E-ISSN 2581 – 7957
P-ISSN 2277 – 3916

DOI: 10.32377/cvrjst2004

An Efficient Selection of Initial Centroids for K-Means Clustering

S. Bhavani[1] and N. Subhash Chandra[2]
[1]Research Scholar, CVR College of Engineering/CSE Department, Hyderabad, India.
E-mail: bhavanisrirangam122@gmail.com
[2]Professor, CVR College of Engineering/CSE Department, Hyderabad, India.
E-mail: subhashchandra.n.cse@gmail.com

Abstract: **One of the most popular unsupervised clustering algorithms is the K-Means clustering algorithm which can be used for segmentation to analyse the data. It is a centroid-based algorithm, where it calculates the distances to assign a point to a cluster. Each cluster is associated with a centroid. The selection of initial centroids and the number of clusters play a major role to decide the performance of the algorithm. In this context, many researchers worked on, but they may not reach a goal to cluster the images in minimum runtime. Existing histogram based initial centroid selection methods are used on grayscale images only. Two methods, i.e., Histogram based initial centroids selection and Equalized Histogram based initial centroids selection to cluster colour images have been proposed in this paper.**

The colour image has been divided into R, G, B, three channels and calculated histogram to select initial centroids for clustering algorithm. This method validated on three benchmark images and compared to the existing K-Means algorithm and K-Means++ algorithms. The proposed methods give an efficient result compared to the existing algorithms in terms of runtime.

Index Terms: **Histogram, Equalized Histogram, Initial Centroids, K-clusters, K-Means clustering, K-Means++ clustering.**

I. Introduction

In computer vision, image segmentation is the process of partitioning an image into multiple segments. The goal of segmenting an image is to change the representation of an image into something that is more meaningful and easier to analyse. This paper presents segmentation of images using clustering techniques. A cluster refers to a collection of data points aggregated together because of certain similarities. K-Means clustering is one of the simplest and popular unsupervised machine learning algorithms. Assuming the number of clusters k and selecting initial centroids play an important role for clustering in K-Means clustering. For the choice of k, that is, number of clusters, a popular method called "Elbow Method" gives optimum value, but the algorithm has to run several times and then only k value can be decided. There are many methods for initial cluster centre choice like "random data points", "K-Means++". The general procedure to determine the best partition and optimal number of clusters is by validation measures like Sum of Squared Error (SSE) [11], Silhouette Score [10], Calinski_Harabasz_Score [12], Davies_Bouldin_Score [2], Clustering Fitness and Run Time [10].

The aim of this paper is to propose an efficient centroids selection for K-Means clustering based on the histogram peaks that are high density data points to be clustered within a single cluster initially, later the next level density etc. The selection of the centroids is chosen by sorting the histogram. After the selection of centroids, the rest of the process is like a random centroids method.

II. Related Works

In K-Means with random initial centroids method k number of random centroids or initial seeds is selected initially for k number of clusters. The algorithms start calculating the distance between a pixel point and all the centroids, and the pixel will be assigned to the cluster with a minimum distance. Once a new point is assigned, then a new centroid is obtained by taking the mean of all data points of that cluster. This will be continued for all the data points. This procedure will be continued until there is no change in the previous centroid and new centroid for all the clusters [4].

D. T. Pham et. al. [9] says, instead of using a single predefined K, a set of values might be adopted. It is important for the number of values considered to be reasonably large to reflect the specific characteristics of the data sets. At the same time, the selected values have to be significantly smaller than the number of objects in the data sets, which is the main motivation for performing data clustering, but this method could be computationally expensive if it is used with large data sets because it requires several applications of the K-Means algorithm before it can suggest a guide value for K.

Haimonti Dutta et. al. [7] presented a semi-supervised K-Means algorithm but the presence of small and noisy clusters in the data made it difficult to find an agreement in the optimal choice of K and says, more sophisticated seeding techniques incorporating machine learning algorithms are required.

Nameirakpam Dhanachandra et. al. [8] proposed subtractive clustering method on medical images to generate centroids based on the potential value of the data points. Here, the author took the number of clusters, k=3.

Zubair Khan et. al. [13] proposed an adaptive histogram-based approach to determine the initial parameters for K-Means on grey images. Here, authors took the initial parameter as a single variable known as grey level is used to assign intensity values to the pixels. It is a 2-step initial parameter estimation procedure to choose proper number of clusters and optimal initial cluster centres will give a better

Received on 04-09-2021, Revised on 30-09-2021, Accepted on 23-11-2021.

analysis on the data or image from which it can be decided that the K value as well as initial seeding of the algorithm, but the initialization problem of K-Means is used only for grey images. This can be extended for the colour images and the task of grouping the individual peaks in the histograms to represent the true colour and the object boundaries in the image.

Rena Nainggolan et. al. [11] said that, in manual choice of K, the algorithm has to be run many times in order to get efficient clustering results. There are also a few methods like Silhouette Method, Hierarchical clustering method etc., exist for the choice of K.

Raja Kishor Duggirala [7] proposed a fuzzy-based clustering; one can know if data objects fully or partially belong to the clusters based on their membership in different clusters. Hybridization of K-Means with the FCM (Fuzzy C-Means) is implemented for the improvement in performance. Hybrid algorithm of KM and FCM is exhibiting a better performance in terms of execution time of the CPU, Clustering Fitness (CF) and Sum of Squared Error (SSE).

Bernad Jumadi Dehotman Sitompul et. al. [2] proposed the clustering method of determining initial centroid of K-Means algorithm based on minimum Sum of Squared Error, able to improve clustering result and enhance DBI value obtained by simple and determine initial centroid of K-Means algorithm. Here, the authors used numerical data like Seeds Dataset and suggested future work for categorical and image data.D. Arthur, [5] proposed the KMeans++ method, where the centroids to be distant from each other, leading to better results than random initialization. In this method, initially the first cluster centre has been chosen at random from data points, then for the next centre, each data point of the nearest cluster centre is chosen using squared distance method by using the probability formula (1). This step will be repeated until the number of centres has been chosen. The rest of the process is like a random centroids method.

$$C_i = \frac{D(x)^2}{\sum_{x\in X} D(x)^2} \qquad (1)$$

Ahmet Esad TOP et. al. [1] proposed K-Means with a method of Naïve Sharding centroid initialization. The algorithm sorts the pixels in ascending order according to their R, G, B value summations and divides them into shards to choose centroids rather than selecting them randomly, which is the case in the traditional K-Means algorithm. The dataset is then horizontally split into k pieces, or shards. Finally, the original attributes of each shard are independently summed, their mean is computed, and the resultant collection of rows of shard attribute mean values becomes the set of centroids to be used for initialization. Sharding initialization is expected to execute quicker than random centroid initialization, especially considering the time needed for randomly initializing centroids for increasingly complex datasets.

Chunhui Yuan et. al. [3] proposed four kinds of K-value selection algorithms, such as Elbow Method, Gap Statistic, Silhouette Coefficient, and Canopy, are used to cluster the Iris data set to obtain the best K value and the clustering result of the data set. For large scale data sets both time and space complexity will be more to run the Gap Statistic algorithm. The computational overhead can be very large; hence the Silhouette Coefficient algorithm is also not used for large-scale data sets. For large and complex data sets, it is obvious that the Canopy algorithm is the best choice. For real-world multidimensional data containing complex fields of information and for experimental verification, there is a necessity to deeply explore the advantages and disadvantages of each algorithm or to improve the performance of the algorithm.

III. The Proposed System

This section presents our approach to find efficient centroids selection with the application of Histogram based K-Means algorithm on 3 different images.

3.1. Pre-Processing of Image:

Downloaded images are resized and each image is converted into a data frame in order to apply the K-Means algorithm.

3.2. K-Means algorithm:

Given a set of observations (x1, x ... xn), where each observation is a d-dimensional real vector, K-Means clustering aims to partition the n observations into k (≤ n) sets S = {S1, S2, ..., Sk}, so, as to minimize the sum of square error (SSE). Formally, the objective is defined with the equation (2).

$$\arg\min S \sum_{i=1}^{k}\sum_{x\in S_i} \|x-\mu_i\|^2 = \arg\min S \sum_{i=1}^{k} |S_i| Var S_i \qquad (2)$$

Where, $\boldsymbol{\mu}_i$ is the mean of points in S_i, x is any data point and S_i is i^th cluster.

3.2.1. Histogram of an image:

An image histogram is a type of histogram that acts as a graphical representation of the intensity distribution in a digital image [6]. It plots the number of pixels for each intensity value. By looking at the histogram for a specific image, a viewer will be able to judge the entire intensity distribution immediately. For pseudocode refer to the Histogram algorithm.

Algorithm: Histogram

```
Input: Grey image/Single Channel of an Image
Output: Histogram of the given image
Step 1: Declare an n dimensional array of histogram with 256 levels
Step 2: Read the shape of image into a variable
Step 3: For each row
        Step 4: For each column
                Step 5: Increment corresponding histogram level count in an n dimensional array of histogram
Step 6: Return an n dimensional array of histogram
```

3.2.2. Histogram Based K-Means:

The given colour image split into the R, G, B channels and for each channel a separate histogram is generated. Each histogram is sorted in descending order to get a high density of intensity. Based on the given k value, those many or several centroids selected from the sorted list of histograms. The selected centroids are given as input to the K-Means algorithm as initial centroids. For pseudocode refer to the Histogram Based K-Means algorithm.

TABLE I.
COMPARISON OF RUNTIMES IN SECONDS

K	Lena Image				Baboon Image				Peppers Image			
	K-Means with Random Centroids	K-Means ++	Histogram based K-Means	Equalized Histogram based K-Means	K-Means with Random Centroids	K-Means++	Histogram based K-Means	Equalized Histogram based K-Means	K-Means with Random Centroids	K-Means ++	Histogram based K-Means	Equalized Histogram based K-Means
2	0.273	0.373	0.17	0.339	0.378	0.556	0.176	0.365	0.235	0.322	0.152	0.374
3	0.72	0.811	0.216	0.381	0.599	0.704	0.184	360.	0.335	0.536	0.174	0.357
4	0.655	0.929	0.195	0.35	0.77	1.15	0.213	0.384	0.577	0.866	0.276	0.393
5	0.842	1.1	0.172	0.393	1.147	1.004	0.207	0.393	0.75	0.897	0.205	0.385
6	1.5	1.548	0.26	0.486	1.565	1.804	0.226	0.436	1.022	1.047	0.225	0.404
7	1.716	1.891	0.284	0.486	1.901	2.169	0.294	0.517	1.986	1.29	0.247	0.477
8	2.054	2.036	0.382	0.863	2.591	3.646	0.319	0.561	1.875	2.371	3360.	0.916
9	2.838	2.308	0.427	0.801	4.366	3.458	0.571	0.705	3.269	5.343	0.639	1.448
10	3.804	2.838	0.876	0.986	3.577	3.786	0.685	0.924	3.875	3.865	0.583	0.785

Algorithm: Histogram Based K-Means

```
Input: Input Image, Number of Clusters
Output: Clustered Image
Step 1: Split Image into B, G, and R channels
Step 2: Call Histogram for each channel
Step 3: Sort histograms of each channel in descending order.
Step 4: Declare an n dimensional array of Initial Centroids
Step 5: for 0 to k clusters
    Step 6: Append centroids for each channel from sorted histograms to Initial Centroids
    Step 7: Call K-Means with Initial Centroids
    Step 8: Output the Clustered Image
```

3.2.3. Equalized Histogram of an image:

Histogram equalization is a technique for adjusting image intensities to enhance contrast [6]. Let f be given an image represented as an RxC matrix of integer pixel intensities ranging from 0 to L-1. L is the number of possible intensity values, often 256. Let p denote the normalized histogram (probability of intensity) of ‘f’ with a bin for possible intensity. Hence; ‘pn’ is defined with the equation (3).

$$p_n = \frac{number\ of\ pixels\ ith\ intensity\ \ n}{total\ number\ of\ pixels} \qquad (3)$$

Where n=0, 1 ... L-1.

The equalized histogram image g will be defined by the equation (4).

$$g_{i,j} = floor((L-1)\sum_{n=0}^{h_{i,j}} p_n \qquad (4)$$

Where floor () rounds down to the nearest integer. This is equivalent to transforming the pixel intensities, l, of ‘f’ by the equation (5).

$$T(l) = floor((L-1)\ \Sigma_{n=0}^{l} p_n \qquad (5)$$

For pseudocode refer algorithm Equalized Histogram.

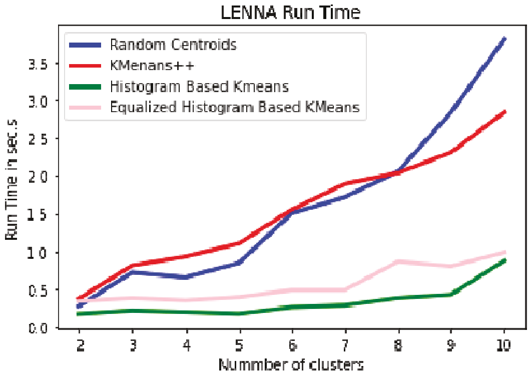


Figure 1. Runtime in Seconds for Lena Image

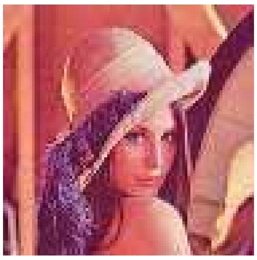


Figure 2. Original Lena Image

E-ISSN 2581 – 7957
P-ISSN 2277 – 3916

DOI: 10.32377/cvrjst2004

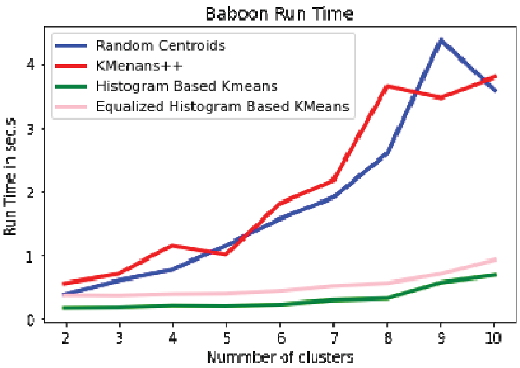


Figure 3. Runtime in Seconds for Baboon Image

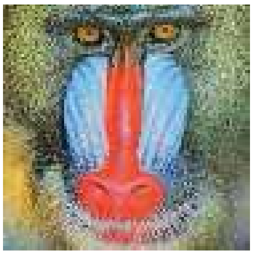


Figure 4. Original Baboon Image

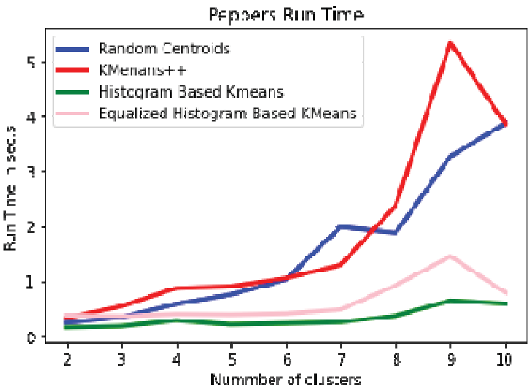


Figure 5. Runtime in Seconds for Peppers Image

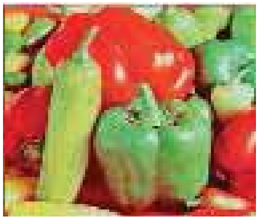
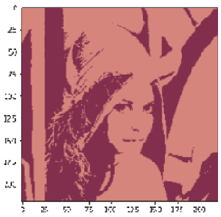
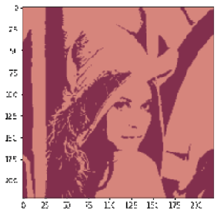
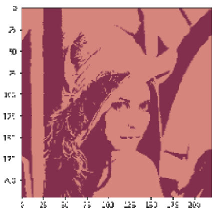
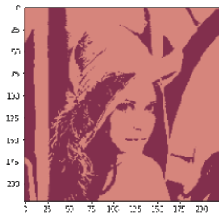
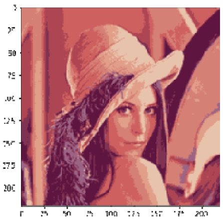
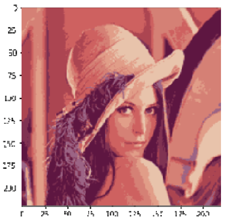
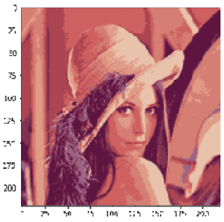
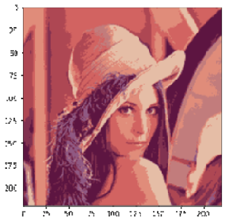


Figure 6. Original Peppers Image

TABLE II.
LENA IMAGE RESULTS

K	K-Means with Random Centroids	K-Means++	Histogram based K-Means	Equalized Histogram based K-Means
2				
8				

Algorithm: Equalized Histogram

```
Input: Grey image/Single Channel of an Image
Output: Equalized Histogram of the given image
Step 1: Call Histogram of channel
Step 2: Read the shape of image into a variable
Step 3: Calculate Probability Distribution Function of Histogram
Step 4: Calculate Cumulative Distribution Function of Histogram
Step 5: Define transformation function
Step 6: For each row
    Step 7: For each column
        Step 8: Apply transformation function and store in resulting an n dimensional array as Equalized Histogram of channel
Step 9: Return Equalized Histogram of channel
```

3.2.4. Equalized Histogram based K-means:

The given colour image split into the RGB channels and for each channel a separate Equalized Histogram is generated. Each Equalized Histogram is sorted in descending order to get high density and intensity values. Based on the given k value, those several numbers of centroids selected from the sorted list of Equalized Histograms. The selected centroids are given as input to the K-Means algorithm as initial centroids. For pseudocode refer Equalized Histogram Based K-Means.

Algorithm: Equalized Histogram Based K-Means

```
Input: Input Image, Number of Clusters
Output: Clustered Image
Step 1: Split Image into B, G, R channels
Step 2: Call Equalized Histogram for each channel
Step 3: Sort histograms of each channel in descending order.
Step 4: Declare an n dimensional array of Initial Centroids
Step 5: For   0 to k clusters
    Step 6: Append centroids for each channel from sorted histograms to Initial Centroids
    Step 7: Call K-Means with Initial Centroids
    Step 8: Output the Clustered Image.
```

TABLE III.
BABOON IMAGE RESULTS

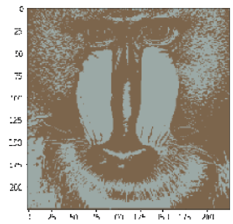
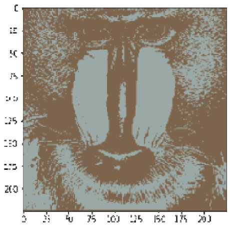
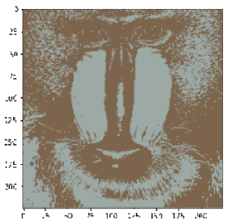
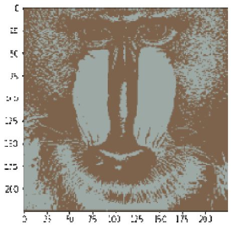
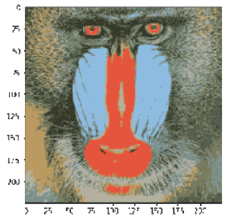
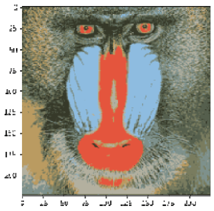
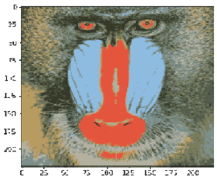
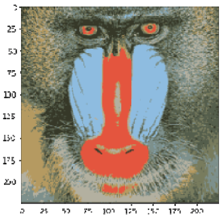
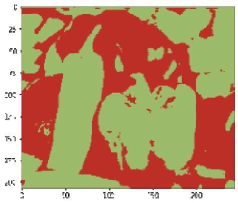
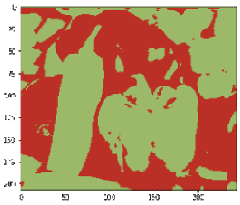
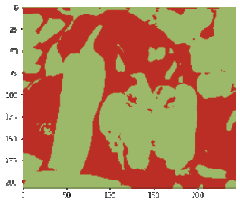
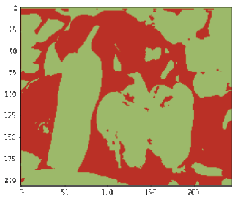
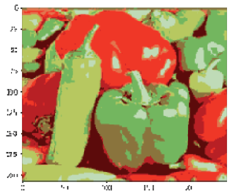
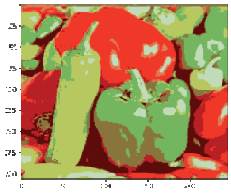
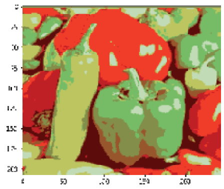
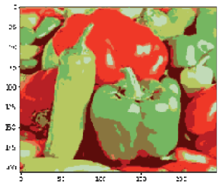
K	K-Means with Random Centroids	K-Means++	Histogram based K-Means	Equalized Histogram based K-Means
2				
8				

TABLE IV.
PEPPERS IMAGE RESULTS

K	K-Means with Random Centroids	K-Means++	Histogram based K-Means	Equalized Histogram based K-Means
2				
8				

IV. Result and Discussion

The proposed system is experimented using https://colab.research.google.com, an open source for python programming.

Performance Analysis:

The present proposed models are experimented on 3 different images and runtimes of all clustering techniques are measured. Comparison of results of runtime in seconds of each clustering technique is listed in Table 1, the corresponding line graphs are also presented in figure 1, figure 3, and figure 5. Original images are presented in figure 2, figure 4, and figure 6. The clustering results for 'k' values 2 and 8 are presented in tables from 2 to 4.

V. Conclusions

Clustering is playing a vital role in image segmentation which is used in many applications. The most commonly used K-Means clustering which takes randomly generated initial centroids is not reaching the local optima. The proposed Histogram based selection of initial centroids to overcome the above drawback. The present paper also proposed equalized histogram-based selection of initial centroids to improve the performance of the algorithm. But, in the above analysis, the equalized histogram method is not performing appreciated results. This can be enhanced by considering spatial values as another dimension in clustering.

It is observed that Histogram based, and Equalized Histogram based K-Means gives better performance in the view of runtime comparatively with Random centroids K-Means and K-Means++. Histogram based K-Means is taking less runtime as compared to the Equalized Histogram based K-Means, but in future, it can be extended that Equalized Histogram based K-Means may perform better in other kind of parameters like Sum of Squared Error, Silhouette Score etc., with the application of spatial dimensions.

References

[1] Ahmet Esad TOP, F. Şükrü TORUN & Hilal KAYA, PARALLEL K-MEANS CLUSTERING WITH NAÏVE SHARDING FOR UNSUPERVISED IMAGE SEGMENTATION VIA MPI, Mühendislik Bilimleri ve Tasarım Dergisi 8(3), 791 – 798, 2020, e-ISSN: 1308-6693, Journal of Engineering Sciences and Design DOI: 10.21923/jesd.748209.

[2] Bernad Jumadi Dehotman Sitompul, Opim Salim Sitompul and Poltak Sihombing, Enhancement Clustering Evaluation Result of Davies-Bouldin Index with Determining Initial Centroid of K-Means Algorithm, Journal of Physics: Conference Series, Volume 1235, The 3rd International Conference on Computing and Applied Informatics 2018 18–19 September 2018, Medan, Sumatera Utara, Indonesia, 1-6.

[3] Chunhui Yuan & Haitao Yang, "Research on K-Value Selection Method of K-Means Clustering Algorithm", (2019), J. 2. 226-235. 10.3390/j2020016, doi: https://doi.org/10.3390/j2020016.

[4] Data Mining – Concepts and Techniques - Jiawei Han & Micheline Kamber, Morgan Kaufmann Publishers, 3rd Edition, 2012.

[5] David Arthur and Serei Vassilvitskii, 2007, k-means++: The Advantages of Careful Seeding, In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, New Orleans, SIAM, pp. 1027-1035.

[6] Digital Image Processing, Rafael C. Gonzalez and Richard E. Woods, Fourth edition. Pearson Education, 2018.

[7] Haimonti Dutta, Rebecca J. Passonneau, Austin Lee, Axinia Radeva, Boyi Xie, David Waltz and Barbara Taranto, Learning Parameters of the K-Means Algorithm from Subjective Human Annotation, Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference, 2011, 465-470.

[8] Nameirakpam Dhanachandra, Khumanthem Manglem, Yambem Jina Chanu, Image Segmentation Using K -means Clustering Algorithm and Subtractive Clustering Algorithm, Procedia Computer Science, Volume 54, 2015, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2015.06.090, 764-771.

[9] Pham, D. & Dimov, Stefan & Nguyen, Cuong, Selection of K in K -means clustering, Manufacturing Engineering Centre, Cardiff University, Cardiff, UK 2004, 103-119, https://doi.org/10.1243/095440605X8298.

[10] Raja Kishor Duggirala, Segmenting Images Using Hybridization of K-Means and Fuzzy C-Means Algorithms, Introduction to Data Science and Machine Learning, Keshav Sud, Pakize Erdogmus and Seifedine Kadry, IntechOpen, (July 10th 2019), 1-27, DOI: 10.5772/intechopen.86374. Available from: https://www.intechopen.com/chapters/68050.

[11] Rena Nainggolan, Resianta Perangin-angin, Emma Simarmata, and Feriani Astuti Tarigan, , Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) optimized by using the Elbow Method, Journal of Physics: Conference Series, 2019, doi: 10.1088/1742-6596/1361/1/012015, 1-6.

[12] Y. Liu, Z. Li, H. Xiong, X. Gao and J. Wu, "Understanding of Internal Clustering Validation Measures," 2010 IEEE International Conference on Data Mining, 2010, pp. 911-916, doi: 10.1109/ICDM.2010.35.

[13] Zubair Khan, Jianjun Ni, Xinnan Fan & Pengfei Shi, An improved K-means clustering algorithm based on an adaptive initial parameter estimation procedure for image segmentation. International Journal of Innovative Computing, Information and Contro, 2017, 1509-1525.