

A Hybrid Agglomerative Method for Improved Image Segmentation

Manish Kumar¹ Meenu Saini²

^{1,2}Student

^{1,2}Department of Electronics & Communication Engineering

^{1,2}Global Research Institute of Management and Technology, Radaur, Haryana, India

Abstract— This paper proposes a hybrid method of image segmentation by using k-means and agglomerative methods of image segmentation. The K-means method is used to find optimum number of clusters with the help gap method and a validity measure. Then this value is used as a limiting value in merge algorithm. The performance of algorithm is measured using a validity index which is measured by two factors. The first factor is intra-cluster distance whose minimum value is desired and another is inter-cluster distance for which a maximum value is required. Once optimum number of cluster is found then k-means clustering algorithm is again applied to generate large number of clusters, then from these large numbers of clusters, pair of clusters with most similar characteristics are merged iteratively until number of clusters are reduced up to optimum number of clusters. The similarity measure is taken from Davies-Bouldin Index. The proposed algorithm is performing better than simple k-means algorithm.

Key words: hybrid method of image segmentation, K-means method, Davies-Bouldin Index

I. INTRODUCTION

Image segmentation divides a digital image into multiple regions in order to analyze them [6]. It is also used to distinguish different objects in the image. Several image segmentation techniques have been developed by the researchers in order to make images smooth and easy to evaluate. Famous techniques of image segmentation which are still being used by the researchers are Edge Detection, Threshold, Histogram, Region based methods, and Watershed Transformation [1]. There are two types of images i.e. gray scale and color images. Image segmentation for color images is totally different from gray scale images. The property of a pixel in an image and proximity of pixels near to that pixel are two basic parameters for any image segmentation algorithm. It can also be representing as similarity of pixels in any region and discontinuity of edges in image. Edge based segmentation is used to divide image on the basis of their edges [2]. Region based methods used the threshold in order to separate the background from an image [3], whereas neural network based techniques used the learning algorithm to train the image segmentation process [4]. The result taken from image segmentation process is the main parameter for further image processing research; this result will also determine the quality of further image processing process. Image segmentation is also used to differentiate different objects in the image, since our image is divided into foreground and background, whereas foreground of image is related to the region of interest, and background is the rest of the image.

While clustering and segmentation algorithms are unsupervised learning processes, users are usually required to set some parameters for these algorithms. These parameters vary from one algorithm to another, but most clustering/segmentation algorithms require a parameter that

either directly or indirectly specifies the number of clusters/segments. This parameter is typically either k, the number of clusters/segments to return, or some other parameter that indirectly controls the number of clusters to return, such as an error threshold. Setting these parameters requires either detailed pre-existing knowledge of the data, or time-consuming trial and error. The latter case still requires that the user has sufficient domain knowledge to know what a good clustering “looks” like. However, if the data set is very large or is multidimensional, human verification could become difficult. To find a reasonable number of clusters, many existing methods must be run repeatedly with different parameters, and are impractical for real-world data sets that are often quite large. We desire an algorithm that can efficiently determine a reasonable number of clusters/segments to return from any hierarchical clustering/segmentation algorithm. We have to identify the correct number of clusters to return from a hierarchical clustering/segmentation algorithm. The paper is divided into five parts. First part introduced with the concept of color segmentation. Next part provides motivation to research work. Third part presents segmentation methodology. Fourth part describes result and analysis followed by conclusion and future scope.

II. MOTIVATION TO WORK

In cluster analysis, a fundamental problem is to determine the best estimate of the number of clusters, which has a deterministic effect on the clustering results. However, a limitation in current applications is that no convincingly acceptable solution to the best-number-of-clusters problem is available due to high complexity of real data sets. Choosing an appropriate clustering method is another critical step in clustering. A large number of clustering methods are available for cluster analysis. However a fundamental problem in applying most of the existing clustering approaches is that the number of clusters needs to be pre-specified before the clustering is conducted. The clustering results may heavily depend on the number of clusters specified [5]. It is necessary to provide educated guidance for determining the number of clusters in order to achieve appropriate clustering results. At the current stage of research, none of the existing methods of choosing the optimal estimate of the number of clusters is completely satisfactory. The gap method was recently proposed by Tibshirani et al [7]. The main idea of the gap method is to compare the within cluster dispersions in the observed data to the expected within cluster dispersions assuming that the data came from an appropriate null reference distribution. Simulation results reported by [7] indicated that the gap method is a potentially powerful approach in estimating the number of clusters for a data set. However, recent studies have shown that there are situation where the gap method may perform poorly. For example, data clusters consisting of objects from well separated exponential populations. The

correct choice of k is often ambiguous, with interpretations depending on the shape and scale of the distribution of points in a data set and the desired clustering resolution of the user. In addition, increasing k without penalty will always reduce the amount of error in the resulting clustering, to the extreme case of zero error if each data point is considered its own cluster (i.e., when k equals the number of data points, n). Intuitively then, the optimal choice of k will strike a balance between maximum compression of the data using a single cluster, and maximum accuracy by assigning each data point to its own cluster. If an appropriate value of k is not apparent from prior knowledge of the properties of the data set, it must be chosen somehow. There are several methods of finding the optimal number of clusters such as Davies Bouldin Index, Dunn's Index and many other Validity Measures based on similarity and dissimilarity concept [8].

Our research objective will be to propose a better clustering algorithm for color image segmentation. The k -means algorithm is mainly used for clustering and image segmentation but its main problem is that it uses random initial centers which greatly affect its performance. Another problem of this method is that it can't decide about the number of cluster in advance for the unknown data. K -means method can find out optimum number of clusters in image segmentation if used with gap method. Another algorithm of image segmentation is agglomerative top-down method which initially segments an image in large number of clusters and then similar clusters can be merged to reduce number of clusters but its major problem is that it cannot find optimum number of cluster up to which clusters can be merged.

We need to find the method to find the optimal number of cluster for K -means clustering then validate the validity of the clusters, after that we need to find the criteria through which we can merge the similar cluster which can be similar in terms of the variance or the distance measure. We have to assume each pixel of the input image as a data point, hence there will be n data points or we can say n patterns $x_1, x_2, x_3, \dots, x_n$ generated from the input image which can have n number of pixels. After that we need to apply K -means clustering algorithm to the whole set of patterns so that we can generate the large number of clusters, where each cluster contains the similar type of data points, then we need to find the method to select the two clusters from these large number of clusters for merging. The merging will be based on some similarity between each of the existing cluster pairs, to find the similarity between the existing cluster we need to consider the variance of each individual clusters and also the distance between the two cluster centers. We need to develop an algorithm which can merge the similar looking clusters based on similarity criteria.

III. RESEARCH METHODOLOGY

For image segmentation of image we have used an agglomerative bottom-up algorithm in which similar pixels are merged to one cluster. Colored images have three color-features; red green and blue. For image segmentation, first of all, the image is flattened in 2-dimensional array having columns representing three basic colors. Then flattened

image is clustered using some clustering algorithm such as k -means.

Although K -means is most used algorithm for image segmentation but the main disadvantage of the k -means algorithm is that the number of clusters, K , must be supplied as a parameter. There are several ways to find the optimal number of clusters like Davies-Bouldin index or some other validity measure [9]. To find the optimal number of cluster we have used a validity measure which is based on intra-cluster and inter-cluster distance measures. An intra-cluster distance measure is the distance between a point and its cluster center. An inter-cluster distance is the distance between clusters' center. Smallest of these inter-cluster distances is considered. A validity index is taken as

$$\text{Validity index} = \frac{\text{Avg(Intra - cluster Distance Measures)}}{\text{min(Inter - cluster Distance Measures)}}$$

Clustering which gives a min value for validity measure will give the ideal value of k .

Once optimum number of clusters is selected then merge algorithm is followed in which first k number of clusters is formed using k -means. Now out of these k clusters two most similar clusters are selected and then merged reducing the total number of clusters to $k-1$. The similar procedure is formed till number of cluster reduces to optimum number which we have already found using k -means algorithm and validity index. The complete methodology can be explained in the following steps.

A. Algorithm:

Step 1: First optimum number of clusters is found using k -means algorithm and validity index. The number of cluster giving lowest value of validity index is optimum number of clusters.

Step 2: Segment the image in large number of clusters using k -means. If n is optimum number of cluster given by step1 then $k >> n$.

Step 3: Then two clusters are selected for merging from these k number of clusters by using the similarity or criterion function. The criterion function measures the similarity between each existing cluster pairs, by considering the variance of each individual cluster, the variance of the merged cluster and the distance between two cluster centers as given in the following section. It reduces the number of clusters to $k-1$.

Step 4: Repeat step3 until $k=n$.

The Davies-Bouldin index introduced by David L. Davies and Donald W. Bouldin in 1979 is a metric for evaluating clustering algorithms [8]. This is an internal evaluation scheme, where the validation of how well the clustering has been done is made using quantities and features inherent to the dataset. The Davies-Bouldin index can be calculated by the following formula:

Let C_1, C_2, \dots, C_k be the cluster then DB measure is defined as

$$DB = \frac{1}{k} \sum_{i=0}^k R_i \quad (1)$$

Where $R_i = \max_{i \neq j} \{R_{ij}\}$ and $j = 1, 2, \dots, k$

$$R_{ij} = \left[\frac{(\sigma_i^2 + \sigma_j^2)}{d_{ij}^2} \right] \quad (2)$$

Where σ_i^2 & σ_j^2 are the variance of cluster C_i and C_j .

Since algorithms that produce clusters with low intra-cluster distances (high intra-cluster similarity) and high inter-cluster distances (low inter-cluster similarity) will have a low Davies–Bouldin index, the clustering algorithm that produces a collection of clusters with the smallest Davies–Bouldin index is considered the best algorithm based on this criterion[9]. This is value is considered as the validity index as explained in the previous section.

IV. RESULTS & ANALYSIS

For finding the performance & result analysis, four jpeg images are selected for demonstration. The original images are given in table 1. These images are selected due to their

varying features such as background image, number of colors etc. If an image which is represented with less number of colors should not be segmented in high number of clusters and an image which is made of high numbers of colors should be represented with sufficient number of clusters so that its outlines is visible clearly. The test image1 is modeled image made of only 5 colors without any background. Test Image 2 is a cartoon image also has less colors but with a clearly visible and outlined background. The test image 3 is having large colors with clear visible background image while last image has many varying colors but with less visible background.

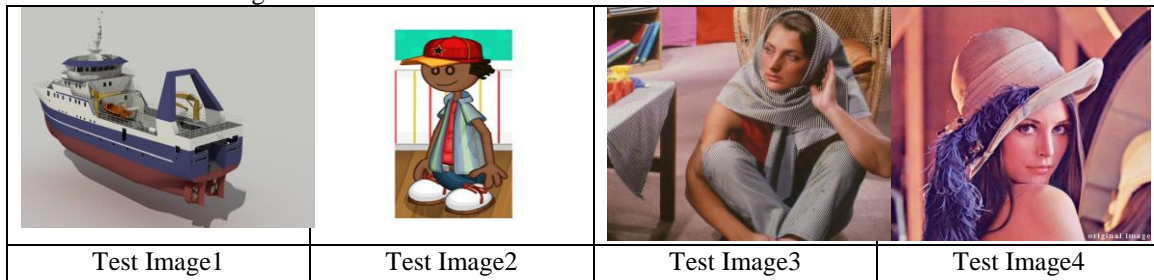


Table 1: Test Images

In first step as explained in algorithm in the previous section, k-means is used with different k values. We have taken range of k from 3 to 20 in our experimentation. The affectivity of clustering is measured by a validity index. The least value of this is desired. At the end of this step these validity index is plotted and the k-value which is giving least value is taken as the optimum number of clusters. In the second steps k-value is taken very high. In our experimentation this value is fixed as 30. Higher the value of k is there better segmentation can be achieved. This segmentation is created using k-means algorithm. Finally out of these 30 clusters pair of clusters with maximum similarity is chosen for merging. Before this if there is any empty cluster, then it is deleted. This process of merging goes on until we remain with optimum number of clusters we have calculated in the first step. Then after merging again performance is clustered using validity index. The results are shown in table 2

3	Test Image3	7-8	.1867	0.0466-0.1051
4	Test Image4	5-7	0.1722-0.2451	0.0162-0.0517

Table 2: Experimentation Results

If we analyze the result from table, we can easily find out that merge algorithm outperform k-means algorithm if validity index criterion is followed. First image is the simplest one which is made of only few colors and is without any background. Both method k-means as well as merge algorithm performs well but later outperform k-means in all ten experiments. Similar is the case with test image2.

Sr. No.	Image	Optimum number of clusters range	K-means Validity range in 10 experiments.	Validity after merging range in 10 experiments
1	Test Image1	5	0.0121-0.0758	0.0069-0.0454
2	Test Image2	6-8	0.1021-.1779	0.0128-.1079

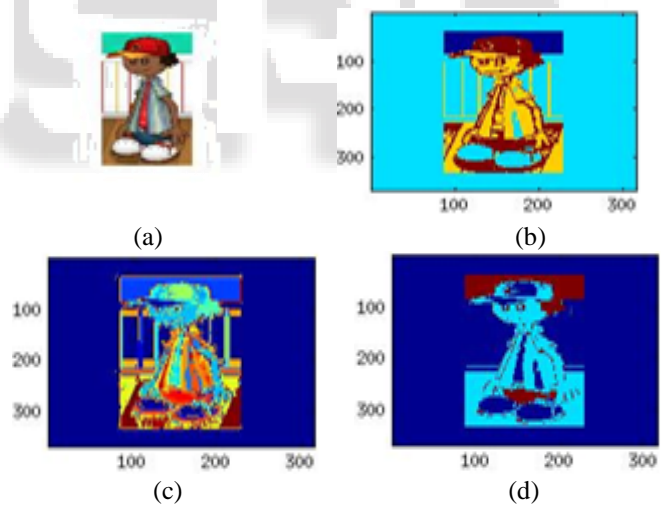
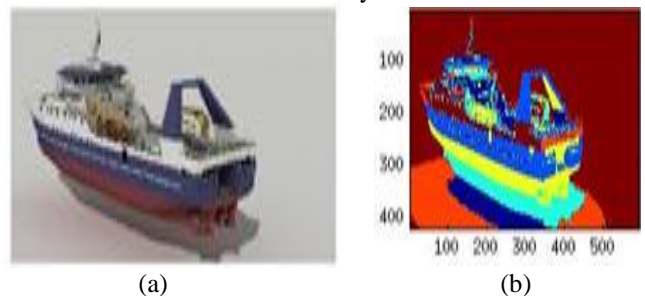


Fig. 1: Experimentation Results of Beaver image (a) Original Image (b) k-means clustering with optimum cluster 5 with validity index 0.0452 (c) k-means clustering with cluster 30 (d) merging based clustering with optimum cluster 5 with validity index 0.0069



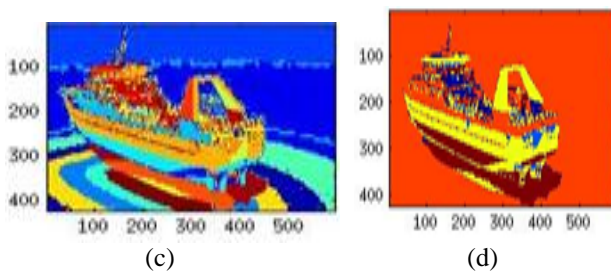


Fig. 2: Experimentation Results of Car image (a) Original Image (b) k-means clustering with optimum cluster 5 with validity index 0.0716 (c) k-means clustering with cluster 30 (d) merging based clustering with optimum cluster 5 with validity index 0.0194

In test image3 experiments, one time out of ten experiments, k-means performs better only when number of optimum cluster is more than eight. This may be due to unclear background in the image. The last image is complex where more than 10 clusters are chosen for segmentation. In every experiment, merge based algorithm perform far better as compared to k-means algorithm. So, through our experiments we have found that merging based algorithm is better than k-means if we use a Davies-Bouldin Similarity measure.

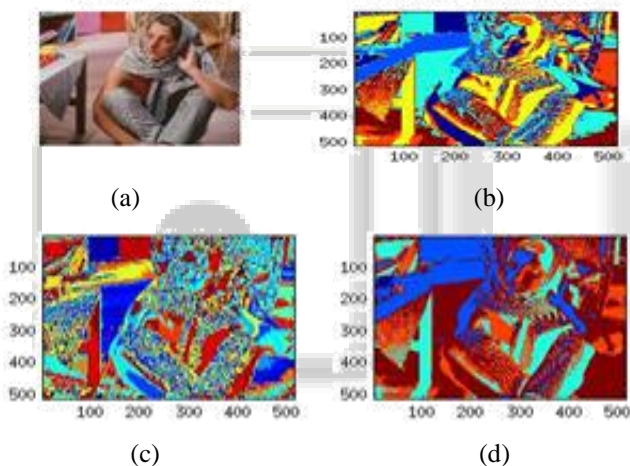


Fig. 3: Experimentation Results of Rose image (a) Original Image (b) k-means clustering with optimum cluster 7 with validity index 0.1867 (c) k-means clustering with cluster 30 (d) merging based clustering with optimum cluster 7 with validity index 0.0354

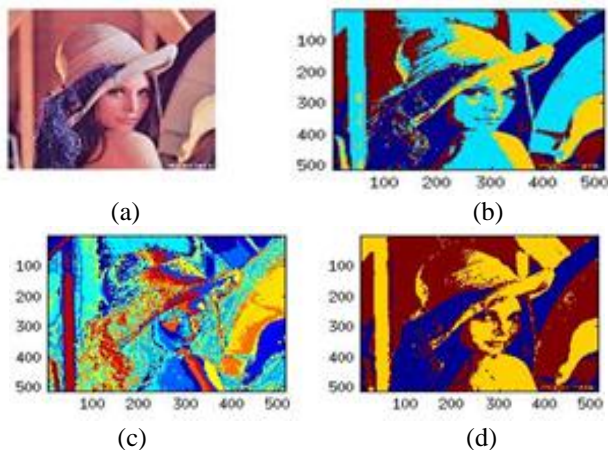


Fig. 4: Experimentation Results of Rose image (a) Original Image (b) k-means clustering with optimum cluster 5 with validity index 0.1722 (c) k-means clustering with cluster 30 (d) merging based clustering with optimum cluster 5 with validity index 0.0194

(d) merging based clustering with optimum cluster 5 with validity index 0.0191

V. CONCLUSION

The dissertation proposes an improved method of image segmentation by using k-means and merge algorithm of image segmentation where similarity measure is taken from Davies-Bouldin Index. The k-means algorithm is used to find optimum number of clusters and then this value is used as a limiting value in merge algorithm. The performance of algorithm is measured using a validity index which is measured by two factors. The first factor is intra-cluster distance whose minimum value is desired and another is inter-cluster distance for which a maximum value is required. Once optimum number of cluster is found then k-means clustering algorithm is again applied to generate large number of clusters, then from these large numbers of clusters, pair of clusters with most similar characteristics are merged iteratively until number of clusters are reduced up to optimum number of clusters. Four images have been taken for experimentations. More than 95% of times the proposed merge algorithm is performing better than simple k-means algorithm.

REFERENCE

- [1] Rajesh Dass, Priyanka, Swapna Devi, "Image Segmentation Techniques", IJECT Vol.3, Issue 1, Jan-March 2012.
- [2] S.K Somasundaram, P.Alli," A Review on Recent Research and Implementation Methodologies on Medical Image Segmentation", Journal of Computer Science 8(1): 170-174, 2012.
- [3] Salem Saleh Al-amri, N.V Kalyankar, Khamitkar S.D, "Image Segmentation by using threshold Techniques", Journal of computing, Volume 2, issue 5, may 2010.
- [4] Bo Peng, Lei Zhang , David Zhang, "Automatic Image Segmentation by Dynamic Region Merging", IEEE Transactions on image processing, Vol.20, No. 12 December 2011.
- [5] R. Patil and K. Jondhale, "Edge based technique to estimate number of clusters in k-means color image segmentation," in Proc. 3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT), pp. 117-121, 2010.
- [6] K. Jain, Fundamentals of Digital Image Processing, Englewood Cliffs, NJ: Prentice-Hall, 1989.
- [7] Robert Tibshirani, Guenther Walther and Trevor Hastie, "Estimating the Number of Data Clusters via the Gap Statistic" J.R. Statist. Soc. B (2001), 63, pp. 411—423
- [8] Deborah L, Baskaran R, Kannan A (2010) A survey on internal validity measure for cluster validation. International Journal of Computer Science & Engineering Survey 1 (2) 85–102. doi: 10.5121/ijcses.2010.1207
- [9] Davies, David L.; Bouldin, Donald W. (1979). "A Cluster Separation Measure". IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-1(2):224–227. doi:10.1109/TPAMI.1979.4766909