A NOVEL FOOD IMAGE SEGMENTATION BASED on HOMOGENEITY TEST of K-MEANS CLUSTERING

SALWA KHALID ABDULATEEF

TIKRIT UNIVERSITY, COLLEGE OF COMPUTER SCIENCEAND MATHEMATICS, IRAQ

Abstract

Data clustering is an important machine-learning topic. It is useful for variety of applications one of them is image segmentation. A given divided image into regions homogenous additional to certain features is the image segmentation process, which matches real objects of an actual scene. FIS (Food Image Segmentation) is important for calories estimation. K-means has been used for performing such task. However, in order to conclude the food items number in the image, it requires interacting with the application. This article, presents a novel approach based dependently on k-means named Hk-means (Homogeneity test of k-means) is developed to calculate k value and applied for FIS for the purpose of assuring full autonomy in the calories estimation system. This approach uses the homogeneity test so as to compensate the new item existence in the image. The suggested method Hk-means is tested on food images and show accuracy 96%. The experimental results has achieved 1.5 second execution time when compare with benchmark method.

Keywords: Image segmentation, k-means, clustering, computer vision, statistical processes

INTRODUCTION

Image segmentation is an emerging computer visualization application. It has found interest in a variety of resources such as medical field, diagnostics based on vision data [1], [2], industrial processes such as assembly line calculations [3], [4], agronomy [5], [6] and forensic image analysis data for operations of crime [7]. Image segmentation is the process of segmenting a given image into related locations in relation to specific features, and in the hope of harmonizing with the realities of the scene. In fact, image classification is used to provide simple and easy-to-follow data for the stages of a standard pattern recognition program [8]-[11], [30].

Classification is still a challenging topic for research in computer vision [12]-[15]. In addition to the existence of various types of image classification methods in the literature, there are drawbacks and challenges that can be addressed. Another difficulty in using the components of the algorithms is the lack of information about the specific range of specific chemical algorithms. This is handled in some cases by enabling the segmentation algorithm to work under certain assumptions. Unfortunately, this does not correspond to the requirements of the utility and the type of implementation of component algorithms. Therefore, there is a great need to develop existing segmentation algorithms to handle a portion of unknown parameters in the application context by incorporating additional algorithmic processes.

Data integration is one of the most important and popular methods of data analysis. Data records are grouped into non-tagged sets over similarities [16], [17]. Integration is used in many fields such as [18], [19]. The k -tested algorithm was proposed by MacQueen [20] which was used to classify images into several locations as calories measured [21]. In this research article, we deal with the classification of food images in the context of an unknown number of foods. In the classical literature on segmentation, methods have been adopted to perform the segmentation after assigning it to a feed number. However, the plate may contain a voluntary amount of food items depending on the user's preference or desire. Thus, the k-means are at greater risk of not knowing about the parameter k that represents the number of food items. Because of this, it is important to develop a new food separation method to manage this component. Also, this method should not be costly depending on the processing time considering that the amount of data is large in images. The purpose of this research article is to add statistical processing to the k-means to make it useful in distinguishing the food plate from the unknown food number. The analysis of the calculations will depend on the color distribution of the materials by considering the different color space according to their significance. It is important to extend the feature vectors of features to include color features that allow for additional information about the color distribution of the object and therefore help in avoiding local minima.

The organization of the article is as follows. In Section 2, the domain is assigned. The related activities are given in section 3. The methodology is given in section 4. The results are given in section 5. Finally, a summary and conclusion are given in section 6.

II. BACKGROUND

In this section, a summary about segmentation based on k-means is provided. Next, the limitations of k-means based segmentation are discussed with some visual examples.

A. K-means based Segmentation

Let X={x_i},i=1,…n represents a set of n-dimensional points. In image segmentation each x_i denotes one pixel in the image. While the dimension of x_i denotes the number of colour spaces used in the segmentation. The image contains k food items. We denote each cluster as c_n and the mean of each cluster as u_n. The image segmentation is performed to minimize the error that is represented in the “Eq.(1)”.

$$J(C) = \sum_{n=1}^{k} \sum_{x_i \in c_n} \|x_i - u_n\|^2 \quad (1)$$

The problem is then formulated as how to find the points u_n; n=1,..k where the value of J(C) is minimum. The following steps are performed to resolve the k-means problem.

- Select initial k clusters. u_n;n=1,..k
- Generate a new cluster by placing each pixel to its closest cluster centre.
- Compute new cluster centres. If (converged) go to 4, else go to 2.
- Result is current clusters.

B. Limitations of k-means

One of the problematic limitations of k-means in the context of image segmentation is lacking of pre-knowledge about the number of clusters, which denotes the items that are to be segmented. k is the most critical choice when performing k-means clustering; wrong values of k might lead to local minima [22], [23]. In the literature numbers of heuristic approaches have been proposed for selecting k. A brief discussion is provided in the related work section.

III. RELETED WORK

In the literature, various methods have used image classification methods. Some of them deal with the problem of determination k and provide some solutions to those problems. Ray and Turi [24] introduced a new classification algorithm based on the intra-cluster and inter-cluster distance resulting in the automatic determination of clusters. However, the problem with this method is its complexity of the attribute as it requires producing all images separated into 2 clusters until they reach a large number of clusters when using a validation scale to determine which combinations are best when the validation rate gets the smallest value.

According to the authors [25] the authors propose new algorithms that efficiently investigate the spatial location of overlapping regions and the number of clusters in order to optimize the terms of the Bayesian information criterion (BIC) or the Akaike information criterion (AIC) measures. This method performs a series of experiments on the performance of k-data starting from a low probability and continues to be tied up (taking every opportunity) by calculating the means of each case for the best choice. As a result, this adds a computational overhead to an algorithm such as [24]. In addition, other researchers have shown BIC dysfunction in comparison to other methods. For example, Hamerly and Elkan [26] have run multiple k-means in the data and performed statistical tests to make sure that each group follows a Gaussian distribution. This method works well compared to the BIC-based method; however, it suffers from some limitations. First, in order to have a reliable test of the hypothesis that the data are Gaussian distributions, it is important to have an adequate number of samples in each cluster. Unfortunately, this doesn't always work especially when the data corresponds to a small object size and the image resolution is low. Also, performance-based methods have been found to address the problem of finding the sum of clusters.

The authors in [27] proposed an edge detection as a way to determine the number of clusters when performing a k-segmentation. Only long edges are used to determine the collection value based on color matching. The average color of each edge is calculated along with the Euclidean distance. The number of edges remaining after the merging of the edges is assumed to be the number of sets of images. This value is used for the k-value for the k-image separation. The result of the detection limit is less than noise, which may affect the accuracy of the calculation. In addition to the k-k determination problem, other researchers have faced other issues. In [28] the authors addressed the problem of the implementation of k-methods or implementation methods (IM). Also, they compared eight widely used methods in different types of information. To overcome the difficulty of obtaining a knee point index method used by Zhao [29] for BIC and the sum of squares in the study-based correlation coefficients has been proposed. A small amount is considered the most appropriate number of clusters.

IV. METHODOLOGY

There are four key steps of our proposed system: (a) image acquisition for food images (b) extracted twelve colour features [30] and (c) k-means clustering (d) modified k-means clustering for connected food image segmentation. “Fig.1” shows the illustration of the flowchart for modifies k-means applied. The detailed description of the proposed method steps will be discussed in the following subsections.

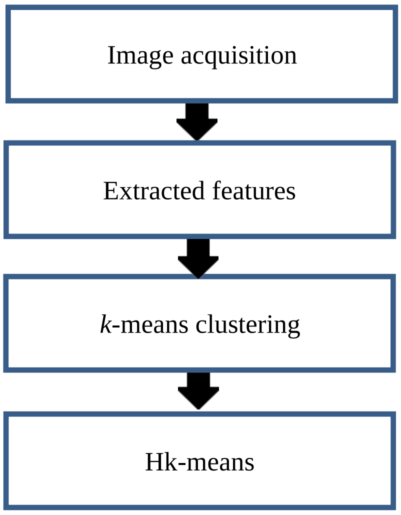


Fig .1. Steps of proposed methodology

A. Image acquisition

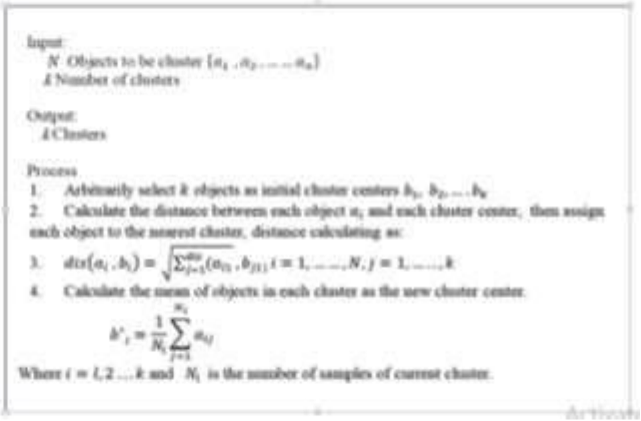
The first step in this process is to take a photo. It uses different types of food items in preparation. The program, using a smartphone with an 8-mega pixel camera, stores 25 images in the Joint Photographic Scientists Group (JPEG) by default. Light adjustments do not require more than lighting as usual in any indoor environment. There is no limit to the angle of acquisition of an angle to an additional corner of the dining table for the final purpose of capturing precise information. The food is arranged on a round plate without any restrictions on the color of the plate.

B. Features for segmentation

In order to segment regions, each region is dealt separately. Classical k-means is applied on each region separately has to pass to it the set of color features. The selected features were mentioned in [30].

C. Classical k-means clustering

The k-means clustering algorithm is a split-based integration [31]. According to the algorithm, first select the k objects as the starting point for the database, then the distance center must be calculated for each season between each cluster center and each object and should be assigned to the nearest group. Thereafter, the ratings of all qualifications should be reviewed. This process should be repeated until the criterion function is changed. The details of the algorithm are shown in "Fig. 2".



```
Input:
    N Objects to be cluster {$a_1, a_2, \ldots, a_n$}
    k Number of clusters

Output:
    k Clusters

Process
1.  Arbitrarily select k objects as initial cluster centers $b_1, b_2, \ldots, b_k$
2.  Calculate the distance between each object $a_i$ and each cluster center, then assign
each object to the nearest cluster, distance calculating as:
3.  $dis(a_i, b_i) = \sqrt{\sum_{j=1}^{dis}(a_{i1}, b_{j1})}$ $i = 1, \ldots, N, j = 1, \ldots, k$
4.  Calculate the mean of objects in each cluster as the new cluster center.
        $b'_i = \frac{1}{N_i}\sum_{j=1}^{N_i} a_{ij}$
Where $i = 1, 2 \ldots k$ and $N_i$ is the number of samples of current cluster.
```

Fig.2. K-means algorithm

D. Homogeneity test of k-means

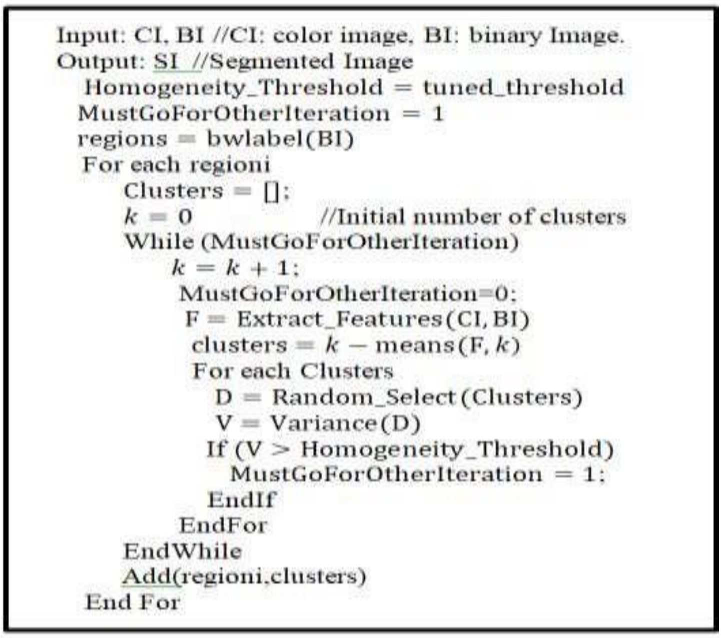
Our aim in this paper is to modify the k-methods to fit the image classification. The most common problem in food classification is compiled from the two problems below: first, the standard way of organizing food items is to put them in the way they are linked. This can result in k-ends ending with local mining or poor assembly results because the data points are subdivided into other clusters (pixels can be partially separated from the nearest component). Second, there is no prior knowledge about the diet. To avoid such problems it is important to do two things [30]:

First, the features should represent each pixel. Second, assuming that k is selected indiscriminately, the easiest way to verify that k is correct is to add to the algorithm a mathematical test called: homogeneity test. The role of this experiment is to ensure that the results of the cluster analysis are consistent to determine the optimal amount of food items. This is done by selecting a random number of samples in a fraction and calculating the difference in "Eq. 2" of those points. In this case, it is important to distinguish between two cases: the first case is when k is lower than the true value of food and the second case is k is higher than the true value of food. and k should be increased.When we are in the final case or the test gives a good decision, the results of the associated classification can lead to overlapping. To avoid the second case and save the first case k should be started from 0 and greater presented by the procedural process. The suspension index is taken when the homogeneity test gives the correct decision for the first time. In the meantime, the modified k methods should be discontinuous and the results of the linear combination should give the correct result for the separation. The Hk-coupled food separation algorithm is shown in “Fig. 3 ”.

$$Variance = \frac{1}{N}\sum_{i=1}^{N}(\quad _i - \mu)^2 \qquad (2)$$

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad (3)$$

where: x_i is indicates to the value of the i th colour component of the image pixel i, and N denotes to the number of pixels in the image.



```
Input: CI, BI //CI: color image, BI: binary Image.
Output: SI //Segmented Image
 Homogeneity_Threshold = tuned_threshold
 MustGoForOtherIteration = 1
 regions = bwlabel(BI)
 For each regioni
     Clusters = [];
     $k$ = 0            //Initial number of clusters
     While (MustGoForOtherIteration)
         $k = k + 1$;
         MustGoForOtherIteration=0;
         F = Extract_Features(CI, BI)
          clusters = $k$ − means(F, $k$)
          For each Clusters
            D = Random_Select(Clusters)
            V = Variance(D)
           If (V > Homogeneity_Threshold)
             MustGoForOtherIteration = 1;
           EndIf
         EndFor
     EndWhile
     Add(regioni,clusters)
 End For
```

Fig.3. Pseudo code of Hk-means

In more details, to perform the homogeneity test random points of the region are generated. They are considered to be a testing cluster. The value of the variance is compared with pre-determined threshold (homogeneity threshold). In case the value is lower, the region is considered to be as homogenous and to consist of one food items. Otherwise, the region is considered to be non-homogenous and as combined of more than one food item.

V. RESULTS and DISCUSSION

To confirm the proposed method, As seen from "Fig. 4 "the number of groceries varies in sequence, three or four. Also, the composition of the food items is different indicating the different distribution of the points in each. Experiments were applied to a total of 25 images, without regard to the plate (which was removed with an active contour). Notably, the Hk-pathway classification provided the results of good classification with the ability to determine the optimal number of food items according to the homogeneity test. All of these tests were performed on a personal computer with Intel (R) Core (TM) memory and 4.00GB using Matlab R2018a.

The performance of the Hk algorithm has been compared with the argument [29]. Two types of experiments are performed.

- The first is responsible for generating an estimated value of k and comparing it to reality.
- Secondly it is responsible for determining the output time of the method and comparing it with the output from the bench.

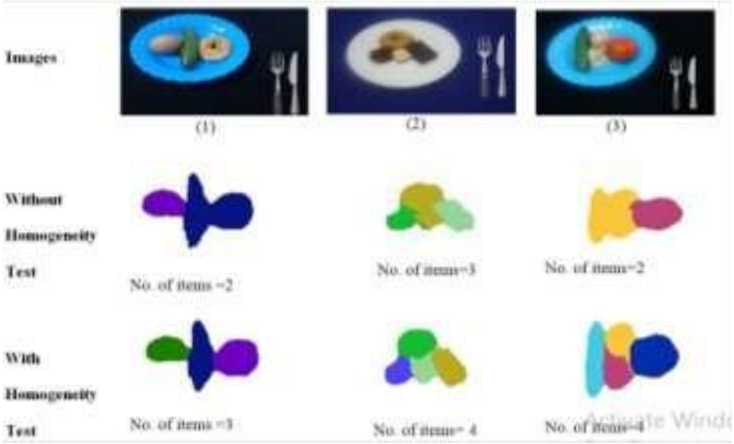


Fig.4. Example of comparison segmentation method with and without homogeneity test

As shown in “Fig. 4 ", Hk-paths succeeded in determining the true value of k for all images. While the bench failed in most cases.

In the results of the reported value states that the correct regions for estimating the k value of Hk-path are 24, this method failed in the case of 3 factors. While the bench failed for 3 items in 3 pictures and 4 items failed in 10 pictures to estimate k value. Testing the proposed benchmark method using an accuracy measure is calculated based on "Eq. (4)".

$$Accuracy(\%)= \left(\frac{Number\ of\ correct\ results}{Total\ number\ of\ images}\right)*100\% \qquad (4)$$

A total of 25 images were examined; the overall accuracy of the Hk-methods was 96%, but the accuracy of the bench was 60% for the estimated value of k.

In addition to validating the value of k, the extraction time is made up of 25 images and compared to the marking method. Fig.5 shows the results of the execution time. In “Fig. 5 ", the y-axis sees the time in seconds, and then the x-axis represents the images. Obviously our improved approach has received less competition time than the bench.

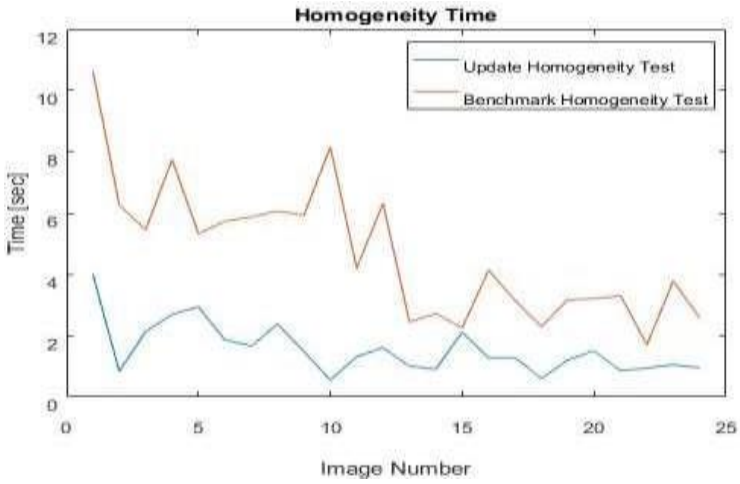


Fig.5. Result of comparison between develop method and benchmark by execution time

VI. CONCLUSION and FUTURE WORK

In this article a new image algorithm has been developed based on k-methods and named Hk-methods. This algorithm is effective in finding the correct number of clusters using a statistical test. This method was tested on a set of food items and provided an accuracy of 96% in predicting the number of food items and our improved method received less time limit than the measurement name. Future work is to incorporate this approach into an overall calorie-based diet plan to ensure full program independence.

REFERENCES

[1] M. C. Christ, and R. M. Parvathi, “Segmentation of medical image using K-means clustering and marker controlled watershed algorithm,” *European Journal of Scientific Research* , 71(2), pp. 190–194, 2012.

[2] E. Abdel-Maksoud, M. Elmogy, and R. Al-Awadi, “Brain tumor segmentation based on a hybrid clustering technique,” *Egyptian Informatics Journal*, 16(1) , pp. 71–81, 2015.

[3] C. T. Yiakopoulos, K. C. Gryllias, and I. A. Antoniadis, “Rolling element bearing fault detection in industrial environments based on a K-means clustering approach, ” *Expert Systems with Applications*, 38(3), pp. 2888–2911, 2011.

[4] T. Velmurugan, “Performance based analysis between K-Means and Fuzzy C-Means clustering algorithms for connection oriented telecommunication data, ”*Applied Soft Computing*,19 ,pp.134–146, 2014.

[5] A. B. Payne, K. B. Walsh, P. P. Subedi, and D. Jarvis, “Estimation of mango crop yield using image analysis–segmentation method, ”*Computers and Electronics in Agriculture*, 91, pp.57–64, 2013.

[6] Q. Wang, S. Nuske, M. Bergerman, and S. Singh, S. “Automated crop yield estimation for apple orchards. *In Experimental robotics,*” *Springer*, pp. 745–758, 2013.

[7] M. Urschler, A. Bornik, E. Scheurer, K. Yen, H. Bischof, and D. Schmalstieg, “Forensic-case analysis: from 3D imaging to interactive visualization,” *IEEE Computer Graphics and Applications*, 32(4), pp.79–87, 2012.

[8] M. Awad, “An Unsupervised Artificial Neural Network Method for Satellite Image Segmentation,” *Int. Arab J. Inf. Technol*, 7(2), pp.199–205, 2010

[9] S. Alpert, M. Galun, A. Brandt, and R. Basri, “Image segmentation by probabilistic bottom-up aggregation and cue integration,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(2), pp.315–327, 2012.

[10] K. Krishan, and S. Singh, “Color Image Segmentation Using Improved Region Growing and K-Means Method,” *IOSR Journal of Engineering*, 4(5), pp.43–46, 2014.

[11] P. Sujatha, and K. K. Sudha, “Performance analysis of different edge detection techniques for image segmentation,” *Indian Journal of Science and Technology* , 8(14) , 2015.

[12] H. Zhang, J. E. Fritts, and S. A. Goldman, “Image segmentation evaluation: A survey of unsupervised methods,” *Computer Vision and Image Understanding*, 110(2) , pp.260–280, 2008.

[13] S. Muthamizhselvi, D. Jeyakumari, and R. Kannan, “A novel predicate for active region merging in automatic image segmentation,” *IJRET,* 2(4), pp. 542–548, 2013.

[14] Q. Zhang, Y. Chi, and N. He, “Color image segmentation based on a modified k-means algorithm,” *In Proceedings of the 7th International Conference on Internet Multimedia Computing and Service*,2015, 46.

[15] A. Fakhry, H. Peng, and S. Ji, “Deep models for brain EM image segmentation: novel insights and improved performance,” *Bioinformatics*, 32(15), pp.2352–2358, 2016.

[16] M. C. Chandhok, S. Chaturvedi, and A. A. Khurshid, “An approach to image segmentation using K-means clustering algorithm,” *International Journal of Information Technology (IJIT),* 1(1), pp.11–17, 2012.

[17] A. Hatamlou, “Black hole: A new heuristic optimization approach for data clustering,” *Information Sciences*, 222, pp.175–184,2013.

[18] S. R. Meenakshi, A. B. Mahajanakatti, and S. Bheemanaik, “Morphological image processing approach using K-means clustering for detection of tumor in brain, ” *International Journal of Science and Research*, 3(8), pp.24–29, 2014.

[19] L. K. Vincent, and N. M. Philip, “Combined difference image and k-means clustering For SAR image change detection,” *International Journal of Advanced Research in Biology, Ecology, Science and Technology*, 1(2), pp.61–65, 2015.

[20] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” *In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1997, 281–297.

[21] P. Pouladzadeh, S. Shirmohammadi, and T. Arici, “ Intelligent SVM based food intake measurement system,” *In Proceedings of CIVEMSA 2013: The IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications*, 2013, 87–92.

[22] A. K. Jain, “Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), pp. 651–666, 2010.

[23] D. T. Pham, S. S. Dimov, and C. D. Nguyen, “Selection of K in K-means clustering,” Part C: *Journal of Mechanical Engineering Science*, 219(1), pp. 103–119, 2005.

[24] S. Ray, and R. H.Turi, “Determination of number of clusters in K -Means clustering and application in colour image segmentation,” *In Proceedings of the 4th international conference on advances in pattern recognition and digital techniques*, 1999, pp.137–143.

[25] D. Pelleg, and A. W. Moore, " X-means: Extending K-means with efficient estimation of the number of clusters, " *In proceedings of the Seventeenth International Conference on Machine Learning ICML,* Vol. 1, 2000, pp. 27–734.

[26] G. Hamerly, and C. Elkan, "Learning the k in k-means," *In NIPS*, Vol. 3, pp. 281–288, 2003.

[27] R. V. Patil, and K. C. Jondhale, " Edge based technique to estimate number of clusters in k-means color image segmentation, "*In 3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT)*, Vol. 2, pp. 117–121, 2010.

[28] M. E. Celebi, H. A. Kingravi, and P. A. Vela, "A comparative study of efficient initialization methods for the k-means clustering algorithm," *Expert Systems with Applications*, 40(1), pp. 200–210, 2013.

[29] Q. Zhao, *Cluster validity in clustering methods*. Publications of the University of Eastern Finland, 2012.

[30] S. K. Abdulateef, M. Mahmuddin, and N. H. Harun, "Developing a new features approach for colour food image segmentation," *ARPN Journal of Engineering and Applied Sciences*, 12(23), pp. 6904-6910, 2017.

[31] D. J. Bora, and A. K. Gupta, "A novel approach towards clustering based image segmentation," *International Journal of Emerging Science and Engineering (IJESE)*, 2(11), pp. 6–10, 2015.