I.J. Image, Graphics and Signal Processing, **2016, 9, 10-16**
Published Online September 2016 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2016.09.02

Effect of Reducing Colors Number on the Performance of CBIR System

Abbas H. Hassin Alasadi
Computer Science Department, Science College, Basra University, Basra, Iraq.
ACIT, IJPRAI, IJCT, UJCT member.
Email: abbashh2002@gmail.com

Saba Abdual Wahid
Computer Science Department, Science College, Basra University, Basra, Iraq.
Email: sabalobody@yahoo.com

Abstract—Taking inspiration from the fact that a human can distinguish only a limited number of colors, reducing the number of colors is an interesting task to be incorporated in image retrieval systems that is based on using only the most discriminative colors, which most of the time yields better results.

Accordingly, the main goal of this paper is to study the influence on performance of reducing the colors number contained in images. Accomplishing this task poses an extra overhead on the system, which requires more computation time, but, on the other hand, can accelerate the comparison process. Due to their popularity and success, we specifically concentrate this study on histogram indexing methods, using both Euclidean distance and histogram intersection to assess consequently the distance and the similarity between images. Some simple, pertinent ideas related to the way we compare a pair of images using Euclidean Distance are given in the end of the paper, supported by preliminary obtained results**.**

***Index Terms*—**CBIR, image retrieval, histogram indexing, colors reduction, histogram intersection**.**

I. Introduction

Image retrieval falls under the area of information retrieval field, with the specificity that the information for which the user is seeking, in this case, bears the image form. Many interesting efforts have been made in this field [1], ranging from using methods coming from textual retrieval (TBIR: Text Based Image Retrieval), e.g. [2, 3] to employing content based image retrieval (CBIR) methods, e.g. QBIC system [4, 5]. Although all this significant received attention, image retrieval is still an open problem because the up to now proposed systems failed to satisfy completely the user needs.

Taking inspiration from the fact that a human can distinguish only a limited number of colors, reducing the number of colors is an interesting task to be incorporated in image retrieval systems that is based on using only the most discriminative colors, which most of the time yields better results.

Accordingly, the main objective of this paper is to study the influence on performance of reducing the colors number contained in images. Accomplishing this task poses an extra overhead on the system, which requires more computation time, but, on the other hand, can accelerate the comparison process. In this work, we are interested in histogram indexing methods, using both Euclidean distance and histogram intersection to assess consequently the distance and the similarity between images. Some simple ideas related to the way we compare a pair of images using Euclidean distance are given in the end of the paper, supported by preliminary obtained results.

The rest of this paper is arranged as follows. Firstly, survey the different processes of which an image retrieval system consists. Histogram indexing methods are covered in section two, whereas section three is devoted to how to assess the closeness between a pair of images using a histogram as a feature. In section four, we put under light reducing colors number process. We show, in section dive, the experimentation conducted and we try to discuss the results obtained. Finally, section six indicates the conclusion with some perspectives.

II. Content-Based Image Retrieval System

As an information retrieval system, an image retrieval system consists of two main processes (Fig. 1). The first one is the indexing process, the aim of which is to encode an image in some simpler form, generally consisting in extraction of particular features. The second one, which receives the output of the first process as input, is intended to assess the similarity between two images representation. Accordingly, in order to ameliorate the global performance of such system, effort has to be put into one of these two processes.

III. Histogram Indexing Method

Many indexing methods for CBIR are reported in the literature. However, a color histogram based technique reported in [5-8] proved to be an effective one through the years. It has been used in many works and, although it is one of the oldest methods, many researchers admit it as a basic method for CBIR. It has influenced many works, although, Histograms hold some major shortcomings [9-11]. For instance, and because Histograms are unable to spatially locate colors in images, many efforts have been achieved to enhance this approach to enable taking into account the location information of colors.

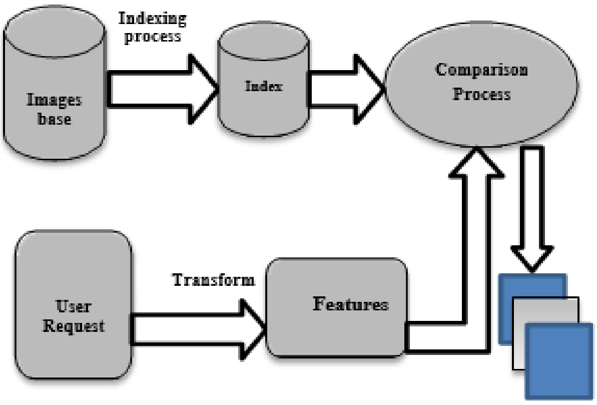

Fig.1. Schema Representing Different Compounds of a generic CBIR System.

The histogram itself is a statistic vector, the elements of which hold the pixels count for each color in the image. We also report the following well known pros and cons for Histograms:

- Histograms are sensitive to noisy interferences such as illumination changes and quantization errors.
- Large dimension of histogram involves large computation on indexing.
- Histograms do not take into consideration color similarity across different bins.
- Histograms cannot locate objects within an image.
- Two perceptually very different images with similar color distribution will be considered similar by a color histogram based retrieval system (Fig. 2).

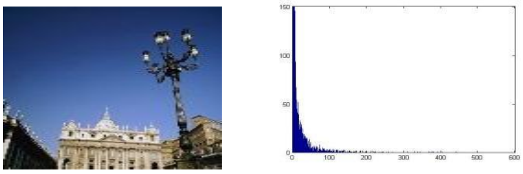

Fig.2. An Image from the Wang Images Database and Its Histogram Distribution.

IV. Comparison of Histograms

The very important task of a CBIR system is to assess the degree of resemblance (similarity measure) or dissemblance (distance measure) between two images. Indexing the image using some features extracted from the images aims to represent images in much simpler form. In this work, we use the Euclidean distance and the histograms intersection similarity to compare pairs of images.

The Euclidean distance is defined by (1):

$$l_2\left(H,H'\right)=\left(\sum_{m=1}^{M}\left(H_m,H'_m\right)^2\right)^{½} \qquad (1)$$

Where H_m and H'_m are the two histograms whose similarity is being assessed and M is their common bins number.

The histogram intersection formula is given by (2):

$$d\cap\left(F^q,F^d\right)=\sum_{j=1}^{n}\min\left(F^q,F^d\right) \qquad (2)$$

Where F^q and F^d are the two histograms to compare and n is the number of bins.

V. Reducing Colors Number process

Reducing colors number is the key process of our approach. The aim is to take into consideration, during the comparison process, only a limited number of perceptually significant (dominant) colors within images. Accordingly, we split the image colors (histogram colors) into two classes: colors, which have important occurrences in the image (colors that have highest frequencies in the global color histogram of the image) and colors, which have fewer occurrences. The formers are considered as dominant colors, while the latter are designated as non- dominant. The process of reducing the colors number implies a re-ordering of colors in terms of their occurrence rate in the image. The top k-best colors are kept to encode the image, k being the pre-set value. In a later step, the non- dominant colors in the image are replaced by their respective nearest dominant colors in the RGB space in the Euclidean distance sense. Fig. 3 shows an illustration case, with an image containing originally 54103 colors (left) and on the right, its transform when reducing the color numbers to 16 dominant colors. The occurrence number of each dominant color is updated.

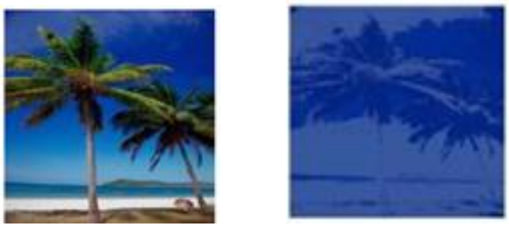

Fig.3. An image without and with undergoing the reducing process.

The color reduction step is useful at least for the following reasons:

- Taking into account all existing colors in the image during the comparison process leads to a system that might not comply with real time requirements.

It is known that an RGB image, where each pixel is encoded on 24 bits yields a huge Histogram of 16 777 216 positions (potential colors).

- The incorporation of the colors reduction has little impact on the response time because indexing is an offline task.
- Not all colors existing in the image are important from a discrimination viewpoint. The discrimination power is mainly concentrated within a few dominant colors.
- In our view, taking into account only some colors of the image makes the system closer to the human perception system, which can distinguish objects within images relying on a small number of colors.

VI. Experimentation and Results

As mentioned above, we use two methods helping to assess the similarity or the dissimilarity between a pair of images; we obtain then four cases. The number of dominant colors to keep (k) is another arguable question. In the following experiments, we keep 16, 8, 4 or even just 2 colors.

Another question that arises is how to assess a CBIR system? There are mainly two widely used measures: the precision and the recall [12, 13].

The precision is defined as the ratio of relevant images retrieved to all images retrieved, while the recall is defined as the ratio of relevant images retrieved to all relevant images in a database, or the probability given that an image is relevant that it will be retrieved.

$$\text{Precision} = \frac{\text{Number of relevant images retrieved}}{\text{Total number of images retrieved}} \quad (3)$$

$$Recall = \frac{\text{Number of relevant images retrieved}}{\text{Total number of relevant images in the Database}} \quad (4)$$

Another question to ask here is what is the number of images retrieved to take into account for computing these metrics? It is known that the precision decreases when we increase the number of retrieved images. This is not the case for the recall, which has a tradeoff with the precision. As a compromise, and, although the choice of the number of images does not affect the desired comparison, but, instead, affects the evaluation of the histogram indexing method and the searching algorithm, namely, the Euclidean distance and histogram intersection, the computation of both metrics is done taking into account 10, 20 and 30 first retrieved images. Therefore, the precision and the recall are computed in all these cases; the results are shown in the tables below.

Note that, the experiments were carried out on a collection of 100 images selected from the Wang database [http://Wang.ist.psu.edu/docs/related.shtml], the precisions and the recalls, depicted in the Tables 1 – 3 , are the average values and the queries have been chosen randomly.

Table 1. Taking into account the 10 first images.

Without reducing				Keeping 16 dominant colors				Keeping 8 dominant colors				keeping 4 dominant colors				Keeping 2 dominant colors			
ED		HI		ED		HI		ED		HI		ED		HI		ED		HI	
P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R
10 %	10 %	10 %	10 %	23 %	23 %	40 %	40 %	21 %	21 %	30 %	30 %	22 %	22 %	30 %	30 %	20 %	20 %	/	/

Table 2. Taking into account the 20 first images retrieved.

Without reducing				Keeping 16 dominant colors				Keeping 8 dominant colors				keeping 4 dominant colors				Keeping 2 dominant colors			
ED		HI		ED		HI		ED		HI		ED		HI		ED		HI	
P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R
20 %	40 %	5 %	10 %	15,5 %	31 %	35 %	70 %	45 %	29 %	/	/	14 %	28 %	/	/	14,5 %	29 %	/	/

Table 3. Taking into account 30 first images retrieved

Without reducing				Keeping 16 dominant colors				Keeping 8 dominant colors				keeping 4 dominant colors				Keeping 2 dominant colors			
ED		HI		ED		HI		ED		HI		ED		HI		ED		HI	
P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R
16,66 %	50 %	10 %	30 %	13,33 %	40 %	/	/	12,66 %	38 %	/	/	12,66 %	38 %	/	/	12,33 %	37 %	/	/

Where:

P: is the average precision found.
R: is the average Recall found.
ED: is the Euclidean distance.
HI: is the histogram intersection method.
/: designs that we cannot compute the precision or the recall.

As can be seen from Table 1, the high value of precision and recall we can reach is 40% when employing only 16 colors and employing the intersection histogram method. It is clear that the performance has been improved using the reducing process and so is the case, no matter the number of the kept colors.

The results outlined in Table 2 reveal that the values of precision and recall found in the case when using the colors reduction process are higher than those found without reduction, but, in this last case, the highest values are found separately. Indeed the best value of precision (45%) is provided when keeping 8 colors and making use of the Euclidean distance while the best value of recall (70%) is reached when keeping 16 colors and using the histogram intersection method.

As described in Table 3, usage of the Euclidean distance without reducing colors achieved better results than the other cases with a little difference compared to those found when keeping 16 colors and using the Euclidean distance.

We then can claim that the reduction of the number of colors is advised when using the Euclidean distance (looking for the dissimilarity). According to the results, 16 is the best choice for the number of colors to keep. In contrast, using the histogram intersection requires a little computation time and the incorporation of the reduction process in this case increase the chance to get zero as a value of the intersection that makes the re-ordering of the images returned difficult and as consequence, we cannot compute the recall and the precision. In summary, if we choose reduction when using the histogram intersection method, the number of colors to keep has to be inferior to the number of retrieved.

In the light of the above comments, we conclude that the best case is when taking into consideration 16 dominant colors and using the Euclidean distance. Fig. 4 and Fig. 5 show with high accuracy this situation by giving the precision and recall for all queries conducted rather than the average given in the pre-cited tables.

The colors reduction consists in encoding the image relying on the dominant colors without removing the non-dominant ones. Indeed, this process adds the value of each non-dominant color to the value of the nearest dominant one. In the following experiment, we rely just on the dominant colors, that is to say we do not take into account the non-dominant ones. By doing so, we will certainly ameliorate the complexity and, as a consequence, the computation time. Fig. 6 to Fig. 8 show the average precision and recall in the two cases: reduction with and without taking into account the non-dominant colors.

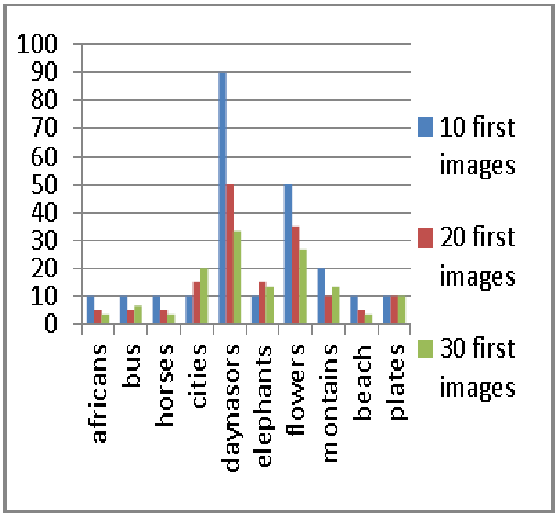

Fig.4. Precision when keeping 16 colors using the Euclidean distance with different number of images retrieved taken into account.

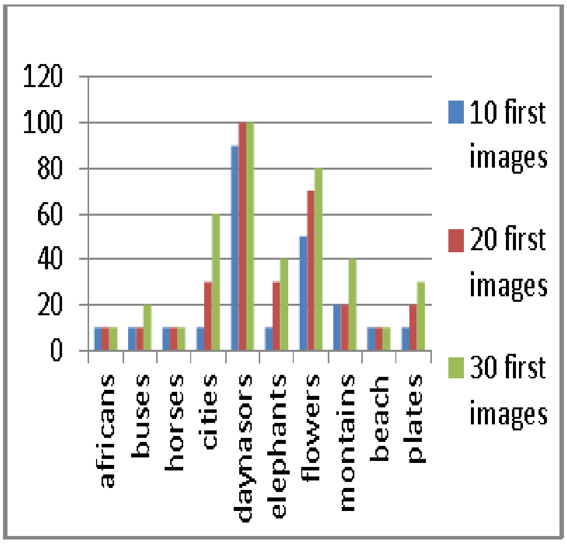

Fig.5. The recall when keeping 16 colors using the Euclidean distance with different number of first images retrieved taken into account.

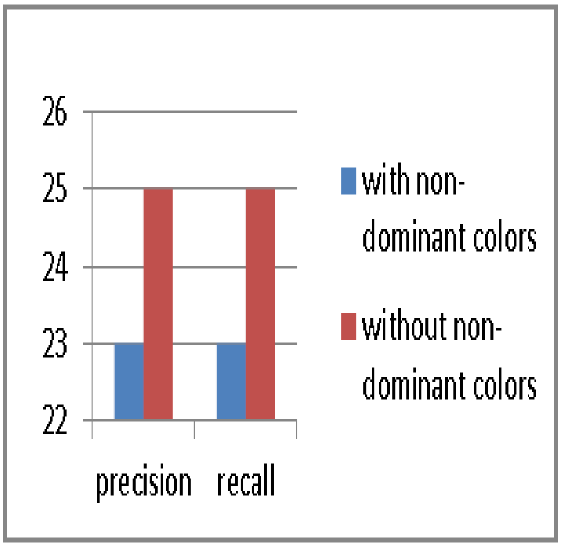

Fig.6. Precision and recall when taking into account 10 first images retrieved.

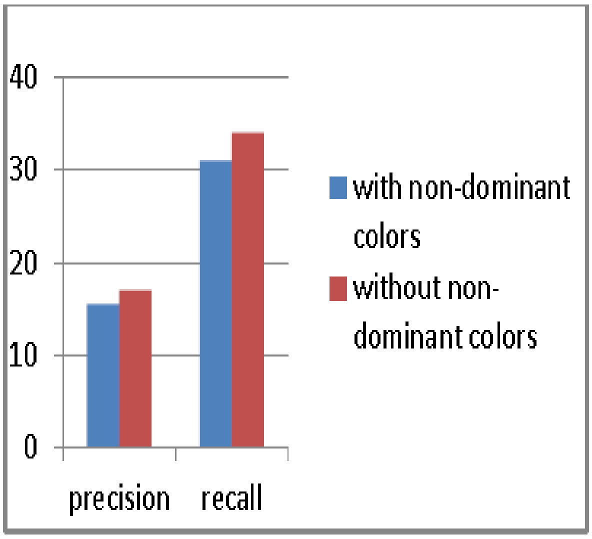

Fig.7. Precision and recall when taking into account 20 first images retrieved.

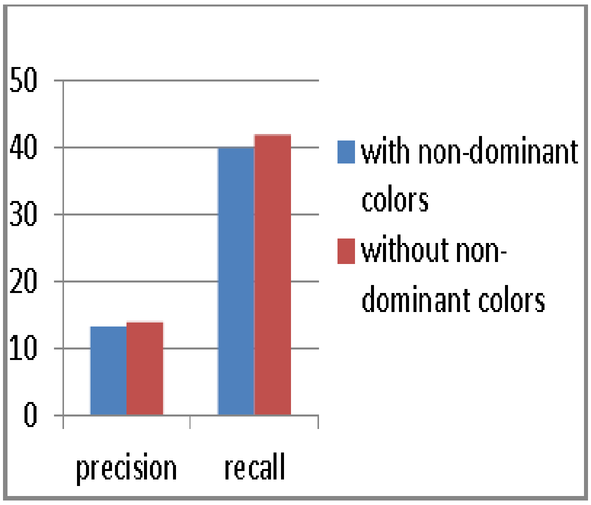

Fig.8. Precision and recall when taking into account 30 first images retrieved.

Based on these results, we claim that removing non-dominant colors when encoding the image based on dominant ones provides a good performance in addition to the gain reaped in terms of decreasing the complexity.

Now, we change the way we compute the distance between the 16 dominant colors. This distance, as described above, is calculated between the 16 bins of each image, taking consideration that the compared bins should be of the same color. This time, we will perform the distance differently. We first order the bins in terms of their occurrence number. Then, the bins are compared without checking the nature of the bin. That is, the top ranked bins will be compared together, and then the bins ranked second and so on. Applying this new distance, we get the results depicted in the Fig. 9 to Fig. 11.

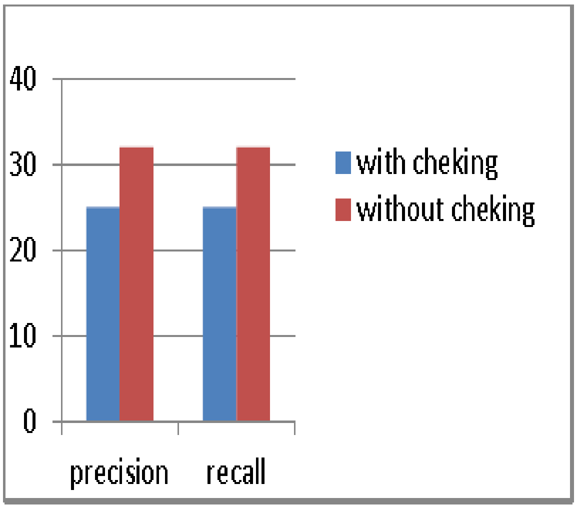

Fig.9. Precision and recall with and without cheking the nature of the bin to compare taking into account 10 first images retrieved.

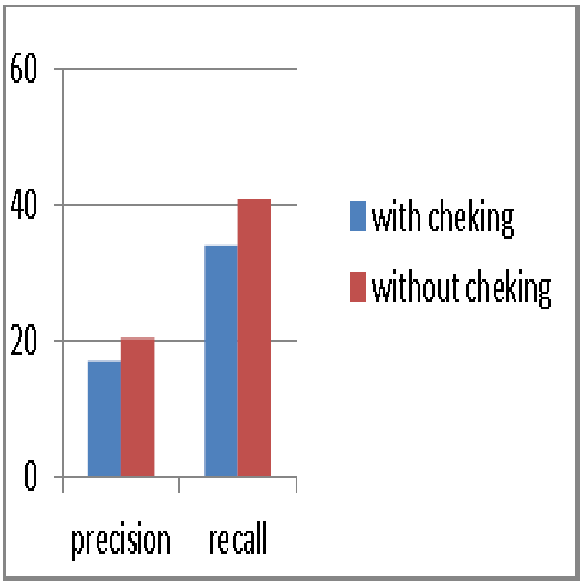

Fig.10. Precision and recall with and without cheking the nature of the bin compared taking into account 20 first images retrieved.

Even with this second modification, the results have been improved, and so during all the three scenarios. According to that, we claim that among the three cases: reducing colors to 16 colors using the Euclidean distance (case1), reducing colors to 16 colors without taking into account the non-dominant colors (case2), and reducing colors to 16 bins without taking into account the non-dominant colors and comparing the bins in terms of their order number rather than the equality of their colors (case3), the best one is the last case (case 3). The following precision-recall curves shown in the Fig. 12 illustrate this reality.

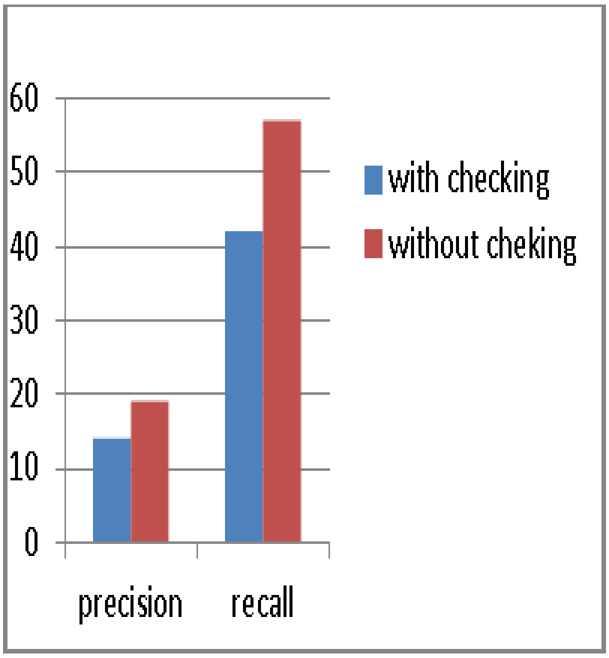

Fig.11. Precision and recall with and without checking the nature of the bin compared taking into account the 30 first images retrieved.

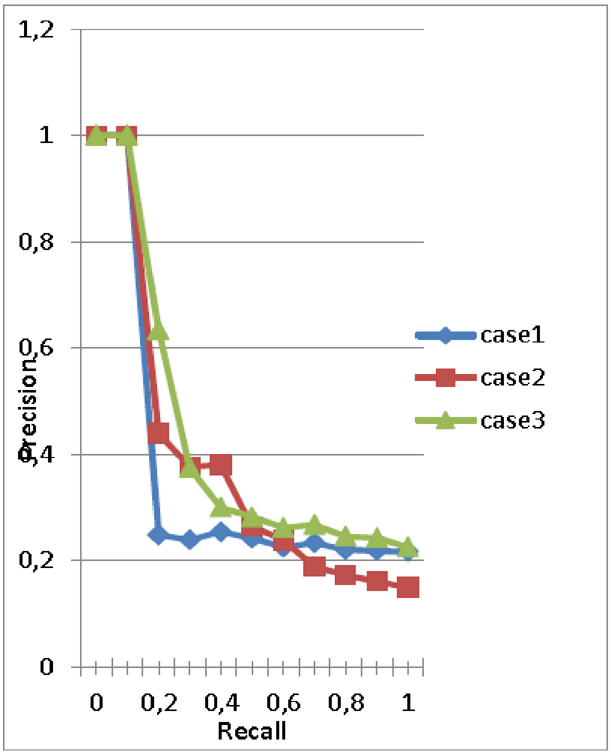

Fig.12. Precision vs. Recall curves for the tree cases.

VII. Conclusion

In this paper, we considered the reduction of the colors number in images and its influence on the performance of a CBIR system. In the light of the obtained results, we claim that the case when keeping 16 colors combined to the Euclidean distance without taking into account the non-dominant colors and comparing bins in terms of their order number rather than the equality of their colors is the best case. On the other hand, the histogram intersection method does not need the colors reduction process owing to its small computation time. Adopting the reduction in this case may lead to the situation resulting in the Boolean model, which makes the ordering of the images retrieved not possible to perform.

References

[1] Abbas H. Hassin and Ali B. Yousif (2013). Content-based Image Retrieval using Texture and Color Features. *Journal of Thi-Qar Science*, 3(4):142-149

[2] Barya, N., & Jaiswal, H. (2015). Survey on Content based Image Retrieval to Deal with Rapid Growth of Digital Images. *International Journal of Computer Applications*, *124*(12).

[3] Dharani, T., & Aroquiaraj, I. L. (2013, February). A survey on content based image retrieval. In *Pattern Recognition, Informatics and Mobile Engineering (PRIME), 2013 International Conference on* (pp. 485-490). IEEE.

[4] Flickner, M., Sawhney, H., Niblack, W., *et al* (1995). Query by image and video content: The QBIC system.*Computer*, *28*(9), 23-32.

[5] Jaworska, T. (2016). Query Techniques for CBIR. In Flexible Query Answering Systems 2015 (pp. 403-416). Springer International Publishing.

[6] Ibrahim S. I. Abuhaiba,Ruba A. A. Salamah (2012). *Efficient Global and Region Content Based Image Retrieval*, IJIGSP, 4 (5),pp.38-46.

[7] Sharma, N. S., Rawat, P. S., & Singh, J. S. (2011). Efficient CBIR using color histogram processing. *Signal & Image Processing*, *2*(1).

[8] Shrivastava, N., & Tyagi, V. (2015). An efficient technique for retrieval of color images in large databases. *Computers & Electrical Engineering*, *46*, 314-327.

[9] Juneja, K., Verma, A., Goel, S., & Goel, S. (2015, February). A survey on recent image indexing and retrieval techniques for low-level feature extraction in CBIR systems. In *Computational Intelligence & Communication Technology (CICT), 2015 IEEE International Conference on* (pp. 67-72). IEEE.

[10] Singha, M., & Hemachandran, K. (2012). Content based image retrieval using color and texture. *Signal & Image Processing: An International Journal (SIPIJ)*, *3*(1), 39-57.

[11] Kekre, H. B., & Sonawane, K. (2013). Histogram Bins Matching Approach for CBIR Based on Linear grouping for Dimensionality Reduction. *International Journal of Image, Graphics and Signal Processing*, 6(1), 68.

[12] Jasmine, K. P., & Kumar, P. R. (2014). Color Histogram and DBC Co-Occurrence Matrix for Content Based Image Retrieval. *International Journal of Information Engineering and Electronic Business* (IJIEEB), 6(6), 47.

[13] Rahmani, M. K. I., Ansari, M. A., & Goel, A. K. (2015, February). An Efficient Indexing Algorithm for CBIR. In *Computational Intelligence & Communication Technology (CICT), 2015 IEEE International Conference on* (pp. 73-77). IEEE.

[14] Liu, G. H., & Yang, J. Y. (2013). Content-based image retrieval using color difference histogram. *Pattern Recognition*, *46*(1), 188-198.

[15] Patil, C. G., Kolte, M. T., Chatur, P. N., & Chaudhari, D. S. (2014). Optimum Features selection by fusion using Genetic Algorithm in CBIR. *International Journal of Image, Graphics and Signal Processing* (IJIGSP), 7(1), 25.

Authors' Profiles

Abbas H. Hassin Alasadi is Assistant Professor and Postgraduate Program Coordinator of the Department of Computer Science at Basra University. He received his PhD degree from School of Engineering and Computer Science / Harbin Institute of Technology, China. He spent more than ten years as Assistant Professor at different Universities abroad the current position. His research interests include Medical Image processing, Biometrics, Information retrieval, and Human-computer interaction. His research work have been published in various international journals and conferences. Dr. Abbas is an active reviewer in many journals of the areas of computer science and software engineering. He is one of ACIT, UJCS, and IJPRAI members.

Saba Abdual Wahid is a lecturer at the Computer Science Department, Science College, Basra University, Basra, Iraq. She received her BSc from Basra University in Computer Science major. In 200**7**, she finished her MSc from Basra University. Her research interests include soft computing and image processing.

How to cite this paper: Abbas H. Hassin Alasadi, Saba Abdual Wahid,"Effect of Reducing Colors Number on the Performance of CBIR System", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.8, No.9, pp.10-16, 2016.DOI: 10.5815/ijigsp.2016.09.02