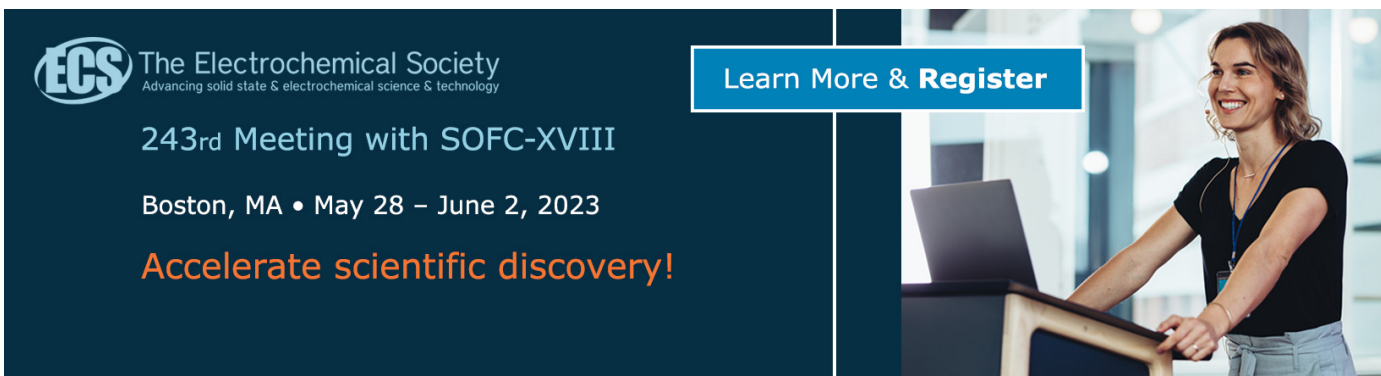**PAPER • OPEN ACCESS**

Car shape clustering using sobel edge detection with divisive average linkage and single linkage algorithm (case: bus, sedan, citycar, mpv, and truck)

To cite this article: Victor Hadi Saputra *et al* 2020 *IOP Conf. Ser.: Mater. Sci. Eng.* **1007** 012136

View the article online for updates and enhancements.

Car shape clustering using sobel edge detection with divisive average linkage and single linkage algorithm (case: bus, sedan, citycar, mpv, and truck)

Victor Hadi Saputra*, Dyah Erny Herwindiati and Tri Sutrisno
Informatics Engineering Department, Faculty of Information Technology
Universitas Tarumanagara Jakarta Indonesia 11440

* victor.535160035@stu.untar.ac.id

Abstract. This study aims to make an application that able to make clustering for car type based on shape by using edge detection and clustering algorithm. The car types being used in this study are the image of the bus, sedan, city car, mpv, and truck. In this study, the method used to classify is with the divisive average linkage and single linkage method. The goal is to prove whether the divisive average linkage and single linkage can be used to classify the shape of the car image correctly. The shape obtained by using the Sobel edge detection method because the Sobel method is an edge detector that has advantages in reducing the noise before doing calculations so that the resulting edges form an image like the original.

1. Introduction
Clustering is a process to group a physical or non-physical object into groups or classes with the same similarity. Clustering is said to be good if it produces a high degree of similarity in a class. The clustering method is divided into two categories, namely hierarchical clustering, and non-hierarchical clustering, where hierarchical clustering is a group analysis method that attempts to build a hierarchy of data groups, grouping strategies are divided into two types namely Agglomerative and Divisive. In this research, car shape grouping will be done using the Sobel edge detection with divisive average linkage and single linkage clustering methods. The car types being used are the image of the bus, sedan, city car, mpv, and truck.

2. Method and materials
2.1. Sample preparation
The data sample to be used is car images taken from the Google search engine. The image data used has a size of 216x216 pixels and has a .jpg or .jpeg format with a white background. The car types being used are the image of the bus, sedan, city car, mpv, and truck.

Figure 1. Sample cars image data

2.2. Method

In this application, the Sobel edge detection method is used to get the shape of the car's image that the result will be followed by the clustering process using divisive average linkage and single linkage method. After the object is grouped into multiple clusters, it will be evaluated on the quality of the cluster produced through a classifying process that aims to test whether the cluster formed in the research is good or not by using the Evaluation Sillhouette.

2.2.1. Sobel Edge Detection

Sobel Operator is the development of Robert's method by using the HPF filter which is giving one zero number buffer. This method takes the principle of the Laplacian and Gaussian functions known as functions for generating HPF. The advantages of this Sobel method are the ability to reduce noise before doing edge detection calculations [1].

1. A review of the pixel settings around its pixels (x, y) is

$$\begin{bmatrix} a_0 & a_1 & a_2 \\ a_7 & (x, y) & a_3 \\ a_6 & a_5 & a_4 \end{bmatrix}$$

2. With constants C = 2. In the form of kernel (mask), Sx and Sy can be expressed as follows:

$$S_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \quad S_y = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}$$

The x coordinate is defined as increasing in the right-direction and the y coordinate is defined as increasing in the down-direction

3. The formula to find the gradient from the Sobel operator is

$$S_x = (a_2 + ca_3 + a_4) - (a_0 + ca_7 + a_6) \tag{1}$$

$$S_y = (a_0 + ca_1 + a_2) - (a_6 + ca_5 + a_4) \tag{2}$$

$$M = \sqrt{Sx^2 + Sy^2} \tag{3}$$

Information:
M = Magnitude
Sx = X-axis Gradient
Sy = Y-axis Gradient

Figure 2. Transformation from RGB image into sobel image

2.2.2. Divisive

The divisive method is a clustering method with a top-down principle [2]. Clustering with the top-down principle is initiated from all data objects in a single cluster division cluster running on every iteration. The Divisive method belongs to the category hierarchical clustering on each step there is a cluster addition until finally all the elements in the data set are included in the cluster. This means that if there is N data then the Divisive step method runs as much as N-1 times.

2.2.2.1. Divisive Single Linkage

The Single Linkage Method is a classifying process based on the closest distance between objects.If two objects are separated by a short distance, then both objects will merge into a single cluster and so on.To better understand how this method works note the following algorithms [3]:
1. Euclid distance matrix form for given sample data matrix.
2. Assume each data is considered as a cluster, then determine which cluster has the closest distance, for example U cluster and cluster V have the closest distance then combine, the combined result is the UV cluster.
3. From the UV cluster that has been formed look for the minimum distance between the UV cluster and other clusters (objects) that have not yet joined, the new distance matrix obtained is called D (2). For example d (uv) w = min (duw, dvw), then the newly formed cluster is (UVW).
4. Repeat step 2 until all objects are joined in one group.

2.2.2.2. Divisive Average Linkage

Average Linkage treats the distance between two clusters as the average distance between all data in a cluster. The Divisive Average method uses averages to calculate the distance of data to an existing cluster and then breaks up a cluster based on the average value of the distance between each data in the cluster. If a data has the largest average distance then the data is broken down into new clusters or called splinter. The algorithm with the method of Avarage Linkage is [4]:
1. Find data that has the largest mean value of other data. This data creates a new cluster (splinter).
2. For any data outside the calculated splinter group At = [average d (i, j) j∈Rsplinter group] - [average d (i, j) j Rsplinter group].
3. Find the data that has the largest positive Dh value, the data means the closest distance to the splinter group. Move the data to the splinter group.
4. Repeat steps 2 and 3 until all Dh differences are negative. At that time the data has been divided into 2 clusters.
5. Select the cluster to be in the cluster. This cluster is broken down in the next iteration.
6. Repeat steps 1-5 until all clusters have only one member or up to the desired number of clusters.

2.2.3. Eucledian Distance

Euclidean Distance method is used to calculate the distance between data. This measurement is based on the value of objects in each k dimension in learning. Euclidean Distance uses theorem Pythagoras and object measurement is not limited to 2 dimensions or more. For Euclidean

Distance with 2 objects (d12), the distance between the 2 objects is not more than the length of the triangle hypotenuse[5].

Euclidean Distance formulas for more than 2 dimensions [6]:

$$D(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{4}$$

Information:
D (x, y) = many dimensions of the object
x_i = first dimension of object x
y_i = first dimension of object y

2.2.4. Silhouette evaluation

Silhouette Evaluation is one method for evaluating a cluster. This method is used to test the quality of clusters produced through the clustering process. The steps to calculate the Silhouette value are as follows [7]:

1. For each object i, calculate the average distance from object i with all objects in one cluster. The result of the average value is called ai.
2. For each object i, calculate the average distance from object i to objects in other clusters. Of all the average distances, the smallest value is taken. This value is called bi.
3. After calculating in stages 1 and 2, the silhouette coeficient values are obtained:

$$S_i = (b_i - a_i)/\max(a_i, b_i) \tag{5}$$

Silhouette Coeficient value calculation results can vary between -1 to 1. Clustering results are said to be good if the value of the silhouette is positive. A positive value silhouette indicates that the data contained in the Cluster is in the right Cluster where the average distance between objects in one cluster is smaller than the average distance between objects with objects that are in another Cluster. Subjective Table Criteria of quality groupings based on silhouette coefficients [8]:

Nilai Silhouette	Interpretasi
0,71-1,00	*Strong Cluster*
0,51-0,70	*Good Cluster*
0,26-0,50	*Weak Cluster*
0,00-0,25	*Bad Cluster*

3. Results and discussion

3.1. Divisive Average Linkage Testing

The following is a average linkage test using 2 types of image data that are different from each type of car, the image data of the car used is a type of image of a sedan, truck, bus, city car and mpv using 30 images. Examples of the following test results

Figure 3. Average Linkage test with 30 data images

3.2. Divisive Single Linkage Testing

The following is a single linkage test using 2 types of image data that are different from each type of car, the image data of the car used is a type of image of a sedan, truck, bus, city car and mpv using 30 images. Examples of the following test results
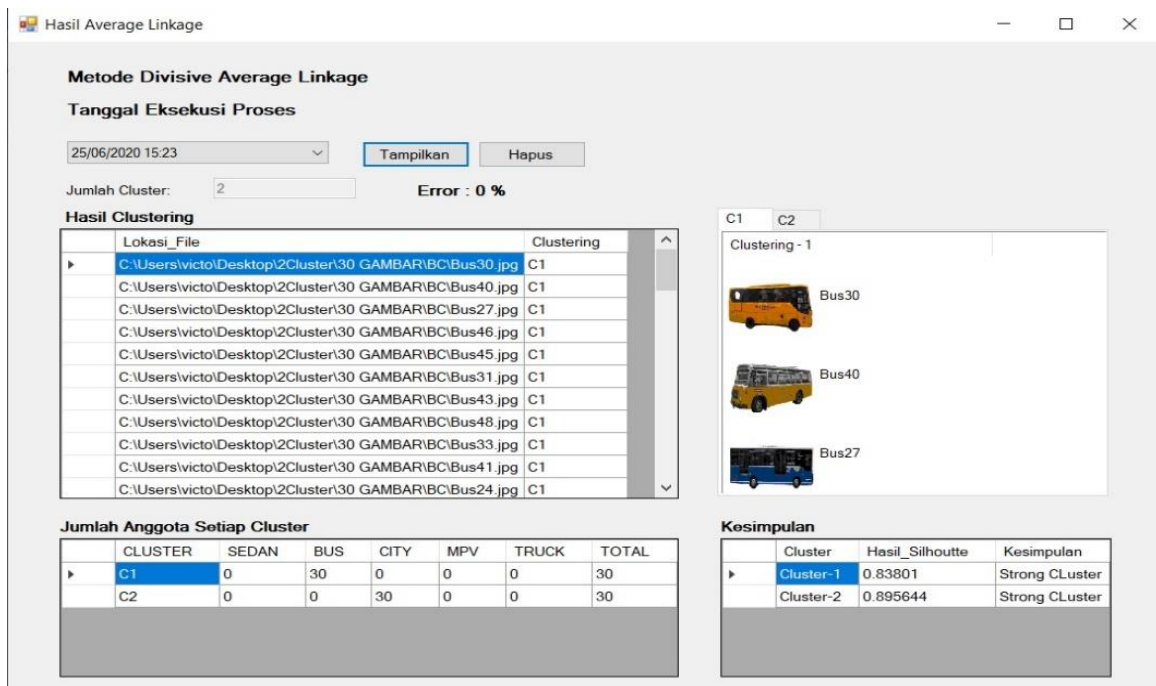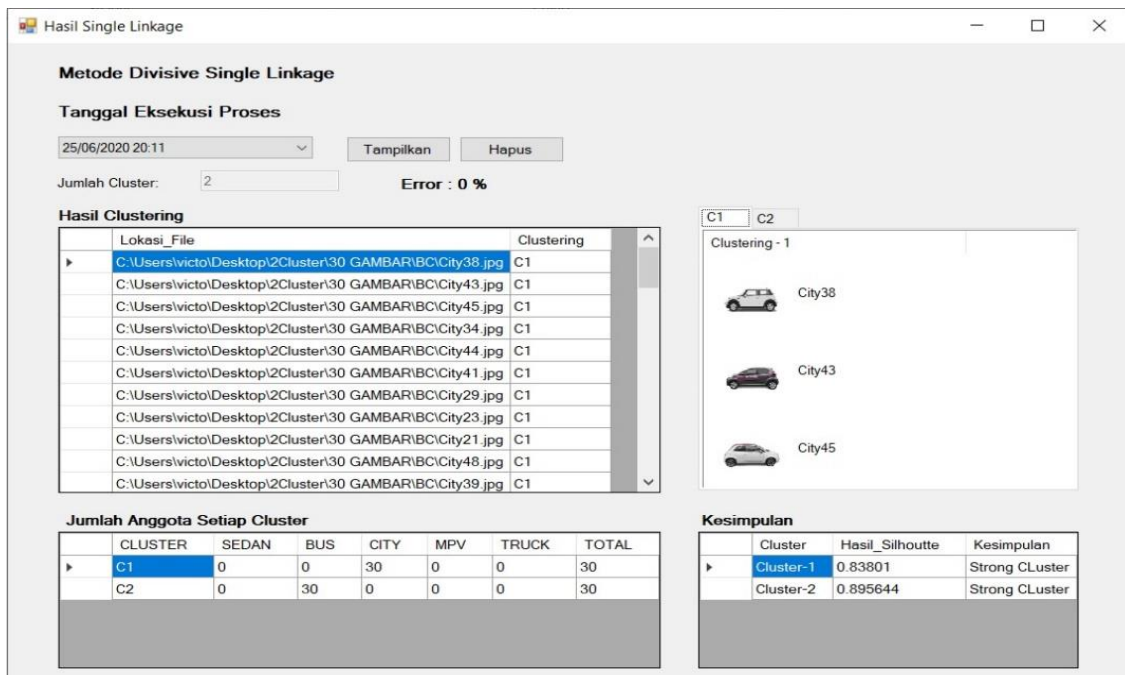

Figure 4. Single linkage test sample with 30 data images

3.3. Analysis of Divisive Clustering Using a Different Data

The process of clustering is divided into 3 stages of experiment by using a different amount of data that is using 30 image data, 50 image data and 60 image data. The goal is to find out the lowest and highest error rates in the cluster.

Table 1. The result of testing with a different data with single linkage method

Cluster	30 Images	50 Images	60 Images	Minimum
2 Cluster	17,49%	15,6%	16,08%	**50 Images**
3 Cluster	38,95%	31,02%	28,11%	**60 Images**
4 Cluster	38,78%	46%	51,54%	**30 Images**
5 Cluster	60,6%	58,4%	59,6%	**50 Images**

From the results in **Table 1**. it can be seen that the best use of data for the single linkage method among 30 image data, 50 image data and 60 image data is the use of 50 image data.

Table 2. The result of testing with a different data with average linkage method

Cluster	30 Images	50 Images	60 Images	Minimum
2 Cluster	5,6%	3,4%	4,4%	**50 Images**
3 Cluster	11,3%	8,7%	9,2%	**50 Images**
4 Cluster	16,64%	14%	16,06%	**50 Images**
5 Cluster	21,3%	21,3%	25,67%	**30 Images**

From the results in **Table 2.** it can be seen that the best use of data for the average linkage method among 30 image data, 50 image data and 60 image data is the use of 50 image data.

3.4. Analysis of Divisive Clustering Using a Different Cluster

The following is an analysis of the difference in results between using average linkage and single linkage clustering using different number of clusters. The goal is to find out the lowest error rate in a cluster by using the single linkage and average linkage methods.

Table 3. The result of testing with a different cluster beetwen average and single linkage

Cluster	Average Linkage	Single Linkage	Minimum Error
2 Cluster	4,56%	16,39%	**Average**
3 Cluster	9,78%	32,69%	**Average**
4 Cluster	15,57%	45,44%	**Average**
5 Cluster	22,73%	59,3%	**Average**

From the results in Table 3. it can be seen that the best cluster usage for the average linkage method between 2 clusters, 3 clusters, 4 clusters and 5 clusters is on the use of 2 clusters with an error percentage of 4.5%. The best method for clustering that using sobel edge detection is divisive average linkage method with an average error rate under 23%.

3.5. Clustering Result Using the Silhouette Method

The Silhouette method is used to evaluate where if the value is close to -1, then the cluster is declared unfavorable (Bad Cluster) and if the value is close to +1 then it is considered good, then the values obtained as follows:

Table 4. The result testing of single linkage with silhouette method

Cluster	30 Images	50 Images	60 Images	Silhouette
2 Cluster	0,586605	0,606254	0,594233	**Good**
3 Cluster	0,117733	0,264952	0,272264	**Bad**
4 Cluster	0,130813	0,010446	-0,00875	**Bad**
5 Cluster	0,04282	-0,03492	-0,2152	**Bad**

In the **Table 4.** based on the result that have been done, the rate of divisive single linkage method has so much bad evaluation results. The results are quite good by using only 2 clusters.

Table 5. The result testing of single linkage with silhouette method

Cluster	30 Images	50 Images	60 Images	Silhouette
2 Cluster	0,76549	0,74918	0,74090	**Strong**
3 Cluster	0,571271	0,496403	0,527584	**Good**
4 Cluster	0,387429	0,402373	0,371211	**Weak**
5 Cluster	0,116906	0,1594046	0,135816	**Bad**

In the **Table 5.** based on the result of the tests that have been done explain that the divisive average linkage method has an average with good evaluation results. The results are quite good between using 2 until 4 clusters.

3.6. Average Time Processing of clustering test between single and average linkage
The following is the results of clustering based on the length of time used for the program to clustering using the divisive single and average linkage method. The test data used is using 30 images, 50 images and 60 images.

Table 6. The average time processing between single and average linkage

Cluster	Average Linkage	Single Linkage	Minimum Time
2 Cluster	19,05 Sec	16 Sec	**Single**
3 Cluster	24,99 Sec	21,93 Sec	**Single**
4 Cluster	31,09 Sec	28,18 Sec	**Single**
5 Cluster	38,48 Sec	33,02 Sec	**Single**

From the result on Table 6. the fastest processing time is to use are single linkage method, because in a single linkage the calculation process uses only the smallest value, whereas in average linkage, the values must be averaged first.

4. Conclusion

Based on the testing result that has been done using a car shape grouping application program with divisive average and divisive single lingkage, it can be concluded that average linkage are better in grouping shapes by using the sobel edge method but more longer than single linkage in processing time because in a single linkage the calculation process uses only the smallest value, whereas in average linkage, the values must be averaged first. Also by using more clusters causing the result more to be worse, because if the shape of data is similar it will cause more errors.

5. References

[1] Subchan Ajie Ari Bowo, Achmad Hidayanto, R.Rizal Isnanto, 2016 Analisis Deteksi Tepi Untuk Mengidentifikasi Pola Daun.

[2] Astrid Alfira, Farini Hemin, Eti Dwi Wiraningsih, 2006 Analisis Hybrid Mutual Clustering menggunakan Jarak Square Eucledian.

[3] Dewi Rachmatin, 2014 Aplikasi Metode-Metode Agglomerative Dalam Analisis Klaster Pada Data Tingkat Polusi Udara.

[4] Silvia Ningsih, Sri Wahyuningsih, Yuki Novia Nasution, 2016 Perbandingan Kinerja Metode Complete Linkage dan Average Linkage Dalam Menentukan Hasil Analisis Cluster Vol **1.**

[5] Adi Pamungkas, 2018 Pencocokan Citra.

[6] M.Nishom, 2016 Perbandingan Akurasi Eucledian Distance, Minkowski Distance, dan Manhattan Distance pada Algorimta K-Means Clustering Berbasis Chi-Square.

[7] Herwin, 2015 Sistem Prediksi Customer Loyality Dengan Metode K-Means.

[8] Andrea Tri Rian Dani, Sri Wahyuningsih, Nanda Arista Rizki, 2019 Penerapan Hierarchial Clustering Metode Agglomerative Pada Data Runtun Waktu Vol **1.**