Fast and Efficient K-means based Algorithm to Content-based Image Clustering

Basel Hafiz
Faculty of Computers and InformaticsSuez Canal University, Egypt

Mohamad Mousa
Faculty of Computers and Informatics Suez Canal University, Egypt

Mohamad Waheed
Faculty of Computers and Informatics Suez Canal University, Egypt

ABSTRACT

Some of the present approaches compare the user's query image against all of the database images; as a result, the computational complexity and search space will boost, respectively. The fundamental purpose of the research presented in the present paper is to evolve a general purpose clustering method that can efficiently and effectively handle large capacity image databases. It can be fluently embedded into distinct CBIR systems. In this paper, we developed a novel content-based image clustering technique based on an effective k-means based algorithm. The co-occurrence matrix features and color moments are utilized to evolve an effective and innovative image clustering framework. The texture and color features are integrated to enhancing results obtained using individual descriptors. The introduced k-means based clustering algorithm has been proposed as a preprocessing procedure to accelerate image retrieval and to enhance image retrieval accuracy. The experimental outcomes based on COREL images have been investigated and indicated considerable refinement in terms of quality of image clustering, retrieval accuracy, and speed compared against the conventional k-means method.

Keywords

Content-based image clustering, Feature extraction, K-means clustering

1. INTRODUCTION

Regarding the swift development of multimedia technology and information technology, huge amounts of the multimedia data are produced and conserved each day. The numerous video, audio, photo, and animation data have deduced in computer vision and image processing becoming very widespread research fields. All these video/image data required being stored into a considerable video/image database for management and conservation; thus the database's capacity becomes larger. Deciding how to effectively and quickly locate images requested is of significant priority. The conventional image searching methods are text-based retrieval, which explores images depending on the keywords assigned by user. There are many situations in which the user's query cannot be described through keywords easily. To resolve this problem, the CBIR was considered to search images depending on the given user's image. Large quantity of CBIR system architectures, methods, and techniques for descriptor extraction and image matching were developed and investigated [1].

In the CBIR system, all images should be indexed using their visual signification as features. These involve the image characteristics like shape, texture, color, and spatial relationships. The image features are conserved within a particular database for the future employment. When an input image is assigned, its features are extracted for matching the restored features within the aforementioned database using a specific pre-established matching algorithm. Thus, a list of comparable results to the assigned image could be returned [1]. Many retrieval systems based on image features have been developed. Despite the wide-ranging applications of textures [5, 7, 9, 11-14, 16, 17], shapes [2, 14, 20-22], colors [2-16], and spatial relationships [3, 6, 18, 19], the image retrieval and discrimination outcomes are unsatisfactory yet.

With a view to enhance the retrieval efficiency and minimize the execution time, many researchers utilize the clustering algorithms to manage the database images before retrieving. The image clustering process is grouping images depending on their likeness. Intuitively, the members of a specific cluster are more similar than the members of the other clusters. Using clustering, the retrieval procedure doesn't require to examine database images one after another to compare with the given query image. The CBIR system just requires matching the input image with the clusters centroids, and then it restores all database images that belong to matched image cluster [23].

To execute image clustering, it is important to pass several phases. The first phase is selecting suitable representation space. The second is matching every image to chosen representation space through suitable similarity measure (distance measure). The last step is performing image clustering either within a supervised manner or within an unsupervised manner. The images within the supervised manner are clustered with the human involvement. Furthermore, the clustering in the unsupervised process depends on the measure of similarity among the diverse cluster centroids and the images. The unsupervised techniques are qualified to perform clustering in completely automatic manner. Thus, they provide more adaptable ways for image clustering.

K-means algorithm is the most suitable technique for data clustering, which is used in CBIR systems to minimize the features compared against the query image when it is compared only against the centroid of clusters [23]. Color features are widely used in researches [4-7, 11, 12, 16, 24-31], as these are widely efficient while the k-means is used. The remainder of the paper is well-organized as follow. Section 2 provides a review of related works in content-based image clustering. The developed content-based image clustering method using an effective k-means based algorithm is introduced in Section 3. Section 4 reports the outcomes of the experiments for evaluating the proposed method. Finally, the paper is summarized and concluded.

2. RELATED WORK

The suggested framework in Mahishwari et al. [24] focuses on texture and color as feature. The Gabor filter and Color moment are employed to extract database image features. The hierarchical clustering approach and k-means algorithm are employed to group image database into several clusters. Ramraj and Narasimhan [25] demonstrated the participation depending on clustering method to group the dataset for CBIR. The introduced method is utilized in historical, normal view, and medical databases.

Shrivestava et al. [26] implemented two clustering methods such as C-means and K-means clustering for CBIR. In these methods, the distance similarity notion is utilized for examination of cluster outcomes in the initial seed selections. After comparing these two clustering methods, K-means performs faster than C-means since it demands high execution complexity. But these clustering methods have a problem in the selection of optimum results.

The introduced CBIR system in Daves et al. [27] retains the semantic concepts of the users in several clusters. The formative semantically clusters assist the retrieval framework to accurately estimate the similarity among dataset images. They also assist the retrieval framework to retrieve matched images within a few clusters and thence minimize the investigation space. This article [28] discusses the comparative approach employed in the color histogram depending on two fundamental methods employed extremely in image retrieval which are; color histogram through k-means and color histogram through GLCM. A group of 9960 experimental images are utilized to evaluate the precision and accuracy of each algorithm. The similarity among the experimental images and the input image is computed using the Euclidean distance. The experimental results demonstrates that the color histogram through K-Means algorithm had high precise and accuracy compared against GLCM. Distinct feature extraction methods together with a contribution-based clustering method are utilized in [29] to retrieve the homogeneous images from dataset. The extraction methods are local binary pattern, RGB color-histogram, and RGB color-histogram together with Cany edge detection. LBP is considered as a fresh rotation-invariant texture measure that is theoretically quite simple but powerful.

In [30], a comparative study on the efficiency of the hierarchical clustering methods application and the classification of imaging context for CBIR; is presented. The purpose of presented study is comparing the obtained outcomes from using several hierarchical clustering methods with various configurations and input parameters using two kinds of comparison approaches. It also aims to demonstrate the performance refinements and the cost boosted through the combination of such clustering methods in CBIR. Younus et al. [31] proposed a new CBIR system which uses the k-means algorithm combined with Particle Swarm Optimization algorithm for clustering the image features. Four image features are introduced for estimating the similarity as follows: two color features: the color moment and color histogram; two texture features: the wavelet moment and co-occurrence matrix.

The precision of image clusters is not adequate in the existing image clustering approaches and furthermore derivative complex calculations will actually damage the overall execution time. With a view to dominate the mentioned in the existing image clustering approaches, a fast and efficient k-means based clustering approach is proposed for preferable retrieval outcomes.

3. PROPOSED METHOD

3.1 Dataset

In the following experiments, a lot of images from Corel dataset [32] have been employed. This dataset consists of massive amount of semantic images of different significations ranging from natural images to animals to sports. Corel photos have been categorized into distinct kinds each of quantity 100 using domain professionals. Many researchers consider that it confronts all the demands to evaluate the image retrieval applications, due to its heterogeneous signification and large size. Concerning our experiments, 1000 images have been collected to form our experimental dataset. This dataset is formed from 10 distinct domains namely Dinosaurs, Buses, Buildings, Foods, Mountains, Africans, Horses, Flowers, Elephants, and Beach. Each image category has 100 images. The image resolution is either 384 × 256 or 256 × 384.

3.2 Feature Extraction

The excellent choice of image descriptors basically influences the performance of the CBIR systems. The present techniques employed to describe gray level and color information involve the color correlograms, co-occurrence matrix, histogram, color moments, and dominant colors [33]. In the present paper, the color and texture features are employed to construct an effective and innovative image clustering framework. As such, co-occurrence matrix features and color moments can characterize diverse image properties. The texture and color features are integrated to enhancing results obtained using individual descriptors. They are ideal choice for integration because they possess quite dissimilar characteristics of the images examined, and therefore are heavily independent on each other.

In this paper, we utilized color moments method [33] because of its minimum computational complexity and its smallest vector dimension. The mean, standard deviation, and skewness color moments features have been demonstrated to be effective and efficient in representing the color distributions within images [33]. To restore the color moments from the image content, we require choosing a specific color space to utilize its properties within the extraction process. In common, the colors are presented in 3-dimensional space. The HSV color-space describes a particular color through its brightness, saturation, and hue values. It is very beneficial for interactive color manipulation and selection.

The mean denotes an average value which presents ideas about the brightness in general. The brighter image holds a great mean whilst the dark image holds a small mean. Mathematically, the definition of mean [33] is presented as follow:

$$\mu_a = \frac{1}{T}\sum_{b=1}^{T} f_{ab}, \tag{1}$$

where fab denotes the value of a-th color component of the pixel b and T denotes pixel number. Also, the variance estimates variation of image pixel and the intensity within an image. Its square root measures the spread of data and the image contrast. It can be identified as the standard deviation. The high contrast photo has a great STD whilst low contrast one holds a minimum STD. The standard deviation and variance can be presented by [33] respectively as follow:

$$\sigma_a = \left(\frac{1}{T}\sum_{b=1}^{T}(f_{ab}-\mu_a)^2\right)^{1/2}, \tag{2}$$

$$\sigma_a^2 = \frac{1}{T}\sum_{b=1}^{T}(f_{ab}-\mu_a)^2, \tag{3}$$

where f_{ab} denotes the value of a-th color component of the pixel b, μ_a denotes the mean of color a, and T denotes the amount of features. The skewness estimates the symmetry over mean and is presented mathematically by [33] as follow:

$$s_a = \left(\frac{1}{T}\sum_{b=1}^{T}(f_{ab} - \mu_a)^3\right)^{1/3} \qquad (4)$$

By computing the three color moments of each channel (H,S,V), the feature vector should be a 9 dimension vector.

The co-occurrence matrices listing display how frequently collections of the pixel values changed occur within an image. As well, it is so massive in extracting helpful image features. The features restored are well-known as Haralick features. It includes homogeneity, energy, correlation, and contrast. According to [33], the co-occurrence matrix Cθ°,d(a, b) is employed to calculate the co-occurrence for the pixels of gray-values a and b as distance d, which is polar-coordinates together with orientation and discrete length. Practically, θ holds the values 315°, 270°, 225°, 180°, 135°, 90°, 45°, and 0°, respectively. The matrix Cθ°,d(a, b) can be presented as follow:

$$C_{\theta°,d}(a,b) = \#\begin{cases}1, if\ I(x_1,y_1) = a, I(dcos\theta\ + x_1, dsin\theta + y_1) = b \\ 0, otherwise\end{cases}, \qquad (5)$$

where # indicates the amount of elements within the set, a, b = 0.255 amount of potential gray-levels within the image, d denotes the distance from gray-level value b to a of the image, and the dimension of co-occurrence matrix C(a, b) is M×N.

Energy: This measure is well-known as uniformity, homogeneity, and angular second moment, which are the image consistency measurements. This value will be low if the gray-level intensities were near for each other, and it is high when all of the elements are very irregular.

$$Energy = \sum_a\sum_b C^2_{\theta,d}(a,b) \qquad (6)$$

Entropy: This measure is considered as the energy opposite. The entropy value is low when the elements are irregular, while it has the highest peak if they are uniform.

$$Entropy = -\sum_a\sum_b C_{\theta,d}(a,b)\log C_{\theta,d}(a,b) \qquad (7)$$

Contrast: This measure is well-known as inertia and estimates the matrix difference moment and the contrast value is large when the image has large local variation.

$$Contrast = \sum_a\sum_b(a - b)^2\, C_{\theta,d}(a,b) \qquad (8)$$

Correlation: This measure estimates the linear dependency for gray-level value within the matrix. The high or low correlation value guides to no present conclusion concerning the image.

$$Correlation = \frac{\sum_a\sum_b(ab)\, C_{\theta,d}(a,b) - \mu_x\mu_y}{\sigma_x\sigma_y} \qquad (9)$$

Inverse difference moment: This measure is well-known as local homogeneity and it is considered as the contrast opposite. The uncertainty of matrix elements precedes large values of comparable gray-levels subsequent for each other; whist the function value is large.

$$Inverse\ Difference\ Moment = \sum_a\sum_b\frac{C_{\theta,d}(a,b)}{|a-b|^2}, a \neq b \quad (10)$$

where the means μ_y, μ_x and the standard deviations σ_y, σ_x can be presented as follow:

$$\mu_x = \sum_a a\sum_b C_{\theta,d}(a,b) \qquad (11)$$

$$\mu_y = \sum_b b\sum_a C_{\theta,d}(a,b) \qquad (12)$$

$$\sigma_x = \sum_a(a-\mu_x)^2\sum_b C_{\theta,d}(a,b) \qquad (13)$$

$$\sigma_y = \sum_b\left(b-\mu_y\right)^2\sum_a C_{\theta,d}(a,b) \qquad (14)$$

By computing the previous texture features, the texture feature vector should be a 5 dimension vector and the total feature vector should be a 14 dimension vector.

3.3 Efficient K-means Based Algorithm

One of the widely significant factors to efficient CBIR is the time. Many traditional CBIR systems compare every image of database images against the user's query image in order to retrieve the top relevant images. This process is quite computationally ineffective when the image database is large. Consequently, image categorization or clustering has been employed as a preprocessing procedure to accelerate image retrieval and to enhance image retrieval accuracy. It aims to minimize the features compared against the query image when it is compared only against the centroid of clusters.

3.3.1 Traditional K-means Algorithm

K-means clustering algorithm is a simple unsupervised learning algorithm. It is exceedingly used in clustering large datasets. This clustering algorithm attempts to partition *n* given data objects into a set of diverse clusters whereas every data object belongs to a specific cluster that has the nearest mean. K-means algorithm includes two separate stages. In the assignment stage, the algorithm chooses *k* centers at random, whereas the value of *k* is constant in advance. Then, it attempts to assign every data object or observation to the closest center. The Euclidean distance is commonly considered to compute the distance between the cluster centroids or centers and each observation. When all of the observations are assigned to the given *k* clusters, the primary stage is accomplished and the early clustering is done. In the next stage, the algorithm recalculates the means of the prior created clusters to be the centers of the new clusters. This procedure continues iteratively until the convergence of the criterion function. The criterion function is presented as follow:

$$E = \sum_{i=1}^{k}\sum_{d\in Ci}|d - x_i|^2 \qquad (15)$$

where *E* denotes the sum of the squared error (SSE) for all data objects, *d* denotes a data object or observation, and x_i denotes the average of the cluster C_i. The SSE tries to make the created clusters as detached as possible. The procedure of k-means clustering algorithm is given as follow:

Algorithm 1: Traditional k-means algorithm

```
Input: A database D containing N objects and a number of required clusters K
Output: K clusters created
Method:
1. Choose k objects at random from D to be the initial cluster centroids
2. While there is change in the cluster centroids
    a. Compute the Euclidean distance of each object x from D to each cluster centroid $c_i$
    b. Assign x for the closest one
    c. Re-calculate the cluster centroid for every cluster as the average of all objects in the cluster
3. End while
```

3.3.2 Proposed Clustering Algorithm

One of the main drawbacks of the conventional k-means is that the outcomes of clustering basically depend on the initial chosen cluster centroids or seeds. K-means++ [34] proposes an efficient seeding procedure to enhance the k-means clustering quality by choosing an initial group of centroid seeds that has superior prospect of being nearer to the ideal solution. The fundamental idea of the k-means++ clustering is to select the group of initial cluster centroids or seeds for the k-means algorithm one after another in a consecutive style, where the present group of selected seeds will bias the determination of the next seed. While the initial seeding of k-means++ demands extra running time, the traditional k-means section converges so quickly after the seeding stage of k-means++ and thus k-means++ actually reduces the running time [34].

But, k-means++ as the traditional k-means requires computing the distance of each data object to all of the k centroids when it performs the iteration every time, which occupies extensive execution time, particularly for large capacity databases. Especially, for the aforementioned shortcomings, this paper presents an enhanced k-means++ clustering method. The major idea of proposed clustering method is to specify two common arrays to possess the distance of each data object to the closest cluster through each iteration and certain labels of the clusters. These retained values can be employed in the next iterations as follows: we compute the distance of the present data object to the fresh cluster centroid, if the calculated distance is equal to or less than its distance with the old centroid, this data objects remains in place that was allocated to in the previous iteration. Thence, we don't need to compute the distance of this object to remaining k-1 cluster centroids, saving the execution time to these centroids. Otherwise, we need to compute the distance of the present object to k cluster centroids, and locate the closest cluster centroid and assign the object to it. Then, the specific label of closest centroid and the computed distance to this centroid are recorded. Because in all of iterations of the proposed method, many data objects still reside in their original clusters; it implies that many groups of the database objects would not be executed, saving the overall time of computing the distance, that way enhancing the effectiveness of the k-means++ method.

The enhanced algorithm demands two arrays (Closest_cluster and Minimum_distance) to hold a few pieces of information in every iteration that is used within the following iterations. The array Closest_cluster is used to hold the specific label of closest centroid, while the array Minimum_distance keeps the distance of a specific database object to its closest centroid. In the present research, an efficient clustering algorithm that is depending on the k-means clustering method is proposed. The proposed method can effectively enhance the clustering speed and quality and minimize the overall time complexity. It can be considered as an addition to k-means++. The procedure of proposed clustering algorithm is given as follow:

Algorithm 2: Proposed clustering algorithm

```
Input: A database D containing N objects and a number of required clusters K
Output: K clusters created

Method:
1. Choose a single object at random uniformly from D to be the first seed s in S
2. While $\|S\| < K$
    a. For each object $d_i \in D$, compute the minimal distance from $d_i$ to the closest $s_j \in S$, $D(d_i , s_j)$
    b. Choose $d_i \in D$ with probability $\frac{D^2(d_i,s_j)}{\sum D^2(d_i,s_j)}$
    c. $S \leftarrow S \cup \{d_i\}$
3. End while
4. Compute the distance from each object $d_i \in D$ to each cluster seed $s_j \in S$, $D(d_i,s_j)$
5. Assign $d_i$ for the closest cluster $s_j$
6. Save label j of cluster $s_j$ in which $d_i$ is, into closest_cluster[i] and save $D(d_i,s_j)$ into minimum_distance[i]

7. Re-calculate the cluster seed for every cluster as the average of all data objects in the cluster
8. Repeat
    a. For each object $d_i \in D$, calculate the distance from $d_i$ to the seed of the current closest cluster:
        i. If $D(d_i , s_j) \leq$ minimum_distance[i]
              $d_i$ remains in this cluster
        ii. Else
            1. Calculate the distance from each $d_i \in D$ to each cluster seed $s_j \in S$, $D(d_i , s_j)$
            2. Assign $d_i$ for the closest cluster $s_j$
            3. Save label j of cluster $s_j$ in which $d_i$ is, into closest_cluster[i] and save $D(d_i , s_j)$ into minimum_distance[i]
    b. Re-calculate the cluster seed for every cluster as the average of all data objects in the cluster
9. Till the convergence is met
```

4. EXPERIMENTAL RESULTS

To present a more objective comparison and evaluation, the proposed clustering algorithm is tested through a group of COREL images, formed by ten image semantic categories, each one containing hundred color images. The semantic categories are Dinosaurs, Buses, Buildings, Foods, Mountains, Africans, Horses, Flowers, Elephants, and Beach with corresponding IDs denoted using integer numbers from one to ten respectively. Within COREL images, it is well-known whether 2 images with the same category. Thus we can compare and evaluate the execution and performance of our proposed method in terms of quality of retrieval accuracy, image clustering, and speed. Particularly, the quality of image clustering can be measured through distribution of the semantics within the formed cluster, and the retrieved image can be considered a true match when it is with same class as the user's query image.

4.1 Quality of Clustering

The proposed clustering algorithm would be capable to generate a set of image clusters. Each cluster includes images of identical or even similar semantics. Furthermore, the confusion matrices are good way to evaluate clustering performance. Whatever the case, to calculate these matrices, the amount of clusters must be the same as the amount of different semantics, that is anonymous in practice. Though we can impose our proposed clustering algorithm to always produce ten clusters for our particular experiments, the experiment process would be distinct to the actual applications. So, we utilize entropy and purity to estimate the quality of our image clustering.

The cluster entropy is a beneficial cluster goodness measure. It observes the allocation of semantic categories in the cluster. The entropy of cluster C_i can be presented as follow:

$$E(C_i) = \frac{1}{\log k}\sum_{j=1}^{k}\frac{|C_{i,j}|}{C_i}\log\frac{|C_{i,j}|}{C_i}, \qquad (16)$$

where $|C_{i,j}|$ denotes the number of images at cluster C_i which belongs to class j, and $|C_i|$ denotes the total amount of images at C_i. Every cluster may include images of diverse categories. The value of cluster entropy near 0 implies that it is comprised at most of one category; whilst the value near 1 means that it consists of a mix of all classes.

Another valuable cluster goodness measure is cluster purity. It observes the proportion of the predominant semantic category size within a cluster to its size oneself. The purity of cluster C_i can be presented as follow:

$$P(C_i) = \frac{1}{|C_i|}\max_{j=1,\dots,k}|C_{i,j}| \qquad (17)$$

Some further notations employed in the system performance evaluation are presented. For the query image *j*:

1. m_j denotes the *total number of clusters retrieved*.
2. *E*(*j*) denotes the *average cluster entropy of clusters retrieved,*

$$E(j) = \frac{1}{m_j}\sum_{i=1}^{m_j} e(C_i), \qquad (18)$$

3. *P*(*j*) denotes the *average cluster purity of clusters retrieved,*

$$P(j) = \frac{1}{m_j}\sum_{i=1}^{m_j} p(C_i), \qquad (19)$$

Each image within our experimental database is examined as a user query. For the query images of the same semantic class, the next statistics are calculated: the average of E(j) and the average of P(j). As shown in Table 1 and Table 2, the proposed clustering algorithm produces ideal quality clusters. Compared against the entropy and purity of subsets of images produced by the traditional k-means method, the goodness and validity of image clusters produced by our proposed method is in average much enhanced for all classes.

Table 1. Purity of the clusters. (Larger values denote purer clusters)

ID.	Class Name	Average $P_{k\text{-}means}$	Average $P(j)$
1	Dinosaurs	1	1
2	Buses	0.77	0.89
3	Buildings	0.46	0.8
4	Foods	0.52	0.9
5	Mountains	0.34	0.76
6	Africans	0.42	0.81
7	Horses	0.71	0.9
8	Flowers	0.83	0.96
9	Elephants	0.36	0.75
10	Beach	0.35	0.76

Table 2. Entropy of the clusters. (Smaller values indicate superior cluster quality)

ID.	Class Name	Average $E_{k\text{-}means}$	Average $E(j)$
1	Dinosaurs	0	0
2	Buses	0.45	0.14
3	Buildings	0.76	0.28
4	Foods	0.71	0.15
5	Mountains	0.81	0.3
6	Africans	0.78	0.24
7	Horses	0.53	0.16
8	Flowers	0.36	0.08
9	Elephants	0.82	0.3
10	Beach	0.81	0.3

4.2 Retrieval Accuracy

With regard to image retrieval, entropy and purity may not present a comprehensive evaluation of the retrieval system performance although they estimate the goodness of formed image clusters. For what could occur is a group of semantically clear and pure clusters, while none of these image clusters sharing the very same image semantics together with the users' query image. Thus, one wants to observe the relationship among the query and clusters. For this reason, we utilized the average precision. It is calculated as follow:

$$P = \frac{Pertinent\ images\ retrieved}{Overall\ number\ of\ database\ images\ retrieved} \qquad (20)$$

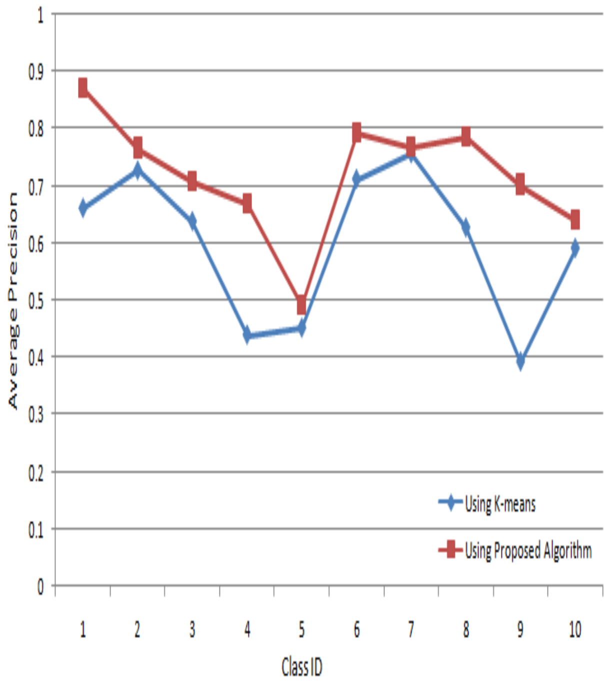

Fig 1: Comparison of average retrieval precision

Figure 1 presents the average retrieval precision for every category for all experimental images through the proposed clustering algorithm. As shown, our system operates superior than implementing the conventional k-means algorithm.

4.3 Speed

The conventional k-means method always converges to the local minimum. In k-means method, before it converges, the calculations of cluster centers and distance are completed while numbers of iterations are performed *t* times. The value of *t* changes relying on the initial cluster centroids. The allocation of data objects has an important relation with the fresh cluster centers. The computational complexity for the conventional k-means method is O(n × k × t), where n denotes the total amount of the data objects, t denotes the amount of iterations, and k denotes the amount of image clusters.

This paper proposed a new rapid and influential clustering algorithm based on k-means++ clustering method, to acquire helpful initial cluster centroids. The computational complexity for the proposed clustering method is O(n × k). In the suggested method, some data objects stay in their original clusters, whilst the others transmit to different clusters. So, if the object stays in the original position, this demands O(1), else it demands O(k). Together with the algorithm convergence, the amount of data objects transmitted from their clusters will decrease. If 50% of the objects transmit from their original clusters, the computational complexity will be $O(^{n}/_{2} \times k)$. As a consequence, the overall time complexity is O(n × k), while the conventional k-means clustering method demands O(n × k × t).

The proposed clustering algorithm has been tested on Intel(R) Core(TM) i3 2.4 GHz PC running Windows operating system. To compare and contrast the speed of our suggested method with the conventional k-means clustering method, which is tested and implemented on the same PC, 100 queries are implemented at random. For similarity evaluation and clustering, the suggested clustering method takes in average 0.6 second per query, whilst the traditional k-means takes 1.15 second in order to estimate image similarities and cluster the results. The implementation time is quite satisfactory within the toleration of the real time content-based image retrieval.

5. CONCLUSIONS

One of the main drawbacks of the standard k-means algorithm is that, it requires computing the distance of each data object to all of the k centroids when it performs the iteration every time, which occupies extensive execution time, particularly for large capacity databases. So we developed a public purpose clustering technique particularly to treat this shortcoming. It can efficiently and effectively handle large capacity image databases. It can be fluently embedded into distinct CBIR systems. The co-occurrence matrix features and color moments are utilized to evolve an effective and innovative image clustering framework. The texture and color features are integrated to enhancing results obtained using individual descriptors.

The experimental outcomes indicate that, the proposed clustering algorithm produces ideal quality clusters. Compared against the entropy and purity of subsets of images produced by the traditional k-means clustering method, the goodness of the image clusters produced by our proposed method is in average much enhanced for all classes. Also, it takes in average 0.6 second per query, whilst the traditional k-means takes 1.15 second in order to estimate image similarities and cluster the results. The implementation time is quite satisfactory within the toleration of real time content-based image retrieval.

6. REFERENCES

[1] Datta R., Joshi D., Li J., and Wang J.Z., "Image Retrieval: Ideas, Influences, and Trends of the New Age," *ACM Comput Surv*, vol. 40, no. 2, pp. 5:1-60, 2008.

[2] Alnihoud J., “Content-based Image Retrieval System Based on Self Organizing Map, Fuzzy Color Histogram and Subtractive Fuzzy Clustering,” *The International Arab Journal of Information Technology*, vol. 9, no. 5, pp. 452-458, 2012.

[3] Hurtut T., Gousseau Y., and Schmitt F., "Adaptive Image Retrieval Based on the Spatial Organization of Colors," *Comput Vis Image Und*, vol. 112, pp. 101–113, 2008.

[4] Karthikeyan M. and Aruna P., "Probability Based Document Clustering and Image Clustering using Content-based Image Retrieval," *Appl Soft Comput*, vol. 13, no. 2, pp. 959–966, 2013.

[5] Lin C. and Lin W., "Image Retrieval System Based on Adaptive Color Histogram and Texture Features," *Comput J*, vol. 54, no. 7, pp. 1136–1147, 2010.

[6] Lin C., Chan Y., Chen K., Huang D., and Chang Y., "Fast Color Spatial Feature Based Image Retrieval Methods," *Expert Syst Appl*, vol. 39, no. 9, pp. 11412–11420, 2011.

[7] Lin C., Chen R., and Chan Y., "A Smart Content-based Image Retrieval System Based on Color and Texture Feature," *Image Vision Comput*, vol. 27, no. 6, pp. 658–665, 2009.

[8] Min R. and Cheng H., "Effective Image Retrieval using Dominant Color Descriptor and Fuzzy Support Vector Machine," *Pattern Recognit*, vol. 42, no. 1, pp. 147–157, 2009.

[9] Młynarczuk M., Gorszczyk A., and S´lipek B., "The Application of Pattern Recognition in the Automatic Classification of Microscopic Rock Images," *Comput Geosci*, vol. 60, pp. 126–133, 2013.

[10] Nezamabadi-Pour H. and Kabir E., "Image Retrieval using Histograms of Unicolor and Bi-color Blocks and Directional Changes in Intensity Gradient," *Pattern Recognit Lett*, vol. 25, no. 14, pp. 1547–1557, 2004.

[11] Serrano-Talamantes J., Aviles-Cruz C., Villegas-Cortez J., and Sossa-Azuela J., "Self Organizing Natural Scene Image Retrieval," *Expert Syst Appl*, vol. 40, no. 7, pp. 2398–2409, 2013.

[12] Subrahmanyam M., Jonathan Wu Q., Maheshwari R., and Balasubramanian R., "Modified Color Motif Co-occurrence Matrix for Image Indexing and Retrieval," *Comput Electr Eng*, vol. 39, no. 3, pp. 762–774, 2013.

[13] Subramanian M. and Sathappan S., "An Efficient Content Based Image Retrieval using Advanced Filter Approaches," *The International Arab Journal of Information Technology*, vol. 12, no. 3, pp. 229-236, 2015.

[14] Tajeripour F., Saberi M., and Fekri-Ershad S., "Developing a Novel Approach for Content Based Image Retrieval Using Modified Local Binary Patterns and Morphological Transform," *The International Arab Journal of Information Technology*, vol. 12, no. 6, pp. 574-581, 2015.

[15] Wang H., Mohamad D., and Ismail N.-A., "An Efficient Parameters Selection for Object Recognition Based Colour Features in Traffic Image Retrieval," *The International Arab Journal of Information Technology*, vol. 11, no. 3, pp. 308-314, 2014.

[16] Yildizer E., Metin Balci A., Jarada T., and Alhajj R., "Integrating Wavelets with Clustering and Indexing for Effective Content-based Image Retrieval," *Knowl-Based Syst*, vol. 31, pp. 55–66, 2012.

[17] Hafiane A. and Zavidovique B., "Local Relational String and Mutual Matching for Image Retrieval," *Inform Process Manag*, vol. 44, no. 3, pp. 1201–1213, 2008.

[18] Fonseca M.J. and Jorge J.A., "Towards Content-based Retrieval of Technical Drawings Through High-dimensional Indexing," *Comput Graph*, vol. 27, no. 1, pp. 61–69, 2003.

[19] Raveaux R., Burie J.-C., and Ogier J.-M., "Structured Representations in a Content-based Image Retrieval Context," *J Vis Commun Image Represent*, vol. 24, no. 8, pp. 1252–1268, 2013.

[20] Frosini P. and Landi C., "Persistent Betti Numbers for a Noise Tolerant Shape-based Approach to Image Retrieval," *Pattern Recognit Lett*, vol. 34, pp. 863–872, 2013.

[21] Lee S., "Symmetry-driven Shape Description for Image Retrieval," *Image Vision Comput*, vol. 31, no. 4, pp. 357–363, 2013.

[22] Mohd Anuar F., Setchi R., and Lai Y.-K., "Trademark Image Retrieval using an Integrated Shape Descriptor," *Expert Syst Appl*, vol. 40, no. 1, pp. 105–121, 2013.

[23] Haridas K. and Thanamani A., "An Efficient Image Clustering and Content-based Image Retrieval using Fuzzy K-means Clustering Algorithm," *International Review on Computers and Software (IRECOS)*, vol. 9, no. 1, pp. 147-153, 2014.

[24] Maheshwari M., Silakari S., and Motwani M., "Image Clustering using Color and Texture," *Proc. 1st Conference on Computational Intelligence, Communication Systems, and Networks (CICSYN)*, pp. 403–408, 2009.

[25] Narasimhan H. and Ramraj P., "Contribution-based Clustering Algorithm for Content-based Image Retrieval," *Proc. 5th Conference on Industrial and Information Systems (ICIIS)*, pp. 442–447, 2010.

[26] Shrivastava R., Upadhyay K., Bhati R., and Mishra D., "Comparison between K-Mean and C-Mean Clustering for CBIR," *Proc. 2nd Conference on Computational Intelligence, Modeling, and Simulation (CIMSIM)*, pp. 117-118, 2010.

[27] Davis R.A., Zhongmiao Xiao, and Xiaojun Qi, "Capturing Semantic Relationship among Images in Clusters for Efficient Content-based Image Retrieval," *Proc. 19th IEEE Conference on Image Processing (ICIP)*, pp. 1953–1956, 2012.

[28] Rasli R., Muda T., Yusof Y., and Bakar J., "Comparative Analysis of Content-based Image Retrieval Techniques using Color Histogram: A Case Study of GLCM and K-Means Clustering," *Proc. 3rd Conference on Intelligent Systems, Modeling, and Simulation (ISMS)*, pp. 283–286, 2012.

[29] Mahajan S. and Patil D., "Image Retrieval using Contribution-based Clustering Algorithm with Different Feature Extraction Techniques," *Proc. Conference on IT in Business, Industry, and Government (CSIBIG)*, pp. 1-7, 2014.

[30] Stefan R.A., Szoke I.-A., and Holban S., "Hierarchical Clustering Techniques and Classification Applied in Content-based Image Retrieval," *Proc. 10th IEEE Jubilee Symposium on Applied Computational Intelligence and Informatics (SACI)*, pp. 147–152, 2015.

[31] Younus Z.S., Mohamad D., Saba T., Alkwaz M.H., Rehman A., Al-Rodhaan M., and et al., "Content-based Image Retrieval using PSO and K-means Clustering Algorithm," *Arab J Geosci*, vol. 8, pp. 6211-6224, 2015.

[32] Corel 1000 and Corel 10000 image database, available at: http://wang.ist.psu.edu/, last visited 2015.

[33] Choras R., Andrysiak T., and Choraś M., "Integrated Color, Texture, and Shape Information for Content-based Image Retrieval," *Patt Anal Appl*, vol. 10, no. 4, pp. 333–343, 2007.

[34] Arthur D. and Vassilvitskii S., "K-means++: The Advantages of Careful Seeding," *Proc. Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2007.