An Efficient Technique using Text & Content Base Image Mining Technique for Image Retrieval

Mahip M.Bartere[*], Dr.Prashant R.Deshmukh[**]

*(Department of Computer Science, Sipna College of Engineering & Technology, Amravati)
** (HOD of Computer Science , Sipna Colllege of Engineering & Technology, Amravati)

ABSTRACT

Image mining presents special characteristics due to the richness of the data that an image can show. Effective evaluation of the results of image mining by content requires that the user point of view is used on the performance parameters. Comparison among different mining by similarity systems is particularly challenging owing to the great variety of methods implemented to represent likeness and the dependence that the results present of the used image set. Other obstacle is the lag of parameters for comparing experimental performance. In this paper we propose an evaluation framework for comparing the influence of the distance function on image mining by color and also a way to mine an image from its name. Experiments with color similarity mining by quantization on color space and measures of likeness between a sample and the image results have been carried out to illustrate the proposed scheme. Important aspects of this type of mining are also described.

Keywords: **image mining, color space, color similarity. String Patterns.**

I.INTRODUCTION

Image are generated at increasing rate by sources such as military reconnaissance flights; defense and civilian satellites; fingerprinting devices and criminal investigation;scientific and biomedical imaging; geographic and weather information systems; stock photo databases for electronic publishing and news Agency; fabric and fashion design; art galleries and museum management; architectural and engineering design; and WWW search engines. Most of the existing image management systems are based on the verbal descriptions to enable their mining. A key-word description of the image content, created by some user on input, in addition to a pointer to the image data is the base of this system. Image mining is then based on standard mining. However, verbal descriptions are almost always inadequate, error prone and time consuming. The majority of pictorial information in real world images (as that in figure cannot be fully captured by text and numbers due to the limitation power of languages. A more efficient approach is gathered when image example is given by the user on input to the mining process. Automatically generate matching is required then for an efficient image mining. The basic idea is to extract characteristic features similar to that of object recognition schemes. After matching, images are ordered with respect to the query image according to their similarity measures and displayed for viewing. in this work, we present a framework for considering the influence of this distance function on color mining. This framework assesses a system's quality from the viewpoints of users; it provides a basic set of attributes to characterize the ultimate utility of systems. Then we analyze examples of mining by color and present some conclusions.

II. MINING BY COLOR

Mining in visual database is quite different from standard alphanumeric mining. On current approaches, feature vectors per image are computed for evaluation distance function on the feature space. Then this function is used to retrieve images from a given set. Images with distance less than a predefined threshold or within a predefined number are retrieved. These feature vectors facilitate mining by color, texture, geometric properties, shape, volume, spatial constraints,etc.

Experimental results show that image mining based on color provides high discrimination power. Querying by color similarity has been proposed in several systems. Although, these search engines support querying based on color, each system has special characteristics and limitations. For example, the color space, the used color quantization algorithms, the distance functions and indexing methods are different. When one search for images that contain colors similar to an example, matching is usually performed by evaluating distance in the used color space. The implementation usually return a fixed-size set of nearest neighbor without regard to actual threshold of similarity. In practice, determination of an appropriate threshold of similarity is difficult; frequently it involves multiple characteristics and arbitrary weightings. Whether it really performs a good work on mining similar colors is complicated by the fact that the human perception of colors is mainly psychological and does not have suitable mathematical definition.

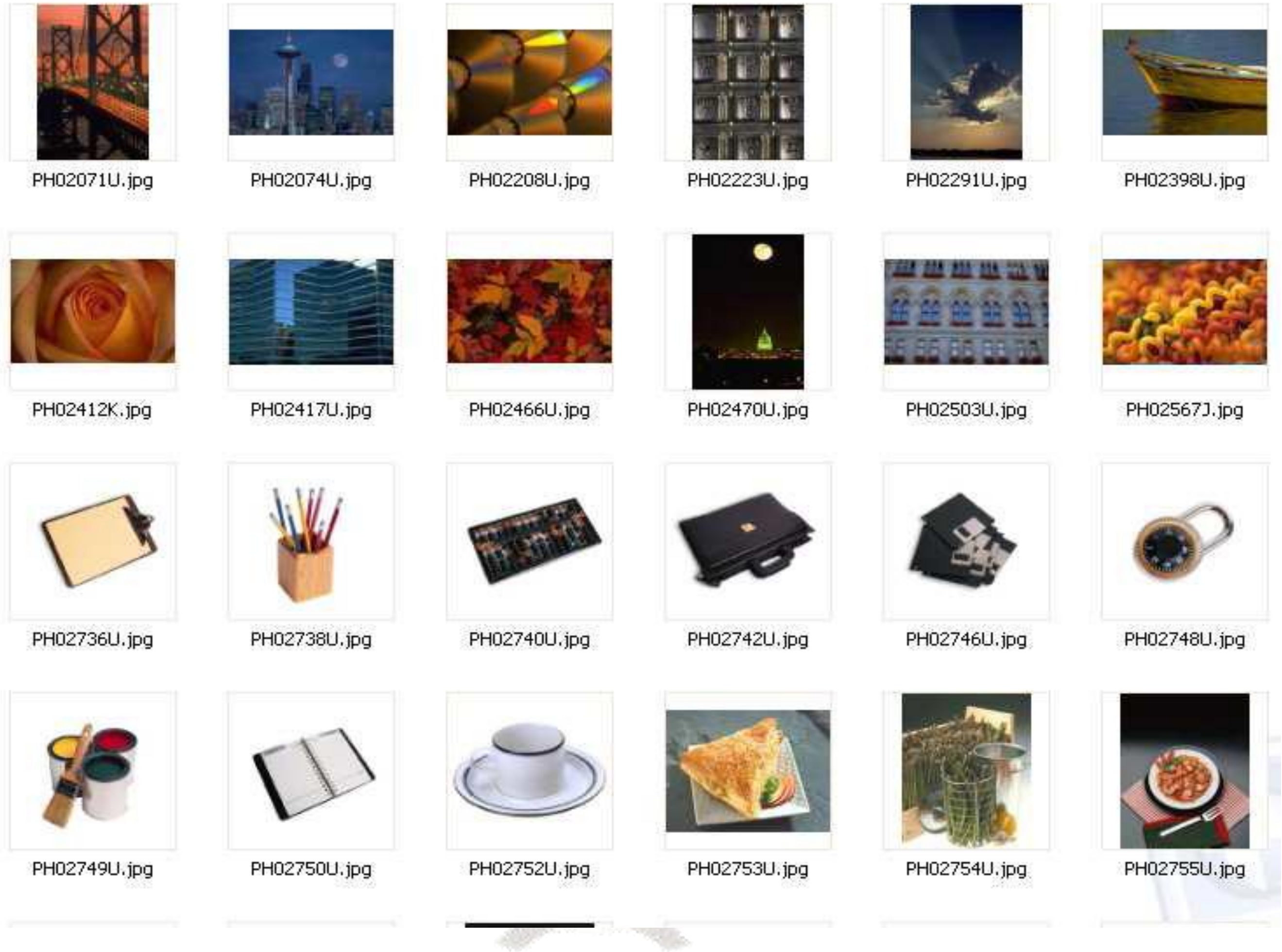


Figure 1:Example of Dataset Images

Well-known distance measures do not exactly matches what a user fells as a similar color. They work in ad-hoc manner, but no one pays much attention to their real efficiency. The lack of effective evaluation parameters or benchmarks for retrieval systems are identifies as a critical issue. Without a common technique each system uses individual evaluation procedures and image-match scores are not consistently compared among the various systems. Moreover, in visual information systems, this must be defined in terms of simple human perception aspect to preserve its real objective of efficiency. User satisfaction is the most important consideration for evaluating software's effectiveness.

III. PROPOSED METHOD

• Retrieval by Image

Complexity is a useful point in comparing software systems .This aspect is normally obtained from the source code, but it is completely irrelevant for the user. The mining result is a more important aspect for the user. The factor concerning to the mining result are: the underlying color space used to represent the color features; the quantization approach used; the number of bins on the histogram space (its dimension or digital color resolution); the distance function used to represent the notion of nearness on the color space (histogram representation); the fixed number of images to be retrieved and the threshold used for matching similarity Several color spaces have been used for color representation based on the perceptual concepts.

There is no agreement on which is the best choice. Anyway, its desirable characteristics are completeness, uniformity, compactness, and user oriented. Completeness means that it must include all perceptible different colors. Uniformity means that the measured proximity among the colors must be directly related to the psychological similarity among them. Compactness means that each color presents a perceptual difference from the other colors. Color quantization transforms a continuous tone picture into a discrete image. The digitalization process maps each component of a continuous color signal into a series of limited number of (fewer) colors.

This process inevitably introduces distortion. The visible distortion is a subjective and psychological notion. The questions are how to choose the colors to reproduce the original (not necessarily colors that appear in the original image). A quantization algorithm should distribute any visible distortion throughout the image so that none stands out to be found particularly objectionable by an average human observer. Empirical algorithms (as the popularity algorithms and the median-cut algorithms) present cases where significant color shifts can be found. One of the numerical criteria for color image quantization is to minimize the maximum variance between original pixel color and the corresponding quantified color, which provides better results than empirical algorithms. Another numerical criterion is to minimize the maximum discrepancy between original and quantified pixel values. Recent works use adaptive quantifiers. The basic strategy employed by these is a two-step approach.

The first step group original colors into clusters that are as small as possible. The second step computes a quantified color for each cluster. This means that each image is associated with two types of histograms in the mining process. The used color space is the HSV, where H (hue) is the attribute associated with the dominant wavelength. The HSV model is based on psychophysical data. For images already expressed in the RGB space the transformation into the hexagonal cone of HSV is performed by the well known transformations.

The H axis is more sensitive to color variation than S (saturation) and V (color intensity or value). S and V are more sensitive to lighting variation from shadows and distance from the light source. Thus, the H axis was used to be sampled more than the other two. S and V were divided into 3 sections each. The hue values range from 0o to 360o. Channel H was quantified in two forms: the first into 18 sections of 20o each, and the second into 24 sections of 15o each. Five distance functions are used.

They are "city-block" metric, Euclidean metric, histogram intersection, average color distance, and the quadratic distance form. Denoting *he* the histogram of the example image and *hp* the histogram of each image to be compared, then the "city block" metric or (d1), and the Euclidean metrics or (d2) are given by:

$$d^{\tau}_{e,p} = \left[\sum_{m=0}^{M-1} \left| h_e[m] - h_p[m] \right|^{\tau} \right]^{\frac{1}{\tau}}$$

Where if τ = 1, it represents "city-block" metric and if τ = 2, it corresponds to Euclidian metric. If the images has the same number of pixels, {*he*} = {*hp*} Where

$$\{h_e\} = \sum_{m=0}^{M-1} h[m]$$

Then the distance function based on histogram intersection Or (d3) is given by:

$$d_{e,p} = 1 - \frac{\sum_{m=0}^{M-1} \min(h_e[m], h_p[m])}{\{h_e\}}$$

The average colors distance or (d4) uses the average magnitude along the three channel of the space color .The Euclidean distance between their average colors defines the distance between two images. The quadratic distance measure form or (d5) use the expression: Where *A* is a matrix of similarity weights, *A= [aij]*, 0≤ *aij* ≤1 and *aii* =1 . Each entry is given by *aij = (1- dij / dmax)*, where *dij* is the Euclidean distance between colors *i* and *j*, and *dmax* is the greater distance between colors on the normalized HSV space. That is, the coefficient *aij* for two colors:
*m*0 *he se ve* =(, ,) and *m h s v* 1 *p p p* =(, ,) That determines each element of *A= [aij]*, is given by [3]:

$$a_{m_e m_p} = 1 - \frac{\left[(v_e - v_p)^2 + (s_e \cos(h_e) - s_p \cos(h_p))^2 + (s_e sen(h_e) - s_p \right.}{}$$

The possibilities of find all the relevant content of database are an important aspect for interpreting the queries results and also for classifying the quality of each metric. The possibility of "no-show" an image characterizes false negative results, i.e. not all images on the set with similar color composition can be retrieval by the environment, because it does not take color similarity into account adequately.

On this case some significant image can never be mining and the user concludes that such image does not exist. False negative can be related to deficient consideration of color similarity by the metric. The parameter named Retrieval Robustness (RE) was built [24] to show the ability of mining all images on the set that are of the same type of a given sample. On a query, each time a correct image (with color in the same group of the image query) appears it is considered a significant answer. The maximum number of significant images, *Ns*, that can appear depends on the number of images of the group on consideration, *Ng*, and the number that the user request, *n*. A measure of the completeness of the inquiry is then defined by:

$$RE = Ns / (Ng \times n) \%$$

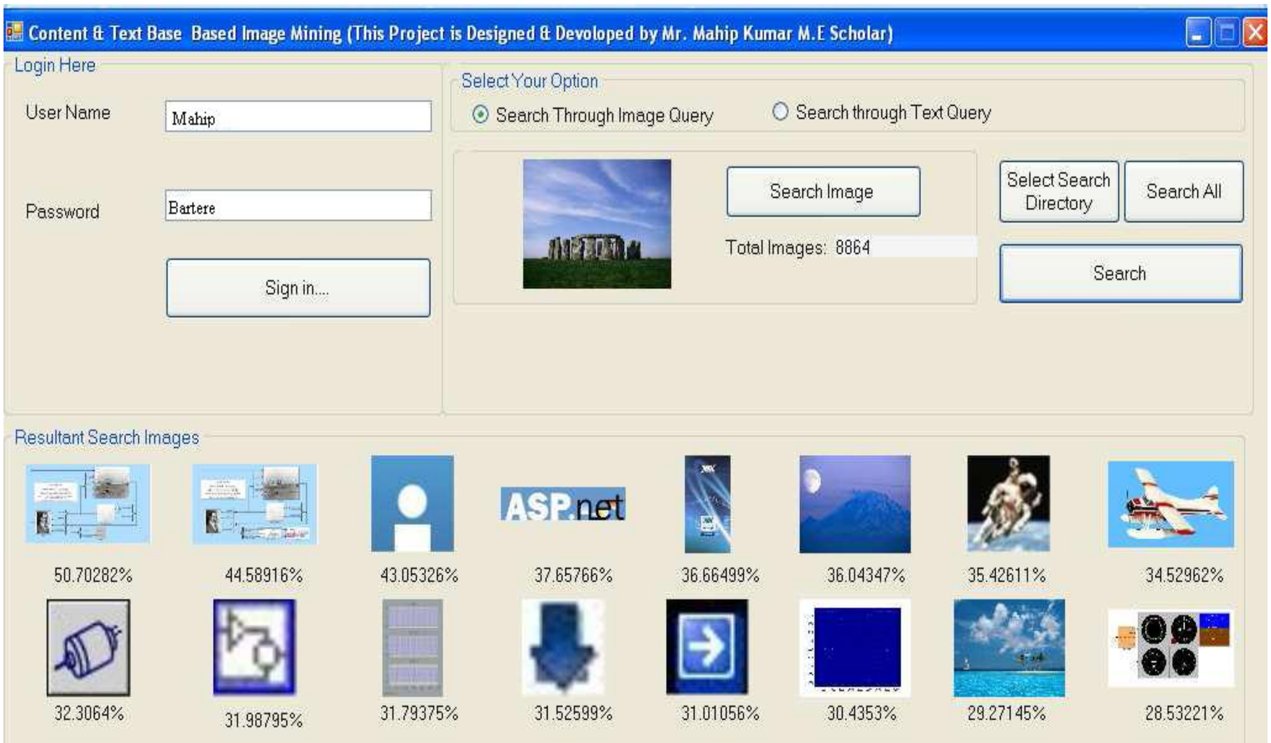
On RE evaluation, the significant mining results are considered over the first *n* request number of results [8,24-28]. It is presented on percentage for easy comparison between each combination of possibility.

- **Image Retrieval By Text**

Text retrieval is a subfield of information retrieval, which is the art and science of searching for information in documents or for documents themselves. Text retrieval is focused on image name. Commonly, the user places a query in form of some keywords and the text retrieval engine returns the image that matches his query best. Well-known examples of text retrieval engines are search engines on the World Wide Web. Text retrieval is an active field of research and there are many different approaches to this problem.

IV. EXPERIMENTAL RESULT

In this section we analyze the performance of the two different colors quantization and the five distance functions. Each combination is evaluated with respect to the percentage of Retrieval Robustness (RE) using the same image set and the similar color groups. Below figures and the others considering all image color groups show that the results are related with the images color composition more than with the number of bins on the histogram. They depend also of the number of images the user wants to find. On average, better results concerning retrieval robustness (RE) were obtained using Euclidian metric, which is also an easy computed value. Great difference of metric performance can be seen if few numbers of images are asked. On increasing the number of wanted images almost all metric presents quite the same performance



Content & Text Base Based Image Mining (This Project is Designed & Devoloped by Mr. Mahip Kumar M.E Scholar)

Login Here

User Name: Mahip

Password: Bartere

Select Your Option

● Search Through Image Query ○ Search through Text Query

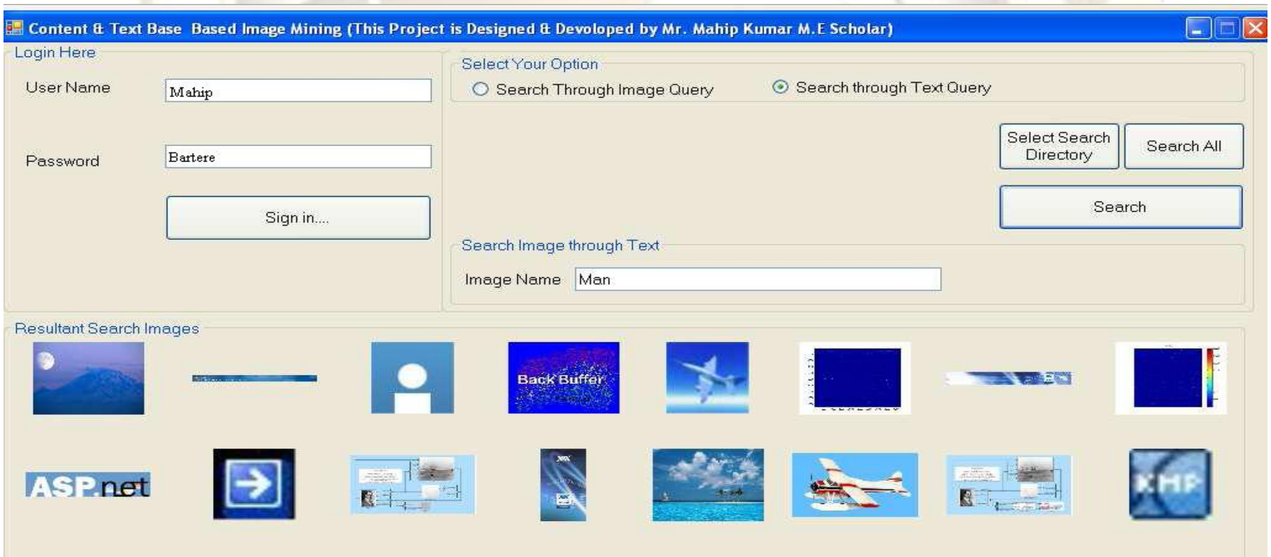
Total Images: 8864

Resultant Search Images

50.70282% 44.58916% 43.05326% 37.65766% 36.66499% 36.04347% 35.42611% 34.52962%

32.3064% 31.98795% 31.79375% 31.52599% 31.01056% 30.4353% 29.27145% 28.53221%

Figure 2: Search by an Image



Content & Text Base Based Image Mining (This Project is Designed & Devoloped by Mr. Mahip Kumar M.E Scholar)

Login Here

User Name: Mahip

Password: Bartere

Select Your Option

○ Search Through Image Query ● Search through Text Query

Search Image through Text

Image Name: Man

Resultant Search Images

Figure 2: Search by Text

REFERENCES

[1] V. N. Gudivada, V. V. Raghavan, "Content-Based Image Retrieval Systems", IEEE Computer,September, 18-22, 1995.

[2] S. A. Stricker, "Bounds for discrimination power of color indexing techniques", Proc. SPIE, pp. 15-24, 1994.

[3] J. Hafner, H. S. Sawhney, W. Equitz, M. Flickner and W. Niblack, “Efficient Color Histogram Indexing for Quadratic Form Distance Functions”, IEEE Trans. Pattern Analysis Machine Intell., Vol. 17, No.7, pp. 729-736, 1995.

[4] A. Pentland, R. W. Picard, S. Sclaroff, “Photobook: content-based manipulation of databases”, Int. J. Computer Vision, Vol. 18 , No. 3, 233-254, 1996. http://wwwwhite. media.mit.edu/~tpmink/photobook

[5] M. J. Swain, D. H. Ballard, “Color Indexing”, Int. J.Comp.Vision, Vol. 7 , No. 1, 11-32, 1991.

[6] M. Flickner, H. Sawhner, W. Niblack, J. Ashley, Q.Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petrovic, D. Steele, P. Yanker, “Query by image video content: the QBIC system”, IEEE Computer, September, 23-32,1995.http://wwwqbic.almaden.ibm.com/~qbi

[7] P. K. Kaiser, R. M. Boyton, Human Color Vision,Second Ed., Washington, D.C.: Optical Society of America, 1996.

[8] J. R. Bach, C. Fuller, A Gupta, A Hampapur, B. Horowits, R. Humphrey, R. C. Jain, C. Shu, “Virage image search engine: an open framework for image management”, Symposium on Electronic Imaging: science and technology storage & retrieval for image and video databases IV,IS&T/SPIE, 76-87, 1996. - http://www.virage.com

[9] B. Hill, Th. Roger, F. W. Vorhagen, “Comparative analysis of the quantization of color spaces on the basis of the CIELAB color-difference formula”,ACM Transaction on Graphics, Vol. 16, No. 2, April, 109-154, 1997.

[10] H. Zhang, Y. Gong , C. Y. Low, S. W. Smoliar, “Image retrieval based on color features: an evaluation study:, Proc. of SPIE, 2606, pp. 176-187,1997.

[11] C. Z. Ren, R. W. Means, “Context Vector Approach To Image Retrieval”, Proc. IEEE ICIP, Vol. 1, No 407, 1997.