

A Comparison of Image Grouping Techniques of Content Based Image Retrieval Using K-Means Clustering Algorithm

Muhamadainanshah Adnan, N.M. Nik Arni, M.O. Balkish

Abstract: Content Based Image Retrieval (CBIR) system is an alternative approach to Text Based Image Retrieval (TBIR) system in retrieving the images. The system consists of three phases which are feature extraction, image grouping and image retrieval. This study focused on colour feature for feature extraction process, image grouping for grouping images according to their characteristic similarities. For image retrieval, several well-known clustering techniques were introduced and applied to CBIR system. The clustering technique of K-Means type is the most preferable clustering technique since it is easy to be implemented and also fast computation. However, because of many improvement that have been done towards this technique, there exist variations of K-Means clustering algorithms. Thus, in this research, a comparison performance among three types of K-Means clustering algorithms, namely the basic K-Means, Fuzzy K-Means and K-Harmonic Means algorithms is performed. Four validation techniques are used for determining the most efficient algorithm in retrieving the images, which were Davies-Bouldin index (DB), Calinski-Harabasz index (CH), Dunn index (Dunn) and Silhouette width (SC). Based on these four validation techniques, the K-Harmonic Means clustering algorithm was found to be the best clustering algorithms in grouping image dataset.

Keywords: Image Grouping Techniques, Content Based Image Retrieval, K-Means clustering algorithm.

I. INTRODUCTION

As the technology keeps growing, all information can be stored in the internet. Facebook is one of the places where people store their information. All this includes image, videos and personal data. It has been reported that about 6 billion images and 72 hours of videos is being uploaded monthly to face book. Because of the large database, sometimes it is hard to retrieve information that had been stored. Therefore, a lot of research had been done to overcome this problem. According to Torres and Falcao [1],

some systems need to be developed in order to manage the large database especially in searching the images. So, one of the systems that have been invented is the Image Retrieval System.

Image retrieval is the process of searching and retrieving the images from large database of digital images [2]. There are two approaches for this system which are the traditional approach known as Text Based Image Retrieval (TBIR) and Content Based Image Retrieval (CBIR). With TBIR, images are retrieved by using keywords [1] while with CBIR images are retrieved by content description by the user.

The text based is commonly used by the search engine such as Yahoo and Google. However, this approach consume more time and very laborious to retrieve the images [1], [3]. Also, since this approach uses text annotation, an image is difficult to be described precisely [3]. This means that different people may use different keyword to describe the same image [4]. Sometimes, people might also misspell the keyword which will affect the results of the retrieval images [4]. Therefore, in order to overcome these weaknesses, the CBIR system has been introduced where this approach is directly retrieve the image based on its visual content such as colour, shape and texture.

Basically CBIR system is a technology that distinguishes different regions from an image based on their equality in colour, pattern, texture and shape. Sometimes, there exists some noise either in the image database or in the query image itself. This noise affects the quality of the image. Therefore, this research considers the presence of noise in both image database and query image in order to measure the performance of the clustering algorithms. The main goal of doing clustering is to gain insight of data which include detecting anomalies and identifying salient features [5]. Clustering will also classify and compress the data. The similarity between the two images is measured by the closeness of one region to another. By extracting the features of an image, it makes CBIR system perform better in the process of searching, browsing and content mining compared to TBIR system [4]. There are many different approaches that can be used to perform CBIR system. However, according to Chang et. al [4], clustering techniques is commonly used to perform the CBIR system for easier retrieval of the images. The most preferable algorithm compared to other clustering algorithm is the K-Means algorithm since it is easy to be implemented and fast computation [6].

Revised Manuscript Received on February 05 2019.

Muhamadainanshah Adnan, Centre for Statistical and Decision Science Studies, UiTM, Faculty of Computer and Mathematical Sciences, University Teknologi MARA (UiTM) 40450 Shah Alam, Selangor, Malaysia

N.M. Nik Arni, Centre for Statistical and Decision Science Studies, UiTM, Faculty of Computer and Mathematical Sciences, University Teknologi MARA (UiTM) 40450 Shah Alam, Selangor, Malaysia

M.O. Balkish, Centre for Statistical and Decision Science Studies, UiTM, Faculty of Computer and Mathematical Sciences, University Teknologi MARA (UiTM) 40450 Shah Alam, Selangor, Malaysia



However, due to its high computational complexity for large data sets, a lot of researches have been done to improve this algorithm. As a result, many variations of K-Means Clustering algorithms were developed to retrieve the image. This research chooses and focuses only on three types of K-Means Clustering algorithms in order to determine the best K-Means Clustering algorithm to be used in CBIR system. These algorithms are basic K-Means algorithm, Fuzzy K-Means algorithm and K-Harmonic Means algorithm.

In CBIR system, several features extraction methods used are colour, pattern, texture and shape. However, this research only used image colour feature to extract the similarities in each image. It is hoped that the results of this

research benefit the academic communities by laying out the theoretical and statistical aspects of image retrieval, where statistical analysis can be applied to image data.

II. METHODOLOGY

In this study, the database used was the Wang database. This database consists of 1000 images with 10 classes of subset from Corel stock database. The 10 classes in this database are African people and villages, beaches, buildings, buses, dinosaurs, elephants, flowers, horses, mountains and glaciers, and food. Some example of images that contains in the database will be show in Figure 1 below:










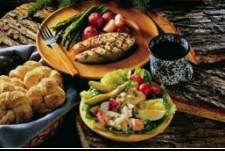
Class	Image	Class	Image
African People and Villages		Beaches	
Buildings		Buses	
Dinosaurs		Elephants	
Flowers		Horses	
Mountains and Glaciers		Food	

Fig. 1 Example of images that contains in the WANG image database

Process Flow in Content Based Image Retrieval

This research consists of three phases which started with the images features extraction. All features of the images in the database and also the query image will be extracted to represent the images. The second phase would be the clustering analysis. In this phase, the image database will be clustered according to their natural grouping. This natural grouping was obtained by comparing the features that represents the images. The clustering method used in this research was the K-means Algorithms. There are three types of K-means clustering algorithm that had been compared. These three types of algorithms are the Basic K-means, Fuzzy K-means and K-harmonic Means.

The last phase for this research was the retrieving process. In this phase, the query image will be assigned to the cluster that it naturally belongs to. Then, the process of retrieving the most similar image with the query image will be started by calculating the distance between a query image and each images in the cluster that it belongs to. The most similar image will be retrieved based on the smallest distance. The performance of the Content Based Image Retrieval by using the best K-means algorithms will be measured by using Accuracy and Redundancy Factors.

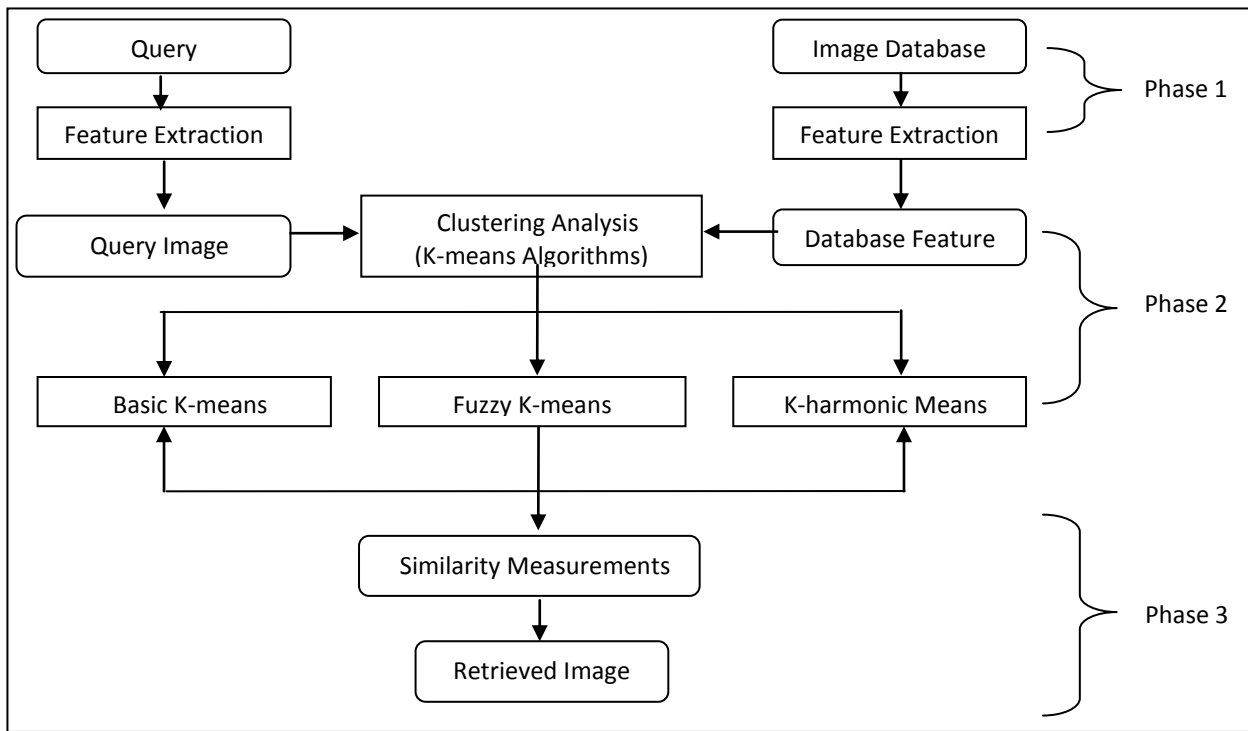


Fig. 2 Process Flow for Content Based Image Retrieval

Process Feature Extraction Method

This research will be using colour moment for representing the features of each image. This technique is very effective in doing colour based image analysis [7]. Since the distribution of any colour can be characterized and the information is concentrated on low order moments, only the first moment which is the mean, second moment which is the variance and the third moment which is the skewness are used [8].

1st moment (mean)

$$Mean(\mu) = \sum_{j=1}^N \frac{1}{N} P_{ij} \quad (3.1)$$

where,

N : the total number of pixels in the image
2nd moment (variance)

$$Variance(\sigma^2) = \left(\frac{1}{N} \sum_{j=1}^N (P_{ij} - E_i)^2 \right) \quad (3.2)$$

where,

N : the total number of pixels in the image
 E_i : the mean value for the i th colour channel of the image
3rd moment (skewness)

$$Skewness(s) = \left(\frac{1}{N} \sum_{j=1}^N (P_{ij} - E_i)^3 \right) \quad (3.3)$$

where,

N : the total number of pixels in the image

E_i : the mean value for the i th colour channel of the image

K-Means Algorithm Procedures

In this study, the three different types of K-means algorithms used are Basic K-Means algorithm, the Fuzzy K-Means algorithm and the K-Harmonic Means algorithm. K-means clustering algorithm is one of the traditional and unsupervised learning algorithms in solving fundamental clustering problem by trying to find possible natural grouping in the data according to their similarities [9]. Thus, this method is useful in grouping the data according to their similarities. Improved algorithms based on the K-mean algorithm that been developed are the Fuzzy K-mean algorithm and the K-harmonic means algorithm.

Measure of Performance for K-Means Clustering Algorithms

In this study, there are four measures that will be used for measuring the performance of the above algorithms. The measures are Davies-Bouldin Index, Calinski-Harabasz Index, Dunn Index and Silhouette Width.

Measure of Performance for Content Based Image Retrieval

Accuracy Level

In order to evaluate the performance of the system that been proposed in this study, the level of the accuracy for the result will be calculated. The formula for the accuracy of the result in CBIR is



$$Accuracy = \frac{r}{R} \times 100 \quad (3.4)$$

where,

- r : number of relevant images
- R : total number of image retrieved

Redundancy Factor

In order to support the result for validation using accuracy level, the redundancy factor will also be calculated. This redundancy factor will be used to measure the performance of the K-means algorithms in retrieving image.

The formula to calculate this measurement is:

$$RF = \frac{(R - C)}{C} \quad (3.5)$$

where:

- RF : redundancy factor
- R : total number of images retrieved
- C : the number of images in a class

According to Chadha et.al [8], the system is said to be over-worked so that the RF value is greater than 0. If the RF value is lower than 0, the system is said to be under-worked which leads to underperforming. In other words, the system is said to get a good results when the value of the redundancy factor (RF) is approaching to 0.

III. RESULTS

Grouping Image Data

There were four validation techniques that are used to measure the clustering performance. The table below shows the results of the validation.

Table. 1 Cluster Validation Results for Images without Noise Database

Validation	Methods		
	K-Means	Fuzzy K-Means	K-Harmonic Means
DB	0.8810	0.9584	0.9060
CH	580.6155	510.7748	574.0201
Dunn	0.7140	0.5825	1.1448
SC	0.4246	0.3417	0.4568

Based on Table1, for DB index, the best clustering algorithm was the K-means clustering algorithm because this algorithm has the lowest validation value. This technique measured the average similarity between each cluster and it is almost similar. However, the lowest value for this validation shows that the cluster were compact and well separated. For CH index, this technique measured the rate between the dispersion between the cluster and dispersion within the group or cluster. Thus, based on this validation value, the best clustering algorithm was the K-means clustering algorithm since it has the highest validation value.

For Dunn index, for this technique, the dataset was clustered well when the distance among the cluster is large and the diameters of the cluster are expected to be small. Therefore, the higher the value for this technique, the better the clustering algorithm which means that the K-Harmonic means was the best algorithm. For the last validation technique, it shows that the best algorithm was the K-Harmonic since it has the highest average Silhouette width.

The Silhouette width measured how well the observation was clustered. If the value approaches to 1 it means that the observation has been well clustered. Therefore, by average, the images in the database have been well clustered by K-Harmonic Means algorithm compared to the other two clustering algorithms.

Based on all four performance values, the single best K-Means clustering algorithm in grouping image dataset was not obtained. This is because two of the cluster validation which were Davies-Bouldin index (DB) and Calinski-Harabasz index (CH) indicated that the K-Means clustering algorithm was the best while the other two cluster validations which were index and Silhouette width (SC) indicated that the K-Harmonic Means clustering algorithm was the best. Consequently, in order to determine the best K-Means clustering algorithm, the present dataset were split into two parts which were the evaluation part and validation part. The evaluation part consists of 60% of the image database that has been selected randomly according to classes while the remaining 40% of the image database is reserved for the validation part.

Table. 2 Cluster Validation Results for Evaluation and Validation Part

Validation	Evaluation			Validation		
	K-Means	Fuzzy K-Means	K-Harmonic Means	K-Means	Fuzzy K-Means	K-Harmonic Means
DB	0.8885	0.9937	0.8753	0.8921	0.8718	0.8795
CH	345.4314	322.8943	308.2569	222.1827	225.9265	225.1934
Dunn	0.8493	0.7011	0.8726	0.7593	0.7480	0.9448
SC	0.4217	0.3800	0.4281	0.4275	0.4055	0.4416



Table 2 shows the clustering performance for the image dataset that has been split into two parts. In both evaluation and validation section, K-Harmonic Means was the best algorithm. Conclusively, this result proves that the best K-Means clustering algorithm in grouping image database based on colour moment features is the K-Harmonic Means clustering algorithm. Thus, this clustering algorithm has been implemented in the CBIR system.

The Performance of Content Based Image Retrieval System

The first system performance is the accuracy level. Based on Table 3, it is apparent that the accuracy level for the system varies according to image classes. The highest accuracy level is from retrieving images from original database for class 7 (Dinosaurs) and class 8 (Horses) with both nearly reach 50% of accuracy level respectively.

Table. 3 Accuracy Level of Content Based Image Retrieval System

Image Class	K-Harmonic Means
class 1	27.520
class 2	24.102
class 3	15.983
class 4	35.282
class 5	34.626
class 6	17.573
class 7	48.411
class 8	47.317
class 9	18.188
class 10	26.509

Whether the CBIR system is under-worked or over-worked is measured by the redundancy factor. The system was categorized as under-worked which also can be denoted as underperformed if the value of redundancy factor is negative. In other words, some of the images that are supposed to include in retrieval images may not be included. The system was categorized as over worked or over performed when the value of redundancy factor is bigger than zero. This means that there are other images that are not supposed to be in the class were included in the retrieval process. For example, in this study, there are 100 images in each class in the database. Therefore, the system was categorized as a good system when the redundancy factor approached zero.

Table. 4 Redundancy Factor of Content Based Image Retrieval System

Image Class	K-Harmonic Means
class 1	0.830
class 2	0.493
class 3	0.120
class 4	-0.037
class 5	0.203
class 6	0.377
class 7	-0.313
class 8	0.483
class 9	0.150
class 10	0.150

Based on Table 4, the results for retrieving images using original database varied according to the image classes. The system performance in retrieving image from class 4 (Buses) was the most optimum since the redundancy factor value was near to zero. This means that the total retrieving images for these classes were nearly equal to the total images for it classes.

IV. CONCLUSIONS

As a formative conclusion for this study, it is proven that statistical analysis can also be applied on image data by extracting the features of the image. Since the presence of noise in an image may affect the features that represent the images, all the images in the database were grouped according to their natural grouping by using clustering algorithms. The clustering algorithms that were used in this study were K-Means clustering algorithm, Fuzzy K-Means clustering algorithm and K-Harmonic Means clustering algorithm. Based on the analysis performed, the best K-Means clustering algorithms in grouping image dataset was the K-Harmonic Means clustering algorithms since this method gives compact and small dispersion value within a cluster. The images were also well clustered.

REFERENCES

- da Silva Torres, R., &Falcão, A. X. (2006). "Content-Based Image Retrieval: Theory and Applications", RITA, 13(2), 161-185.
- Murthy, V. S. V. S., Vamsidhar, E., Kumar, J. S., &Rao, P. S. (2010). "Content based image retrieval using Hierarchical and K-means clustering techniques", International Journal of Engineering Science and Technology, 2(3), 209-212.
- Huu, Q. N., Thu, H. N. T., &Quoc, T. N. (2012). "An efficient content based image retrieval method for retrieving images", International Journal of Innovative Computing, information and Control, 8(4).
- Chang, R. I., Lin, S. Y., Ho, J. M., Fann, C. W., & Wang, Y. C. (2012). "A novel content based image retrieval system using K-means/KNN with feature extraction", Computer Science and Information Systems/ComSIS, 9(4), 1645-1661.
- Celebi, M. E., et al. (2013). "A comparative study of efficient initialization methods for the k-means clustering algorithm", Expert Systems with Applications, 40(1), 200-210.
- Subitha, S., &Sujatha, S. (2013). "Survey paper on various methods in content based information retrieval", IMPACT: International Journal of Research in Engineering & Technology, 1(3), 109-120.
- Patil, J. K., & Kumar, R. (2011). "Color Feature Extraction of Tomato Leaf Diseases", International Journal of Engineering Trends and Technology, 2(2), 72-74.
- Chadha, A., Mallik, S., &Johar, R. (2012). "Comparative study and optimization of feature-extraction techniques for Content Based Image Retrieval", International Journal of Computer Applications, 52(20), 35-42.
- Zhou, H., & Liu, Y. (2008). "Accurate integration of multi-view range images using k-means clustering", Pattern Recognition, 41(1), 152-175.
- D. Davies and D. Bouldin, "A cluster separation measure", IEEE PAMI, vol. 1, no. 2, pp. 224-227, 1979.
- M. Halkidi and M. Vazirgiannis, "Clustering validity assessment using multi-representatives", in SETN, 2002.
- P. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis", J. Comput. Appl. Math., vol. 20, no. 1, pp. 53-65, 1987.
- Handl J, Knowles J, Kell DB (2005). "Computational Cluster Validation in Post-Genomic Data Analysis", Bioinformatics, 21(15), 3201-3212.

