A Comparative Study of K-Means, Fuzzy C-Means and Possibilistic Fuzzy C-Means Algorithm on Noisy Grayscale Images

Sachin Singla
M.Tech IT Student
Department of Computer Science & IT
BBSBEC, Fatehgarh Sahib

Baljit Singh Khehra, PhD
Associate Professor
Department of Computer Science & IT
BBSBEC, Fatehgarh Sahib

ABSTRACT
Clustering is used to arrange the graphic data in the cluster unsupervised learning methods. Clustering is used in the field of image processing to identifying objects that have same features in an image. Clustering can be categorized into Hard and Fuzzy clustering scheme. This article discusses the study of hard clustering based Standard K-Means and different soft (fuzzy) clustering algorithm exits such as Fuzzy C-Means (FCM) and Possibilistic Fuzzy C-Means (PFCM). These algorithms are used to segment and analyse the standard and coloured images but this research work deals with noisy grayscale images. PSNR, MSE and SSIM are used as evaluation parameter to compare the K-Means, FCM and PFCM results. Finally, the experimental results proved that PFCM favorable over FCM and K-Means.

Keywords
Image segmentation, Clustering, K-means, FCM, PFCM.

1. INTRODUCTION
Image segmentation is an essential step behind image understanding and image analysis, such as key to locate objects and boundaries, machine vision and medical imaging. The goal of segmentation is to split an image into a set of separate regions that are visually different, homogenous and meaningful on the basis of some features such as intensity, color and texture etc. Many distinct segmentation techniques have been discovered and can be seen in [1-3]. Image segmentation categorized into four techniques i.e. threshold, edge detection, clustering and region extraction, but this article only discusses image segmentation based clustering method.

Clustering is used to handle similar items based on their individual data in the data set is divided into groups. Assembling data elements into clusters depends upon the principle of maximizing the similarity interlayer and reduces the degree of similarity of things. Data item is assigned to a closet matching cluster based on their minimum distance. Quality of cluster depends upon low inter-class and high intra-class similarity [4]. Clustering can be divided into hard and fuzzy clustering scheme, everyone has their own features. Hard clusters to limit each data point belong to exactly one cluster. As a consequence, using hard clustering is very difficult task where image have poor contrast, overlapping intensities, noise etc.

Another clustering scheme is fuzzy clustering which is based on membership of each data items. In Fuzzy Clustering, Fuzzy C-Means is widely used algorithm where each data item has some degree of membership value which is used to determine the closeness of data item to a cluster [5, 6]. In Fuzzy C- Means each data item may belong to one or more clusters. FCM has problem that it creates the noise points.

To avoid this problem, Krishnapuram and Keller suggested a new fuzzy clusters model named possibilistic c-means (PCM) [7, 8]. PCM uses typicality values rather than membership value but PCM problem with overlapping clusters. The PFCM better clustering algorithm as it has the potential to give more value to membership or typicality values [9]. PFCM inherit the properties of PCM and FCM that often avoids various problems like cluster coincident and noise sensitivity. In rest of the paper section 2 discuss KM, Section 3 does FCM, Section 4 discuss PFCM, Section 5 presents Experimental Results includes some evaluation parameter which compare KM, FCM and PFCM and finally Section 6 has conclusion.

2. K-MEANS (KM)
K-Means is simple unsupervised learning method, which belongs to hard clustering technique [10]. K-means algorithm is used to partition the standard images into two or more clusters [11]. In K-Means, the Euclidean distance is used to calculate the distance between the data item and cluster centroid. Each cluster has its own centroid and data item assign to that cluster which has minimum distance among all the clusters. It is an iterative procedure to find out minimized sum of distances from each data item to its cluster centroid, over all clusters. Main idea behind is to find the K i.e. number of cluster and randomly assigns the centroid value to each cluster. Each centroid value is different from one another. Find the distance between each data item and its centroid and assign each data item to cluster based on minimum distance. Calculate the average of each cluster and selects the average value as new centroid. Iterative process has generated. Note that you are allowed to change the place k prototypes, little by little until beyond a certain number of iterations or changes. The purpose of this algorithm is to minimize the following objective function [12]:

$$J = \sum_{i=1}^{c} \sum_{k=1}^{n} \left\| z_k^{(i)} - v_i \right\|^2 \quad (1)$$

Where in $\left\| z_k^{(i)} - v_i \right\|^2$ is the Euclidean distance calculated between a data point $z_k^{(i)}$ and the centroid v_i, is the distance between the first n data points from their indicator respective central cluster.

Let Z is the dataset, $Z_k = \{z_1, z_2, \dots, z_n\}$ and list of cluster centers represented by $V_i = \{v_1, v_2, \dots, v_c\}$. Algorithmic steps for K-Means:

1. Initialize the number of cluster i.e. c.

2. Randomly set the clusters centroids.
3. For each data point in image:
   a. Calculate the distance between each cluster centroids using Euclidean distance norm.
   b. Move the data closer to the cluster that has minimum distance compare to another.
4. Recalculate the new centroids of clusters using

$$V_i = (1/n)\sum_{i=1}^{n} z_i \quad (2)$$

Here 'n' represents, total number of data items in i^{th} cluster.

5. Repeat step 3, until no more centroid is changing or fixed no. of iteration.

3. FUZZY C-MEANS (FCM)

Fuzzy C-Means clustering is a method that allows data objects belonging to two or more clusters and it is developed by Dunn in 1973 [13]. This is widely used for pattern recognition and standard images like medical, geological and satellite images. In K-Means each data point does or does not belongs to cluster i.e. belongs to only one class of cluster but Fuzzy Clustering extends this notion to assign each data point to two or more clusters based on their membership function. In K-Means each data point has value 0 or 1 but in fuzzy each data point has membership value between 0 and 1 which shows the how much closeness of data point to cluster. It is bound by the sum of information about each data point to the value of all cluster members must be one [14, 15].The membership function is calculated and associates each data item to that cluster who have the highest membership value among all clusters. The purpose of this algorithm is to minimize the following objective function:

$$J(Z;U;V) = \sum_{i=1}^{c}\sum_{k=1}^{n}(\mu_{ik})^m \|z_k - v_i\|^2 \quad (3)$$

Where,

- J is the objective function
- μ_{ik} is the membership value
- m is the fuzziness factor and m value must be greater than 1
- z_k is the k^{th} data point or pixel of image
- v_i is the i^{th} cluster centroid
- $\|z_k - v_i\|^2$ is Euclidean distance

Let Z is the dataset, $Z_k = \{z_1, z_2, \ldots, z_n\}$ and list of cluster centers represented by $V_i = \{v_1, v_2, \ldots, v_c\}$. Algorithmic steps for Fuzzy C-Means:

1. Initialize the number of cluster i.e. c
2. Randomly set the clusters centroids.
3. Randomly assign membership value to $U = [\mu_{ik}]$ between 0 and 1.
4. Calculate the fuzzy membership $[\mu_{ik}]$ using:

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{\|z_k - v_i\|}{\|z_k - v_j\|}\right)^{\frac{2}{m-1}}}, \quad (4)$$

   Here, $1 \le i \le c, 1 \le k \le n$.

5. Calculate the center Vector v_i using:

$$v_i = \frac{\sum_{k=1}^{n}\mu_{ik}^{m} z_k}{\sum_{k=1}^{n}\mu_{ik}^{m}}, \quad (5)$$

   Here, $1 \le i \le c$.

6. Stop if $\|U_{k+1} - U_k\| < \delta$, otherwise go to step 3.

Here 'δ' is termination criteria between [0,1] and $U = [\mu_{ik}]$ is a fuzzy membership matrix.

4. POSSIBILISTIC FUZZY C-MEANS (PFCM)

In 1997, J.C. Bezdek and N.R. Pal suggested a FPCM on which adhesion values of FCM and typicality PCM values with the constraints that the sum of all data points of the typicality of a cluster is the one that leads to unrealistic typicality data values for large data sets [16]. To solve this problem PAL, suggested Possibilistic Fuzzy C-Means, which also overcome the problem of FCM, PCM and K-Means [9,12]. The purpose of this algorithm is to minimize the following objective function:

$$\underbrace{\min}_{(U,T,V)} \{J(Z,U,T,V) = \sum_{i=1}^{c}\sum_{k=1}^{n}\left(a\mu_{ik}^{m} + bt_{ik}^{\eta}\right)\|z_k -$$

vi2 +i=1cγik=1n 1−tikη (6)

Subject to the constraint $\sum_{i=1}^{c}\mu_{ik} = 1 \; \forall \; k$, and $\mu_{ik} \ge 0$, $t_{ik} \le 1$. Here b>0, a>0, η>1, m>1 and γ_i>0 used as user defined constants. Constants a, b is used to define the relative proportion of typicality and membership values. Here, $U = [\mu_{ik}]$ is a membership matrix similar as in FCM and $T = [t_{ik}]$, a typicality matrix similar as in PCM algorithm. If give more importance to membership values than PFCM work closer to FCM algorithm and if give more importance to typicality values than PFCM work closer to PCM. Let b=0 and $\gamma_i = 0$ for all i then Equation 6 becomes equivalent to FCM problem. Now let a=0 then Equation 6 equivalent to PCM problem and let b=0 and $\gamma_i \neq 0$ then Equation 6 also equivalent to FCM problem.

Let Z is a dataset, $Z_k = \{z_1, z_2, \ldots, z_n\}$ and list of cluster centers represented by $V_i = \{v_1, v_2, \ldots, v_c\}$. Set the various parameters b>0, a>0, η>1, m>1. Algorithmic steps for Possibilistic Fuzzy C-Means:

1. Initialize the number of cluster i.e. c.
2. Randomly set the clusters centroids.
3. Run the FCM Algorithm described in section 3.
4. With the help of these results, the penalty parameter γ_i is calculated for every cluster using following equation and put K=1.

$$\gamma_i = K\frac{\sum_{k=1}^{n}\mu_{ik}^{m}\|z_k - v_i\|^2}{\sum_{k=1}^{n}\mu_{ik}^{m}} \quad (7)$$

5. Calculate the membership values $U = [\mu_{ik}]$ if distance between image pixel point and centroid is greater than 0. Membership values calculated using (4).

   Here, $1 \le i \le c, 1 \le k \le n$.

6. Calculate the typicality values T= $[t_{ik}]$ if distance between image pixel point and centroid is greater than 0. Typicality values calculated using following equation:

$$t_{ik} = \frac{1}{1 + \left(\frac{b}{\eta}\|z_k - v_i\|^2\right)^{\frac{1}{(\eta-1)}}}, \quad (8)$$

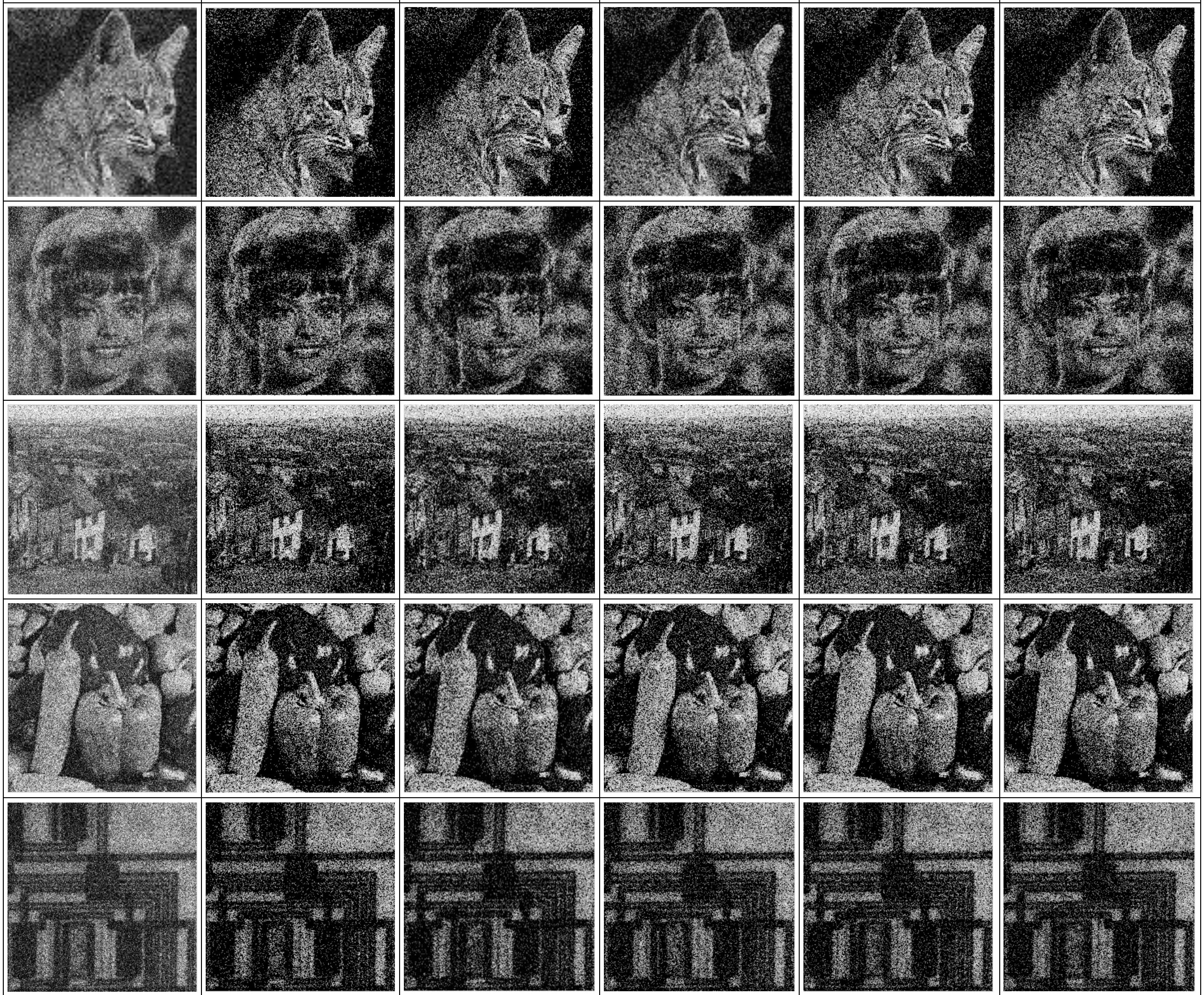
7. Calculate the center Vector v_i using:

$$v_i = \frac{\sum_{k=1}^{n}\left(a\mu_{ik}^{m} + bt_{ik}^{\eta}\right)z^k}{\sum_{k=1}^{n}\left(a\mu_{ik}^{m} + bt_{ik}^{\eta}\right)} \quad (9)$$

8. Stop, if the error is less or equal to $\|V_{k+1} - V_k\| < \delta$, otherwise go to step 6.

Now, it is possible to identify the cluster by using both typicality and membership values.

Original Image	KM result	FCM Result (For m=2)	PFCM Results(m=2,η=2)		
			For a=1 and b=1	For a=2 and b=1	For a=1 and b=2



a)	b)	c)	d)	e)	f)

Figure 1. Comparison of algorithmic results for different noisy grayscale images named *cat, zelda, house, pepper* and *circuit*, a) The original Images, b) KM results, c) FCM results, (d-e) PFCM results for various parameter values.

5. EXPERIMENTAL RESULTS

In this paper, several noisy gray scale images are tested in order to show results obtained from clustering algorithm. Images are taken from database of static images. Noise are added to original gray scale images with the help of default Gaussian noise function of matlab. Matlab has also function of parameters PSNR, MSE and SSIM for checking the image quality of grayscale images [17]. On the basis of above parameter this research concludes which clustering algorithm is better for segmentation of noisy grayscale images.

In this work, the size of noisy grayscale images are 255 * 255 and tested with different initial condition for K-Means, FCM and PFCM algorithm is given below.

5.1 K-Means

- Total No. of cluster taken: 2.
- Centroid (random): (0.13 0.564).
- Max Iteration: 100.

5.2 FCM

- For Total No. of cluster taken: 2.
- Centroid(random): (0.13 0.564).
- Max Iteration: 100.
- Membership matrix initialize randomly.
- Weighting component m=2, Epsilon $\delta = 0.0001$.

5.3 PFCM

- Total No. of cluster taken: 2.
- Centroid (random) : (0.13 0.564).
- Max Iteration: 100.

- Membership and typicality matrix initialize randomly and Epsilon δ=0.0001.
- PFCM checked for various parameter values are :
  - m=2, η=2, a=1, b=1.
  - m=2, η=2, a=2, b=1.
  - m=2, η=2, a=1, b=2.

After applying the K-means, FCM and PFCM algorithm on various noisy gray scale images output shown in Figure 1.

5.4 Evaluation Parameters

5.4.1. MSE

Mean Squared Error is calculated pixel by pixel by summing the squared differences of all pixels and dividing by the number of total pixels. Let's assume an Image $P = \{p_1, p_2, \dots\dots, p_m\}$ and Image Q= $\{q_1, q_2, \dots\dots, q_m\}$ with 'm' no. of pixels then,

$$MSE(P,Q) = \frac{1}{m}\sum_{i=1}^{m}\|p_i - q_i\|^2 \tag{9}$$

Lesser the MSE value, the better the image quality. Table 1 shows the computed results for K-Means, FCM and PFCM between truth image and reconstructed image for respective algorithm.

5.4.2 PSNR

Peak Signal-to-Noise Ratio is described as maximum value of maximum signal power with respect to MSE assumed as noise power, given in decibels (dB). Let's assume an Image $P = \{p_1, p_2, \dots\dots, p_m\}$ and Image Q= $\{q_1, q_2, \dots\dots, q_m\}$ with 'm' no. of pixels then PSNR is given by,

$$PSNR(P,Q) = 10Log_{10}\left(\frac{Max\ Signal\ Power^2}{MSE(P,Q)}\right) \tag{10}$$

For 8 bit gray scale image max value is 255. Higher Peak Signal to Noise Ratio value, the better the image quality. Table 2 represents the computed results for K-Means, FCM and PFCM between truth image and reconstructed image for respective algorithm.

5.4.3 SSIM

The Structural Similarity Index Measuring the structural similarity between truth and reconstructed image [21]. Its value between -1 and 1.When the both images are equal then SSIM close to 1.

The SSIM assessment index based on computation of three terms i.e. luminance, contrast and structural term. Then overall SSIM index as a multiplicative of above three terms as follow:

$$SSIM(P,Q) = \frac{(2\mu_P\mu_Q + K_1)(2\sigma_{PQ} + K_2)}{(\mu_P^2 + \mu_Q^2 + K_1)(\sigma_P^2 + \sigma_Q^2 + K_2)} \tag{11}$$

Where, $K_1 = (0.01 * L).\hat{}2$ and $K_2 = (0.03 * L).\hat{}2$ are regularization constants and L is dynamic range value, μ_P, μ_Q are the local means and σ_P, σ_Q are the standard deviation, σ_{PQ} are cross variance.

Higher the SSIM value, the better the image quality. In this work, matlab Function SSIM are used for checking the quality of images. Table 3 shows the computed results for K-Means, FCM and PFCM between truth image and reconstructed image for respective algorithm.

Table (1-3) shows PFCM have lower values for MSE, higher values for PSNR and SSIM which proves that PFCM produces the better results for noisy gray scale images.

Table 1. MSE values for segmented images by KM, FCM and PFCM algorithm

Original Image Name	K-Means	FCM (m=2)	PFCM (m=2,η=2)		
			For a=1 and b=1	For a=2 and b=1	For a=1 and b=2
Cat	0.0321	0.0293	0.0223	0.0231	0.0224
Zelda	0.0350	0.0327	0.0281	0.0283	0.0285
House	0.0487	0.0461	0.0409	0.0428	0.0420
Pepper	0.0374	0.0355	0.0316	0.0316	0.0307
Circuit	0.0284	0.0260	0.0191	0.0205	0.0192

Table 2. PSNR values for segmented images by KM, FCM and PFCM algorithm

Original Image Name	K-Means	FCM (m=2)	PFCM (m=2,η=2)		
			For a=1 and b=1	For a=2 and b=1	For a=1 and b=2
Cat	63.2860	63.4675	64.6429	64.4870	64.6232
Zelda	62.6922	62.9790	63.6491	63.6197	63.5750
House	61.2559	61.4936	62.0107	61.8198	91.9016
Pepper	62.4026	62.6247	63.1403	63.1390	63.2545
Circuit	63.5965	63.9811	65.3209	65.0041	65.3025

Table 3. SSIM values for segmented images by KM, FCM and PFCM algorithm

Original Image Name	K-Means	FCM (m=2)	PFCM (m=2,η=2)		
			For a=1 and b=1	For a=1 and b=1	For a=1 and b=1
Cat	0.9975	0.9977	0.9983	0.9982	0.9983
Zelda	0.9973	0.9974	0.9978	0.9977	0.9978
House	0.9959	0.9961	0.9965	0.9964	0.9965
Pepper	0.9969	0.9971	0.9975	0.9974	0.9975
Circuit	0.9977	0.9978	0.9984	0.9984	0.9985

6. CONCLUSION

In this work, a Gaussian noise function is used to add the artificial noise into different gray scale images in order to prove that PFCM provide the better result on noisy gray scale images than K-Means and FCM. After analysis of the obtained results of PSNR, MSE and SSIM on noisy gray scale images shows that PFCM is effective and more robust. Results obtained from PFCM algorithm is nearer to FCM and K-Means algorithm and it require more computational time as compare to K-Means and FCM. So, in future work need to develop new methodology or to improve which is good for noisy gray scale images and provide better quality of images.

7. ACKNOWLEDGMENTS

The author would like to give sincerest gratitude towards all faculty and staff of Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib for their valuable advice and healthy criticism throughout for the research work.

8. REFERENCES

[1] Gonzalez, R.C., Woods, R.E. and Eddins, S.L., 2004. Digital image processing using MATLAB.

[2] Fu, K.S. and Mui, J.K., 1981. A survey on image segmentation. Pattern recognition, 13(1), pp.3-16.

[3] Pal, N.R. and Pal, S.K., 1993. A review on image segmentation techniques. Pattern recognition, 26(9), pp.1277-1294.

[4] Ghosh, S. and Dubey, S.K., 2013. Comparative analysis of k-means and fuzzy c-means algorithms. International Journal of Advanced Computer Science and Applications, 4(4), pp. 34-39.

[5] Bezdek, J.C., 2013. Pattern recognition with fuzzy objective function algorithms. Springer Science & Business Media.

[6] Park, D.C., 2010, April. Intuitive fuzzy C-means algorithm for MRI segmentation. In Information Science and Applications (ICISA), 2010 International Conference on (pp. 1-7). IEEE.

[7] Krishnapuram, R. and Keller, J.M., 1993. A possibilistic approach to clustering. Fuzzy Systems, IEEE Transactions on, 1(2), pp.98-110.

[8] Almeida, R.J. and Sousa, J.M.C., 2006, September. Comparison of fuzzy clustering algorithms for classification. In Evolving Fuzzy Systems, 2006 International Symposium on (pp. 112-117). IEEE.

[9] Pal, N.R., Pal, K., Keller, J.M. and Bezdek, J.C., 2005. A possibilistic fuzzy c-means clustering algorithm. Fuzzy Systems, IEEE Transactions on, 13(4), pp.517-530.

[10] Santhi, M.V.B.T., Sai Leela, V.R.N., Anitha, P.U. and Nagamalleswari, D., 2011. Enhancing K-Means Clustering Algorithm. International Journal of Computer Science & Technology, IJCST, 2(4), pp.73-77.

[11] Dehariya, V.K., Shrivastava, S.K. and Jain, R.C., 2010, November. Clustering of image data set using K-means and fuzzy K-means algorithms. In Computational Intelligence and Communication Networks (CICN), 2010 International Conference on (pp. 386-391). IEEE.

[12] Quintanilla-Dominguez, J., Ojeda-Magaña, B., Cortina-Januchs, M.G., Ruelas, R., Vega-Corona, A. and Andina, D., 2011. Image segmentation by fuzzy and possibilistic clustering algorithms for the identification of microcalcifications. Scientia Iranica, 18(3), pp.580-589.

[13] Dunn, J.C., 1973. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters.

[14] Zadeh, L.A., 1965, Fuzzy sets, Information and control, 8(3), 338-353.

[15] Yang, Y. and Huang, S., 2012. Image segmentation by fuzzy C-means clustering algorithm with a novel penalty term. Computing and Informatics, 26(1), pp.17-31.

[16] Pal, N.R., Pal, K. and Bezdek, J.C., 1997, July. A mixed c-means clustering model. In Fuzzy Systems, 1997., Proceedings of the Sixth IEEE International Conference on (Vol. 1, pp. 11-21). IEEE.

[17] AI-Najjar, A.Y. and Soong, D. C., 2012. Comparison of image quality assessment: PSNR, HVS, SSIM, UIQI. International Journal of Scientific a& Engineering Research, 3(8), p.1.