Exploring the Relationship Between Feature and Perceptual Visual Spaces

Abebe Rorissa
Department of Information Studies, University at Albany, State University of New York, Draper Hall, Room 113, 135 Western Avenue, Albany, NY 12222. E-mail: arorissa@albany.edu

Paul Clough
Department of Information Studies, University of Sheffield, Sheffield, S1 4DP, United Kingdom. E-mail: p.d.clough@sheffield.ac.uk

Thomas Deselaers
Human Language Technologies and Pattern Recognition Group, RWTH Aachen University, Computer Science Department, 52056 Aachen, Germany. E-mail: deselaers@informatik.rwth-aachen.de

The number and size of digital repositories containing visual information (images or videos) is increasing and thereby demanding appropriate ways to represent and search these information spaces. Their visualization often relies on reducing the dimensions of the information space to create a lower-dimensional feature space which, from the point-of-view of the end user, will be viewed and interpreted as a perceptual space. Critically for information visualization, the degree to which the feature and perceptual spaces correspond is still an open research question. In this paper we report the results of three studies which indicate that distance (or dissimilarity) matrices based on low-level visual features, in conjunction with various similarity measures commonly used in current CBIR systems, correlate with human similarity judgments.

Introduction

The ubiquity of computers and digital cameras has lead to rapid growth in collections of digital texts and images (Lyman & Varian, 2003), requiring appropriate methods to access relevant information. In the last years, researchers and system developers have focused on developing search and browse tools with information visualization capabilities. Research has shown that effective browsing and information visualization can assist users in finding relevant images (Laine-Hernandez & Westman, 2006; Rodden, Basalaj, Sinclair, & Wood, 1999; Rodden, Basalaj, Sinclair, & Wood, 2000; Rodden, Basalaj, Sinclair, & Wood, 2001).

Received August 23, 2007; revised October 18, 2007; accepted October 19, 2007

 DOI: 10.1002/asi.20792

There are various types, as well as levels, of features that can be used to represent an image identified in the literature of both concept-based and content-based image retrieval (see, e.g., Enser, 2000; Greisdorf, & O'Connor, 2002a; Jörgensen, 1998; Jörgensen, Jaimes, Benitez, & Chang, 2001; Laine-Hernandez & Westman, 2006). Concept-based retrieval uses assigned free-text or terms from a vocabulary (assigned manually or more recently automatically) to index and retrieve images; whereas Content-Based Image Retrieval (CBIR) uses low-level features derived from the visual content of an image itself (e.g. color, shape and texture).

Content-based image retrieval (CBIR) systems routinely use multidimensional scaling (MDS) and hierarchical clustering for the visualization of both stored and retrieved images (see, e.g., Deselaers, Keysers, & Ney (in press); Fauqueuer & Boujema, 2003; Rubner, 1999; Stan & Sethi, 2003). They do so based upon using matrices to represent and store low-level features, such as color, shape, and texture, in conjunction with various similarity measures to determine interdocument proximity within the visualization space. Goodrum (2001) was among the first to use MDS to study human similarity judgments in an effort to map users' cognitive representations for image similarity by task.

There is a general agreement that humans perceive all levels of image features, from the primitive/syntactic to the highly semantic (Jörgensen et al., 2001). The amount of information contained in an image, as a source of information, or the meaning it conveys to different viewers depends on several factors and is difficult to measure. While low-level image features carry certain information and can easily be extracted using computer vision methods, they are no match for the information a human observer perceives. This has

contributed to the complexity of image indexing and retrieval, in general, and to information visualization, in particular. The disparity in the types of features perceived by humans and the predominantly low-level features utilized by CBIR systems lead some researchers to term this problem the "semantic gap" (Datta, Li, & Wang, 2005; Dori, 2000; Lew, Sebe, Djeraba, & Jain, 2006; Neumann & Gegenfurtner, 2006; Smeulders, Worring, Santini, Gupta, & Jain, 2000).

Despite many CBIR systems providing visualizations of stored and/or retrieved images for human browsing, there are few studies that investigate and compare the visualization of image collections in a perceptual space with the visualization of image collections by CBIR systems in the feature space (Fauqueur & Boujema, 2003; Gupta, Santini, & Jain, 1997; Santini & Jain, 1999; Zhu & Chen, 2000). For many years, calls have been made to investigate this gap, specifically the relationship between visualizing an image collection based on low-level visual features (the feature space), with the judgment of perception and image similarity by human users (the perceptual space; see, e.g., Chen, Gagaudakis, & Rosin, 2000). Rogowitz, Frese, Smith, Bouman, & Kalin (1998) argue that there is a correlation between the visual features of images and their semantic content. If so, we posit that visual features of images and similarity measures used in CBIR should provide similar results as human similarity judges, and the difference between the feature and perceptual spaces of an image collection should not be significant. This study therefore aims at addressing the following important question: To what extent do low-level visual image features and similarity measures used by current CBIR systems correspond to human similarity judgments? This question has become particularly important given the recent interest in visualization and browsing of image collections, as well as the need to bridge the semantic gap.

Another important topic in CBIR is the combination of features. Most descriptors model a particular property of images, and to obtain optimal results, the combination of features is often required. In Yavlinsky, Pickering, Heesch, and Rüger (2004), an automatic learning approach based on known relevance is proposed to obtain a suitable combination of features, and in Müller, Müller, Squire, Marchand-Maillet, and Pun (2000), features weights are obtained from user feedback to an image retrieval system. The approach presented here does not require any relevance judgments, but rather it learns a combination of visual features for similarity comparisons that resembles human perception as closely as possible. We envisage that the results of this study will contribute toward advances being made regarding the nature of human image perception, specifically perceived similarity judgments, and thereby leading to a more informed design of image indexing, retrieval, and visualization.

This article is structured as follows: Relevant literature is presented followed by our research methodology, results and discussion, and possible implications of our findings on the design of image retrieval systems.

Literature Review

General Similarity

Similarity is one of the most important and well-researched constructs in information science because it plays an important role in human perception (Goldstone, 1999; Melara, 1992; Tversky, 1977; Tversky & Gati, 1978) and information organization and retrieval (Santini & Jain, 1999; Zhang & Korfhage, 1999a, 1999b). A major component of any information retrieval (IR) system is similarity matching to determine interdocument similarity and the degree of similarity between a user's information need (represented verbally or visually for image retrieval) and the documents (or surrogates) in a repository. Image similarity from a computational standpoint is investigated in Vasconcelos & Lippmann (2000).

Humans group or categorize objects based upon their degree of similarity, and this judgment by a human is, in part, based on the perception and cognition of an object's features or attributes. Thus, in order to understand similarity as a construct, research should be anchored in the human perception of an object's features or attributes (Melara, 1992) and cognition. This is because our "ability to assess similarity lies close to the core of cognition" (Goldstone, 1999, p. 757).

Geometric models of similarity that equate observed dissimilarities between objects to the metric distances between the points representing these objects on a coordinate space constitute many of similarity measures used in image retrieval. Even though some recent CBIR systems compare images based on models that may use geometric relationships between parts of an image and for visualization purposes, it has been shown that most human similarity judgment data violate the metric axioms of these similarity models (Tversky, 1977). This motivates the need to consider not only the disparities between the feature and perceptual spaces but also continue to find ways and means to bridge them.

Similarity Measures and Information Visualization

Similarity measures are metrics used to quantify interdocument similarity and the relevance of documents in a collection to queries based on proximities between their feature representations. They are widely used for both text retrieval (Qin, 2000; Zhang & Korfhage, 1999a; Zhang & Korfhage, 1999b; Zhang & Rasmussen, 2001) and image retrieval (Deselaers, Keysers, & Ney, 2004; Gupta et al., 1997; Santini & Jain, 1999; Zachary, 2000; Zachary, Iyengar, & Barhen, 2001). While similarity measures used in text retrieval mainly involve term frequencies and weighting schemes, most similarity measures for image retrieval are applied on their low-level features. Many of the similarity measures used in IR are based on the vector-space model (Salton, Wong, & Yang, 1975).

Some of the most popular similarity measures used in text retrieval based on geometric models are the cosine (angle)-based and the distance-based measures (Zhang & Rasmussen, 2001; Zhang & Korfhage, 1999b). In contrast, distance-based similarity measures are the most widely used by CBIR

DOI: 10.1002/asi

systems, the cosine (angle)-based measure having limited use (Gupta et al., 1997). The most widely-used distance-based similarity measures are the City-Block distance[1] and the Euclidean distance (or the L2-norm), two special cases of the Minkowski metric. Other popular similarity measures are the Kullback-Leibler divergence and the Jensen-Shannon-divergence (JSD). In this article, we compare popular visual descriptors and similarity measures used in many of the current CBIR systems to human similarity judgments in order to ascertain whether any of these correspond to a user's perceptual space. We do this by comparing the manual grouping of images in three tasks with clusters generated automatically through the use of low-level features and similarity metrics. Furthermore, we propose a method to determine a combination of visual features that match human perception as closely as possible.

Information visualization has different connotations or meanings to different people from various disciplines. In this article, we refer to the graphical presentation or visual depiction (usually in the form of an *n*-dimensional map through MDS, or a tree map through clustering) of a large document collection (or their surrogates) as information visualization. In this sense, it is mainly based on one or more of the above-mentioned similarity measures. Both *n*-dimensional and tree maps place similar documents close to each other while placing dissimilar documents further apart from each other. The document collection in the visualization is sometimes referred to as an "information space," even though the term, according to the cognitive IR theory, is meant to include other components of an IR system such as the representation of documents and information needs and indexing systems (Ingwersen, 1992, 1996).

Cognitive Theory for Information Retrieval

A comprehensive view of user interaction with an IR system has been addressed by cognitive IR theory (Ingwersen, 1992, 1996). This theory draws on a number of ad hoc IR theories and approaches from all facets of information science. This theory views the creation and reception of information, by both human and machine alike, as acts of information processing and is contrary to the view that only humans are recipients of data and information (Ingwersen, 1996). A cognitive view of information also goes beyond just meaning. For instance, an image may carry different semantic value to different viewers or recipients who may provide as many different interpretations as the semantic values depending on their situation and context (Ingwersen, 1996). In other words, from the point of the cognitive view of the user, the image presents a message—information—and the meaning varies for different viewers because an image may carry different semantic values.

According to cognitive IR theory, user interaction consists of cognitive processes on the part of the user. Within this general framework, Ingwersen (1996) formulates the global model of polyrepresentation in IR with two sets of elements: the cognitive space and the information space (for a recent, albeit slightly different take on these two elements, please see Newby, 2001). Elements in the cognitive space include the user's information need, problem space, work task or interest, and dominant work domain(s), while elements in the information space are mainly various representations of semantic entities (e.g., documents and their surrogates). In this article, we refer to the visualization of a collection of images (e.g., a an MDS configuration or map of the information space, representation or collection of images using feature vectors) that depicts similarity between the images as the feature space (closer in meaning to document space); while we use the term perceptual space to refer to a graphical map of similarity for an image collection as judged by human subjects. We do not use the term "cognitive space" because it is much broader than our definition of perceptual space.

Perceptual Image Similarity

Investigating the relationships between human image similarity and approaches used in CBIR is by no means new; on the contrary, this has long been recognized as a core problem in image retrieval. Methods for extracting and comparing low-level features that correspond more closely to human image similarity are more likely to satisfy the end users of image retrieval systems (Neumann & Gegenfurtner, 2006). However, as Li, Chang, and Wu (2003, p. 512) state, "Quantifying perceptual similarity is a difficult problem. Indeed we may well be decades away from fully understanding how human perception works." Our study aims to complement the existing literature on image retrieval and contribute to an understanding of visual perception.

Rogowitz et al. (1998) conducted two psychological scaling experiments on a set of 97 digital photos (on a wide range of topics), comparing human similarity perception with two image similarity metrics. MDS techniques were used to investigate the characteristics of human similarity perception based on two tasks: (a) arranging images so that those perceived more similar were placed physically closer together (table scaling) and (b) assigning a numeric value to a pair of images to indicate perceived similarity (computer scaling). Results from these experiments showed that humans use many dimensions to evaluate image similarity, including color, luminance and semantic information, and similarity values, were used to inform the use of MDS to create an intuitive navigation space for images.

Li et al. (2003) report a perceptual distance function for measuring image similarity which is independent from human observers. Their distance function, the dynamic partial function (DPF), seeks to activate different low-level features for different object pairs, which they argue relate strongly to the findings of cognitive psychology. Their measure uses the assumption that similar images may be represented by different weightings of image features (i.e., not all features

[1]Alternatively referred to as the Manhattan distance, Hamming distance or L1-norm.

DOI: 10.1002/asi

contribute equally between similar images). They make use of six image features and compare DPF with a number of existing distance functions (Euclidean, Cosine and L_1). Their evaluation consists of applying transformations (that human perception is known to be invariant) to a collection of images, and they measure success based on retrieval of as many transformed images as possible. The aim of this approach to evaluating perceptual image similarity was to reduce the effects of subjective decisions that are inherent in performing human similarity judgments.

Neumann and Gegenfurtner (2006) evaluated a simple CBIR system, developed based on an understanding of known properties in human vision. Their evaluation consisted of a two-alternative forced-choice (2AFC) design in which 900 query images were selected from the Corel database and two best matching images (retrieved automatically) presented to the user (15 undergraduates) to select the image most similar to the query image. Results showed that the psychologically based image indexes retrieved images judged to be more similar to the query than other approaches. Squire and Pun (1998) also compared the human clustering of images with feature-driven machine clustering of images and found that the human clusters differed strongly among each other, but that the methods for automatic clustering disagreed to an even higher degree. Greisdorf and O'Connor (2002b) also found high disagreement among individuals asked to make piles of images.

Our study is similar to this previous work in that we also aim to explore the relationships between the feature and perceptual space. Our work is most similar to that of Neumann and Gegenfurtner (2006) in that we measure human similarity directly. However, in addition, we use a larger set of image features and specifically quantify the correlation between human image similarity and computed similarities from the feature space from a larger number of participants (and tasks). In addition to evaluating the contribution of individual features, we also propose a method to combine several feature-based similarity measurements to obtain one that matches human similarity judgments as closely as possible. To the best of our knowledge, this has not been reported in past literature.

Methodology

Three studies were conducted between March 2003 and November 2006. *Studies 1 & 2*, conducted between March 2003 and November 2006, used an approach of free-sorting (Coxon, 1999) for data collection. Participants were asked to categorize two separate random samples of 50 images into groups of similar images without constraints on the time taken for categorization and the number of categories created. *Study 3*, conducted between June and October 2004, was different from the first two in that instead of free-sorting, human similarity judgment data were obtained through direct magnitude estimation of pair-wise similarity for a random sample of 30 images.

Materials

A total of 130 images were used in the three studies. For Studies 1 & 2, a separate, random sample of 50 color images was selected from disc number 6 of the *Hemera Photo Objects Volume I*, a stock photo collection (http://www.hemera.com). These images are from the "people" category and each one was printed on a 4-by-5-inch (10.2-by-12.7-cm) card and given to participants. A random sample of 30 color photographs of varying subjects taken by O'Connor and Wyatt (2004) served as materials for Study 3.

Participants

Participants in the three studies were 180 volunteer graduate students at two major U.S. universities (one in the Southwest and the other in the Northeast). Thirty of those participated in Study 1 (16 female and 14 male), 75 in Study 2 (59 female and 16 male), and the remaining 75 in Study 3 (49 female and 26 male). All participants were between the ages of 21 and 60 years old.

Procedure

Human Similarity Judgments and Similarity/Dissimilarity Matrices. Participants of Studies 1 & 2 were instructed to first inspect the images and then to sort them into as many groups (or categories) as they wished, using their own general criteria for similarity. Participants were free to rearrange, break, or remake the groups until they reached an arrangement (or visualization) that was satisfactory to them. The cards were reshuffled before they were given to the next participant. Participants of Study 1 formed between 3 and 7 groups and the mean, median, mode, and standard deviation of the number of groups of images formed was 8, 7, 7 and 3.3, respectively. Participants of Study 2 formed a minimum of 2 and a maximum of 24 groups, while the mean, median, mode, and standard deviation of the number of groups of images formed by them were 8.79, 8, 9 and 4.1, respectively. Results from analyzing the manual clustering of images and terms used to label groups of images revealed that people tend to use superordinate and interpretive terms more than terms that are at the basic level of abstraction as well as those that describe perceptual image features (Rorissa & Iyer, in press).

Sorting data were aggregated, over all participants, to a similarity matrix using a widely used measure of similarity for sorting data, namely *percent overlap* (Dunn-Rankin, Knezek, Wallace, & Zhang, 2002). The *percent overlap* for two images i and j is simply the ratio of the number of participants who put both i and j in the same group during sorting to the total number of participants. Because percent overlap is a measure of similarity (the higher the value the more similar the pair of stimuli are), entries of the corresponding dissimilarity matrix were computed using $\delta_{ij} = max - S_{ij}$, where max is 1, and S_{ij} is the percent overlap for images i and j. To measure the reliability (internal consistency) of the participants' sortings, we used Jaccard's Coefficient. In order to compute the coefficient, we randomly

DOI: 10.1002/asi

divided the number of participants in each of the two studies into two groups. Calculated Jaccard's Coefficient values range from 0 (or no consistency) to 1 (or maximum consistency) and coefficients for the two studies were 0.76 and 0.79, respectively, an indication of a strong internal consistency (or reliability).

An e-mail message, with the URL for a similarity judgment task and a unique identifier, was sent to each of the participants of Study 3 between June and October 2004. A follow-up e-mail message was sent to participants who did not complete the task after two weeks from the date the first e-mail message was sent. After reading the instructions, participants were presented with a Web-based form for each of two sets of pairs of the 30 images (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, 1975). Magnitude estimation (with no modulus) was used where participants used a horizontal line (5 inches long and 1/5 inch thick: in data analysis, a length of 1 inch represents 100 units) to indicate the degree of similarity of pairs of images.

Two sets of 435 pairs of the 30 images (we will refer to them as *SIMAB & SIMBA*, where A and B are two images; the second set, SIMBA, was obtained by reversing the order of pairs in the first set as well as the order of images in each pair) were judged by the participants of the task and pairs of images were presented in the same order for all the participants. The participants took a mandatory five-minute break between the two sets in order to minimize the fatigue effect due to the large number of pairs of images. 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.

Reliability (internal consistency of similarity judgments by participants of the two sets of images, SIMAB and SIMBA) was assessed using Cronbach's (1951) coefficient alpha (α). Alpha values were 0.965 and 0.963, respectively, for the two sets (SIMAB and SIMBA), which are well above the recommended threshold (0.70). Two similarity matrices (one for each set of 435 pairs) were formed. Each entry or element of the two similarity matrices was determined by taking the logarithms of the raw magnitude estimations provided by all participants of the similarity judgment task and then aggregated using the geometric means of the logarithms of the magnitude estimations. Entries or elements of the corresponding dissimilarity matrices were computed using $\delta_{ij} = max - S_{ij}$, where max is 2.54714 and 2.48158 for SIMAB and SIMBA, respectively, and S_{ij} is the corresponding entry or element in the similarity matrices for the pair of images i and j.

Feature Extraction and Similarity/Distance Matrices

The same images that were used for the studies described above were also compared using eight low-level visual descriptors. These descriptors were extracted automatically and an appropriate distance metric computed based on pairwise comparison between all images. In the following text, we provide a short description of each low-level descriptor used in these experiments, describe the corresponding similarity/dissimilarity measures used, and refer to related work on visual features. An overview of the used descriptors and corresponding distance measures is given in Table 2.

Appearance-based image descriptor. The simplest approach is to directly use the pixel values of an image as features: Images are scaled to a common size and compared using Euclidean distance. In optical character recognition (OCR) and for medical data, improved methods based on these image features usually obtain excellent results (Keysers, Deselaers, Gollan , & Ney , 2007). In this work, we have used 32×32 pixel versions of the images, compared using Euclidean distance. It has been observed that for classification and retrieval of medical radiographs, this method serves as a reasonable baseline (Deselaers, Müller, Clough, Ney, & Lehmann, 2007).

Color histograms. Widely used in image retrieval (Deselaers et al., (in press); Faloutsos et al., 1994; Puzicha, Rubner, Tomasi, & Buhmann, 1999; Smeulders et al., 2000; Swain & Ballard, 1991), color histograms are among the most basic of approaches. To demonstrate performance improvements in algorithms for image retrieval, systems using only color histograms are often used as a baseline. The color space is divided into partitions, and for each partition, pixels with a color within its range are counted. This results in a representation of the relative frequencies of occurring colors. We use the Red, Green, and Blue (RGB) color space for histograms and observed only minor differences with other color spaces (also observed in Smith and Chang (1996). In accordance with Puzicha et al. (1999), we used the Jensen Shannon divergence to compare histograms.

Global texture descriptor. In Deselaers et al. (2004), a texture descriptor consisting of several parts is described. Fractal dimension measures the roughness or the crinkliness of a surface and it is calculated using the reticular cell counting method (Haberäcker, 1995). Coarseness characterizes the grain size of an image and is calculated depending on the variance of the image. Entropy of pixel values is used as a measure of the probability of information content in an image. The spatial gray-level difference statistics describes the brightness relationship of pixels within neighborhoods. It is also known as co-occurrence matrix analysis (Haralick, Shanmugam, & Dinstein, 1973). The circular Moran autocorrelation function measures the roughness of the texture. For the calculation a set of autocorrelation functions is used (Gu et al., 1998).

Monomial invariant feature histogram. A feature is called invariant with respect to certain transformations if it does not change when these transformations are applied to the image. The transformations considered here are translation

DOI: 10.1002/asi

and rotation. In this work, invariant feature histograms are used (as presented in Siggelkow, Schael, & Burkhardt, 2001). These descriptors are based on constructing features invariant with respect to certain transformations by integration over all considered transformations. The resulting histograms are compared using the Jeffrey divergence (Puzicha et al., 1999). The histograms take into account monomial functions of the pixel values in a certain area around each pixel and are known to perform similar to color histograms.

Relational invariant feature histograms. These are constructed in the same way as the monomial invariant feature histograms, described in the previous paragraph. However, instead of using a monomial function, these histograms take into account the differences in the brightness of neighboring pixels and are therefore relatively invariant with respect to changes in lighting while maintaining good performance in discriminating between images.

Tamura features. In Tamura, Mori, & Yamawaki (1978), the authors propose six texture features corresponding to human visual perception: coarseness, contrast, directionality, line-likeness, regularity, and roughness. From experiments testing the significance of these features with respect to human perception, it was concluded that the first three features are very important. Thus, in our experiments, we use coarseness, contrast, and directionality to create a histogram describing the texture (Deselaers et al., 2004) and compare these histograms using the Jensen Shannon divergence (Puzicha et al., 1999). In the QBIC system (Faloutsos et al., 1994), histograms of these features are also used.

Patch histograms using a learned dictionary. Currently in object recognition and detection, the common assumption is that objects consist of parts that can be modeled independently. This has led to a wide variety of a bag-of-features approach (Deselaers, Keysers, & Ney, 2005; Dorkó, 2006). In this article, we follow this approach to generate histograms of image patches for retrieval. The creation is a three-step procedure:

1. The sub-images are extracted from all training images and the dimensionality is reduced to 40 dimensions using PCA transformation.
2. The sub-images of all training images are jointly clustered using the EM algorithm for Gaussian mixtures to form 2000-8000 clusters.
3. All information about each sub-image is discarded except its closest cluster center. Then, for each image, a histogram over the cluster identifiers of the respective patches is created, thus effectively coding which "visual words" from the code-book occur in the image. These histograms are then compared using the JSD measure.

Sparse patch histograms using general dictionary. Here, the images are represented by image patches that are extracted at each position and then efficiently stored in a histogram. In addition to the patch appearance, the positions of the extracted patches are considered and provide a significant increase in the recognition performance. Using this method, we create sparse histograms of 65,536 (8^4) bins, which are compared using the JSD measure (a detailed description of the method is given in Deselaers, Hegerath, Keysers, and Ney, 2006). In comparison to the histograms of patches described in the previous paragraph, here the bins are setup to effectively cover the complete feature space of patches ,whereas the previous histograms cover only the part of the patch space that is actually covered by the images. The advantage of the general dictionary is that it is very easy to obtain while creating the learned dictionary is a computationally expensive task.

Data Analysis

There is no single best method or measure to assess the degree of correspondence between distance/dissimilarity matrices of sets of images (and, by extension, their respective MDS configurations/maps and clusterings). Hence, to assess the difference between the feature and perceptual spaces of an image collection, we used Mantel's (1967) test, which is widely used by researchers in fields such as ecology and zoology. Mantel's test was preferred to other methods and measures (e.g., *adjusted Rand index* (Hubert & Arabie, 1985) to compare partitions and to test the similarity between two multidimensional scaling (MDS) configurations (e.g., *procrustes analysis* (Legendre & Legendre, 1998). This is because Mantel's test utilizes the original similarity/dissimilarity matrices, while the other two rely on transformations of the original similarity/dissimilarity matrices.

Mantel's test provides a measure, Z (Mantel's statistic), of the significance of the correlation between elements of two distance/dissimilarity matrices. The test involves computation of several values of Mantel's statistic, Z, and a randomization procedure to see whether the observed correlation (as measured by Z) is significantly different from random correlation (random values of Z; Manly, 2005). The Mantel test statistic, Z, is given by

$$Z = \sum_{i,j} X_{ij} Y_{ij}$$

where X_{ij} and Y_{ij} ($i \neq j$) are the ith and jth off diagonal elements of the two distance/dissimilarity matrices. The null hypothesis tested is as follows:

H_0: there is no association between elements in the two distance/dissimilarity matrices.

The standardized Mantel's test statistic, r, (its values ranging between -1 and 1) is given by

$$r = \frac{1}{n^2 - n - 1} \sum_{i,j} \frac{X_{ij} - \overline{X}}{S_X} \cdot \frac{Y_{ij} - \overline{Y}}{S_Y}$$

where n is the number of rows (columns/cases) in one of the distance/dissimilarity matrices, $\overline{X}$ and $\overline{Y}$ are the average of the elements in the two distance/dissimilarity matrices, and S_X and S_Y are their standard deviations. In order to test

DOI: 10.1002/asi

the significance of the Mantel statistic (either Z or r), randomization of the elements of one of the distance/dissimilarity matrices (while holding the other constant) is used to create a randomized distribution of Z (or r) values. The p-value of the test of significance is

$$p = \frac{NGE + 1}{N + 1}$$

where NGE is the number of Z values obtained through randomization that are greater than or equal to the observed Z value, and N is the number of randomizations.

Results

The three studies (summarized in Table 1) yielded four dissimilarity matrices constructed as described above, based on human similarity judgment data obtained through free-sorting and magnitude estimation. The corresponding set of 8 dissimilarity matrices for the 8 types of descriptors with their according similarity measures (Table 2) were also constructed. Pairs of dissimilarity matrices for each study (one each for human similarity judgments and visual descriptors) were analyzed with *zt*, a computer program to conduct Mantel's test (Bonnet & Van de Peer, 2002). Computed values of the standardized Mantel's test statistic, r, together with their respective p-values are presented in Table 3.

Table 3 shows that the correlation coefficients (standardized Mantel statistic, r) between the dissimilarity matrices for human similarity judgments of all three studies and dissimilarity matrices based on six of the eight visual descriptors were significantly different from zero with p-values smaller than 0.005. What is more, human similarity judgments from Studies 1 & 2 have moderate correlations ($p < 0.005$) with all except one of the visual descriptors (descriptor 3—GTF). The fact that dissimilarity matrices for human similarity judgments of Study 3 (obtained through magnitude estimation) were not significantly associated with most of the eight dissimilarity matrices for visual descriptors raises an interesting question regarding the effect of mode/method of human similarity judgment. Human similarity judgment data collected through free-sorting tasks produced significant correlations with almost all visual descriptors; human similarity judgment data obtained through direct magnitude estimation did not. Although the highest correlation is <0.3, on the basis of these results there is enough evidence for us to conclude that a statistically significant positive relationship exists between human similarity judgment and similarity measures for the majority of visual image features. We believe this to be evidence for a correspondence between the feature and perceptual spaces, thereby supporting the argument of Chen et al. (2000) and the use of low-level features for visualizing images.

Table 3 also shows that despite most of the descriptors having significant correspondence to human similarity judgments, none of the descriptors alone correlates very strongly with human perception. One interesting topic in CBIR is the

TABLE 1. A summary of the three studies.

	Study		
Study details	1	2	3 (ab & ba)
No. of images	50	50	30
Type of images	People	People	Misc.
Method of similarity judgment	Free-sorting	Free-sorting	Magnitude estimation
No. of participants	30	75	75*
Total number of groups formed	240	659	N/A
Min. No. of groups	3	2	N/A
Max. No. of groups	7	24	N/A
Mean No. of groups	8	8.79	N/A
Median No. of groups	7	8	N/A
Mode No. of groups	7	9	N/A
SD (No. of groups)	3.3	4.1	N/A

* The 75 participants of Study 3 judged two sets of 435 pairs (ab & ba) of the same set of 30 images.

TABLE 2. Visual features and similarity measures used to construct the feature matrices.

	Feature	Similarity measure
1	32 × 32 image	Euclidean distance
2	Color histogram	Jensen Shannon Divergence
3	Global texture feature	Euclidean distance
4	Monomial invariant feature histogram	Jensen Shannon Divergence
5	Relational invariant feature histogram	Jensen Shannon Divergence
6	Tamura texture histogram	Jensen Shannon Divergence
7	4096 bin patch histogram (learned)	Jensen Shannon Divergence
8	65536 bin sparse patch histogram	Jensen Shannon Divergence

DOI: 10.1002/asi

TABLE 3. Standardized Mantel statistic (r) values for the association between dissimilarity matrices for the three studies and visual image features (1-8).

		Study 1		Study 2		Study 3 (SIMAB)		Study 3 (SIMBA)	
	Feature	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*
1	32 × 32 image	0.230*	0.0009	0.162*	0.0009	0.103	0.066	0.070	0.148
2	Color histogram	0.269*	0.0009	0.160*	0.0009	0.062	0.168	0.076	0.118
3	GTF	0.015	0.3027	0.047	0.0569	0.055	0.199	0.079	0.128
4	Monomial IFH	0.232*	0.0009	0.162*	0.0009	0.043	0.2478	0.055	0.1888
5	Relational IFH	0.185*	0.0009	0.090**	0.0029	−0.037	0.315	−0.047	0.271
6	Tamura histogram	0.184*	0.0009	0.107*	0.0009	0.021	0.343	0.042	0.262
7	**4096 bin patch histogram**	**0.295***	**0.0009**	**0.237***	**0.0009**	**0.214***	**0.0009**	**0.267***	**0.0009**
8	65536 bin patch (sparse) histogram	0.281*	0.0009	0.245*	0.0009	0.121	0.015	0.143	0.008

*$p < .001$, **$p < .005$, one-tailed (1000 Randomizations).

combination of features. Because most features correspond to particular properties of an image (e.g., color histograms describe only the color distribution of images and GTF and Tamura describe only textural properties), for most scenarios a combination of features is typically the most successful approach.

Therefore, we propose a method that finds the combination of features that best matches human perception. Similarity measures in CBIR systems are commonly combined linearly, i.e., each feature/similarity measure is assigned a weight and then the weighted sum is calculated to obtain a similarity measure accounting for different properties. Similarly, the method proposed here calculates descriptor weights such that their combination best matches with human perception.

Given the eight descriptors and corresponding similarity matrices, it is possible to find the linear combination of these descriptors that leads to the similarity matrix best resembling the similarity matrices from the human studies. Given a pair of images, we find feature weightings which lead to the same similarity score as obtained from the three studies. Considering all images from each study at once, a strongly over-determined system of linear equations is obtained comprising eight variables (the weights for each of the descriptors) and as many equations as pairs of images considered (1225 in studies 1 and 2; 435 in study 3).

These systems of equations are solved using singular value decomposition (SVD) and the solutions lead to a set of weights for study 1 and 2, and one set of weights for study 3AB and study 3BA, respectively. These weights are used to create a new set of similarity matrices by calculating the weighted sum of the dissimilarity matrices of the individual descriptors. A particularly interesting result is whether the findings (i.e., weights) from one study can be applied to the other studies to find a good combination of descriptors. The results from Mantel's test with these feature combinations are given in Table 4.

The weights obtained from each of the studies were used to create a new similarity matrix for each study. As expected, creating a feature combination for a particular study leads to very high *p*-values. These values can be seen as a very optimistic estimate of how well visual descriptors can be combined to match human perception. In fact, using a linear combination of the features used, no better match is possible for the task at hand. However, because the combined descriptors also lead to high correspondences for the other studies, we can conclude that we can learn how to combine features from one dataset and apply the combination to other tasks. In particular, we can conclude that the method leads to a feature combination that generalizes well over different sets of images; i.e., it is possible to consider one set of images, execute a study with human subjects, obtain the optimal feature combination using our proposed method, and use this combination of features with another, possibly much larger, set of images. From Tables 3 and 4, it can be observed that the feature combinations perform almost as good as the best single

TABLE 4. Results from Mantel's test for the combinations of features for the three studies.

		r values ($p = 0.000999$)			
		Study 1	Study 2	Study 3 (SIMAB)	Study 3 (SIMBA)
Weights from study	1	0.393648	0.244613	0.201997	0.207113
	2	0.257257	0.269415	0.265245	0.294700
	3 (SIMAB)	0.256781	0.167512	0.298475	0.350045
	3 (SIMBA)	0.245353	0.163617	0.290779	0.355902

Note. The weights were obtained from solving the system of linear equations from each particular study and were applied to the other studies.

DOI: 10.1002/asi

descriptors for each study, but the combinations of features are more robust in the sense that they perform equally well for all of the studies whereas a high variance in correspondence can be observed for the individual descriptors in Table 3.

Much work exists on feature combination in the CBIR literature; however, as far as we know, this is the first attempt at creating a weighted combination of features for CBIR to match human similarity judgments.

Figures 1(a)–(d) show the weights for the three studies (a = Study 1, b = Study 2, c = Study 3/AB, d = Study 3/BA). It can be clearly seen that the weights obtained are similar among all studies and, in particular, the weights obtained for studies 3AB and 3BA are almost identical (because they involved the same sample of images and participants). For all studies, the weight for the patch histogram with learned vocabulary is by far the highest, and this is consistent with the fact that this descriptor has the highest correspondence as an individual descriptor. Furthermore, combining the weights for color histograms and monomial IFH produced a positive and a negative weight. Because it is well known that these features have very similar properties (Deselaers et al., 2004), either one can be replaced by the other. Therefore, the fact that, in Study 2, the monomial IFH has a positive weight and the color histogram has a negative weight cannot be considered a major

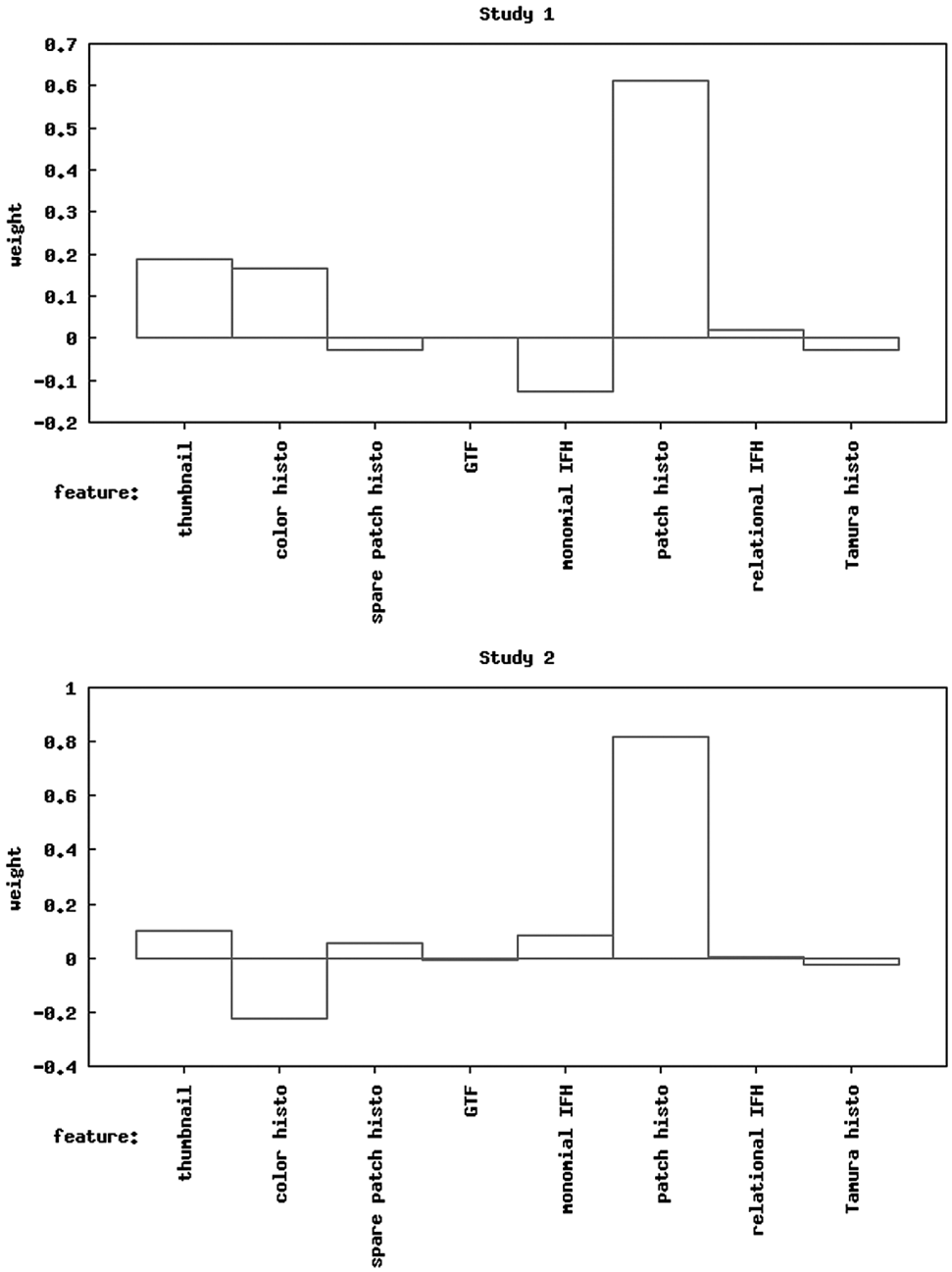

FIG. 1. Weights obtained for the three studies by solving the system of linear equations to find the combination of features that best resembles human perception.

DOI: 10.1002/asi

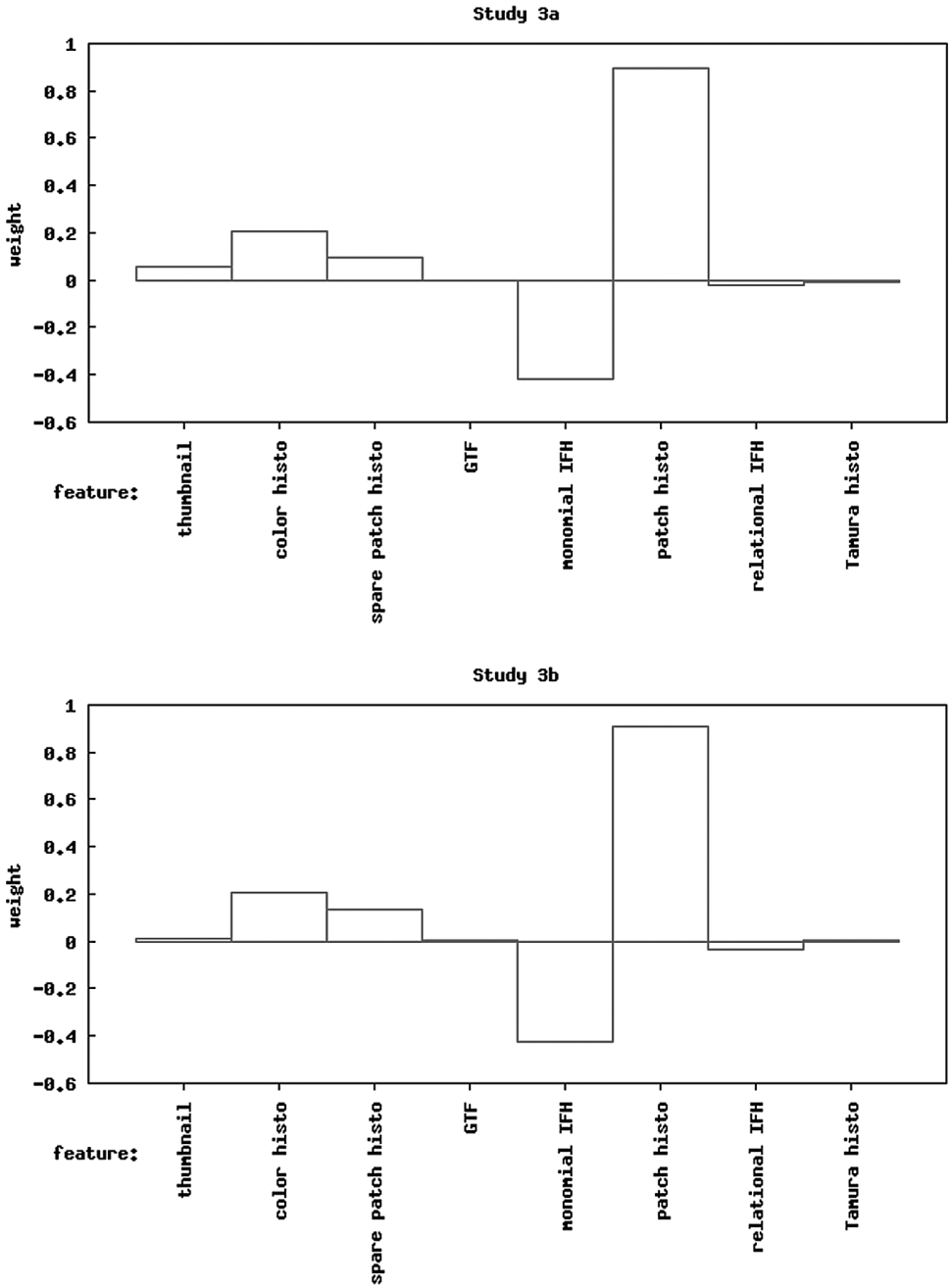

FIG. 1. *Continued*

difference to the other studies in which the color histogram is weighted positively and the monomial IFH is weighted negatively. Interestingly, the above-mentioned difference between Studies 1, 2, and 3AB/BA respectively, due to the different human similarity judgment methods used, does not play a role anymore. The weights obtained from any of the studies can be used for any other study leading to high correspondence with human perception.

Tables 5–7 show the correlations between the different descriptors used to represent the images. The correlations are very similar for all three studies and it is clear that the color histogram and the monomial invariant feature histogram are strongly correlated because they represent nearly the same information. Interestingly, both of these are also strongly correlated with the two types of patch histograms that show that the patch histograms do not only capture local texture information but also color texture information. The different textural descriptors (global texture feature, relational invariant feature histogram, and Tamura texture feature) show only a moderate correlation, which is a strong hint that the texture representations differ in what they represent and that probably none of them is sufficient to give a complete description of the textures in an image.

DOI: 10.1002/asi

TABLE 5. Correlation between image descriptors for study 1.

		2		3		4		5		6		7		8	
	Descriptor	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*
1	32 × 32 image	0.19*	0.0009	−0.02**	0.2997	0.14**	0.0109	0.026**	0.277	−0.01**	0.429	0.047**	0.169	0.306*	0.0009
2	Color histogram			−0.15*	0.0009	0.78*	0.0009	−0.005**	0.48	−0.05**	0.20	0.37*	0.0009	0.33*	0.0009
3	Global texture feature (GTF)					0.25**	0.0039	0.426*	0.0009	0.496*	0.0009	0.27*	0.0009	0.30*	0.0009
4	Monomial invariant feature histogram							0.30*	0.0009	0.315*	0.0009	0.588*	0.0009	0.57*	0.0009
5	Relational invariant feature histogram									0.78*	0.0009	0.588*	0.0009	0.58*	0.0009
6	Tamura texture histogram											0.74*	0.0009	0.69*	0.0009
7	4096 bin patch histogram (learned)													0.77*	0.0009
8	65536 bin sparse patch histogram														

*$p < .001$, **$p < .005$, one-tailed (1000 Randomizations).

TABLE 6. Correlation between image descriptors for Study 2.

		2		3		4		5		6		7		8	
	Descriptor	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*
1	32 × 32 image	0.19*	0.0009	0.03**	0.258	0.18**	0.0019	0.16**	0.015	0.19**	0.03	0.243*	0.0009	0.59*	0.0009
2	Color histogram			0.3*	0.0009	0.96*	0.0009	0.52*	0.0009	0.52*	0.0009	0.79*	0.0009	0.65*	0.0009
3	Global texture feature (GTF)					0.408*	0.0009	0.58*	0.0009	0.66*	0.0009	0.34*	0.0009	0.25*	0.0009
4	Monomial invariant feature histogram							0.58*	0.0009	0.605*	0.0009	0.77*	0.0009	0.64*	0.0009
5	Relational invariant feature histogram									0.87*	0.0009	0.47*	0.0009	0.44*	0.0009
6	Tamura texture histogram											0.54*	0.0009	0.50*	0.0009
7	4096 bin patch histogram (learned)													0.69*	0.0009
8	65536 bin sparse patch histogram														

*$p < .001$, **$p < .005$, one-tailed (1000 Randomizations).

TABLE 7. Correlation between image descriptors for Study 3.

		2		3		4		5		6		7		8	
	Descriptor	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*
1	32 × 32 image	0.104**	0.42	0.30*	0.0009	0.12**	0.0339	−0.08**	0.14	−0.12**	0.041	0.017**	0.390	0.375*	0.0009
2	Color histogram			−0.01**	0.465	0.98*	0.0009	0.09**	0.082	0.08**	0.123	0.65*	0.0009	0.51*	0.0009
3	Global texture feature (GTF)					0..001**	0.478	0.063**	0.1298	0.15**	0.048	0.15**	0.0109	0.01**	0.376
4	Monomial invariant feature histogram							0.12**	0.039	0.11**	0.0529	0.64*	0.0009	0.53*	0.0009
5	Relational invariant feature histogram									0.68*	0.0009	0.24*	0.0009	0.12**	0.02
6	Tamura texture histogram											0.40*	0.0009	0.06**	0.145
7	4096 bin patch histogram (learned)													0.36*	0.0009
8	65536 bin sparse patch histogram														

*$p < .001$, **$p < .005$, one-tailed (1000 Randomizations).

Implications for Image Retrieval

The visual elements of an image are directly related to perceptual aspects along with high-level concepts that define its meaning. The results from this study suggest that representing low-level features of images using real-valued attributes and using a suitable distance function to compare them does allow various perceptual aspects of visual content to be represented and visualized according to human similarity judgments, supporting existing literature such as Chen et al., 2000). Such a representation, together with proper distance measures and learning, can effectively help to reduce the semantic gap.

Two fundamental approaches for accessing information are search and browse. The use of browsing has shown to be a very effective technique in the domain of image retrieval (Combs & Bederson, 1999; Chang et al., 2004; Laine-Hernandez & Westman, 2006) and combined with text searching based on descriptive metadata (e.g., text assigned to an image to represent high-level semantic content), then users are able to select their preferred interaction mode (content or concept-based) and move between the two (Combs & Bederson, 1999). Jörgensen and Jörgensen's (2005) study of image professionals revealed that 85.6% of the searches involved browsing of results, implying that this behavior is important in finding and selecting relevant images.

Visualization techniques are typically utilized in image retrieval to either visualize the results of a targeted search or allow the exploration of an entire collection of images. Visualizing an entire image collection (called a collection overview) is different from visualizing the results of a targeted search because rather than trying to locate a specific item, the goal is often to obtain a general understanding of the underlying theme of a collection (Chang et al., 2004; Combs & Bederson, 1999). Various visualization approaches using mainly low-level visual features have been suggested for image retrieval. For example, Janecek and Pu (2004) advocate the use of semantic "fisheye" views to enable focusing in on relevant parts of a wide set of results. This type of visualization helps users to examine local details while still maintaining a view of the broader context. Liu, Xie, Tang, Li, and Ma (2004) developed a similarity-based results presentation that was meant to graphically depict the closeness of relationships between images based on "regions of interest" within the images. The items were then arranged in a way so that closely related pictures were situated near and overlapped each other. Park, Baek, and Lee (2005) took the top 120 images and clustered these using hierarchical agglomerative clustering methods (HACM). Clusters are then ranked based on the distance of the cluster from the query. The effect is to group together visually similar images in the results.

Other visualization approaches have combined both visual and textual information to cluster sets of images into multiple topics. For example, Cai, He, Li, Ma, and Wen (2004) use visual, textual, and link information to cluster Web image search results into different types of semantic clusters.

Barnard and Forsyth (2001) organize image collections using a statistical model which incorporates both semantic information extracted from associated text and visual data derived from image processing. During a training phase, they train a generative hierarchical model to learn semantic relationships between low-level visual features and words. The resulting hierarchical model associates segments of an image (known as blobs) with words and clusters these into groups which can then be used to browse the image collection.

Given the importance of using low-level features for visualizing images, similarity models which more closely fit with human similarity judgement must be investigated if effective and intuitive information access is to be provided to image repositories (Del Bimbo, 1999). The results of these experiments indicate that as a single visual feature (cf. Table 3), the 4096 bin patch histogram with the JSD dissimilarity measure provides the most consistent correlation across different image organization tasks. This single feature would appear to encapsulate some degree of perceptual information and therefore be most likely a good candidate for visualizing images—without using any associated semantic information—in more general tasks. For example, browsing/navigating a visual space or organizing the results of a general image search engine (e.g., Google Images). It is also evident that using a combination of features (rather than single ones) will result in a higher correlation with human image similarity. This suggests, like previous work, that feature combination should be used in visualization as different features are likely to be important to different users and for different tasks.

Conclusions

In this article, we attempted to tackle the bigger problem of the gap in the feature and perceptual features and spaces of image collections. We set out to investigate if low-level visual image features correlate to human similarity perception and whether it is possible to find a combination of features that closely resembles human similarity perception. In order to achieve this, we collected and analyzed human similarity judgment data from three studies, involving 130 images and 180 participants, using free-sorting (two studies) and magnitude estimation (one study). Human similarity judgment data were aggregated into dissimilarity matrices, using various data summary techniques, and correlated with dissimilarity matrices based on eight visual features of the 130 images. The final analysis through Mantel's (1967) test revealed that most of the visual features had moderate positive correlations with human perceived features, as evidenced by the statistically significant standardized Mantel statistic values, *r*-values, for correlations between the majorities of the relevant pairs of dissimilarity matrices.

The fact that these results were obtained in three separate studies suggests that they may not be due to chance and we can safely conclude that there is a reasonable degree of correspondence between the feature and perceptual spaces of collections of images and more so when combinations of visual features are considered in representation/indexing of images and the construction of their visualizations (feature space). Furthermore, combinations of features were created to maximize the correspondence to the three studies and it could be observed that these combinations also lead to high correspondences for the respective other studies, which underlines the observation that these findings are not by chance. Although we recognize that image perception will be influenced by such factors as context, the content of an image collection itself and a user's task, the results of more general (or context-free) studies are required to understand the nature of image perception in task-independent situations (such as a general image search on the Web). We believe that the results of these studies will lead to a better understanding of the nature of human perception of images and better design of image visualization and browsing systems. There is sufficient evidence to suggest that current CBIR systems and the various techniques they utilize have come a long way in bridging the feature and perceptual gap when it comes to image features and visualization.

We acknowledge that the current approach has limitations as results are not based on a large collection of images and participants who are actual image users in order for them to be generalized. A total of 130 images from two different datasets have been employed: two random samples of 50 images from the "people" subset of a stock photographic collection for Studies 1 and 2 and a random sample of 30 general images for Study 3. Although smaller than other collections used in visualization experiments, we were limited by the amount of time it would take human participants to sort the images. Larger sample sizes, although arguably preferable, could have introduced fatigue and affected the reliability of the data. We also acknowledge that the use of people images could have affected the results of Studies 1 and 2, and the images and the method for assessing similarity used was different for Study 3.

In order to address limitations of the current work, future research needs to look into methods and means to study the nature of the feature and perceptual space gap. The possible areas of focus are as follows: (a) applying the same methodology to different sets of images to validate initial findings, (b) finding better ways to combine features, the quantitative evaluation of the feature combinations for large-scale CBIR experiments, and (c) investigation into whether it is possible to directly extract features that resemble human perception.

Acknowledgements

The first author is grateful to the School of Library and Information Sciences, University of North Texas, for financial and academic support as well as colleagues at the University at Albany. We are also indebted to the 180 participants of the three studies. We thank the three anonymous reviewers for their thoughtful and helpful suggestions.

DOI: 10.1002/asi

References

Barnard, K., & Forsyth, D. (2001). Learning the semantics of words and pictures. In Proceedings of the 8th International Conference on Computer Vision (Vol. II) (pp. 408–415). Vancouver, Canada: IEEE.

Bonnet, E., & Van de Peer, Y. (2002). Zt: A software tool for simple and partial Mantel tests. Journal of Statistical software, 7(10), 1–12.

Cai, D., He, X., Li, Z., Ma, W-Y., & Wen, J-R. (2004). Hierarchical clustering of WWW image search results using visual, textual and link information. Proceedings of the 12th Annual ACM International Conference on Multimedia, 952–959.

Chang, M., Leggett, J. J., Furuta, R., Kerne, A., Williams, J. P., Burns, S. A. et al.(2004). Collection understanding. Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries (pp. 334–342). Tucson, AZ, USA. New York, NY: ACM Press.

Chen, C., Gagaudakis, G., & Rosin, P. (2000). Content-based image visualisation. Proceedings of the Fourth IEEE International Conference on Information Visualisation (pp. 13–18). Los Alamitos, CA: IEEE Computer Society Press.

Combs, T.T.A., & Bederson, B.B. (1999). Does zooming improve image browsing? Proceedings of Digital Library (pp. 130–137). New York: ACM.

Coxon, A. P. M. (1999). Sorting data: Collection and analysis. Thousand Oaks, CA: Sage Publications.

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

Datta, R., Li, J., & Wang, J.Z. (2005). Content-based image retrieval: approaches and trends of the new age. In H. Zhang, J. Smith, & Qi Tian (Eds.), Proceedings of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval (pp. 253–262). New York: ACM.

Del Bimbo, A. (1999). Visual Information Retrieval. San Francisco, CA: Morgan Kaufmann.

Deselaers, T., Hegerath, A., Keysers, D., & Ney, H. (2006). Sparse patch-histograms for object classification in cluttered images. Pattern Recognition, Proceedings of the 26th DAGM Symposium, Berlin, Germany. Lecture Notes in Computer Science, 4174, 202–211.

Deselaers, T., Keysers, D., & Ney, H. (2004). FIRE—Flexible image retrieval engine: ImageCLEF 2004 evaluation. Lecture Notes in Computer Science, 3491, 688–698.

Deselaers, T., Keysers, D., & Ney, H. (2005). Discriminative training for object recognition using image patches. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (Vol. 2, pp. 157–162). San Diego, CA: IEEE.

Deselaers, T., Keysers, D., & Ney, H. (in press). Features for image retrieval: An experimental comparison. Information Retrieval.

Deselaers, T., Müller, H., Clough, P., Ney, H., & Lehmann, T. (2007). ImageCLEF 2005 medical automatic image annotation task. International Journal of Computer Vision, 74, 51–58.

Dori, D. (2000). Cognitive image retrieval. Proceedings of the 15th International Conference on Pattern Recognition (pp. 1042–1045). Barcelona, Spain.

Dorkó, G. (2006). Selection of discriminative regions and local descriptors for generic object class recognition. Unpublished doctoral dissertation, Institut National Polytechnique de Grenoble, France.

Dunn-Rankin, P., Knezek, G., Wallace, S., & Zhang, S. (2002). Scaling methods (2nd prepublication ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

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.

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

Fauqueur, J., & Boujemaa, N. (2003). Logical query composition from local visual feature thesaurus. Third International Workshop on Content-Based Multimedia Indexing, September 22–24, 2003. Retrieved August 14, 2007, from http://www.eng.cam.ac.uk/~jf330/papers/fauqueur_CBMI03.pdf

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.A. (2001). Multidimensional scaling of video surrogates. Journal of the American Society for Information Science and Technology, 52(2), 174–182.

Greisdorf, H., & O'Connor, B.C. (2002a). 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.

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

Gu, Z.Q., Duncan, C.N., Renshaw, E., Mugglestone, M.A., Cowan, C.F.N., & Grant, P. M. (1989). Comparison of techniques for measuring cloud texture in remotely sensed satellite meteorological image data. Radar and Signal Processing, 136(5), 236–248.

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

Haberäcker, P. (1995). Praxis der digitalen bildverarbeitung und mustererkennung (Practice of digital image processing and pattern recognition). München, Wien: Carl Hanser Verlag.

Haralick, R.M., Shanmugam, B., & Dinstein, I. (1973). Texture features for image classification. IEEE Transactions on Systems, Man, and Cybernetics, 3(6), 610–621.

Hubert, L., & Arabie, P. (1985). Comparing partitions. Journal of Classification, 2, 193–218.

Ingwersen, P. (1992). Information retrieval interaction. London: Taylor Graham.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: elements of cognitive IR theory. Journal of Documentation, 52(1), 3–50.

Janecek, P., & Pu, P. (2004). Opportunistic search with semantic fisheye views. EFPL Technical Report: IC/2004/42.

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

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.

Jörgensen, C., & Jörgensen, P. (2005). Image querying by image professionals. Journal of the American Society for Information Science and Technology. 56(12), 1346–1359.

Keysers, D., Deselaers, T., Gollan, C., & Ney, H. (2007). Deformation models for image recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Laine-Hernandez, M., & Westman, S. (2006). Image semantics in the description and categorization of journalistic photographs. In A. Grove & J. Stefl-Mabry (Eds.), Proceedings of the 69th Annual Meeting of the American Society for Information Science and Technology. Retrieved June 11, 2007, from http://www.asis.org/Conferences/AM06/proceedings/papers/48/48_paper.html

Legendre, P., & Legendre, L. (1998). Numerical ecology (2nd English ed.). Amsterdam: Elsevier.

Lew, M.S., Sebe, N., Djeraba, C., & Jain, R. (2006). Content-based multimedia information retrieval: State of the art and challenges. ACM Transactions on Multimedia Computing, Communications, and Applications, 2, 1–19.

Li, B., Chang, E., & Wu, Y. (2003). Discovery of a perceptual distance function for measuring image similarity. Multimedia Systems, 8(6), 512–522.

Liu, H., Xie, X., Tang, X., Li, Z., & Ma, W. (2004). Effective browsing of Web image search results. Proceedings of MIR (pp. 84–90). New York, USA.

Lyman, P., & Varian, H.R. (2003). How much information 2003? Retrieved June 11, 2007, from http://www.sims.berkeley.edu/research/projects/how-much-info-2003

Manly, B. F. J. (2005). Multivariate statistical methods: A primer (3rd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. Cancer Research, 27, 209–220.

DOI: 10.1002/asi

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

Müller, H, Müller, W, Squire, D.M., Marchand-Maillet, S, & Pun, T. (2000). Learning features weights from user behavior in content-based image retrieval. In S. Simoff & O. Zaiane (Eds.), ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Workshop on Multimedia Data Mining MDM/KDD2000). Boston, MA, USA.

Neumann, D., & Gegenfurtner, K.R. (2006). Image retrieval and perceptual similarity. ACM Transactions on Applied Perception, 3, 31–47.

Newby, G.B. (2001). Cognitive space and information space. Journal of the American Society for Information Science and Technology, 52(12), 1026–1048.

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.

Park, G., Baek, Y., & Lee, H-K. (2005). Re-ranking algorithm using post-retrieval clustering for content-based image retrieval. Information Processing and Management, 41(2), 177–194.

Puzicha, J., Rubner, Y., Tomasi, C., & Buhmann, J. (1999). Empirical evaluation of dissimilarity measures for color and texture. Proceedings of the Seventh IEEE International Conference on Computer Vision (Vol. 2, pp. 1165–1173). Corfu, Greece: IEEE.

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.

Rodden, K., Basalaj, W., Sinclair, D., & Wood, K. (1999). Evaluating a visualisation of image similarity as a tool for image browsing. Proceedings of the IEEE Symposium on Information Visualisation (pp. 36–43). San Francisco, CA: IEEE.

Rodden, K., Basalaj, W., Sinclair, D., & Wood, K. (2000). A comparison of measures for visualising image similarity. Proceedings of Challenges of Image Retrieval. Retrieved March 6, 2007, from http://www.rodden.org/kerry/cir2000.pdf

Rodden, K., Basalaj, W., Sinclair, D., & Wood, K. (2001). Does organisation by similarity assist image browsing? ACM Conference on Human Factors in Computing Systems (pp. 190–197). Seattle, WA: ACM.

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.

Rorissa, A., & Iyer, H. (in press). Theories of cognition and image categorization: What category labels reveal about basic level theory. Journal of the American Society for Information Science and Technology.

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

Salton, G., Wong, A., & Yang, C.S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11), 613–620.

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

Siggelkow, S., Schael, M., & Burkhardt, H. (2001). SIMBA—search IMages by appearance. Pattern Recognition, Proceedings of the 23rd DAGM Symposium, Munich, Germany. Lecture Notes in Computer Science, 2191, 9–17.

Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(12), 1349–1380.

Smith, J.R., & Chang, S-F. (1996). Tools and techniques for color image retrieval. Proceedings of the Storage & Retrieval for Image and Video Databases IV (Vol. 2670, pp. 426–437). San Jose, CA: IS&T/SPIE.

Squire, D., & Pun, T. (1998). Assessing agreement between human and machine clusterings of image databases. Pattern Recognition, 31(12), 1905–1919.

Stan, D., & Sethi, I. K. (2003). eID: A system for exploration of image databases. Information Processing and Management, 39, 335–361.

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

Swain, M.J., & Ballard, D.H. (1991). Color indexing. International Journal of Computer Vision, 7(1), 11–32.

Tamura, H., Mori, S., & Yamawaki, T. (1978). Textural features corresponding to visual perception. IEEE Transaction on Systems, Man, and Cybernetics, 8(6), 460–472.

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.

Vasconcelos, N, & Lippman, A. (2000). A unifying view of image similarity. Proceedings of the 15th International Conference on Pattern Recognition (pp. 1038–1041). Barcelona, Spain.

Yavlinsky, A., Pickering, M.J., Heesch, D., & Rüger, S. (2004). A comparative Study of Evidence Combination Strategies. IEEE International Conference on Acoustics, Speech, and Signal Processing (pp. 1040–1043.). Montreal, Canada.

Zachary, J. (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., & 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.

DOI: 10.1002/asi