IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-4, Apr 2019 E-ISSN: 2347-2693

An Improved K-Medoids Partitioning Algorithm for Clustering of Images

M. Kiruthika[1*], S. Sukumaran[2]

[1,2]Dept. of Computer Science, Erode Arts and Science College, Erode-638 009, Tamilnadu, India

[*]*Corresponding Author: kiruthikact7@gmail.com*

DOI: https://doi.org/10.26438/ijcse/v7i4.759764 | Available online at: www.ijcseonline.org

Accepted: 17/Apr/2019, Published: 30/Apr/2019

Abstract— Clustering is an unsupervised classification of patterns into clusters (groups). Image clustering is a system of partitioning image data into clusters on the basis of similarities. It is used in many practical areas like Medical Diagnosis, Military, Remote sensing and etc. It is one type of image indexing where images are categorized into different groups based on their features, such as shape, color, or texture. The purpose of this paper is clustering of visually similar images from the image database using clustering algorithms. The proposed method uses the GLCM (Gray-Level Co-Occurrence Matrix) texture features. The extracted GLCM features are then clustered applying different clustering algorithms such as K-Means, K-Medoids and Improved K-Medoids partitioning clustering techniques. In this work, Corel-1k database is used. This work presents a comparative analysis of various clustering algorithms for image clustering with GLCM feature extraction technique. The experimental outcome of this work shows performance of different clustering algorithms.

Keywords— Image Clustering, Feature Extraction, K-Means, K-Medoids

I. INTRODUCTION

In recent times, the rapid growth of high dimensional data, the deployment of huge image databases supporting a wide range of applications has now become achievable. Databases have a great potential in attracting more users in different fields like environmental, design, marketing, medicine, arts and publishing. Accessing the required and relevant images from large image databases in an efficient manner is now a great necessity.

Clustering is the process of unsupervised classification of data to groups which are called clusters. These data can be data items, observations or feature vectors. Grouped data in each cluster are similar to each other and different from other clusters. Clustering used in many applications and its efficiency is very important as it is considered a basic step in many applications. Clustering can be used for information retrieval, biology, compression, climate, physiology and medicine and business [16].

Traditionally data mining techniques are developed mainly for structured data types and the image data type does not fit in to the structured category. Hence the mining of image data is a challenging problem. The process of Image mining is by extracting meaningful image content and pattern from huge image dataset. The image mining handles extraction of knowledge and relationship among the images for image retrieval, image processing, machine learning databases and data mining. The information representation of an image can be at different levels, namely, pixel, object, semantic concept, and pattern levels [18]. Conventional image mining techniques include object recognition, image retrieval, image indexing, image classification and clustering, and association rule mining.

Image clustering is a process where database of images is given and grouped to clusters using any clustering technique. After clustering, each image in the database has class label, where images with same class label are similar to each other [6]. Main goal of image clustering is to group image data bases or archives to specific number of clusters or groups to extract knowledge or prediction from it, it also provides summarization and visualization for the content of images.

Image clustering is used in many applications as image segmentation, content based image retrieval system, image categorization and unsupervised clustering of image set or database. It is categorized into supervised clustering and unsupervised clustering. The supervised clustering provides collections of pre-classified images. In unsupervised clustering, there are no predefined class label exists for the data points.

Image clustering goal is to organize large image repositories to make them easy to mine and search, to extract knowledge or get useful information from these repositories there are basically three steps which will be discussed in details which are: Pre-processing step which is used to resize, enhances

and improves quality of the image to make it ready for further investigation, Feature extraction of images which is used to extract meaningful information from the image and represent each image with a feature vector to be ready for clustering step, Clustering step which is the final step to group similar images with similar features together [20].

This paper presents a comparative analysis of K-means, K-Medoids and Improved K-Medoids clustering techniques for clustering corel images. This rest of the paper is organized as follows: Section 2 discusses some related work. Section 3 describes the detailed working of K-means, K-Medoids and Improved K-Medoids clustering algorithms. Section 4 analyses the performance of above mentioned clustering algorithms. Section 5 concludes the paper by standing the best clustering algorithm among the three.

II. RELATED WORK

Nanthini et al., (2017) proposed a feature extraction method is implemented by using spectral, SOM and K-means clustering algorithms [8]. All 4 feature extraction techniques are applied to 3 clustering algorithms, totally 12 combinations of image retrieval are done. The experiments results show that the precision values of the proposed combination of texture, color histogram and SIFT feature extraction is better than exclusive histogram, texture and SIFT features separately. Among the three clustering algorithms, it is observed that self-organising map gives high performance than spectral clustering and K-Means clustering algorithms.

Maria Fayez et al., (2016) developed two proposed systems for clustering medical images are implemented on mainly two types of medical images which are X-rays and CT-scans [9]. The two proposed methods can be suitable for other types of medical images. In the first proposed method GLCM was used to extract texture features from the images and k-means clustering algorithm is used to cluster the features extracted, this proposed method gives overall performance of 67.2%. In the second proposed method, 2D wavelet transform texture feature extraction is used to extract feature vector from each medical image, then feature vector is reduced and clustering is applied using K-Means clustering algorithms. The proposed system showed overall performance of 86.8%. This showed that the second proposed method gives overall performance more than the first proposed method.

Seema Wazarkar et al., (2018) presented various image feature extraction and clustering techniques used in various domain for an image analysis [20]. It is carried out and future scope for each domain is provided such as Medical image, 3D imaging, oceanography, industrial automation, remote sensing, mobile phones, security and traffic control are considered applicative areas. The characteristics of an images, clustering approaches for each domain, challenges and future research directions for image clustering are discussed.

III. METHODOLOGY

There are two modules employed in the proposed system. They are feature extraction and image clustering. Fig.1 shows the block diagram of proposed image clustering system.

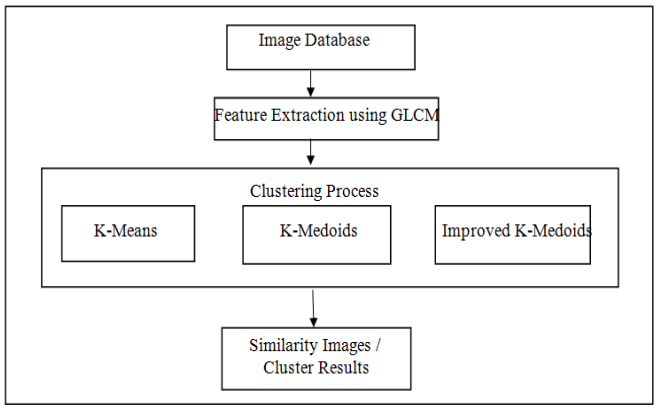

Figure.1 Block Diagram of the Proposed System

A. Corel Images

Image database is a set of image data and those images were taken from UCI repositories in jpeg format. Here Corel datasets are used to evaluate the performance of our method. Corel-1k dataset is classified into ten categories such as bus, buildings, dinosaurs, flowers, etc. The similar images were grouped using feature extraction methods. Clustering methods were applied for grouping the similar images.

B. Feature Extraction using GLCM

The second module of this work deals with feature extraction. Feature extraction is that the method by which certain features of interest inside an image are detected and represented for additional processing. The feature is outlined as a function of one or more measures, every of that specifies some measureable property of an object, and is computed. Texture analysis could be a technique used to measure the patterns in images that are simple for humans to see, but prove more difficult for computers [2].

Texture analysis aims to find distinctive method of indicating the underlying characteristics of textures and represent them in some simpler but unique form, so that they can be used for accurate classification, robust and segmentation of objects [3]. In this paper, Gray level co-occurrence matrix (GLCM) is formulated to find statistical texture features. It provides

the relative frequencies of occurrence of gray level combinations with pairs of image pixels.

The GLCM considers the spatial relationships between two pixels in the image at a time (the reference and the neighbor pixel). The distance between the reference and neighbor pixel can also be chosen [18]. The matrix is built such that each row represents a pixel (reference) in the image and each column represents a pixel (neighbor). The entries of the matrix incorporate the number of the times each gray level in a reference position occurs with each other gray level within the neighbor position. The matrix is then added to its transpose to make a symmetrical matrix. There are 14 kinds of GLCM parameters. Here, seven second order parameters are used namely Energy, Entropy, Contrast, Correlation, Homogeneity, Shade and Prominence. Following are the formulas used for extracting second order parameters.

$$Energy = \sum_{i,j=0}^{N-1} \left(P_{ij}\right)^2 \quad (1)$$

$$Entropy = \sum_{i,j=0}^{N-1} -\ln\left(P_{ij}\right) P_{ij} \quad (2)$$

$$Contrast = \sum_{i,j=0}^{N-1} P_{ij}(i-j)^2 \quad (3)$$

$$Homogeneity = \sum_{i,j=0}^{N-1} \frac{P_{ij}}{1+(i-j)^2} \quad (4)$$

$$Correlation = \sum_{i,j=0}^{N-1} P_{ij} \frac{(i-\mu)(j-\mu)}{\sigma^2} \quad (5)$$

$$Shade = \mathrm{sgn}(A)|A|^{1/3} \quad (6)$$

$$Prominence = \mathrm{sgn}(B)|B|^{1/4} \quad (7)$$

C. Image Clustering

The next module in the proposed work is image clustering. The clustering techniques are used to cluster individual pixels into groups that exhibit homogeneous properties, so that image within each cluster is similar in content. Clustering algorithms provide a useful tool to explore data structures [6]. In this work, three different clustering algorithms are used for image clustering. To select the suitable clustering algorithm for image clustering, K-Means, K-Medoids and Improved K-Medoids partitioning clustering techniques are applied to cluster the extracted features.

(i) K-Means Clustering

K-Means algorithm is the most popular clustering algorithm. It iteratively computes the clusters and their centroids. It is a top down approach to clustering. It is used for creating and analysing the clusters with 'n' number of data points point is divided into 'K' clusters based on the similarity measurement criterion [1]. The results generated using the algorithm mainly depends on initial cluster centroids chosen.

Algorithm

Input:
D= {$t_1, t_2, \ldots, t_n$} // set of elements
K clusters
Output:
K clusters
Algorithm:
assign initial values for means $m_1, m_2, \ldots, m_k$;
repeat
assign each item t_i tothe cluster which has the closest mean;
until convergence criteria is met;

It is an iterative clustering algorithm in which items are stimulated among sets of clusters until the required set is reached. As such, it may be viewed as a type of squared inaccuracy algorithm, although the convergence criteria require not be distinct based on the squared inaccuracy [7]. A high degree of relationship among elements in clusters is obtained, while a high degree of variation among elements in dissimilar clusters is achieved simultaneously.

(ii) K-Mediods Clustering

The K-Medoids algorithm also termed as PAM (Partitioning Around Medoids) algorithm mean a cluster by medoid. Mostly, a random set of k items is taken to be the collection of medoids. Then at every step, all items from the input dataset that are not presently medoids are examined separately to ascertain if they ought to be medoids [18]. That is, the algorithm determines whether or not there is an item that ought to replace one in all the prevailing medoids. Pam is a lot of robust than K-Means within the presence of noise and outliers as a result of a medoid is less influenced by outliers or alternative extreme values than a mean. PAM works efficiently for small data sets, however does not scale well for huge data sets.

Algorithm

Input:
D= {$t_1, t_2, \ldots, t_n$} // set of elements
A // adjacency matrix
K clusters
Output:
K clusters
Algorithm:
arbitrarily select k medoids from D;
repeat
for each t_h not a medoid do

for each medoid t_i do
calculate TC_{ih};
find i, h where TC_{ih} is the smallest;
if $TC_{ih} < 0$, then
replace medoid t_i with t_h;
until $TC_{ih} \geq 0$;
for each $t_i \in D$ do
assign t_i to K_j, where dis(t_i, t_j)is the smallest over all medoids;

(iii) Improved K-Medoids Clustering

In this improved K-Medoids algorithm, the density of each object is calculated first and then the smallest k density values are selected as the initial medoids, which improves the clustering performance [4]. In the improved K-Medoids algorithm, the similarity between pairs of objects is computed once and stored, then a new updating medoids method which employs significantly improves the K-Medoids clustering efficiency. However, the initial medoids optimized by the improved K-Medoids algorithm usually appear in the same cluster, which reduces the final clustering performance [11].

Algorithm

Input:
D= {$t_1, t_2, \ldots, t_n$} // dataset
K // cluster number
Output:
K clusters
Algorithm:
Step 1: Calculate the distance between every pair of all objects based on the chosen dissimilarity measure.
Step 2: Calculate V_j for object j as follows:

$$V_j = \sum_{i=1}^{n} \frac{dist(x_i x_j)}{\sum_{i=1}^{n} dist(x_i x_j)}, \qquad j = 1, \ldots, n$$

Step 3: Sort V_j's in ascending order. Select K objects having the first K smallest values as initial medoids.
Step 4: Obtain the initial cluster result by assigning each object to the nearest medoid.
Step 5: Calculate the sum of distance from all objects to their medoids.
Step6: Find a new medoid of each cluster, which is the object minimizing the total distance to other objects in its cluster. Update the current medoid in each cluster by replacing with the new medoid.
Step 7: Assign each object to the nearest medoid and obtain the cluster result.
Step8: Calculate the sum of distance from all objects to their medoids. If the sum is equal to the previous one, stop the algorithm. Otherwise, go back to the step 6.

IV. RESULTS AND DISCUSSION

The proposed method is implemented in R tool. In this work, experiments are conducted on the feature extraction of GLCM texture feature and three different clustering algorithms are used to cluster the extracted features of the images. In this scheme, 50 images form 5 categories (each category contains 10 images) are shown in Fig.2 tested against the proposed method.

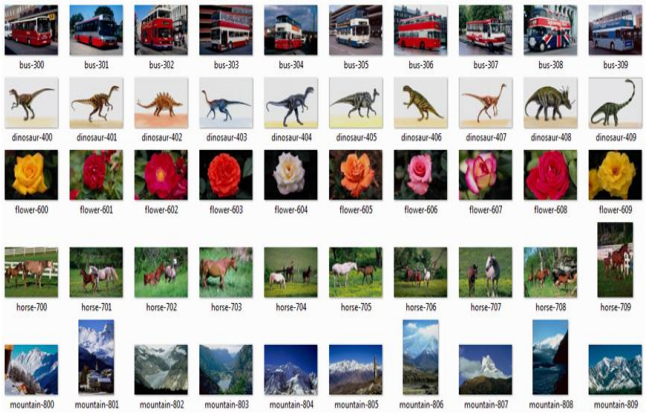

Figure 2. Sample of Unclustered Dataset

The five categories are Bus, Dinosaur, Flower, Horse and Mountain. Each image named its corresponding categories in order to ease the process of accuracy measurement. The accuracy measurement based on the number of images that is probably clustered. An image said to be in proper cluster if its category is the dominant category belong to the cluster. Fig.3 shows the GLCM calculation for images in the dataset and it includes the value of GLCM features.

	glcm_cProminence	glcm_cShade	glcm_contrast	glcm_correlation	glcm_energy	glcm_entropy	glcm_homogeneity1
1	134499.70	1965.14362	10.395935	0.9068979	0.023698721	4.871859	0.6579654
2	99615.30	-1262.13121	6.522345	0.9512134	0.023944007	4.655928	0.6940252
3	134996.17	1155.96606	13.676614	0.8958865	0.013954426	5.266915	0.6131886
4	119056.31	1428.24830	14.196793	0.8769472	0.041031867	4.786314	0.6537533
5	115386.56	-205.74748	7.628268	0.9362207	0.031305851	4.556511	0.7408033
6	132984.81	57.25918	13.397120	0.9004288	0.016200465	5.188061	0.6128639
7	157671.89	694.84922	12.977594	0.9153811	0.024244158	5.114336	0.6223826
8	126028.55	658.89381	18.151215	0.8603737	0.007125877	5.712007	0.4934138
9	158402.53	-929.54096	12.937766	0.9158173	0.010856369	5.388099	0.6126790
10	152038.10	-463.71643	7.836254	0.9488782	0.017510648	4.929853	0.7165965
11	251605.41	-6145.04000	4.593719	0.9473337	0.388586813	2.289306	0.8732326
12	259591.46	-6432.30694	4.414726	0.9555551	0.405747497	2.529405	0.8473467
13	165588.84	-4453.91182	2.625541	0.9701092	0.276361706	2.952152	0.8248065

Showing 1 to 14 of 50 entries

Figure 3. GLCM Value Matrix

The image dataset using the K-Means clustering algorithm, K-Medoids clustering algorithm and Improved K-Medoids clustering algorithm are shown below which returned the five clusters and those results are given in the Fig.4, Fig.5 and Fig.3 respectively.

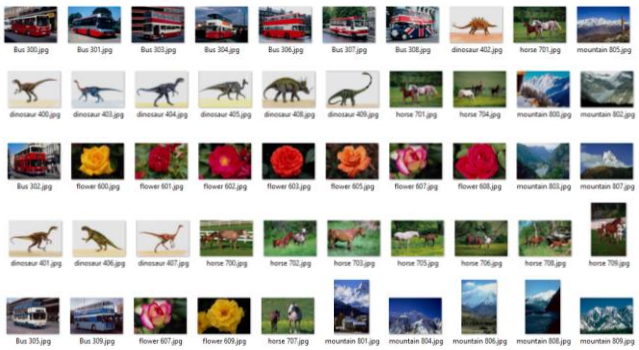

Figure 4. Result of K-Means Clustering

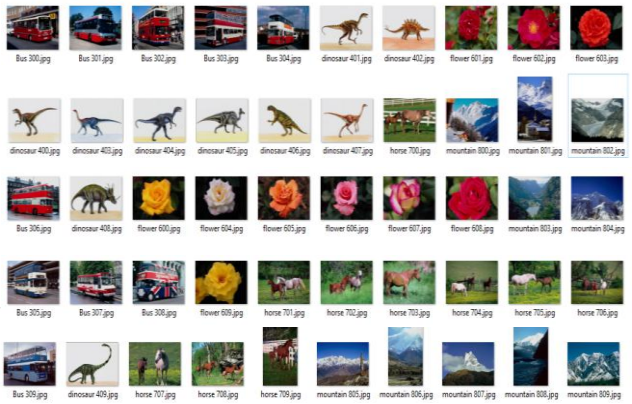

Figure 5. Result of K-Medoids Clustering

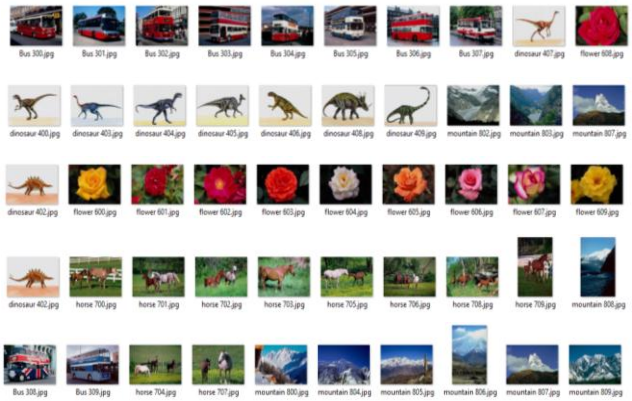

Figure 6. Result of Improved K-Medoids Clustering

Dataset as corel images with pre-labeled classes have been used and then the clustering accuracy measured as the rate of the number of image is correctly clustered form the total number of image, of each category separately based on the features using different clustering algorithms.

$$\text{Accuracy} = (\text{no. of properly clustered images} / \text{total no. of images}) * 100 \quad (8)$$

From the above Equation, measuring the experimental values for the clustering accuracy and it is denoted as 'The clustering accuracy is measured in terms of percentage (%). Higher clustering accuracy ensures the better performance of the method.

Table 1. Tabulation for Clustering Accuracy

Category	**Clustering Accuracy (%)**		
	K- Means	**K-Medoids**	**Improved K-Medoids**
Buses	50%	70%	80%
Dinosaurs	60%	60%	70%
Roses	60%	70%	90%
Horses	60%	70%	80%
Mountains	50%	50%	60%

In this paper, K- Means, K-Medoids and Improved K-Medoids clustering algorithms were applied and compared. The experimental analysis shows Improved K-Medoids clustering algorithm provide better results when compared to K- Means and K-Medoids clustering algorithms is shown in Table 1.

V. CONCLUSION

Clustering techniques are mostly unsupervised methods which will be used to categorize image data into groups based on image similarities. In this paper, the clustering of corel images uses K-Means, K-Medoids and Improved K-Medoids partitioning clustering algorithms had been done. This study compared the efficiency of K-means, K-Medoids and Improved K-Medoids clustering techniques for clustering corel images. From the experiments, it can be concluded that the accuracy of Improved K-Medoids for image dataset is having good evaluation much better than the K-Means and K-Medoids clustering algorithms. A good clustering technique produces high-quality clusters to ensure that the intra-cluster similarity is high and the inter-cluster similarity is low. This work is that the primary step for developing a system for image clustering using clustering techniques.

REFERENCES

[1] Annesha Malakar and Joydeep Mukherjee, "*Image Clustering using Color Moments, Histogram, Edge and K-means Clustering*", International Journal of Science and Research (IJSR), Vol.**2**, No.**1**, **2013.**

[2] Azzam Sleit, Abdel latif Abu dalhoum, Mohammad Qatawneh, Maryam Al-Sharief, Rawa'a Al-Jabaly and Ola Karajeh, "*Image Clustering using Color, Texture and Shape Features*", KSII Transactions on Internet and Information Systems, Vol.**5**, No.**1**, 2**011**.

[3] Dong ping Tain, "*A Review on Image Feature Extraction and Representation Techniques*", International Journal of Multimedia and Ubiquitous Engineering, Vol.**8**, No.**4**, **2013**.

[4] Donghua Yu, Guojun Liu, Maozu Guo and Xiaoyan Liu, "*An Improved K-medoids Algorithm Based on Step Increasing and Optimizing Medoids*", Expert Systems with Applications, **2017**.

[5] Gaurav Mandloi," *A Survey on Feature Extraction Techniques for Color Images*", International Journal of Computer Science and Information Technologies, Vol.**5 (3)**, **2014**.

[6] Kannan. A, Dr.V.Mohan, Dr.N.Anbazhagan, "*Image Clustering and Retrieval using Image Mining Techniques*‖", IEEE International Conference on Computational Intelligence and Computing Research, **2010**.

[7] Khalid Imam Rahmani, Naina Pal and Kamiya Arora, "*Clustering of Image Data Using K-Means and Fuzzy K-Means*", International

Journal of Advanced Computer Science and Applications, Vol.**5**, No.**7**, **2014**.

[8] Nanthini. N, M.Vadivukarassi, N. Puviarasan and P. Aruna, "*Analysis of Clustering Techniques for Retrieval of Images using Proposed Feature Extraction Method*", International Journal of Innovative Research in Computer and Communication Engineering, Vol.**5**, Issue.**3**, **2017**.

[9] Maria Fayez, Soha Safwat and Ehab Hassanein, "*Comparative Study of Clustering Medical Images*", SAI Computing Conference **2016**.

[10] Patil A. J, C.S.Patil, R.R.Karhe and M.A.Aher, "*Comparative Study of Different Clustering Algorithms*", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol.**3**, Issue.**7**, **2014**.

[11] Raghuvira Pratap A, K Suvarna Vani, J Rama Devi and Dr.K Nageswara Rao, "*An Efficient Density based Improved K- Medoids Clustering algorithm*", International Journal of Advanced Computer Science and Applications, Vol.**2**, No.**6**, **2011**.

[12] Rama Kalaivani.E, Suganya. G and Kiruba. J, "*Review on K-Means and Fuzzy C Means Clustering Algorithm*", Imperial Journal of Interdisciplinary Research (IJIR), Vol.**3**, Issue.**2**, **2017**.

[13] SijiT.Mathew and Nachamai M, "*Clustering of Brain MRI Image Using Datamining Algorithm*", International Journal of Advanced Computational Engineering and Networking, ISSN: **2320-2106**, Vol.**3**, Issue.**4**, **2015**.

[14] Sriparna Saha, Abhay Kumar Alok and Asif Ekbal, "*Brain Image Segmentation using Semi-Supervised Clustering*", Expert Systems with Applications, **2016**.

[15] Sukhvir Kaur, "*Survey of Different Data Clustering Algorithms*", International Journal of Computer Science and Mobile Computing, Vol.**5** Issue.**5**, pg. **584-588**, **2016**.

[16] Sukhdev Singh Ghuman, "*Clustering Techniques- A Review*", International Journal of Computer Science and Mobile Computing, Vol.**5** Issue.**5**, pg. **524-530**, **2016**.

[17] Sunil Chowdary, D. Sri Lakshmi Prasanna and P. Sudhakar, "*Evaluating and Analyzing Clusters in Data Mining using Different Algorithms*", International Journal of Computer Science and Mobile Computing, Vol.**3** Issue.**2**, pg. **86-99**, **2014**.

[18] Varsha Kundlikar and Meghana Nagori, "*Image Mining Using Image Feature*", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Vol.**3**, Issue **1**, **2014**.

[19] VikasTondar and Pramod S. Nair, "*A Comparative Study on Clustering Algorithms using Image Data*", International Journal of Computer Applications ISSN: **0975 – 8887**, Vol.**133**, No.**17**, **2016**.

[20] Wazarkar S and B.N. Keshavamurthy, "*A Survey on Image Data Analysis through Clustering Techniques for Real World Applications*", J. Vis. Commun. Image R, **2018**.

[21] Zeynel Cebeci and Figen Yildiz, "*Comparison of K-Means and Fuzzy C-Means Algorithms on Different Cluster Structures*", Journal of Agricultural Informatics, Vol. **6**, No. **3**, **2015**.

Authors Profile

M.Kiruthika received the Bachelor of Computer Science (B.Sc - CT) degree from the Anna University, in 2014 and the Master of Computer Applications (M.C.A.) degree from the Anna University, in 2016. She also recieved the M.Phil degree from the Bharathiar University, Coimbatore, in 2018. She is pursuing Ph.D degree in Computer Science at Bharathiar University. Her research interests include Data Mining.

Dr. S. Sukumaran graduated in 1985 with a degree in Science. He obtained his Master Degree in Science and M.Phil in Computer Science from the Bharathiar University. He received the Ph.D degree in Computer Science from the Bharathiar University. He has 30 years of teaching experience starting from Lecturer to Associate Professor. At present he is working as Associate Professor of Computer Science in Erode Arts and Science College, Erode, Tamilnadu. He has guided for more than 55 M.Phil research Scholars in various fields and guided 13 Ph.D Scholars. Currently he is Guiding 3 M.Phil Scholars and 6 Ph.D Scholars. He is member of Board studies of various Autonomous Colleges and Universities. He published around 80 research papers in national and international journals and conferences. His current research interests include Image processing, Network Security and Data Mining.

