measures are metrics used to determine the relevance of documents in a collection to queries based on proximities between their feature representations or to determine inter-document similarity. Similarity measures play important roles in both text retrieval (Qin, 2000; Zhang & Korfhage, 1999a; Zhang & Korfhage, 1999b; Zhang & Rasmussen, 2001) and content-based image retrieval (CBIR) (Gupta, Santini, & Jain, 1997; Santini & Jain, 1999; Zachary, 2000; Zachary, Iyengar & Barhen, 2001). The most popular of the similarity measures used in text retrieval are the cosine (angle)-based and distance-based measures (Zhang & Rasmussen, 2001; Zhang & Korfhage, 1999b). While the distance-based similarity measures are the most widely used measures in content-based image retrieval (CBIR) systems, the cosine (angle)-based measure has had limited applications in this area (Gupta, Santini & Jain, 1997). This is an indication that geometric/spatial models of similarity have been predominantly applied in both text and image retrieval compared to the other models. Among the geometric/spatial models of similarity, those that use the two Minkowski metrics, namely the Euclidean distance (or L_2 norm) and the city-block distance (or L_1 norm) (discussed earlier under the section "Psychological models of similarity" in this chapter), to measure similarity/dissimilarity are more popular in content-based image retrieval systems (Rubner, 1999; Stricker & Orengo, 1995; Zachary, 2000).

Most content-based image retrieval (CBIR) systems use the color feature of images as the basis for their representation and similarity matching. What is more, while similarity measures used in text retrieval mainly use term frequency and weights in computing inter-document similarity; CBIR systems rely on the color histogram. Figure 2 shows the basic model for a content-based image retrieval (CBIR) system. The image database contains feature vectors of images, which are their feature representations, extracted using appropriate feature extraction algorithms. When a user submits a query image (to a system using a "query by example" method of retrieval), its feature vector is extracted using the same algorithm as the one used to represent images in the database (Zachary, 2000). Images (or their feature vectors) in the database are evaluated for relevance to the user query through a similarity measure and results of this evaluation are produced as retrieved sets of images (Zachary, 2000).

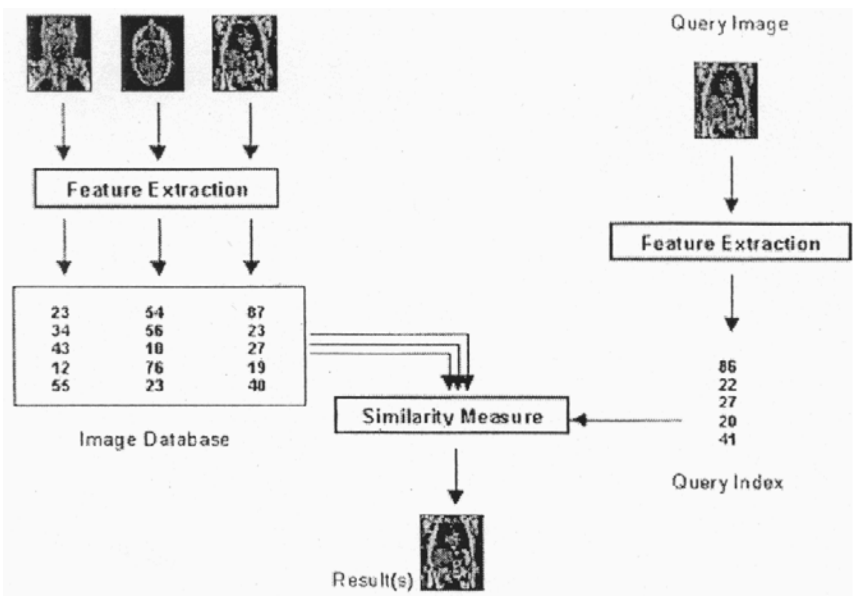


Figure 5. A model for content-based image retrieval systems (Zachary, 2000).

Lately, similarity measures based on the contrast model (e.g. Santini & Jain, 1999) and a combination of the geometric/spatial and transformational models (e.g. Rubner, 1999; Rubner, Guibas, & Tomasi, 1997; Rubner, Puzicha, Tomasi, & Buhmann, 2001; Rubner, Tomasi, & Guibas, 1998, 2000) have been proposed. However, none of these similarity measures have been based on research that tested corresponding models of similarity using humans. The similarity measure Santini and Jain (1999) proposed is the fuzzy features contrast model which uses fuzzy logic. To demonstrate how this similarity measure derived from Tversky's (1977) contrast model is applied, Santini and Jain (1999) used images of human faces as well as predicates such as "the mouth of this person is wide" for which the truth value can be determined through measurement. These measurements make up feature sets of the faces, which Santini and Jain (1999) call fuzzy sets of true predicates. The saliency of the fuzzy sets (i.e., the function f in Tversky's contrast model) is equated with their cardinality (the number of elements in the fuzzy sets).

Rubner and his colleagues (Rubner, 1999; Rubner, Guibas, & Tomasi, 1997; Rubner, Puzicha, Tomasi, & Buhmann, 2001; Rubner, Tomasi, & Guibas, 1998, 2000) developed a similarity measure which has characteristics of both geometric/spatial and transformational models of similarity, called the earth mover's distance (EMD). According to Rubner, Tomasi, & Guibas (2000), the development of this measure was instigated by a transportation problem solved through linear optimization that requires minimizing the cost for transporting goods from several suppliers to several consumers. Instead of global features such as color histograms of images, the earth mover's distance (EMD) is applied on their color or texture signatures (clusters of pixels carrying the same colors/textures with weights equal to the number of pixels in the cluster). Given two distributions (or signatures of two images) in the same space, one considered to be a mass of earth and the other a collection of holes, the earth mover's distance (EMD) is a measure of the amount of work required to transform one distribution to the second distribution (i.e., the amount of work required to fill the holes with the earth) (Rubner, Tomasi, & Guibas, 1998).

Despite the fact that there are CBIR systems that use various types of similarity measures, the majority of which are based on the geometric/spatial models of similarity, to determine inter-document similarity or similarity of documents to users' queries and provide visualizations of stored and/or retrieved image documents as results of users' queries for human browsing, there are not thorough investigations that compare human similarity judgments of images and the various similarity measures (Gupta, Santini & Jain, 1997; Zhu & Chen, 2000). In judging the relevance of a document or information source to a particular query, users inherently use their own measure of similarity. Unless retrieved sets of documents are examined by the user and judged for their relevance, there is no direct way of determining how similar two documents are either to each other or the document the user had in mind when submitting the query to the system. Therefore, there is a need for a similarity measure that predicts the relevance and similarity judgments of documents by users (see Santini & Jain, 1999). A similarity measure based on Tversky's (1977) contrast model seems to fit the criterion, provided that the model is tested and provides a good fit for a sample of images. Testing the contrast model is one of the main purposes of this study.

Features/Attributes and Representation of Images

Image features/attributes and their indexing and representation go hand in hand. As Jörgensen (1995) and Rasmussen (1997) pointed out, almost all forms of indexing and representation, in both concept-based and content-based image representation and retrieval, is based on one or more features/attributes of the images. That there is a mismatch between traditional indexing tools such as subject headings and thesauri and terms used by end users has been well documented. Results of a study that looked at the Art and Architecture Thesaurus (AAT) suggest that only about 16% of entries in the AAT could be used for the purpose of indexing a general image collection (Jörgensen, 2003).

There are various types as well as levels of image features identified in the literature of both concept-based and content-based image retrieval. Some divide image features into three categories (see Markkula et al., 2001) while some group them into two general categories, namely, low-level/syntactic/primitive and high-level/semantic (see Jörgensen , Jaimes, Benitez, & Chang, 2001). However, there is a general agreement that humans perceive all levels of features of images. According to Greisdorf and O'Connor (2002a), categories of features perceived by users of images are color (sometimes even when the image is black and white), shape, texture, objects in the image as well as "implied" by the image, location, action, and/or affect. Jörgensen and her colleagues (Jörgensen, Jaimes, Benitez, & Chang, 2001) provide a conceptual framework for categories of visual attributes in the form of a "pyramid." They divide image attributes into 10 levels and two general categories. The first category, syntax, consists of four mainly perceptual (or basic) features/attributes, while the second category, semantics, is made up of six conceptual (or higher level - semantic) features/attributes.

Several content-based image retrieval (CBIR) systems such as QBIC (Faloutsos et al, 1994; Flickner et al, 1995; Holt et al, 1997), VisualSEEk (Smith & Chang, 1996), MARS (Rui, Ortega, Huang, & Mehrotra, 1999), use the color feature of images as the basis for their representation and similarity matching (Rui, Huang, & Chang, 1999; Zachary, 2000; Zachary & Iyengar, 2001; Zachary, Iyengar, & Barhen, 2001). As a result, the color histogram, which depicts the distribution of pixels carrying the various colors in an image, has become the most useful feature representation and the most researched. The most widely used color space is the RGB (red, green, and blue, the three additive primary colors) space (Jain & Vailays, 1996; Zachary, 2000) due to the fact that the human eye “perceives color as linear combinations of [*the*] three primary colors” (Zachary, 2000, p. 19), even though it is known to not have perceptual uniformity. As a result, some researchers proposed other cost-effective types of features based on color such as image entropy (Zachary, 2000; Zachary & Iyengar, 2001; Zachary, Iyengar, & Barhen, 2001) and color signatures (Rubner, 1999), which are clusters of pixels carrying the same colors with weights equal to the number of pixels in the cluster.

Other features of images such as shape and texture are also used for indexing and representation (Faloutsos, 1994; Flickner et al., 1995; Jain & Vailaya, 1996; Rubner, 1999; Rui, Huang, & Chang, 1999; Zachary, 2000). While the shape feature of images is useful for automatic object recognition (Gudivada & Raghavan, 1997), the texture feature is applicable to pattern recognition and computer vision (Rui, Huang, & Chang, 1999).

The color histogram of an image denotes “the joint probability of the intensities of the three color channels [*Red, Green, & Blue*]” (Rui, Huang, & Chang, 1999, p. 41) and given an image I, its histogram, H(I), is a vector with elements h_{c1}, h_{c2}, h_{c3}, …, h_{cn}, where h_{cj} is the number of pixels carrying color c_j (Stricker & Orengo, 1995). A color image with the three channels (R, G, B) may have three different histograms, one for each channel. Figure 6 shows an image with its color histograms, extracted using CVIPtools (Umbaugh, 1998), for the red, green, and blue channels, respectively.

Content-based feature representation of images, especially those based on the color histogram, are global in nature. In other words, they represent the entire image. However, efforts are also being made to represent images using local features such as individual objects, parts or regions of images. An example of a system that uses such approach is Blobworld, where images are automatically segmented into regions called "blobs" and their color and texture features extracted (Carson et al, 1999).

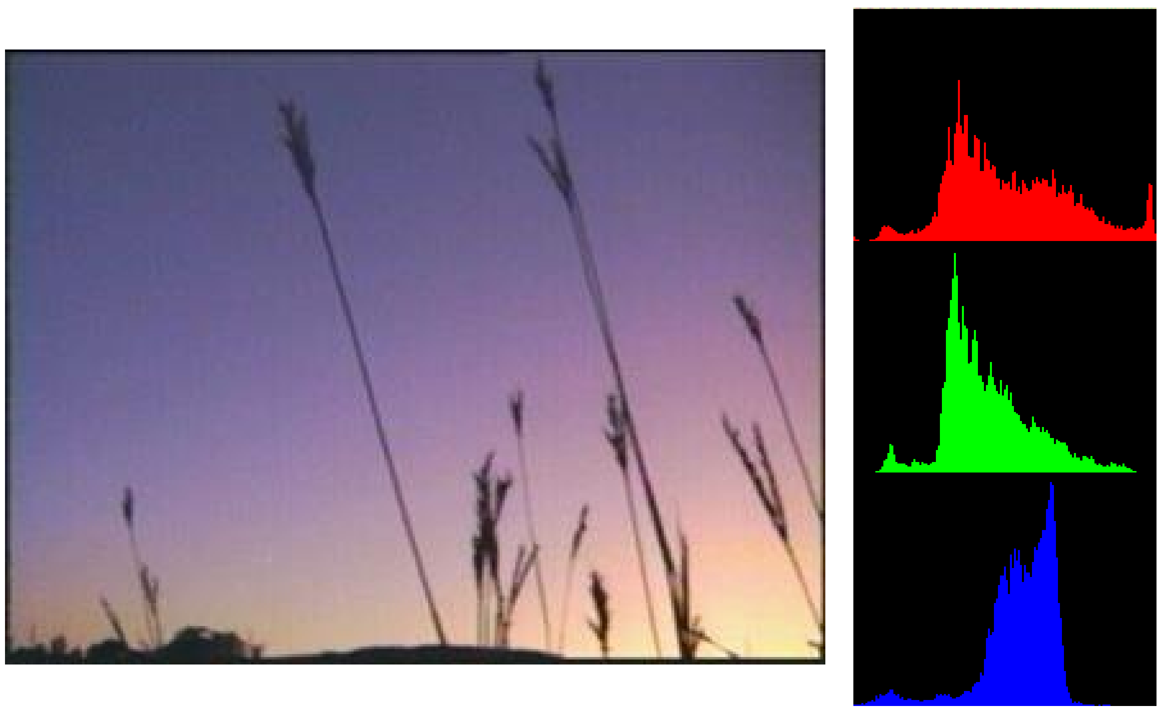


Figure 6. An image with its RGB color histograms - (O'Connor & Wyatt, 2004, used with permission).

So far, the discussion has focused on image features/attributes and which types of features are used to represent images by the various types of content-based image retrieval (CBIR) systems. One of the major problems in information storage, organization, access and retrieval is related to representation of the documents. Blair (1990) argues that representation is "primarily a problem of language and meaning" (p. 1). It is even more so in the case of concept-based image representation where the documents being represented are visual images, rather than structured text documents, and the task of representation and retrieval involves feature description and matching text descriptions of images with users' query terms. O'Connor, O'Connor, and Abbas (1999) suggest that representation of images using terms generated through users' reactions might address this problem. However, the cost of indexing images by human indexers, let alone using user generated terms, especially in large image collections, is prohibitive. In an ideal situation, automatic extraction and indexing of images based on low-level/primitive as well as high-level/semantic features would be possible. However, most of the current content-based image retrieval (CBIR) methods could only enable the extraction of low-level/primitive features such as color, shape, texture, etc. Even though content-based image retrieval (CBIR) mechanisms offer cost effective alternatives, the computational cost of some of these mechanisms could also be high (Zachary & Iyengar, 2001). Given the fact that there is still continued dependence on concept-based image indexing and retrieval, mostly with the help of human indexers (Enser, 2000), the need for CBIR systems capable of extracting high-level/semantic features has increased (Stan & Sethi, 2001). A solution, suggested by researchers, that could improve image representation and retrieval is the use of a combination of both concept and content-based methods (Djeraba, Bouet, Briand, & Khenchaf, 2000; Enser, 2000).

The amount of information contained in an image or the meaning it conveys to different viewers depends on several factors and it is difficult to extract or measure, hence the difficulty in representing and indexing images for effective and efficient retrieval. If a picture is worth a thousand words to a single user, then "it is worth a million words" to a thousand of them (Greisdorf & O'Connor, 2002a, p. 7). This is not to say that words are not suitable for indexing images. Actually, textual descriptions of image attributes are still popular methods of indexing (Enser, 2000; Jörgensen, 1995, 1996, 1998, 2003). Besides, most image users formulate their queries using words though this is neither always the necessary nor the only way (O'Connor, O'Connor, & Abbas, 1999). Therefore, there is a need for more investigation into what features and categories of features are generally perceived by image users and into ways to incorporate them into the traditional tools for image indexing and representation.

A handful of investigators have looked into this problem through user reactions to images and image description tasks (Greisdorf & O'Connor, 2002a; Greisdorf & O'Connor, 2002b; Jörgensen, 1995, 1996, 1998; O'Connor, O'Connor, & Abbas, 1999; Turner, 1994, 1995) as well as through solicitation of queries from image users (Chen, 2001a, 2001b; Choi & Rasmussen, 2002, 2003; Goodrum & Spink, 2001; Hastings, 1995). Subject analysis of images and image indexing has also been addressed by a few authors (Layne, 1994; Mostafa, 1994; Rasmussen, 1997; Shatford, 1986). The meaning and interpretation of visual arts was the subject of investigation by Panofsky (1955) long before these researchers attempted to address the issue of perceived features/attributes, subject analysis, and indexing of images. Jörgensen (2003) asserts that Panofsky's work on the meaning of visual arts "forms the basis for much of the theoretical work that has been done on the classification of art images" (p. 117). Rosch and others (Rosch, 1975;

Rosch, & Mervis, 1975; Rosch et al., 1976) have also studied which categories of attributes of objects carry more information.

Panofsky (1955) classified attributes of visual arts into three levels, namely preiconographical, iconographical, and iconological. Preiconographical attributes are basic or natural characteristics, usually names of objects; iconographical level attributes have meaning attached to them as a result of interpretation; while iconological attributes involve deeper syntheses and multiple interpretations (Jörgensen, 2003).

One researcher who studied the nature of perceived features of images extensively is Jörgensen (1995, 1996, 1998, 2003). In her dissertation research, which required participants to do a description, a sorting, and a searching task, Jörgensen (1995) identified three general levels/categories of image attributes. They are perceptual (P), interpretive (I), and, to a lesser extent, reactive (R) attributes. According to Jörgensen (1995), perceptual attributes relate to physical characteristics of images including objects in the images, image color, and other visual elements; whereas interpretive attributes are in the eyes of the viewer and require more than just perceiving.

Results from Jörgensen's (1995) three tasks performed by participants produced 12 categories (or classes) of attributes, namely objects, people, color, story, spatial location, description, visual elements, art historical information, people-related, external relation, viewer response, and abstract. Generally, this order of categories of attributes represents their frequency of assignment (i.e., "objects" attributes were assigned most, while "abstract" attributes were assigned least) (Jörgensen, 1995).

In summary, content-based image representation and retrieval systems enable only the extraction and representation of primitive/basic/low-level features of images, while concept-based image representation tools include an insignificant number of terms used by image users. However, concept-based methods remain the dominant image representation mechanisms.

Summary

This chapter presents a detailed and critical review of the literature relevant to the topic of the study, which is the nature of perceived features and similarity of images. The study was conducted within the context of image representation and retrieval using Tversky's (1977) contrast model as a conceptual framework. The literature review is organized under four sections. The first section deals with the four psychological models of similarity, namely the geometric/spatial, set-theoretical, alignment-based, and transformational models together with their uses and applications. The second section is a detailed presentation of past research that either used or tested Tversky's (1977) contrast model, one of the set-theoretical models. The third section summarizes literature on similarity measures used by content-based image retrieval (CBIR) systems to determine inter-document similarity and/or the relevance of documents to a user's query, while the last section presents the various types and categories of features/attributes of images used for their indexing and representation.

Weaknesses in the most widely used models of similarity, the geometric/spatial models, led to the formulation and testing of Tversky's (1977) contrast model which seems to explain human similarity judgments better. While content-based image retrieval (CBIR) systems can only extract basic/primitive/low-level features of images and their similarity measures are based on geometric models of similarity, the reviewed literature supports an investigation into alternative mechanisms for representation and similarity matching of images. Therefore, this study attempts to fill the gap in the literature and assist in the efforts to develop such mechanisms. The following chapter, in the light of the literature review, presents detailed descriptions of data collection and analysis procedures and methods.

CHAPTER 3

MATERIALS AND METHODS

Introduction

The main goal of this chapter is to present detailed description of the materials, participants, and the various instruments and methods used for data collection and analyses, including statistical techniques and procedures used to test the contrast model, in order to fulfill the purposes of the study. Even though Tversky's contrast model was used as a framework to guide this study, the study is mainly exploratory in nature. Both qualitative (content analysis) and quantitative (correlational) methods were used to seek answers to the four research questions concerning the nature of perceived features and similarity of images (see chapter 1).

Two web-based forms for two different tasks were used to collect data for the study. The first form is for an image description task where participants describe (list features/attributes of) images and the second is for a similarity judgment task where they judge the degree of perceived similarity of pairs of images on a ratio scale using magnitude estimation (Stevens, 1956, 1966, 1975).

Materials

In the absence of a standard test collection of images, a set of color photographs taken by a number of photographers and published on a CD with a book by O'Connor & Wyatt (2004) served as the population of images. A simple random sample of 30 color digital images (Appendix A) was selected from this collection. Each image was 375X250 pixels in dimension. In order to ensure heterogeneity and sufficient variations of features of the sample of images selected randomly from the collection, 15 volunteer participants were asked to describe them, that is list features of the images. Feature data (list of features) for all the 30 images were subjected to analysis of spread (variability) and the Index of Qualitative Variation (IQV), a measure of qualitative variability (Weisberg, 1992) was used. The IQV measure, calculated using the formula:

$IQV = \dfrac{(1 - p_1^2 - p_2^2 - ... - p_k^2)}{(k-1)/k}$, where p_i (for i=1, 2, 3, …, k), is the proportion of distinct features (features not shared by others) attributed to image i and k is the number of images in the sample, that is, 30. The calculated value of the IQV is 0.99965, a value close to 1, which signifies greater diversity among the sample of 30 images.

Participants

The use of humans as judges of similarities between images for research purposes (Mojsilović et al., 2000; Mojsilović & Rogowitz, 2001; Rogowitz et al., 1998) and as participants in stimulus description tasks involving perceptual/visual stimuli such as images (Gati & Tversky, 1984; Jörgensen, 1995; Tversky, 1977) as well as conceptual/semantic stimuli (Dopkins & Ngo, 2001; Gati & Tversky, 1984; Johnson, 1986; Tversky, 1977) is common. Participants in this study were asked to voluntarily serve as similarity judges of images, through paired comparison (where the degree of similarity of images is rated in pairs), and asked to provide descriptions of images in terms of their features/attributes. The convenience sampling method was adopted for selecting participants. The sample consisted of 150 graduate students at the School of Library and Information Sciences, University of North Texas (population size: 1100).

They were all beginning masters students taking two of the three core courses at the School. Beginning students were chosen in order to minimize the possibility of them providing subject-headings-like terms, which students of LIS who are advanced in their programs tend to do. Convenience sampling can be used for testing models provided that the "model is correctly specified" (Schonlau, Fricker, & Elliott, 2001, p. 34). The model being tested in this study, Tversky's (1977) contrast model, has a strong and well-documented theoretical foundation and well-defined structural relations among the constructs (perceived features and similarity).

The sample size of 150 participants is well over the recommended minimum number of participants for correlational studies (30) (Gay & Diehl, 1992) and it is a sufficient number for structural equation modeling (Schumacker & Lomax, 1996). Participants were randomly assigned to the two tasks with half of them (75) providing similarity judgments and the remaining half (75) taking part in the image description task. A similar procedure was used by Dopkins and Ngo (2001) and Tversky (1977) and is assumed to produce feature listing and similarity judgment data that is free of interactive influences (Dopkins & Ngo, 2001).

Data Collection

Data collection for the study is based on some assumptions related to scaling (measurement) of the degree of perceived similarity of images as well as the ability of participants to describe/list features of images and judge the similarity of pairs of images. These assumptions are:

1. Participants are not only able to perceive and identify features/attributes of images but they are also able to describe/list

them. In other words, “it is possible to elicit from subjects detailed features of stimuli” (Tversky, 1977, p. 339).

2. The degree of perceived similarity of images can be scaled on a ratio level of measurement using magnitude estimation procedures (Stevens, 1956, 1966, 1975). Magnitude estimation has produced useful results in the study of human perception both within and outside psychophysics-a field that investigates the quantitative relations between physical and perceptual magnitudes–and the investigation of perceived similarity of perceptual stimuli such as images (Stevens, 1966; Nunnally & Bernstein, 1994). Magnitude estimation has also been used by information science researchers with encouraging results for scaling users’ perception of relevance (Bruce, 1994; Eisenberg, 1986, 1988) and satisfaction with their information seeking on the Internet (Bruce, 1998).
3. Participants are able to judge the degree of perceived similarity of pairs of images, through magnitude estimation, and produce ratio level data. Research has already shown that participants “can make consistent quantitative appraisals of their subjective experiences” (Stevens, 1956, p.5).

Image Description Task

The main purpose of this task was to solicit lists of features/attributes of the 30 images from participants. Half of the 150 participants, randomly assigned to complete the image description task, were involved in the description (listing of features/attributes) of the 30 images. An email message was sent to each of the students taking two required courses at the School of Library and Information Sciences, University of North Texas, during the Summer 2004 and Fall 2004 semesters. Students were randomly assigned to this task. In the email message, the URL (web address) for the image description task and a unique identifier (a random number assigned by the researcher to identify the participants for authentication purposes) was sent to each participant. A follow-up email message was sent to participants who did not complete the task after two weeks from the date the first email message was sent.

When the participants visited the website for the task, a web-based form (Appendix C) displayed each image individually. The images were presented randomly; that is, no two participants saw the 30 images in the same order. The first page displays instructions for participants on how to complete the task (Appendix B). To ensure internal validity (Ericsson & Simon, 1980), a time constraint was enforced and each of the 30 images was displayed for a maximum of 90 seconds (one and a half minutes) and participants were asked to type as many features/attributes as possible to describe the particular image. Two images (not in the sample) were included (without the participants' knowledge) at the beginning of the task to familiarize participants with the image description task. Demographic data on participants (e.g., gender, age, educational level/background, school/faculty/department) were also collected.

Similarity Judgment Task

The purpose of the similarity judgment task was to obtain human similarity judgment data from participants for the sample of 30 images. The traditional paired comparison design (where the degree of perceived similarity of images is rated in pairs) of the 30 images was used to collect similarity judgment data. The other half of the 150 participants, once again randomly assigned to this task, completed the similarity judgment task. An email message, similar to the one sent to participants of the image description task, was sent to them as well. A follow-up message was also sent two weeks after the first one. After reading the instructions for this task (Appendix D), participants were presented with a web-based form (Appendix E) for each of two sets of pairs of the 30 images [30(30-1)/2=435 pairs in each set] and were asked to judge the degree of perceived similarity of pairs of images on a ratio scale using magnitude estimation (Stevens, 1956, 1966, 1975).

There are two designs/forms of tasks involving magnitude estimation: with and without a standard stimulus or a modulus (Stevens, 1975; Engen, 1971). In a task involving magnitude estimation with a standard stimulus or a modulus, participants are presented with the standard stimulus or modulus together with the magnitude estimation of the modulus, usually an integer multiple of 10, predetermined by the researcher. Participants are then asked to estimate the magnitude, relative to the magnitude of the modulus, of a series of stimuli that vary in intensity of the attribute/continuum being scaled/measured. For instance, if the magnitude of the modulus was given as 20 and the participant judges the magnitude of another stimulus to be twice that of the modulus, he/she will assign the number 40 to this particular stimulus. Similarly, if he/she judges the magnitude of the presented stimulus to be a quarter of the magnitude of the modulus, then he/she will assign the number 5.

The second form of magnitude estimation uses no standard stimulus or modulus. Participants are presented with the stimuli in random orders and they are asked to assign numbers to each stimulus, proportional to their perceived intensity of the attribute/continuum being scaled. A variation of the second form of magnitude estimation instructs participants to mark a point on a horizontal line of a certain length so that the distance between the beginning of the line and the marked point is equal to the magnitude of the intensity of the attribute/continuum being scaled.

The first method of magnitude estimation (with a modulus) was the first to be used in psychometric scaling. However, it was later found that the choice of a standard stimulus by the researcher interferes with the freedom of participants to produce their own magnitude estimations and that the second method (with no modulus) is preferred as it facilitates unconstrained judgments (Stevens, 1975). Therefore, in this study, a variation of the second method of magnitude estimation (with no modulus) was used where participants used a horizontal line (five inches long and a fifth of an inch thick) to indicate the degree of their perceived similarity of pairs of images.

The perceived similarity of the first set of 435 pairs of the 30 images was judged by all the 75 participants of the similarity judgment task first and the second set of 435 pairs (obtained by permuting the positions of the images in the first set of pairs, i.e., a pair *a, b* in the first set would be presented as *b, a* in the second and reversing the order of pairs in the first set) was judged by the same participants after a mandatory five-minute break in order to minimize the fatigue effect due to the large number of pairs of images they had to judge. Pairs of images were presented in the same order for all the participants. The order of pairs of images in both sets of 435 pairs as well as the order of presentation of images in a pair were determined using an optimum order and presentation method for paired comparisons suggested by Ross (1934). Ross (1934) argues that the method eliminates errors due to time and space and that other methods that use random orders are open to such errors. Because the sample size of images for this study is an even number (30), n=31 was used to produce the pairs as suggested by Ross (1934). Pairs involving image number 31 were dropped.

As a familiarization and calibration exercise in magnitude estimation, participants were presented with five lines of varying lengths (two to eight inches) and asked to judge their apparent length. Three practice pairs of images (not included in the sample) were also presented at the beginning of the similarity judgment task (participants were not aware of this fact) in order to familiarize participants with the similarity judgment task using magnitude estimation. Demographic data on participants (e.g. gender, age, educational level/background, school/faculty/department, etc.) were also collected.

Data Analysis

Image Description and Similarity Judgment Tasks

Content analysis was used to analyze data from the image description task. Content analysis data constituted sets of features of the sample of images. Data derived from sets of features were in turn used for determining the relationship between perceived similarity of images and their common and unique/distinctive features and to determine the relative weights given to common and unique/distinctive features in similarity judgments, as well as for testing Tversky's contrast model.

One of the methods of data analyses for verbal/written tasks such as the image description task is content analysis through the creation and testing of a coding scheme that involves the definition of recording units & categories, assessment of the accuracy of coding, revision of coding rules, and coding the entire text (Weber, 1990). For the purpose of this study, the “word” and “word sense” recording units (basic unit of text to be classified or categorized) were used to represent individual features of images and make up sets of features of images. According to Weber (1990), word sense could be a phrase constituting a semantic unit. It includes idioms such as “taken for granted” as well as proper nouns like “the Empire State Building” (Weber, 1990, p. 22).

A dictionary based on the recording units (words or word senses representing features of images) was created for the purpose of coding the list of features supplied by participants for each image into categories whereby each recording unit is assigned to a category where it shares a similar meaning with units already assigned to the category. Weber (1990) recommends categories of synonyms or categories of “words sharing similar connotations” (p. 12). As is customary in content analysis, intercoder agreement was used to measure reliability (the consistency of the coding scheme, more specifically the assignment of features to the categories in the dictionary). Two popular measures, percent agreement and Cohen’s (1960) Kappa, were calculated between the researcher and each of two other coders, using the list of features supplied for all 30 images by all participants of the image description task. Once an acceptable level of the intercoder agreement was achieved in constructing the dictionary, the computer software TEXTPACK (Mohler & Zuell, 1998) was used to determine the frequencies for categories of features (that is, the number of features assigned by all participants) for each of the 30 images.

Validity (how well an instrument/scale measures what it was meant to measure) is another major issue in content analysis. Even though reliability is not a sufficient indicator of validity, it is one of the necessary conditions. Content (face) validity is sufficient for most content analyses and it requires results of the content analysis to be consistent with characteristics of the objects under study (Holsti, 1969). Evaluating the descriptions provided by the participants for over-elaborations and long stories ensured the content validity and usefulness of the image description task and content analysis data derived from their descriptions. Construct validity, a measure of the agreement between a theoretical construct and a procedure used to measure the construct, was also determined for the three constructs measured by using data from the image description task (i.e., the common and distinctive features). This was achieved through testing measurement models for the common and distinctive features, that is, the three independent latent variables/constructs – common features (*commnfet),* distinctive features of image a (in a pair of images a and b) *(distfeta),* distinctive features of image b *(distfetb)* (see Figure 7), using structural equation modeling (SEM).

The paired comparison design of the similarity judgment task is a commonly used method of collection of human similarity judgment data (Dunn-Rankin, Knezek, Wallace, & Zhang, 2004; Ross, 1934). The appropriate measure of reliability for the similarity judgment task is Cronbach's (1951) Coefficient Alpha (α) that measures how well two or more variables/items/scales measure the dependent latent construct (i.e., the perceived similarity of images–*sim* in Figure 7). Construct validity for the similarity (*sim*) construct in the contrast model, or how well the two observed dependent (or Y) variables, *SIMAB & SIMBA*, measure the dependent latent variable/construct similarity (*sim*), was also determined by testing the appropriate measurement model in Figure 7.

Linear correlation techniques were employed to find out the nature of the relationships between perceived similarity (the dependent/criterion variable) and common and unique/distinctive features (the independent/predictor variables) of images, using the observed x (independent) and y (dependent) variables. Structural equation modeling (SEM) (Schumacker & Lomax, 1996) was used to fit the contrast model to similarity judgment and image description (feature listing) data. Linear regression and structural equation modeling were used to determine the relative weights (θ, α, and β in the contrast model $S(a,b)= \theta f(A \cap B) - \alpha f(A-B) - \beta f(B-A)$) of common and unique/distinctive features in similarity judgment. The unit of analysis for all variables in the model is pairs of images and for the sample of 30 images, the total number of cases (N) is 435(which is 30(30-1)/2).

Structural Equation Modeling (SEM)

Structural equation modeling (SEM) is a set of statistical techniques that include path analysis (to study direct/indirect effects); factor analysis (to study measurement models, or how well a set of observed variables measure latent variables/constructs); regression analysis (to study prediction and amount of variance explained); and structure (covariance structure) analysis (to study relationships between latent variables) (Schumacker & Lomax, 1996). It provides a single comprehensive means for data analysis, especially testing complex theoretical models. Structural equation modeling (SEM) is appropriate for studies that investigate research questions related to the relationships between latent variables/constructs and when researchers are seeking to test theoretical model fit to sample data. All four research questions in this study deal with either relationships between latent constructs or the testing of a theoretical model, hence the choice of structural equation modeling as an appropriate set of statistical techniques. In general, a structural equation model has two parts: the measurement model and the structural model (Schumacker & Lomax, 1996). The common and conventional names, notations and symbols for the various types of variables, path coefficients, error terms (measurement & prediction errors) are summarized in Table 1 (Jöreskog & Sörbom, 1993).

Table 1

Conventional Symbols, Names of Variables and Coefficients in Structural Equation Modeling

Symbol	Name	Variable, path or coefficient it stands for
ξ	Ksi	Exogenous (independent) latent variable
η	Eta	Endogenous (dependent) latent variable
γ	Gamma	Path coefficients for a path connecting an exogenous latent variable (ξ) to an endogenous latent variable (η)
β	Beta	Path coefficients for a path connecting an endogenous latent variable (η_1) to another endogenous latent variable (η_2)
Y	Y-variable	Observed variables which depend on the endogenous (dependent) latent variables (η)
X	X-variable	Observed variables which depend on the exogenous (independent) latent variables (ξ)
$\lambda^{(y)}$	Lambda-Y	Path from an endogenous (dependent) latent variable (η) to a Y-variable
$\lambda^{(x)}$	Lambda-X	Path from an exogenous (independent) latent variable (ξ) to an X-variable
ζ	Zeta	Error terms in the structural equations
ε	Epsilon	Measurement errors in the observed Y-variables
δ	Delta	Measurement errors in the observed X-variables

The path diagram in Figure 7 is a graphical representation of Tversky's contrast model [$S(a,b)= \theta f(A\cap B) - \alpha f(A-B) - \beta f(B-A)$] tested in this study with a sample of 30 images and 150 participants where the endogenous (dependent) latent variable, sim (η), is the degree of perceived similarity of images a and b [$S(a,b)$], commnfet (ξ_1) is one of the exogenous (independent) latent variables and is a measure of the common features of a and b [$f(A\cap B)$], distfeta (ξ_2) is the second exogenous (independent) latent variable and it is a measure of the unique/distinctive features of a when compared to b [$f(A-B)$], and distfetb (ξ_3) is the third exogenous (independent) latent variable and is a measure of the unique/distinctive features of b when compared to a [f(B-A)]. Table 2 is a summary of the variables, coefficients, and parameters in Figure 7 as well as how each observed variable in the model is measured.

Use of more than one measure of the latent variables ensures construct validity and minimizes measurement error. The endogenous (dependent) latent variable (perceived similarity or sim) has two measures, SIMAB & SIMBA, operationalized using the similarity judgment task described earlier. SIMAB is similarity judgments data (obtained using magnitude estimation) for the first set of 435 pairs of the 30 images while SIMBA is similarity judgments data for the second set. Like any data obtained using magnitude estimation, values of the two variables, SIMAB & SIMBA, for the 435 cases (pairs of images) were determined by taking the logarithms of the raw magnitude estimations provided by all participants of the image description task and then aggregated using their geometric means. The three independent latent variables (common features, unique/distinctive features of a, and unique/distinctive features of b) have three measures each and these measures are three different methods of operationalization (outlined below) of the function f in the contrast model, some of which (methods 1 & 2) take into account the number of participants who assign a particular feature to a stimulus.

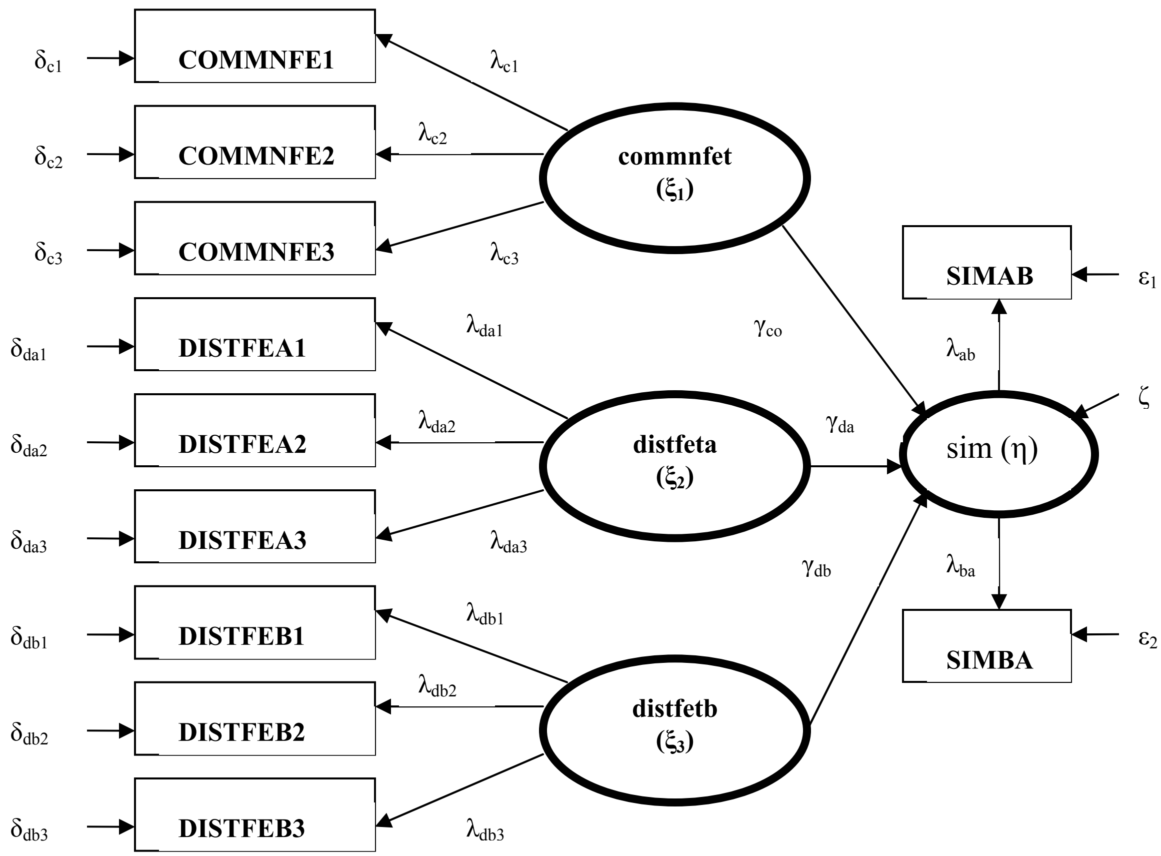


Figure 7. Path diagram for Tversky's contrast model.

Table 2

Variables, Coefficients, and Parameters in Tversky's Contrast Model (Figure 7)

Symbol	Variable, path or coefficient it stands for
ξ_1	Common features (*commnfet*)–exogenous latent variable(1)
ξ_2	Distinctive features of a (*distfeta*) – exogenous latent variable (2)
ξ_3	Distinctive features of b (*distfetb*) – exogenous latent variable (3)
η	Perceived similarity (*sim*) - endogenous latent variable
γ_{co}	Path coefficient for the path from ξ_1 to η
γ_{da}	Path coefficient for the path from ξ_2 to η
γ_{db}	Path coefficient for the path from ξ_3 to η
SIMAB, SIMBA	Observed Y-variables (of η) – similarity judgments for image sets 1 & 2, respectively
COMMNFE1, COMMNFE2, COMMNFE3	Observed X-variables (of ξ_1) – measures of common features using methods 1, 2, & 3, respectively

(*Table continues*)

Table 2 (*Continued*)

Variables, Coefficients, and Parameters in Tversky's Contrast Model (Figure 7)

Symbol	Variable, path or coefficient it stands for
DISTFEA1, DISTFEA2, DISTFEA3	Observed X-variables (of ξ_2) – measures of distinctive features of a using methods 1, 2, & 3, respectively
DISTFEB1, DISTFEB2, DISTFEB3	Observed X-variables (of ξ_3) – measures of distinctive features of b using methods 1, 2, & 3, respectively
λ_{ab}, λ_{ba}	Path from η to SIMAB and SIMBA, respectively
λ_{c1}, λ_{c2}, λ_{c3}	Path from ξ_1 to COMMNFE1, COMMNFE2, & COMMNFE3, respectively
λ_{da1}, λ_{da2}, λ_{da3}	Path from ξ_2 to DISTFEA1, DISTFEA2, & DISTFEA3, respectively
λ_{db1}, λ_{db2}, λ_{db3}	Path from ξ_3 to DISTFEB1, DISTFEB2, & DISTFEB3, respectively
ζ	Error term in the structural equation
ε_1, ε_2	Measurement errors in SIMAB & SIMBA, respectively
δ_{c1}, $\delta_{c2,}$ δ_{c3}	Measurement errors in COMMNFE1, COMMNFE2, & COMMNFE3, respectively
δ_{da1}, $\delta_{da2,}$ δ_{da3}	Measurement errors in DISTFEA1, DISTFEA2, & DISTFEA3, respectively
δ_{db1}, $\delta_{db2,}$ δ_{db3}	Measurement errors in DISTFEB1, DISTFEB2, & DISTFEB3, respectively

The three methods of operationalization of the function f in the contrast model are discussed below.

Method 1: The first method of measurement of common and unique/distinctive features was used by Tversky (1977) and it requires the number of participants who attribute a specific common or unique/distinctive feature to be determined. Let X_i be the proportion of participants who assigned feature X to image i and N_x be the total number of images that feature X was assigned to. According to this method (*method 1*), the measure of the common features of images a and b is computed using:

$f(A \cap B) = \sum X_a X_b \big/ N_x$, for all X in A∩B. Similarly, let Y_i and Z_j be the proportion of participants who assigned unique/distinctive features Y and Z to images i and j, respectively. The measures of the unique/distinctive features of images a and b are $f(A-B)=\sum Y_a$ and $f(B-A)=\sum Z_b$, respectively, for all Y in A-B and all Z in B-A.

Method 2: The second method of measurement of common and unique/distinctive features was developed by Johnson (1986). Let C_i be the number of common features attributed to images a and b by participant i (i=1, 2, 3, …, n) and n be the total number of participants who assigned at least one common feature to a and b. The measure of the common features of images a and b is the mean number of common features and is given by $f(A\cap B)=\sum C_i \big/ n$. To compute the measure of the unique/distinctive features of a and b, let D_i and D_j be the number of unique/distinctive features assigned to images a and b, respectively, by participants i (i=1, 2, 3, …, n) and j (j=1, 2, 3, …, m) and n and m be the total number of participants who assigned at least one unique/distinctive feature to a and b, respectively. The measures of the unique/distinctive features of images a and b are $f(A-B)=\sum D_i \big/ n$ and $f(B-A)=\sum D_j \big/ m$, respectively.

Method 3: The third method of measurement of common and unique/distinctive features was used by Tversky (1977) as well. This measure "assigns equal weight to all features regardless of their frequency of mention" (Tversky, 1977, p. 338). It is determined by simply counting the number of common and unique/distinctive features assigned by participants. Let C_i be the number of common features attributed to images a and b by participant i (i=1, 2, 3, …, n) and n be the total number of participants who assigned at least one common feature to a and b. The measure of the common features of images a and b is the sum of the number of common features assigned by all the participants and is given by $f(A \cap B) = \sum C_i$. To determine the measure of the unique/distinctive features of a and b, let D_i and D_j be the number of unique/distinctive features assigned to images a and b, respectively, by participants i (i=1, 2, 3, …, n) and j (j=1, 2, 3, …, m) and n and m be the total number of participants who assigned at least one unique/distinctive feature to a and b, respectively. The measures of the unique/distinctive features of images a and b are $f(A - B) = \sum D_i$ and $f(B - A) = \sum D_j$, respectively.

Summary

This chapter presents the nature of the population and selection of samples of participants and materials (images) for the study. The various data collection methods, namely the image description and similarity judgment tasks, methods of operationalizations of the measured variables including the three methods (1, 2, and 3) for measuring the common and distinctive features of pairs of images, as well as data analysis methods such as content analysis, correlation analysis and regression analysis, and structural equation modeling are discussed. A path diagram depicting Tversky's (1977) contrast model is presented and issues of validity and reliability of the data collection instruments, scales, and measured variables are addressed.

CHAPTER 4

ANALYSIS OF DATA, RESEARCH FINDINGS, AND DISCUSSION

Introduction

The main purpose of this study is to investigate the nature of perceived features (common and unique/distinctive) and similarity of images, including their measurements and relationships, by way of four research questions (see chapter 1) using Tversky's (1977) contrast model as a framework. Data were collected through image description and similarity judgment tasks performed by 150 participants using 30 images (435 pairs). Qualitative methods (content analysis) and quantitative methods (correlation & regression analyses, structural equation modeling) of data analysis were employed to seek answers to the four research questions. Structural equation modeling including the testing of the measurement models for all four latent variables (common features, distinctive features of a & b, where a & b form a pair of images, and similarity) as well as the testing of the contrast model was done using LISREL© 8.54 software (Jöreskog & Sörbom, 1993).

The chapter also presents a summary of the demographic data on participants, results of the analysis of data from the two tasks (image description and similarity judgment tasks), and a detailed discussion of the results vis-à-vis the four research questions as well as findings of similar research. A summary of the results and research findings is included at the end of the chapter.

Description of Participants

A total of 150 participants took part in two tasks: an image description task (75 participants randomly assigned to this task) and a similarity judgment task (the other 75 participants randomly assigned to the task). Table 3 shows the distribution of participants of both tasks by gender, age, major, highest degree completed, programs of study, whether they had a degree or background in art, and their frequency of use of images. There were more female participants (77%) than male participants, a proportion comparable to that of the total number of female and male students in the school from which the sample of participants was drawn, an indication of the representativeness of the sample, in terms of gender. Approximately three quarters of the participants are between 26 and 50 years old and over 95% of them major in either library or information science. Only about a quarter of the participants indicated masters and Ph.D. as their highest degrees completed at the time of participation, while almost all (97%) are enrolled in the various masters program tracks at the school. Few participants (13%) have either a degree or some background in art, while most of them (88%) either use or work with images at least once a month.

Table 3

Demographic Data on Participants of the Study

Demographics	Image Description Task		Similarity Judgment Task		Total	
	Count	%[a]	Count	%[a]	Count	%[b]
Gender						
M	19	25.33	15	20.00	34	22.67
F	56	74.67	60	80.00	116	77.33
Age						
<21	0	0.00	1	1.33	1	0.67
21-25	12	16.00	9	12.00	21	14.00
26-30	12	16.00	10	13.33	22	14.67
31-35	14	18.66	9	12.00	23	15.33
36-40	11	14.67	10	13.33	21	14.00
41-45	12	16.00	12	16.00	24	16.00
46-50	6	8.00	14	18.67	20	13.33
51-55	6	8.00	7	9.33	13	8.67
56-60	2	2.67	2	2.67	4	2.67
>60	0	0.00	1	1.34	1	0.66
Major						
Journalism	1	1.34	0	0.00	1	0.67
Info. Science	4	5.33	9	12.00	13	8.66
Lib. Science	66	88.00	64	85.34	130	86.67
Info. Systems	0	0.00	1	1.33	1	0.67
Other	4	5.33	1	1.33	5	3.33
Highest degree completed						
Bachelors	57	76.00	51	68.00	108	72.00
Masters	14	18.66	21	28.00	35	23.33
PhD	2	2.67	1	1.33	3	2.00
Other	2	2.67	2	2.67	4	2.67
Program of study						
Masters	73	97.32	72	96.00	145	96.67
PhD	1	1.34	1	1.33	2	1.33
Other	1	1.34	2	2.67	3	2.00
Background/Degree in art						
Yes	9	12.00	11	14.67	20	13.33
No	66	88.00	64	85.33	130	86.67
Frequency of use of images						0.00
Daily	16	21.33	20	26.66	36	24.00
3 times a week	9	12.00	13	17.33	22	14.67
Twice a week	7	9.33	5	6.67	12	8.00
Once a week	10	13.34	8	10.67	18	12.00
once every two weeks	6	8.00	8	10.67	14	9.33
Once a month	18	24.00	12	16.00	30	20.00
Never	9	12.00	9	12.00	18	12.00

Note. [a]n=75. [b]n=150

Analysis of Image Description Task Data

As described in chapter 3, content analysis was used to build a dictionary of all feature-bearing words and word senses supplied by participants of this task. Once the dictionary, with 196 mutually exclusive categories, was built by the researcher and three others through consensus, two more coders were asked to assign features to the categories. The two coders examined a random set of features (23% of the total number of features) and assigned them to 39 randomly selected categories (20%). The coders were told to leave the features that did not fit into any of the categories unassigned. Instructions for coders and a set of features assigned to a sample feature category are in Appendix F. The inter-coder agreement (percent agreement) was calculated using the general formula: $PA_O = A/n$, where PA_O is the " 'proportion agreement, observed,' A is the number of agreements between two coders, and n is the total number of units the two coders have coded" (Neuendorf, 2002, p. 149). The computed values of the percent agreement between the researcher and the two coders were .90 and .92. This is quite a large value, considering that reliability coefficients, including percent agreement values, over 0.80 are considered to be "acceptable in most situations" (Neuendorf, 2002, p. 143).

Another measure of inter-coder reliability is Cohen's (1960) Kappa computed using the formula: $Cohen's\quad Kappa = \frac{PA_O - PA_E}{1 - PA_E}$, where PA_O is the " 'proportion agreement, observed,' and PA_E stands for 'proportion agreement, expected by chance,' " (Neuendorf, 2002, p. 143). PA_E is computed using the formula: $PA_E = \left(1/n^2\right)\left(\sum pm_i\right)$, where n is the number of units coded in common by coders and pm_i is the product of marginals for category i. The computed values of Cohen's Kappa for the researcher and the two coders were .89 and .91. Once again, this is an acceptable value and shows the reliability of the categories formed and the dictionary built using the categories.

A single file containing all features assigned by 75 participants (15301 total features), together with the dictionary, was submitted to the TEXTPACK (Mohler & Zeull, 1998) text analysis software. The minimum and maximum number of features assigned to a single image were 352 and 652, respectively. Table 4 depicts the distribution of features assigned to the 30 images by 75 participants of this task.

Table 4

Distribution of Features Assigned to 30 Images

Image#[a]	No. of features	*Mean*	*SD*	Image#[a]	No. of features	*Mean*	*SD*
1	550	7.33	2.915	16	475	6.33	2.321
2	555	7.40	3.000	17	405	5.40	2.329
3	569	7.59	3.329	18	407	5.43	2.188
4	609	8.12	3.624	19	643	8.57	4.451
5	578	7.71	3.552	20	502	6.69	3.166
6	611	8.15	2.939	21	486	6.48	2.723
7	579	7.72	3.375	22	534	7.12	2.630
8	510	6.80	2.800	23	563	7.51	3.042
9	429	5.72	2.414	24	652	8.69	3.526
10	550	7.33	2.905	25	495	6.60	3.000
11	571	7.61	2.686	26	517	6.89	2.749
12	587	7.83	3.130	27	420	5.60	2.857
13	380	5.07	2.029	28	409	5.45	2.029
14	405	5.40	1.816	29	511	6.81	3.224
15	352	4.69	2.278	30	447	5.96	2.178
				Total	15301	6.80	2.895

Note. [a]Image numbers refer to arbitrary ID numbers assigned by the researcher. n=75

The unit of analysis for all variables in this study, including measures of common and distinctive measures, is pair of images. For the sample of 30 images, there are 435 unique pairs; hence the total number of cases (N) is 435. Values of the measures of common and distinctive features (that is, values of the variables: COMMNFE1, COMMNFE2, COMMNFE3, DISTFEA1, DISTFEA2, DISTFEA3, DISTFEB1, DISTFEB2, DISTFEB3) for each pair were determined using the three methods discussed in chapter 3. The data were then processed using SPSS© version 12.0.1 for Windows (SPSS Inc., 2003) to produce summary statistics for the nine variables that correspond to the nine measures of common (three) and distinctive (three each for a and b, where a and b make a pair of images) features presented in Table 5.

Table 5

Descriptive Statistics for Measures of Common and Distinctive Features

	Measure	*Mean*	*SD*	Skewness	Kurtosis	Cronbach's α
1.	Common features using method 1 (COMMNFE1)	0.060	0.102	6.627**	61.987**	.841
2.	Common features using method 2 (COMMNFE2)	1.565	0.469	1.387**	4.367**	.988
3.	Common features using method 3 (COMMNFE3)	63.260	46.488	1.376**	3.819**	.988
4.	Distinctive features of a using method 1 (DISTFEA1)	4.236	1.493	-0.230	-0.138	.841
5.	Distinctive features of a using method 2 (DISTFEA2)	5.158	0.969	-0.197	-0.391*	.961
6.	Distinctive features of a using method 3 (DISTFEA3)	386.457	73.428	-0.241	-0.277	.961
7.	Distinctive features of b using method 1 (DISTFEB1)	3.773	1.453	-0.255	-0.339	.841
8.	Distinctive features of b using method 2 (DISTFEB2)	4.561	1.125	0.030	-0.771**	.971
9.	Distinctive features of b using method 3 (DISTFEB3)	340.623	86.508	-0.068	-0.621**	.971

Note. N=435. *$p < .05$, **$p < .01$

All observed x variables, except three (DISTFEA1, DISTFEA3, DISTFEB1), had either skewness or kurtosis or both values significantly different from zero ($p < .05$). The original scores had to be transformed to meet univariate normality criterion for further analyses.

Analysis of Similarity Judgment Task Data

The similarity judgment task involved magnitude estimation by participants. As a calibration exercise and to acquaint them with the process of magnitude estimation, participants were presented with a set of five straight lines of varying lengths (two to eight inches) and asked to assign numbers proportional to the apparent length of the lines. The magnitude estimations of the five lines by the 75 participants of the similarity judgment task were averaged across all participants of the task using geometric mean, after log transformations. The Pearson product moment correlation between the geometric means of the logarithms of the magnitude estimations of the lines and the logarithms of their actual lengths is .999, an almost perfect correlation.

Participants were then shown a set of 435 pairs of the 30 images in the sample (set 1) and then the second set (set 2) of 435 pairs (obtained by reversing the order of presentation of pairs in set 1 as well as the order of images in each pair). Similarity judgments (ratings) or original raw data for the two sets of 435 pairs of the 30 images in the sample, that is data for the two observed Y (dependent) variables (SIMAB for the first set, SIMBA for the second set), consist of magnitude estimations (ranging from 1 to 500, where an inch on the straight line used in the web form represents 100 units/numbers) by all participants for each pair of images which were later aggregated using the geometric mean, after log-transformations. According to Stevens (1966, 1975), magnitude estimation data is log-normal, hence the need for log-transformations. The appropriate measure of central tendency for such data is the geometric mean rather than the arithmetic mean. Hence, the geometric means of the base 10 logarithms of the magnitude estimations make up values of the two variables, SIMAB and SIMBA. Table 6 shows the descriptive statistics for the two variables.

Table 6

Descriptive Statistics for Similarity Judgments

Variable	*Mean*	*SD*	Skewness	Kurtosis	Cronbach's α
Similarity judgments of set 1 (SIMAB)	1.719	0.252	0.813**	0.330	.965
Similarity judgments of set 2 (SIMBA)	1.654	0.227	1.211**	1.565**	.963

Note. N=435. **p < .01

The reliability coefficients, that is Cronbach's alpha values, for both variables are well above the minimum value often cited by the literature, which is 0.70 (Nunnally & Bernstein, 1994). However both SIMAB and SIMBA had skewness values significantly different from zero while SIMBA had a kurtosis value significantly different from zero ($p < .01$), a property of non-normal scores, prompting the researcher to use transformations of the two variables in subsequent analyses.

Research Findings and Discussion

Prior to any kind of analysis, including structural equation modeling, data need to be checked for missing values, outliers and univariate normality (Schumacker & Lomax, 1996; West, Finch, & Curran, 1995). In the data collected for this study, there were no missing data and no apparent extreme outliers. However, the similarity judgment data (the observed Y variables: SIMAB and SIMBA) and some of the common and distinctive feature data (six of the nine observed X variables: COMMNFE1, COMMNFE2, COMMNFE3, DISTFEA2, DISTFEB2, DISTFEB3) did not satisfy the univariate normality criterion. Their skewness and kurtosis values (see Tables 5 and 6) are well above the recommended values of zero (Tabachnick & Fidell, 1989), zero being the skewness and kurtosis for a normal distribution. The recommended courses of action when a distribution of scores is non-normal are: (a) transformations such as log-transformations (log x), taking the square roots ($\sqrt{x}$), and inverse (1/x) of scores, and (b) the use of asymptotically distribution free least squares estimation method for structural equation modeling (including model testing) (Schumacker & Lomax, 1996; West, Finch, & Curran, 1995). Even though the log-transformation of all observed variables in this study yielded scores with improved skewness and kurtosis values (See Table 7), they still failed to satisfy the univariate normality criterion. Besides, the log-transformed scores have different means and standard deviations from the original raw scores. In order to achieve the univariate normality, the raw data were subjected to the "Normal Scores" algorithm in PRELIS (du Toit & du Toit, 2001; Jöreskog & Sörbom, 1999). It is evident from Table 7 that the skewness and kurtosis values for normal scores of almost all variables are either zero or very close to zero (p >.94) and the normal scores have the same mean and standard deviation values as the original scores (Tables 5 and 6).

Table 7

Descriptive Statistics for Normal Scores and Logarithmic Transformations of all Observed Variables

	Normal Scores				Log-Transformed Scores			
Variable	*Mean*	*SD*	Skewness	Kurtosis	*Mean*	*SD*	Skewness	Kurtosis
COMMNFE1	0.060	0.102	0.001*	-0.009*	2.459	0.582	-0.824	1.445
COMMNFE2	1.565	0.469	0.061**	-0.178**	4.169	0.232	-13.300	239.953
COMMNFE3	63.260	46.488	0.005*	-0.025*	1.632	0.471	-1.385	1.871
DISTFEA1	4.236	1.493	0.000*	-0.008*	1.589	0.207	-1.931	5.554
DISTFEA2	5.158	0.969	-0.001*	-0.011*	1.704	0.087	-0.719	0.477***
DISTFEA3	386.458	73.428	-0.001*	-0.011*	2.579	0.089	-0.848	1.127
DISTFEB1	3.773	1.453	0.000*	-0.008*	1.529	0.231	-1.626	2.724
DISTFEB2	4.561	1.125	0.000*	-0.009*	1.645	0.113	-0.470	-0.413
DISTFEB3	340.623	86.508	0.000*	-0.009*	3.517	0.120	-0.722	0.388***
SIMAB	1.719	0.252	0.000*	-0.008*	1.231	0.061	0.478	-0.220***
SIMBA	1.654	0.227	0.000*	-0.009*	1.215	0.056	0.823	0.641

Note. N=435. *$p > .94$, ** $p > .46$, ***$p > .05$

Another crucial task of data screening is checking for pairwise linearity of all variables using scatter plots. An examination of the scatter plots for all pairs of the observed variables in the study shows that they are linearly related.

As part of data screening, the convergent validity of a construct or latent variable needs to be assessed as well. To establish convergent validity, observed variables or measures of a particular latent construct need to correlate (Nunnally & Bernstein, 1994). It is evident from Table 8 that all relevant Pearson's product moment correlation coefficients are significant ($p < .001$). Therefore, it is safe to conclude that all latent variables have good convergent validity with respect to their observed variables.

Table 8

Correlations among Observed Variables for each Latent Construct

Latent Construct		Correlations among observed variables		
		COMMNFE1	COMMNFE2	COMMNFE3
	COMMNFE1	1.000		
Common Features	COMMNFE2	.730**	1.000	
	COMMNFE3	.916**	.861**	1.000
		DISTFEA1	DISTFEA2	DISTFEA3
	DISTFEA1	1.000		
Distinctive features of a	DISTFEA2	.782**	1.000	
	DISTFEA3	.784**	1.000**	1.000
		DISTFEB1	DISTFEB2	DISTFEB3
	DISTFEB1	1.000		
Distinctive features of b	DISTFEB2	.679**	1.000	
	DISTFEB3	.683**	.998**	1.000
		SIMAB	SIMBA	
Similarity	SIMAB	1.000		
	SIMBA	.799**	1.000	

**p < .001, one-tailed

<u>Extent to which Various Methods Measure the Common and Distinctive Features of Images (RQ1)</u>

The first research question is: Which methods measure the common and distinctive features of images well? Three measures have been reported in the literature (Johnson, 1986; Tversky, 1977). In chapter 3, the three different methods for measuring common and distinctive features of pairs of images were discussed. The main purpose of this research question (RQ1) is to determine how well the three methods measure the three independent latent variables. This was assessed, using output from LISREL©, through measurement model fit indices and loadings of those measures (or estimated parameters called $\lambda^{(x)}$'s) on corresponding independent latent variables/constructs (common and distinctive features). A covariance matrix (Table 9) consisting of all observed x and y variables (three observed x variables for each of the independent latent variables and two observed y variables for the dependent latent variable), produced from feature data (obtained through an image description task and summarized using three methods for measuring common and distinctive features), and the two observed y variables, SIMAB & SIMBA, was produced using LISREL©.

Convergent validity of the observed variables was already assessed (see the Research Findings section above) by examining the correlations of observed variables for each independent latent variable or construct. A quick glance at Table 8 suggests that all nine observed x variables satisfy the convergent validity criterion of the correlations being statistically significant (Nunnally & Bernstein, 1994).

The relevant elements of the covariance matrix in Table 9 (only elements for the nine observed x variables) were used in the analysis. A single LISREL© program (using the maximum likelihood estimation method) produced results of tests of model fit of the measurement models for the three independent latent variables measured using the three methods and these results are shown in Figure 8. The independent latent constructs (common and distinctive features) were allowed to correlate. How well a given observed variable measures a latent variable or construct is dependent on the statistical significance of the relevant loadings or parameter estimates (Schumacker & Lomax, 1996), provided the measurement model fits the sample data well.

Table 9

Covariance Matrix for all Observed Variables in the Study (Normal Scores)

		1	2	3	4	5	6	7	8	9	10	11
1.	COMMNFE1	0.010										
2.	COMMNFE2	0.035	0.220									
3.	COMMNFE3	4.360	18.768	2161.113								
4.	DISTFEA1	-0.093	-0.346	-43.526	2.230							
5.	DISTFEA2	-0.041	-0.127	-19.102	1.132	0.940						
6.	DISTFEA3	-3.147	-9.716	-1458.167	86.012	71.167	5391.722					
7.	DISTFEB1	-0.070	-0.223	-29.867	0.354	0.205	15.866	2.112				
8.	DISTFEB2	-0.016	0.010	-5.330	-0.003	0.024	1.985	1.110	1.265			
9.	DISTFEB3	-1.353	0.529	-442.821	0.399	1.949	164.606	85.895	97.137	7483.558		
10.	SIMAB	0.017	0.061	7.896	-0.191	-0.072	-5.539	-0.128	-0.024	-2.117	0.063	
11.	SIMBA	0.014	0.051	6.605	-0.156	-0.060	-4.616	-0.124	-0.023	-2.019	0.046	0.051

Note. N=435

In other words, in addition to the assessment of the significance of the loadings ($\lambda^{(x)}$s) on independent latent constructs, the extent of fit of the measurement models need to be assessed as well. The main goal of assessing model fit is to see whether the model produces the original sample covariance matrix (of observed variables) with minimum or no residuals. This is achieved through examination of several model fit indices. In other words, the various model fit indices are indicators of how well the model specified by the researcher fits sample data.

In order to assess model fit, several fit indices (criteria) are recommended instead of a single index since there is no single best index (Schumacker & Lomax, 1996). Consequently, the most widely used indices such as the Chi-square (χ^2), the Standardized Root Mean Square Residual (SRMR), Root Mean Square Error of Approximation (RMSEA), the Goodness-of-fit index (GFI), and the Adjusted-goodness-of-fit index (AGFI) were used to assess the fit of the measurement models to sample data.

Chi-square (χ^2) is a measure of overall model fit and "measures the distance (difference, discrepancy, deviance) between the sample covariance (correlation) matrix and the fitted covariance (correlation) matrix" (Jöreskog, 1993, p. 308). The larger the value of the Chi-square (χ^2), the worse the model fit to the data as it is an indication of a significant discrepancy between the sample covariance (correlation) matrix and the reproduced (or model implied) covariance (correlation) matrix.

The Root Mean Square Residual (RMR) is a measure of the mean difference between the sample (or observed) and the reproduced (model implied) covariance (correlation) matrices. It is the square root of the mean of the squared differences (residuals) between the sample (observed) and the implied covariance (correlation) matrices. The Standardized Root Mean Square Residual (SRMR) is preferred to the Root Mean Square Residual (RMR) since it is a standardized summary of average covariance discrepancies and its values lie between zero and one. A value closer to one is an indication of a good fit (Kelloway, 1998).

The Root Mean Square Error of Approximation (RMSEA), which is a "measure of discrepancy per degree of freedom" (Jöreskog, 1993, p. 310), is a model fit measure that takes the population error of approximation into account. The degrees of freedom for a model with k parameters to be estimated and q observed variables is q(q+1)/2 – k.

The Goodness-of-fit index (GFI) and the Adjusted-goodness-of-fit index (AGFI) were developed as alternatives to the Chi-square (χ^2) which is "N-1 times the minimum value of the fit function" and it is affected by N, the sample size (Jöreskog, 1993, p. 309). Both the Goodness-of-fit index (GFI) and the Adjusted-goodness-of-fit index (AGFI) "do not depend on sample size explicitly and measure how much better the model fits compared with no model at all" (Jöreskog, 1993, p. 309) and their values range between zero and one, with values closer to one indicating better fit of the model to the data.

The acceptable levels or thresholds of these model fit indices are: a Chi-square (χ^2) value that is non-significant (in other words, the observed and estimated covariance matrices need to be found to not be statistically different), values of RMSEA and SRMR below 0.05, and values of GFI and AGFI at least 0.90 (Hu & Bentler, 1999; Jöreskog, 1993).

Table 10 shows values of the model fit indices for the measurement models of common and distinctive features using all observed x variables (measurement model 1 – Figure 8), and pairs of observed x variables (measurement models 2, 3, and 4 – Figures 9, 10, and 11, respectively). Measurement model 1 (all observed x variables – Figure 8) is what has been proposed, based on the measures of common and distinctive features reported in the literature on the contrast model, when the contrast model was specified in chapter 3 (Figure 7). However, none of the values of the fit indices were below or above the recommended thresholds (χ^2=548.48, *df*=27, *p*<.01, *RMSEA*=.211, *SRMR*=.125, *GFI*=.781, *AGFI*=.635).

Since the measurement model with all nine observed x variables is not a good fit to the data, the model had to be modified. Three modified measurement models with pairs of the observed x variables for each independent latent variable were considered. Measurement model 2 (using observed x variables measured using methods 2 and 3 discussed in chapter 3, Figure 9) was the first to be tested. Once again, the model fit indices for this modified model showed a poor fit (χ^2=136.24, *df*=9, *p*<.01, *RMSEA*=.180, *SRMR*=.0435, *GFI*=.905, *AGFI*=.779).

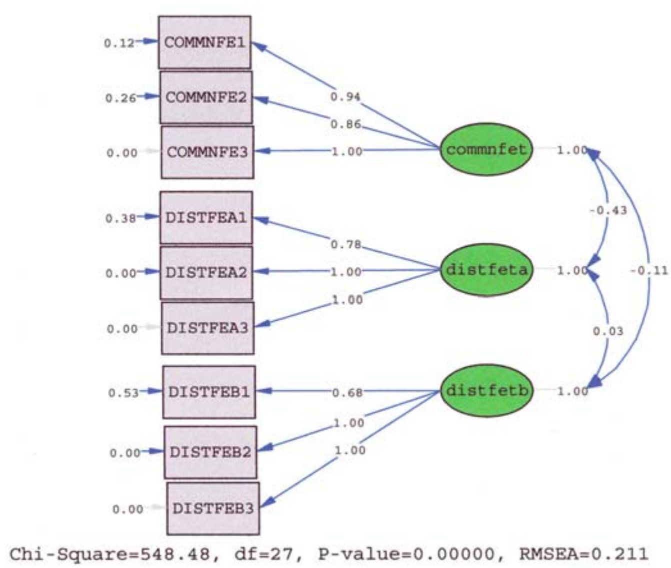


Figure 8. Measurement model for common and distinctive features (all observed x variables – measurement model 1).

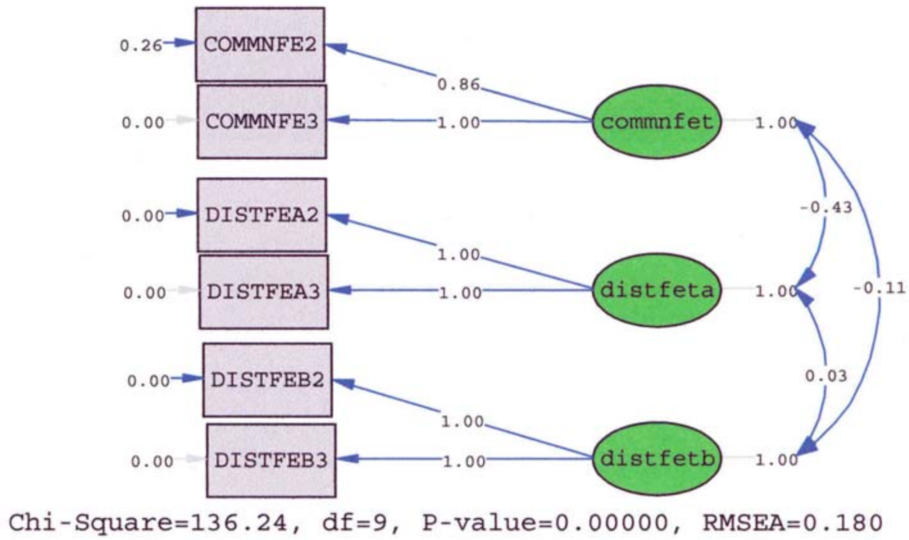


Figure 9. Measurement model for common and distinctive features (observed x variables measured using methods 2 & 3 – measurement model 2).

It is evident from Table 10 that measurement models with a combination of observed x variables measured using methods 1 & 3 (Figure 10) and methods 1 & 2 (Figure 11) fit the data well. Measurement model 3 (observed x variables measured using methods 1 & 3) had all five model fit indices that exceeded their recommended minimum levels or that are less than their recommended maximum levels (χ^2=7.78, df=4, *p*>.05, *RMSEA*=.047, *SRMR*=.0207, *GFI*=.994, *AGFI*=.969). Likewise, measurement model 4 (observed x variables measured using methods 1 & 2) had all model fit indices that either exceeded their recommended minimum levels or that are less than their recommended maximum levels (χ^2=4.75, *df*=4, p>.3, *RMSEA*=.021, *SRMR*=.0234, *GFI*=.996, *AGFI*=.981). Measurement model 4 (observed x variables measured using methods 1 & 2) fits the data slightly better than measurement model 3 (observed x variables measured using methods 1 & 3).

Table 10

Model Fit Statistics for Measurement Models of Common and Distinctive Features

Model	χ^2	*df*	*p*	*RMSEA*	*SRMR*	*GFI*	*AGFI*
Measurement model 1 (all observed x variables)	548.48	27	.00	.211	.125	.781	.635
Measurement model 2 (observed x variables measured using methods 2 & 3)	136.24	9	.00	.180	.0435	.905	.779
Measurement model 3 (observed x variables measured using methods 1 & 3)	7.78	4	.0998	.047	.0207	.994	.969
Measurement model 4 (observed x variables measured using methods 1 & 2)	4.75	4	.3138	.021	.0234	.996	.981

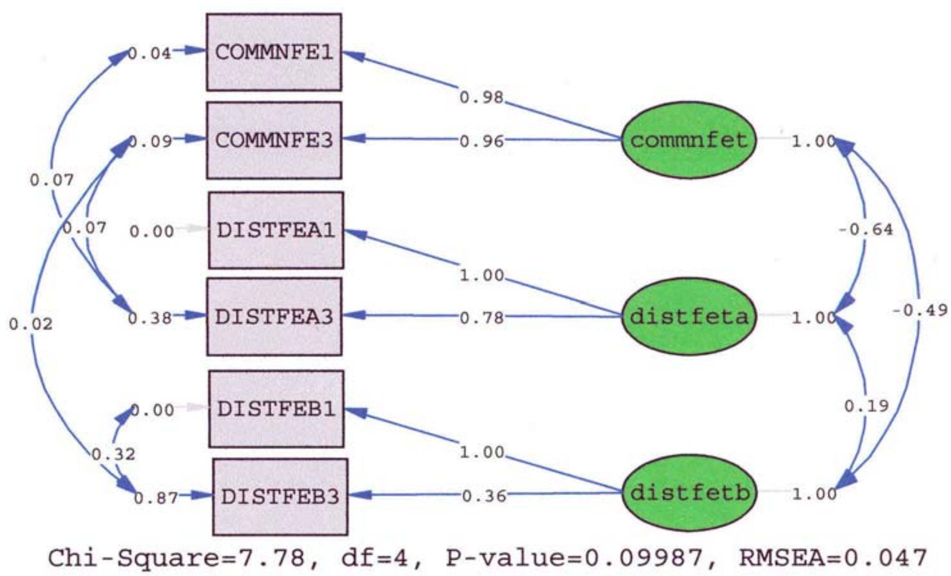


Figure 10. Measurement model for common and distinctive features (observed x variables measured using methods 1 & 3 – measurement model 3).

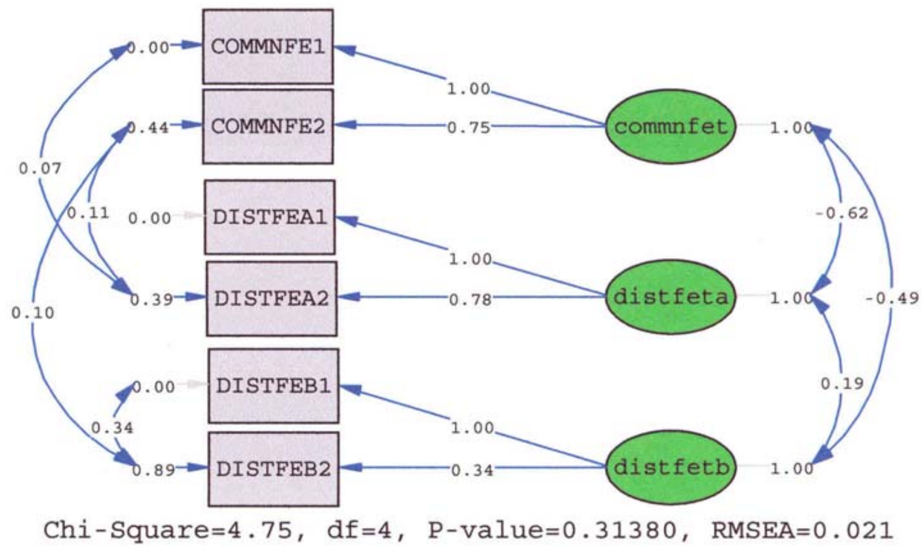


Figure 11. Measurement model for common and distinctive features (observed x variables measured using methods 1 & 2).

Measurement models that have observed x variables measured using methods 2 and 3 together did not fit the data well consistently. This appears to be due to multicollinearity, which is a characteristic of two or more measures of the same latent construct (or predictors of the same dependent variable in regression analysis) that are highly correlated, with correlations above .90 and due to singularity, which is a characteristic of two or more measures of the same latent construct that are perfectly correlated (Tabachnick & Fidell, 1989). The correlation between DISTFEA2 and DISTFEA3, measures of distinctive features of a (where a and b form a pair) measured using methods 2 and 3 is .998 while the correlation between DISTFEB2 and DISTFEB3 is 1.00. The source of this multicollinearity and singularity is the fact that observed x variables measured using method 2 are simply observed x variables measured using method 3 divided by the number of participants who attributed the common (for measures of common features) or distinctive (for measures of distinctive features) features to pairs of or to individual images.

Despite the fact that measurement models 1 (all observed x variables) and 2 (observed x variables measured using methods 2 and 3) did not fit the sample data well, all loadings of the observed x variables onto their corresponding independent latent constructs in all the four measurement models were significant (p<.01).

Extent of Fit of the Contrast Model to Sample Data (RQ2)

The second research question is: To what extent does the contrast model fit human similarity judgments and features/attributes data for a sample of images? The two-step approach of model testing (Jöreskog & Sörbom, 1993; Schumacker & Lomax, 1996) was followed. In the two-step approach, the measurement model is tested first to see if it holds for the set of observed and latent variables and to see if the observed x variables measure their respective latent variables. The structural model is then tested once the measurement model holds.

The measurement models for the independent latent variables (common and distinctive features) with four different combinations of the observed x variables have already been tested. Measurement models of independent latent variables with observed x variables measured using methods 1 & 3 and methods 1 & 2 fit the data well. Therefore, only two measurement models, one consisting of the observed x variables measured using methods 1 & 3 and the two observed y variables (SIMAB and SIMBA) (Figure 12) and another one consisting of observed x variables measured using methods 1 & 2 and the two observed y variables (Figure 13) were tested.

A quick look at Table 11 indicates that the measurement model with a combination of observed x variables measured using methods 1 & 3 (Figure 12) (χ^2=22.18, *df*=10, $p < .05$, *RMSEA*=.053, *SRMR*=.0185, *GFI*=.987, *AGFI*=.955) is a poor fit compared to the measurement model with a combination of observed x variables measured using methods 1 & 2 (Figure 13) (χ^2=16.97, *df*=10, p>.05, *RMSEA*=.040, *SRMR*=.0205, *GFI*=.990, *AGFI*=.965). The latter has all five model fit indices that exceeded their recommended minimum levels (*GFI* >.90, and *AGFI* > .90) or that are less than their recommended maximum levels (non-significant χ^2 with $p > .05$, *RMSEA* <.05, *SRMR* <.05). Therefore, the measurement model with observed x variables measured using methods 1 & 2 was used to test the contrast model.

Table 11

Model Fit Statistics for Measurement Models of Common Features, Distinctive Features, and Similarity

Model	χ^2	*df*	*p*	*RMSEA*	*SRMR*	*GFI*	*AGFI*
Measurement model with observed x variables measured using methods 1 & 3 and the two observed y variables	22.18	10	.01422	.053	.0185	.987	.955
Measurement model with observed x variables measured using methods 1 & 2 and the two observed y variables	16.97	10	.07508	.040	.0205	.990	.965

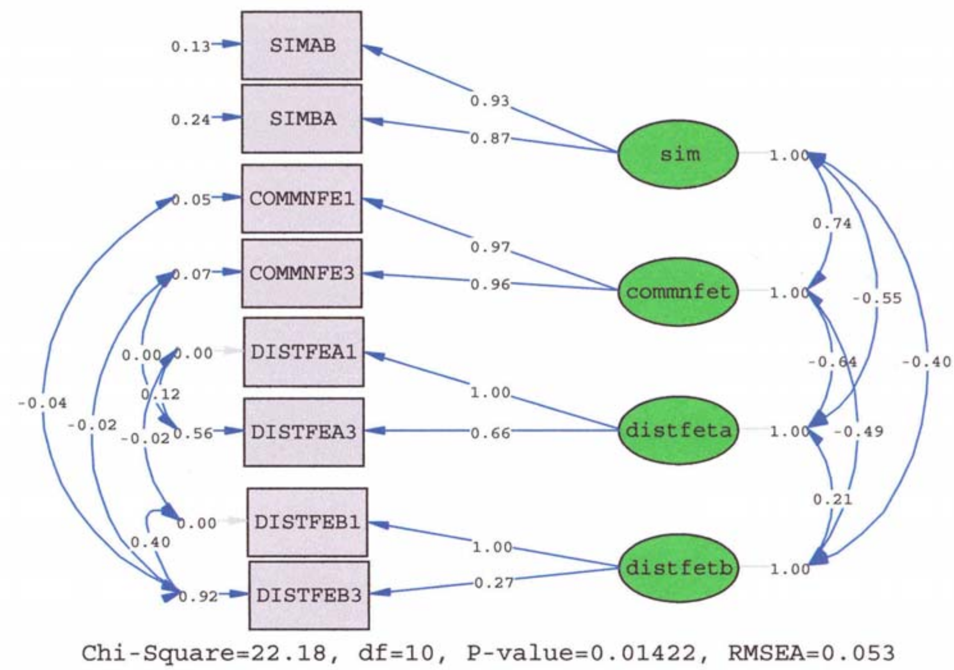


Figure 12. Measurement model for common and distinctive features (observed x variables measured using methods 1 & 3) and similarity.

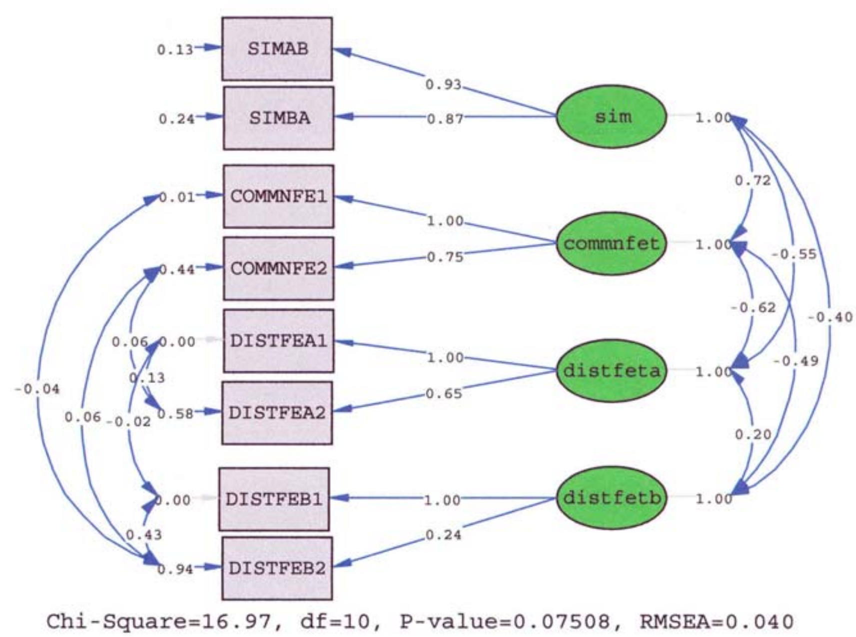


Figure 13. Measurement model for common and distinctive features (observed x variables measured using methods 1 & 2) and similarity.

A modified version of the contrast model, with observed x variables measured using methods 1 & 2 and the two observed y variables (SIMAB and SIMBA), was tested and not only did the model fit the sample data well (χ^2=16.97, *df*=10, *p* >.05, *RMSEA*=.040, *SRMR*=.0205, *GFI*=.990, *AGFI*=.965) (see Figure 14), all the loadings (parameter estimates) for the observed x and y variables onto their respective latent variables were statistically significant ($p < .05$). Table 12 presents the standardized loadings of the observed variables on the latent variables for the modified contrast model.

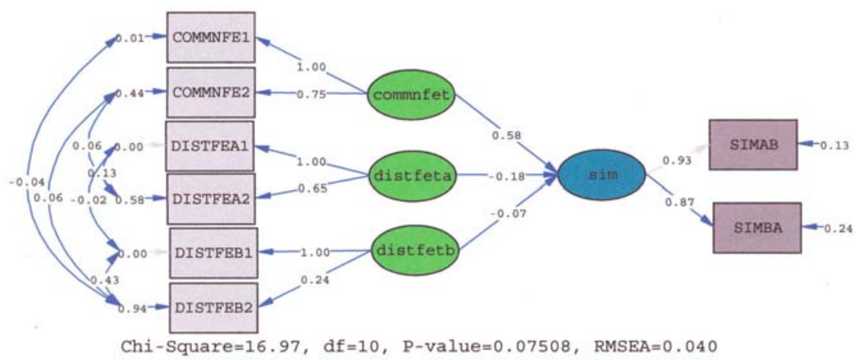


Figure 14. The modified contrast model.

Table 12

Standardized Loadings of Observed x and y Variables on the Latent Variables in the Modified contrast model

Construct	Observed Variable	Latent Variable Loadings
Common features	COMMNFE1	1.00**
	COMMNFE2	.75**
Distinctive features of a	DISTFEA1	1.00**
	DISTFEA2	.65**
Distinctive features of b	DISTFEB1	1.00**
	DISTFEB2	.24*
Similarity	SIMAB	.93**
	SIMBA	.87**

*$p < .05$, **$p < .01$

The path coefficients (except for the path coefficient for the structural relationship between distfetb to sim, with $p < .06$) in the structural model with the three independent latent variables (commnfet, distfeta, and distfetb) and the latent dependent variable (sim) were statistically significant ($p < .05$). Even though the error variance of the latent dependent variable (sim) was .45 and significant (t=9.99), the major portion of its variance is explained (R^2=.55) by the combination of the three independent latent variables (commnfet, distfeta, distfetb). Based on the correlation matrix for the standardized scores of the dependent latent variable (sim) and the three independent latent variables (commnfet, distfeta, distfetb) generated by LISREL©, the proportion of the total amount of variance in the dependent latent variable (sim) explained by each of the independent latent variables is 53.29% (r=.73), 30.58%(r= -.553), and 16.00%(r= -.40), respectively. Furthermore, the direction of the path coefficients (positive path coefficients for common features and negative path coefficients for distinctive features) is proof that Tversky's (1977) statement that common features increase while distinctive features decrease the degree of similarity of pairs of objects. Figure 15 presents the structural model of the modified contrast model with t-values for the path coefficients.

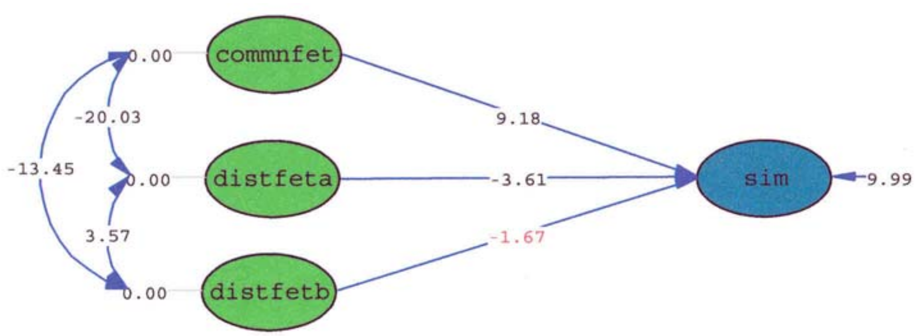


Figure 15. The structural model of the modified contrast model.

Therefore, it is fair to conclude that Tversky's (1977) contrast model, in its modified form (Figure 14), provides a reasonably good fit to the data for a sample of images with path coefficients that are in the right direction and with the right magnitudes.

Relationships between Perceived Features and Similarity of Images (RQ3)

The third research question is: What is the relationship between perceived similarity of images, as judged by humans, and their features/attributes identified and described/listed by humans? The main purpose of the study, by seeking an answer to this question, is to see whether more common features and less distinctive features result in higher similarity judgments. Pearson's product moment correlation was used to determine the relationships between perceived features and similarity of images. The correlations between measures of common and distinctive features and similarity judgments are given in Table 13.

As originally argued by Tversky (1977) and subsequently confirmed by Tversky and Gati (1978) and Johnson (1981, 1986), common features contribute to higher similarity judgments. Correlations between similarity judgments of the first set of pairs of images (SIMAB) and the three measures of their common features were found to be positive and statistically significant (r=.648, .513, and .675; $p < .01$). Likewise, the correlations between similarity judgments of the second set of pairs of images (SIMBA) and the three measures of their common features were positive and statistically significant as well (r=.624, .480, and .626; $p < .01$).

In contrast, distinctive features contribute to lower similarity ratings with correlations between similarity judgments of the first set of pairs of images (SIMBA) and measures of their distinctive features negative and statistically significant (r=-.507, -.296, and -.300; $p < .01$, for the three measures of distinctive features of a, when a and b form a pair, respectively, and r=-.351, -.084, and -.097, $p < .05$, for the three measures of distinctive features of b). Similar results hold for the correlations between similarity judgments of the second set of pairs of images (SIMBA) and the three measures of their distinctive features (r=-.460, -.274, and -.277; $p < .01$, for distinctive features of a, and r=-.375, -.091, and -.103, $p < .05$, for distinctive features of b).

Results of the structural equation analysis (model testing) under research question 2 (RQ2) are in accord with those of the correlations where path coefficients between the independent latent variables (common and distinctive features) and the dependent latent variable (similarity) match correlation coefficients in both magnitude and direction.

Table 13

Means, Standard Deviations, and Correlations of Measures of Common and Distinctive Features and Similarity Judgments of Images

		Mean	*SD*	SIMAB	SIMBA
1.	COMMNFE1	0.060	0.102	.648**	.624**
2.	COMMNFE2	1.565	0.469	.513**	.480**
3.	COMMNFE3	63.260	46.488	.675**	.626**
4.	DISTFEA1	4.236	1.493	-.507**	-.460**
5.	DISTFEA2	5.158	0.969	-.296**	-.274**
6.	DISTFEA3	386.458	73.428	-.300**	-.277**
7.	DISTFEB1	3.773	1.453	-.351**	-.375**
8.	DISTFEB2	4.561	1.125	-.084*	-.091*
9.	DISTFEB3	340.62	86.508	-.097*	-.103*
10.	SIMAB	1.719	0.252	1.000	.799**
11.	SIMBA	1.654	0.227	.799**	1.000

Note. N=435, *$p < .05$, **$p < .01$, one-tailed

These results suggest that the more common features and the fewer distinctive features two images have, the more similar they are. Conversely, the fewer common features and the more distinctive features they have, the less similar they are. This is in total agreement with Tversky's (1977) assertion in formulating the contrast model and they are consistent with results from a similar research in marketing by Johnson (1986), who used consumer products as materials.

Relative Weights of Common and Distinctive Features of Images on their Similarity (RQ4)

The fourth and final research question is: What are the relative weights given to common and distinctive features in human similarity judgments of images? The main purpose of the study regarding this research question is to determine, through structural equation modeling and regression analysis, the relative weights of common and distinctive features (independent variables) on similarity (dependent variable).

In order to estimate the relative weights of common and distinctive features on similarity judgments, on top of results from the testing of the contrast model (RQ2), a regression analysis of similarity judgments on measures of common and distinctive features was conducted. In a way, the regression analysis results are used to validate results from structural equation modeling.

Tables 14 and 15 present the independent variables (measures of common and distinctive features), unstandardized (B) and standardized (Beta) regression coefficients as well as t values associated with the regression coefficients for the two observed Y (dependent) variables SIMAB (similarity judgments of the first set of 435 pairs of images) and SIMBA (similarity judgments for the second set), respectively. The coefficients of determination, R^2 values, with their corresponding F ratios are also presented in the two tables. The regression analyses results in Tables 14 and 15 confirm (except in the case of distinctive features measured using method 3) the fact that common and distinctive features influence similarity of images in the expected direction and with the expected magnitudes (weights) as stated in Tversky's (1977) contrast model. What is more, common and distinctive features reliably predict similarity of images.

In terms of their prediction power (weight), common features have the largest followed by distinctive features of a (where a and b form a pair). Distinctive features of b had the least weight. In other words, in the contrast model ($S(a,b)= \theta f(A \cap B) - \alpha f(A-B) - \beta f(B-A)$, for the sample of images in this study, $\theta > \alpha > \beta$. This is the case despite the fact that the participants were not asked to judge "how similar is image a to image b". Tversky's (1977) assertion that features of the subject (the first object in the pair) are weighed more than those of the referent (the second object in the pair) holds.

In the case of independent variables (common and distinctive features) measured using method 3, even though the explained portions of variability in similarity (dependent variable) for both sets (.46 and .40, respectively) were significant, weights of distinctive features were non-significant (p >.05). This is consistent with the findings of the tests of the measurement models for the independent latent variables where models with observed x variables measured using method 3 did not fit the sample data well.

Furthermore, results of the regression analysis are in agreement with those of the structural equation analysis. Both the directions and magnitudes of loadings of independent latent constructs (common and distinctive features) on the dependent latent construct (similarity) called gammas (γ) are comparable to corresponding standardized regression weights (βs).

Table 14

Regression Analysis of Similarity Judgments of the First Set (SIMAB) on Measures of Common and Distinctive Features of Images

Independent Variable	*B*	*Beta*	*t*
COMMNFE1	1.196	.486	9.505**
DISTFEA1	-.033	-.197	-4.311**
DISTFEB1	-.016	-.090	-2.181*
R^2=.45, F(3,431)=116.101**			
COMMNFE2	.252	.469	11.136**
DISTFEA2	-.042	-.163	-3.865**
DISTFEB2	-.020	-.089	-2.210*
R^2=.30, F(3,431)=60.423**			
COMMNFE3	.004	.666	16.838**
DISTFEA3	-.00005	-.015	-.370
DISTFEB3	-.00007	-.024	-.658
R^2=.46, F(3,431)=120.3**			

*p < .05, **p < .01

Table 15

Regression Analysis of Similarity Judgments of the Second Set (SIMBA) on Measures of Common and Distinctive Features of Images

Independent Variable	*B*	*Beta*	*t*
COMMNFE1	1.042	.470	8.912**
DISTFEA1	-.023	-.154	-3.268**
DISTFEB1	-.020	-.129	-3.036**
R^2=.41, F(3,431)=100.913**			
COMMNFE2	.213	.440	10.199**
DISTFEA2	-.035	-.148	-3.438**
DISTFEB2	-.019	-.096	-2.310*
R^2=.26, F(3,431)=50.604**			
COMMNFE3	.003	.617	14.79**
DISTFEA3	-.00004	-.013	-.303
DISTFEB3	-.00009	-.035	-.917
R^2=.40, F(3,431)=93.285**			

*p < .05, **p < .01

Summary

This chapter presents results of the analysis of data, research findings, and discussions of the results and findings with respect to the four research questions. Participants are described by demographic variables such as gender, age, major, etc. Results of the analysis of the image description and similarity judgment tasks, including descriptive statistics (*mean*, *SD*, skewness, kurtosis, Cronbach's α) for all observed variables are presented. Data screening procedures followed and corrective measures taken before subjecting the data to further statistical analyses are explained.

All four research questions were explored. Research question 1 (RQ1) deals with the measurement models of the independent latent variables (common and distinctive features). Measurement models with observed x variables measured using methods 1 & 3 and 1 & 2 fit the data well. The second research question (RQ2) deals with testing the fit of the contrast model to sample data. A modified version of the contrast model with observed x variables measured using methods 1 & 2 is a good fit to the data with significant path coefficients (or loadings).

The third research question (RQ3) concerns the relationships (associations) between common and distinctive features (independent variables) and similarity (dependent variable) of images. Pearson's product moment correlation coefficients for the relationships between measures of common features and similarity of images are found to be positive and statistically significant ($p<.01$), while coefficients for the relationships between measures of distinctive features and similarity of images are negative and statistically significant ($p<.05$). The fourth and last research question (RQ4) deals with the predictive power (weights) of the common and distinctive features on similarity of images. Results obtained under research question 2 (RQ2), that is, structural coefficients and their direction and magnitude are confirmed: common features have larger weights on similarity than distinctive features.

CHAPTER 5

SUMMARY AND CONCLUSIONS

Introduction

The main purpose of this study was to investigate the nature of the relationships between common and distinctive features of images and their similarity, using Tversky's (1977) contrast model as a theoretical framework, and to test the contrast model within the context of image representation and retrieval. Four research questions were formulated and explored to address the main purpose of the study. The first two research questions address issues related to measures of common and distinctive features of images and measurement as well as structural model fit to the sample data. The remaining two research questions address the nature of the relationships between common and distinctive features (independent variables) and their similarity (dependent variables), including the prediction power (weights) of the independent variables.

Data were collected from 150 participants who were randomly assigned to two tasks (75 participants per task), an image description and a similarity judgment task, using a random sample of 30 images (435 pairs). The image description task data were summarized through content analysis and three measures of common and distinctive features. After initial screening and appropriate corrective measures, a set of 11 observed (nine observed x and two observed y) variables and four latent constructs/variables were subjected to analysis using LISREL© 8.54 (Jöreskog & Sörbom, 1993).

Results of the analysis together with research findings and discussions with respect to the four research questions are presented in chapter 4. This chapter summarizes the results and research findings, points out limitations of the study, and presents concluding remarks as well as implications of research findings for similar research and practice. Finally, recommendations are made as to possible considerations for future research.

Summary of the Findings

To reiterate, the main purpose of this study was to explore the nature of the relationships between features (common and distinctive) and similarity of images, using Tversky's (1977) contrast model as a theoretical framework. This was achieved through correlation and regression analysis, and structural equation modeling. Four research questions were considered.

The first research question dealt with three different methods used in the literature to measure common and distinctive features of objects. Four measurement models for the independent latent variables (common and distinctive features) were tested to find out which observed x variables, measured using the three methods, measure the independent latent variables well. In a way, this is also an indirect test of the appropriateness of the three methods. Two of the measurement models, one with the observed x variables measured using all three methods and the other with observed x variables measured using methods 2 and 3 did not fit the data well. Values of the model fit indices for the first measurement model (all observed x variables) were: χ^2=548.48, *df*=27, *p*<.01, *RMSEA*=.211, *SRMR*=.125, *GFI*=.781, *AGFI*=.635. Values of the fit indices for the second measurement model were: χ^2=136.24, *df*=9, *p*<.01, *RMSEA*=.180, *SRMR*=.0435, *GFI*=.905, *AGFI*=.779. The remaining two measurement models, one with observed x variables measured using methods 1 and 3 (χ^2=7.78, *df*=4, *p*>.05, *RMSEA*=.047, *SRMR*=.0207, *GFI*=.994, *AGFI*=.969) and the other with observed x variables measured using methods 1 and 2 (χ^2=4.75, *df*=4, *p*>.3, *RMSEA*=.021, *SRMR*=.0234, *GFI*=.996, *AGFI*=.981) fit the sample data well. It turned out that the reason why measurement models involving observed x variables measured using methods 2 and 3 had poor fit to the sample data is due to multicollinearity (highly correlated observed variables measuring the same construct) and signgularity (perfectly correlated observed variables measuring the same construct) (Tabachnick & Fidell, 1989).

The second research question concerns the extent of fit of the contrast model to sample data. The two-step approach of model testing (Jöreskog & Sörbom, 1993; Schumacker & Lomax, 1996) was used, where the measurement model is tested first and then the test of the structural model proceeds once the measurement model holds for a set of observed and latent variables. The measurement model with a combination of observed x variables measured using methods 1 and 2 and the two observed y variables (SIMAB and SIMBA) is a good fit to the data (χ^2 =16.97, *df*=10, $p > .05$, *RMSEA*= .040, *SRMR*= .0205, *GFI*= .990, *AGFI*= .965). A modified version of the contrast model, with this measurement model, was tested and found to be a good fit to the data (χ^2 =16.97, *df*=10, $p > .05$, *RMSEA*= .040, *SRMR*= .0205, *GFI*= .990, *AGFI*= .965) with statistically significant structural coefficients between the independent latent variables (common and distinctive features) and the dependent latent variable (similarity). The structural coefficients have the right directions and magnitudes and results confirm Tversky's (1977) contrast model. The final structural equation for the contrast model fitted to the data for the sample of images in the study is: sim= .58*commnfet - .18*distfeta - .07*distfetb.

The third research question of the study was meant to investigate the relationships/associations between common and distinctive features and similarity of images. Some researchers have already found that while common features contribute to higher similarity judgments, distinctive features have the opposite effect (Johnson, 1981, 1986; Tversky, 1977; Tversky & Gati, 1978). This was substantiated by results of this study where the Pearson's product moment coefficients for the correlations between measures of common features and similarity of images are positive and statistically significant ($p < .01$, one-tailed). The correlations between measures of distinctive features and similarity of images are negative and significant ($p < .05$, one-tailed).

The fourth research question raised the issue of the prediction power (weights) of common and distinctive features as predictors of similarity of images. Both regression analysis and structural equation modeling results confirm Tversky's (1977) contrast model in that common and distinctive features can reliably predict similarity of images and common features have more predictive power (weights) than distinctive features. The structural equation for the contrast model (sim= .58*commnfet - .18*distfeta - .07*distfetb) compares well with the regression equations for the Model in terms of magnitudes and directions of structural coefficients and standardized regression weights. The regression equations for the contrast model (where SIMAB and SIMBA are the two observed y variables measuring similarity) are:

SIMAB= .486*COMMNFE1 - .198*DISTFEA1 - .090*DISTFEB1

SIMAB= .469*COMMNFE2 - .163*DISTFEA2 - .089*DISTFEB2

SIMBA= .470*COMMNFE1 - .154*DISTFEA1 - .129*DISTFEB1

SIMBA= .440*COMMNFE2 - .148*DISTFEA2 - .096*DISTFEB2

The numbers are standardized regression (β) weights.

Limitations of the Study

This study has some inherent limitations. Some of these limitations could not be avoided due to lack of access to well-defined populations of materials (images) and participants (image users). As a result, it was not practical to select a random sample of image users because such a population is not clearly defined and known.

Lack of a standard test collection of images forced the researcher to select a sample of images from a publicly available collection of images (included in a published book). The relatively smaller size of the sample of images might be a limitation even though the selection of a smaller sample size is enforced due to the scaling procedures (paired comparisons) used for data collection to scale/measure similarity. Even for the sample of 30 images, each participant had to look at 870 pairs of images.

In terms of the setting for data collection, it was not practical to bring participants to a common room and setting. Therefore, they were allowed to complete the image description and similarity judgment tasks on their own using their office or home computers and settings. This could be a potential limitation, although it was not evident from results of internal consistency measures, which are high (Cronbach's $\alpha > .84$).

Content analysis was used to build feature sets of images. Individual features were assigned to categories of features. Even though the inter-coder agreements were high, the process of assigning features to categories through content analysis might have had a bearing on the results. As a result of these limitations (assignment of features to categories), findings of this study may have limited generalizability. However, in spite of the above mentioned limitations, results of this study are valuable for future research.

Concluding Remarks

Attributes/features of documents are the basis for representation and indexing of both image and text documents. Not only does similarity play a central role in human perception and learning, psychological models of similarity have also been adopted for information retrieval purposes in the form of similarity measures used for determining inter-document similarity or similarity between representations of documents and users' queries. However, few researchers devote their time and energy in trying to understand the nature of perceived features and similarity of documents, including image documents. This study is the first to test a psychological model of similarity, other than geometric models, in the context of document representation and retrieval.

Results of the study point to the fact that a linear combination of common and distinctive features of images can reliably predict their similarity, an assertion made by Tversky (1977) when he formulated the contrast model of similarity between objects. The contrast model fits data for a sample of images well with the structural relationships between common and distinctive features (independent latent variables) and similarity (dependent latent variable) in the expected directions and with the expected magnitudes. Correlations between common and distinctive features and similarity of images are positive (and significant, $p < .01$) and negative (and significant, $p < .05$), respectively. This is due to the fact that participants tend to pay more attention to common features than distinctive features in their similarity judgments (Tversky, 1977).

Regression analysis results for the sample of images in this study indicate that common and distinctive features of images (measured using methods 1 and 2) are significant predictors of their similarity, with common features having more weights than distinctive features. Furthermore, common features have larger relative weights than distinctive features, an observation originally made by Tversky (1977). However, distinctive features measured without taking into account the number of times a feature is attributed to an image (e.g. method 3) were not significant predictors of similarity. In general, there is empirical and theoretical support for these results in the literature (Johnson, 1981, 1986; Tversky, 1977; Tversky & Gati, 1978).

Implications of Research Findings

Research findings in this study have implications for both researchers trying to better understand the nature of the relationships between perceived features of objects and their similarity as well as designers of information representation and retrieval systems. The current study is the first to test Tversky's (1977) contrast model, in the context of representation and retrieval of image documents, using images as materials and structural equation modeling techniques. Results of the study will provide the foundations for future research that will attempt to test the Model using not only images, but also other types of objects as stimuli.

The study has methodological implications as well, especially the scaling/measurement of common and distinctive features and of perceived similarity. All observed variables in the study had high reliability coefficients (Cronbach's $\alpha > .84$). The fact that reliable scaling procedures were developed and used in this study is a significant methodological contribution to the literature. These procedures should be useful for future research that involves the scaling/measurement of common and distinctive features as well as similarity of images and other types of stimuli.

Results of the study confirmed that the contrast model explains human similarity judgments of images well. This has image representation and retrieval (more specifically similarity matching) implications. Given the fact that most of the current vector-space-model-based information retrieval systems use similarity measures derived from geometric models of similarity, despite their weaknesses, a similarity measure based on the contrast model would be a viable alternative. Moreover, a representation model mainly based on sets of features of documents would serve as an alternative to the vector-space model.

Two characteristics set a retrieval system where documents are represented by sets of features and similarity matching of document representations (or document and query representations) based on the contrast model and systems where the vector-space model document representation and similarity measures based on geometric models of similarity (e.g. angle (cosine) measure, Euclidean distance, etc.) are adopted apart. The first distinction is while the first type of retrieval systems considers individual features of documents as discrete elements of a feature set; the latter consider them as an array of numbers or a vector. The second distinction is that while similarity measures based on the contrast model would give more weight to common features than distinctive features of documents, similar to how people rate similarity of objects, and similarity is determined through a linear combination of measures of common and distinctive features, similarity measures based on geometric models of similarity do not weigh common and distinctive features the same way.

Both alternative methods of representation and similarity matching may prove to be useful and effective. However, any similarity measure based on the contrast model needs to take into account the fact that common features carry more weights than distinctive features. Alternative image representation methods and similarity measures have been explored in the past (Rubner, 1999; Santini & Jain, 1999). However, extensive evaluations of these methods will be required. In order for these evaluations to be fruitful, the role of the end user must be considered.

Recommendations for Future Research

This study used the minimum required number of two observed variables for each latent variable in the model. Even if this is not a major handicap, researchers who attempt to test the contrast model need to develop and introduce more measures of common and distinctive features as well as more scales for the measurement of similarity of pairs of objects.

Subsequent studies should attempt to test the contrast model using samples of materials (images and other types of information objects) and participants (users of images and other information objects) selected from well-defined populations (for instance, a standard test collection of images and a community of image users such as photo journalists, etc.) in order to validate results of this study.

The next logical step seems to be the design, development, and testing/evaluation of document representation methods and similarity measures based on the contrast model in a functional information retrieval environment. Testing may involve comparisons of the contrast model-based representation methods and similarity measures to existing methods and measures such as the vector-space model and the angle (cosine) and distance-based similarity measures. Traditional retrieval effectiveness and efficiency measures such as recall and precision as well as alternative metrics could be used for testing/evaluation purposes. Furthermore, if feature set based document (especially image document) representations are to be adopted by retrieval systems, there is a need for thorough investigations into the types and levels of attributes/features not only perceived but also used during searching by people/users.

Summary

A short summary of results and research findings is presented in this chapter. All research findings of the study support Tversky's (1977) contrast model, which depicts similarity judgment as a feature contrast task and equates it to a linear combination of common and distinctive features. Common features of images were found to have a higher predictive power (weight) on their similarity than did distinctive features.

The chapter also presents limitations of the study, concluding remarks, implications of research findings, and recommendations for future research. Limitations of the study are mainly related to the number of observed variables of each latent variable in the contrast model and to sampling of materials (images) and participants (image users). Concluding remarks concern the nature of the relationships between common and distinctive features of images and their similarity as well as the relative weights of common and distinctive features as predictors of similarity of images.

Implications of research findings are discussed in terms of alternative image representation methods and similarity measures based on the contrast model. Methodological implications of the study are also discussed. Finally, recommendations are made for future research with respect to testing the contrast model and evaluation of representation methods and similarity measures based on the contrast model, including comparisons of these methods and similarity measures with other methods and similarity measures, in a functional information retrieval environment.

APPENDIX A

IMAGES USED IN THE STUDY

1*

2*

3*

4*

5*

6*

7*

8*

9*

10*

11*

12*

13*

14*

15*

***O'Connor & Wyatt (2004) (Used with permission**

16*

17*

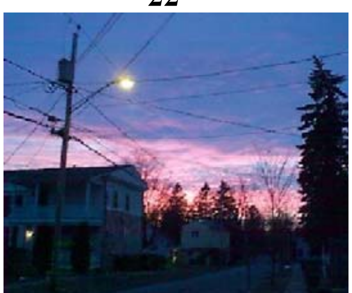
18*

19*

20*

21*

22*



23*

24*

25*

26*

27*

28*

29*

30*

***O'Connor & Wyatt (2004) (Used with permission**

APPENDIX B

INSTRUCTIONS FOR THE IMAGE DESCRIPTION TASK

INSTRUCTIONS

Thank you for taking your time to participate in our study. The study is trying to answer questions regarding perceived features and similarity of images/pictures. In this session, your task will be to describe a set of 32 images/pictures by listing all possible features/attributes/things that you see (perceive) in each image. Please separate adjacent features/attributes/things in your list by a semicolon (*e.g. pretty woman; dog; jumping girl; etc.*). You will have a maximum of 90 seconds (one and half minutes) for each image and if you cannot think of anything more to describe in the image before the 90 seconds lapse, go to the next image by clicking on the 

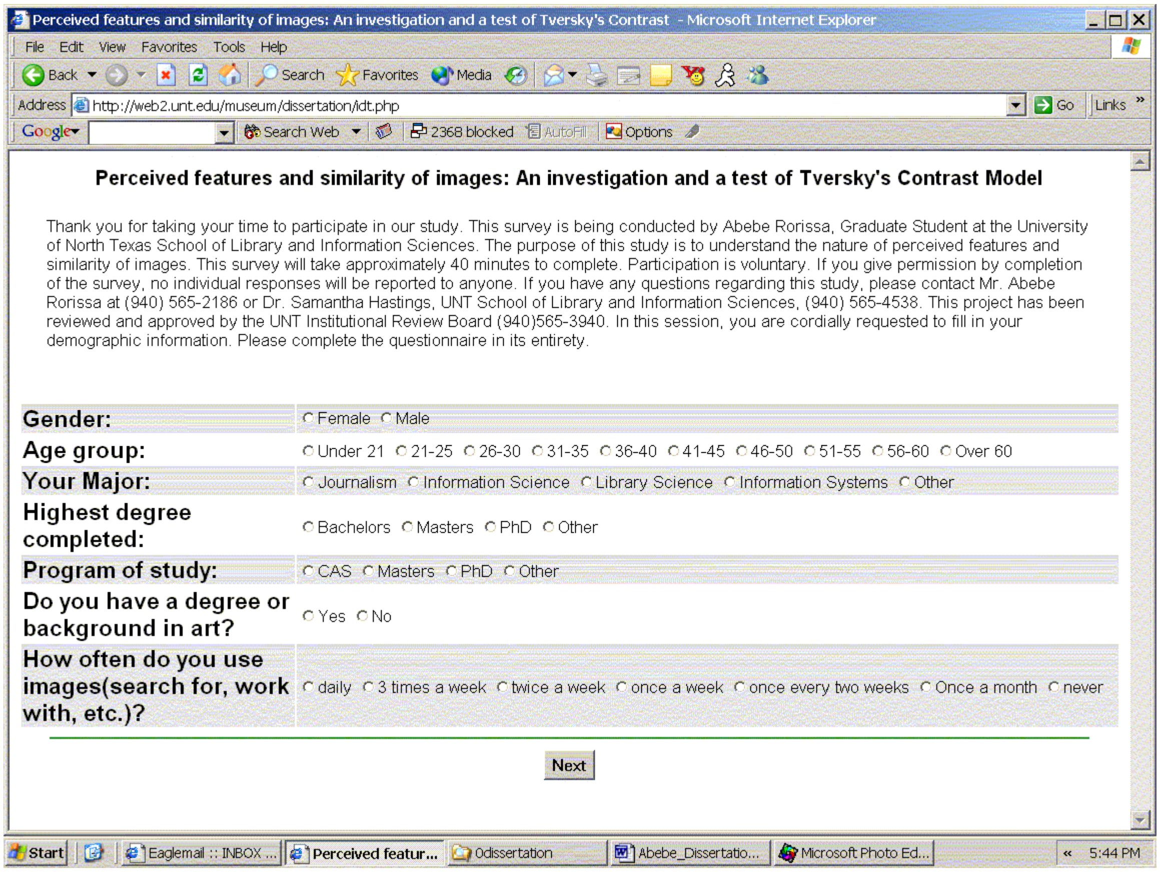
button. Please note that there are no right or wrong answers. Let me know if you need more explanation. The Institutional Review Board (IRB) of the University of North Texas has approved this study and your responses will be anonymous and confidential. Let me know if you have any questions. When you are ready to start the image description task, please click on

.

APPENDIX C

WEB-BASED FORM FOR THE IMAGE DESCRIPTION TASK

IMAGE DESCRIPTION TASK



Perceived features and similarity of images: An investigation and a test of Tversky's Contrast - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://web2.unt.edu/museum/dissertation/idt.php

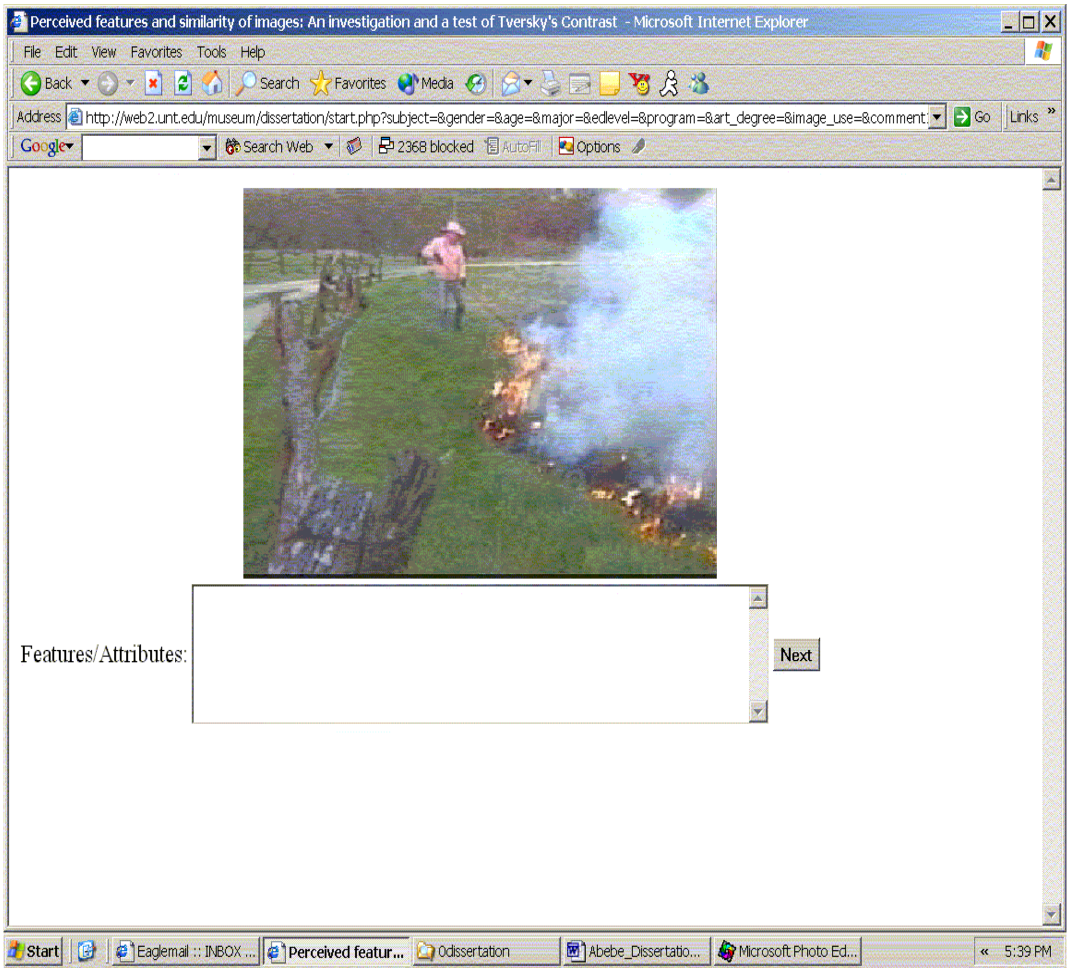
Perceived features and similarity of images: An investigation and a test of Tversky's Contrast Model

Thank you for taking your time to participate in our study. This survey is being conducted by Abebe Rorissa, Graduate Student at the University of North Texas School of Library and Information Sciences. The purpose of this study is to understand the nature of perceived features and similarity of images. This survey will take approximately 40 minutes to complete. Participation is voluntary. If you give permission by completion of the survey, no individual responses will be reported to anyone. If you have any questions regarding this study, please contact Mr. Abebe Rorissa at (940) 565-2186 or Dr. Samantha Hastings, UNT School of Library and Information Sciences, (940) 565-4538. This project has been reviewed and approved by the UNT Institutional Review Board (940)565-3940. In this session, you are cordially requested to fill in your demographic information. Please complete the questionnaire in its entirety.

Gender:	Female · Male
Age group:	Under 21 · 21-25 · 26-30 · 31-35 · 36-40 · 41-45 · 46-50 · 51-55 · 56-60 · Over 60
Your Major:	Journalism · Information Science · Library Science · Information Systems · Other
Highest degree completed:	Bachelors · Masters · PhD · Other
Program of study:	CAS · Masters · PhD · Other
Do you have a degree or background in art?	Yes · No
How often do you use images(search for, work with, etc.)?	daily · 3 times a week · twice a week · once a week · once every two weeks · Once a month · never

Next

.


Perceived features and similarity of images: An investigation and a test of Tversky's Contrast - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://web2.unt.edu/museum/dissertation/start.php?subject=&gender=&age=&major=&edlevel=&program=&art_degree=&image_use=&comment
Go Links
Google Search Web 2368 blocked AutoFill Options
Features/Attributes:
Next
Start Eaglemail :: INBOX ... Perceived featur... 0dissertation Abebe_Dissertatio... Microsoft Photo Ed...
5:39 PM

APPENDIX D

INSTRUCTIONS FOR THE SIMILARITY JUDGMENT TASK

INSTRUCTIONS

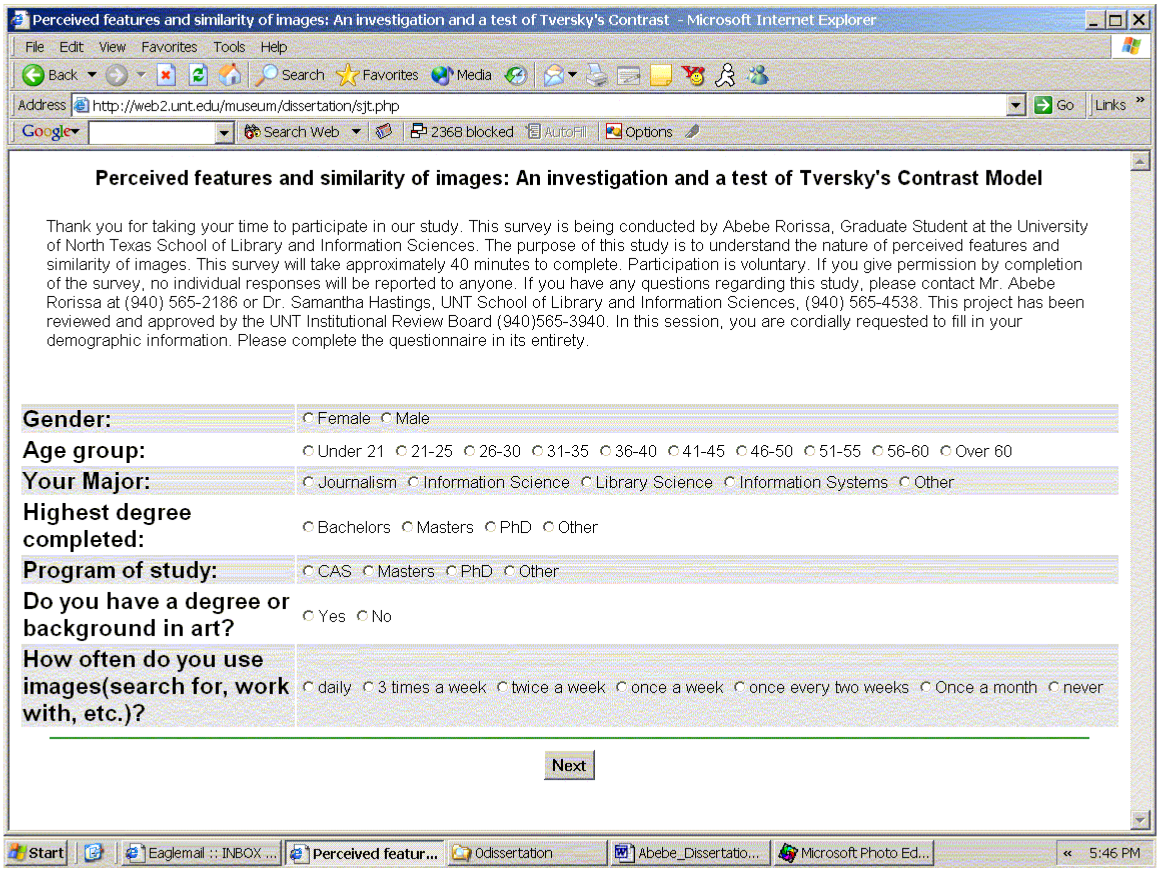
Thank you very much for taking your time to participate in our study. In this session, you are requested to judge the degree of similarity of pairs of images. Before judging the degree of similarity of pairs of images, you will be shown five (5) lines of various lengths. Imagine small, medium and large lines. Imagine also small, medium and large numbers. I would like you to assign numbers to each line in such a way that the larger the number the larger the line, and vice versa. Please assign any positive number (including fractions and decimals) and do not think of any specific unit of measurement (e.g. inches, centimeters). Please try to judge the length of each line independently without comparing it to preceding lines. Your response should be as spontaneous as possible.

After the lines, you will be presented with a series of pairs of images. I would like you to judge how similar the two images in each pair are (using your own criteria for similarity) by clicking on the horizontal line in such a way that the length of the line up to the point where you clicked matches your judgment of the degree of similarity of the pair of images. Please be spontaneous in your response and complete the task in its entirety. You will take a five(5) minutes break halfway through the task. The Institutional Review Board (IRB) of the University of North Texas has approved this study and your responses will be anonymous and confidential. Let me know if you have any questions. When you are ready to start the similarity judgment task, please click on Next.

APPENDIX E

WEB-BASED FORM FOR THE SIMILARITY JUDGMENT TASK

SIMILARITY JUDGMENT TASK


Perceived features and similarity of images: An investigation and a test of Tversky's Contrast - Microsoft Internet Explorer

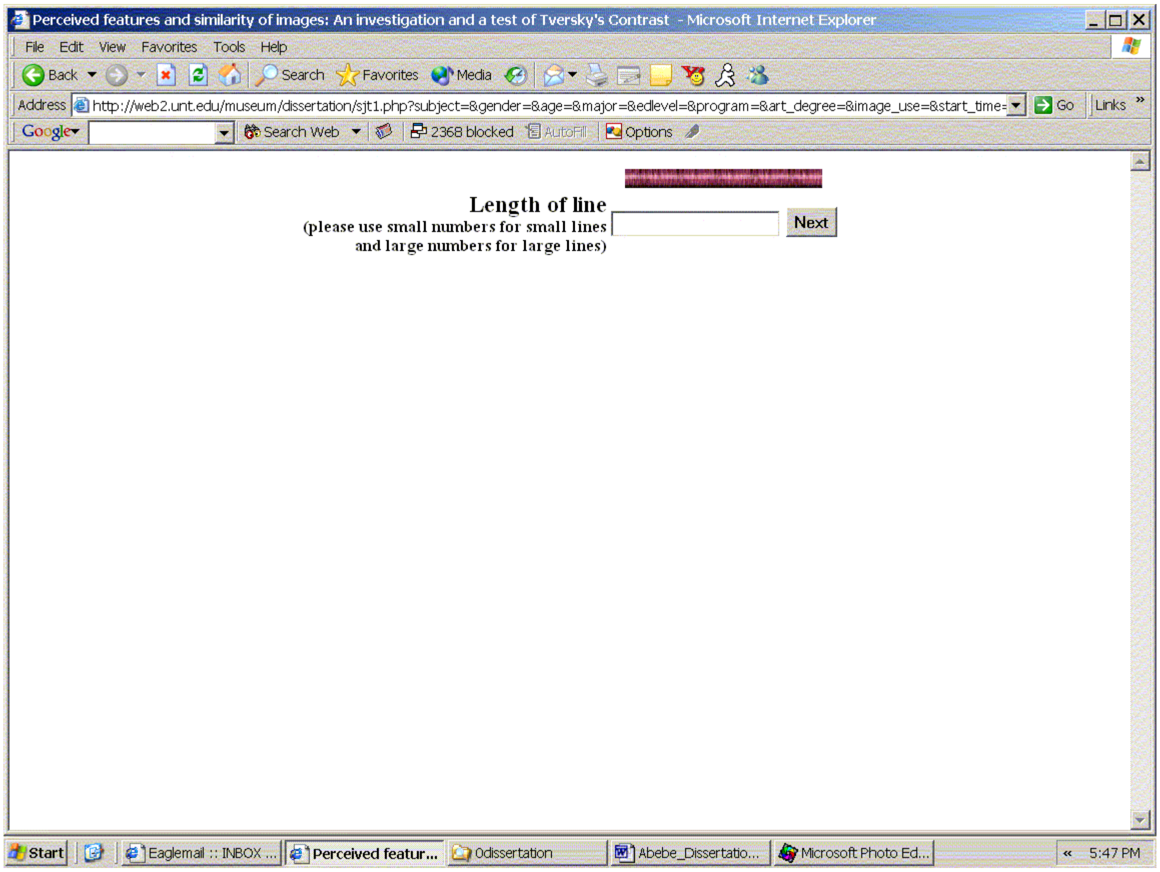
http://web2.unt.edu/museum/dissertation/sjt.php

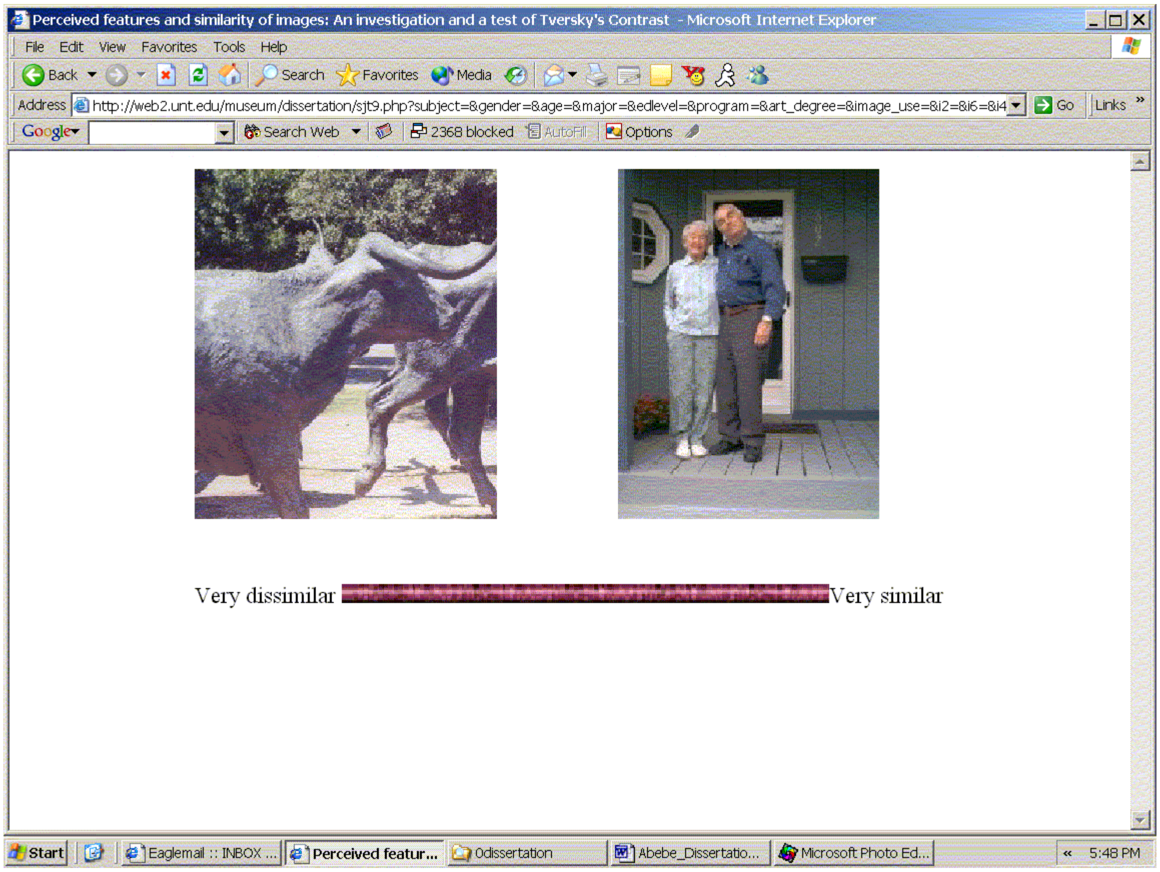
Perceived features and similarity of images: An investigation and a test of Tversky's Contrast Model

Thank you for taking your time to participate in our study. This survey is being conducted by Abebe Rorissa, Graduate Student at the University of North Texas School of Library and Information Sciences. The purpose of this study is to understand the nature of perceived features and similarity of images. This survey will take approximately 40 minutes to complete. Participation is voluntary. If you give permission by completion of the survey, no individual responses will be reported to anyone. If you have any questions regarding this study, please contact Mr. Abebe Rorissa at (940) 565-2186 or Dr. Samantha Hastings, UNT School of Library and Information Sciences, (940) 565-4538. This project has been reviewed and approved by the UNT Institutional Review Board (940)565-3940. In this session, you are cordially requested to fill in your demographic information. Please complete the questionnaire in its entirety.

Gender:	Female Male
Age group:	Under 21 21-25 26-30 31-35 36-40 41-45 46-50 51-55 56-60 Over 60
Your Major:	Journalism Information Science Library Science Information Systems Other
Highest degree completed:	Bachelors Masters PhD Other
Program of study:	CAS Masters PhD Other
Do you have a degree or background in art?	Yes No
How often do you use images(search for, work with, etc.)?	daily 3 times a week twice a week once a week once every two weeks Once a month never

Next


Perceived features and similarity of images: An investigation and a test of Tversky's Contrast - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://web2.unt.edu/museum/dissertation/sjt1.php?subject=&gender=&age=&major=&edlevel=&program=&art_degree=&image_use=&start_time=
Go Links
Google Search Web 2368 blocked AutoFill Options
Length of line
(please use small numbers for small lines
and large numbers for large lines)
Next
Start Eaglemail :: INBOX ... Perceived featur... 0dissertation Abebe_Dissertatio... Microsoft Photo Ed...
5:47 PM


Perceived features and similarity of images: An investigation and a test of Tversky's Contrast - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://web2.unt.edu/museum/dissertation/sjt9.php?subject=&gender=&age=&major=&edlevel=&program=&art_degree=&image_use=&i2=&i6=&i4
Go Links
Google Search Web 2368 blocked AutoFill Options
Very dissimilar
Very similar
Start Eaglemail :: INBOX ... Perceived featur... 0dissertation Abebe_Dissertatio... Microsoft Photo Ed...
5:48 PM

APPENDIX F

INSTRUCTIONS FOR CODERS AND A SAMPLE OF FEATURES ASSIGNED TO A FEATURE CATEGORY

Instructions for Coders

Attached are three documents:

- A set of 30 images (pictures)
- A list of 954 feature bearing terms or phrases which were assigned to the 30 images (image numbers are included together with this list of feature terms) by 75 participants as part of an image description task.
- A list of 39 categories of features.

I would like you to examine each feature term, determine which category it belongs to in the list of categories of features, and then assign the corresponding category number in the "category#" column.

Please do not assign any "category#" if you think that the feature term does not belong to any of the 39 categories.

Please let me know if you have any questions.

Thank you.

Abebe Rorissa

Sample features assigned to a category

Category: Art(ist)/Museum/Sculpture

Art
Art-Gallery
Artist
Art-Museum
Artsy
Artwork
Gallery
Interaction-With-Art
Modern-Art
Modern-Artsy
Modern-Sculpture
Museum
Museum-Or-Public-Park
Portrait
Sculpted
Sculpture
Sculpture-In-The-Park
Sculptures

APPENDIX G

IRB APPROVAL LETTER

UNIVERSITY of NORTH TEXAS

Office of Research Services

April 14, 2004

Abebe Rorissa
School of Library and Information Sciences
University of North Texas

RE: Human Subjects Application No. 04-107

Dear Mr. Rorissa,

Your proposal titled "Perceived Features and Similarity of Images: An Investigation and a Test of Tversky's Contrast Model" has been approved by the Institutional Review Board and is exempt from further review under 45 CFR 46.101. **Federal policy 45 CFR 46.109(e) stipulates that IRB approval is for one year only.**

It is your responsibility according to U.S. Department of Health and Human Services regulations to submit annual and terminal progress reports to the IRB for this project. Please mark your calendar accordingly. The IRB must also review this project prior to any modifications.

Please contact Shelia Bourns, Compliance Administrator, ext. 3940 or Boyd Herndon, Assistant Director for Compliance, ext. 3941, if you wish to make such changes or need additional information.

Sincerely,

Scott Simpkins, Ph.D.
Chair
Institutional Review Board

SS:sb

P.O. Box 305250 • Denton, Texas 76203-5250 • (940) 565-3940
Fax (940) 565-4277 • TTY (800) RELAY TX • www.unt.edu

REFERENCES

Attneave, F. (1950). Dimensions of smilarity. *American Journal of Psychology, 63*, 516-556.

Ben-Shakhar, G., & Gati, I. (1992). The effect of common versus distinctive stimulus features on electrodermal orientation to significant stimuli. *Psychophysiology, 29,* 306-314.

Blair, D.C. (1990). *Language and representation in information retrieval*. Amsterdam: Elsevier.

Bruce, H. (1994). A cognitive view of the situational dynamism of user-centered relevance estimation. *Journal of the American Society for Information Science, 45*(3), 142-148.

Bruce, H. (1998). User satisfaction with information seeking on the Internet. *Journal of the American Society for Information Science, 49*(6), 541-556.

Carson, C., Thomas, M., Belongie, S., Hellerstein, J.M., & Malik, J. (1999). Blobworld: A system for region-based image indexing and retrieval. In D.P. Huijsmans & A.W.M. Smeulders (Eds.), *Proceedings of the Third International Conference on Visual Information Systems.* (pp. 509-516). Amsterdam: Springer.

Chang, S., Smith, J.R., Beigi, M., & Benitez, A. (1997). Visual information retrieval from large distributed online repositories. *Communications of the ACM, 40*(12), 63-71.

Chen, H. (2001a). An analysis of image queries in the field of art history. *Journal of the American Society for Information Science and Technology, 52*(3), 260-273.

Chen, H. (2001b). An analysis of image retrieval tasks in the field of are history. *Information Processing & Management, 37*(5), 701-720.

Choi, Y., & Rasmussen, E.M. (2003). Searching for images: The analysis of users' queries for image retrieval in American history. *Journal of the American Society for Information Science and Technology, 54*(6), 498-511.

Choi, Y., & Rasmussen, E.M. (2002). Users' relevance criteria in image retrieval in American history. *Information Processing & Management, 38*(5), 695-726.

Chu, H. (2001). Research in image indexing and retrieval as reflected in the literature. *Journal of the American Society for Information Science and technology, 52*(12), 1011-1018.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

Djeraba, C., Bouet, M., Briand, H., & Khenchaf, A. (2000). Visual and textual content based indexing and retrieval. *International Journal on Digital Libraries, 2*(4), 269-287.

Dopkins, S., & Ngo, C.T. (2001). The contribution of category relationships to similarity ratings. *Canadian Journal of Experimental Psychology, 55,* 244-252.

Dunn-Rankin, P., Knezek, G., Wallace, S., Zhang, S. (1990). *Scaling methods*. (2nd ed.). Mahwah, NJ; London: Lawrence Erlbaum Associates.

du Toit, M., & du Toit, S. (2001). *Interactive LISREL: User's guide.* Lincolnwood, IL: Scientific Software International.

Eisenberg, M.B. (1986). *Magnitude estimation and the measurement of relevance*. Unpublished Doctoral Dissertation, Syracuse University, Syracuse, NY

Eisenberg, M.B. (1988). Measuring relevance judgments. *Information Processing & Management, 24*(4), 373-389.

Engen, T. (1971). Psychophysics II. Scaling methods. In J. W. Kling & L. A. Riggs (Eds.), *Woodworth & Schlosberg's experimental psychology*. (pp. 47-86). New York: Holt, Rinehart & Winston.

Enser, P.G.B. (2000). Visual image retrieval: Seeking the alliance of concept-based and content-based paradigms. *Journal of Information Science, 26*(4), 199-210.

Ericsson, K.A., & Simon, H.A. (1980). Verbal reports as data. *Psychological Review, 87*(3), 215-251.

Faloutsos, C., Barber, R., Flickner, M., Hafner, J., Niblack, W., Petkovic, D., & Equitz, W. (1994). Efficient and effective querying by image content. *Journal of Intelligent Information Systems, 3*(3/4), 231-262.

Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., & Yanker, P. (1995). Query by image and video content: The QBIC system. *Computer, 28*(9), 23-32.

Gati, I., & Tversky, A. (1984). Weighting common and distinctive features in perceptual and conceptual judgments. *Cognitive Psychology, 16*(3), 341-370.

Gay, R. L., & Diehl, P. L. (1992). *Research methods for business and management.* New York: Macmillan.

Goldstone, R.L. (1999). Similarity. In R. A. Wilson & F. C. Keil (Eds.), *The MIT encyclopedia of the cognitive sciences*. (pp. 757-759). Cambridge, MA: MIT Press.

Goodrum, A., & Spink, A. (2001). Image searching on the Excite Web search engine. *Information Processing & Management, 37*(2), 295-311.

Greisdorf, H., & O'Connor, B.C. (2002a). Modeling what users see when they look at images: A cognitive viewpoint. *Journal of Documentation, 58*(1), 6-29.

Greisdorf, H., & O'Connor, B.C. (2002b). What do users see? Exploring the cognitive nature of functional image retrieval. In E. G. Toms (Ed.), *Proceedings of the 65th Annual Meeting of the American Society for Information Science and Technology*. (pp. 383-390). Medford, NJ: Information Today.

Gudivada, V. N., & Raghavan, V. V. (1997). Modeling and retrieving images by content. *Information Processing & Management, 33,* 427-452.

Gupta, A., Santini, S., & Jain, R. (1997). In search of information in visual media. *Communications of the ACM, 40*(12), 35-42.

Hastings, S.K. (1995). Query categories in a study of intellectual access to digitized art images. In T. Kinney (Ed.), *Proceedings of the 58th Annual Meeting of the American Society for Information Science*. (pp. 3-8). Medford, NJ: ASIS.

Holsti, O.R. (1969). *Content analysis for the social sciences and humanities*. Reading, MA: Addison-Wesley.

Holt, B., Weiss, K., Niblack, W., Flickner, M., & Petkovic, D. (1997). The QBIC project in the Department of Art and Art History at UC Davis. In C. Schwartz & M. Rorvig (Eds.), *Proceedings of the 60th Annual Meeting of the American Society for Information Science.* (pp. 189-195). Medford, NJ: ASIS.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1-55.

Hyperdictionary. (2003). Retrieved October 12, 2003 from: http://www.hyperdictionary.com/.

Jain, A.K., & Vailaya, A. (1996). Image retrieval using color and shape. *Pattern Recognition, 29*(8), 1233-1244.

Johnson, M.D. (1981). Context effects in product perception. In K. B. Monroe (Ed.), *Advances in Consumer Research*. (pp. 112-115). Chicago, IL: Association for Consumer Research.

Johnson, M.D. (1986). Consumer similarity judgments: A test of the contrast model. *Psychology & Marketing, 3*(1), 47-60.

Johnson, M.D., & Horne, D.A. (1988). The contrast model of similarity and comparative advertising. *Psychology & Marketing, 5*(3), 211-232.

Jöreskog, K. G. (1993). Testing structural equation models. In K.A. Bollen & J. S. Long (Eds.), *Testing structural equation models.* (pp. 294-316). Newbury Park, CA: Sage Publications.

Jöreskog, K., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Jöreskog, K. G., & Sörbom, D. (1999). *PRELIS 2.3.* Chicago: Scientific Software International.

Jörgensen, C. (1995). *Image attributes: An investigation*. Unpublished doctoral dissertation, Syracuse University, Syracuse, NY.

Jörgensen, C. (1996). Indexing images: Testing an image description template. In S. Hardin (Ed.), *Proceedings of the 59th Annual Meeting of the American Society for Information Science*. (pp. 209-213). Medford, NJ: Information Today.

Jörgensen, C. (1998). Attributes of images in describing tasks. *Information Processing & Management, 34*(2/3), 161-174.

Jörgensen, C. (2003). *Image retrieval: Theory and research*. Lanham, Md.; Oxford: Scarecrow Press.

Jörgensen, C., Jaimes, A., Benitez, A.B., & Chang, S.F. (2001). A conceptual framework and empirical research for classifying visual descriptors. *Journal of the American Society for Information Science and Technology, 52*(11), 933-947.

Kelloway, E. K. (1998). Using LISREL for structural equation modeling: A researcher's guide. Thousand Oaks, CA: Sage Publications.

Layne, S.S. (1994). Some issues in the indexing of images. *Journal of the American Society for Information Science, 45*(8), 583-588.

Lee, J.Y. (2001). *An investigation of modes of access for the retrieval of emotional content of pictures*. Unpublished doctoral dissertation, Syracuse University, Syracuse, NY.

Lynch, C.A. (1991). The technologies of electronic imaging. *Journal of the American Society for Information Science, 42*(8), 578-585.

Markkula, M., Tico, M., Sepponen, B., Nirkkonen, K., & Sormunen, E. (2001). A test collection for the evaluation of content-based image retrieval algorithms-A user and task-based approach. *Information Retrieval, 4*(3-4), 275-293.

Medin, D.L., Goldstone, R.L., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100*(2), 254-278.

Melara, R.D. (1992). The concept of perceptual similarity: From psychophysics to cognitive psychology. In Algom. D. (Ed.), *Psychophysical Approaches to Cognition*. (pp. 303-388). Amsterdam: North-Holland.

Mohler, P. Ph., & Zeull, C. (1998). *TEXTPACK PC.* Mannheim: ZUMA.

Mojsilović, A., Kovacevic, J., Kall, D., Safranek, R.J., & Ganapathy, S.K. (2000). The vocabulary and grammer of color patterns. *IEEE Transactions on Image Processing, 9*(3), 417-431.

Mojsilović, A., & Rogowitz, B. (2001). Capturing image semantics with low-level descriptors. In: *Proceedings of the International Conference on Image Processing, ICIP 2001*. Thessaloniki, Greece.

Mostafa, J. (1994). Digital image representation and access. In M. E. Williams (Ed.), *Annual Review of Information Science and Technology*. (pp. 91-135). Medford, NJ: Learned Information.

Neuendorf, K. A. (2002). *The content analysis guidebook.* Thousand Oaks, CA: Sage Publications.

Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric theory*. New York: McGraw-Hill.

O'Connor, B.C. (1996). *Explorations of indexing & abstracting: Pointing, virtue, and power*. Englewood, CO.: Libraries Unlimited.

O'Connor, B.C., O'Connor, M.K., & Abbas, J.M. (1999). User reactions as access mechanism: An exploration based on captions for images. *Journal of the American Society for Information Science and Technology, 50*(8), 681-697.

O'Connor, B.C., & Wyatt, R.B. (2004). *Photo provocations: Thinking in, with, and about photographs*. Lanham, Md.; Oxford: Scarecrow Press.

Panofsky, E. (1955). *Meaning in the visual arts: Papers in and on art history*. Garden City, N.Y.: Doubleday.

Qin, J. (2000). Semantic similarities between a keyword database and a controlled vocabulary database: An investigation in the antibiotic resistance literature. *Journal of the American Society for Information Science, 51*(3), 166-180.

Rasmussen, E.M. (1997). Indexing images. In M. E. Williams (Ed.), *Annual Review of Information Science and Technology*. (pp. 169-196). Medford, NJ: Learned Information.

Rogowitz, B.E., Frese, T., Smith, J.R., Bouman, C.A., & Kalin, E. (1998). Perceptual image similarity experiments. In B. E. Rogowitz & T. N. Pappas (Eds.), *Human Vision and Electronic Imaging III, Proceedings of the SPIE*. (pp. 576--590 San Jose, CA:

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General, 104*(3), 192-233.

Rosch , E., & Lloyd, B.B. (1978). *Cognition and categorization*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Rosch, E., & Mervis, C. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*(4), 573-605.

Rosch, E., Mervis, C.B., Gray, W.D., Johnson, D.M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*(3), 382-439.

Ross, R.T. (1934). Optimum orders for the presentation of pairs in the method of paired comparisons. *Journal of Educational Psychology, 25*, 375-383.

Rubner, Y. (1999). *Perceptual metrics for image database navigation*. Unpublished doctoral dissertation, Stanford University, Stanford, CA.

Rubner, Y., Guibas, L.J., & Tomasi, C. (1997). The earth mover's distance, multi-dimensional scaling, and color-based image retrieval. In: *Proceedings of the ARPA Image Understanding Workshop*. (pp. 661-668). New Orleans, LA.

Rubner, Y., Puzicha, J., Tomasi, C., & Buhmann, J.M. (2001). Empirical evaluation of dissimilarity measures for color and texture. *Computer Vision and Image Understanding, 84*(1), 25-43.

Rubner, Y., Tomasi, C., & Guibas, L.J. (2000). The Earth Mover's Distance as a metric for image retrieval. *International Journal of Computer Vision, 40*(2), 99-121.

Rubner, Y., Tomasi, C., & Guibas, L.J. (1998). A metric for distributions with applications to image databases. In *Proceedings of the 1998 IEEE International Conference on Computer Vision*. (pp. 59-66). Bombay, India.

Rui, Y., Huang, T.S., & Chang, S.F. (1999). Image retrieval: current techniques, promising directions and open issues. *Journal of Visual Communication and Image Representation , 10*(1), 39-62.

Rui, Y., Ortega, M., Huang, T.S., & Mehrotra, S. (1999). Information retrieval beyond the text document. *Library Trends, 48*(2), 455-474.

Santini, S., & Jain, R. (1999). Similarity measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 21*(9), 871-883.

Schonlau, M., Fricker, R. D., & Elliott, M. N. (2001). *Conducting Research Surveys via E-mail and the Web.* Santa Monica, CA: Rand.

Schumacker, R.E., & Lomax, R.G. (1996). *A beginner's guide to structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum Associates.

Shatford, S. (1986). Analyzing the subject of a picture: A theoretical approach. *Cataloguing and Classification Quarterly, 6*(3), 39-62.

Shepard, R.N. (1962a). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika, 27*(2), 125-140.

Shepard, R.N. (1962b). The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika, 27*(3), 219-246.

Shepard, R.N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology, 1*, 54-87.

Sloman, S.A., & Rips, L.J. (1998). Similarity as an explanatory construct. In S. A. Sloman & L. J. Rips (Eds.), *Similarity and symbols in human thinking*. (pp. 1-15). Cambridge, Massachusetts: MIT Press.

Smith, J.R., & Chang, S.F. (1996). Visualseek: A fully automated content-based image query system. In *Proceedings of ACM Multimedia 96*. (pp. 87-98). New York: ACM Press.

Stan, D., & Sethi, I.K. (2001). Mapping low-level image features to semantic concepts. In: *Proceedings of SPIE: Storage and Retrieval for Media Databases*. (pp. 172-179). San Jose, CA: SPIE.

Stevens, S.S. (1956). The direct estimation of sensory magnitudes - loudness. *American Journal of Psychology, 69*(1), 1-25.

Stevens, S.S. (1966). A metric for the social consensus. *Science, 151*(3710), 530-541.

Stevens, S.S. (1975). *Psychophysics: Introduction to its perceptual, neural, and social prospects*. New York: Wiley.

Stricker, M., & Orengo, M. (1995). Similarity of color images. In W. Niblack & R. Jain (Eds.), *Proceedings of SPIE Conference on Storage and Retrieval for Image and Video Databases III, volume 2420*. (pp. 381-392). San Diego, CA:

Tabachnick, B. G. & Fidell, L. S. (1989). Using multivariate statistics. (2nd ed.). New York: HarperCollins.

Thurstone, L.L. (1927). A law of comparative judgment. *Psychological Review, 34*, 273-286.

Turner, J.M. (1994). *Determining the subject content of still and moving image documents for storage and retrieval: An experimental investigation*. Unpublished doctoral dissertation, University of Toronto, Toronto, Canada.

Turner, J.M. (1995). Comparing user-assigned terms with indexer-assigned terms for storage and retrieval of moving images: Research results. In T. Kinney (Ed.), *Proceedings of the 58th Annual Meeting of the American Society for Information Science*. (pp. 9-12). Medford, NJ: ASIS.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*(4), 327-352.

Tversky, A., & Gati, I. (1978). Studies of similarity. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization*. (pp. 79-98). Hillsdale, NJ: Lawrence Erlbaum Associates.

Tversky, A., & Gati, I. (1982). Similarity, separability, and the triangle inequality. *Psychological Review, 89*(2), 123-154.

Ulhaque, E., & Bahn, K.D. (1992). A spreading activation model of consumers' asymmetric similarity judgment. *Advances in Consumer Research, 19*(1), 782-786.

Umbaugh, S.E. (1998). *Computer vision and image processing: A practical approach using CVIPtools*. Upper Saddle River, NJ: Prentice Hall.

Weber, R.P. (1990). *Basic content analysis*. (2nd ed.). Newbury Park, CA.: Sage Publications.

Weisberg, H. F. (1992). *Central tendency and variability.* Newbury Park: Sage Publications.

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R.H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues and applications.* (pp. 56-75). Thousand Oaks, CA: Sage Publications.

Zachary, J.M. (2000). *An information theoretic approach to content based image retrieval*. Unpublished doctoral dissertation, Louisiana State University and Agricultural & Mechanical College, Baton Rouge, Louisiana.

Zachary, J., & Iyengar, S.S. (2001). Information theoretic similarity measures for content based image retrieval. *Journal of the American Society for Information Science and Technology, 52*, 856-867.

Zachary, J., Iyengar, S.S., & Barhen, J. (2001). Content based image retrieval and information theory: A general approach. *Journal of the American Society for Information Science and Technology, 52*, 840-852.

Zhang, J., & Korfhage, R.R. (1999a). A distance and angle similarity measure method. *Journal of the American Society for Information Science, 50*(9), 772-778.

Zhang, J., & Korfhage, R.R. (1999b). DARE: Distance and angle retrieval environment: A tale of the two measures. *Journal of the American Society for Information Science, 50*(9), 779-787.

Zhang, J., & Rasmussen, E.M. (2001). Developing a new similarity measure from two different perspectives. *Information Processing & Management, 37*(2), 279-294.

Zhu, B., & Chen, H. (2000). Validating a geographical image retrieval system. *Journal of the American Society for Information Science, 51*(7), 625-634.