The utilization of human color categorization for content-based image retrieval

Egon L. van den Broek[a], Peter M. F. Kisters[b], and Louis G. Vuurpijl[a]

[a]Nijmegen Institute for Cognition and Information,
Postbox 9104, 6500 HE Nijmegen, The Netherlands;
[b]Department of Artificial Intelligence, University of Nijmegen,
Montessorilaan 3, 6525 HR Nijmegen, The Netherlands.

ABSTRACT

We present the concept of intelligent Content-Based Image Retrieval (iCBIR), which incorporates knowledge concerning human cognition in system development. The present research focuses on the utilization of color categories (or focal colors) for CBIR purposes, in particularly considered to be useful for query-by-heart purposes. However, this research explores its potential use for query-by-example purposes. Their use was validated for the field of CBIR by two experiments (26 subjects; stimuli: 4 times the 216 W3C web-safe colors) and one question ("mention ten colors"). Based on the experimental results a Color LookUp Table (CLUT) was defined. This CLUT was used to segment the HSI color space into the 11 color categories. With that a new color quantization method was introduced making a 11 bin color histogram configuration possible. This was compared with three other histogram configurations of 64, 166, and 4096 bins. Combined with the intersection and the quadratic distance measure we defined seven color matching systems. An experimentally founded benchmark for CBIR systems was implemented (1680 queries were performed measuring relevance and satisfaction). The 11 bin histogram configuration did have an average performance. A promising result since it was a naive implementation and is still a topic of development.

Keywords: intelligent Content-Based Image Retrieval, CBIR, color categories, focal colors, color matching, human color perception

1. INTRODUCTION

With an average of 14.38 images on each web page[1] the World Wide Web consists to a great extent of images.[2] Unfortunately, most image retrieval engines are text-based and do not provide the means for searching on image content. Given the exploding market of digital photo and video camera's, the fast growing amount of image content further increases the need for image retrieval engines. This holds in particular for home applications, where the naive user requires support for browsing and searching image material.

Although "classic" text-based information retrieval (IR) methods are fast and reliable when images are well-named or annotated, they are incapable of searching in unannotated image collections. Content-Based Image Retrieval (CBIR)[3,4] methods are capable of searching in such collections. However, present CBIR methods do not perform well, nor fast enough to handle image databases beyond a closed domain and the techniques used are still subject to development.

As both retrieval methods can complement the weaknesses of the other, current research on image retrieval tries to combine IR and CBIR.[5] The system proposed here combines text-based and content-based techniques while exploiting knowledge from human color perception. It is therefore named intelligent CBIR (iCBIR). The control flow of our iCBIR system comprises four processes that are typical in image retrieval: (i) definition of the search query, (ii) matching of the query to an image database, (iii) presentation of the retrieval results and (iv) re-iterating, allowing the user to limit or expand the results. The development of the iCBIR system is part of Eidetic, a joint Dutch project within the NWO ToKeN2000* program. Eidetic is targeted on the integration of

Further author information: (Send correspondence to E. L. van den Broek)
E. L. van den Broek: E-mail: e.vandenbroek@nici.kun.nl, Telephone: +31 (0)24 3615476
P. M. F. Kisters: E-mail: citrofyl@hotmail.com, Telephone: +31 (0)24 3615476
L. G. Vuurpijl: l.vuurpijl@nici.kun.nl, Telephone: +31 (0)24 3615981
*URL: *http://www.ToKeN2000.nl*

Human Vision and Electronic Imaging IX, edited by Bernice E. Rogowitz, Thrasyvoulos N. Pappas, Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 5292 © 2004 SPIE and IS&T · 0277-786X/04/$15

five query paradigms: (i) Query-by-text, a query is specified by a textual description of the image; (ii) Query-by-example, an example (object present in the) image is used and its features (e.g. color, texture, and shape), are extracted to facilitate CBIR; (iii) Query-by-sketch, this query paradigm was based on the findings of Schomaker et al.,[6] who state that users can sketch the shape of objects present in the image and use this as a query; (iv) Query-by-color, the color (of an object in) the image, is defined and used for querying; (v) Query-by-texture, the texture (of an object) in the image is defined and used for querying. The query paradigms query-by-text and query-by-sketch are implemented in the current system Vind(x).[7]

In the proposed system, an initial query is defined using text, subsequently employing traditional IR techniques to yield a first set of images, or no images. If the set of retrieved images satisfies the information need of the user, the image query is terminated successfully. Else, if one or more of the retrieved images matches the image the user imagined, this can be used to re-iterate on the results by employing well-known query-by-example techniques. If no retrieved images are contained in the retrieval set, the user has to either rephrase his textual query, or fall back to another technique, called query-by-heart, which incorporates query-by-sketch, query-by-color, and query-by-texture. The latter technique requires the user to manually define the content of the image, based on the target image that merely exists in his mind.

Note that in all phase of iCBIR, the human has a central role. This is conform the advice of Rui et al.,[8] who already emphasized the importance of the user. In our opinion, this advice should be reinforced: the user should get a central position in the development of new CBIR engines. Especially, available information concerning the users' perceptive and cognitive abilities should be incorporated in an iCBIR system.

For the present paper we focus on query-by-color. We introduce a new color matching method, that takes human cognitive capabilities into account. This is of particular interest when considering the afore-mentioned query-by-heart and query-by-example techniques, as will be explained in Section 2. Theory on human color perception is presented that points in the direction that humans perceive color in so-called focal colors (or color categories). Furthermore, their importance in the distinction between query-by-example and query-by-heart is explained. Section 3 briefly describes our experiments that sustain the theory presented in Section 2. With the experiments we prove the difference between color categorization by color discrimination and by color memory. The experimental results are used for the segmentation of a color space, which is described in Section 4. The segmented color space is used as color histogram for histogram matching purposes. It is compared with other color quantization methods in Section 5. Section 6 describes how retrieval performance can be measured. In addition, the components of the benchmark are explained: four histogram configurations and two distance measures. Next, the experimental setup of the benchmark is described. The Section ends with the results of the benchmark. In section 7 advantages and disadvantages of the color quantization as proposed, are discussed and conclusions are drawn.

2. HUMANS DO PERCEIVE FOCAL COLORS

Human color perception is a complex function of context, for example: illumination, memory, object identity, culture, and emotion can all take part.[9–11] As already 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. Human competence at color constancy is surprisingly poorly understood. The main experiments on humans[13,14] do not explore all circumstances and it is not known, for example, how robust color constancy is or the extend to which high-level cues contribute to our color judgments." So, we are not even close to having a good model for human color perception.

2.1. Eleven colors!?

In our opinion, one should consider color in CBIR from another perspective, that of the focal colors or color categories (see also the World Color Survey[†]): Black, white, red, green, yellow, blue, brown, purple, pink, orange, and gray.[15–17] People use these categories when thinking, speaking, and remembering colors. Research from diverse fields of science emphasize the importance of them in human color perception. The use of this knowledge may provide a solution for the problems of color matching in CBIR. So, we have chosen to exploit the fact that

[†]URL: *http://www.icsi.berkeley.edu/wcs/*

all humans tend to think and perceive colors in 11 basic color categories. Summarizing, we can state that there are two main advantages in using the 11 basic color categories: (i) They are robust to variability between people (i.e., All people are different and so is their color perception.). (ii) They are robust to variability within people (i.e. People have changing moods and, for example, also changing perceptual abilities.).

2.2. Query-by-Example versus Query-by-Heart

Most CBIR-engines distinguish two forms of querying, in which the user uses either an example image (query-by-example) or defines features by heart, such as: shape, color, texture, and spatial characteristics (query-by-heart). In the latter case, we are especially interested in the use of the feature color. In the remaining part of this article we therefore define query-by-heart as query-by-heart utilizing color. At the foundation of both query-by-example and query-by-heart, lies a cognitive process, respectively color discrimination and color memory. Let us illustrate the importance of the distinction between query-by-example and query-by-heart by a simple example. Imagine a user wants to find images of brown horses.

Suppose the user possesses one such image and uses it to query-by-example. Images found 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 own image and with each other. This comparison we call the process of color discrimination. So, in this process the colors are (directly) compared to each other.

In the case of query-by-heart the user is required to retrieve the color brown from memory. Probably, this will not be one particular color, but rather a fuzzy notion of some set of colors: a color category, based on color memory. Each of the elements of this brown set (or category) are acceptable colors. There is no need for several types of brown. Providing the keyword "brown" or pressing a button resembling the fuzzy set brown is sufficient.

In both forms of querying the CBIR-system can use a Color Look-Up Table (CLUT) for the determination of the elements of this set, described by R, G, and B-values. The set is fuzzy due to the several influences on the color (of the object of interest), such as the color of the surrounding and the semantic context in which the object is present.

However, it is clear that a distinction should be made between color categorization by discrimination and color categorization by memory. An important distinction because humans are capable of discriminating millions of colors but when asked to categorize them by memory, they use a small set colors: focal colors or color categories.[15–17] Despite the fact that the importance of such a distinction is evident, this differentiation is not made in CBIR-systems.

We propose to use the 11 color categories for query-by-heart purposes in CBIR. For this purpose the front end of a CBIR engine was already extended with an eleven color pallet, as described in.[18] The 11 color matching engine perfectly fits this interface. However, we wanted to explore the use of the 11 color categories further and extend their use to query-by-example. But before this was done an endeavor was toward experimental evidence for the 11 color categories.

3. EXPERIMENTAL PROOF FOR THE 11 COLOR CATEGORIES

Until now, literature provided little clues for the theoretical and experimental evidence for the 11 color categories. For computer environments no evidence at all was present. Therefore, we have conducted an inquiry and two experiments.[19] The results prove that:

- The use of color categories is valid in a CBIR context,
- Color space can be described using color categories,
- There is a difference in color categorization using color discrimination or color memory.

Moreover, this research shows that it is possible for humans to quantize a color space into a limited set of clusters, representing a color look up table that can be used for CBIR. And as the process of color quantization is performed by humans, a new model of human color categorization is introduced.

3.1. The inquiry and experiments

Twenty-six subjects with normal or corrected-to-normal vision and no color deficiencies, participated to the experiments which were set up in an average office environment. The first task was to write down the 10 colors that arose from memory first.

The stimuli comprised the full set of the 216 web-safe colors. Each color stimulus was presented in the center of the screen, on an gray background. Below the stimulus, 11 buttons were placed. In the color memory experiment the buttons were labeled with the names of the 11 focal colors; in the color discrimination experiment each of the buttons did have one of the 11 focal colors. The 11 focal colors were presented conform the sRGB standard of the World Wide Web consortium (W3C)[‡].

Half of the participants started with the color discrimination experiment, the other half started with the color memory experiment. Each experiment consisted of 4 blocks of repetitions of all 256 stimuli (in a different order), preceded by a practice session. In addition, the 11 buttons were also randomized for block and for each participant. The practice session consisted of 10 stimuli. Block, stimulus, and button order was the same for both experiments.

3.2. Results and conclusions

The 10 colors written down by the participants confirmed the existence of the 11 focal colors (or color categories). All subjects named red, green, blue, and yellow. The other focal colors were also mentioned far more frequently than the non focal colors. Illustrative is that despite the fact that, with 11 occurrences, pink was the least mentioned focal color, it was mentioned almost twice as much than the most frequently mentioned non-focal color: violet (6).

The main result of both experiments is a Color LookUp Table (CLUT), distinguishing the discrimination and memory experiment[§]. No consistent color categorization was found over the experiments. This is due to the cognitive processes of discrimination and memory that influence color categorization strongly.

4. COLOR SPACE SEGMENTATION

Our aim was to bridge the gap between the large amount of colors in images and the 11 color categories. In order to do so, a quantization system that fits the 11 color categories was developed (see Figure 1 for its processing scheme). It segments the color space in the 11 color categories, using the CLUT resulting from the two experiments.

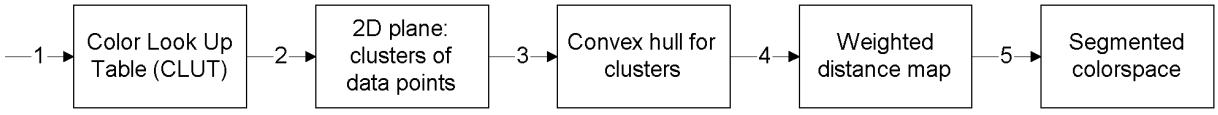


Figure 1. The phases of processing. (1) The Color LookUp Table is read. (2) Conversion of RGB data to HI-space and (3) Connection of the points belonging to the same cluster, resulting in a convex hull for each cluster. Next, (4) Weighted distance mapping[20] is done, using hexadecagonal region growing.[21] This results in a probability space that describes the complete data space. (5) The exact edges between the clusters / classes are extracted, resulting in a segmented color space.

First, the RGB coordinates of the categorized web-safe colors were converted to HSI (hue, saturation, and intensity) coordinates using conversion formulas defined in.[22] The HSI color space in particular was useful for color space segmentation since its hue-axis spreads out the RGB values along to its central diagonal. Hence, all color categories were separated, which made segmentation possible and a more intuitive color space was determined.

As the HSI model has three axis, 3D segmentation is required. To simplify computational complexity the saturation axis is ignored. This decision can be justified, as when the saturation values of the CLUT are left

[‡]http://www.w3.org/Graphics/Color/sRGB.html

[§]The CLUT can be found at: *http://eidetic.ai.kun.nl/egon/demos/vindx_colorselector/* .

out, the clusters of categorized colors can still be distinguished, ignoring the achromatic colors: black, gray, and white. The 8 remaining color categories are: red, green, blue, yellow, orange, purple,pink, and brown. For the remaining colors the segmentation of the HSI color space can thus be reduced to a 2D problem, being a xy-plane with isolated points representing each of the data points.

In,[20] a weighted distance mapping technique, based on hexadecagonal region growing,[21] is described that maps a limited set of data points (i.e., the 216 web-safe colors categorized by humans) onto a fully specified color space (see also Figure 1). As we restrict the space to 2 dimensions, 2D image processing techniques can be employed. For each category, the data points belonging to a cluster (color category) were fully connected by using a line generator. This also resulted in convex hulls for each of the clusters. Next, weighted distance mapping was applied on the data. This involves the region growing techniques described in.[20] The resulting process is visualized in Figure 2.

Using curve fitting techniques the extracted edges were converted to Fourier functions that express the borders between the color categories in segmented HSI color space.

4.1. Comparison of CLUT and segmented color space

So, the 2D color space was segmented by weighted distance mapping, using the experimental results. This was done using non-fuzzy colors. With non-fuzzy colors we mean the colors categorized by at least 10 of the 26 subjects to one category. The remaining colors are the fuzzy colors: the colors categorized to two categories by at least 10 subjects, for each category. So, the fuzzy and non-fuzzy colors together make up the CLUT, derived from the experiments.

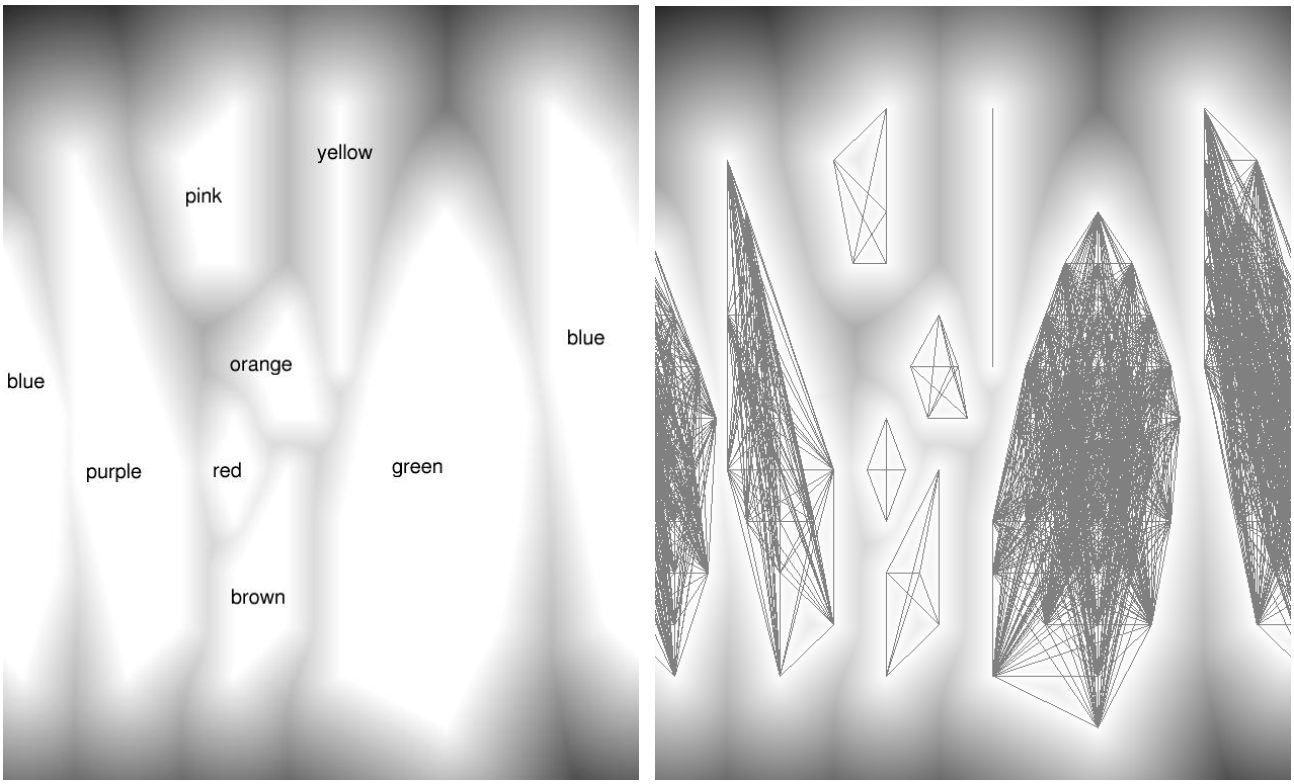
CLUT Category	**Segmented color space**							
name	Purple	Pink	Orange	Red	Brown	Yellow	Green	Blue
Purple	X	X	-	X	-	-	-	X
Pink	X	X	X	X	-	-	-	-
Orange	-	X	X	-	X	X	-	-
Red	X	-	X	X	X	-	-	-
Brown	-	-	X	X	X	-	-	-
Yellow	-	-	-	-	-	X	X	-
Green	-	-	-	-	-	X	X	X
Blue	X	-	-	-	-	-	-	X

Table 1. Colors and their neighbor colors in the segmented color space. The neighbor colors for each color category were found after analysis of the fuzziness of the experimental results.

The validation of the segmented color space consisted of two tests and the analysis of their results: (i) Categorization of the non-fuzzy colors and (ii) Categorization of the fuzzy colors. The segmented color space is valid iff it categorizes the stimuli used in the experiments to the same categories as the subjects did. So, first this was done for the non-fuzzy colors. For each of the 8 color categories a 100% match was found on the non-fuzzy colors, using the segmented color space. However, this result is not surprising since the segmented color space is based on the non-fuzzy colors.

In the second phase the fuzzy colors were categorized using the segmented color space. For all fuzzy colors held that they were placed into one of the color categories to which they were assigned, by the subjects in the experiments. However, due to the exact edges instead of fuzzy boundaries, the fuzzy character of the CLUTs vanished.

The consequences of the segmentation by exact edges of the color space on the original fuzzy boundaries of the color categories was analyzed. In all cases, the two or more color categories to which different subjects assigned a stimulus were neighboring color categories in the segmented color space. So, one could label these colors as being perceptual neighbors. In Table 1 the 8 color categories are listed with their neighbors. The segmented color space has a 100% match with the experimental results. All non-fuzzy colors are categorized correctly. All fuzzy colors are mapped in one of the categories to which they were assigned to in the experiments, where the other categories appeared to be a neighbor color category in the segmented color space.



(a) (b)

Figure 2. (a) and (b) Present the Hue (horizontal-axes) and Intensity (vertical-axes) dimensions of the HSI-color space as a probability space. The intensity of the pixels represent the confidence value that it belongs to its nearest segment. This weighted distance map categorizes the 2-dimensional color space completely. In (a) each segment is labeled with its color category name. The wire-models in (b) represent non-fuzzy CLUT color categories. These are clustered, using a line-generator. Their boundary is its convex hull.

4.2. The achromatic colors: Black, white and gray

The Hue axis is no linear axis but a value expressing the angle to a central rod. Gray values lie around this rod since the Saturation value of the central rod is 0. So, if the color space is considered to be a 2D plane without a Saturation axis, the Gray values are omni-present (throughout the hue/intensity plane). Thus a second step is required for the chromatic/achromatic separation.[23] Leaving out the hue axis of the HSI color space is a possible way to estimate a border between chromatic and achromatic colors. In other words, a Saturation-Intensity function will be computed expressing the boundary where the colors have lost so much of its color information that people tend to call it gray (white, or black respectively for high and low intensity levels). This seems a promising (2D) approach but it has proved to be suboptimal when categorization results were evaluated. From the experimental data using this approach one can conclude though that people tend to agree when a color turns gray. In other words, no significant difference in chromatic / achromatic separation is observed between fuzzy and non fuzzy color sets.

However, two problems arise when estimating the chromatic / achromatic separation: (i) The CLUT based on the W3C set of web-safe colors, does not provide enough achromatic values to make good estimations and moreover, (ii) the border between achromatic values and chromatic values changes along the hue axis. Because the boundary is fluctuant, using a 2D approach optimal segmentation based on CLUT values isn't possible. So segmentation has to be computed otherwise.

Current research explores a $2\frac{1}{2}$D approach to solve the problem of chromatic / achromatic separation. For each of the 8 (already segmented) color categories the chromatic / achromatic separation will be calculated in the saturation/intensity plane, dividing the 3D segmentation problem into 8 problems with 2D complexity. So along the hue axis, several chromatic / achromatic separations will be computed, representing the fluctuant separation.

Segmentation within the achromatic value set (to separate white, gray and black) suffers from the same problems mentioned above. In the current implementation however, we use approximations for gray, black and white values, dividing the intensity axes in three sections of equal length.

5. COLOR QUANTIZATION

At the core of many color matching algorithms lies a technique based on histogram matching. Most CBIR engines extend color histogram matching with spatial and statistical algorithms to boost color search. In order to use histogram matching, a quantization of the colors available in an image is applied to reduce computational complexity and provide a means of generalization. The latter depends on the number and proper choice of bins that are used and the color space that is employed. The color space determines which colors are close to each other, and thus influence color grouping during the quantization process.

Histogram matching on a large number of bins (weak quantization), has a big advantage: Regardless of the color space used during the quantization process, the histogram matching will have a high precision (see equation 1 in the next Section). Even quantization based on non-intuitive color spaces like RGB leads to a system with acceptable color matching performance when weak quantization is applied. Disadvantage of such a system (when it is naively implemented) are its high computational complexity and poor recall (see equation 1 because of poor generalization. Another disadvantages is the problematic mapping to a (easy to use) color definition interface in a query-by-heart context.

When strong color quantization is performed, these disadvantages can be solved. Nevertheless, to ensure acceptable precision, it is of decisive importance that human color perception is respected during quantization (such as using a color space that resembles human color perception). Hence, the combination of color space and selection of proper bins is crucial for the acceptance of the histogram matching technique by the user.

The results provided in the previous sections indicate that the small number of 11 bins and the chosen HSI color space may very well be used for color histogram matching. The strong point of our method is that it does not rely on an algorithm to yield the linear or non-linear mapping of colors into bins, but rather that it exploits human color perception for this task. Below, a number of color histogram matching techniques (distance measures) are described that will be used to assess this statement in a first set of experiments described in Section 6.

6. THE BENCHMARK

In order to assess the validity of our approach for image retrieval, we have made a first comparison study with the three color matching algorithms described above. The field of IR provides two metrics for estimating retrieval effectiveness: recall and precision. Recall signifies the relevant images in the database that are retrieved in response to the query. Precision is the proportion of the retrieved images that are relevant to the query.

$$recall = \frac{\#relevant\ retrieved}{\#relevant} \qquad precision = \frac{\#relevant\ retrieved}{\#retrieved} \tag{1}$$

The key issue is to determine which images are relevant. Merriam-Webster's dictionary[¶] defines relevance as " *the ability (as of an information retrieval system) to retrieve material that satisfies the needs of the user* ". So, relevance concerns the satisfaction of the user. The judgment of users is the only way to determine the recall and precision of the matching algorithms.

As we are unable to a priori approximate the number of relevant images for a given query, it is not possible to determine the recall of the systems. We can only examine their precision. The number of retrieved images

[¶]Online available at: *http://www.m-w.com*

follows from the retrieval and is fixed to 15 for this experiment. So, it is required to know the number of relevant retrieved images, for which the experiments described in this section are used.

6.1. The benchmark system

The CBIR system used for the present experiment contains 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 histogram configurations we have chosen for this purpose are presented in the next Section. The second parameter describes the histogram matching technique to be used. The techniques we have used are described in Section 6.3. Matching is performed to the 60,000 images of the Corel image database[‖].

6.2. Histogram configurations

Four histogram configurations were used (11, 64, 166, 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 configuration containing 64 is inspired by the PicHunter[24] image retrieval engine, which uses a HSV(4x4x4) quantization method. The quantization method used for the 166 bins is similar to the approach described in.[25] We call the configuration HSV(18x3x3)+4, meaning that the quantization was performed for 18 hues, 3 saturation, 3 values, and 4 achromatics (representing the central rod in the HSV color space). The last histogram configuration is the QBIC configuration using 4096 bins.[26, 27] The quantization method used for this configuration is RGB(16x16x16). This (computational heavy) configuration is picked to show the insignificance of the color space (used for quantization) when a large number of bins is used. Please note that the color histogram matching methods described in the previous Section have been implemented for this experiment and that no efforts have been made to exactly copy the optimized matching algorithms of the PicHunter, QBIC, and the system described by.[25]

6.3. Distance measures

Two histogram matching functions are used in our benchmark: the Histogram intersection distance and the quadratic distance. For other histogram matching functions we refer to works of Gonzales,[28] or Puzicha.[29] We have chosen for these two measures because: (i) the intersection distance is one the most used and widely excepted measures, (ii) the quadratic distance is reported as performing good,[26] and (iii) we had to limit the number of measure since our focus lies on quantization and a benchmark should be workable. Exhaustive testing of all distance measures was therefore not conducted.

Swain's[30] color-indexing algorithm identifies an object by comparing its colors to the colors of each of the potential target objects. This is done by matching the color histograms of the images via their histogram intersection. The intersection of the histogram of the query image h_q and the histogram of a potential target image h_t is defined as[30, 31]:

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

where q is the query image and t is the target image. m represents a bin and M is the total number of bins. This holds iff the histograms are normalized such that $|h_q| = |h_t|$.

The quadratic distance is defined as[31]:

$$D_q(q,t) = (h_q - h_t)^T \mathbf{A}(\mathbf{h_q} - \mathbf{h_t}), \tag{3}$$

where $\mathbf{A} = [\mathbf{a_i j}]$ and $a_i j$ denotes the similarity between elements with indexes i and j.

This distance measures is used in QBIC.[27] Since it is computationally very expensive in its naive implementation optimizations are proposed, such as.[32]

[‖]URL: *http://www.corel.com*

6.4. Design

For the benchmark two distance measures are chosen: the intersection distance (see equation 2) and the quadratic distance (see equation 3). We have used the four histograms, consisting of respectively 11, 64, 166, and 4096 bins (as described in 6.2. Each distance measure was applied on each number of bins, with one exception. The combination of 4096 bins with the quadratic distance measure was found computationally much too expensive to use.[32] So, in total seven systems that are compared in this benchmark.

For each system, 20 query results had to be judged by human subjects, making a total of 140 per subject. Each set of 140 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 done in the experiment the subjects 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 5x3 matrix. On the left side of this matrix the query image was shown. The layout (4:3) and the size of the images were chosen in such a way that the complete retrieval result was viewable and no scrolling was necessary (see Figure 3).

6.5. Subjects, instructions and data gathering

12 subjects, both men and women in the age of 20-60, participated in the benchmark, making a total of 1680 query-results (one of them did not finish the experiment). The subjects 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 judgment of the subjects was two-fold. 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 Figure 3).

We recorded for each query of each participant: the image ID, the query number, the distance measure used, the number of bins used, satisfaction rate, and images judged as relevant.

Both the number of selected images and the rating for each query were normalized per person. This was necessary since the range of the number of selected images as well as the rating of satisfaction varied strongly between subjects. The normalized values were used for the analysis.

6.6. Results

We have analyzed the data using six one-way ANOVAs. For both the number as relevant retrieved indicated (by way of selection) images and the overall rated satisfaction of the query-result, the systems were compared as well as their compounds: their histogram configuration (11, 64, 166, and 4096 bins) and their distance measure (quadratic distance (QD) and the intersection distance ID).

For both the relevant retrieved images as the satisfaction rate, strong differences were found between the seven systems (resp. $F(6,1673) = 10.39$, $p < .001$ and $F(6,1673) = 12.72$, $p < .001$). For both the relevant retrieved images and the rated satisfaction, the systems could be classified in three groups. For the as relevant indicated images the groups were: (i) QD-11, QD-64, QD-166, and ID-11, (ii) ID-64 and ID-166, and (iii) ID-166 and ID-4096 ($p < .05$). For the measured satisfaction the groups were: (i) QD-11, QD-64, QD-166, and ID-11, (ii) ID-64 and ID-166, and (iii) ID-4096 ($p < .05$). On both judgments the systems were ranked as (with their means between brackets: $QD-11$ [3.44] $\sim$ $QD-64$ [3.56] $\sim$ $QD-166$ [3.61] $\sim$ $ID-11$ [3.78] $<$ $ID-64$ [4.41] $\sim$ $ID-166$ [4.65] $\sim<$ $ID-4096$ [5.14].

A more detailed analysis, revealed a clear influence of both the distance measure ($F(1,1678) = 38.09$, $p < .001$) and the number of bins ($F(3,1676) = 12.75$, $p < .001$), on the number of relevant images. The same results were shown on the satisfaction for both the distance measure ($F(1,1678) = 45.74$, $p < .001$) and the number of bins ($F(3,1676) = 15.73$, $p < .001$).

In order to determine the variability between subjects in their judgments two additional one-way ANOVAs were done. Their outcome was that the subjects differ in a large extent on both their satisfaction ($F(11,1668) = 38.77$, $p < .001$) and the as relevant judged images ($F(11,1668) = 39.03$, $p < .001$). For satisfaction we identified five groups and for the number of relevant images we were able to identify seven groups of subjects ($p < .05$). Since 12 subjects participated, this illustrates the enormous inter personal differences in rated satisfaction and in the judgment of when an image is relevantly retrieved.

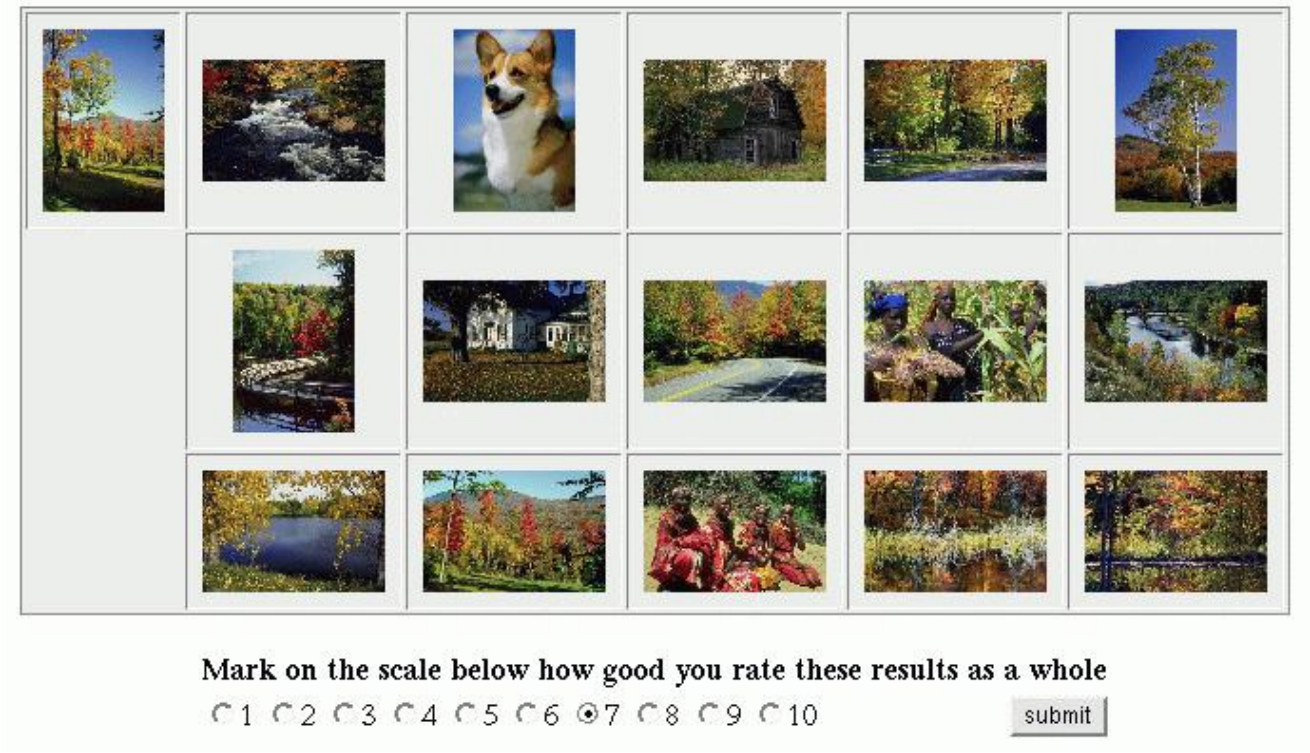


Figure 3. The interface of a query such as was presented to the subjects. They were asked to select the best matching images and to rate their satisfaction.

7. DISCUSSION

We have introduced the concept of intelligent Content-Based Image Retrieval (iCBIR), which is based on human cognition instead of on techniques. The use of cognition for color, one of the most prominent features used in CBIR, was discussed. The phenomenon of 11 color categories (or focal colors) was introduced. Its importance for color matching and query-by-heart was discussed and evaluated in a benchmark with other techniques.

The 11 color categories were proved to be valid by conducting an inquiry and two experiments. The experimental data provided us with a color lookup table. This was used to segment the HSI color space, using hexadecagonal region growing that resulted in a weighted distance map. From this map the 11 color categories were derived.

The 11 segments were used as the bins of a histogram in color matching. It was compared with histograms consisting of 64, 166, and 4096 bins. This was performed with two distance measures: the intersection distance and the quadratic distance. The seven resulting systems were tested in a benchmark.

The system that combined the 11 bin histogram with the intersection distance measure performed better than all systems using quadratic measures, but it performed not as good as systems using a stronger quantization of color space (i.e., used histograms with resp. 64, 166, 4096 bins) combined with the intersection distance measure. So, our naive implementation of the 11 bin concept should be boosted to be able to compare with systems using histograms with more bins.

A few explanations can be given for the lack of performance. Since, we have used an extreme weak quantization relatively much performance can be gained with incorporating statistical techniques, such as within-bin color distributions. However, if we would have used such techniques the comparison with the other systems would not have been fair anymore. In addition, we did not have a good segmentation of achromatic colors operational, which without any doubt negatively influenced the retrieval performance of the 11 bin histogram configuration.

However, an advantage of the 11 bin approach is its low computational complexity. On this topic the 11 bin concept outperforms the other histograms by far. Taking in consideration that the latter is of extreme importance[33] in the field of CBIR, the results were very promising.

It is our believe that the combination of human cognition and statistical image processing, will yield image retrieval systems that have a better chance to bridge the semantic gap. Although the 11 color category matching approach was initially developed for query-by-heart purposes, the benchmark proved that it works well for query-by-example purposes. With that this paper has introduced a general purpose, computationally inexpensive way of color quantization. In addition, a thoroughly designed experimental benchmark was presented for CBIR testing. With that two components were defined essential for the future development of CBIR techniques.

ACKNOWLEDGMENTS

The Dutch organization for scientific research (NWO) is gratefully acknowledged for funding Eidetic (project-number: 634.000.001), a project within the ToKeN2000 research line, in which this research was done. Furthermore, we would like to thank Frans Gremmen for advice on the methodological design and statistical analysis of the experiments and Thijs Kok for his assistance. Last, we would like to thank the volunteers who participated in one of our experiments.

REFERENCES

1. P. Lyman and H. Varian, "How much information." http://www.sims.berkeley.edu/research/ projects/how-much-info, 2000.
2. M. Lew, "Next generation web searches for visual content," *Computer* **33**(11), pp. 46–53, 2000.
3. Y. Rui, T. S. Huang, and S.-F. Chang, "Image retrieval: Past, present, and future," *Journal of Visual Communication and Image Representation* **10**, pp. 1–23, 1999.
4. 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 Transactions on Pattern Analysis and Machine Intelligence* **22**(12), pp. 1349–1380, 2000.
5. K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan, "Matching words and pictures," *Journal of Machine Learning Research* **3**, pp. 1107–1135, 2003.
6. L. R. B. Schomaker, L. G. Vuurpijl, and E. de Leau, "New use for the pen: outline-based image queries," in *Fifth International Conference on Document Analysis and Recognition*, pp. 293–296, IEEE, 1999.
7. L. G. Vuurpijl, L. R. B. Schomaker, and E. L. van den Broek, "Vind(x): Using the user through cooperative annotation," in *Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition*, S. N. Srihari and M. Cheriet, eds., pp. 221–226, (Ontario, Canada), 2002.
8. Y. Rui, T. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: A power tool for interactive content-based image retrieval," *IEEE Transactions on circuits and systems for video technology* **8**(5), pp. 644–655, 1998.
9. P. Kay, "Color," *Journal of Linguistic Anthropology* **1**, pp. 29–32, 1999.
10. B. Saunders and J. van Brakel, *Theories, technologies, instrumentalities of color: Anthropological and historical perspectives*, Lanham, Maryland: University Press of America Inc., 2002.
11. A. Schulz, *Interface design - Die visuelle gestaltung interaktiver computeranwendungen*, Rohrig Verlag, St. Ingbert, 1998.
12. D. Forsyth and J. Ponse, *Computer Vision: A modern approach*, Prentice Hall, 2002.
13. J. McCann, S. McKee, and Taylor, "Quantitative studies in retinex theory," *Vision Research* **16**, pp. 445–458, 1976.
14. L. Arend and A. Reeves, "Simultaneous colour constancy," *Journal of the Optical Society of America* **3**, pp. 1743–1751, 1986.
15. B. Berlin and P. Kay, *Basic color terms: Their universals and evolution*, Berkeley: University of California Press, 1969.
16. R. Goldstone, "Effects of categorization on color perception," *Psychological Science* **5**(6), pp. 298–304, 1995.

17. E. R. Heider, "Universals in color naming and memory," *Journal of Experimental Psychology* **93**(1), pp. 10–20, 1972.
18. E. L. van den Broek, L. G. Vuurpijl, P. Kisters, and J. C. M. von Schmid, "Content-based image retrieval: Color-selection exploited," in *Proceedings of the Dutch-Belgian Information Retrieval Workshop (DIR2002)*, M.-F. Moens, R. de Busser, D. Hiemstra, and W. Kraaij, eds., **3**, pp. 37–46, University of Leuven, Belgium, (Belgium, Leuven), December 2002.
19. 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," in *Proceedings of the 15th Belgian-Netherlands Conference on Artificial Intelligence (BNAIC2003)*, T. Heskes, P. Lucas, L. Vuurpijl, and W. Wiegerinck, eds., pp. 59–68, Nijmegen: SNN, University of Nijmegen, October 2003.
20. E. L. van den Broek, T. E. Schouten, and P. M. F. Kisters, "Weighted distance mapping," in *Proceedings of the International Conference on Computing Science 2004*, submitted, ed., 2004.
21. E. Coiras, J. Santamaria, and C. Miravet, "Hexadecagonal region growing," *Pattern Recognition Letters* **19**, pp. 1111–1117, 1998.
22. T. Gevers and A. W. M. Smeulders, "Color based object recognition," *Pattern Recognition* **32**, pp. 453–464, 1999.
23. N. Ikonomakis, K. N. Plataniotis, and A. N. Venetsanopoulos, "A region-based color image segmentation scheme," in *Proceedings of Visual Communications and Image Processing '99*, K. Aizawa, R. L. Stevenson, and Y.-Q. Zhang, eds., **3653**, pp. 1202–1209, 1999.
24. I. Cox, M. Miller, S. Omohundro, and P. Yianilos, "Pichunter: Bayesian relevance feedback for image retrieval," in *Proceedings of International Conference on Pattern Recognition*, pp. 361–369, Vienna, Austria, August 1996.
25. J. R. Smith and S.-F. Chang, "Single color extraction and image query," in *Proceedings of the 2nd IEEE International Conference on Image Processing*, B. Liu, ed., pp. 528–531, IEEE Signal Processing Society, IEEE Press, 1995. http://www.ee.princeton.edu/ icip95/.
26. 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," in *Proceedings of Storage and Retrieval for Image and Video Databases*, W. Niblack, ed., **1908**, pp. 173–187, February 1993.
27. 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**, pp. 23–32, september 1995.
28. R. C. Gonzales and R. E. Woods, *Digital image processing*, Prentice-Hall, Inc., New Jersey, 2nd ed., 2002.
29. J. Puzicha, Y. Rubner, C. Tomasi, and J. Buhmann, "Empirical evaluation of dissimilarity measures for color and texture," in *Proceedings the IEEE International Conference on Computer Vision*, **2**, pp. 1165–1173, (Corfu, Greece), september 1999.
30. M. J. Swain and D. H. Ballard, "Color indexing," *International Journal of Computer Vision* **7**(1), pp. 11–32, 1991.
31. J. R. Smith, *Integrated spatial and feature image systems: Retrieval, analysis, and compression*. PhD thesis, Columbia University, 1997.
32. J. L. Hafner, H. S. Sawhney, W. Equitz, M. Flickner, and W. Niblack, "Efficient color histogram indexing for quadratic form distance functions," in *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, **17**(7), pp. 729–736, 1995.
33. F.-D. Jou, K.-C. Fan, and Y.-L. Chang, "Efficient matching of large-size histograms," *Pattern Recognition Letters* **25**(3), pp. 277–286, 2004.