Performance comparison of distance metrics in content-based Image retrieval applications

A Vadivel
Dept. of Computer Science and Engineering, Indian Institute of Technology, Kharagpur, India
vadi@cc.iitkgp.ernet.in

A K Majumdar
Dept. of Computer Science and Engineering, Indian Institute of Technology, Kharagpur, India
akmj@cse.iitkgp.ernet.in

Shamik Sural
School of Information Technology, Indian Institute of Technology, Kharagpur, India
shamik@cse.iitkgp.ernet.in

***Abstract* - Color is always considered to be an important attribute in content-based retrieval of images. One of the standard ways of extracting a signature from an image is to generate a color histogram. During retrieval, the histogram of a query image is compared with the histogram of each of the database images using a standard distance metric. The retrieved result is dependent both on the color histogram and the distance metric. We have done a detailed study of the performance of different distance metrics for a number of color histograms on a large database of images. We use Manhattan distance, Euclidean distance, Vector Cosine Angle distance and Histogram Intersection distance for performance comparison. Five standard and well-known color histograms were considered for evaluation and the results show that the Manhattan distance performs better than the other distance metrics for all the five types of histograms.**

Keywords: Color Histogram, Euclidean Distance, Manhattan Distance, Vector Cosine Angle Distance, Histogram Intersection, Precision.

I. INTRODUCTION

Content based image retrieval (CBIR) is a lively discipline, expanding in both depth and breadth and an important research topic covering a large number of research domains like image processing, computer vision, very large databases and human computer interaction [6, 10]. Digital image databases have seen an enormous growth over the last few years. Since many image collections are poorly indexed or annotated, there is a great need for developing automated, content-based methods that could help users to retrieve images from these databases. There are usually three kinds of search strategies. One of them is, search by association in which the user has no specific aim other than finding interesting things. Another requirement is to search for a specific image. The user has a specific image in mind and the target is interactively specified by refining similarity to a group of given examples. This is useful for art, industrial components or catalogue searches. The third type of searching is category search. Here the user aims to retrieve an image from a specific class (catalogues of varieties, for example).

Distance metric or matching criteria is the main tool for retrieving similar images from large image databases for all the above categories of search. Several distance metrics, such as the L1 metric (Manhattan Distance), the L2 metric (Euclidean Distance) and the Vector Cosine Angle Distance (VCAD) have been proposed in the literature for measuring similarity between feature vectors [1]. In content-based image retrieval systems, Manhattan distance and Euclidean distance are typically used to determine similarities between a pair of images. In document retrieval systems, on the other hand, distance metrics based on cosine angle are more commonly used for determining similarity between two documents. Jain and Vailaya have used two distance metrics, namely, Euclidean distance and Histogram Intersection for an image retrieval system using color histogram and shape [2]. Even though the Euclidean distance and the cosine angle based distances coincide when the components of the feature vectors are normalized by the norm of the vector, they differ when they are normalized otherwise. In image processing applications, components of a feature vector (e.g., color histogram) are usually normalized by the size of the image and as a result, the Manhattan, Euclidean, the cosine angle based distance and Histogram Intersection distance metrics produce different ordering of retrieved images.

The main contribution of this paper is to demonstrate the performance of these distance metrics for ordering of similar images in an image retrieval system. To the best of our knowledge, such a performance comparison has not been done on large image databases for the standard color histograms. We present experimental results on a database of 28,168 images to compare these distance metrics applied on five different color histograms. The histograms considered in this work include a standard RGB histogram RGB [9], Jain and Vailaya's histogram implementation (JV) [2], an HSV-based histogram (HSV) [8], an HSV-based histogram with soft decision (HSVSD) [7] and the histogram used in the QBIC system (QBIC) [3]. We have developed a content-based image retrieval application as part of our research. Queries can be made on an image database for the retrieval of similar images. The current capability of the system is handling of color histograms with a number of distance metrics. However, other features like shape and texture are being included for multi-feature retrieval of images. The system is accessible on the web for interested readers at http://www.imagedb.iitkgp.ernet.in. The results of the current paper are based on the experiments done on this system.

In the next section of the paper, we describe the different distance metrics used for comparison. In section 3, we explain the details of the web-based application. Section 4 contains the experimental results and we draw conclusions from our work in the last section of the paper.

II. DISTANCE METRICS

In the domain of image retrieval from large databases using signatures like color histogram, each 'n' dimensional feature vector may be considered as a point in the 'n' dimensional vector space. Thus, a feature vector $\mathbf{F} = \{f_1, f_2, \ldots, f_n\}$ is mapped to a point $P(f_1, f_2, \ldots, f_n)$ in the n-dimensions. This mapping helps us to perceive the images (represented by their feature vectors) as high-dimensional points. The advantage of this representation is that one can now use different distance metrics for (i) finding similarity between two images and (ii) ordering a set of images based on their distances from a given image. This enables us to do a nearest neighbor search on a large database of images and retrieve a result set containing images that are closest matches to a user-specified query. It is evident that the images and their ordering depend both on the feature extraction method as well as on the distance metric used. In this work we consider Manhattan distance, Euclidean distance, Vector Cosine Angle distance as well as Jain and Vailaya's implementation of Histogram Intersection and Euclidean distance metrics and their properties.

Manhattan Distance

It is also called the L_1 distance. If $u = (x_1, y_1)$ and $v = (x_2, y_2)$ are two points, then the Manhattan Distance between u and v is given by

$$MH(u, v) = |x_1 - x_2| + |y_1 - y_2| \quad (1)$$

Instead of two dimensions, if the points have n- dimensions, such as $a = (x_1, x_2, \ldots\ldots, x_n)$ and $b = (y_1, y_2, \ldots\ldots, y_n)$ then, eq. 1 can be generalized by defining the Manhattan distance between a and b as

$$MH(a,b) = |x_1 - y_1| + |x_2 - y_2| + \ldots + |x_n - y_n|$$

$$= \sum_{i=1}^{n} |x_i - y_i| \quad (2)$$

Euclidean Distance

It is also called the L_2 distance. If $u = (x_1, y_1)$ and $v = (x_2, y_2)$ are two points, then the Euclidean distance between u and v is given by

$$EU(u,v) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (3)$$

Instead of two dimensions, if the points have n- dimensions, such as $a = (x_1, x_2, \ldots\ldots, x_n)$ and $b = (y_1, y_2, \ldots\ldots, y_n)$ then, eq. 3 can be generalized by defining the Euclidean distance between a and b as

$$EU(a,b) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots\ldots + (x_n - y_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (4)$$

Vector Cosine Angle Distance

If we consider two vectors $\mathbf{X}$ and $\mathbf{Y}$ where $\mathbf{X} \equiv (x_1, x_2, \ldots, x_n)$ and $\mathbf{Y} \equiv (y_1, y_2, \ldots, y_n)$, then $\cos\theta$ may be considered as the Cosine of the vector angle between $\mathbf{X}$ and $\mathbf{Y}$ in n dimension. Formally, we define VCAD as follows.

$$VCAD(\mathbf{X}, \mathbf{Y}) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2} \sqrt{\sum_i y_i^2}} = \frac{\mathbf{X}.\mathbf{Y}}{\|\mathbf{X}\|\|\mathbf{Y}\|} \quad (5)$$

In Fig. 1, we show the Vector cosine angle distance between two 2-dimensional vectors (x_1, x_2) and (y_1, y_2). One important property of vector cosine angle is that it gives a metric of similarity between two vectors unlike Manhattan distance and Euclidean distance, both of which give metrics of dissimilarities. Also $VCAD(\mathbf{X}, \mathbf{Y}) \in [0,1]$. This makes it easy to combine distance between two images using multiple features.

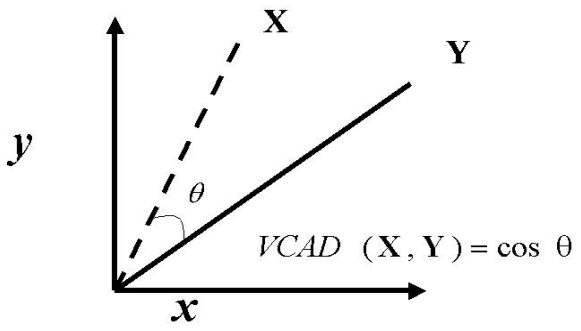


Figure 1. Vector Cosine Angle Distance Definition.

Histogram Intersection Distance proposed by Jain and Vailaya

Let I_R, I_G and I_B be the normalized color histograms of an image in a database and let Q_R, Q_G and Q_B be the normalized color histograms of a query image. Here, R, G and B denote the three color channels in the RGB color space. The similarity between the query image and a stored image in the database, $S_c^{HI}(I, Q)$, is given by

$$S_c^{HI}(I,Q) = \frac{\sum_r \min(I_R(r), Q_R(r)) + \sum_g \min(I_G(g), Q_G(g)) + \sum_b \min(I_B(b), Q_B(b))}{\min(|I|, |Q|) * 3} \quad (6)$$

Note that the value of $S_c^{HI}(I,Q) \in [0,1]$. If the histograms I and Q are identical, then $S_c^{HI}(I,Q) = 1$. If either of the two images (query or database) is completely contained in the other, then $S_c^{HI}(I,Q) = 1$.

Modified Euclidean Distance proposed by Jain and Vailaya

In their work**,** Jain and Vailaya have redefined Euclidean distance in a way that gives normalized metrics of similarity. They define the similarity between a query image and a stored database image, $S_c^{ED}(I,Q)$, as

$$S_c^{ED}(I,Q) = 1.0 - \frac{\sqrt{\sum_r (I_R(r) - Q_R(r))^2 + \sum_g (I_G(g) - Q_G(g))^2 + \sum_b (I_B(b) - Q_B(b))^2}}{2 * 3} \quad (7)$$

Note that the value of $S_c^{ED}(I,Q) \in [0,1]$. If images I and Q are identical, then $S_c^{ED}(I,Q) = 1$.

III. CONTENT BASED IMAGE RETRIEVAL APPLICATION

We have developed a web-based application for large-scale study of image retrieval algorithms as shown in Fig. 2. The system retrieves images that are similar to a user-specified query from an image database. The system has been developed on Linux OS, Apache as Web Server, and PHP for server side, JavaScript and HTML for client side, and C Language for CGI script. The number of images currently in the database is 28,168. We are expanding the database by adding more images from time to time. The image features i.e. color histograms are stored in a file structure. The number of bins is 64 for all the histogram types. When accessed, the application system displays 20 randomly picked images from its database in 4 rows (with selected options and upload facility). The default options are 20 for the number of images, RGB as the histogram type and Euclidean as the distance metric. We next describe some of the important components of the application.

Query Specification

A query in the application is specified by an example image. The example image is selected either by clicking on any of the displayed images or by uploading an external image. Initially, a random set of 20 images is displayed. Any of these images may be selected as an example image. Refreshing the page on the web browser displays a new set of 20 random images. The number of images to be retrieved and displayed is selected as an input parameter. The histogram type and the distance metric can also be chosen for retrieval. Under the option to display number of images, one can select 10, 20 and 30. The Histogram type can be any of the five color histograms mentioned in section 1. For the RGB based histogram, the two HSV based histograms and QBIC histogram, the possible choices of distance metric are (i) Euclidean Distance (ii) Manhattan Distance and (iii) Vector Cosine Angle Distance. For Jain and Vailaya's histogram, the possible distance metrics are (i) Histogram Intersection and (ii) Euclidean distance.

Display of Result Set

The nearest neighbor result set is retrieved from the image database based on the query image and is displayed as shown in Fig. 3. The distance value from the query image is printed below each image. The retrieval process considers the parameter values selected in the options boxes for the number of nearest neighbors, histogram and distance metric.

Histogram Display

One of the objectives of our long term research goal is to study the properties of different color histograms and how they affect nearest neighbor query for a variety of distance metrics. To get an insight into this aspect, we have made a provision for displaying the histograms. The " Histogram" hyper link on the result page displays all the retrieved histograms as shown in Fig 4. On each of these result set histograms, we also display the query image histogram for effective comparison.

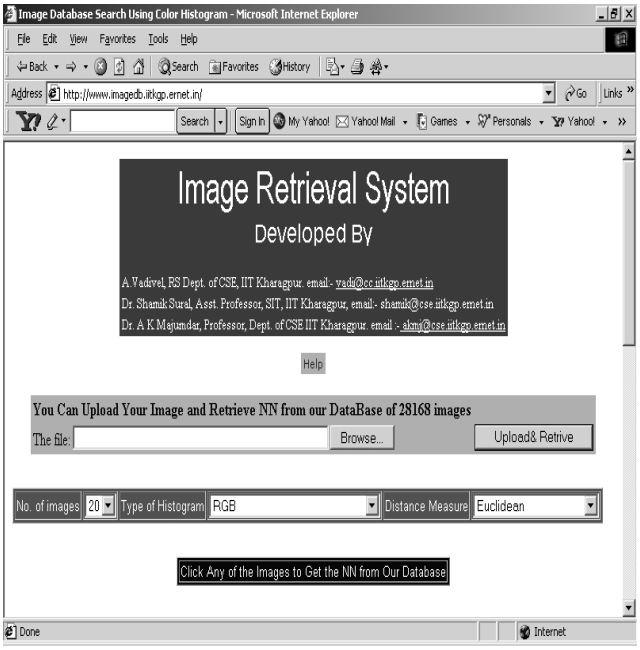


Figure 2. Web based Image Retrieval System.

Figure 3. Result Set display in the application.

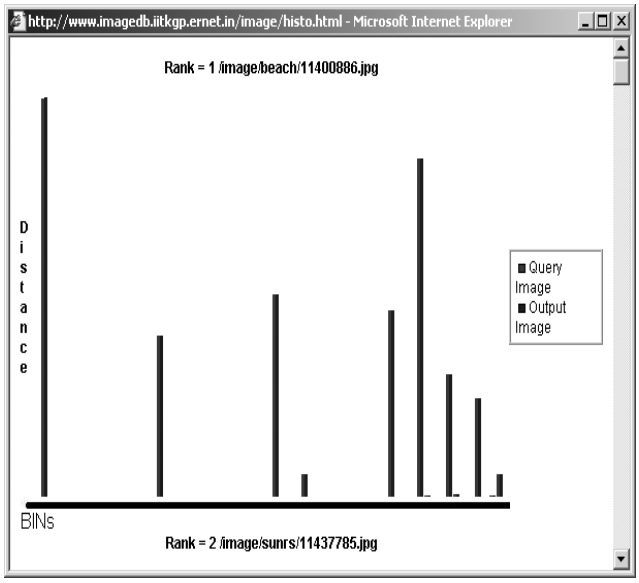


Figure 4. Histogram Display in the application.

External Image Upload

Users are often interested in retrieving images similar to their own query image. To facilitate this, we provide a utility to upload an external image file and use the image as a query on the database. By this facility, readers can also evaluate the performance of our application.

User Guide and Help

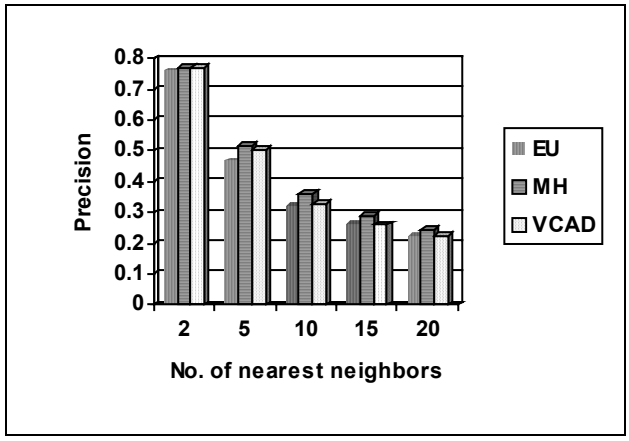
We provide a Help Page to guide the user effectively through the application. All the different options, details about the histograms and the distance metrics including the formula are explained in this page.

IV. EXPERIMENTAL RESULTS

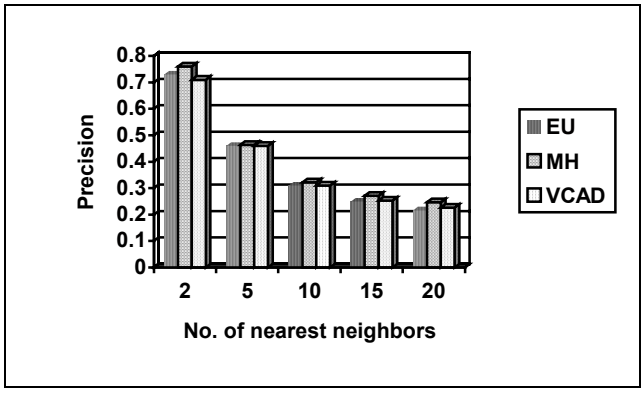
Performance comparison of content-based image retrieval systems is a non-trivial task since it is very difficult to determine the relevant sets. Ground-truthing of a large image database is a resource and cost intensive work. In the absence of ground truth, determining Recall as a performance metric is not possible. The only way to present performance objectively is through precision, which is defined as follows.

$$\Pr ecision = \frac{No.of\ relevant\ images\ retrieved.}{Total\ no.of\ images\ retrieved.}$$

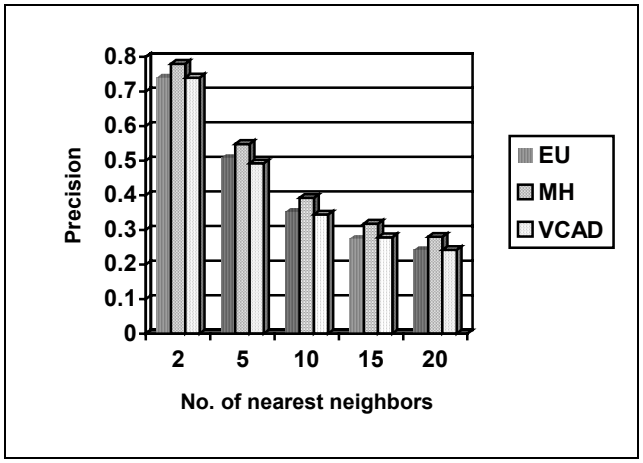
Even though we do not exactly know the relevant set, the observer's feeling about relevant images in the retrieved set is what is used as a measure of precision. In each of the experiments performed, we have calculated the metric of precision for 50 randomly selected images and taken the average. Precision is shown for the first 2, 5, 10, 15 and 20 nearest neighbor images of the result set. The experiment is carried out for all the five histograms mentioned in the section 1. For each histogram, all the supported distance metrics are used for the calculation. In figure 5(a)-(e) we have shown the results of retrieval in terms of precision for the five histograms. In each figure, the results are displayed for all the relevant distance metrics.



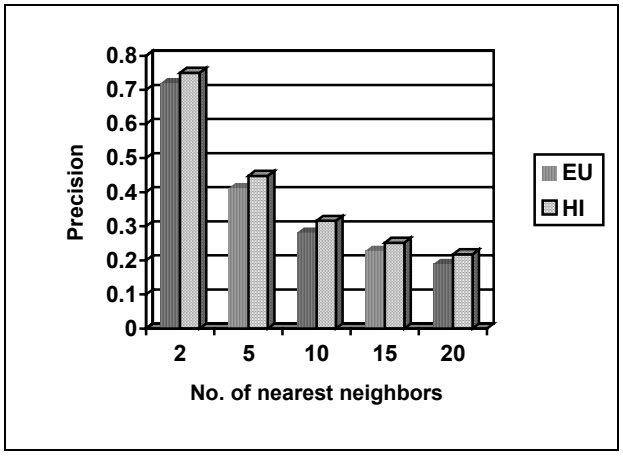
5(a)



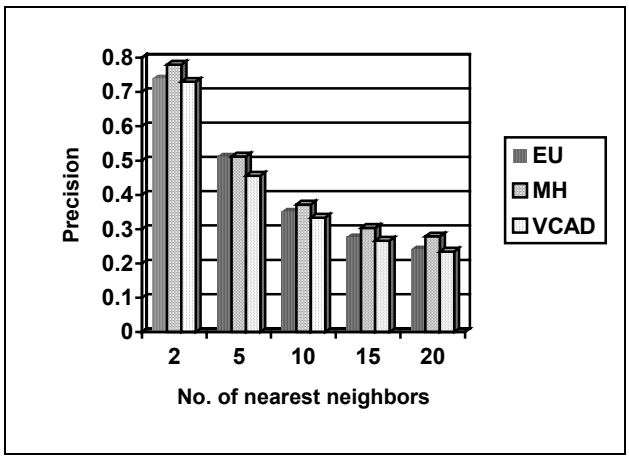
5(b)



5(c)



5(d)



5(e)

Figure 5. Performance comparison of the distance metrics. Precision variation with nearest neighbor for (a) RGB Histogram (b) HSV Histogram (c) HSVSD Histogram (d) JV Histogram and (e) QBIC Histogram.

From the figures, we see that with increase in the number of nearest neighbors, the precision goes down for each observation. The reason is that all the distance metrics tend to bring fairly close images correctly as the first few elements in the retrieval set. However, with higher number of nearest neighbors spurious images keep coming in the result set. This feature is well reflected in our application. In all the figures except 5(d), Manhattan distance gives the best-retrieved result. For the Jain and Vailaya's histogram (fig. 5(d)), Histogram Intersection performs better than the Euclidean distance. However, the precision obtained by the Histogram Intersection distance is less than the precision obtained by Manhattan distance for all the other histograms.

V. CONCLUSION

Color histograms and distance metric definitions have a lot of impact on the performance of image retrieval systems. However, there is no existing research work that does a comparative analysis of distance metrics on a variety of color histograms using large image databases. In our work, we have considered some of the standard and recently proposed color histograms as well as four commonly used distance metrics. A very large database of more than 28,000 images was used for detailed study. From the results we conclude that Manhattan distance gives the best performance in terms of precision of retrieved images. It should be noted that the web-based application developed by us is being continuously improved. A supervisory learning module is being planned that will provide relevance feed back [4] to the system regarding correctly retrieved images to bring further objectivity in our observations. We may include provision of feedback from multiple users. Another important observation is that Vector Cosine Angle distance, which considered as the do-facto distance metric in text retrieval, performs almost the same as Euclidean distance for content-based image retrieval applications also. Due to its two important properties mentioned in section 2 of the paper, we feel that VCAD is a natural choice while combining multiple features like color, texture and shape. We would like to perform analysis of other distance metrics like Earth Mover's distance [5] and weighted Euclidean distance [1,3] as a continuation of our work.

VI. REFERENCES

[1] J. Hafner, H. Sawhney, W. Equitz, M. Flickner, and W. Niblack, "Efficient Color Histogram Indexing for Quadratic Form Distance Functions", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 7, pp. 729-736, 1995.

[2] A. Jain and A. Vailaya. "Image Retrieval using Color and Shape", *Pattern Recognition*, 29(8), pp. 1233–1244, 1996.

[3] W. Niblack, R. Barber,W. Equitz, M. Flickner, E. Glasman, D. Pektovic, P. Yanker, C. Faloutsos, and G. Taubin,, "The QBIC Project: Querying Images by Content using Color Texture and Shape", *Proc. SPIE Int. Soc. Opt. Eng., in Storage and Retrieval for Image and Video Databases*, vol. 1908, pp. 173-187, 1993.

[4] M.Ortega , Y. Rui, K. Chakrabarti, K. Porkaew, S. Meharotra and T. S. Huang "Supporting Ranked

Boolean Similarity Queries in MARS", *IEEE Transactions on Knowledge and Data Engineering*, vol. 10, no. 6, pp. 905-925, 1998.

[5] Y. Rubner, L.J.Guibas and C.Tomasi, "The Earth Mover's Distance, Multi-dimensional Scaling, and Color-based Image Retrieval", *Proc. DARPA IU Workshop, New Orleans, LA*, pp. 661-668, May 1997.

[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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349-1380, December 2000.

[7] S. Sural, "Histogram Generation from the HSV Color Space using Saturation Projection", in "Multimedia Systems and Content-based Image Retrieval", *S.Deb eds, Idea Group Publishing, Hershey, PA, USA*, 2003 (in press).

[8] S.Sural, G. Qian and S.Pramanik,"Segmentation and Histogram Generation using the HSV Color Space for Content Based Image Retrieval", *IEEE Int. Conf. on Image Processing, Rochester, NY*, 2002.

[9] M.J. Swain and D. H. Ballard, " Color Indexing,"*International Journal of Computer Vision* vol. 7, no. 1, pp. 11-32, 1991.

[10] R. B.Yates and B. R. Neto, Modern Information Retrieval," *ACM Press*, pp. 27-28, 1999.