Periodicals of Engineering and Natural Sciences ISSN 2303-4521
Vol. 9, No. 1, January 2021, pp.59-70

Improve a technique for searching and indexing images utilizing content investigation

Suad Kakil Ahmed [1], Huda Abdalkareem Mardan [2]
[1,2] Computer System Department, Technical Institute of Kirkuk, Northern Technical University

ABSTRACT

In this research, algorithms were developed to assess the similarities between two or more images and reduce the time spent on searching for them based on the analysis of color intensity diversity and (histogram analysis), besides analyzing the partial proportion of the color component of the images. These algorithms can be used to search for images in the database, systems, and computer networks. This experiment was carried out using a computer program that was programmed using C Builder. The results were analyzed and illustrated with multiple examples. The results showed that the time spent in the searching process depends on several criteria. The most important of these criteria is the number of images used in the search process, number and method of processing the images involved in the search process, in addition to evaluating the time spent in the search process of the three algorithms whose effectiveness has been evaluated. (The best algorithm according to results and performance).

Keywords: Histogram, digital processing of the images, modeling research, the intensity of images' pixels

Corresponding Author:
Suad Kakil Ahmed
Computer System Department
Northern Technical University
Kirkuk, Iraq
E-mail: suadkakil@ntu.edu.iq

1. Introduction

At present, the process of data and search engines are widely required such as searching for images in the databases, similar and original images, clips, and others. There is a database dealing with images, artwork, satellite images, and photography, as well as a collection of international photographs. The databases used to store these images can be very large as they contain hundreds of thousands or even millions of images. In most cases, such databases are indexed by keywords which are stored by humans or users who classify images under specific categories, e.g., images are stored in the database in words that describe them. The most important previous studies used the various methods of searching for images based on their contents. The first study is scientific study [1]. In this research paper, the image was fragmented using a staph algorithm K-Mean which is used for pattern recognition research [2] to enable researching and retrieving to the image fragment that categorized into points containing objects and other points that do not contain objects to use proportionate properties to the type. The points containing objects have also aggregated parts, while the points that do not contain objects spread over the entire image scene as determined by the mathematical equation(x2). The performance of indexing four areas was assessed by histogram and three other areas were assessed by wavelet. The results showed that the indexing of areas based on wavelet provides a good comparison for the areas that do not contain objects[3]. While the indexing areas using histogram provides a good comparing for the areas that. This research paper focuses on indexing and searching for images based on (Wavelet) and the algorithm used in indexing and retrieving of images by laying out the image into parts to be capable of searching for them within the databases; besides the algorithm used in (Daubechies' Wavelet) that turns the color of each element of the image into small waves to accelerate the retrieving process. The results of the research were carried out by comparing the selected image and the image involved in the databases. The researcher assessed the results by searching for an image in a database that is consists of 10,000 images; the best 100 images were taken in 3.3 seconds.

1.1. The research problem

The searching process for images based on indexing and storing words in the database requires a long time and therefore does not provide the necessary flexibility to search for the required image, where several images may resemble the same label.

1.2. The aims of the research

The research aims to reduce the time spent on searching for similar images in the databases by searching for images based on (search by model) that's when the user selects a part of an image or a complete image, searches for similar images in databases, places it as a research guide in the database and compare the similarities between these images, to ensure the accuracy of the search results.

1.3. Research methodology

Mathematical modeling methods, programming, digital signal processing techniques, and imagery, were used in this research.

2.1. **Digital images**

Any image can be defined as (two-dimensional equation), d(x and y) where x and y are equal coordinates. The value of the equation (d), on the coordinate axes, is called (grey level) or (intensity) of a point. We can call it a digital image when x and y values belong to the group of discrete (quantities). These elements are called (picture elements) or (pixels)[4] .

Here we can clarify the meaning of Image Processing, image analysis, vision through computers.

- ***Image Processing:*** It is the process where the inputs are images and the outputs are other images such as image enhancement algorithms.
- ***Vision through computers***: It concerns with simulating (it designed as rule-based in which required a deep understanding of function's behavior to be designed[5], the human ability to see, including the ability to learn, deduct and react according to the visual inputs.
- ***Image analysis:*** It is an intermediate phase between the vision through the computer and image processing. It is somewhat difficult to find a dividing line between these three subjects (Image processing, analysis, and computer vision) However, the processes, in which the computer is used, can be divided into three levels[7 ,6] :

1- Low-level processes that include deformation, contrast enhancement, and quality assurance of the image.

2- Medium-level processes that include image segmentation into areas or elements and describe these elements for reduction to represent computer processing, as well as identification of image-specific elements.

3- High-level processes that include understanding and recognition (making sense) of a set of identified elements. At the top of this level, there are learning and acquiring knowledge associated with computer vision processes (1), (2).

2.2. **Histogram**

It is a descriptive function (diagram) that shows the distribution of color levels and their proportions in the image while the (normalized histogram equation) measures the intensity of each pixel in the image. The method of analyzing by using the histogram is one of the most effective methods for comparing and indexing images, so each image has its histogram [3]. But if the images duplicate each other, their histograms are also duplicated.

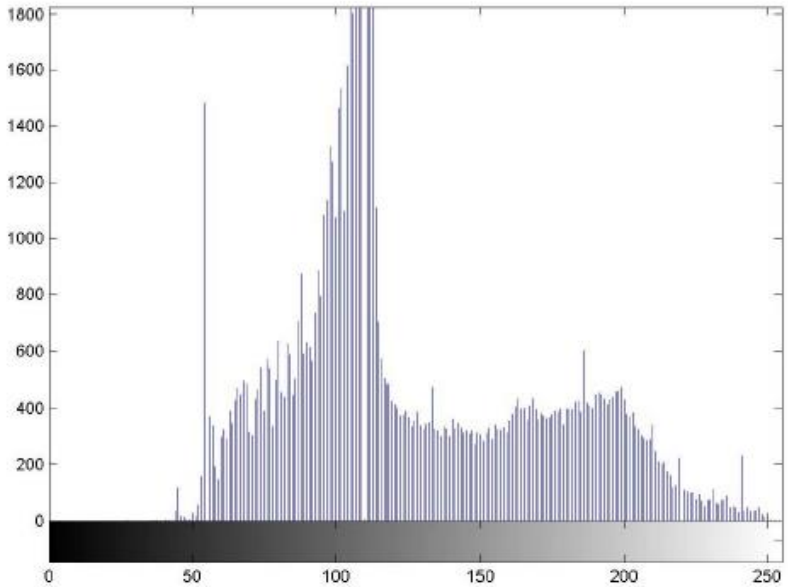

Figure 1. The image histogram

3. Mathematical modeling

3.1. Method of separating colors (RGB) and analyzing color intensity variation

On each byte of a 24-bit image, three bits can be encoded into each pixel (because each pixel is represented by three bytes[8]. If we call the image color (c) then the intensity of colors can be formulated as follows:

(1) $0 \leq I(C) < 256$ or $0 \leq I(C) < I\,max$

The color intensity variation of the image can be calculated by the following equation:

(2) $I(C) = 0.3.\ R(C) + 0.59.\ G(C) + 0.11.\ B(C)$

For comparing images based on analyzing of color intensity, we use the concept of distance between intensity. To compare two images and calculate their similarities, we use the following equation:

(3) $\boldsymbol{d_p}\ (^{1}h,^{2}h) = \sqrt[p]{\frac{1}{N}} \cdot \sum_{i=0}^{n-1} \mathbf{1}\boldsymbol{h_i} - \mathbf{2}\,\boldsymbol{h_i}$

Whereby:

1hi – 2hi: the color intensity variation of the two compared images.

P: represents numbers of images used in the comparing process, here we used two images.

N: represent the color extension of (R, G, B) each of these has a chromatic extension from 0 to 25.

To find out the extent of similarities between the two images used in the analysis process according to the previous equation is that when dp (1h,2 h) decreases, the similarities between the two images increases.

3.2. The method of assessing the similarities between images based on the histogram

In reviewing method 1, the histogram analysis of the image's intensity is a general description of the elements. This approach does not take into account the specificities of each of its color components. So researches have been done to improve the RGB color division and to analyze their intensity taking into account the surrounding [9].

$$R_{r1}(i) = |H1r[i] - H2_r\,[i-1]|,$$

The intensity of red color $R_{r2}(i) = |H1r[i] - H2_r\,[i]|,$ …… (4)

$$R_{r3}(i) = |H1r[i] - H2_r\,[i+1]|$$

$$R_{g1}(i) = \left|H1_g[i] - H2_g\,[i-1]\right|,$$

The intensity of green color $R_{g2}(i) = \left|H1_g[i] - H2_g\,[i]\right|,$ …… (5)

$$R_{g3}(i) = \left|H1_g[i] - H2_{\,g}\ [i+1]\right|$$

$$R_{b1}(i) = |H1_b[i] - H2_b\,[i-1]|,$$

The intensity of blue color $R_{b2}(i) = |H1_b[i] - H2_b\,[i]|,$ ……. (6)

$$R_{b3}(i) = |H1_b[i] - H2_{\,b}\ [i+1]|$$

In the previous three equations (4, 5, 6) we find that h1, h2 is the histogram of the involved image and the image to be compared with it. To calculate the difference between the two images (the least between the histograms of each color component as follows [9]:

The difference between the red color: $S_r = \sum_{i=0}^{n-1} min\ (R_{rk}\ (i)\,)$ $\quad 1 \leq k\ \leq 3$ ….. (7)

The difference between the green color : $S_g = \sum_{i=0}^{n-1} min\ \left(R_{gk}\ (i)\right)$ $\quad 1 \leq k\ \leq 3$ …...(8)

The difference between the blue color: $S_b = \sum_{i=0}^{n-1} min\ (R_{bk}\ (i))$ $\quad 1 \leq k\ \leq 3$ …. (9)

The formula calculates the difference between the three components RGB between the two images can be written as follows:

$$d(H1,H2) = \sqrt{S_r^2 + S_g^2 + S_b^2} \quad \ldots\ldots (10)$$

The general formula for calculating the similarities of two images besides calculating the difference between the three components RGB is as follows:

$$d(H1,H2) = \left(\sum_{i=0}^{N-1} min(|H1_r[i] - H2_r[i-1]|, |H1_r[i] - H2_r[i]|, |H1_r[i] - H2_r[i1+1]|)\right)^2 + \left(\sum_{i=0}^{N-1} \min\left(\left|H1_g[i] - H2_g[i-1]\right|, \left|H1_g[i] - H2_g[i]\right|, \left|H1_g[i] - H2_g[i+1]\right|\right)^2 + \left(\sum_{i=0}^{N-1} min(|H1_b[i] - H2_b[i-1]|, |H1_b[i] - H2_b[i]|, |H1_b - [i]H2_b[i1+1]|)\right)^2 \quad \ldots\ldots (11)$$

8.3. Method of assessing the similarities of images based on color ratio analysis

One of the methods that make the search more flexible is that it is possible to enter specific proportions of colors and search databases for images similar to these proportions (for example, finding an image with a percentage of colors, for example, a blue color 30% red color 20% [10].
Suppose that a set of colors of image pixels is K=NumP[I] and all the images' pixels are NumAll. So it is obvious representing the image as follows :
SumAll = width -height
Whereby, width and height represent the width and height of the image. To calculate the image's color ratio, the equation is as follows:

$$\text{Pr os[i]} = \frac{NumP[i]}{NumAll}.100\%,\ i \in K$$

The previous equation is used for calculating one component of color, but if we want to calculate the three components RGB, we shall use the following equation:

$$Proc[i] = \begin{pmatrix} \frac{NumP\left[i_{GB(R-1)}\right]}{NumAll} + \frac{NumP\left[i_{RB(G-1)}\right]}{NumAll} + \\ \frac{NumP\left[i_{RG(B-1)}\right]}{NumAll} + \frac{NumP[i_{RGB}]}{NumAll} + \\ \frac{NumP\left[i_{GB(+1)}\right]}{NumAll} + \frac{NumP\left[i_{RB(G+1)}\right]}{NumAll} + \\ \frac{NumP[i_{RG(+1)}]}{NumAll} \end{pmatrix}.100\ \%, \mathrm{I} \in K \quad \ldots\ldots \quad (12)$$

To compare the similarities of two images, their mathematical equivalent is as follows:

$$d(A,B) = \sqrt{\frac{1}{K}.\sum_{i=0}^{K-1} (\text{Pr ocA[i]} - \text{Pr ocB[i]})^2} \quad \ldots. (13)$$

Whereby: A and B represent the value of color ratios of the images used in the comparison process.

3.3. Algorithms for assessing similarities and searching for images

According to formulas of equations (1), the following algorithms have been initiated and implemented in these formulas to assess the similarities and search for images.

3.4. Image similarity algorithm based on analyzing the diversity of color intensity

After loading up the image into the program and calculating the height and width of the image, the pixel values are stored in a matrix so that equation formula (2) is used to calculate the color intensity variation as follows:

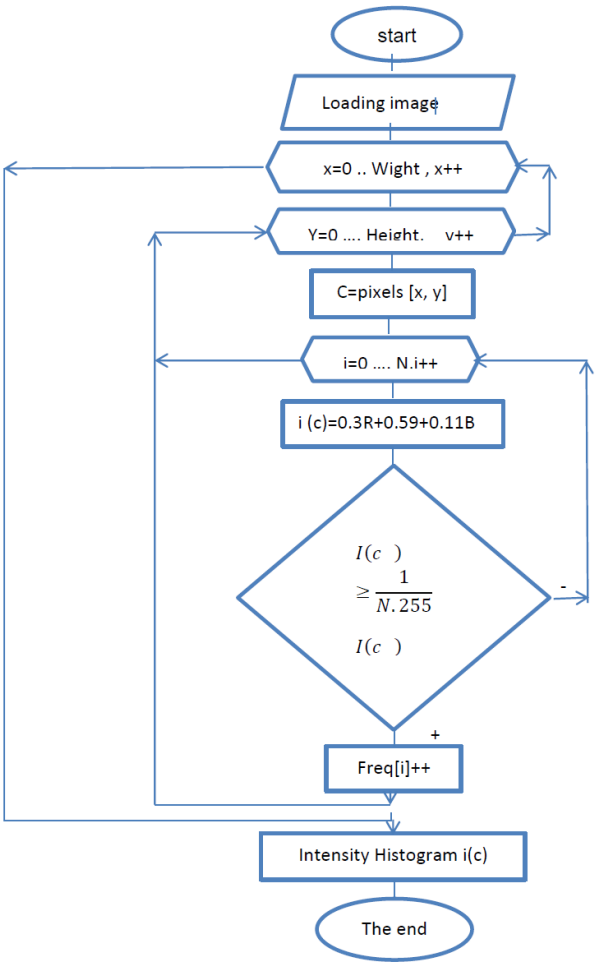

Figure 2. The method of calculating the diversity of color intensity

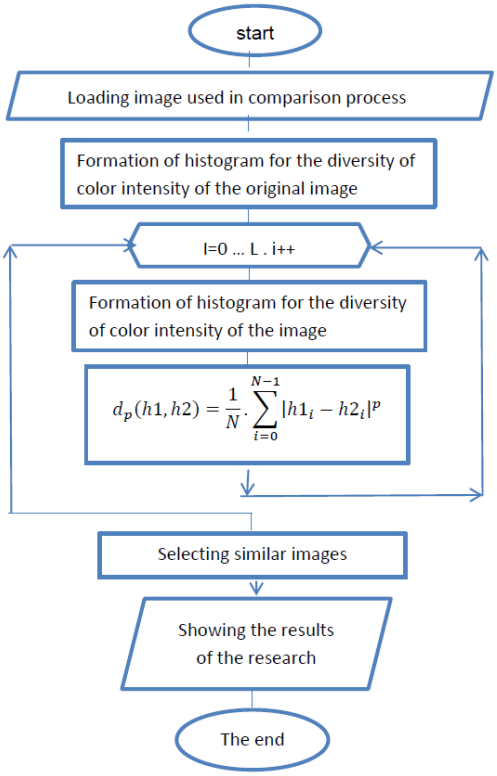

Figure 3. Algorithm of image similarities based on calculating the diversity of color intensity

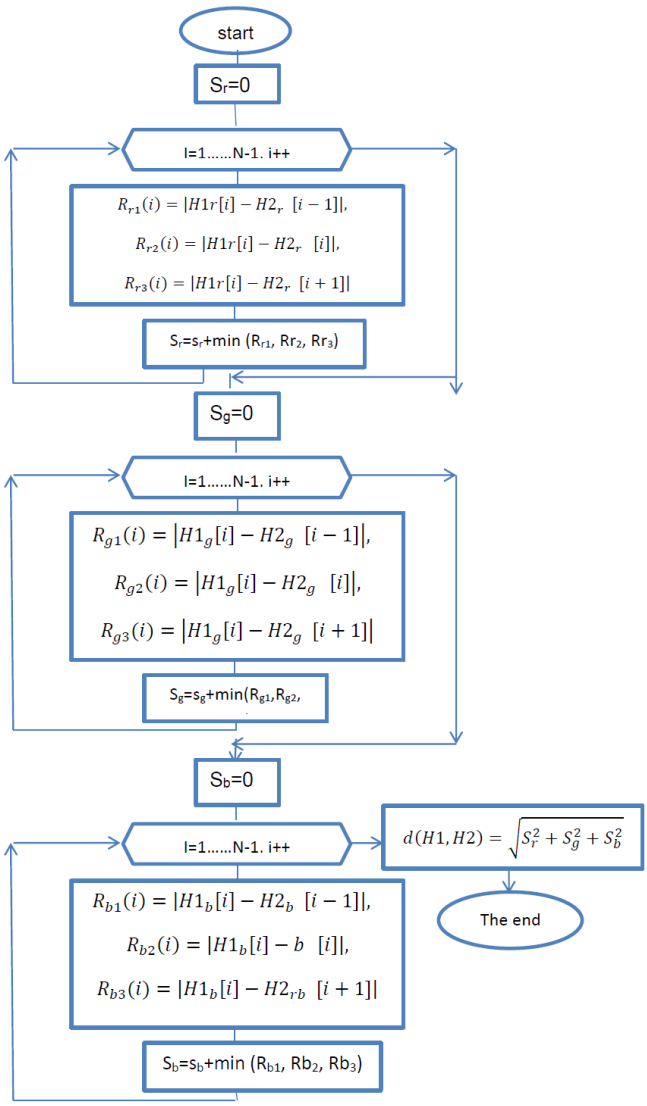

Figure 4. Image similarity algorithm based on the histogram

3.4. Experimental study of the effectiveness of image similarity assessment algorithms

Comparative programs are designed and programmed to use the **C++ Builder** language, the following figures show images of the programs used and the results

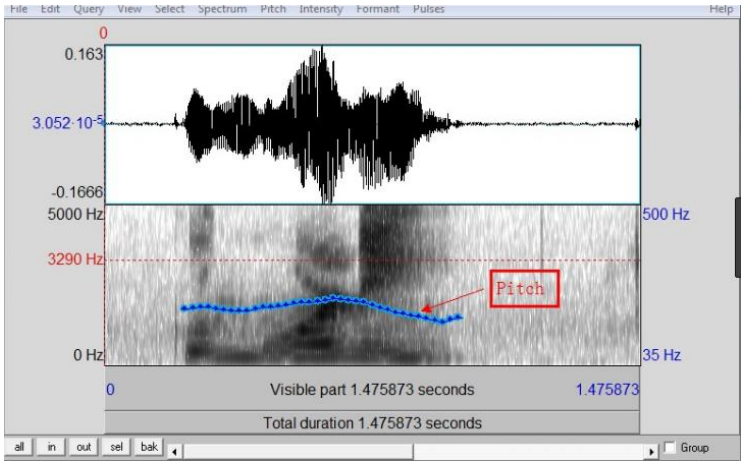

Figure 6. The program interface and comparison of an image with a set of listed images using color intensity variation analysis

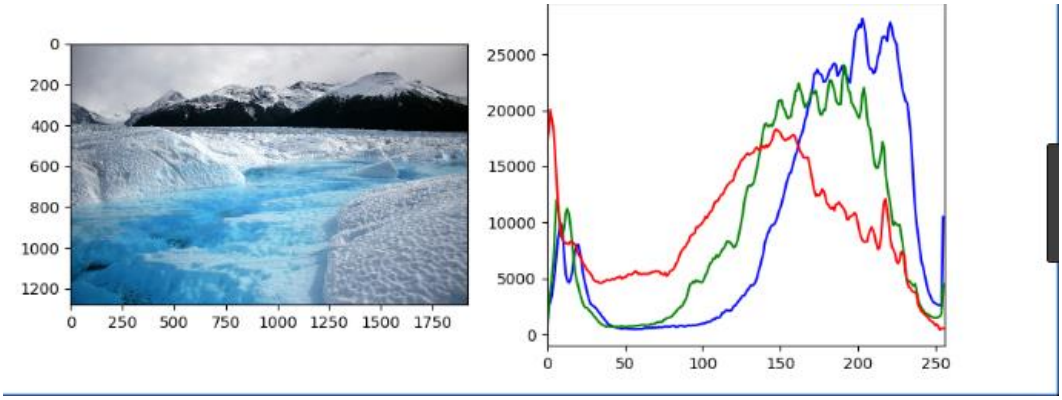

Figure 6. The program interface and compare an image with a set of listed images using histogram analysis

Whereby, the numbers used in the previous figure represent the following:
1 – The original image.
2 – An image of the histogram of the original image.
3 – A list of the images compared with the original image.
4 – Properties of the original image.
5 – Histogram of similarities and differences between images in the list.
6 – Finding four images similar to the original image.

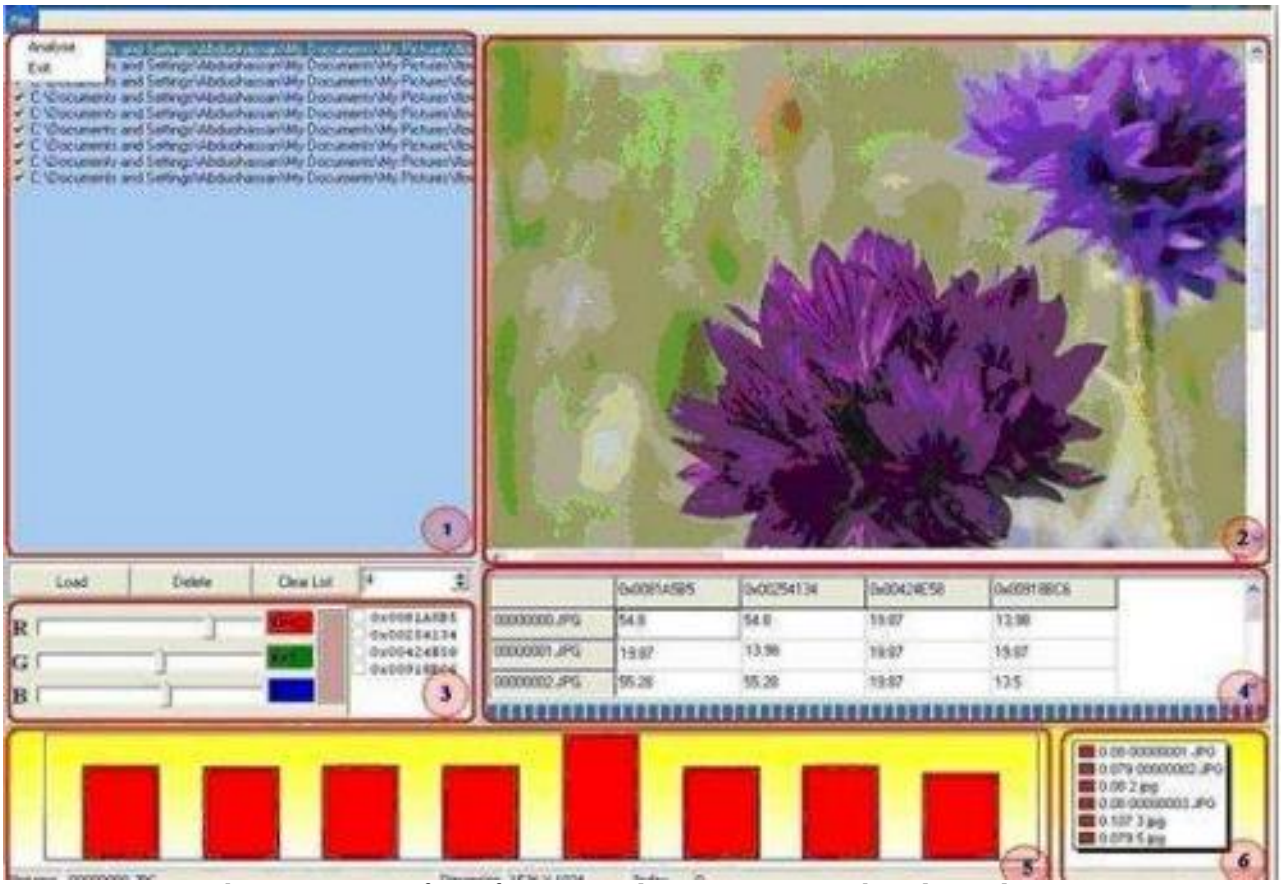

Figure 7. The program interface of image similarities assessment based on color intensity variation analysis

Whereby, the numbers used in the previous figure represent the following:
1 - Image analysis list.
2 – The selected image.
3 – Selecting some colors of the image.
4 – A result of color ratio analysis (the percentage of colors selected in each image of the list).

5 - The result of image comparison (The diagram).
6 - The resemblance to the original image.

In the three designed programs, a set of images was used. Figure (9) represents the original image, represents the set of images involved in the analysis process, which will be compared with the original image, which is available in the following sizes: **10x10, 50x50, 100x100, 200x200, 500x500, 1024x1024, 2048x2048.**

The images involved in the analysis process were used as a single set with different sizes. Each set of different sizes in each process was tested with a set of images ranging from one to five pictures. Besides, we could measure the time needed to search for similar images in each process. The calculation of time starts from the moment the original image histogram is created [12 ,11].

Figure 8. The involved images in the analysis process

The study was carried out using a computer with the following specifications:

4. Results

4.1. The results of the research taken time to assess the similarities of images based on color intensity variation analysis

Table 1 represents the results of the experimental studies of the research spent time (a bit of second) using different numbers and sizes of images to assess color intensity variation analysis according to equations 1,2, and 3.

Table 1. The taken time in searching for images using the color intensity variation analysis algorithm

Image size / **Number of**	**10 * 10**	50 * 50	100 * 100	200 * 200	500 * 500	1024 * 1024	2048 * 2048
1	**0.0733**	0.1799	0.2396	0.4596	0.9996	3.1998	6.9996
2	**0.1632**	0.358	0.4396	0.9996	1.9998	4.9998	12.9996

10 * 10	50 * 50	100 * 100	200 * 200	500 * 500	1024 * 1024	2048 * 2048	**Image size** **Number of**
0.2332	0.527	0.8996	1.3996	3.1998	7.9998	18.9996	**3**
0.3232	0.708	1.2096	1.7996	3.996	9.9996	25.9998	**4**
0.4042	**0.8798**	**1.5196**	**2.2496**	**4.99**	**12.9996**	**34.9998**	**5**

4.2. The results of the research taken time to assess the similarities of images based on histogram analysis

Table [2] represents the results of the experimental studies of the research taken the time (a bit of second) using different numbers and sizes of images and histogram analysis according to equations (4).

Table 2.The taken time in searching for images using the color histogram analysis algorithm

10 * 10	50 * 50	100 * 100	200 * 200	500 * 500	1024 * 1024	2048 * 2048	**Image size** number of involved images
0.3998	0.9996	1.3998	1.6998	1.9998	3.1998	12.996	**1**
0.5998	1.0533	1.9998	3.1996	3.3998	6.9996	24.996	**2**
0.9996	2.6998	3.6996	3.9996	5.1798	9.9996	36.996	**3**
1.9998	3.6998	4.6998	4.9998	7.9998	12.9996	49.998	**4**
2.6666	**5.9998**	**6.6996**	**6.9996**	**7.9998**	**15.9996**	**61.998**	**5**

4.3. The results of the research spent the time to assess the similarities of images based on color ratio analysis

Table 3. represents the results of the experimental studies of methods of searching for images using the image similarities algorithm based on color ratio analysis and it also took time

10*10	50*50	100*100	200*200	500*500	1024*1024	2048*2048	**Image size number of involved images**
0.998	1.7	4.098	5.0196	6.9996	9.996	26.0196	**1**
1.54	2.98	4.9998	6.9998	7.9998	15.3996	45.9996	**2**
2.23	3.99	6.09	7.598	9.1998	21.3996	66.9996	**3**
3.128	4.9998	7.088	8.7998	9.9798	26.0196	87.9798	**4**
4.09	5.98	7.98	9.498	10.9998	30.9996	109.9998	**5**

The results were represented in the diagrams; the Coordinate axes as follows:

The ax "X": represents different sizes of images by pixel.

The ax "y": represent the research taken time.

These diagrams show the results of the research spent time by using three different algorithms in two cases as follows:
*The first case: comparing the original image with the involved images as one set with different sizes; that's when numbers of images equal one image for each size.
*The second case: comparing the original image with the involved images as one set with different sizes; that's when the numbers of images equal five images for each size [13].

Table 4. The research spent time for a set of images when their numbers equal one image for each size

2048 *2048	1024 *1024	500 *500	200 *200	100 *100	50 * 50	10 * 10	algorithm	
6.9996	3.1998	0.9996	0.4596	0.2396	0.1799	0.0733	Color intensity diversity analysis	**1**
12.996	3.1998	1.9998	1.6998	1.3998	0.9996	0.3998	Histogram analysis	**2**
26.0196	**9.996**	**6.9996**	**5.0196**	**4.098**	**1.7**	**0.998**	**Color ratio analysis**	**3**

Table 5. The research took the time of a set of images when their numbers equal five images for each size

2048x 2048	1024x1024	500x 500	200x200	100x100	50x 50	10x10	algorithm	
34.9998	12.9996	4,99	2.2496	1.5196	0.8798	0.4042	Color intensity diversity analysis	**1**
61.998	15.9996	7.9998	6.9996	6.6996	5.9998	2.6666	Histogram analysis =	**2**
109.9998	**30.9996**	**10.9998**	**9.498**	**7.98**	**5.98**	**4.09**	**Color ratio analysis**	**3**

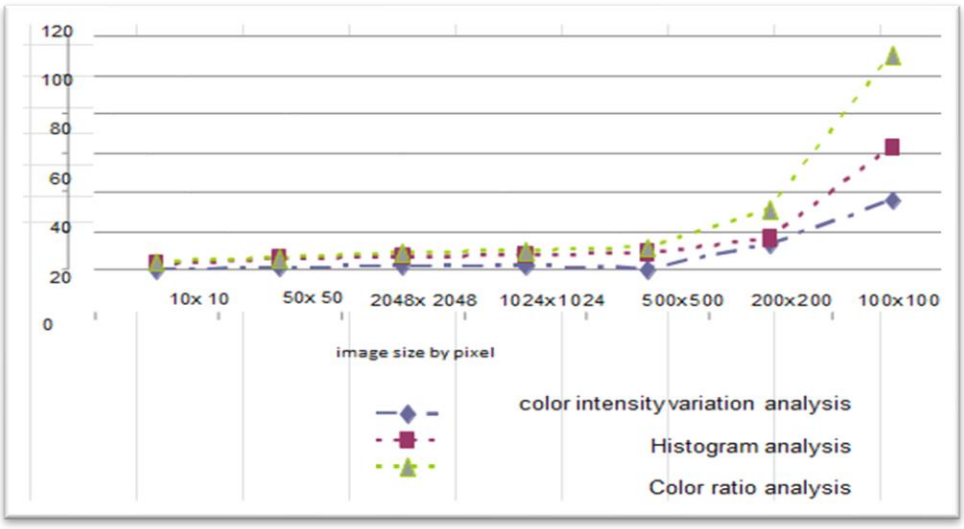

Figure 9. The diagram of comparing different algorithms of searching for images
(Numbers of images in the list: 5 images)

From the diagrams, it's clear that the researchers spent time is related to the sizes and numbers of involved images. It's also obvious that searching for images on basis of color intensity variation algorithm is the least one in the research spent time, but it is almost the same time for small sizes of images. We also noticed that the researchers spent time is related to the method of processing of the involved images in the research. It means that in case of involved images processing with the original image, the spent time increases; and in case of the preprocessing of involved images and storing the results in the database, the spent time, in comparing with the original image, decreases [14]. Besides the evaluation of the spent time in the search of the three algorithms, the effectiveness of these algorithms was evaluated (the best algorithm in terms of results and performance). For this purpose, 75 identical images and 75 completely different images were provided (landscape of the trees at different times of the day and images of the sky at daytime), for the three programs the results were as follows.

Table 6. The results of searching for images of different algorithms

Percentage of faults in images (75 different images)	Percentage of images (75 identical images)	Algorithm
4,6	64	Searching for images on basis of color intensity variation
4	80	Searching for images on basis of Histogram analysis
16	45,3	Searching for images on basis of color ratio analysis

It's noted that the best results of searching for identical images are the result of using histogram analysis of images, but the absolute result of 100% in the search process was not reached due to some reasons that cause the difference in color brightness of some images of the scene at different periods (day and night)) These are the same reasons that cause the low performance of the algorithm to search for identical images based on the partial analysis of the color ratio, but these can be eliminated by taking multiple images for the same scene at different periods [15].

5. Conclusions

This research provided to develop some image search algorithms based on content analysis and reduces the taken time in the search and similarities between the image of the search and the listed images that it will be compared with, where the results can be summarized as follows:

1 - To search for images and analyze their contents, the best way is to take advantage of the statistical characteristics of the image, which allows for the creation of algorithms to analyze and evaluate the similarities of images. In the study, the comparison and histogram methods were used on basis of the percentage of colors in the image.

2- The development of an image similarities assessment algorithm based on histogram analysis and comparison on basis of colors ratio in the image which leads to fast performance of searching for images.

3-. One of the future works of this study is the possibility of adding and comparing other algorithms to select the best of these algorithms.

References

[1] M. A. Ameen, "Content-Based Image Search Engine," in *The 17the national conference on Computer Science, King Abdul Aziz University*, 2004: Citeseer.

[2] W. Peizhuang, "Pattern recognition with fuzzy objective function algorithms (James C. Bezdek)," *SIAM Review*, vol. 25, no. 3, p. 442, 1983.

[3] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Color-and texture-based image segmentation using em and its application to image querying and classification," in *Proc. ICCV*, 1998, pp. 675-682.

[4] J. C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*. Springer Science & Business Media, 2013.

[5] H. A. Mardan and S. K. Ahmed, "Using AI in wireless communication system for resource management and optimisation," *Periodicals of Engineering and Natural Sciences (PEN)*, vol. 8, no. 4, pp. 2068-2074, 2020.
[6] E. R. Dougherty, *Digital image processing methods*. CRC Press, 2020.
[7] N. O'Mahony *et al.*, "Deep learning vs. traditional computer vision," in *Science and Information Conference*, 2019: Springer, pp. 128-144.
[8] L. E. George and S. K. Ahmad, "Hiding image in image using iterated function system (IFS)," in *Proceedings of the European conference of systems, and European conference of circuits technology and devices, and European conference of communications, and European conference on Computer science*, 2010: World Scientific and Engineering Academy and Society (WSEAS), pp. 68-74.
[9] M. Veluchamy and B. J. O. Subramani, "Image contrast and color enhancement using adaptive gamma correction and histogram equalization," vol. 183, pp. 329-337, 2019.
[10] J. R. Smith and S.-F. Chang, "Automated binary texture feature sets for image retrieval," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, 1996, vol. 4: IEEE, pp. 2239-2242.
[11] J. Z. Wang, G. Wiederhold, O. Firschein, and S. X. Wei, "Content-based image indexing and searching using Daubechies' wavelets," *International Journal on Digital Libraries*, vol. 1, no. 4, pp. 311-328, 1998.
[12] F. H. M. Al-Kadei, "Two-level hiding an encrypted image," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 18, no. 2, pp. 961-969, 2020.
[13] T. Wang *et al.*, "Deep learning-based image quality improvement for low-dose computed tomography simulation in radiation therapy," vol. 6, no. 4, p. 043504, 2019.
[14] J. Lee, K.-C. Lee, S. Cho, and S.-H. J. S. Sim, "Computer vision-based structural displacement measurement robust to light-induced image degradation for in-service bridges," vol. 17, no. 10, p. 2317, 2017.
[15] U. A. Nnolim, "Smoothing and enhancement algorithms for underwater images based on partial differential equations," *Journal of Electronic Imaging*, vol. 26, no. 2, p. 023009, 2017.