

Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015)

## Image Segmentation using $K$ -means Clustering Algorithm and Subtractive Clustering Algorithm

Nameirakpam Dhanachandra\*, Khumanthem Manglem and Yambem Jina Chanu

*National Institute of Technology, Manipur 795 001, India.*

---

### Abstract

Image segmentation is the classification of an image into different groups. Many researches have been done in the area of image segmentation using clustering. There are different methods and one of the most popular methods is  $k$ -means clustering algorithm.  $K$ -means clustering algorithm is an unsupervised algorithm and it is used to segment the interest area from the background. But before applying  $K$ -means algorithm, first partial stretching enhancement is applied to the image to improve the quality of the image. Subtractive clustering method is data clustering method where it generates the centroid based on the potential value of the data points. So subtractive cluster is used to generate the initial centers and these centers are used in  $k$ -means algorithm for the segmentation of image. Then finally medial filter is applied to the segmented image to remove any unwanted region from the image.

© 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license

(<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Peer-review under responsibility of organizing committee of the Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015)

**Keywords:** Image segmentation;  $K$ -means clustering; Median filter; Partial contrast stretching; Subtractive clustering.

---

### 1. Introduction

Image segmentation is one of the mostly used methods to classify the pixels of an image correctly in a decision oriented application. It divides an image into a number of discrete regions such that the pixels have high similarity in each region and high contrast between regions. It is a valuable tool in many field including health care, image processing, traffic image, pattern recognition etc. There are different techniques for image segmentation like threshold based, edge based, cluster based, neural network based<sup>1</sup>. From the different technique one of the most efficient methods is the clustering method. Again there are different types of clustering:  $K$ -means clustering, Fuzzy  $C$ -means clustering, mountain clustering method and subtractive clustering method.

One of most used clustering algorithm is  $k$ -means clustering. It is simple and computationally faster than the hierarchical clustering. And it can also work for large number of variable. But it produces different cluster result for different number of number of cluster. So it is required to initialize the proper number of number of cluster,  $k^2$ . Again, it is required to initialize the  $k$  number of centroid. Different value of initial centroid would result different cluster. So selection of proper initial centroid is also an important task.

Nowadays image segmentation becomes one of important tool in medical area where it is used to extract or region of interest from the background. So medical images are segmented using different technique and process outputs are

---

\*Corresponding author. Tel.: 07308294378.

E-mail address: [dhana.namei@gmail.com](mailto:dhana.namei@gmail.com)

used for the further analysis in medical. But medical images in their raw form are represented by the arrays of numbers in the computer<sup>3</sup>, with the number indicating the values of relevant physical quantities that show contrast between different types of body parts. Processing and analysis of medical images are useful in transforming raw images into a quantifiable symbolic form, in extracting meaningful qualitative information to aid diagnosis and in integrating complementary data from multiple imaging modalities. And one of the fundamental problems in medical analysis is the image segmentation which identifies the boundaries of objects such as organs or abnormal region in images. Results from the segmentation make it possible for shape analysis, detecting volume change, and making a precise radiation therapy treatment plan.

## 2. Related Work

There have been many works done in the area of image segmentation by using different methods. And many are done based on different application of image segmentation. *K*-means algorithm is the one of the simplest clustering algorithm and there are many methods implemented so far with different method to initialize the centre. And many researchers are also trying to produce new methods which are more efficient than the existing methods, and shows better segmented result. Some of the existing recent works are discussed here.

Pallavi Purohit and Ritesh Joshi<sup>4</sup> introduced a new efficient approach towards *K*-means clustering algorithm. They proposed a new method for generating the cluster center by reducing the mean square error of the final cluster without large increment in the execution time. It reduced the means square error without sacrificing the execution time. Many comparisons have been done and it can conclude that accuracy is more for dense dataset rather than sparse dataset.

Alan Jose, S. Ravi and M. Sambath<sup>5</sup> proposed Brain Tumor Segmentation using *K*-means Clustering and Fuzzy *C*-means Algorithm and its area calculation. In the paper, they divide the process into three parts, pre-processing of the image, advanced *k*-means and fuzzy *c*-means and lastly the feature extraction. First pre-processing is implemented by using the filter where it improves the quality of the image. Then the proposed advance *K*-means algorithm is used, followed by Fuzzy *c*-means to cluster the image. Then the resulted segment image is used for the feature extraction for the region of interest. They used MRI image for the analysis and calculate the size of the extracted tumor region in the image.

Madhu Yedla, Srinivasa Rao Pathakota, T. M. Srinivasa<sup>6</sup> proposed Enhancing *K*-means clustering algorithm with improved initial center. A new method for finding the initial centroid is introduced and it provides an effective way of assigning the data points to suitable clusters with reduced time complexity. They proved their proposed algorithm has more accuracy with less computational time comparatively original *k*-means clustering algorithm. This algorithm does not require any additional input like threshold value. But this algorithm still initializes the number of cluster *k* and suggested determination of value of *k* as one of the future work.

K. A. Abdul Nazeer, M. P. Sebastian<sup>7</sup> proposed an enhanced algorithm to improve the accuracy and efficiency of the *k*-means clustering algorithm. They present an enhanced *k*-means algorithm which combines a systematic method consisting two approaches. First one is finding the initial centroid and another is assigning the data point to the clusters. They have taken different initial centroid and tested execution time and accuracy. From the result it can be conclude that the proposed algorithm reduced the time complexity without sacrificing the accuracy of clusters.

## 3. Contrast Enhancement using Partial Contrast Stretching

Medical images which have been used for the analysis may have their own weakness such as blurred or low contrast. So a contrast enhancement technique such as Partial Spatial Starching (PCS) is used to improve the image quality and contrast of the image<sup>8</sup>. It is done by stretching and compression process. By applying this technique, the pixel range of lower threshold value and upper threshold value will be mapped to a new pixel range and stretched linearly to a wide range of pixels within new lower stretching value, and the remaining pixels will experience compression (Fig. 1).

## 4. Subtractive Clustering Algorithm

Subtractive clustering is a method to find the optimal data point to define a cluster centroid based on the density of surrounding data points<sup>9</sup>. This method is the extension of Mountain method, proposed by Chiu<sup>10</sup>. Mountain method is

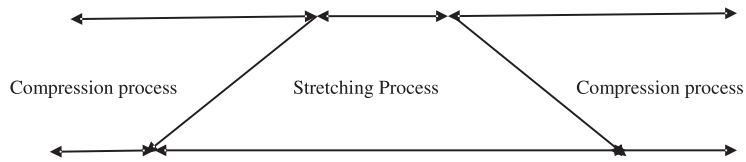


Fig. 1. Partial contrast stretching process.

very simple and effective. It estimates the number and initial location of the cluster centers. It distribute the data space into gridding point and compute the potential for each data point base on its distance to the actual data point. So the grid point with many data point nearby will have high potential value. And so this grid point with highest potential value will be choose as first cluster centre. So after selecting the first cluster centre we will try to find the second cluster centre by calculating the highest potential value in the remaining grid points. As grid points near the first cluster center will reduce its potential value, the next cluster center will be grid with many data point nearby other than first cluster center grid point. So this procedure of acquiring new cluster center and reducing the potential of surrounding grid point repeat until potential of all grid points falls below a threshold value. So this method is one of the simplest and effective methods to find the cluster centers. But with increase in the dimension of data, its computation complexity grows exponentially. So, subtractive clustering algorithm solves the computational method associated with mountain method. It uses data points as the candidates for cluster centre and the computation of this method is proportional to the problem size.

Consider a collection of  $n$  data points:  $X = \{x_1, x_2, x_3 \dots x_n\}$ . Then each point is considered as a potential cluster center. The potential of data point's  $x_n$  is defined as:

$$P_n = \sum_{i=1}^n e^{\frac{-4x_n - x_i^2}{r_a^2}} \quad (1)$$

where  $r_a$  is hyper sphere cluster radius in data space and it is a positive constant which is used to define the neighbourhood. The symbol  $\|\cdot\|$  denotes the Euclidean distance. So the measure of the potential for the data point is a measure of function of distance to all other data points.

After finding the potential of each data points, select the data point with maximum potential as the first cluster centre. Let us consider  $x_1$  and  $p_1$  as first cluster centre and its corresponding potential respectively. Then revise the potential of each data point by using the formula given below.

$$P_n = P_n - P_1 e^{\frac{-4x_n - x_1^2}{r_b^2}} \quad (2)$$

$r_b$  is the hyper sphere penalty radius in data space and it is a positive constant. Here an amount of potential is subtracted from each data point as a function of distance from the first cluster center. So the data points near the first cluster center will have greatly reduced potential, and therefore it have less chance to select as next cluster center. After calculating the revise potential of each data points, find the next highest potential as the next cluster center. So these processes continue until a sufficient number of cluster centre are obtained.

## 5. K-Means Clustering Algorithm

Clustering is a method to divide a set of data into a specific number of groups. It's one of the popular method is  $k$ -means clustering. In  $k$ -means clustering, it partitions a collection of data into a  $k$  number group of data<sup>11,12</sup>. It classifies a given set of data into  $k$  number of disjoint cluster.  $K$ -means algorithm consists of two separate phases. In the first phase it calculates the  $k$  centroid and in the second phase it takes each point to the cluster which has nearest centroid from the respective data point. There are different methods to define the distance of the nearest centroid and one of the most used methods is Euclidean distance. Once the grouping is done it recalculate the new centroid of each

cluster and based on that centroid, a new Euclidean distance is calculated between each center and each data point and assigns the points in the cluster which have minimum Euclidean distance. Each cluster in the partition is defined by its member objects and by its centroid. The centroid for each cluster is the point to which the sum of distances from all the objects in that cluster is minimized. So  $K$ -means is an iterative algorithm in which it minimizes the sum of distances from each object to its cluster centroid, over all clusters.

Let us consider an image with resolution of  $x \times y$  and the image has to be cluster into  $k$  number of cluster. Let  $p(x, y)$  be an input pixels to be cluster and  $c_k$  be the cluster centers. The algorithm for  $k$ -means<sup>13</sup> clustering is following as:

1. Initialize number of cluster  $k$  and centre.
2. For each pixel of an image, calculate the Euclidean distance  $d$ , between the center and each pixel of an image using the relation given below.

$$d = \|p(x, y) - c_k\| \quad (3)$$

3. Assign all the pixels to the nearest centre based on distance  $d$ .
4. After all pixels have been assigned, recalculate new position of the centre using the relation given below.

$$c_k = \frac{1}{k} \sum_{y \in c_k} \sum_{x \in c_k} p(x, y) \quad (4)$$

5. Repeat the process until it satisfies the tolerance or error value.
6. Reshape the cluster pixels into image.

Although  $k$ -means has the great advantage of being easy to implement, it has some drawbacks. The quality of the final clustering results is depends on the arbitrary selection of initial centroid. So if the initial centroid is randomly chosen, it will get different result for different initial centers. So the initial center will be carefully chosen so that we get our desire segmentation. And also computational complexity is another term which we need to consider while designing the  $K$ -means clustering. It relies on the number of data elements, number of clusters and number of iteration.

## 6. Median Filter

Median filtering is used as a noise removal in order to obtain a noise free image. After segmentation is done, the segmented image may still present some unwanted regions or noise. So to make the image a good and better quality, the median filter is applied to the segmented image. We can use different neighborhood of  $n \times n$ . But generally neighborhood of  $n = 7$  is used because large neighborhoods produce more severe smoothing.

## 7. Proposed Algorithm

The proposed algorithm consists of partial contrast stretching, subtractive clustering,  $k$ -means clustering and median filter. Mostly the medical images which are used for segmentation have low contrast. So contrast stretching is used to improve the quality of the image. After improving the quality of image, subtractive clustering algorithm is used to generate the centers, based on the potential value of the image. Number of centre is generated based on number of cluster  $k$ . This centre is used as initial centre in  $k$ -means algorithm. Using the  $k$ -means algorithm, the image is segmented into  $k$  number of cluster. After the segmentation of image, the image can still contain some unwanted region or noise. These noises are removed by using the median filter. The proposed algorithm is followed as below (Fig. 2).

1. Load the image to be segmented.
2. Apply partial contrast stretching.  
Initialize number of cluster,  $k$ .
3. Use equation (1) to calculate the potential for every pixel value of the image.
4. Find maximum potential in step 3 and set that point be first center cluster and its corresponding potential as maximum potential.
5. Use equation (2) to update the potential value of other remaining pixels based on the first cluster center.

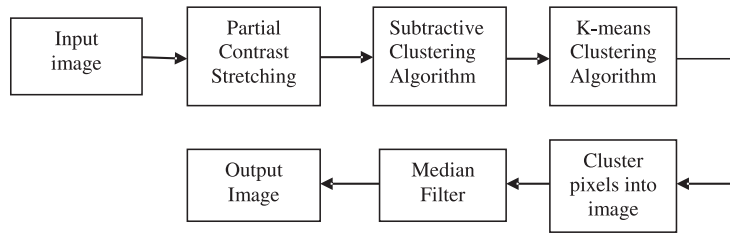


Fig. 2. Block diagram of the proposed algorithm.

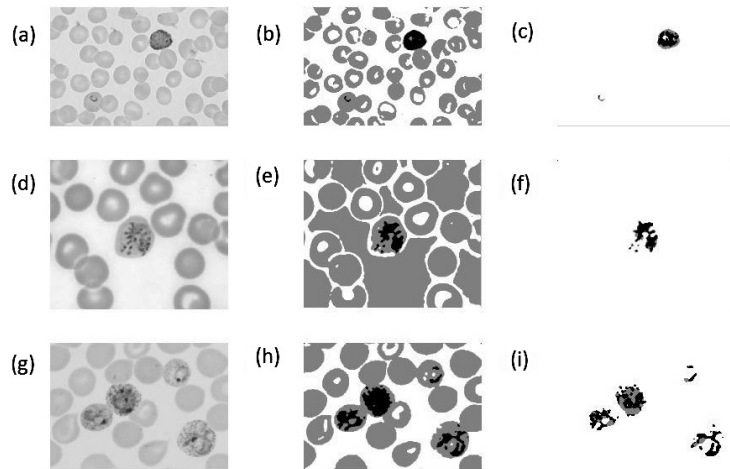


Fig. 3. (a), (d), (g) Original image; (b), (e), (h) *K*-means algorithm; (c), (f), (i) Proposed algorithm.

6. Again find the maximum potential in the step 4 and let that point be second point.
7. Continue the process until it finds the  $k$  number of cluster.
8. Used  $k$  centre as initial centre in the  $k$ -means clustering algorithm.
9. Find the Euclidean distance of each centroid from every pixel of the image using the relation (3).
10. Assign the pixel with minimum distance with respect to centroid to its respective cluster of the centroid.
11. Recalculate the new center location by using the equation (4).
12. Repeat the steps 10–12, until it satisfies the tolerance or error value.
13. Reshape the cluster into image.
14. Median filter is applied to the segmented image to remove any unwanted noise or region.

## 8. Results

We used different types of medical images for the analysis. In medical image analysis, mostly the infected areas or area of interest are segmented from background. We used infected blood cell like malaria infected blood cell for the analysis. Matlab is used to implement the proposed algorithm. We compare the result of  $k$ -means algorithm with proposed algorithm and it is shown in the Fig. 3.

In the subtractive algorithm we can find the number of cluster  $k$ . But instead of finding the number of cluster we have given user defined number of cluster  $k$  based on the type of the image. Here we have taken the number of cluster,  $k = 3$ . Now in the subtractive cluster we can tune the output image by using the different value of hyper sphere cluster radius,  $r_a$  and hyper sphere penalty radius,  $r_b$ . We know that  $r_a$  defines the neighborhood distance, so changing this constant will affect the output. In the Fig. 4 we have shown output images for different value of  $r_a$ .

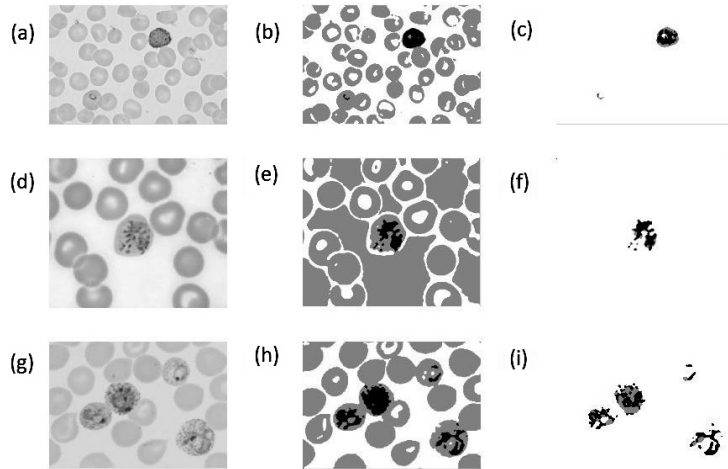


Fig. 4. (a), (e), (i) Original image; (b), (f), (j) Proposed algorithm with  $r_b = 0.1$ ; (c), (g), (k) Proposed algorithm with  $r_b = 0.5$ ; (d), (h), (l) Proposed algorithm with  $r_b = 1.2$ .

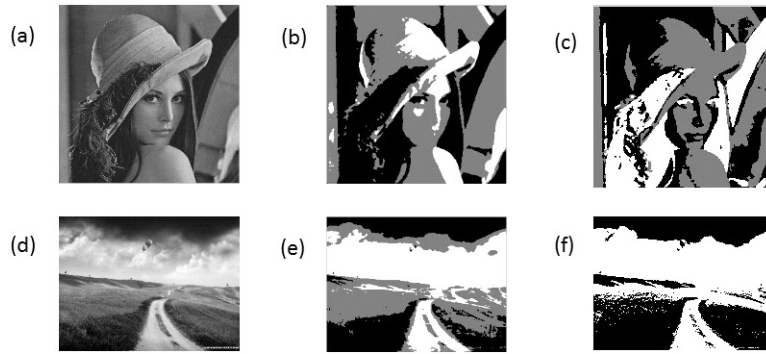


Fig. 5. (a), (d) Original image; (b), (e)  $K$ -means algorithm; (c), (f) Proposed algorithm.

We have taken different value of  $r$  and according to the value of  $r$  we get different result as observed from the figure. So we need to be very careful while taking the value of  $r_a$ . And we can also tune the output result by varying the value of  $r_a$ . In the same way we can also tune the output by varying the value of the sphere penalty radius,  $r_b$ .

And lastly we have check the result in some other images and the segmented output result are shown above in the Fig. 5. From the figure we observed that the output result using proposed algorithm has better segmentation result as compared to the classical  $k$ -means algorithm.

The quality of the segmented image is analyzed using the measurement value of Root Mean Square Error (RMSE) and Peak to Signal Noise Ration (PSNR)<sup>14</sup>.

1. Root Mean Square Error: It has been used as a standard performance measurement of the output image. It gives how much output image is deviated from the input image.

$$RMSE = \sqrt{\frac{1}{n_x n_y} \frac{\sum_{x=0}^{n_x-1} \sum_{y=0}^{n_y-1} [r(x, y)]^2}{\sum_{x=0}^{n_x-1} \sum_{y=0}^{n_y-1} [r(x, y) - t(x, y)]^2}} \quad (5)$$

Table 1. RMSE and PSNR values.

| Image               | RMSE<br>(proposed algorithm) | PSNR<br>(proposed algorithm) | RMSE<br>( <i>K</i> -means algorithm) | PSNR<br>( <i>K</i> -means algorithm) |
|---------------------|------------------------------|------------------------------|--------------------------------------|--------------------------------------|
| First Blood Sample  | 0.0077                       | 34.80                        | 0.0081                               | 34.63                                |
| Second Blood Sample | 0.0084                       | 34.79                        | 0.0144                               | 32.44                                |
| Third Blood Sample  | 0.0073                       | 34.43                        | 0.0079                               | 33.99                                |
| Lena Sample         | 0.0017                       | 35.77                        | 0.0041                               | 31.9                                 |
| Scenery Sample      | 0.0011                       | 38.235                       | 0.0036                               | 32.94                                |

2. Peak to Signal Noise Ratio: The peak to signal noise ratio is the proportion between maximum attainable powers and the corrupting noise that influence likeness of image. It is used to measure the quality of the output image.

$$PSNR = 10 \cdot \log_{10} \left[ \frac{\max(r(x, y))^2}{\sqrt{\frac{1}{n_x n_y} \frac{\sum_{x=0}^{n_x-1} \sum_{y=0}^{n_y-1} [(r(x, y))]^2}{\sum_{x=0}^{n_x-1} \sum_{y=0}^{n_y-1} [r(x, y) - t(x, y)]^2}}} \right] \quad (6)$$

where  $r(x, y)$  is the input image and  $t(x, y)$  is the segmented image. And the smaller value of RMSE means the image is of good quality and smaller value of PSNR means the image of poor quality. The value of RMSE and PSNR of the segmented image is given below in the Table 1.

The RMSE and PSNR value are calculated for classical *K*-means Algorithm as well as proposed method. The values of RMSE are getting very low and the value of PSNR is getting above 30 and when both methods are compared it is found that the proposed method has better result. So we can conclude that the output image resulted from the proposed algorithm are of good quality.

## 9. Conclusion

We have segmented an image by using *k*-clustering algorithm, using subtractive cluster to generate the initial centroid. At the same time partial contrast stretching is used to improve the quality of original image and median filter is used to improve segmented image. And the final segmented result is compare with *k*-means clustering algorithm and we can conclude that the proposed clustering algorithm has better segmentation. The output images are also tune by varying the hyper sphere cluster radius and we can conclude from that result that by varying the hyper sphere cluster radius we can get different output. And so we should take the value of hyper sphere cluster very carefully. Finally RMSE and PSNR are checked and observed that they have small and large value respective, which are the condition for good image segmentation quality. And comparison for RMSE and PSNR are done for proposed method and classical *K*-means algorithm and it is found that the proposed method have better performance result. In the future, we can improve the quality of the output image more by using the morphological operation and get better performance measurement. We can also implement different clustering method using subtractive clustering algorithm. And lastly we can implement and analyze in different areas of image segmentation.

## References

- [1] Khaled Hammouda, A Comparative study of Data Clustering technique. Department of System Design Engineering, University of Waterloo, Canada.
- [2] Koheri Arai and Ali Ridho Barakbah, Heirarchical *K*-means: An algorithm for Centroid initialization for *K*-means, Saga University, (2007).
- [3] D. L. Pham, C. Y. Xu, J. L. Prince, A Survey of Current Methods in Medical Image Segmentation, In *Annual Review of Biomedical Engineer*, (2000).
- [4] Pallavi Purohit and Ritesh Joshi, A New Efficient Approach towards *k*-means Clustering Algorithm, In *International Journal of Computer Applications*, (0975-8887), vol. 65, no. 11, March (2013).
- [5] Alan Jose, S. Ravi and M. Sambath, Brain Tumor Segmentation using *K*-means Clustering and Fuzzy *C*-means Algorithm and its Area Calculation. In *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 2, issue 2, March (2014).



- [6] Madhu Yedla, Srinivasa Rao Pathakota and T. M. Srinivasa, Enhanced  $K$ -means Clustering Algorithm with Improved Initial Center, In *International Journal of Science and Information Technologies*, vol. 1(2), pp. 121–125, (2010).
- [7] K. A. Abdul Nazeer and M. P. Sebastian, Improving the Accuracy and Efficiency of the  $k$ -means Clustering Algorithm, In *Proceedings of the World Congress on Engineering, London, WCE*, vol. 1, July (2001).
- [8] A. N. Aimi Salihah, M. Y. Mashor, N. H. Harun and H. Rosline, Colour Image Enhancement Technique for Acute Leukaemia Blood Cell Morphological Feature, In *IEEE International Conference on System, Man and Cybernatic*, pp. 3677–3682, (2010).
- [9] K. M. Bataineh, M. Naji and M. Saqer, A Comparison Study between Various Fuzzy Clustering Algorithm, In *Jordan Journal of Mechanical and Industrial Engineering*, vol. 5, no. 4, August (2011).
- [10] R. Yager and D. Filev, Generation of Fuzzy Rules in Mountain Clustering, In *Journal of Intelligent and Fuzzy System*, vol. 2(3), pp. 209–219, (1992).
- [11] Shehroz S. Khan and Amir Ahmad, Cluster Centre Initialization Algorithm for  $K$ -means Cluster, In *Pattern Recognition Letters*, pp. 1293–1302, (2004).
- [12] Sorin Israil, An Overview of Clustering Methods, In *With Application to Bioinformatics*.
- [13] Aimi Salihai Abdul, Mohd Yusuff Masor and Zeehaida Mohamed, Colour Image Segmentation Approach for Detection of Malaria Parasiter using Various Colour Models and  $k$ -Means Clustering, In *WSEAS Transaction on Biology and Biomedecine.*, vol. 10, January (2013).
- [14] Jaskirat Kaur, Sunil Agarwal and Renu Vig, A Methodology for the Performance Analysis of Cluster Based Image, In *International Journal of Engineering Research and Application*, vol. 2, (2012).