Automatic Centroids Selection in K-means Clustering Based Image Segmentation

A. Pugazhenthi, Jyoti Singhai

Abstract— **This paper proposes a K-means clustering based image segmentation algorithm which select the centroids automatically. It eliminates the limitations associated with K-means clustering such as selection of initial centroids and dead centers. As image histogram is one of the best ways to represent the distribution of image gray levels, the proposed approach selects centroids as the gray levels corresponding to the peaks of the image histogram. With these initial centroids, K-means clustering is performed. The result is post processed by some morphological operations. The proposed algorithm uniformly segments the regions of interest over randomly selected centroids. The performance of proposed algorithm and random centroids selection is compared with some validity parameters like SSIM, MSE, PSNR, IF, SC and CC. Comparison with the existing algorithms confirms the improvement in qualitative parameters. The tool used in this work is MATLAB R2012a.**

Index Terms— **Image Segmentation, Clustering, K means Clustering, Segmentation Quality Measure.**

I. Introduction

IMAGE segmentation is one of the primary steps in image analysis for object identification. Image segmentation [1] is the process of partitioning the image into a set of meaningful objects or regions. Therefore segmentation techniques divide the digital image into meaningful homogeneous regions with respect to intensity, color or texture. Image segmentation plays an important role in content based image retrieval, machine vision, object detection, traffic control systems and others. Thresholding, edge based, region based, and morphological watershed are some basic segmentation techniques [1]. Clustering [3] is one of the widely used image segmentation techniques in several applications. It is an unsupervised classification which has a variety of applications like image segmentation, object recognition, document retrieval, data mining, and many others. Clustering is the process of grouping objects such that the objects in a group are more similar than the objects in other groups. A good cluster should be compact and separable which means minimum inter cluster distance and maximum intra cluster distance. Inter cluster distance is the distance between the objects and centroid within the cluster. Intra cluster distance is the distance between the cluster centroids. The clustering algorithms [3] are mainly fallen into hierarchical and partitional clustering. In hierarchical clustering the whole image is considered as a single cluster where as in partitional clustering each pixel in the image is considered as a cluster. There are many partitional clustering algorithms such as K-means, ISODATA [6], and Fuzzy c-means (FCM) [5] that are used for image segmentation. K-means algorithm [7] is most widely used owing to its simplicity, quick convergence and handling of large data set. The K-means algorithm has the following limitations.

1. Number of clusters needs to be known.
2. Segmentation depends strongly on initial centroids.
3. Due to bad selection of centroids Dead centers may occur, where the cluster has no members.

Plenty of research has been done to remove the above limitations in K-means clustering. K. A. Abdul Nazeer et al. (2009) [15] computed the distances between each data point and all other data points in the set of data points, based on this they made k sets. Finally the initial centroids were obtained based on minimum distance. Ran Vijay Singh et al. (2011) [16] partitioned the whole data set into different segments and calculates the frequency of data point in each segment. The segment which shows maximum frequency of data point was assigned as the centroid of that cluster. Md. Sohrab Mahmud et al. (2012) [17] proposed improved algorithm in which they found a weighted average score of dataset. Then the entire set of data points was sorted. The sorted list of data points were then divided into k subsets. The nearest possible value of mean from each dataset becomes the initial centroids of the cluster to be constructed. Hong Yao et al. (2012) [2] proposed a random selection of centroids such that the arithmetic mean of centroids was closer to Otsu threshold. Each algorithm is capable of removing two of those three limitations. None of the above algorithms removed all the limitations.

Remaining sections is organized as follows. Section II presents the existing K-mean algorithm and Hong yao's algorithm. Section III explains the proposed algorithm. Experimental results and qualitative analysis are mentioned in section IV. We conclude section V by discussing the results.

A. Pugazhenthi is with the Electronics and Communication Engineering Department, G Pulla Reddy Engineering College, Kurnool, India (e-mail: pugazh1989@ gmail.com).

Dr. Jyoti Singhai is with Electronics and Communication Engineering Department, Maulana Azad National Institute of Technology, Bhopal, India (e-mail: j.singhai@ gmail.com).

978-1-4799-3358-7/14/$31.00 ©2014 IEEE

II. Existing Algorithms

A. K-means Clustering Algorithm

The K-means Clustering algorithm [9] is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. It is a non-fuzzy clustering method which allows one pixel to be a member of one cluster in any one time. It is an iterative technique that is used to partition all the pixels in an image into K groups of clusters and each cluster is characterized by its centroid.

Let X= {x_1, x_2, x_3…x_n} represent a set of pixels of the given image, where n is the total number of pixels in the image. C= {c_1, c_2, c_3…c_k} represent a set of centroids, where k is the number of clusters. The main objective of K-means algorithm is to minimize an objective function J that determines the closeness between the pixel and the cluster centroids, and is calculated as follows,

$$J=\sum_{i=1}^{n}\sum_{j=1}^{k}||x_{ij}-c_j||^2 \tag{1}$$

Where c_j is the j^{th} cluster centroid and x_{ij} is the i^{th} pixel in j^{th} cluster. The detailed K-means algorithm is as given below,

1. Randomly select k centroids.

2. Assign each pixel to its nearest centroid by calculating argmin {| x_i - c_j |}, i=1, 2, 3 …n, j=1, 2, 3 …k

3. Compute the new cluster centroids $C=\{c^*_1, c^*_2, c^*_3....c^*_k\}$

$$c^*_i=\frac{1}{n_i}\sum_{x_j\in c_i} x_j \tag{2}$$

Where, n_i is the number of pixels belonging to the cluster of centroid c_i and x_j is the pixels belonged to cluster c_i .

4. If old centroids=new centroids, then stop, otherwise repeat from step 2.

The time complexity [3] of the algorithm is O(nkl) where n is the number of pixels in the image, k is the number of cluster centroids and l is the number of iterations taken by the algorithm to converge. However, the K-means algorithm suffers from several disadvantages; such as its sensitivity to initial number of clusters, random selection of initial centroids and dead centers, which could lead to poor segmentation. Different number of clusters and combination of centroids will give different segmentation result.

A. Hong Yao's Algorithm

Hong yao et al. proposed an algorithm where the number of clusters is determined automatically [2] from the image histogram peaks. If the image histogram has k peaks, then the number of clusters in K-means algorithm is k. The detailed algorithm is as given below,

1. Firstly, smooth the image histogram, then do the derivative operation and get all peaks and valleys of the image histogram.

2. Calculate the distance threshold value which is the average distance between each pair of adjacent peak and valley points. Represent this threshold as T1.

$$T_1=\frac{1}{M}\sum_{i=1}^{M}Dis(i) \tag{3}$$

Where, Dis(i) is the distance between adjacent peak and valley points, and M is the number of the distance measures. If one peak was selected then the next peak will occur after the threshold (T_1) and the remaining peaks in between them were going to be discarded.

3. Thus the obtained k peaks represents the number of clusters.

III. Proposed approach

As image histogram is one of the best ways to represent the distribution of image gray levels, in the proposed approach the centroids are chosen as the gray levels corresponding to the peaks of the image histogram. As the peaks are selected as centroids the dead centers problem will not occur. Steps to be followed to calculate the centroids as follows,

1. Select the input image.

2. Select number of clusters k from Hong Yao et al. algorithm.

3. Select the initial centroids as the gray levels corresponding to the peaks of the image histogram.

4. Calculate the Otsu threshold of the image proposed by Nobuyuki Otsu [8], which will segment the object from the background and represent that as threshold (T2).

5. Calculate the arithmetic mean (M) of centroids.

$$M=\frac{C_1+C_2+...+C_k}{k} \tag{4}$$

Where k is the number of clusters and C_1, C_2 … C_k are centroids.

6. If |M- T_2|<= T_1, then stop and finalize those gray values as centroids, otherwise go for second maximum in between two centroids. Where T_1 is Hong Yao's distance threshold, T_2 is Otsu's threshold and M is the arithmetic mean of centroids.

7. Perform K-means clustering.

8. Extract the objects in every cluster and add some of them which belonged to the same object needed to be segmented.

9. There are still some noise points. Median filter of order three is used to remove those background noises.

10. Region filling morphological operation is carried out to fill the holes present in the object, where dark pixels are surrounded by lighter pixels.

11. Area opening is done and the final image is multiplied with the original image.

12. Finally the segmented result we got.

Figure 1 shows the flow chart of proposed algorithm.

IV. Results and Analysis

The algorithm is implemented in Matlab version 7.14.0.739(R2012a) on Windows7 Ultimate on core(TM) 2 Duo processor. The proposed segmentation algorithm is tested

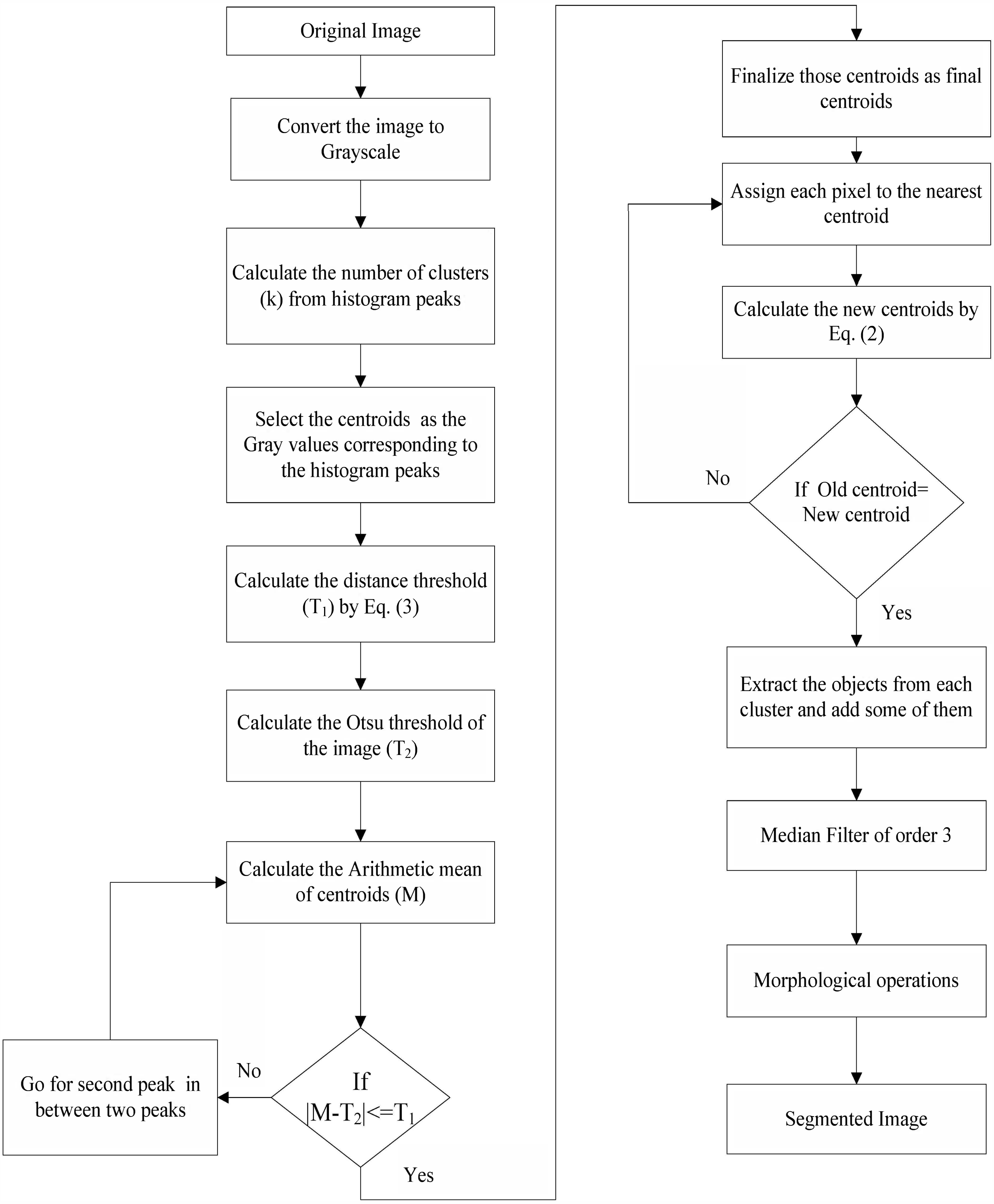
Original Image
Convert the image to Grayscale
Calculate the number of clusters (k) from histogram peaks
Select the centroids as the Gray values corresponding to the histogram peaks
Calculate the distance threshold (T_1) by Eq. (3)
Calculate the Otsu threshold of the image (T_2)
Calculate the Arithmetic mean of centroids (M)
If $|M-T_2|<=T_1$
No
Go for second peak in between two peaks
Yes
Finalize those centroids as final centroids
Assign each pixel to the nearest centroid
Calculate the new centroids by Eq. (2)
If Old centroid= New centroid
No
Yes
Extract the objects from each cluster and add some of them
Median Filter of order 3
Morphological operations
Segmented Image

Fig 1. Flow chart of Proposed Algorithm

on set of seven images i.e. Agassi, Pop, Ball, Fish, Tulip, Hand and Flower. Figure 2 shows the comparisons of original images, segmented images by Otsu's threshold, K-means clustering by random selection of centroids and proposed algorithm.

The agassi.jpg is a color image of size 271x186 that contains single object and two different types of background. The

pop.jpg is a color image of size 290x174 that has one object and three different types of background. The ball.jpg is a color image of size 400x300 that has three objects and a textured background. The fish.jpg is a color image of size 393x224 that contains single object and dark background. The tulip.jpg is a color image of size 400x300 that has two objects and the background contains leafs. The hand.jpg is a color image of size 259x194 that consists of two objects and two different types of background. The flower.jpg is a monochrome image of size 200x200 that has one object and dark background with small variation.

A. *Qualitative Image Analysis*

Qualitative analysis is one of the widely used procedure in which the probabilistic statement about the algorithm's effectiveness and weakness is carried out on the basis of human visual perception. The qualitative analysis is conducted to understand, whether the algorithm segments desired regions of interest from unwanted backgrounds or not. The segmentation comparison of proposed algorithm with the existing algorithms is given in Figure 2.

B. *Quantitative Image Analysis*

Quantitative analysis is a numerically oriented procedure to figure out the performance of algorithms without any human error. It is a digital metric that quantifies how the segmented image appears relative to the original version of the image. The quality of segmented images obtained with the proposed algorithm is evaluated quantitatively by calculating some statistical parameters like Structural Similarity index (SSIM), Image Fidelity (IF), Structural content (SC), Mean Squared Error (MSE), Peak Signal to Noise ratio (PSNR), Correlation Coefficient (CC) as mentioned in [11]-[13] and compared with existing algorithm. MSE, PSNR are pixel difference based measures and easy to evaluate. Assume an MxN image, where M = number of rows and N = number of columns. Let $F_1(i,j)$ represents the original image and $F_2(i,j)$ represents the segmented image, the segmentation quality measures are defined as given below.

1) Structural Similarity Index (SSIM): The structural similarity (SSIM) index [12] is a method for measuring the similarity between two images. Structural information in an image is defined as those attributes that represent the structure of objects in the image. The structural information is used to represent the structure of objects in the image. SSIM compares the structures of original and segmented images. The SSIM is computed within a local square window, which moves pixel-by-pixel over the entire image. At each step, the local statistics and SSIM index are calculated within the local window. Mathematically SSIM is defined by using the following formula

$$\text{SSIM}=\frac{(2\mu_x\mu_y+c_1)(2\sigma_{xy}+c_2)}{(\mu_x{}^2+\mu_y{}^2+c_1)(\sigma_x{}^2+\sigma_y{}^2+c_2)} \qquad (5)$$

Where μ_x, μ_y and σ_x, σ_y represent the mean and standard deviation of original image and segmented image. Desired value of SSIM is closer to 1.

2) Peak Signal to Noise Ratio (PSNR): The peak signal-to-noise ratio (PSNR) [12] is a measure of quality of segmented image based on pixel difference between original and segmented images. PSNR is measured in decibels (dB). Mathematically PSNR is defined by using the following formula

$$\text{PSNR}=10\log\frac{255^2}{\text{MSE}} \qquad (6)$$

Where MSE represents Mean Squared Error. Mean Squared Error (MSE) [12] is the error between the original image and segmented image. It is computed by averaging the squared intensity difference of original and segmented image pixels, and mathematically it can be defined as

$$MSE=\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}[F_1(i,j)-F_2(i,j)]^2 \qquad (7)$$

Desired value of PSNR is as high as possible. Desired value of MSE is as low as possible.

3) Image Fidelity (IF), Structural Content (SC): Image fidelity (IF) [11] refers to the ability of a process to provide an image accurately, without any visible distortion or information loss. Image fidelity metric quantifies the distortion of a processed image relative to its original image. Structural Content (SC) measures how the structural content of the object is preserved in the segmented image when compared to the original image. This metric measures the similarity between two images.

$$IF=1-[\sum_{i=1}^{M}\sum_{j=1}^{N}[F_1(i,j)-F_2(i,j)]^2 \Big/ \sum_{i=1}^{M}\sum_{j=1}^{N}[F_1(i,j)]^2] \qquad (8)$$

$$SC=\sum_{i=1}^{M}\sum_{j=1}^{N}[F_1(i,j)]^2 \Big/ \sum_{i=1}^{M}\sum_{j=1}^{N}[F_2(i,j)]^2 \qquad (9)$$

The dynamic range of IF is [0, 1]. The maximum value occurs when both images are identical. Desired value of IF, SC is closer to 1

4) Correlation Coefficient(CC): Correlation Coefficient (CC) [13] measures how the segmented image is correlated to the original image. Mathematically CC is defined by using the following formula

$$\rho_{x,y}=\frac{\text{cov}(x,y)}{\sigma_x\sigma_y}=\frac{E[(x-\mu_x)(y-\mu_y)]}{\sigma_x\sigma_y} \qquad (10)$$

Where μ_x, μ_y and σ_x, σ_y represent the mean and standard deviation of original image and segmented image. The best value of CC is closer to 1.

TABLE I

QUALITATIVE ANALYSIS OF RANDOM SELECTION OF CENTROIDS AND CENTROIDS SELECTED BY PROPOSED APPROACH

Image	**Mode of centroids selection**	**Initial centroids**	**SSIM**	**MSE**	**PSNR**	**IF**	**SC**	**CC**
Agassi.jpg (k=13)	Randomly selected	60,70,100,120,140,150,160,180, 200,220,235,245,250	0.8516	2991.90	13.3713	0.7843	1.2644	0.9138
		49,71,95,125,130,155,170,190, 210,225,234,249,253	0.8518	2987.70	13.3774	0.7846	1.2639	0.9149
		50,58,89,127,139,147,168,172, 194,223,229,231,244	0.8513	2996.20	13.3651	0.7840	1.2649	0.9137
		34,54,89,124,134,144,164,184, 194,204,224,234,254	0.8519	2986.20	13.3796	0.7847	1.2638	0.9140
		27,47,92,117,147,167,177,187, 217,227, 231,247	0.8518	2987.70	13.3774	0.7846	1.2639	0.9140
	Proposed algorithm	**2,16,51,65,103,125,152,166, 190,210,222,235,255**	**0.8928**	**2154.30**	**14.7977**	**0.8429**	**1.1773**	**0.9445**
Pop.jpg (k=9)	Randomly selected	30,55,86,111,160,200,210,222,240	0.9469	2925.40	13.4690	0.8414	1.1804	0.8831
		44,69,85,125,169,194,200,229,240	0.9429	3085.50	13.2376	0.8331	1.1922	0.8777
		11,40,79,120,167,199,220,230,250	0.9451	2997.60	13.3631	0.8377	1.1857	0.8807
		50,75,100,125,150,175,200,225,250	0.9499	2803.10	13.6544	0.8478	1.1716	0.8874
		5,36,68,95,147,191,216,235,249	0.9633	2237.10	14.6340	0.8774	1.1324	0.9066
	Proposed algorithm	**11,56,75,98,119,135,163,202,236**	**0.9715**	**1881.80**	**15.3851**	**0.8960**	**1.1091**	**0.9185**
Ball.jpg (k=7)	Randomly selected	44,79,130,188,210,240,255	0.6971	4505.40	11.5934	0.6298	1.5691	0.8670
		83,114,167,181,19,222,239	0.5504	6579.30	9.949	0.4628	2.1260	0.7396
		50,75,150,175,200,225,250	0.6761	4789.00	11.3284	0.6070	1.6274	0.8536
		68,148,181,194,21,222,248	0.4794	7354.10	9.4655	0.4005	2.4510	0.6948
		82,129,171,189,20,228,254	0.5240	6859.80	9.7677	0.4403	2.2332	0.7250
	Proposed algorithm	**51,61,76,82,164,208, 219**	**0.7583**	**3488.50**	**12.7044**	**0.7116**	**1.3905**	**0.9225**
Fish.jpg (k=8)	Randomly selected	60,100,120,150,170,199,210,244	0.9056	1720.34	15.7639	0.8652	1.1462	0.9573
		50,69,88,111,140,179,199,200	0.9434	974.19	18.2444	0.9208	1.0777	0.9791
		39,77,100,124,143,167,191,252	0.9375	1080.67	17.7891	0.9128	1.0871	0.9768
		10,40,60,100,134,167,199,222	0.5100	10838.08	7.7815	0.2729	3.5744	0.3281
		99,125,150,170,185,199,210,225	0.8527	2820.93	13.6253	0.7839	1.2641	0.9149
	Proposed algorithm	**21,59,65,118,130,150,182,196**	**0.9543**	**796.39**	**19.1195**	**0.9339**	**1.0627**	**0.9814**
Tulip.jpg (k=5)	Randomly selected	17,31,75,159,166	0.8415	3255.90	13.0041	0.6968	1.4191	0.8068
		57,94,144,210,250	0.8407	2994.90	12.8702	0.6873	1.4283	0.7863
		45,84,175,221,244	0.1671	9578.00	8.3181	0.1234	7.617	0.3381
		20,69,185,232,255	0.9273	1406.40	16.6497	0.8646	1.1462	0.9302
		25,90,150,170,220	0.8414	3255.90	13.0041	0.6968	1.4191	0.8068
	Proposed algorithm	**50,140,170, 230,250**	**0.938**	**1068.20**	**17.8441**	**0.8953**	**1.1073**	**0.9554**
Hand.jpg (k=7)	Randomly selected	99,120,154,190,200,217,240	0.2387	9884.60	8.1812	0.2055	4.6864	0.5686
		77,81,99,140,190,222,240	0.5707	5704.70	10.5685	0.5381	1.8314	0.8902
		55,79,95,150,196,214,244	0.5051	6663.20	9.8940	0.4618	2.1288	0.8283
		25,57,78,159,177,235,251	0.5763	5626.50	10.6284	0.5443	1.8108	0.8943
		9,37,49,167,209,240,255	0.5531	5975.40	10.3672	0.5166	1.9066	0.8727
	Proposed algorithm	**41,75,82,94,168,203,225**	**0.6632**	**4392.20**	**11.7039**	**0.6426**	**1.5374**	**0.9509**
Flower.jpg (k=6)	Randomly selected	77,99,110,140,166,180	0.8625	2568.30	14.0344	0.8225	1.2058	0.9537
		50,80,113,176,217,246	0.8515	2829.40	13.6139	0.8015	1.2316	0.9439
		90,120,140,160,199,200	0.8376	3178.70	13.1083	0.7819	1.2678	0.9292
		100,140,190,220,240,255	0.3823	10800.00	7.7965	0.2754	3.5428	0.5172
		30,90,150,220,240,255	0.3745	10876.00	7.7662	0.2704	3.6071	0.5137
	Proposed algorithm	**40,52,61,168,188,217**	**0.8996**	**1834.70**	**15.4952**	**0.8712**	**1.1389**	**0.9708**

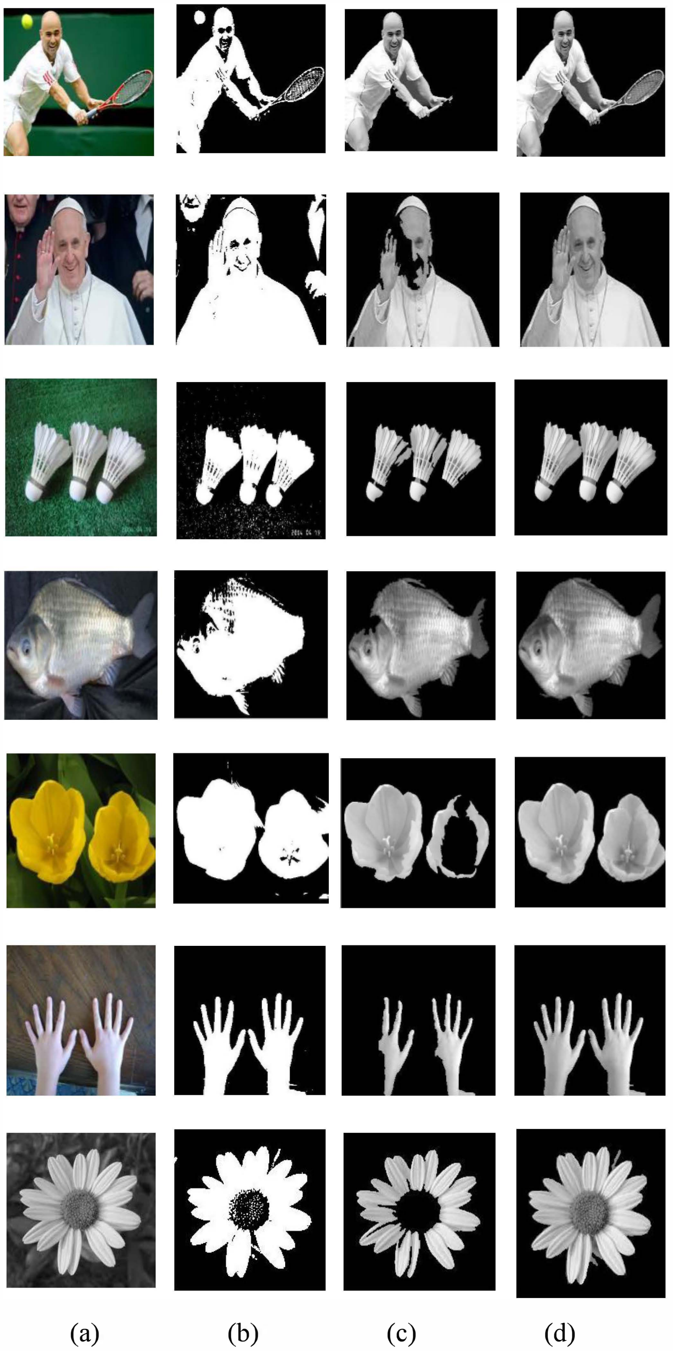

Fig 2. (a) Original image, (b) Otsu threshold, (c) K-Means clustering by random centroids and (d) Proposed Algorithm

From Table I. the image segmentation quality measures such as SSIM, IF, SC, MSE, PSNR and CC were improved for proposed algorithm.

V. Conclusion

In this paper, we propose an enhanced K-means clustering algorithm for image segmentation that selects the centroids automatically. The performance of proposed algorithm is evaluated qualitatively and quantitatively. The qualitative result shows that the proposed algorithm uniformly segments the objects over the randomly selected centroids and also it segments the objects from noisy, texture and complex background. The segmented result is compared quantitatively using quality measures such as SSIM, MSE, PSNR, IF, SC, CC. The MSE was improved by 38.75%. The PSNR was improved by 19.01%. The CC was improved by 14.04%. The SSIM was improved by 16.97%. The IF was improved by 21.96%.

References

[1] Gonzalez R. C., Woods R.E., Digital Image Processing, 2nd ed., Pearson Education, 2000.

[2] Hong Yao, Qingling Duan, Daoliang Li, Jianping Wang, An improved k-means clustering algorithm for fish image segmentation, Mathematical and Computer Modelling, in press.

[3] Jain A. K., Murthy M. N., Flynn P. J., Data clustering: a review, ACM Comput. Surv., 1999, 31, (3), pp.264-323.

[4] Gonzalez R. C., Woods R. E, Eddins S. L., Digital Image Processing using MATLAB, Pearson Education, 2008.

[5] Punya Thimajshima, A new modified fuzzy c-means algorithm, IEEE Transations,2000,pp.1684-1686.

[6] Nargess Memarsadehi, David M. Mount, Nathan S. Netanyahu, Jacqueline Le Moigne, A fast implementation of the ISODATA clustering algorithm, IJCGA, 2005, Source of Acquisition NASA Goddard Space Flight Center.

[7] RuiXu, DonaldWunsch, Survey of Clustering Algorithms, IEEE transactions on neural networks, vol.16, no.3, May 2005.

[8] Otsu N., A Threshold Selection Method from Gray-Level Histogram, IEEE Transactions on Systems, Man and Cybernetics, 1979 vol.9, No.1, pp.62-66.

[9] MacQueen J. B., Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. 1967 pp. 281–297.

[10] Nasrabadi N. M., King R. A., Image coding using vector quantization: A review, IEEE Trans. Communications, vol.36, no.8,pp. 957-971, August 1988.

[11] Ahmet M. Eskicioglu, Paul S. Fisher Image Quality Measures and their performance IEEE Transactions on Communications, Vol.43, No.12, December 1995.1. R. C. Gonzalez, R. E. Woods, “Digital Image Processing”, 2nd ed., Pearson Education, 2000.

[12] Zhou Wang, Alan Conrad Bovik., et.al., Image Quality Assessment: From Error Visibility to Structural Similarity, IEEE Transactions on Image processing, vol. 13, no.4, April 2004.

[13] en.wikipedia.org/wiki/Correlation_coefficient.

[14] Dinesh Kumar V. P., Thomas T., Clustering of invariance improved Legendre moment descriptor for content based image retrieval, IEEE International Conference on Signal Processing, Communications and Networking, pp. 323-327, 2008.

[15] Abdul Nazeer K. A. et al., Enhancing the k-means clustering algorithm by using a O(n logn) heuristic method for finding better initial centroids, 2011 Second International Conference on Emerging Applications of Information Technology, 2011.pp 261-264.

[16] Ran Vijay Singh, Bhatia M. P. S., Data Clustering with Modified K-means Algorithm, IEEE International Conference on Recent Trends in Information Technology, June 2011.pp 717-712.

[17] Sohrab Mahmud Md., Mostafizer Rahman Md., Nasim Akhtar Md., Improvement of K-means Clustering algorithm with better initial centroids based on weighted average, 7th International Conference on Electrical and Computer Engineering, December 2012. pp 647-650.

[18] Jegatha Deborah L., Baskaran R., Kannan A., A Survey on Internal Validity Measure for Cluster Validation, International Journal of Computer Science and Engineering Survey (IJCSES), vol.1, no.2, November 2010. “Synthetic structure of industrial plastics (Book style with paper title and editor),” in Plastics, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.G