

Medical Images Retrieval using Clustering Technique

Enas M.F. El Houby

Systems & Information Dept., Engineering Division
National Research Centre
Dokki, Cairo, Egypt
em.fahmy@nrc.sci.eg

Abstract— The past few years have witnessed an increasing of media rich digital data such as images, videos, and audios. Medical domain is one of the rich images data domains. Retrieving images from large and varied collections of medical images databases is a challenging and important problem. So, image retrieval is one of the fastest and most exciting growing research areas in the field of medical imaging. In this research, clustering technique has been used to group similar medical images in the database together to decrease the retrieval time of the searched image. The searched image has been searched through different clusters to find the nearest cluster, and then the matched image or the nearest set of images has been searched inside that cluster. Different experiments had been applied to study the effect of changing number of clusters in the retrieval time and the correctness of the retrieved images, also the effect of changing threshold value on the correctness of the retrieved images has been studied.

Keywords- Image retrieval; Medical image; K-means; Clustering; Manhattan Distance.

I. Introduction

Due to the developments in various imaging technologies, the number of images produced from different sources especially in medical field increase in alarming rate. The internet is an excellent example of a distributed database containing several millions of images. Therefore it is very necessary to develop an appropriate information system to manage large collection of images [1, 2, 3]. One of the key issues in image management system is to locate a desired image in a large and varied collection of images. Patient-to-patient search, which can compare multiple patients and retrieve relevant cases among them, can be used as a training tool for medical students. The accuracy in diagnosis also can be improved by comparing similar images of patients [4].

Magnetic resonance (MR) images play a vital role in identifying various brain related problems. Some of the diseases of the brain show abnormalities predominately at a particular anatomical location which on MR appears at a slice at defined level. The first step in reporting of the magnetic resonance images is reviewing of the cross sectional images at various levels. Generally, the inter-patient search which can compare multiple patients and retrieve relevant cases among them will especially help the expert in diagnosis of structure-specific diseases, such as hippocampus or basal ganglia disorders. The retrieval of images of the same subject at a particular level from a large database is also quite useful in the decision-diagnosis process as the information available in the images is complementary [4].

Clustering is a process of grouping the data objects such that all objects in the same group are similar and object of other group are dissimilar. In literature, many categories of cluster analysis algorithms are presented. Partitioning methods are one of efficient clustering methods, where database is partition into groups in iterative relocation procedure. K-means is widely used partition method [5].

The clustering approach can be used to reduce the number of comparisons and to improve retrieval time when searching images database. It partitions the image space by

clustering the dataset. When a new target image is added to the database, the distance between this image and all clusters is computed and the image is associated to its nearest cluster. When a query image is probed, one step involves determining the nearest cluster and another step involves the computation of the distances between the query image and the target images assigned to the corresponding cluster [6].

In this research, the clustering technique has been applied on medical databases to improve retrieval time of the searched image. The proposed technique is based on partitioning of the images dataset into a set of clusters, such that minimizing the between-class similarities while maximizing the within-class similarities. K-means clustering method is the used clustering technique. To find a match for a searched image, the search will be through clusters till finding the nearest cluster, then searching for the matched image inside that cluster. The rest of this paper is organized as follows. In section II, an overview of previous related works is presented. In section III, the proposed K-means clustering method is described. In section IV, experimental results are provided. Conclusions and future works are given in section V.

II. RELATED WORK

A lot of previous work was studied for the subject of clustering and image retrieval. Varghese Abraham et al. [4] proposed a technique to locate desired slice using Rotational, Scaling and Translational (RST) invariant features derived from a ternary encoded local binary pattern (LBP) image. The LBP image is obtained by labeling each pixel with a code of the texture primitive based on the local neighborhood. The distance function based on the RST features extracted from LBP between query and database image is used to retrieve similar images corresponds to the query image. Azhar Qudus & Otman Basir [7] considered similar slice retrieval problem on the same class of images by giving importance to similarity of anatomical structures. Unay et al. [8] addressed brain image retrieval problem using Local Binary Patterns (LBP) in which similarity attributes of anatomical structures are extracted using fiducial points

which are obtained using a Kanade–Lucas–Tomasi (KLT) transform.

Cai, Deng, et al., [9] considered the problem of clustering Web image search results returned by an image search engine. The hierarchical clustering method using visual, textual and link analysis had been used to cluster Web images into different semantic clusters to facilitate users' browsing. Spectral techniques had been applied to find a Euclidean embedding of the images which respects the graph structure. In [10, 11] they considered the problem of clustering image search results. They claimed that in web image search, a good organization of the search results is as important as the search accuracy. Mishra, Pranjul, et al., [12] developed Content Based Image Retrieval (CBIR) systems which are capable to use image as the input query and retrieve images based on visual contents of the images. It is used for browsing and retrieving images from large database. Many clustering techniques with some feature of image have been used. Georgios Petkos, et al., [13] presented a scalable graph-based multimodal clustering which utilizes example relevant clustering to learn a model of the "same event" relationship between two items in the multimodal domain and subsequently to organize the items in a graph.

P. Chatur, P. Chouragade, [14] presented an image ranking and retrieval technique known as visual re-ranking which reordering of visual images based on their visual appearance. This approach relies on analyzing the distribution of visual similarities among the images and image ranking system that finds the multiple visual themes and their relative strengths in a large set of images. The major advantages of this approach are that, it improves the search performance by reducing the number of irrelevant images acquired as the result of image search and provides quality consistent output. Matthieu Guillaumin, et al., [15] presented a method for face recognition using a collection of images with captions. All faces of a particular person in a data set have been retrieved, and the correct association between the names in the captions and the faces in the images has been established.

III. THE PROPOSED MEDICAL IMAGES CLUSTERING TECHNIQUE

In this research, clustering technique has been used to group medical images dataset into groups, such that all images in the same group are similar and images of other group are dissimilar, K-means is the used partition method. The purpose of clustering the images is to group similar cases together and decrease the needed time to retrieve required image and all relevant images as possible. It enables to retrieve relevant cases and compare multiple patients, which help in diagnosis of diseases. After clustering the medical images dataset into k clusters, the retrieval of the searched image will need two steps. The first step is to find the nearest cluster to the searched image, then finding the needed image or set of the nearest images inside that cluster.

A. Material

Samples of medical images datasets have been used for the analysis of the proposed technique. The datasets have been collected from Pixmeo [16] website which is a Swiss company specialized in medical imaging software development. 233 slices of cross sectional images of magnetic resonance at various levels had been collected for brain tumor.

B. The k-means method

The K-means method has been used to group a given dataset D of n images into K clusters, Where $K \leq n$. The mean of the image is used as a representative of the image. The partition of the data images into K-clusters is performed such that each image belongs to exactly one cluster and each cluster contains at least one image. The Manhattan distance has been used to minimize the distance between the means of the images inside the same cluster, and maximize the distance between those images inside one cluster and those in the other clusters. The following equations can formulate the images clustering:

Equation (1) calculates the mean of the pixels of each image

$$m_j = \frac{1}{s} \sum_y \sum_x p_j(x, y) \quad (1)$$

Where m_j is the mean of the image number j, $P_j(x, y)$ is the pixel of the image number j at x, y dimensions and s is the total number of pixels of the image. Equation (2) computes the mean value of the images in a cluster which can be called as cluster's centroid or center of gravity. It can be calculated as follow:

$$\mu_i = \frac{1}{T} \sum_{j \in i} m_j \quad (2)$$

μ_i is cluster's centroid i.e., the mean of the means of the images grouped in the cluster number i, m_j is the mean of the image number j where image number j belongs to cluster number i and T is the number of images in the cluster i.

Equation (3) iteratively relocates the images into different clusters to minimize the difference between images and the centroid of clusters using Manhattan distance between the mean of the allocated image m_j , and the mean of clusters μ_i as follows:

$$D(j) = \min(|m_j - \mu_i|) \quad (3)$$

Where m_j is the mean of the image number j and μ_i is the centroid of the cluster number i and $D(j)$ is minimum distance between image m_j and centroids of different clusters. Using (1) – (3) a large dataset D of n images is grouped into k clusters. Each cluster is represented by a centroid of the cluster and a set of means of the images belonging to that cluster [17]. Algorithm1 depicts the different steps of K-means algorithm for partitioning given images into K-clusters.

Algorithm: K-means, algorithm for partitioning given images, where each cluster center is represented by mean value of images in the cluster.

Inputs:

1. k : number of clusters
2. D : a dataset containing n images.

Output: a set of k clusters.

Procedure:

- (1) Randomly choose k images from D dataset as initial cluster centers;
- (2) Repeat
 - (3) For each remaining data images repeat this step using Eq. (3):
 - Calculate distance between image mean (1) and each cluster centroid (2) ;
 - Compare the distances and assign the image to the most similar cluster whose distance is the most minimum;
 - (4) Update the new mean of each cluster to represent new cluster centroid using Eq. (2);
- Until there is no change in clusters.

Algorithm1: K-means algorithm for partitioning given images into K-clusters

C. The Retrieval of the Searched Image

After clustering the images dataset into k clusters, the retrieval of the searched image will need two steps. In the first step, the Manhattan distance between the mean of searched image (W), and the mean (μ_i) of each cluster of images is calculated. The nearest cluster to the searched image which has the minimum distance between its centroid and the mean of the searched image is the selected cluster. This is done as follows.

$$W = \frac{1}{s} \sum_y \sum_x p_w(x, y) \quad (4)$$

Where W is the mean of the searched image, $p_w(x, y)$ is the pixel of the searched image at x , y dimensions and s is the total number of pixels of the image.

$$D\mu(i) = (|W - \mu_i|) \quad (5)$$

Where W is the mean of the searched image, μ_i is the centroid of cluster number i and $D\mu(i)$ is the Manhattan distances between W and centroid μ_i of the cluster number i . The minimum value of the $D\mu(i)$ defines the cluster C which has the closest mean to the mean of the searched image W as follows:

$$C = \text{Minimum}(D\mu(i)) \quad (6)$$

Where C is the cluster that has the minimum Manhattan distances $D\mu(i)$ between its centroid μ_i and the mean of the searched image W among different k clusters. In the second step, the Manhattan distance between the mean of the searched image W and the mean m_j of each image in the selected cluster number C are calculated as follows:

$$Dm(j) = (|W - m_j|) \quad (7)$$

Where W is the mean of the searched image, m_j is the mean of image j in the selected cluster C and $Dm(j)$ is the Manhattan distances between W and m_j the images in the selected cluster C . The image that has distance value $Dm(j)$ is equal to zero represents the image that matches the searched image as follows:

$$Dm(j) = 0 \quad (8)$$

Where Dm is a list of the Manhattan distances between W and m_j the set of images in the selected cluster C . If (8) has satisfied then from (7):

$$W = m_j \quad (9)$$

Where W is the mean of the searched image and m_j is the mean of image j in the selected cluster number C that matched the searched image. In the case that there is not a true match of the searched image in the cluster, or the need for a set of images which is the nearest to the searched image, then a threshold value has been proposed to return a set of the nearest images to the searched image whose distance to the searched image are less than or equal the specified threshold (T). This can be done using the following equation:

$$Dm(j) \leq T \quad (10)$$

Where $Dm(j)$ is the set of the Manhattan distances between W and the mean m_j of the images in cluster C and T is a threshold specified by the user. This equation returns a set of images m_j that has Manhattan distances to W is smaller than or equal to T . Algorithm2 depicts the different steps of the retrieval of the searched image.

Algorithm: *retrieval*, algorithm for finding a matched image or nearest images to the searched image, where each searched image is represented by its mean value.

Inputs:

1. A set of k clusters of images
2. The searched image W
3. The threshold value

Output: The matched image or set of nearest images.

Procedure:

- (1) Calculate the mean value of the searched image using Eq(4);
- (2) For each cluster k
Calculate the distances between the searched image mean and the cluster centroid using Eq(5);
- (3) Find the nearest cluster C to the searched image that has the minimum distance as in Eq(6);
- (4) For each image inside cluster C
Calculate the distances between the mean of searched image and the mean of each image inside cluster C Using Eq(7);
- (5) Find the matched image inside cluster C whose distance to the searched image $=0$ as in Eq(8);
- (6) Otherwise find the nearest set of images whose distance to the searched image \leq Threshold as in Eq(10).

Algorithm2: The retrieval algorithm of the searched image

IV. EXPERIMENTAL RESULTS AND DISCUSSION

This section shows an empirical performance evaluation of the proposed technique. Extensive experimental studies had been tried on the collected medical datasets in order to test the proposed technique. Fig. 1 shows samples of brain tumor images from the collected medical dataset. All of our experiments are carried out using processor Intel® Core (i7) CPU@1.6 GHz, 4 GB of RAM with the Matlab® platform.

The images in the dataset are distributed on K clusters. The number of clusters K has been changed to study the effect of its changing on the efficiency of the proposed technique. Fig. 2 shows the relation between changing number of clusters and the times needed to create clusters. It is shown that the time to allocate the images of the dataset on clusters increases as the number of clusters increases. It ranges from 0.005 to 0.12 second to cluster the images into 1 to 120 clusters. Although there is a bias around the best fit line, but in general with the increasing number of clusters the time to allocate images increases. This bias may due to the k -means depends on a random initial selection of k images from dataset as initial clusters centroids, as mentioned before in algorithm1. So according to the initial selection, the number of iterations needed to reach the stable clusters are changed and so the time.

As all MR images are centered, take most of the area of the image with black background. So these features have no effect on the calculations of the mean of each image. The probability that two images would have the same mean is expected to be zero. Fig. 3 shows the effect of changing number of clusters on the number of correctly retrieved images. As shown from Fig. 3 with changing number of clusters, the number of correctly retrieved images is constant and equals 231 out of 233. While the number of incorrectly retrieved image is also constant and equals 2 as shown in fig. 4. So the number of clusters hasn't any effect on the correctness of the retrieval of images.

Different number of clusters has been used to study the

effect of changing number of clusters on the retrieval time. As shown in Fig. 5, with increasing the number of clusters, the time needed to find the correct cluster for the searched image increases, as the search is through more number of clusters. The time ranges from almost 0.5×10^{-4} to 6×10^{-4} seconds to find the cluster of the searched image. While with increasing number of clusters, the time needed to search inside the selected cluster decrease, because the number of images per cluster decreases. As shown in Fig. 6, the time ranges from about 0.05×10^{-3} to 1.3×10^{-3} seconds to find a match to the searched image in the selected cluster. It is clear that the time of finding a match to the searched image inside the selected cluster is more important than the time of finding a cluster. So the selection of the best number of clusters should depend mainly on the time of finding a matched image in the cluster. It seems that it is suitable to select number of cluster start from 22 clusters.

Threshold can be used to retrieve a set of images which are closest to the searched image instead of a single matched image. Comparing these set of images may help in diagnosis and following the patient's cases. By assuming threshold value is 0.3, and searching for the nearest images to image (a) in Fig. 7, the set of images from (b) to (f), in addition to (a) are retrieved as shown in fig.7. The retrieved images have mean values that differ from the searched image by value less than or equal the threshold (0.3). The expert can have a chance to select from the returned set of images. By assuming threshold value is 0.2, and searching for the nearest images to image (a) in Fig. 8, the set of images from (b) to (c), in addition to the image itself (a) are retrieved from the dataset, as shown in fig. 8. By increasing threshold to 0.3, image (d) is added to the set of retrieved images, it is visually dissimilar to (a), but it may be of value to the expert. By increasing threshold to value greater than 0.3, extra images are retrieved.

Different images have been retrieved using different threshold values and the precision have been calculated as the percentage of visually similar retrieved images to the total number of retrieved images as in (11). Table 1 shows

the number of retrieved images using different values of threshold. As it is shown in table 1, as the threshold increases the number of retrieved images increases but the precision decreases. In image 1, the number of retrieved images is high ranging from 6 to 26, with changing in threshold from 0.1 to 0.8 but the precision is varied from 83.3% to 61.5%. For images from 2 to 5, the numbers of retrieved images are low, for image 2 the precision is 100% until 0.2 threshold value and is dropped to 75% at threshold 0.3 and remains 75% till threshold 0.8, but for images 3, 4 the precisions remain 100% until threshold 0.7 and are dropped to 71.4 at threshold 0.8. For image 5 the precisions are varied from 66.7% to 75% with changing threshold values. From these results, the precision depends on the

searched images and the rate of existence of similar images relative to the dissimilar images which have mean values close to the mean of searched image in the dataset, but in general the value of 0.3 or less for threshold is recommended to guarantee the retrieval of the nearest images with high precision.

$$\text{Precision} = \frac{\text{visually similar retrieved images}}{\text{total retrieved images}} \tag{11}$$

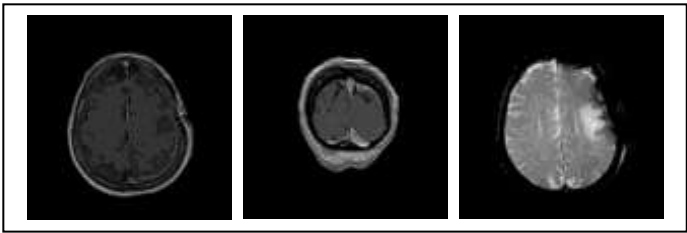


Figure1: A sample of images from brain tumor dataset

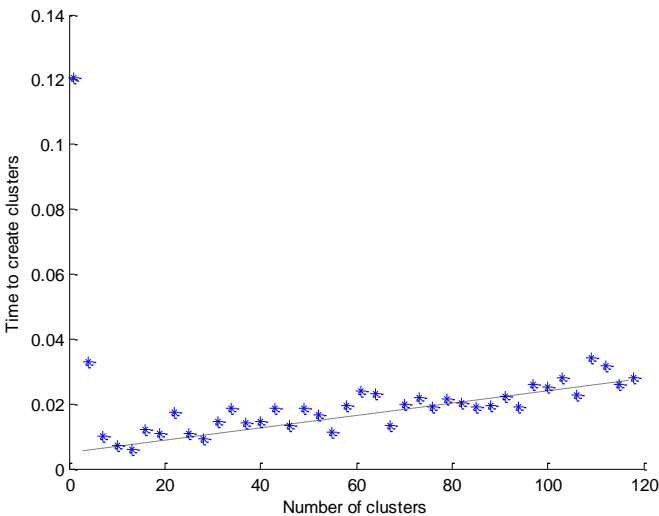


Figure 2: Number of clusters versus times needed to create clusters

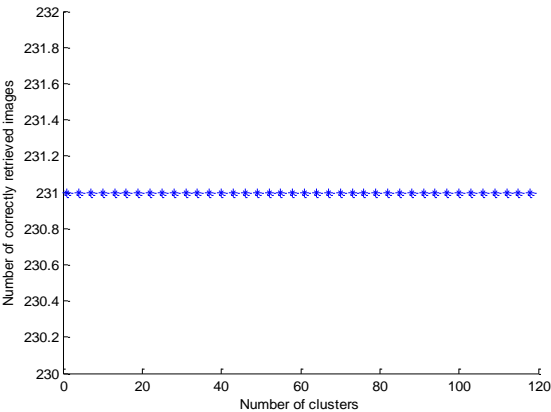


Figure 3: Number of correctly retrieved images versus number of clusters

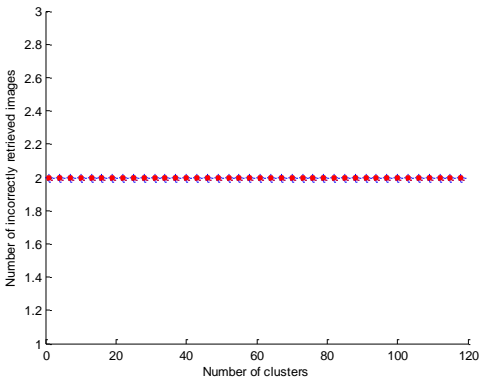


Figure 4: Number of incorrectly retrieved images versus number of clusters

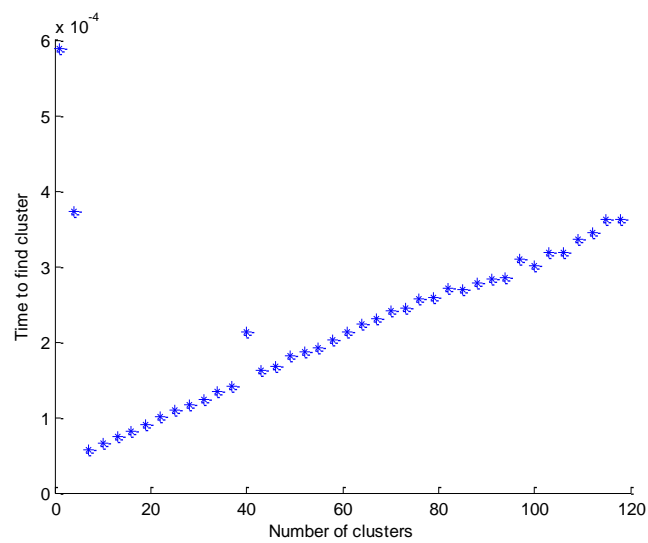


Figure5: The time needed to find the nearest cluster to the searched image.

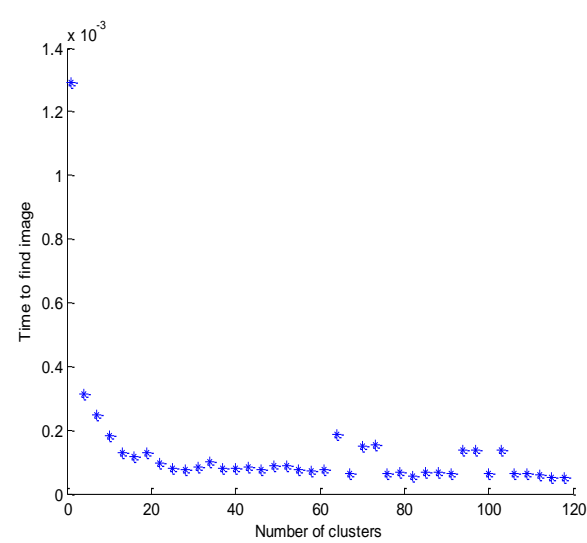


Figure6: The time needed to find the searched image inside the selected cluster

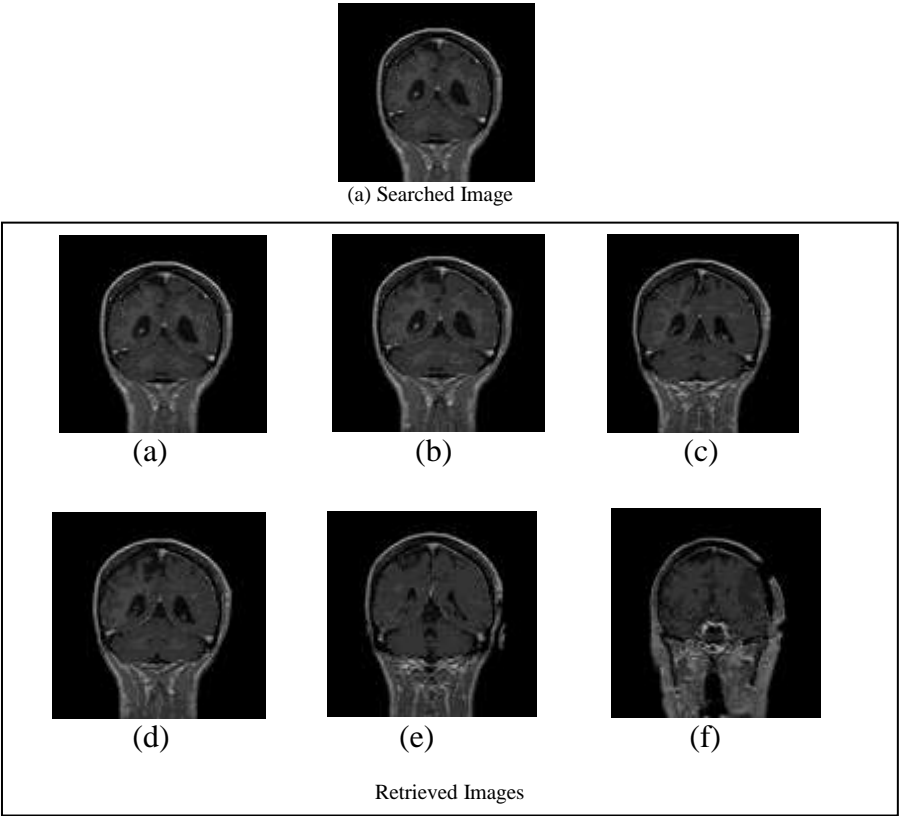


Figure7: A set of images which are retrieved by searching for image (a) using threshold value equal 0.3

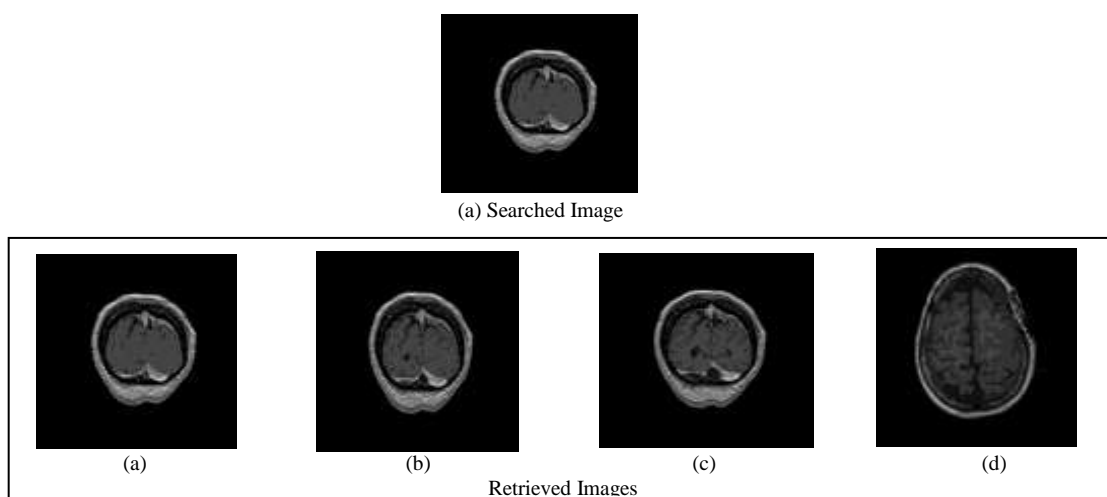



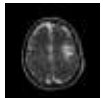
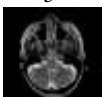


Figure8: A set of images which are retrieved by searching for image (a) using threshold value equal 0.2

Table 1: The number of retrieved images using different values of threshold

image	Threshold	Retrieved image no.	Visually Similar	Visually Dissimilar	Precision %
	0.1	6	5	1	83.3
	0.2	7	5	2	71.4
	0.3	14	10	4	71.4
	0.4	16	11	5	68.75
	0.5	20	13	8	65
	0.6	20	13	8	65
	0.7	26	16	10	61.5
	0.8	26	16	10	61.5
	0.1	2	2	0	100
	0.2	3	3	0	100
	0.3	4	3	1	75
	0.4	4	3	1	75
	0.5	4	3	1	75
	0.6	4	3	1	75
	0.7	4	3	1	75
	0.8	4	3	1	75
	0.1	3	3	0	100
	0.2	4	4	0	100
	0.3	4	4	0	100
	0.4	4	4	0	100
	0.5	4	4	0	100
	0.6	5	5	0	100
	0.7	5	5	0	100
	0.8	7	5	2	71.4
	0.1	3	3	0	100
	0.2	4	4	0	100
	0.3	4	4	0	100
	0.4	4	4	0	100
	0.5	4	4	0	100
	0.6	4	4	0	100
	0.7	5	5	0	100
	0.8	7	5	2	71.4
	0.1	3	2	1	66.7
	0.2	3	2	1	66.7
	0.3	4	3	1	75
	0.4	4	3	1	75
	0.5	6	4	2	66.7
	0.6	7	5	2	71.4
	0.7	7	5	2	71.4
	0.8	7	5	2	71.4

V. CONCLUSION AND FUTURE WORK

In this research, K-means clustering technique had been used to distribute images dataset into k clusters to improve the search performance. The proposed technique had been used to retrieve MR images from medical dataset successfully. The time required to find a searched image from the dataset depends on the number of clusters, as the number of clusters increases, the time required to find the correct cluster increases but the time required to find the correct image inside the selected cluster decreases. Threshold has been used to retrieve a set of related images to the searched image. In the future, different datasets especially medical can be analyzed and different techniques can be tried to retrieve images. Different features of images can be used to retrieve the searched image. And different related images can be retrieved which can help in diagnosis and follow patients cases.

References

- [1] Müller, H., Michoux, N., Bandon, D., Geisbuhler, A. A review content-based image retrieval systems in medical applications—Clinical benefits and future directions. *Int. J. Med. Inf.* 73: pp. 1-23, 2004.
- [2] Lehmann, T.M., Wein, B., Dahmen, J., Bredno, J., Vogelsang, F., Kohnen, M., Content-based image retrieval in medical applications: A novel multi-step approach. *Proceedings SPIE* 3972: pp. 312-320, 2000.
- [3] Y. Jing and S. Baluja, "VisualRank: Applying PageRank to Large-Scale Image Search", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 30, No. 11, 2008.
- [4] Abraham Varghese, Kannan Balakrishnan, Reji R. Varghese, and Joseph S. Paul, "Content Based Image Retrieval of Brain MR Images across Different Classes", *International Scholarly and Scientific Research & Innovation* 7(8) 2013.
- [5] Chatti Subbalakshmi, Rao P. Venkateswara, and S. Krishna Mohan Rao. "Performance Issues on K-Mean Partitioning Clustering Algorithm." *International Journal of Computer (IJC)* 14.1, PP 41-51, 2014.
- [6] Perronnin Florent, Dugelay Jean-Luc, "Clustering face images with application to image retrieval in large databases", *Proceedings of the SPIE- The International Society for Optical Engineering Journal*, Volume 5779, pp. 256-264, 2005.
- [7] Azhar Quddus, and Otman Basir, "Semantic Image Retrieval in Magnetic Resonance Brain Volumes". *IEEE transactions on information technology in biomedicine*, Vol. 16, NO. 3, May 2012.
- [8] D. Unay, A. Ekin, and R. Jasinsch. "Local Structure-Based Region-Of- Interest Retrieval in brain MR Images," *IEEE Transactions on Information technology in Biomedicine*, vol.14, No.4, July 2010.
- [9] Deng Cai, Xiaofei He, Zhiwei Li, Wei-Ying Ma, Ji-Rong Wen, "Hierarchical clustering of WWW image search results using visual, textual and link information." *Proceedings of the 12th annual ACM international conference on Multimedia*. ACM, 2004.
- [10] Vivisimo clustering engine, (2004) <http://vivisimo.com>.
- [11] O. Zamir and O. Etzioni, "Grouper: A Dynamic Clustering Interface to Web Search Results". In *Proceedings of the Eighth International World Wide Web Conferenc*, 1999.
- [12] Mishra, Pranjul, Ms Sonam, and Mrs S. Vijayalakshmi. "Content Based Image Retrieval Using Clustering Technique: A Survey." *International Journal of Research in Computer Engineering & Electronics* 3.2, 2014.
- [13] Georgios Petkos, Symeon Papadopoulos, Emmanouil Schinas, Yiannis Kompatsiaris, "Graph-based multimodal clustering for social event detection in large collections of images." *MultiMedia Modeling*. Springer International Publishing, Volume 8325, pp 146-158, 2014.
- [14] PrashantChatur, Pushpanjali Chouragade, "A soft computing approach for image searching using", *international journal of computer engineering & technology (ijcet)*, Volume 4, Issue 2, March – April, pp. 543-555, 2013.
- [15] Matthieu Guillaumin, Thomas Mensink, Jakob Verbeek, Cordelia Schmid, "Face recognition from caption-based supervision." *International Journal of Computer Vision* 96(1), PP 64-82, 2012.
- [16] <http://www.osirix-viewer.com/datasets/>
- [17] Han, Jiawei. Kamber, Micheline. *Data Mining Concepts and Techniques*, Morgan Kaufmann Publishers, 2nd Edition, 2006.