Content Based Image Clustering Technique Using Statistical Features and Genetic Algorithm

Bushra Kh. AlSaidi
College of Administration and Economics, University of Baghdad, Baghdad, Iraq
dr.bushraalsaidi14@gmail.com

Baydaa Jaffer Al-Khafaji
Computer Science Department, College of Education for Pure Science/Ibn Al-Haitham, University of Baghdad, Baghdad, Iraq
bjkh68@yahoo.com

Suad Abed Al Wahab
Ministry of Education, Baghdad, Iraq
suaad71computer@yahoo.com

Abstract—Text based-image clustering (TBIC) is an insufficient approach for clustering related web images. It is a challenging task to abstract the visual features of images with the support of textual information in a database. In content-based image clustering (CBIC), image data are clustered on the foundation of specific features like texture, colors, boundaries, shapes. In this paper, an effective CBIC) technique is presented, which uses texture and statistical features of the images. The statistical features or moments of colors (mean, skewness, standard deviation, kurtosis, and variance) are extracted from the images. These features are collected in a one dimension array, and then genetic algorithm (GA) is applied for image clustering. The extraction of features gave a high distinguishability and helped GA reach the solution more accurately and faster.

Keywords-content based image clustering; statistical feature; genetic algorithm

I. INTRODUCTION

Content-based image retrieval (CBIR) has become a broad field of research with applications in several fields like scientific research, design, industry, medicine and internet [1]. Image retrieval started by retrieving images based on texts associated with the images. There are several issues connected with retrieving images based on text such as manual keyword annotation, variations in perceptions, explanations and other issues. As a result, researchers applied CBIR where images are retrieved based on auto-derived features [2]. Texture and color together with other features like shape, edge and surface play important roles in CBIC for their ability of providing significant characteristics for automatic computer based image recognition. Texture has appeared in many CBICs, while color has been exploited to improve the clustering performance [3]. Content based means that the retrieving system will analyze real content instead of metadata like image description, keywords or tags. Content refers to the color information and textures that can be detected from the images. Many search engines used image retrieval systems entirely dependent on metadata, leading to many erroneous results while the manual placing/editing of image descriptions in big databases is costly and time consuming. For this reason, this paper provides an effective algorithm for CBIC using color and texture properties to eliminate the above problems. However, CBIR systems face the problem of feature extraction complexity or the high dimensions of feature vectors. These problems made image retrieval results poor in accuracy and response time [4].

The aim of this research is to provide an efficient CBIC technique with a one dimensional vector for color and texture properties. Images are clustered by using statistical features and texture features of the gray image density rather than a color image because gray density takes less memory space and holds significant image properties. So the proposed technique does not require complex computations for feature extraction.

II. IMAGE CLUSTERING

Clustering is the process of assembling data objects in a way that all objects in the same group are similar and differentiating from the objects in other groups. Numerous cluster analysis algorithms have been introduced. Partition is one of the methods of effective clustering, where the database is divided into groups in an iterative process. Clustering can be used to reduce the number of comparisons and improve recovery time when images are searched in a database [5]. Image clustering is considered to be a demanding field. One of the most common problems faced by the image clustering algorithms is that there are many types of image representation, and this representation has different distinguish abilities, so it has different effects on clustering methods [6]. The most significant task of any CBIC algorithm is feature extraction. When feature space rather than pixels is used, the similarity between stored images can be measured by these features [7]. Feature extraction from the visual content of an image plays a significant role in many image processing and pattern recognition applications [8]. One of the significant features is color which plays an important role in CBIC because of its strength and independence of image's orientation and size [3].

A. Texture Feature Extraction

Texture analysis plays an important role in many applications based on color brightness or local spatial diversity of density. Texture analysis is commonly used in image classification, segmentation and retrieval. Furthermore, texture feature extraction is the main component of texture analysis. The approaches of texture feature extraction are fractal approaches, statistical methods, and wavelet transform. Statistical methods are simple, easy to implement, adaptive,

Corresponding author: Bushra Kh. AlSaidi

and robust for texture analysis [9]. In this paper entropy filtering and range filtering are utilized and the local standard deviation of the input images is calculated.

B. Color Feature Vector Extraction

In CBIC the most commonly used visual feature is color, which is simple to represent and robust. Several studies have been proposed to study color perception and color spaces [10]. Color moments are utilized to distinguish images and to measure their similarity. The base of color moments is that the color probability distribution in an image can be interpreted by color distribution. If the color in the images has a specific probability distribution, this distribution can be considered as a feature determining that image [7]. Before extracting color features, the RGB image is converted into a grayscale image. The grayscale image consists of 8-bit pixels, with values ranging from 0 to 255, unlike the color image which sets a value of 8 bits per red, green and blue color. The benefit from using a grayscale image is that it takes less memory and simplifies the amount of information. After extracting the texture feature, the second step is to extract the value of the color moments (mean, skewness, standard deviation, kurtosis, and variance) from the images. These moments are calculated for each image and are stored in one dimension array.

III. GENETIC ALGORITHM

GA is an effective, powerful and integrated search and adaptation technology. Its basic components are:

- Chromosomes are a problem representation strategy. The candidate solutions for the optimization problem are encoded in a population of chromosomes.
- A fitness function is utilized to evaluate each chromosome. In every generation, the individuals are evaluated by the fitness function, and the achievement of reproduction varies with fitness. The important components of the suggested approach are the definition of the fitness function for the problem and its utilization in GA.
- Selection/reproduction: The selection process copies chromosomes into a new population, according to their fitness values. Tournament selection method was used in the proposed algorithm because of its low complexity. The basic notion of the tournament technique is the random choice of a number of individuals from the current population and the copying of the best fitted individuals from them into an intermediate population.
- Elitism, crossover, and mutation, which are called genetic operators. The central aim of crossover is to randomly exchange information between parent chromosomes by exchanging parts of their genetic information. There are many crossover approaches. In this paper two-point crossover was used. In mutation, random alterations in the chromosome structure are done representing genetic diversity into the population. Elitism replaces the worst individuals of the current population with the best individuals of the prior population.

IV. THE PROPOSED CONTENT-BASED GA

The input of the suggested algorithm is a set of images. The output consists of a set of clusters and a set of images in each cluster. There is a group of parameters that should be set by the user. These parameters are the number of chromosomes that represent population size (pop-Size), maximum generations (max-Gen), mutation rate (P-m) and crossover rate (P-x). The steps of the proposed algorithm are:

Algorithm: image clustering.
Input: images to be clustered.
Tune: set of parameters pop-Size, max-Gen, P-m and P-x.
Outputs: the set of images in each cluster.
Begin:
Read the feature vector.
Initiate the population.
For i=1 to pop-size
Create individual.
Calculate fitness function for the initial population.
Current population = initial population.
While generation number less than max-Gen do
Do elitism.
Select the parents of the next generation by Tournament algorithm.
Generate the offsprings by Crossover.
Mutate offsprings.
Calculate fitness function for the new population (offsrings).
Do Elitism.
Current population = new population.
Output: The current population, its fitness and the set of images covered by each individual in the population.

V. POPULATION REPRESENTATION AND INITIALIZATION

GA works on a number of candidate solutions, called population, and contains multiple coding for the parameter specified at the same time [11]. To resolve image clustering, the population was started using 50 chromosomes of equal length in two different ways:

A. First Method to Iinitiate Population

This method to initiate the population was proposed to avoid the bias of the algorithm for some images. The steps of the method, applied to each individual in the population are:

- Generate a random image number. This number must be smaller than the total number of images in the data set.
- Select the mean for the image with the random number generated in the previous step as the mean value of the candidate cluster.
- Generate another random image number.
- Select the standard deviation of this image as the standard deviation value of the candidate cluster.
- Generate another random image number.
- Chose the skewness of this image as the skewness value of the candidate cluster.
- Generate another random image number.

- Chose the kurtosis of this image as the kurtosis value of the candidate cluster.
- Generate another random image number.
- Chose the variance of this image as the variance value of the candidate cluster.
- Repeat the above steps until the pop-size is reached.

After these steps are finished the population has been initialized. Table I shows a snapshot of the initial population. Each row of the table represents a candidate cluster that will be evaluated by the fitness function. The candidate cluster has 5 statistical features of the image (mean, slandered deviation, skewness, kurtosis, and variance). Images that achieve a minimum Manhattan distance with the candidate cluster will be allocated to this cluster. The fitness function of this individual or candidate cluster will be the number of images that have a minimum Manhattan distance from it.

TABLE I. SNAPSHOT OF THE INITIAL POPULATION

Mean	Std.	Skewness	Kurtosis	Var.
6.5398	11.622	13.0082	135.079	2.9488
12.6480	19.6358	5.2894	385.5637	1.7636
5.2544	6.5685	14.2042	43.1454	2.7611
7.4712	7.0551	7.9771	49.7747	1.6558
7.3502	10.7057	11.9982	114.6126	2.7449
4.7325	6.1222	8.9581	37.4817	2.1233
4.5854	5.5170	9.0047	30.4373	2.1259
3.4846	3.9387	10.2564	15.5134	2.2339
7.7545	7.3939	7.8977	54.6694	1.9488
63.8923	5.398	20.3111	29.1445	3.2706
6.0577	6.6078	9.9426	43.6630	2.2871
13.6771	9.5359	5.3893	90.9343	1.2311
12.4672	11.6368	11.0113	135.4144	2.3422
8.9082	8.0444	4.8876	64.7128	1.2228

B. Second Method to Initiate Population

This method to initiate the population is proposed in order to exploit some of the prior knowledge contained in the set of images through which the speed of converging can be increased. The steps of the method, applied to each individual in the population are:

- Chose a random image number.
- Chose the feature vector of the image with the number generated in the first step from the feature vectors resulted from the feature extraction process.
- Repeat the above steps until the pop-size is reached.

After these steps are completed, the population has been initialized (see Table II for a snapshot).

VI. EXPERIMENTAL RESULTS AND DISCUSSION

The experimental part of the suggested method was conducted on a set of images form the internet, a total of 1000 images were obtained. The data collection includes images of nature divided into 8 groups: birds, trees, forests, birds in the sky, coast, birds above the sea, and trees on the coast. All database images were resized to 180x180 resolution in the pre-processing step. Then, images were converted to grayscale. After these steps, the feature extraction phase began then the GA has been applied. Figure 1 shows examples of clusters presented by the proposed algorithm.

TABLE II. SNAPSHOT OF THE INITIAL POPULATION

Mean	Std.	Skewness	Kurtosis	Var.
8.474	10.630	8.2814	113.009	1.9745
10.037	9.137	8.2324	83.4892	1.7096
2.8662	5.238	32.386	27.4369	4.5214
2.592	6.613	57.220	43.7347	6.4893
17.473	12.327	6.4979	151.972	1.6579
12.557	10.476	5.8171	109.755	1.4637
9.594	8.814	16.064	77.6918	3.0411
13.699	8.863	10.197	78.5613	1.8909
12.279	9.216	6.1317	84.9521	1.4843
3.689	6.5839	11.271	43.3473	2.7077
6.997	6.743	15.526	45.4687	2.6535
9.883	10.611	8.1856	112.6048	2.0876
6.388	8.034	14.871	64.5583	2.8186
6.968	9.50	7.7772	90.2652	2.0691
6.974	4.8671	9.4987	23.6883	1.8907
13.943	14.072	4.8165	198.0311	1.3296

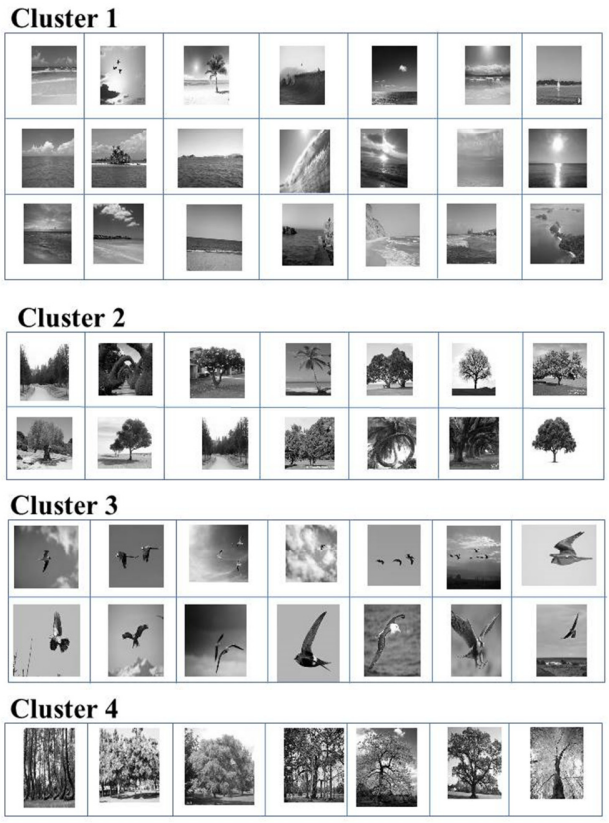

Fig. 1. Examples of clusters presented by the proposed algorithm.

The experiments first studied the parameter sensitivity. One important parameter is population size. The study of the relation between the number of clusters presented by the proposed method and the size of the population showed that increasing population size leads to an increase in the number of clusters and a reduction in the number of images assigned to each cluster. Therefore, determining the best size of the

population depends on the size of the image data set and the number of clusters detected in the data set. Table III shows the effect of population size on the number of clusters and the number of images in each cluster.

TABLE III. EFFECT OF POPULATION SIZE ON THE NUMBER OF CLUSTERS AND THE NUMBER OF IMAGES IN EACH CLUSTER.

	Experiment No.				
	1	2	3	4	5
No. of clusters	Images in cluster	Images in cluster	Images in cluster	Images in cluster	Images in cluster
1	4	17	8	1	14
2	2	5	20	28	54
3	18	3	5	4	3
4	6	9	27	11	28
5	4	22	25	38	31
6	6	2	12	9	
7	2	2	16	19	
8	1	6	9	9	
9	8	2	4	11	
10	0	0	4	1	
11	4	12	8	28	
12	16	6	20	4	
13	1	7	5	11	
14	7	1	27	38	
15	1	4			
16	4	19			
17	6	5			
18	3				
19	1				
20	4				
21	4				
22	8				
23	2				
24	6				
25	4				
26	1				
27	7				
28	4				

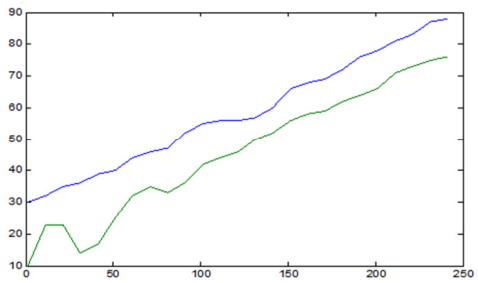

Fig. 2. Performance evaluation of the proposed algorithm for first and second initialization method

The second idea, which was studied through the experiments, is the effect of the method of population initialization on the performance of the algorithm. The performance of the algorithm was measured using precision and recall. The results showed that the second method, which exploits prior knowledge, improved the performance of the algorithm and reduced image clustering time. Figure 2 shows the comparison of the performance of the algorithm using the above mentioned two different methods of population initialization. Feature extraction method and similarity measurement are two main steps in image clustering [2]. In this paper, we tried several methods for the fitness function and Manhattan distance showed the best results.

VII. CONCLUSION

In the subject of image retrieval and image clustering, image feature extraction is a great challenge for the researcher. This paper presented a GA-based image clustering method using texture and statistical features. Experimental results showed high distinguishability of these features. The main advantage of the proposed algorithm is that it does not need the class label to be provided, the image set has not need to be labeled first. Also, the proposed extraction method reduces the feature vector dimension and improves the efficiency of the clustering method.

REFERENCES

[1] N. Ghosh, S. Agrawal, M. Motwani, "A Survey of Feature Extraction for Content-Based Image Retrieval System", Lecture Notes in Networks and Systems, Vol. 34, pp. 305-313, Springer, 2018

[2] J. Wang, L. Wang, X. Liu, Y. Ren, Y. Yuan, "Color-Based Image Retrieval Using Proximity Space Theory", Algorithms, Vol. 11, No. 8, ArticleID 115, 2018

[3] M. Pham, G. Mercier, L. Bombru, "Color Texture Image Retrieval Based on Local Extrema Features and Riemannian Distance", Journal of Imaging, Vol. 3, No. 4, ArticleID 43,2017

[4] K. Kumar, Zain-ul-Abidin, J. P. Li, R. A. Shaikh, "Content Based Image Retrieval Using Gray Scale Weighted Average Method", International Journal of Advanced Computer Science and Applications, Vol. 7, No. 1, pp. 1-6, 2016

[5] E. M. F. El Houby, "Medical Images Retrieval using Clustering Technique", International Journal on Recent and Innovation Trends in Computing and Communication, Vol. 3, No. 5, pp. 3134-3141, 2015

[6] T. Peng, H. Gao, "A Cluster Priority Level Decision Method for Image Features", International Journal of Database Theory and Application, Vol. 9, No. 2, pp. 171-182, 2016

[7] A. Malakar, J. Mukherjee, "Image Clustering using Color Moments, Histogram, Edge and K-means Clustering", International Journal of Science and Research, Vol. 2, No. 1, pp. 532-537, 2013

[8] A. Obulesu, V. V. Kumar, L. Sumalatha, "Content based Image Retrieval Using Multi Motif Co-Occurrence Matrix", International Journal of Image, Graphics and Signal Processing, Vol. 4, pp. 59-72, 2018

[9] X. Zhang, J. Cui, W. Wang, C. Lin,"A Study for Texture Feature Extraction of High-Resolution Satellite Images Based on a Direction Measure and Gray Level Co-Occurrence Matrix Fusion Algorithm", Sensors, Vol. 17, No. 7, ArticleID 1474, 2017

[10] S. Soman, M. Ghorpade, V. Sonone, S. Chavan, "Content Based Image Retrieval Using Advanced Color and Texture Features", International Conference in Computational Intelligence, Guangzhou, China, November 17-18, 2012

[11] B. K. AlSaidi, "Automatic Approach for Word Sense Disambiguation Using Genetic Algorithms", International Journal of Advanced Computer Science and Applications, Vol. 7, No. 1, pp. 41-44, 2016