Image Retrieval and Perceptual Similarity

DIRK NEUMANN and KARL R. GEGENFURTNER
Justus Liebig University Giessen

Simple, low-level visual features are extensively used for content-based image retrieval. Our goal was to evaluate an image-indexing system based on some of the known properties of the early stages of human vision. We quantitatively measured the relationship between the similarity order induced by the indexes and perceived similarity. In contrast to previous evaluation approaches, we objectively measured similarity both for the few best-matching images and also for relatively distinct images. The results show that, to a large degree, the rank orders induced by the indexes predict the perceived similarity between images. The highest index concordance employing a single index was obtained using the chromaticity histogram. Combining different information sources substantially improved the correspondence with the observers. We conclude that image-indexing systems can provide useful measures for perceptual image similarity. The methods presented here can be used to evaluate and compare different image-retrieval systems.

Categories and Subject Descriptors: H.3.3 [**Information Systems**]: Information Storage and Retrieval—*Retrieval models*; H.3.1 [**Information Systems**]: Content Analysis and Indexing—*Indexing methods*; J.4 [**Social and Behavioral Sciences**]—*Psychology*

General Terms: Experimentation

Additional Key Words and Phrases: Color indexing, content-based image retrieval, Fourier spectrum, image search

1. INTRODUCTION

The goal of image-indexing systems is to find a set of images that is similar to the target image for which the user is searching. Depending on the intention of the user, similarity may be defined by objects, configuration, illumination, camera position or zoom, by semantic aspects, or any combination of the above. Ideally, computer vision algorithms should extract all the relevant features from the image in the same way as a human observer. While there has been tremendous progress in recent years [Rui et al. 1999], it is quite clear that this ultimate goal is far beyond our current knowledge of human vision, cognition, and emotion. Here we present a paradigm useful for the evaluation of different image-indexing systems. We construct a relatively simple image-indexing system based on some of the known properties of visual perception and show that such systems can provide measures of visual similarity that correlate highly with human judgments.

Most current algorithms are quite successful in using low-level features of the images. Color, in particular, has proved to be very effective for the calculation of image similarity, since an object's color is, to a large degree, independent of viewing position or viewing distance. The most prominent color

This work was supported by the Deutsche Forschungsgemeinschaft grant DFG Ge/879.

Authors' addresses: (Present address) Dirk Neumann, California Institute of Technology, Pasadena, California 91125. Karl Gegenfurtner, Department of Psychology, Giessen University, Otto-Behaghel-Str. 10, 35394 Giessen, Germany.

ACM Transactions on Applied Perception, Vol. 3, No. 1, January 2006, Pages 31–47.

statistics are histograms with equally sized bins in RGB or HSI color space. In HSI space, the luminance (intensity) axis is often ignored, since it is argued that the overall brightness of an image is irrelevant with respect to image similarity. Hence, luminance-independent indexes should be more robust with respect to different illumination conditions. Other color statistics include correlation or covariance coefficients encoding spatial information about the color distribution [e.g. Huang et al. 1999; Stricker and Dimai 1997]. Features such as texture and shape are other subjects of current research [e.g. Flickner et al. 1995]. The combination of multiple features has received much attention in recent years. Iqbal and Aggarwal [2002] combine linguistic color labels and a Gabor texture index for image-indexing. Rui et al. [1998] combine color, texture, and shape information and use the user's feedback about the relevance of the search results to update the weighting of the histograms.

All these approaches are based on physical, low-level features. However, in the end, the similarity of a given image with a target image will always be judged by a human observer. Therefore, we wanted to test whether the output of such image-indexing systems does correlate with perceptual similarity measures. Until now, the success of indexing algorithms has mostly been judged by the experimenters or the designers of the algorithms. In contrast, we use a strictly quantitative and objective approach to evaluate perceived image similarity, by obtaining forced choice judgments from a group of naive observers. We show that our method leads to consistent results and can be used to compare the effectiveness of different indexes.

For our image-indexing system, we used a representation of color and spatial frequency, similar to the one implemented in the primate visual system. Our color index is based on the opponent color processes known to be implemented in the color opponent retinal ganglion cells of primates. Our spatial index is based on multiresolution, multiorientation frequency filters, similar to the ones implemented by simple and complex cells in primary visual cortex.

This procedure provides us with a "plain vanilla" image-indexing system, which we can then use to evaluate its performance. Most of the earlier systems were evaluated informally. The prevalent method is to have the experimenter decide which images found by the algorithm are similar to the query image. As Cox et al. [2000] criticize, the results obtained by such methods are highly dependent on the strictness of the similarity criteria the observers use, the homogeneity of the images in the database, and the number of images displayed. In any case, it is fairly obvious that such evaluation criteria cannot be used to compare different systems.

A different approach is used by so called relevance feedback systems, where the human user is part of the loop. The human observer interacts with the computer to optimize retrieval performance [Salton and Buckley 1990; Rui et al. 1998; Cox et al. 2000]. However, such fine tuning to an individual observer is not always possible and it would be useful to find criteria that allow performance evaluation for a representative sample of users. This is the goal of the work presented here.

Of course, no single algorithm will work ideally for all observers. A certain image can mean many different things to different observers. Image analysis based on low-level features will probably never be able to predict human emotional responses to emotionally charged images, such as the ones contained in the International Affective Pictures System (IAPS) set [see, for example, Maljkovic and Martini 2005]. However, one can still try to predict average perceived similarity from image content. Work by Rogowitz et al. [1998] and Oliva and Torralba [2001] has shown that low-level image features are often similar for images belonging to different semantic categories.

An interesting alternative should be noted here, which avoids the topic of subjective similarity all together. Li et al. [2003] propose a "perceptual distance function for measuring image similarity" that is entirely independent of any human observers. They take images of objects and apply a group of transformations (translation, rotation, etc.) that human perception is known to be invariant. Then they design nonlinear functions on a large set of low-level features. Their ultimate goal is to retrieve as many of the transformed images as possible. While their results appear to be quite impressive, it has to be kept in mind that their concept of similarity is far narrower than the one we investigate here.

We evaluated our indexing system by comparing the similarity judgments made by the algorithm to judgments made by human observers. Unlike Rogowitz et al. [1998] who calculated a similarity metric for a set of 96 images with a multidimensional scaling technique, we measured similarity directly. By looking at the degree with which the algorithm correlates with judgments of the observers, we can estimate how much the different features (indexes) contribute.

In three experiments, we measured the relationship between judgments of perceived similarity and the computed similarity distance over a broad range of images. The similarity of images was varied from highly similar best matches to relatively distinct images with rank 2000 in the result list.

In the first experiment, we investigated the influence of the degree of similarity between the test images on the similarity judgments. In the second experiment, the three indexes were compared to each other to determine their individual contributions. In the third experiment, the indexes were combined iteratively to measure the improvement in prediction when color, spatial, and luminance information are considered.

In brief, we found good agreement between the images selected with the perception-based indexes and the perceived similarity. The observers' judgments can be best predicted with the chromaticity histogram. The luminance and the Fourier histogram both contribute to the similarity judgments and the percentage of agreement increases considerably if the luminance and the Fourier information are combined with the chromaticity index. We found that the percentage of agreement decreases linearly as a function of the logarithmic rank position, from the first, best-matching image up to the 2000th image in the result list. The correlation was found for each of the three indexes and for the index combinations. While there are certainly better image retrieval systems available in the literature, our paradigm provides an objective and simple way to evaluate and compare these systems.

2. COLOR AND SPATIAL INDEX

Ideally, content-based image retrieval (CBIR) systems should be modeled with respect to the users and return images in good agreement with the perceived similarity. Simple color and texture histograms capture important aspects of the perceived similarity and we here pursue a straightforward variant of the histogram approach. For the construction of the color and texture histogram, we employ established coding principles of human vision. Two color histograms were constructed in the color-opponent space of retinal ganglion and LGN cells (DKL) [Derrington et al. 1984]. The texture information was stored analogous to the local orientation selective simple cells in the primary visual cortex using a two-dimensional discrete Fourier transform.

2.1 Early Vision for Color

At the first step of seeing, light is absorbed and converted into neural signals by the three different classes of cone photoreceptors in the retina. This initial stage of processing is well-understood and the exact shape of the absorption spectra of the cones are now known quite precisely. While the cones are often called red, green, and blue cones, they all absorb light over a wide range of the visible spectrum. Most notably, the representation of color in the cones is quite different from the RGB triplets of modern image sensors. Although the transformation converting between the two color spaces is nonlinear, it can be approximated linearly. Any such conversion process requires careful calibration of the image-acquisition device [Wandell 1993].

Still, at the level of the retina, the signals from the cones get transformed by a complex network of retinal cells into color opponent signals [Wässle and Boycott 1991]. Electrical recordings from single neurons in the retina and the lateral geniculate nucleus (LGN) have shown three different classes of neurons. A “luminance-type” neuron simply takes the sum of the outputs from all three cone classes. “Red–green” opponent neurons take the difference between the red- and the green-cone signal. “Blue-yellow” opponent neurons take the difference between the blue-cone signal and the sum of the red- and green-cone signal [Gegenfurtner and Kiper 2003]. This basically results in a principal components analysis of the cone signals [Buchsbaum and Gottschalk 1983; Ruderman et al. 1998] and thus removes any correlation between the signal channels, resulting in nearly optimal information transmission from the eye to the brain. Note that these color opponent signals are quite different from the ones proposed by Hering [1964] and Hurvich and Jameson [1957], which form the basis of the widely used HSI space.

Since color information, in the end, gets represented in the human brain as a small number of different hue categories, it seems much better, at first, to simply use these categories in a color index. However, the assignment of RGB values to categories is not constant, but changes dependent on the distribution of colors in a scene, which is, to a large degree, dependent on the illuminant. Since human observers can “discount” the illuminant [D’Zmura and Lennie 1986], any computer vision algorithm would have to implement the same degree of color constancy. We chose to use a luminance-independent color opponent space and chose our categories, such that the bins would roughly contain equally discriminable ranges of colors.

It was mentioned in the introduction that there is no simple transform to convert RGB triplets into human cone photoreceptor excitations. However, these receptor excitations are necessary to calculate the DKL cone-opponent coordinates of the colors. We circumvented this problem by using the photoreceptor excitations resulting from the display of the images on a standardized and calibrated display monitor (Sony GDM-F500). Since our observers made their judgments looking at the images displayed on that monitor, it is correct to use these excitations as the basis for the similarity metric.

A first color index encodes only chromaticity and is, therefore, luminance independent. The second index encodes the mean luminance of the color tones. For the color bins, we used bin sizes reflecting the granularity of higher-order color perception. For example, for saturated colors, the resolution for hue is much finer than it is for unsaturated colors while the resolution of saturation decreases, in agreement with the results of Krauskopf and Gegenfurtner [1992], who also showed that the increase of thresholds for detecting colors along one cardinal axis is independent from the other cardinal axis. Therefore, this code will lead to a highly efficient representation that ensures that colors within the same bin are nearly indistinguishable, while, at the same time, minimizing the number of bins. However, it is quite likely that other color representations will lead to results similar to the ones presented here.

2.2 Color Histograms

To create the two color histograms, we converted all images into the DKL color space that is spanned by two orthogonal color-opponent axes and luminance. The resulting color distribution can be visualized by plotting the pixels of the image with respect to their coordinates on the red–green and yellow–blue color axis. In Figure 1a, the pixels of the images with the hats (at the bottom left of the chart) were plotted according to their color-opponent coordinates. The pixels of the differently colored hats fall into separate radial sections. The Cartesian color-opponent values were then transformed into the polar coordinates hue and saturation and this cylindrical space was used to define the histogram bins.

The color histograms were created by dividing the chromaticity plane into logarithmic-radial segments (Figure 1b). The resolution for saturation of these bins decreases with increasing saturation. Six different rings were used to discriminate saturation; the remaining unsaturated color tones in the center were averaged into a single, gray bin. The ring r, to which a color tone belongs, was calculated by the logarithm of the saturation s: $r = \lfloor -\log_2 s \rfloor$. The histogram’s hue resolution n_r is lowest for unsaturated, gray colors and doubles with increasing saturation: $n_r = 2^{\lfloor 6-r \rfloor}$. This yields 127 bins—64 bins for the most saturated colors. Color values exceeding 95% or falling below 5% of the maximally

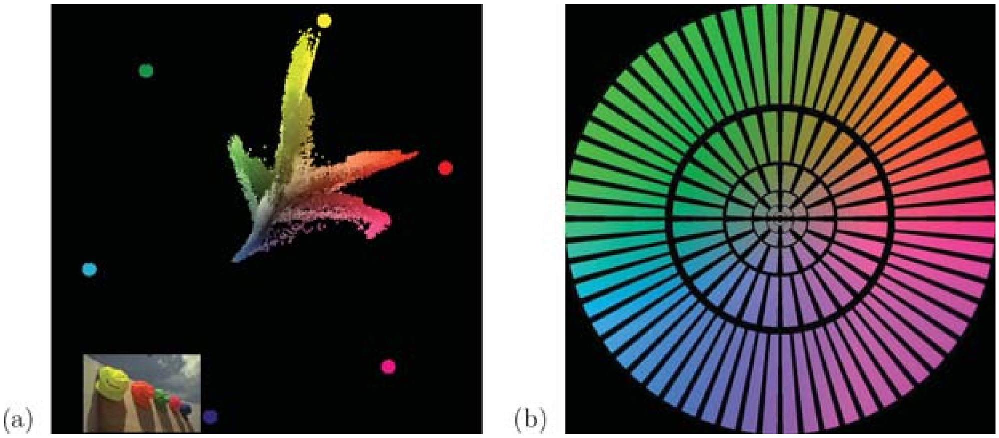

Fig. 1. Color distribution in the DKL space. (a) The color distribution of the image with hats (bottom left) plotted with respect to the coordinates at the color-opponent axes in DKL color space. When multiple pixels had the same position, luminance was averaged. (b) Subdivision of the chromaticity subspace of DKL color space into 127 logarithmic-radial color bins used for the color indexes.

possible luminance value of RGB space were classified as white and black. This is essential, since the hue resolution for very dark and very bright colors is limited.

To calculate the two color indexes, the color distributions of the images were mapped into these 129 bins. For each image two vectors were stored: the frequency of the color tones f and the average luminance level of the pixels in each bin l. If a bin was empty, the luminance level was set to zero.

$$f_{bin} = P\left[pixels \,|\, bin\right] \tag{1}$$

$$l_{bin} = \begin{cases} E\left[lum(pixels) \,|\, bin\right] & \text{if } f_{bin} \neq 0 \\ 0 & \text{if } f_{bin} = 0 \end{cases} \tag{2}$$

Figure 2a shows the frequency of the color bins for the images with the hats. The luminance of the segments in the chart indicates their frequency. The most frequent segment is white. In Figure 2b, the color tones belonging to each color bin were plotted with the average luminance of that bin illustrating the information stored in the luminance vector.

2.3 Texture Histogram

In the visual cortex, spatial information is extracted using an array of linear and nonlinear filters tuned to spatial frequency and orientation [Hubel and Wiesel 1968]. The tuning characteristics of these filters seem to be highly optimized for viewing our typical visual environment [Field 1987]. The efficiency of these spatial filters was modeled by Watson [1987] with a "Cortex" transform. He transformed the grayscale image using the discrete Fourier transformation (DFT) and applied a filter to the coefficient matrix with the same orientation and bandwidth properties found in early spatial vision. In a psychophysical experiment, he found that a reconstructed image is typically indistinguishable from the original at a code size of 1 bit/pixel.

To extract spatial information, we used the luminance dimension only, which, in the DKL color space, is orthogonal to the color-opponent subspace used for the color indexes. For each image, the Fourier power coefficients were calculated using DFT. The two-dimensional (2D) power spectrum was further segmented into bins representing spatial contrasts of distinct orientation and frequency ranges. The resolution of the Fourier index is a function of frequency and orientation, representing the orientation

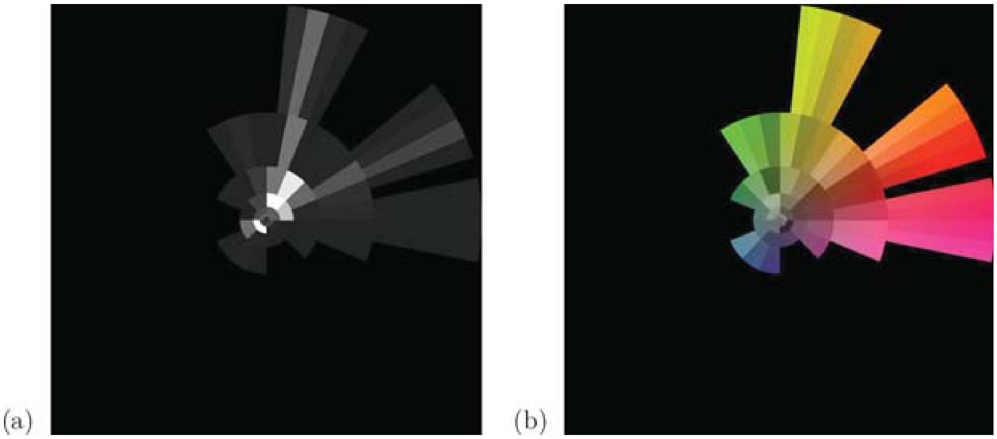

Fig. 2. Color histograms. (a) Frequency of the color bins for the image with the hats (Figure 1a). The brightness of the bins is proportional to their frequency in the image. The most frequent bin is white. (b) Mean luminance in the color bins for the image with the hats. The color tones belonging to each color bin are plotted with the average luminance level of that bin. The frequency for the "black" and "white" bins are not shown.

of high-frequency contrasts more precisely than of lower frequencies, just as in the "Cortex" model by Watson [1987].

The spatial frequency histogram was constructed in analogy to the spatial processing in the primary mammalian visual cortex. Each cell extracts information about local orientation at a particular spatial scale. This is frequently modeled by convolving the image with 2D Gabor patches (Eq. 3). The convolution with a Gabor function of a particular orientation θ and spatial scale ω can be efficiently computed in Fourier space. The Fourier transform of a Gabor function is a (scaled and shifted) Gaussian. The center of the Gaussian in Fourier space (u_0, v_0) corresponds to the optimal spatial frequency ω and the orientation θ of the Gabor patch. Thus, the luminance energy that is filtered by a set of Gabor filters $g_{\sigma,\omega,\theta}$ is contained in a circumscribed region of the 2D Fourier transform of the image. The squared Fourier coefficient, the energy (specifically its distribution the Fourier energy spectrum) was then used to construct the spatial index (Eq. 5).

$$g_{\sigma,\omega,\theta}(x, y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2+y^2}{2\sigma^2}} e^{i\cdot\omega(x\cos\theta+y\sin\theta)} \tag{3}$$

$$F_{\sigma,\omega,\theta}(x, y) = (Lum * g_{\sigma,\omega,\theta})(x, y) \tag{4}$$

$$|\mathcal{F}\circ F_{\sigma,\omega,\theta}|^2 = \left|\int_u\int_v (Lum\cdot G_{\sigma,\omega,\theta})(u, v)\,du\,dv\right|^2 \tag{5}$$

The spectrum was divided into radial-logarithmic bins similarly to the chromaticity segments. Each of the 126 segments represents contrasts of distinct orientation and frequency ranges. The bins at the origin correspond to contrasts with very low or zero frequency and store the mean luminance of the image. The resolution of the histogram for spatial scale is highest for lower frequencies and orientation is best represented in the high-frequency band. The chart of the Fourier energy index for the image with the hats is shown in Figure 3.

To determine the distribution of Fourier energy, we used the 2D discrete Fourier transform for the luminance dimension of the DKL space. Since the grayscale values of the images are real numbers,

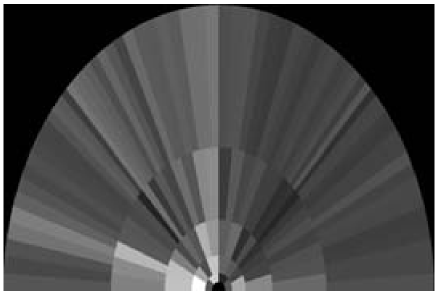

Fig. 3. Fourier index for the image with the hats (Figure 1a). Spatial frequency increases from the center to the edge of the chart. The central bins containing average luminance information were removed. The brightness of a bin is proportional to the logarithm of the average energy within the bin. The segment with the highest energy is white.

the resulting Fourier spectrum is symmetric and for the index creation only the upper half of the spectrum was used. To reduce the FFT computing time, the 768 × 512 sized images were rescaled to 96 × 64 thumbnails using bilinear approximation. The thumbnails were then projected onto the luminance dimension of the DKL space. To correct for artifacts that arise from the rectangular form of the images, the grayscale images were multiplied with a circular binary mask with a radius of 32 pixels prior to Fourier transform. This leads to a more uniform distribution of Fourier energy across orientations, since it removes the artifacts from the image borders. Orthogonal components are still most frequent, because many images contain objects with vertical or horizontal orientations. Since we wanted to construct a luminance independent spatial index, the DC bins were not used for searching.

2.4 Distance Norms

To find similar images for a given query image, it is necessary to compare the feature vectors. Most indexing systems interpret the indexes as points in an Euclidian space and use it to define the distance between two images. We used the Euclidian norm for comparing the luminance and the Fourier index.

For the chromaticity frequency index, the more intuitive intersection norm was used. It defines the similarity between two images by the sum of the minimum of the corresponding bin frequencies [$s_i = \sum_{i=1}^{n} \min(x_i, y_i)$]. The value can be interpreted as the proportion the two color distributions share. For example, a value of 0.75 means that for 75% of the pixels in one image, there exist pixels in the other image, that fall into corresponding color bins. Figure 4 shows the percentage of color concordance for a sample query using the chromaticity index.

2.5 Cue Combination

It is necessary to normalize the distance functions before combining the different indexes. We used the z-transform to normalize the distance values (Figure 5). The use of the z-transform seems reasonable insofar as the distance distributions we inspected manually show the form of a slightly skewed Gaussian density function. The parameters of the z-transform are estimated for each query using a set of 1000 randomly selected images ($z_d = (d - \bar{d}_{1000})/\hat{\sigma}_{d_{1000}}$).

It is necessary to estimate the parameters for each query image separately because the (query) images differ in their mean distances to all images I and in the variance of the distance distributions. These values can be used to characterize the image. If $\bar{d}_s(q, I)$ is large and $\sigma_s(q, I)$ is small, then the image q would be quite distinct with regard to the similarity metric s used for the comparisons. If an image is compared using multiple indexes then the z-score—calculated for each index separately—ranks the indexes in terms of their salience. The distance of an image to itself is zero, by definition. Thus, the z-value of an image to itself is the (negative) distance to the mean of the database in terms of the

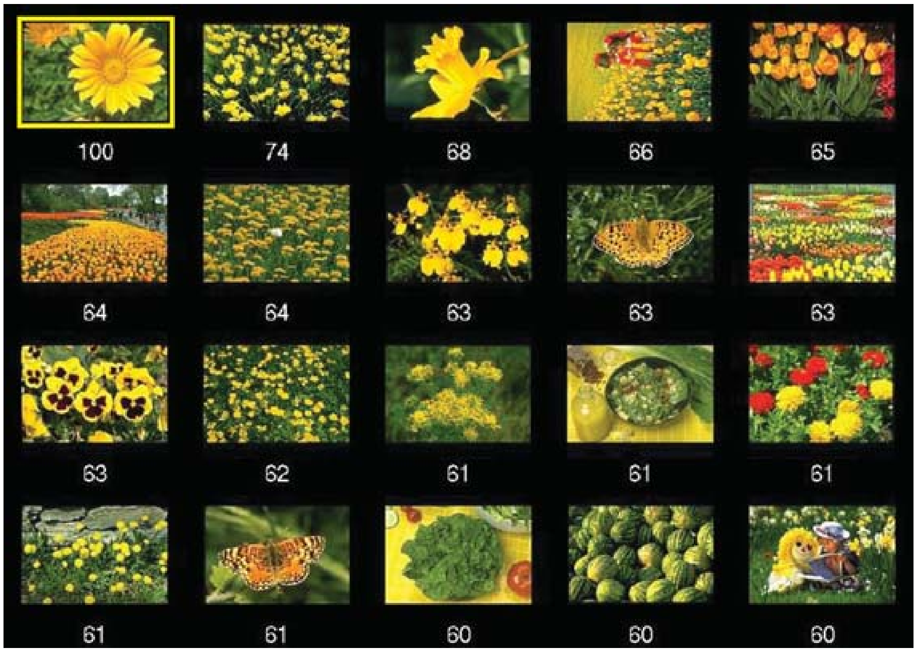

Fig. 4. Sample query results comparing images using the chromaticity histogram and the intersection norm. The number below the images shows the percentage of pixels the images shares with the query image at the top left. The query images (top left) shares 100% of the color distribution with itself.

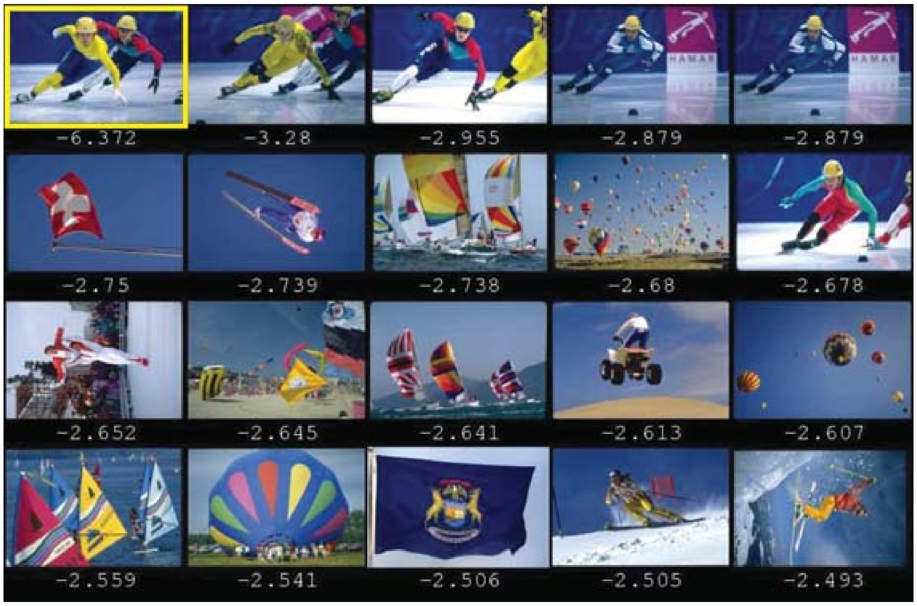

Fig. 5. Sample query results for a combined search using z-transform. Below each image, the average of the z-values of the chromaticity, the luminance, and the Fourier histogram distances is shown. The query image is shown at the top left.

standard deviation. If the image is quite distinct from the images in the database, then the absolute z-distance is large and the z-transform automatically weights the indexes that are more characteristic for a higher image.

$$z_S(q,i) = \frac{1}{|S|}\sum_{s\in S}\frac{d_s(q,i) - \bar{d}_s(q,I)}{\sigma_s(q,I)} \tag{6}$$

2.6 Implementation

The algorithms were implemented into a Java framework. The image database can address image files in a local database and on http or ftp servers. For a query, the user can select the different indexes and combine them with filters (e.g., removal of DC components in the Fourier spectrum) and one of the two distance metrics. The program can be run either as a standalone program or as a client-server combination, e.g., client applets in web browsers and a central search server. For the 2D discrete Fourier transformation and the bilinear rescaling, we used the "hips" image software package [Landy et al. 1984].

For the database of 60,000 images used in the experiments, it took between 1 and 2 s to find the best 100 matches on a 750 MHz Pentium-III-based computer running the standard Java implementation of Linux. When all three indexes were combined, a query was processed in less than 5 s. It should be noted, that we only used a linear search strategy, not taking advantage of possible bounding assumptions and tree-based data structures.

3. PSYCHOPHYSICAL EVALUATION

At present, content-based image retrieval (CBIR) systems are usually characterized by two values: the percentage of relevant images in the database that is returned in query (recall) and the proportion of related images in the result set (precision). Whether an image is considered as relevant to a particular query is highly subjective and, in many studies, only defined by the implicit criteria of one or a few observers. The size of the image databases is usually very small, often not more than a few hundred or thousand images. For these reasons, the recall-precision curves are relatively difficult to interpret quantitatively, although they are frequently used to show that a new feature retrieves a higher percentage of relevant images than another metric, using a particular image database and relevance criteria. However, perceived similarity is highly context dependent and it might not be a desirable goal to develop one universal feature set for image similarity. For application in different contexts and for relevance feedback systems, it will be necessary to reliably measure the relationship between perceived similarity and the feature metrics in a quantitative and strictly objective manner.

3.1 Measuring Similarity

The similarity between two images can be assessed by different methods. Previously, for the evaluation of the PicHunter system [Papathomas et al. 2001], absolute and relative rating scales have been considered. In the absolute similarity case, the user indicated the similarity between two images on a 5-point scale. In the relative case, the user judged the degree of similarity between the query image and two test images on a rating scale. "0" indicated that the left image is most similar and "4" that the right was a better match. The absolute and relative rating methods are highly correlated; the relationship between both measures follows the form of a psychometric function [Papathomas et al. 2001]. Yet, the use of rating scales requires the users to adjust their scale to the content of the database prior to the experiment, e.g., by seeing a list of random images.

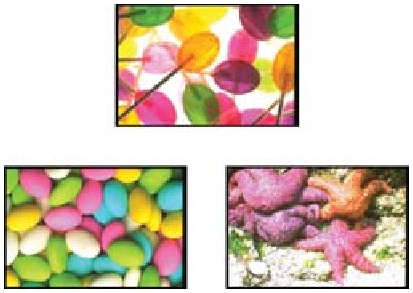

Fig. 6. 2AFC display used in the experiments. At the top, the query image is show; below are two test images (target and distractor positions were randomized).

A simpler method is the two-alternative forced-choice (2AFC) design. We used a configuration with three images: the query image at the top together with two test images below (Figure 6). The subjects were asked to compare the similarity of the test images with the query image and to select the image that is more similar. We defined the similarity between the query image and a target image $p(q, i)$ as the probability of preferring the image i. For a given query image q, the image i can be chosen to be either relatively similar or quite dissimilar to q. If the similarity metric s corresponds to the observers', then varying the computed degree of similarity $s(q, i)$ should have an influence on the selection rate. For weak indexing approaches, however, the observed correlation would be low. The relationship between the computed similarity $s(q, i)$ and the rate of preferring the image $p(q, i)$ can, therefore, be used to evaluate the similarity metric.

The 2AFC design has several advantages over the classical evaluation approaches. First, it is objective and independent of any criteria the observer has to apply in yes/no or rating tasks. We did not instruct the subjects to use any particular comparison criteria. Choosing one of the two images that is more similar was generally considered an easy task and most subjects finished the 900 comparisons in about 45 min, at a self-paced rate of about 3 s per image. Second, the 2AFC design allows the manipulation of the relative similarity between the images. The two test images can be either chosen to be relatively similar, or one or both could be relatively distinct from the query images. Currently most similarity measures are only evaluated for the first best-matched images using precision information. However, little is known whether histogram-based indexing can be used to compare the similarity of relatively distinct images. In the experiments, we varied the similarity between the target and the query images over a broad range, from best matches to rank 2000. Third, the results relate the similarity norm to a probability. The results could be directly incorporated into Bayes models of the user [Cox et al. 2000] and relevance feedback systems. Most systems use arbitrary functions for this purpose.

We evaluated our system in three experiments. In the first, the evaluation procedure itself was validated. One image, the target image, had either rank 1, 2, 20, 200, 2000, or was random. The other image, the distractor image, was chosen to be either very similar (rank 1), moderately similar (rank 2000), or random. For the three distractor conditions, the functional relationship between the similarity rank and the observers' judgment was measured.

In a second experiment, the three indexes were compared. The indexes store distinct information: the chromaticity histogram encodes the frequency of the color bins, the luminance histograms stores average luminance levels, and both color histograms do not contain spatial information. From the Fourier index, the bins containing average luminance information were removed. Therefore, it only stores information about orientation and spatial frequency. We wanted to know which index would be most suitable for image-indexing and to what extent the information sources (chromaticity, luminance, spatial) contribute to the perceived similarity.

In the third experiment the indexes' information were combined. We tested whether the chromaticity histogram can be enhanced by using the spatial information and whether the luminance index would further improve, or worsen the concordance with the observers' judgments.

3.2 Method

For all experiments a large commercial database (Corel Corp.) of 60,000 digitized photographs was used. It contains a wide range of themes each consisting of 100 images. The images show, for example, natural and man-made objects, landscapes and close-ups, and were photographed under natural conditions and artificial illumination. From the database, 900 query images were randomly selected for each experiment. For each of these images, the 2000 best matching images for each relevant index or index combination were retrieved. The images with the rank numbers 1, 2, 20, 200, and 2000 were selected for the experiments. In addition, the target image could be a random image.

To achieve comparability across experiments, the distractor was always determined by the color histogram and the intersection norm. With the exception of experiment one, where the influence of the distractor similarity was investigated, the rank of the distractor image was always 200. The position of the target/distractor (left or right), the order of query images, and of the similarity conditions were randomized per subject (mixed design).

The images were displayed on a 21-in. computer monitor (Sony GDM-F500) with a resolution of 1280 × 1024 pixels on a 50% gray background. The experiments were self-paced without decision time limits and lasted between 45 and 60 min. The subjects were 15 undergraduates who were obliged to participate in experiments for their curriculum and were naive with respect to the experiments. They were instructed to compare the similarity of the two test images with the image at the top and to decide which image is more similar. Subjects were told that there is no "true or false" and to judge intuitively if unsure. The answers were given by clicking the left or right mouse button.

3.3 Experiment 1: Influence of Distractor Similarity

In the first experiment, the relationship between the similarity computed with the chromaticity histogram and the human judgments was measured. As a second factor, the similarity of the distractor image was varied to determine whether the degree of similarity between the two test images would influence the preference rates.

The image similarities were calculated using the chromaticity histogram and the intersection norm. The rank of the distractor image was either 1 (best match), 200, or random. The rank of the target image was 1, 2, 20, 200, 2000, or random. This results in 3 × 5 conditions, the points of equal target and the distractor ranks were not measured.

The results show a strong correlation between the logarithmic rank of the test image and the probability of preference. In Figure 7, the similarity rank of the target image is plotted along the x axis. For each of the three distractor ranks, the rate of preferring the target is shown on the y axis. The hypothetical points of equal similarity between the target and the distractor image are marked by filled circles. If the distractor image is random, then the subjects statistically prefer the histogram-selected target image for all rank positions (top line of Figure 7). The lowest selection rate of 0.5 could theoretically be reached if the distractor and the target image are both random, as indicated by the right-most filled circle. The image evaluated as being most similar by our indexing system was selected as being more similar more than 80% of the time. If a distractor image is chosen that is deemed more similar by the indexing system (rank 200 or rank 1), then the curves are shifted downward: for the middle curve, the distractor had rank 200; for the lower curve, the distractor was always the best match.

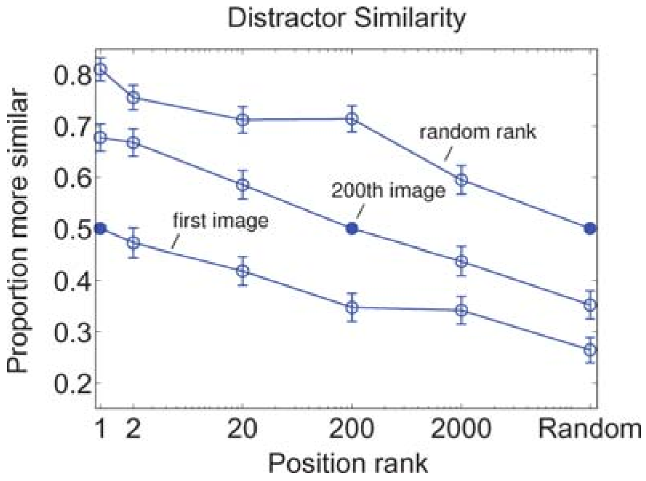

Fig. 7. Probability of preferring the target image against the distractor image as a function of similarity between the target and the query image. The distractor rank was varied from top to bottom: random, 200 and 1 (best match). The filled circles indicate hypothetic points of equal similarity of distractor and target image with expected probability. 0.5 and, therefore, define the rank of the distractor image (from left to right: 1, 200, random). The similarity distance was calculated using the chromaticity index and the intersection metric. The standard error is shown for each point.

Table I. Regression (slope b and intercept a) and Correlation Coefficients for the Relationship between Logarithmic Target Rank and Concordance for Different Distractor Ranks

Distractor Rank	b	a	r
1	−.022	.49	.985
200	−.032	.68	.999
Random	−.027	.80	.965

The correlation coefficient between the logarithmic rank and the percentage of concordance is very high ($r > 0.95$) and significant ($p < 0.05$) for all conditions. Table I shows the regression (b, a) and the correlation (r) coefficients between the logarithmic rank and target preference.

$$p_{subj} = b\log(rank\{z_q[s(q,i)]\}) + a + \xi_{subj} \tag{7}$$

Changing the distractor similarity shifts the functions by a constant amount and does not alter the slope of the linear regression. This is an important finding since it shows that the subjects' do not change their judgments if one image is either very similar or random. Therefore, the distrator image in the 2AFC design is not of critical importance for the evaluation of the indexes. For the following experiments, we kept the distractor similarity constant. We decided to use the medium distractor similarity (rank 200) to optimally use the range.

3.4 Experiment 2: Index Comparison

In experiment 2, we wanted to know to what extent the other information sources (luminance, orientation, and spation frequency) contribute to the similarity judgments. The target images were chosen by using one of the three histograms and their ranks were again 1, 2, 20, 2000, or random. The distractor image was determined with the chromaticity histogram and had rank 200. For the chromaticity histogram the minimum norm was used, the distances between the luminance and Fourier vectors were calculated with the Euclidian norm.

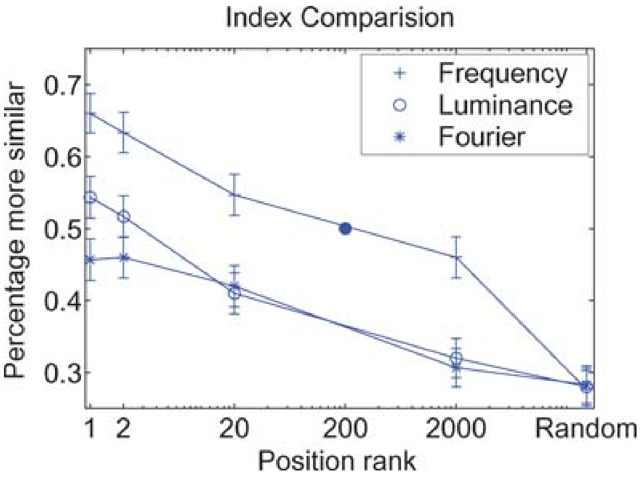

Fig. 8. Relationship between the probability of preferring the target image against the distractor image as a function the similarity between query and target image. The similarity distances were calculated using the chromaticity, luminance, and spatial index. The filled circle indicates the theoretic point where the target and the distractor image are equally similar.

The results show a strong advantage for the chromaticity histogram over the luminance and Fourier information. For relatively distinct images (rank 20, 2000) the luminance and spatial index show similar performance. However, for highly similar images (rank 1, 2) the luminance index provided better matches (Figure 8, Table II).

3.5 Experiment 3: Index Combination

The performance of the luminance and spatial index is clearly worse than the images selected with the chromaticity histogram. However, this does not imply that luminance or spatial information does not contribute to the subjects' perception of similarity. In the third experiment, we examined whether the correspondence with the observers' judgments could be improved if the spatial or the spatial and the luminance information was used in addition to the color frequencies. To combine different features, we computed the average of the z-transformed distance values as described in Section 2.5. All other variables were identical to experiment two.

Figure 9 shows a general improvement of the selection rates of the image retrieved when either spatial, or both spatial and luminance information is combined with the chromaticity histogram. The correspondence with the judgments was improved by 5% if the spatial information is used. The luminance information further enhances the concordance by 4% (see Table III). Therefore, the luminance and the spatial index contribute to image similarity independent of the chromaticity histogram.

The overall improvement of 9% clearly indicates that the z-transformation is a good choice for combining index information.

4. DISCUSSION

4.1 Summary

Our results show that psychophysical methods should be and can be successfully used to evaluate and compare different content-based image retrieval systems. Here, we evaluated our own rather simple system, which is based on some elementary principles of the human visual system. However, the strength of this approach lies in the applicability to different competing algorithms for image retrieval.

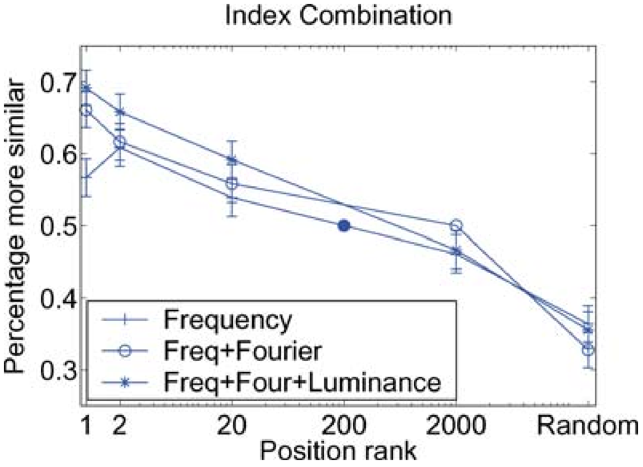

Fig. 9. Relationship between the probability of preferring the target image against the distractor image as a function of the similarity between the query and the target image. The similarity distances were calculated using the chromaticity, the chromaticity and the spatial, and all three indexes. The filled circle indicates the theoretic point where the target and distractor are equally similar. The standard error is shown for each data point.

Table II. Regression (Slope b and Intercept a) and Correlation Coefficients for the Relationship between Logarithmic Target Rank and Concordance for the Different Indexes

Index	b	a	r
Color Frequency	−.034	.66	.976
Luminance	−.025	.52	.978
Fourier spectrum	−.019	.47	.989

Table III. Regression and Correlation Coefficients for the Relationship between Logarithmic Target Rank and Concordance for Combinations of the Indexes

Index Combination	b	a	r
Freq.	−.021	.60	.967
Freq. + Fourier	−.028	.65	.959
Freq. + Fourier + Lum.	−.031	.69	.998

4.2 Perception-Based Image Indexing

Color histograms and their variations are used with great success in many current CBIR systems, ever since Ballard [Swain and Ballard 1991] introduced the concept. Under most circumstances, it probably matters very little what exact color representation is used. Ideally, linguistic labels, corresponding to the universal color names used by human observers [Berlin and Kay 1969], would be used. A straightforward approach is a simple mapping of RGB color space to color names, as was used by PicHunter [Cox et al. 2000]. This approach will fail for ojects that are photographed under different illuminations, or in front of backgrounds with a chromatic bias. Under these circumstances, color constancy algorithms can improve the retrieval quality [Funt and Finlayson 1995; Gevers and Smeulders 1996; Alferez and Wang 1999]. Of course, if camera settings are used that include white balancing, then the camera might already deliver a good approximation to human color constancy.

One has to keep in mind that the assignment of names to colored objects probably happens at a rather late stage of visual processing [De Valois and De Valois 1993]. If an earlier level representation with more than just 7–11 categories is desired, the code used in the primate visual system at the second stage of color vision, color opponency, might be close to optimal. This code removes the correlation between the three different cone photoreceptors [Buchsbaum and Gottschalk 1983; Zaidi 1997]. It has been shown in a variety of psychophysical experiments that the axes of this space are independent for a variety of visual tasks [Krauskopf 2001]. Since the bin sizes we used were proportional to color discrimination thresholds obtained by Krauskopf and Gegenfurtner [1992], it can be argued that the code is optimal in the sense of requiring the least number of bins to represent all possible discriminable colors. It should also be noted that this color space is not identical to other color opponent spaces, such as HSI. However, in all likelihood the exact representation of colors will not make a significant difference.

In addition to the frequency-encoding chromaticity histogram, we stored the average luminance levels of the color bins in a second vector. The separate processing of chromaticity and luminance (contrast) information is part of various models of visual processing. We used the separation of chromaticity and luminance information to compare the contribution of both information sources to the similarity judgments. We found that the luminance index is correlated with the observers' judgments, too. The prediction of the perceived similarity is not as good as with the chromaticity histogram. However, luminance appears to be considered when images are compared, even if the luminance of the color tones seems not to be as important as their frequency.

The spatial information was extracted using the 2D DFT. The index was constructed to represent the distribution of orientation and spatial frequency in the image analogous to the processing of contrast information in the visual cortex. Although many texture feature sets describe similar information, Fourier analysis offers a mathematically profound way to extract the information. The explicit representation of orientation and spatial frequency in the Fourier spectrum allows the simple filtering of these dimensions. In our program, the user could, for example, select a radial filter to use the Fourier index as an orientation index only. By averaging the bins of equal frequency ranges a rotation invariant search can be conducted. Of course, there are more sophisticated ways of encoding the spatial information of images, for example, by multiple histograms [Stricker and Dimai 1997], correlograms [Huang et al. 1999], explicit texture statistics [Liu and Picard 1996] or by wavelet transforms [Manjunath and Ma 1996; Wang et al. 1997; Liang and Kuo 1999]. However, our evaluation clearly demonstrates that our index does capture perceptual similarity.

4.3 Evaluation

Currently, CBIR systems are evaluated by precision and, sometimes, recall measures. Relevant images are, in the majority of the publications, defined by similarity judgments of a single person deciding whether an image belongs to the same category as the query image, or is in some unspecified way "similar" to it. The image databases used for evaluation typically contain between a hundred and some thousand images.

An exception is the well-evaluated PicHunter system [Cox et al. 2000; Papathomas et al. 2001]. The system includes simple features like the image height and width, color histograms, color autocorrelogram, a color-coherence vector, and, for a subset of the images, semantic annotations. The features were tested in a forced-choice design similar to the one we used and then later compared within the target testing paradigm [Cox et al. 2000]. However, the target testing evaluation procedure used in Papathomas et al. [2001] is limited to relevance feedback systems. This is an ideal way to evaluate a complete system, but it is less ideal to measure the relationship between single features and human perception. With target testing the functional relationship between an index similarity space and the similarity space of human perception cannot be determined.

Cox et al. [2000] used a sigmoid function to transform the raw distances to a probabilistic scale. Our results suggest that the relationship can be modeled by a logarithmic function between the similarity rank and the selection probability for 2AFC designs. That relationship holds for all indexes and is not impacted by the similarity of the distractor image.

For relevance feedback systems, it is highly important to measure the relationship between the similarity metric and the perceived similarity over the whole range of similarity, from the most similar image to random images.

4.4 Conclusion

In summary, we have shown that the psychophysically based indexes are effective in finding similar images. The indexes were constructed in accordance with some of the known properties of the early stages of human vision. The color codes in the "red-green" and "blue-yellow" channels were modeled using the color-opponent axes of the DKL color space and a logarithmic-radial scaling for the histogram bins. The luminance information was stored in a separate index. The 2D discrete Fourier transform was used to create an orientation and spatial frequency histogram analogous to similar representations in the visual cortex. Most importantly, we evaluated the indexing system with a strictly quantitative and objective approach using a large, heterogeneous database of 60,000 digitized photographs.

REFERENCES

Alferez, R. and Wang, Y.-F. 1999. Geometric and illumination invariants for object recognition. *IEEE Trans. PAMI 21*, 6 (June), 505–536.

Berlin, B. and Kay, P. 1969. *Basic Color Terms: Their Universality and Evolution*. University of California Press, Berkeley, CA.

Buchsbaum, G. and Gottschalk, A. 1983. Trichromacy opponent color coding and color transmission in the retina. In *Proc. Roy. Soc. Lond. (B). 220*, 89–113.

Cox, I. J., Miller, M. L., Minka, T. P., Papathomas, T. V., and Yianilos, P. N. 2000. The bayesian image retrieval system, pichunter: Theory, implementation, and psychophysical experiments. *IEEE Transactions on Image Processing 9*, 1, 20–37.

De Valois, R. and De Valois, K. 1993. A multi-stage color model. *Vision Research 33*, 1053–1065.

Derrington, A. M., Krauskopf, J., and Lennie, P. 1984. Chromatic mechanisms in lateral geniculate nucleus of macque. *J. Physiol. 357*, 241–265.

D'Zmura, M. and Lennie, P. 1986. Mechanisms of color constancy. *Journal of the Optical Society of America a-Optics Image Science and Vision 3*, 10, 1662–1672.

Field, D. 1987. Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Amer. 4*, 12, 2379–2394.

Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., and Yanker, P. 1995. Query by image and video content: The qbic system. *IEEE Computer Magazine 28*, 23–32.

Funt, B. V. and Finlayson, G. D. 1995. Color constant color indexing. *IEEE Trans. PAMI 15*, 5 (May), 522–529.

Gegenfurtner, K. and Kiper, D. 2003. Color vision. *Annual Review of Neuroscience 26*, 181–206.

Gevers, T. and Smeulders, A. 1996. A comparative study of several color models for color image invariant retreival. In *Proc. 1st Int. Workshop on Image Databases and Multimedia Search*. Amsterdam, Netherlands, 17.

Hering, E. 1964. *Outlines of a Theory of the Light Sense*. Harvard University Press, Cambridge, MA.

Huang, J., Kumar, S., Mitra, M., Zhu, W., and Zabih, R. 1999. Spatial color indexing and applications. *International Journal of Computer Vision 35*, 3, 245–268.

Hubel, D. and Wiesel, T. 1998. Early exploration of the visual cortex. *Neuron 20*, 3, 401–412.

Hurvich, L. M. and Jameson, D. 1957. An opponent-process theory of color-vision. *Psychological Review 64*, 6, 384–404.

Iqbal, Q. and Aggarwal, J. K. 2002. Combining structure, color and texture for image retrieval: A performance evaluation. In *Proceedings of the 16th Intern Conf Pattern Recognition (ICPR)* (Aug 11–15, 2002). 438–443.

Krauskopf, J. 2001. *Color Vision: From Genes to Perception*. K. R. Gegenfurtner & L. T. Sharpe, Eds. Cambridge University Press, London. 303–316.

KRAUSKOPF, J. AND GEGENFURTNER, K. 1992. Color discrimination and adaptation. *Vision Research 32*, 11, 2165–2175.

LANDY, M. S., COHEN, Y., AND SPERLING, G. 1984. Hips: a Unix-based image processing system. *Computer Vision, Graphics, and Image Processing 25*, 3 (Mar.), 331–347.

LI, B. T., CHANG, E., AND WU, Y. 2003. Discovery of a perceptual distance function for measuring image similarity. *Multimedia Systems 8*, 6, 512–522.

LIANG, K.-C. AND KUO, C.-C. J. 1999. Waveguide: A joint wavelet-based image representation and description system. *IEEE Trans. Image Processing 8*, 11.

LIU, F. AND PICARD, R. W. 1996. Periodicity, directionality, and randomness: Wold features for image modeling and retrieval. *IEEE Trans. PAMI 18*, 7, 517–549.

MALJKOVIC, V. AND MARTINI, P. 2005. Short-term memory for scenes with affective content. *Journal of Vision 5*, 3, 215–229.

MANJUNATH, B. S. AND MA, M. Y. 1996. Texture features for browsing and retrieval of image data. *IEEE Trans. PAMI 18*, 8 (Aug.), 837–842.

OLIVA, A. AND TORRALBA, A. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision 42*, 3, 145–175.

PAPATHOMAS, T., COX, I., YIANILOS, P., MILLER, M., MINKA, T., CONWAY, T., AND GHOSN, J. 2001. Psychophysical experiments on the PicHunter image retrieval system. *Journal of Electronic Imaging 10*, 1, 170–180.

ROGOWITZ, B. E., FRESE, T., SMITH, J. R., BOUMAN, C. A., AND KALIN, E. 1998. Perceptual image similarity experiments. *Human Vision and Electronic Imaging III 3299*, 1, 576–590.

RUDERMAN, D. L., CRONIN, T. W., AND CHIAO, C.-C. 1998. Statistics of cone responses to natural images: implications for visual coding. *J. Opt. Soc. America A 15*, 2036–2045.

RUI, Y., HUANG, T. S., ORTEGA, M., AND MEHROTRA, S. 1998. Relevance feedback: A power tool for interactive content-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology 8*, 5, 644–655.

RUI, Y., HUANG, T., AND CHANG, S. 1999. Image retrieval: current techniques, promising directions and open issues. *Journal of Visual Communication and Image Representation 10*, 39–62.

SALTON, G. AND BUCKLEY, C. 1990. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science 41*, 4, 288–297.

STRICKER, M. AND DIMAI, A. 1997. Spectral covariance and fuzzy regions for image-indexing. *Machine Vision and Applications 10*, 2, 66–73.

SWAIN, M. J. AND BALLARD, D. H. 1991. Color indexing. *International Journal of Computer Vision 7*, 1, 11–32.

WANDELL, B. 1993. Color appearance: The effects of illumination and spatial resolution. *Proc. Nat. Acad. Sci.* 90. 1494–1501.

WANG, J. Z., WIEDERHOLD, G., FIRSCHEIN, O., AND WEI, S. X. 1997. Content-based image-indexing and searching using daubechies' wavelets. *Int. J. Digit. Libr. 1*, 311–328.

WATSON, A. B. 1987. Efficiency of a model human image code. *J. Opt. Soc. Am. 4*, 12, 2401–2417.

WÄSSLE, H. AND BOYCOTT, B. B. 1991. Functional architecture of the mammalian retina. *Physiol Rev 71*, 447–480.

ZAIDI, Q. 1997. Decorrelation of L- and M-cone signals. *J. Opt. Soc. America A 14*, 3430–3431.

Received June 2004; revised May 2005; accepted October 2005