Content-Based Image Retrieval Benchmarking: Utilizing Color Categories and Color Distributions

Egon L. van den Broek, Peter M. F. Kisters and Louis G. Vuurpijl
Nijmegen Institute for Cognition and Information, Radboud University Nijmegen, THE NETHERLANDS

From a human centered perspective three ingredients for Content-Based Image Retrieval (CBIR) were developed. First, with their existence confirmed by experimental data, 11 color categories were utilized for CBIR and used as input for a new color space segmentation technique. The complete HSI color space was divided into 11 segments (or bins), resulting in a unique CBIR 11 color quantization scheme. Second, a new weighted similarity function was introduced. It exploits within bin statistics, describing the distribution of color within a bin. Third, a new CBIR benchmark was successfully used to evaluate both new techniques. Based on the 4050 queries judged by the users, the 11 bin color quantization proved to be useful for CBIR purposes. Moreover, the new weighted similarity function significantly improved retrieval performance, according to the users.

Journal of Imaging Science and Technology 49: 293–301 (2005)

Introduction

Digital media are rapidly replacing their analog counterparts. Less than 10 years ago a digital photo camera was solely used in professional environments.[1] In contrast, nowadays many home users own a digital photo camera. This development is accompanied by (i) the increasing number of images present on the internet, (ii) the availability of the internet for an increasing number of people, and (iii) a decline in digital storage costs.

As a result, the need for browsing image collections has emerged. This development gave birth to a new field within Information Retrieval (IR): image retrieval. When images are part of a web page or when images are textually annotated in another form, IR-techniques can be utilized. However, how do we search annotated? We will first discuss quantitative arguments, followed by qualitative arguments that point out the relevance of Content-Based Image Retrieval (CBIR). CBIR uses features of the image itself, i.e., color, texture, shape, and spatial characteristics, which enables us to search for images that are not textually annotated.

Murray[2] determined in his article, “Sizing the internet”, on July 10, 2000, that 2.1 billion unique pages were present on the internet. He further states that “internet growth is accelerating, indicating that the internet has not yet reached its highest growth period.” Currently, estimates of the number of unique pages range from over 50 million[3] up to over 8 billion.[4]

In addition Murray determined the average number of images present on a page to be 14.38. One year later Kanungo et al.[5] drew a sample of internet (consisting of 862 pages) and determined the average number of images per page as 21.07. Unfortunately, neither of these papers report their definition of an image. The latter is of importance since one can make a distinction between images, e.g., cartoons and photos, and web graphics, i.e., backgrounds, bullets, arrows, and dividers. Furthermore, the size of the “invisible web”, i.e., databases available through websites. was not taken into account. From the previous facts we can derive that between 720 million and 168 billion images are present on the internet. Due to a lack of statistical data, we can not make an estimation of the two other sources of images: the “invisible internet’ and individual users’ private image collections. However, it is safe to say that these two latter sources of images will increase the number of images substantially.

Next to the quantitative argument as discussed above, a qualitative argument can be made that illustrates the importance of CBIR. Let P be a square of pixels. P either consists of characters c or is an image i. Let i be a graphic itemizing bullet, typically of size 8^2 to 26^2 pixels. Let c be the word “the”. Using a standard font size the word “the” needs a square of 17^2 pixels, which equals the average size of a bullet. However, a graphic itemizing bullet can, for example, resemble “the footprint of a bear”* using as many pixels but having a much richer semantic content. So, the saying “a picture is worth more than a thousand words” holds when considering the semantics that can be expressed per unit area.

Original manuscript received June 3, 2004

†Corresponding Author: E. L. van den Broek, egon@few.vu.nl; current address: Department of Artificial Intelligence, Vrije Universiteit Amsterdam, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

*Text and bullet are present on: http://www.w3schools.com/graphics/graphics_bullets.asp [Last accessed on April 17, 2005]

Based on these considerations we conclude that CBIR is of significant importance for IR in general, for retrieval of unannotated images in image databases, and for home users that manage their own image, e.g., photo, collections.

However, as Smeulders et al.[6] noted in 2000 "CBIR is at the end of its early years" and is certainly not the answer to all problems. To mention a few, CBIR engines are not capable of searching beyond a closed domain, are computationally too expensive, have a low retrieval performance, and do not yield results that match the needs of the user. Therefore, the CBIR techniques used are still subject of development.

Due to the large differences between users' search strategies, even interactive user-adaptable CBIR engines have been proposed.[7,8] Such an approach is as useful as it is complex. We will attempt to find out whether such an approach is needed at this moment in CBIR development.

Our approach to improve the performance of CBIR systems is through the utilization of knowledge concerning human cognitive capabilities. In our research the distinction is made between query-by-memory and *query-by-example*, each requiring cognitive processes. With *query-by-memory* the user has to define image features by memory, whereas in case of query-by-example an example image is supposed to be present. A CBIR engine uses features such as shape, spatial characteristics, texture, and color to explore the image content. The current research will focus on the last of these features: color. It will be shown in this article that human color categories can be utilized for CBIR techniques.

Human Color Categories

Human color perception is a complex function of context. For example: illumination, memory, object identity, culture, and emotion all take part in the process.[9–11] As mentioned by Forsyth and Ponse,[12] "It is surprisingly difficult to predict what colors a human will see in a complex scene; this is one of the many difficulties that make it hard to produce really good color reproduction systems." We are not even close to having a good model for human color perception.

Therefore, we have chosen to consider color in CBIR from another perspective, that of the focal colors or color categories (see also the World Color Survey[13]): black, white, red, green, yellow, blue, brown, purple, pink, orange, and gray.[14–16] Note that these color names do not resemble a particular color, but rather a fuzzy notion of some set of colors: a color category.

People use these categories when thinking of or speaking about colors or when they recall colors from memory. Research from various fields of science emphasizes the importance of focal colors in human color perception. The use of this knowledge may provide the means for bridging the semantic gap that exists in CBIR. In literature[9,14–16] three advantages of the 11 basic color categories are mentioned:

1. They are robust to variability among people, i.e., different people perceive colors differently;
2. They are robust to variability within individual observers, e.g., the perception of color by a single person changes because of, for example, changing moods or attention; and
3. They are an attractive concept from a computational point of view.

Experimental Proof for the 11 Color Categories

Until recently, little experimental evidence was present for the existence of the 11 color categories. For computer environments no evidence at all was present. However, in van den Broek et al.[17] we have presented the results of our experiments on human color categorization, which form the basis of the present research.

Before the experiments were conducted, twenty-six participants were asked to perform the following task: Write down 10 colors that first come in mind. The results confirmed the existence of the 11 color categories. Next, all twenty-six subjects participated in two experiments. The stimuli for both of them comprised the full set of the 216 web-safe colors.[18] Below the stimulus (the color), 11 buttons were placed. In the so called, color memory experiment the buttons were labeled with the names of the 11 focal colors; in the so called color matching experiment each of the buttons was colored with one of the 11 focal colors. Each experiment consisted of 4 blocks of repetitions of all randomized 216 stimuli.

For a thorough description of this research we refer to Ref. 17. The main result of both experiments is a Color Look-Up Table (CLUT).[†] Summarizing, the results prove that:

- The use of color categories is valid in a CBIR context;
- A color space can be described using color categories; and
- There is a difference between color categorization using color discrimination (or matching) and color categorization using color memory.

Moreover, this research shows that humans consistently quantize colors into a limited set of clusters (the CLUT), which can be considered as an efficient model of human color categorization.

In the next section we will describe the segmentation of color space in 11 color categories, using the CLUT. After that, we will explain how this segmentation is used as a quantization scheme for CBIR and how it can be used for both query-by-memory and query-by-example.

Color Space Segmentation

Our aim was to create a matching engine that uses a color quantization scheme, which compresses all possible image colors into 11 color categories in a manner similar to human color categorization. Therefore, the color quantization scheme should be based on the CLUT (see above). However, the 216 web-safe colors, categorized by humans into the 11 color categories that make up the CLUT, did not provide enough information for using the CLUT directly as a color quantization scheme. In order to tackle this problem, we have developed a new algorithm to successfully map the 11 color categories to the complete HSI (hue, saturation, and intensity) color space.

Before it was possible to use the CLUT as input data for the segmentation process, some coordinates were removed from the CLUT. We distinguish fuzzy and non-fuzzy coordinates in the CLUT. Non-fuzzy coordinates were defined as: coordinates assigned to the same color category by at least 10 of the 26 subjects (see Ref. 17 for further details); fuzzy coordinates were less consistently assigned to a color category and, therefore, removed from the CLUT. The RGB coordinates of the non-fuzzy CLUT were converted to HSI coordinates (see Ref. 19 for the conversion algorithm).

Just like colors, the 11 color categories can be divided into two groups: the chromatic color categories, i.e., blue,

[†] The CLUT can be found at: http://www.few.vu.nl/~egon/CLUT-markers.pdf.

yellow, green, purple, pink, red, brown, and orange, and the achromatic color categories, i.e., black, gray, and white. This distinction was also applied for the segmentation of the HSI color space. Although the HSI color space has three axes, the axes enable us to distinguish between the color categories using two 2D segments, one for the chromatic and one for the achromatic colors, thereby reducing the computational complexity of the segmentation process. One of the segments comprised a hue-intensity plane that provided means to distinguish the chromatic color categories. The other segment, a saturation-intensity plane, provided means to distinguish the achromatic color categories.

The non-fuzzy HSI coordinates of the CLUT were labeled with their category and were plotted in two 2D planes. All coordinates belonging to the same color category were fully connected by a line generator. This resulted in fully connected graphs for each of the color categories. These graphs were converted to filled convex hulls for both 2D planes.

In order to segment the two 2D planes, Fast Exact Euclidean Distance (FEED) transformations[20] were applied on each of the 2D planes, resulting in a weighted distance map.[21] The FEED transformations produced two fully segmented 2D HSI weighted distance maps. Next, a hill climbing algorithm was applied to determine the edges between the color categories. These were converted by curve fitting to Fourier functions, which express the borders between the color categories in the two segmented HSI 2D planes. Henceforth, the HSI color space was segmented in 11 color categories, which defines a quantization scheme for CBIR.

Finally, we needed to validate this quantization scheme. It would be valid if it categorizes the stimuli used in the two experiments in a similar way as the subjects did. For the non-fuzzy colors a 98% match was found, which confirmed expectations since the segmented color space is based on the non-fuzzy colors. In addition, for the fuzzy colors a 98% match was found. Henceforth, we have a validated segmented HSI color space that defines an 11 bin quantization scheme for CBIR.

Query-by-Example versus Query-by-Memory in CBIR

Most CBIR engines distinguish two forms of querying: (i) query-by-example: the user provides an example image, and (ii) query-by-memory, in which the user defines features by memory, such as shape, spatial characteristics, texture, and color. For the present research, we are especially interested in the use of the color feature. In the remaining part of this article we, therefore, define query-by-memory as query-by-memory utilizing color and define query-by-example as query-by-example utilizing color.

At the basis of both query-by-example and query-by-memory, lie two distinct cognitive processes: color discrimination (or matching) and color memory. Let us illustrate the importance of this distinction for CBIR.

Imagine that a user wants to find different images of red cars. Suppose the user already possesses such an image and uses it to conduct a query-by-example. Then, images from the database will be matched to the example image by the CBIR engine. The resulting images are presented to the user. The user compares all retrieved images with his example image and with each other. In this comparison a process of color discrimination is triggered, with which humans can differentiate between millions of colors,[22] since the colors are (directly) compared to each other.

In case of query-by-memory no example image is available. The user is required to retrieve the color red from memory. In general, this will not be one particular color, but rather a fuzzy notion of some set of colors: a color category, based on color memory. Since the user wants to find different kinds of cars, each of the elements of this set (or category) of red colors will be acceptable for the user. No need is present to differentiate between several types of red. Providing the keyword red or pressing a button resembling the fuzzy set of red is sufficient. Therefore, 11 bins fit the query-by-memory paradigm.

Enhanced 11 Bin Color Quantization

The 11 bin quantization of color space was originally developed for query-by-memory. So, the user has to rely on his limited color memory when judging the retrieved images. For the query-by-example paradigm, the drastic reduction of color information to 11 color categories (or bins) is coarse.

However, query-by-example is of importance for CBIR since it has two advantages compared to query-by-memory: (i) it requires a minimal effort of the user, and (ii) it is the most widely used paradigm since all possible features (color, texture, shape, and spatial information) can be analyzed. In query-by-memory the latter is hard and partially impossible. For example, users experience it as difficult to sketch a shape[23] and are not capable of defining complex textures. Since query-by-example is such an important paradigm for CBIR, we should aim to adapt the 11 bin quantization scheme to the query-by-example paradigm.

We will now explain that instead of adopting a more precise quantization scheme, the notion of the 11 color categories should be preserved. However, a higher precision is needed for the 11 bin quantization scheme.

In van den Broek et al.[24] the 166 bin quantization (18 × 3 × 3) of HSV color space was not judged as performing significantly better in query-by-example than the 64 bin quantization (4 × 4 × 4) of HSV color space, despite the fact that the 166 bin quantization is 2.6 times more precise than the 64 bin quantization. Hence, a more precise quantization scheme is not a guarantee for success. In addition, in the same study the 11 bin color quantization performed as well as the more precise, 64 and 166 bin quantizations. So, the 11 bin quantization can be considered as an extremely efficient color quantization scheme.

The success of the 11 bin color quantization scheme can be explained by its origin: human color categorization, where the 64 and 166 bin quantization schemes naively segmented each of the three axes of HSV color space into equal segments.

One way to extend the 11 color histogram would be to divide each color category into a number of segments, for example, relative to the size of the area each category consumes in the HSI color space. However, with such an approach only the number of pixels present in a bin are taken into account; color variations within bins are ignored. Therefore, we chose to incorporate statistical information that describes the distribution of pixels within each bin.

Such an approach is only useful if a segment of color space represented by a bin is perceptually intuitive for the users. The naive 64, 166, and 4096 bin quantizations as used in previous research[24] are not perceptually intuitive for users. For these quantization schemes, the incorporation of statistical data would not make sense.

Since the statistical values can be precomputed and stored, these can be represented as a vector of size $n*a$ where n is the number of bins and a is the number of statistical values per bin. This representation is similar to the vector-representation of a histogram. Therefore, each statistical value can be represented as a virtual bin. Therefore, such an approach is relatively cheap compared to a more precise quantization.

In the next section we will describe the within bin statistical information and how it is used as a similarity measure

Similarity Function Using Within Bin Statistics

The Intersection Similarity Measure

A distance measure calculates the distance between two histograms. A distance of zero represents a perfect match. We use the histogram intersection distance (D) of Swain and Ballard[25] between a query image (q) and a target image (t):

$$D_{q,t} = \sum_{m=0}^{M-1} | h_q(m) - h_t(m) |, \tag{1}$$

where M is the total number of bins, h_q is the normalized query histogram, and h_t is the normalized target histogram.

When combining distance measures (by multiplying them), a single perfect match would result in a perfect match for the total combination. However, this is an unwanted situation since one would expect a perfect match only if all distance measures indicate a perfect match.

Therefore, the similarity, i.e., similarity = (1 – distance), for each variable is calculated, instead of its distance.

In order to determine the intersection similarity (S) we adapt Eq. (1) to give:

$$S_{q,t} = \sum_{m=0}^{M-1} 1 - | h_q(m) - h_t(m) |. \tag{2}$$

Extension of the Intersection Measure

Based on Eq. (2) a new distance measure is developed, incorporating statistical infor-mation of each color category separately. In histogram matching only the magnitudes of the bins are of importance, i.e., the number of pixels assigned to each bin. However, for our new distance measure we will use five values in addition to the number of pixels in the bins.

These values are stored in a *color bucket b*, assigned to every color category (or quantized color space segment):

$$\left\{ \begin{array}{ll} x_1(b) = \#(b), & \text{i.e., the number of pixels in bucket } b\text{; the} \\ & \text{original histogram value } h \\ x_2(b) = \mu H(b), & \text{i.e., the mean hue } H \text{ of bucket } b \\ x_3(b) = \mu S(b), & \text{i.e., the mean saturation } S \text{ of bucket } b \\ x_4(b) = \sigma H(b), & \text{i.e., the standard deviation of the hue val-} \\ & \text{ues } H \text{ in bucket } b \\ x_5(b) = \sigma S(b), & \text{i.e., the standard deviation of the satura-} \\ & \text{tion values } S \text{ in bucket } b \\ x_6(b) = \sigma I(b), & \text{i.e., the standard deviation of the inten-} \\ & \text{sity values } I \text{ in bucket } b, \end{array} \right.$$

where $x_i(b)$ denotes value i of color bucket b of either query image q: $q_i(b)$ or of target image t: $t_i(b)$.

For each pair q_i and t_i (with $i \in \{1, 6\}$), of each bucket b the similarity S_{q_i,t_i} is determined, as follows:

$$S_{q_i,t_i}(b) = 1 - | q_i(b) - t_i(b) |, \tag{3}$$

where the range of S_{q_i,t_i} is [0,1].

For the buckets representing the achromatic color categories, no values were calculated for the hue and saturation axis. The achromatic color categories are situated in the central rod of the HSI model. Hue values are represented by the angle around this rod (indicating the basic color). Saturation values refer to the distance of a point to the central rod. The larger the distance, the stronger the color information of a certain hue, is present.

Achromatic colors show very small values for saturation, regardless of their hue angle. Therefore, when referring to achromatic categories, statistical information about the hue and saturation axis does not contribute to the precision of the search algorithm and is, therefore, ignored in the algorithm. To achieve the latter, by definition $\mu H(b) = \mu S(b) = \sigma H(b) = \sigma S(b) = 0$ for buckets b representing achromatic colors. This results in $S_{q_2,t_2}(b) = S_{q_3,t_3}(b) = S_{q_4,t_4}(b) = S_{q_5,t_5}(b) = 1$.

In addition, note that the mean values for the third axis of the HSI color space, the intensity axis, are not used for similarity calculation. With the exclusion of the mean intensity for each bucket, the similarity measure is intensity invariant, which enables generalization in matching. However, this advantage can, for example, become a disadvantage in a setting where solely color levels are compared.

Now that all values of a bucket are described, the total similarity for each color bucket b, i.e., a quantized color category, can be defined as:

$$S_{q,t}(b) = \\ S_{q_1,t_1}(b) \sum S_{q_2,t_2}(\mathrm{b}) \sum S_{q_3,t_3}(b) \sum S_{q_4,t_4}(b) \sum S_{q_5,t_5}(b) \sum S_{q_6,t_6}(b) \tag{4}$$

In addition to the statistical information, extra histogram information is used for determining the similarity. For each color bucket b of the query image q a weight-factor $W_q(b)$ is calculated. The weight is proportional to the number of pixels in a bucket. So, the most dominant color category of the query image, having the most pixels, has the largest weight. The reason to add such a weight is twofold. First, small buckets that represent a relative small number of pixels do not disturb the similarity calculation. Second, empty buckets do not enter into the similarity calculation, because their weight is zero.

$$W_q(b) = \frac{q_{1(b)}}{\sum_{i=0}^{B-1} q_1(i)} \tag{5}$$

where B is the total number of color buckets. Further, note that for a normalized histogram, as is the case in the present research, Eq. (5) can be rewritten as:

$$W_q(b) = q_1(b). \tag{6}$$

The total image similarity is then defined as:

$$S_{q,t} = \sum_{b=0}^{B-1} S_{q,t}(b) \cdot W_q(b). \tag{7}$$

A technical advantage of this similarity measure, which is incorporated in the 11 bin matching engine, is that it can be used or can be ignored when matching. The matching performance in a query-by-example setting will benefit from the additional information. For the query-by-memory paradigm the same engine can be used, but when preferred, the statistical information can be excluded.

Computational Complexity

Each statistical value can be regarded as a virtual bin. For all 11 bins the standard deviation of the intensity (σI) is determined. In addition, for the 8 chromatic colors the mean hue (μH), the mean saturation (μS), the standard deviation of the hue (σH), and the standard deviation of the saturation (σS) are determined. So, for the enhanced 11 bin configuration a total of 11 + 8 • 4 = 43 virtual bins are added. Hence, the computational complexity of the enhanced 11 bin configuration is equal to that of a 11 + 43 = 54 bin histogram.

The CBIR Benchmark

Introduction: The 3 CBIR Systems

The importance of benchmarking CBIR systems is illustrated by the foundation of "The Benchathlon Network",[26] which aims to "develop an automated benchmark allowing the fair evaluation of any CBIR system".

The CBIR systems used for the present benchmark contain two modules: the color matching engine and an interface module. The interface module is concerned with the presentation of the query and retrieval results (images) in HTML. It connects to the matching engine by calling *cgi-bin* scripts that generate the web pages and log the interaction with the user.

The color matching module is configurable with two parameters. The first describes the (pre-indexed) color histogram database to be used. The second parameter describes the histogram matching technique to be used. Matching was performed on the 60,000 images of the Corel image database.

For the first parameter, the color histogram database, two histogram configurations were used (11 and 4096 bins), each having their own quantization method. For the histogram configuration using 11 bins a quantization method was used based on the proposed segmented HSI color space. The second histogram configuration is the QBIC configuration using 4096 (16 × 16 × 16) bins[27,28] determined in RGB color space. This computationally heavy configuration is chosen because it performed best in the benchmark described in van den Broek et al.[24]

For the second parameter, two histogram matching functions were used in our benchmark: (i) the histogram intersection distance[25] (see Eq. (1)), and (ii) the proposed similarity function, which combines intersection similarity (see Eq. (2)) and statistical information (see above). We have used the intersection distance measure because it was judged as performing better than the quadratic distance for all histogram configurations.[24]

The proposed similarity function was only applied on the 11 bin configuration. So, in total three engines, i.e., combinations of color quantization schemes and distance measures, are compared in the benchmark: (i) the 4096 bin configuration, (ii) the 11 bin configuration, and (iii) the enhanced 11 bin configuration, using the similarity function.

Method

For each of the three engines, 30 query results had to be judged by human subjects, making a total of 90 per participant. They were unaware of the fact that three distinct engines were used to retrieve the images. Each set of 90 queries was fully randomized, to control for influence of order. Normally such retrieval results are presented in their ranked order. However, if this would have been the case in the benchmark, the participants would be biased to the first retrieval results after a few queries. Therefore, the ranking of the retrieved images is presented in random order.

Each query resulted in 15 retrieved images, presented in a 5 × 3 matrix. On the left side of this matrix the query image was shown as illustrated in Fig. 2.

The participants were asked to judge the retrieved images solely based on the color distribution hereby ignoring the spatial distribution of the colors. It was emphasized that semantics, shape, etc. should not influence their judgment. The participants were asked to perform two tasks.* On the one hand they were asked to mark the images that they judged as relevant. On the other hand, they were asked to indicate their overall satisfaction with the retrieved results on a scale from 1 to 10, see Fig. 2.

Fifty-one participants, both males and females in the age range 20-60, finished the benchmark; 11 did start with the benchmark but did not finish it. The data of this second group of participants was not taken into account for analysis.

Regrettably, six of the 51 participants did not complete the benchmark as instructed. Five of them did not select any of the retrieved images. One participant consistently selected one image for each of the 90 query results. Therefore, these six participants were not taken into account for the analysis either. Hence, in total we did collect usable data from 45 participants, making a total of 4050 queries that were judged.

We recorded for each query of each participant: the image ID, the query number, the number of bins used, whether or not the within bin statistics were used, the selected satisfaction rate, and which and how many images the participant judged as relevant.

Both the number of selected images and the rating for each query were normalized per participant. The normalized values were used for the analysis. How this normalization was done is defined in the next section.

Normalization of the Data

The participants' strategies for selection and rating of the retrieved images varied enormously. On behalf of the analysis, a normalization of the scores was applied for each participant separately.

Normalizing a range of scores takes the maximum and minimum score possible and the maximum and minimum score provided by the participant into account. Such a transformation is defined by:

$$S_n = a \cdot S_p + b, \tag{8}$$

where S_n is the normalized score, S_p is the score provided by the participant, and a and b are defined as:

$$a = \frac{\max - \min}{\max_p - \min_p} \qquad b = \max - a \cdot \max_p \tag{9}$$

with max and min being respectively the maximum, and minimum possible scores and $\max_p$ and $\min_p$ being the maximum and minimum scores provided by the participant. Note that where S_p is an integer, the normalized score S_n is a real number, since both a and b are real numbers. However, this is not a problem for further analysis of the data.

Results

Two dependent variables resulted from the experiments: the number of images selected by the participants as acceptable and the overall rating given by the partici-

*The complete instructions can be found at *http://www.few.vu.nl/~egon/CBIR-benchmark.html*.

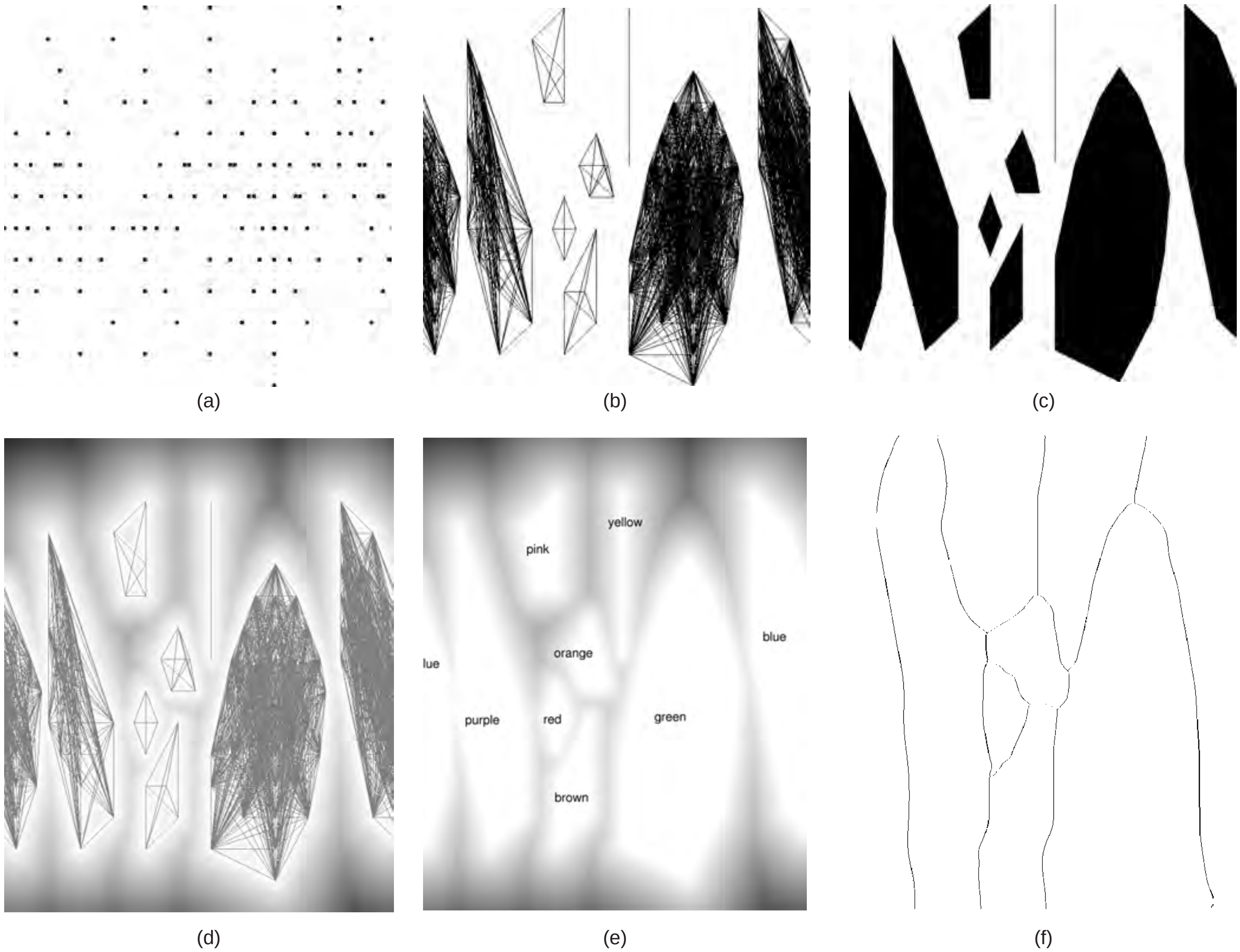


Figure 1. The processing scheme of the HSI color space segmentation using human color categorization data, gathered through two experiments.[17] The resulting RGB-coordinates were converted to HSI (hue-saturation-intensity) coordinates.[19] Next to the HI plane visualized here, a SI plane was used to segment the complete HSI color space. (a) The visualization of 8 of the 11 color categories present in the Color Look-Up Table (CLUT); (b) All CLUT coordinates belonging to the same color category were transformed by a line connector into fully connected graphs; (c) The graphs were converted to filled convex hulls; (d) The weighted distance map (WDM)[21] created using Fast Exact Euclidean Distance (FEED) transformations,[20] on the filled convex hulls, with the graphs visualized in it; (e) WDM labeled; and (f) The edges between the color categories, determined by a hill climbing algorithm and described by Fourier functions.

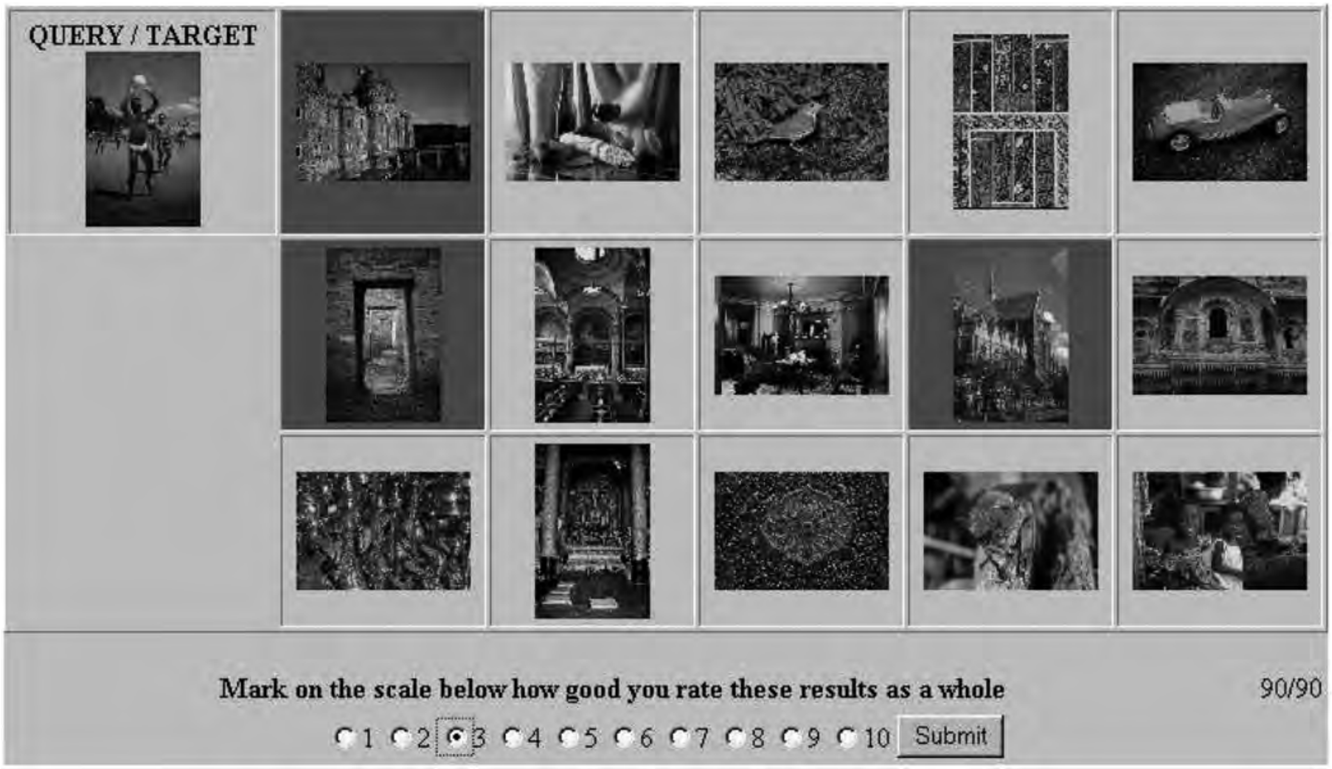
pants. Both measures were analyzed. The aim of the first phase of analysis was to determine whether a difference was present between the three engines. Therefore, for each measure a one-way ANOVA was applied. A strong and highly significant general difference was found between the engines for both the number of selected images (F(2,4047) = 60.29; p < .001) and for the overall rating provided by the participants (F(2,4047) = 97.60; p < .001).

A Duncan post hoc test on the number of selected images, revealed two homogeneous subsets within the group of three engines. According to the number of selected images, the 11 bin engines with and without within bin statistics did not differ (p < .01). Yet, this finding was not confirmed by the values on overall rating of the retrieval performance. A complete description of the statistical data can be found in Table I.

Based on the latter result we conducted an additional analysis. Six additional ANOVAs were applied: each of the three engines was compared with the two others for both measures. According to the overall rating the within bin statistics had improved the performance of the 11 bin quantization engine (F(1,2698) = 15.15; p < .001). In contrast, on the number of selected images a non-significant result (F(1,2698) = 3.00; p < .084) was found. The complete results of the six ANOVAs can be found in Table II.

Further, we were interested in the variability between participants. To determine a general effect of variability between participants, we applied a Multivariate ANOVA, which revealed, for both measures, a strong variability between participants (number of selected images: F(1,4046) = 10.23; p < .001 and rating: F(1,4046) = 6.61; p < .010).

Three Multivariate ANOVAs were done to determine for each of the three engines and for each of the two measures, how much participants differ in their scores. A complete overview of the variability between the participants for each of the three engines is provided in Table III.



Mark on the scale below how good you rate these results as a whole

Figure 2. The interface of a query as was presented to the participants. They were asked to select the best matching images and to rate their overall satisfaction, with respect to their color distribution only.

TABLE I. Descriptive statistics of the benchmark. Each engine is defined by its color quantization scheme (#bins) and whether or not statistical data on bin level (stats.) was taken into account. In addition, the number of queries (#queries) performed by each engine is mentioned. For the number of selected images (#images selected) as well as for the overall rating the mean (μ) value, the standard deviation (σ), and the confidence interval (min, max) at 99% is provided, for each engine.

			#images selected			rating		
#bins	stats.	#queries	μ	σ	(min, max), p = 99%	μ	σ	(min, max), p = 99%
11	no	1350	3.67	3.51	3.42-3.93	4.76	2.29	4.60–4.92
	yes	1350	3.91	3.48	3.65–4.17	5.10	2.28	4.94–5.26
4096	no	1350	5.13	4.06	4.87–5.39	5.96	2.36	5.80–6.12

TABLE II. Strength and significance of the difference found between the 11 bin, the enhanced 11 bin (including within bin statistical information: + stats.), and the 4096 bin engine, on both the number of selected images and the overall rating.

		strength and significance of difference	
engine 1	engine 2	#images selected	rating
11 bins	11 bins + stats.	F(1,2698) = 3.00 (p < .084)	F(1,2698) = 15.15 (p < .001)
11 bins	4096 bins	F(1,2698) = 99.02 (p < .001)	F(1,2698) = 181.23 (p < .001)
11 bins + stats.	4096 bins	F(1,2698) = 70.27 (p < .001)	F(1,2698) = 93.44 (p < .001)

TABLE III. Strength and significance of the variability between participants for the 11 bin, the enhanced 11 bin (including within bin statistical information: + stats.), and the 4096 bin engine, on both the number of selected images and the overall rating.

	strength and significance of variability between participants	
engine	#images selected	rating
11 bins	F(1,1348) = 7.00 (p < .008)	F(1,1348) = 3.31 (p < .069)
11 bins + stats.	F(1,1348) = 5.83 (p < .016)	F(1,1348) = 2.42 (p < .120)
4096 bins	F(1,1348) = 0.47 (p < .493)	F(1,1348) = 1.19 (p < .276)

Discussion

A large amount of data was collected through the benchmark, which is permanently available for participants at *http://www.few.vu.nl/~egon//CBIR-benchmark.html*. The results of this research comprise 4050 queries that were judged by 45 participants. Two measures were used: the number of selected images and the overall rating, indicating the satisfaction of the participant.

Without being told, the participants judged three distinct engines. Each engine can be defined by a combination of a color quantization measure and a distance measure. Two engines used the 11 bin quantization of color space and the third used the 4096 bin quantization of color space. The latter was judged as performing best in a previous pilot study.[24] The 4096 bin quantization and one of the 11 bin quantizations, which we call the standard 11 bin engine, employed the intersection distance measure. The other 11 bin quantization was equipped with a newly developed similarity function, based on within bin statistical information, which we therefore name the enhanced 11 bin engine.

Similar to the results presented in van den Broek et al.[24] the 4096 bin quantization scheme performed best, for both the number of selected images and for the overall rating. In addition it was found for the 11 bin quantization that the similarity function boosted the performance significantly, compared to the intersection distance measure. However, the latter result was only confirmed by the overall rating, not by the number of selected images; see also Tables I and II).

In the future, CBIR engines will probably be used by all computer users. Therefore, we were interested in whether the participants agreed in their judgment of the three engines.

Since the participants differ enormously in many ways, they might also be expected to differ in judging CBIR engines. If these differences are significant one would need interactive user-adaptable CBIR engines.[7,8] The latter would increase the complexity of CBIR system development enormously.

A strong variability between the participants was found for all three engines, with respect to the number of images they selected (see Table III). In contrast, the overall rating did not show a significant variability between the participants for any of the engines. So, a strong discrepancy was present between these two measures with respect to the variability between participants.

Moreover, the participants reported that judging whether a retrieved image should be considered as relevant is a particularly difficult process. This was mainly due to the fact that they were asked to judge the images based solely on their color distribution and to ignore their semantic content. Therefore, we have strong doubts concerning the reliability of the number of selected images as a dependent variable. The overall rating should be considered as the only reliable variable. For a benchmark such as ours, the number of selected images should, therefore, not be included in future research nor in further analysis of the current research.

Conclusion

We have explained our approach to Content-Based Image Retrieval (CBIR), which exploits human cognition instead of just image processing techniques. The importance of the 11 color categories (or focal colors) for CBIR was discussed. Using experimental data, the HSI color space was segmented, which resulted in a color quantization scheme for CBIR. Using this 11 bin color quantization scheme and the 4096 bin color quantization scheme of QBIC, combined with the intersection distance measure, two CBIR engines were developed. In addition, a third engine using the 11 bin quantization, combined with a new similarity function, was developed. The new similarity function incorporates within bin statistical information.

The 4096 bin engine performed best, according to the participants. However, in addition we found that the new similarity function boosted the performance of the 11 bin color quantization scheme significantly.

The advantage of the standard 11 bin approach in combination with the new similarity measure is its low computational complexity, where it outperforms the 4096 bin histogram by far. Taking in consideration that the latter is of extreme importance[29] in the field of CBIR, the results were very promising.

The work in this line of research continues. Other distance measures, the influence of the color space chosen for segmentation, the FEED algorithm, and differences between various types of images, e.g., photos, cartoons, and paintings, will be a topic of future research. In addition, we are working on texture analysis, shape extraction from image content, and the use of spatial information.

With these topics, and even in general, it is our belief that the combination of human cognition and statistical image processing will yield better performance for CBIR systems. But even more important, such a combination will enable us eventually to bridge the semantic gap present in CBIR.

Acknowledgment. The Dutch organization for scientific research (NWO) is gratefully acknowledged for funding Eidetic (project-number: 634.000.001), a project within the ToKeN research line, in which this research was done. In addition, we would like to thank Janna von Schmid for thoroughly reviewing this manuscript. Last, we would like to thank the volunteers who participated in our benchmark.

This work was performed while E. L. van den Broek was a PhD. student at the Nijmegen Institute for Cognition and Information, Radboud University, Nijmegen, and P. M. F. Kisters was an MSc. student at the Department of Artificial Intelligence, Radboud University, Nijmegen.

References

1. J. R. Janesick, *Scientific Charge-Coupled Devices*, SPIE, The International Society for Optical Engineering, Bellingham, WA, 2001.
2. B. H. Murray, *Sizing the internet*, Technical report, Cyveillance, Inc., Arlington, VA 2000.
3. NetCraft Ltd., *Web Server Survey Statistics*, http://news.netcraft.com/archives/web_server_survey.html, [Last accessed on April 17, 2005].
4. Google Inc., Google Blog Wednesday, November 10, 2004: Google's index nearly doubles, http://www.google.com/googleblog/2004/11/googles-index-nearly-doubles.html, [Last accessed on April 17, 2005].
5. T. Kanungo, C. H. Lee and R. Bradford, What fraction of images on the Web contain Text?, in *Proc. First International Workshop on Web Document Analysis (WDA2001)*, A. Antonacopoulos and J. Hu, Eds., World Scientific Publishing Co. Seattle, WA, USA, 2001.
6. A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, Content-based image retrieval at the end of the early years, *IEEE Trans. Pattern Analysis and Machine Intelligence* **22,** 1349–1380 (2000).
7. G. Frederix, G. Caenen and E. J. Pauwels, PARISS: Panoramic, Adaptive and Reconfigurable Interface for Similarity Search, in *Proc. ICIP 2000 International Conference on Image Processing*, R. K. Ward, Ed., vol. III, 2000, pp. 222–225.
8. G. Giacinto and F. Roli, Bayesian relevance feedback for content-based image retrieval, *Pattern Recognition* **37,** 1499–1508 (2004).
9. G. Derefeldt, T. Swartling, U. Berggrund, and P. Bodrogi, Cognitive color, *Color Res. Appl.* **29,** 7–19 (2004).
10. P. Kay, Color, *J. Linguistic Anthropol.* **1,** 29–32 (1999).
11. B. Saunders and J. van Brakel, in *Theories, Technologies, Instrumentalities of Color: Anthropological and Historical Perspectives*, B.

Saunders and J. van Brakel, Eds., University Press of America Inc., Lanham, MD, 2002.

12. D. Forsyth and J. Ponse, *Computer Vision: A modern approach*, Pearson Education, Inc., Upper Saddle River, NJ, 2002.
13. P. Kay, *The World Color Survey*, http://www.icsi.berkeley.edu/wcs/, [Last accessed on April 17, 2005].
14. B. Berlin and P. Kay, *Basic Color Terms: Their Universals and Evolution,* University of California Press, Berkeley, CA, 1969.
15. R. Goldstone, Effects of categorization on color perception, *Psychol. Sci.* **5,** 298–304 (1995).
16. E. R. Heider, Universals in color naming and memory, *J. Exp. Psychol.* **93,** 10–20 (1972).
17. E. L. van den Broek, M. A. Hendriks, M. J. H. Puts, and L. G. Vuurpijl, Modeling human color categorization: Color discrimination and color memory, *Proc. 15th Belgian-Netherlands Conference on Artificial Intelligence (BNAIC2003)*, T. Heskes, P. Lucas, L. Vuurpijl, and W. Wiegerinck, Eds., University of Nijmegen, Nijmegen, NL, 2003, pp. 59–68.
18. M. Stokes, M. Anderson, S. Chandrasekar, and R. Motta, *A Standard Default Color Space for the internet - sRGB*, Technical report, The World Wide Web Consortium (W3C), http://www.w3.org/Graphics/Color/sRGB.html [Last accessed on April 17, 2005].
19. T. Gevers and A. W. M. Smeulders, Color based object recognition, *Pattern Recognition* **32,** 453–464 (1999).
20. Th.E. Schouten and E. L. van den Broek, Fast Exact Euclidean Distance (FEED) Transformation, J. Kittler, M. Petrou, and M. Nixon, Eds., *Proc. 17th International Conference on Pattern Recognition*, 2004, vol. 3, p. 594-597. IEEE Computer Society Press, Los Alamitos, CA.
21. E. L. van den Broek, Th.E. Schouten and P. M. F. Kisters, Weighted distance mapping, (2005). Weighted distance mapping (WDM), *Proc. IEE International Conference on Visual Information Engineering: Convergence in Graphics and Vision* [in press]. IEE Publishing, Stevenage, UK.
22. *Munsell Color Science Laboratory*, Rochester Institute of Technology, http://www.cis.rit.edu/mcsl/, [Last accessed on April 17, 2005].
23. L. G. Vuurpijl, L. R. B. Schomaker and E. L. van den Broek, Vind(x): Using the user through cooperative annotation, *Proc. Eighth International Workshop on Frontiers in Handwriting Recognition,* S. N. Srihari and M. Cheriet, Eds., IEEE Computer Society Press, Los Alamitos, CA, 2002, pp. 221–226.
24. E. L. van den Broek, P. M. F. Kisters and L. G. Vuurpijl, The utilization of human color categorization for content-based image retrieval, *Proc. SPIE*, **5292,** 351–362 (2004).
25. M. J. Swain and D. H. Ballard, Color indexing, *Int'l. J. Computer Vision* **7,** 11–32 (1991).
26. S. Marchand-Maillet and G. Beretta, *The Benchathlon Network*, http://www.benchathlon.net/, [Last accessed on April 17, 2005].
27. W. Niblack, R. Barber, W. Equitz, M. Flickner, E. Glasman, D. Petkovic, P. Yanker, and C. Faloutos, The QBIC project: Querying images by content using color, texture, and shape, *Proc. SPIE* **1908,** 173–187 (1993).
28. M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele and P. Yanker, Query by Image and Video Content: The QBIC System, *IEEE Computer* **28,** 23–32 (1995).
29. F.-D. Jou, K.-C. Fan and Y.-L. Chang, Efficient matching of large-size histograms, *Pattern Rec. Lett.* **25,** 277–286 (2004).