INTERNATIONAL JOURNAL
ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Hierarchical and K-means Clustering in the Line Drawing Data Shape Using Procrustes Analysis

Ridho Ananda[a*], Agi Prasetiadi[b]

[a] *Faculty of Industrial Engineering and Design, Institut Teknologi Telkom Purwokerto, Purwokerto, 53147, Indonesia*
[b] *Faculty of Informatic, Institut Teknologi Telkom Purwokerto, Purwokerto, 53147, Indonesia*
*Corresponding author: *ridho@ittelkom-pwt.ac.id*

***Abstract*— One of the problems in the clustering process is that the objects under inquiry are multivariate measures containing geometrical information that requires shape clustering. Because Procrustes is a technique to obtaining the similarity measure of two shapes, it can become the solution. Therefore, this paper tried to use Procrustes as the main process in the clustering method. Several algorithms proposed for the shape clustering process using Procrustes were namely hierarchical the goodness-of-fit of Procrustes (HGoFP), k-means the goodness-of-fit of Procrustes (KMGoFP), hierarchical ordinary Procrustes analysis (HOPA), and k-means ordinary Procrustes analysis (KMOPA). Those algorithms were evaluated using Rand index, Jaccard index, F-measure, and Purity. Data used was the line drawing dataset that consisted of 180 drawings classified into six clusters. The results showed that the HGoFP, KMGoFP, HOPA and KMOPA algorithms were good enough in Rand index, F-measure, and Purity with 0.697 as a minimum value. Meanwhile, the good clustering results in the Jaccard index were only the HGoFP, KMGoFP, and HOPA algorithms with 0.561 as a minimum value. KMGoFP has the worst result in the Jaccard index that is about 0.300. In the time complexity, the fastest algorithm is the HGoFP algorithm; the time complexity is 4.733. Based on the results, the algorithms proposed in this paper particularly deserve to be proposed as new algorithms to cluster the objects in the line drawing dataset. Then, the HGoFP is suggested clustering the objects in the dataset used.**

***Keywords*—Hierarchical; K-means; clustering; data shape; Procrustes.**

Manuscript received 26 Feb. 2021; revised 8 May 2021; accepted 18 Aug. 2021. Date of publication 30 Sep. 2021.

I. INTRODUCTION

Today is the big data era where various activity about anything is mostly saved as a data. Not only does the data become evidence of events that occurred, but it also can be utilized to get more information about the description or prediction of a particular phenomenon. That information is certainly able to be used for deciding. To obtain those, we need the data analysis methods; one of them is the clustering method. Clustering is part of data mining techniques where it is quite popular. In clustering, objects are grouped by maximizing similarity measures among objects in the same group and minimizing similarity among objects in the different groups [1]. The procedure of clustering is divided into two, namely hierarchy and non-hierarchy. The examples of hierarchical clustering are single linkage, complete linkage, average linkage, and ward linkage. Whereas for non-hierarchical clustering, K-means algorithm is the most popular example. Until now, clustering has been widely used in various fields; for instance, it has been used in education[2]–[6], environment [7], [8], health [9], and technology fields [10], [11].

Sometimes the objects under inquiry are observed by features which are angular measures such as degree and direction. Furthermore, there are certain conditions in which the gathered data are multivariate measures containing geometrical information. These all are instances of the shape dataset. Therefore, it is needed shape analyses to get information on the data. The analyses are performed by taking a finite number of object points that can represent the shape of that object. A finite number of object points are called landmarks [12]. Two shapes of objects have the same shape if, after their landmarks are translated, rotated, and dilated to each other, that shapes match exactly. Procrustes is one of the tools used to obtain dissimilarity measures of that landmark in statistical shape analysis.

Procrustes refers to a procedure of matching two landmarks and producing a measure of the dissimilarity. Supposed that X and Y are the matrices of landmarks, to measure the

difference between those landmarks, Procrustes utilize the sum of the squared distance E given by $E(\mathrm{X},\mathrm{Y}) = \|\mathrm{Y}-\mathrm{X}\|_F^2$. Geometrically, Procrustes will minimize $E(\mathrm{Y},\mathrm{X})$ by using series of Euclidean similarity transformations, i.e., translation, rotation, and dilation[13]. The first formula of Procrustes was the ordinary Procrustes analysis (OPA). Then, Procrustes had been developed into the generalized Procrustes analysis (GPA)[12] and the Goodness-of-fit of Procrustes (GoFP)[14].

GoFP is the Procrustes procedure to measure the similarity of X and Y by using optimal translation-normalization-rotation-dilation, denoted by $GoFP(\mathrm{X},\mathrm{Y})$. The value of $GoFP(\mathrm{X},\mathrm{Y})$ is in [0,1]. If $GoFP(\mathrm{X},\mathrm{Y}) \approx 1$, then it means that X and Y have an excellent match. Conversely, if $GoFP(\mathrm{X},\mathrm{Y}) \approx 1$, then X and Y have the poor match. One of the advantages of GoFP is that it has the symmetrical property, $GoFP(\mathrm{X},\mathrm{Y}) = GoFP(\mathrm{Y},\mathrm{X})$. Today, $GoFP(\mathrm{X},\mathrm{Y})$ not only has been utilized to measure the match of two configurations but also evaluate the performance of biplot analysis[15], [16], process the variables selection algorithm[17], measure the quality of imputation data[18], [19], and detect outliers[20].

Based on the description above, we want to try using GoFP in the shape clustering process on shape data. As a result, we propose a shape clustering algorithm by using GoFP. Results of the proposed algorithm are then compared with the shape clustering algorithm by using OPA[21]. Datasets of line drawings are used in this paper to see the performance of the algorithm proposed.

This paper provides a brief review of Procrustes analysis in section 2. The basic idea of the shape clustering algorithm by using GoFP and the validity measures are provided in section 3. Then, the brief description of data used, simulation study, and discussion are given in section 4. In the last section, we provide the conclusion of the result in this paper.

II. Material and Method

A. The Goodness-of-fit of Procrustes

In ancient Greek, Procrustes was a bandit who offered inn on an iron bed to any traveler on the road from Eleusis to Athens. If the traveler did not fit Procrustes's bed, he would torture that guest to make a perfect fit with his bed by stretching their limbs or cutting them off. In mathematics, Procrustes was a technique of matching two shapes and producing a measure of the match. Those shapes were provided as configuration matrices of the same size. Suppose X and Y are n-by-p and m-by-p configuration matrices, respectively. If $n > m$ then Y needs to be optimally matched to X by adding l-by-p matrix where $l = n - m$. Procrustes utilizes the squared of the Euclidean norm E between the points in Y and the corresponding points in X, also known as Procrustes distance, provided by Equation 1.

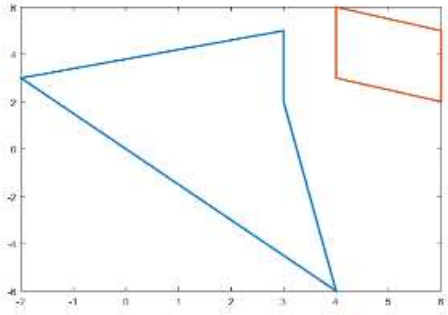
$$E(\mathrm{Y},\mathrm{X}) = \|\mathrm{Y}-\mathrm{X}\|_F^2 \qquad (1)$$

Geometrically, Procrustes works to minimize $E(\mathrm{Y},\mathrm{X})$ by using series of Euclidean similarity transformations, namely translation, rotation, and dilation. The optimal translation of X and Y are $\mathrm{X}_T = \mathrm{X} - \frac{1}{n}1_n 1_n'\mathrm{X}$ and $\mathrm{Y}_T = \mathrm{Y} - \frac{1}{n}1_n 1_n'\mathrm{Y}$, where n and 1_n are the number of rows and n-by-1 vector having each component equal to 1, respectively. The optimal rotation is derived by using the complete form of singular value decomposition (CFSVD) of $\mathrm{X}'\mathrm{Y}$, i.e., $\mathrm{X}'\mathrm{Y} = \mathrm{U}\Sigma\mathrm{V}'$, we get the orthogonal matrix $\mathrm{Q} = \mathrm{V}\mathrm{U}'$ that is utilized to achieve optimal rotation. Optimal dilation is provided by scalar $c = \frac{trace(\mathrm{X}'\mathrm{Y})}{trace(\mathrm{Y}'\mathrm{Y})}$ [13].

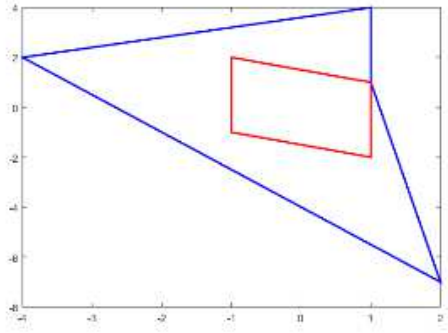
Suppose that there are two configurations X and Y in n-by-p. The goodness-of-fit of Procrustes between X and Y is obtained using Equation 2.

$$\mathrm{GoFP}(\mathrm{X},\mathrm{Y}) = \mathrm{GoFP}(\mathrm{Y},\mathrm{X}) = \left(\sum_{i=1}^{r}\sigma_{ii}\right)^2 \qquad (2)$$

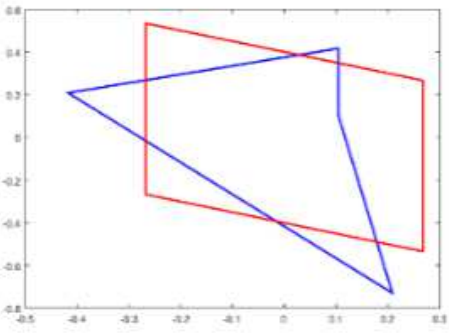
where r and σ_{ii} are rank and singular value of $\overline{\mathrm{X}}_T'\overline{\mathrm{Y}}_T$ or $\overline{\mathrm{Y}}_T'\overline{\mathrm{X}}_T$ with $\overline{\mathrm{X}}$ and $\overline{\mathrm{Y}}$ are matrices after normalization process by using formula $\overline{\mathrm{X}} = \frac{\mathrm{X}}{\|\mathrm{X}\|_F}$ and $\overline{\mathrm{Y}} = \frac{\mathrm{Y}}{\|\mathrm{Y}\|_F}$. The value of $\mathrm{GoFP}(\mathrm{X},\mathrm{Y})$ belongs to the interval [0,1]. If $\mathrm{GoFP}(\mathrm{X},\mathrm{Y}) \approx 1$ then the configurations have the excellent match. Conversely if that value approximate 0 then those have the poor match. Illustration of the GoFP is shown in Fig 1.



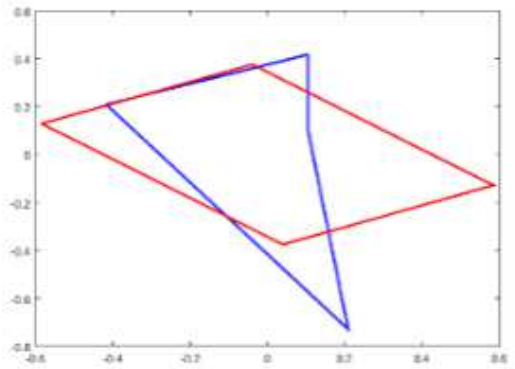
(a)



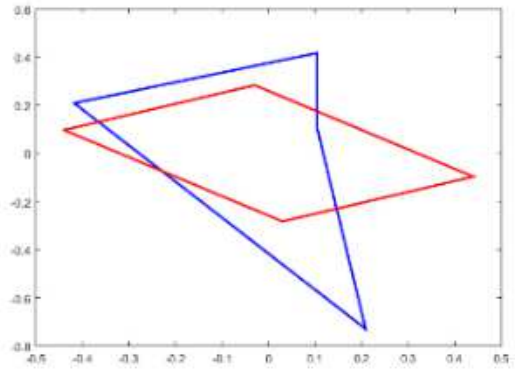
(b)



(c)



(d)



(e)

Fig. 1 Procrustes process include: (a) before transformation, (b) translation, (c) translation-normalization, (d) translation-normalization-rotation, and (e) translation-normalization-rotation-dilation.

B. Ordinary Procrustes Analysis

The procedure of ordinary Procrustes analysis (OPA) implicates the least squares matching of two configurations, suppose X and Y, using the similarity transformations. Parameter Q and c is exploited to minimize the OPA that is given by Equation 3.

$$E_{OPA}(\mathrm{Y},\mathrm{X}) = \|\mathrm{X}_T - c\mathrm{Y}_T\mathrm{Q}\|_F^2 \tag{3}$$

Where $\mathrm{Q} = \mathrm{VU}'$ from $\mathrm{Y}_T'\mathrm{X}_T = \|\mathrm{X}_T\|\|\mathrm{Y}_T\|\mathrm{USV}'$ by using the procedure of CFSVD, whereas $c = \frac{\mathrm{trace}(\mathrm{Y}_T'\mathrm{X}_T\,\mathrm{Q})}{\mathrm{trace}(\mathrm{X}_T'\mathrm{X}_T)}$ and X_T and Y_T are gained using the optimal translation procedure [12].

C. Generalized Procrustes Analysis

The generalized Procrustes analysis (GPA) of shape is the calculation of the average shape of all similar shapes to obtain one shape that can represent all those shapes. Suppose that $S = \{\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n\}$ is a set of configuration matrices of similar shapes. An algorithm to compute the GPA for S is as follows:

- Translation. Match the centroid of all configurations using Equation 4.

$$\overline{\mathrm{X}}_i = \mathrm{X}_i - \frac{1}{n}1_n 1_n'\mathrm{X}_i \tag{4}$$

- Initialize W using $\mathrm{W} = \frac{1}{n}\sum_{i=1}^{n}\overline{\mathrm{X}}_i$.
- Rotations and Dilation. For the i th configuration ($\forall i = 1,2,\ldots,n$), rotate and dilate the configuration using Equation 5.

$$\overline{\mathrm{X}}_i^* = c_i\overline{\mathrm{X}}_i\mathrm{Q}_i \tag{5}$$

where Q_i and c_i are defined as $\mathrm{Q}_i = \mathrm{V}_i\mathrm{U}_i'$ and $\mathrm{c}_i = \frac{\mathrm{trace}(\mathrm{W}'\overline{\mathrm{X}}_i\mathrm{Q}_i)}{\mathrm{trace}(\overline{\mathrm{X}}_i'\overline{\mathrm{X}}_i)}$. And U and V are given by Equation 6.

$$\mathrm{W}'\overline{\mathrm{X}}_i = \|\overline{\mathrm{X}}_i\|\|\mathrm{W}\|\mathrm{U}_i\mathrm{S}_i\mathrm{V}_i' \tag{6}$$

- Modify W using Equation 7.

$$\mathrm{W} = \frac{1}{n}\sum_{i=1}^{n}\overline{\mathrm{X}}_i^*. \tag{7}$$

- Repeat steps 3 and 4 until the Procrustes' sum of squares cannot be reduced further. The calculation uses Equation 8.

$$\sum_{i=1}^{n}\|\mathrm{W} - c_i\overline{\mathrm{X}}_i\mathrm{Q}_i\|_F^2, \tag{8}$$

W is the result of GPA[12].

D. Hierarchical Procrustes Clustering

The basic idea of the hierarchical Procrustes clustering is the capability of GoFP to measure the similarity of two configurations. Using that capability, we can collect those configurations with the highest similarity measure into one cluster. It shows that GoFP has the potential to carry out the shape clustering process on shape data. So, we intend to utilize GoFP optimally in the shape clustering procedure where it has not been addressed in previous works. The algorithm that we propose is the hierarchical Procrustes clustering. The procedure of that algorithm is described in the following steps:

- Suppose that $S = \{\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n\}$ is a set of n configurations in which these configurations are the initial cluster. We get n clusters.
- $\forall i \in \{1,2,\ldots,n-1\}$ and $\forall j \in \{i+1, i+2, \ldots, n\}$, the $\mathrm{GoFP}(\mathrm{X}_i, \mathrm{X}_j)$ is determined using Equation 8, where r and σ_{ii} are rank and singular value of $\overline{\mathrm{X}}_{iT}'\overline{\mathrm{X}}_{jT}$ or $\overline{\mathrm{X}}_{jT}'\overline{\mathrm{X}}_{iT}$ with $\overline{\mathrm{X}}_i$ and $\overline{\mathrm{Y}}_j$ are matrices after normalization process.
- Choosing the optimal GoFP by using Equation 9.

$$\max_{i,j}\mathrm{GoFP}(\mathrm{X}_i, \mathrm{X}_j) \tag{9}$$

- Combining X_i and X_j into one cluster. If X_i or X_j has entered a certain cluster with some other configurations, then those configurations are also included in that new cluster.
- Repeating procedures 3 and 4 until the desired number of clusters is formed.

E. K-means Procrustes Clustering

The idea of k-means Procrustes clustering appeared when we tried to utilize GPA and GoFP in the shape clustering process simultaneously. In our assumptions, the procedure of the GoFP can be utilized to gain the distance between the two configurations. At the same time, the GPA is possible to correct the centroid for each k-means iteration. Based on those

assumptions, we propose a shape clustering algorithm by using GoFP and GPA, which have not been addressed in previous works. The procedure of that algorithm is described in the following steps:

- Suppose that $S = \{X_1, X_2, \dots, X_n\}$ is a set of n configurations. Partitioning those configurations into k initial clusters arbitrarily.
- Computing centroid each cluster by using GPA.
- Assigning a configuration to the cluster whose centroid is nearest by using GoFP.
- Recalculate the centroid for the cluster receiving the new configuration and for the cluster losing the configuration.
- Repeating the second step until no more cluster member changes.

Based on the two algorithms previously proposed, we also try to implement the ordinary Procrustes analysis (OPA) algorithm for the shape clustering process. We will perform hierarchical clustering and k-means clustering on shape data using the previous procedure, where the GoFP algorithm is replaced with OPA.

F. Clustering validity

Cluster validity is a technique that provides a quantitative measure that can be utilized to evaluate certain clustering algorithms [22]. There are two types of cluster validity, namely internal and external cluster validity. This paper uses external cluster validity because there is information of the initial cluster in the shape data used, where four cluster validity measures are considered here. There are Rand index, Jaccard coefficient, F-measure, and Purity[23], [24]. A brief mathematical representation of these cluster validity used is given in the next paragraph.

Suppose that a dataset with n objects had been clustered using a certain algorithm where the result of the clustering process was m clusters collected in $P = \{P_1, P_2, \dots, P_m\}$. Think that $G = \{G_1, G_2, \dots, G_n\}$ was the set of the initial cluster structure of the data. Then, it is defined constants *a, b, c,* and *d* whose values are gained based on a comparison of the cluster of each pair of objects in G and P with the following conditions.

- a is the number of pairs of objects in the dataset that belongs to the same cluster in G, as well as in P.
- b is the number of pairs of objects in the dataset that belongs to the same cluster in G but different cluster in P.
- c is the number of pairs of objects in the dataset that belongs to the different cluster in G but the same cluster in P.
- d is the number of pairs of objects in the dataset that belongs to the different cluster in G, as well as in P.

Based on those values, the definition of the rand index is given in Equation 10.

$$\text{Rand Index} = \frac{a+d}{a+b+c+d}. \tag{10}$$

Rand index provides weight to those objects that were simultaneously clustered in the two clustering results[25]. The Jaccard coefficient is defined by Equation 11.

$$\text{Jaccard coefficient} = \frac{a}{a+b+c}. \tag{11}$$

Jaccard coefficient differs from Rand index by deleting d from both the numerator and denominator, placing the importance on a. However, because b or c must be increasing when d decreases, then d is implicitly in Equation 11[26]. As for F-measure and Purity, their calculations do not utilize the constants above. Suppose that n_{ij} is the number of objects that belong to cluster i in the initial cluster structure G and cluster j in the results of clustering process P. n_i is the number of objects that belong to cluster i in G and n_j is the number of objects that belong to cluster j in P. By using those definition, we can calculate the recall and precision of pairs of P and G clusters by using Equations 12 and 13.

$$\text{Recall}\ (i,j) = \frac{n_{ij}}{n_i}. \tag{12}$$

and

$$\text{Precision}\ (i,j) = \frac{n_{ij}}{n_j}. \tag{13}$$

Then, the F-measure of pairs of P and G clusters is given by Equation 14.

$$\text{F}(i,j) = \frac{2.\text{Recall}(i,j).\text{Precision}(i,j)}{\text{Precision}(i,j)+\text{recall}(i,j)}. \tag{14}$$

F-measure is got from the average of all $\text{F}(i,j)$ obtained. Similar to the F-measure, in the first step, a purity measure is got by calculating the purity value of each cluster in P by using Equation 15.

$$P_j = \frac{1}{n_j} \max_i n_{ij}. \tag{15}$$

Then, the overall Purity of the clustering is obtained as a weighted sum of each cluster purities and given as Equation 16.

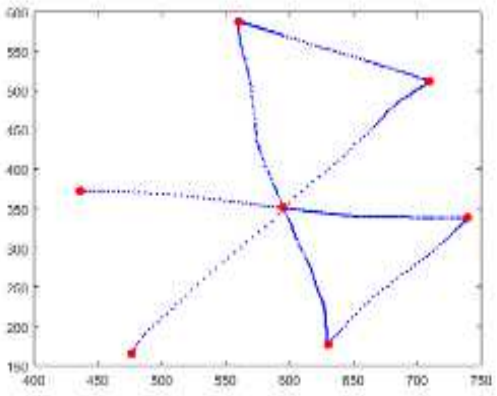
$$\text{Purity} = \sum_{j=1}^{m} \frac{n_j}{n} P_j. \tag{16}$$

Where n_j is the number of objects that belong to cluster j in P. m is the number of clusters, and n is the total number of objects in dataset used. For all cluster validities used in this paper, the larger values of each cluster validity indicate better clustering quality.

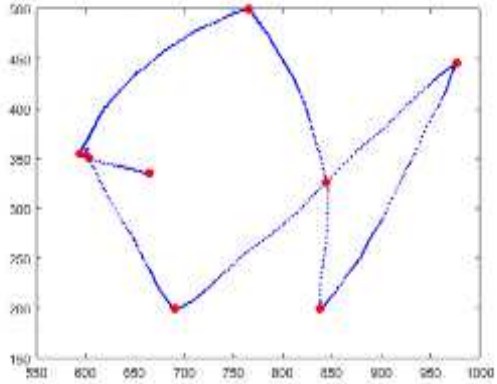
III. Result and Discussion

A. The Shape Dataset

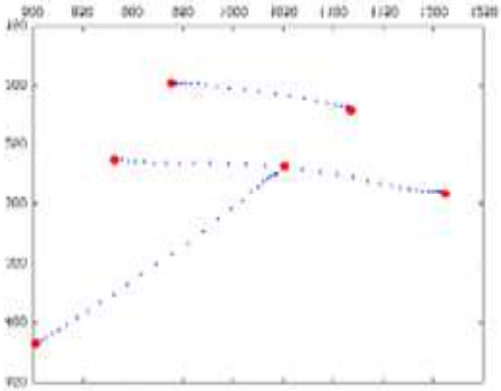
The shape dataset used in this research is a line drawing dataset that consists of 180 drawings classified into six clusters[21]. In full, the line drawing dataset contains Bluetooth, fish, Japanese postal mark, root in mathematics, alphabet T, and alphabet X shapes. Each of those shapes totals thirty. The visualization of the dataset is given in Fig 2. Those figures also provide information about a landmark that is used in this paper. That landmark is gained from the corner points of each data. The number of points on the Bluetooth, fish, Japanese postal mark, root in mathematics, alphabet T, and alphabet X landmark is 8, 8, 5, 4, 4, and 5. Because the number of landmark points for each shape is not the same, one landmark needs to be optimally matched to another by adding the centroid of the landmark. The additions are done before the GoFP, OPA, or GPA calculation and are not made permanent.



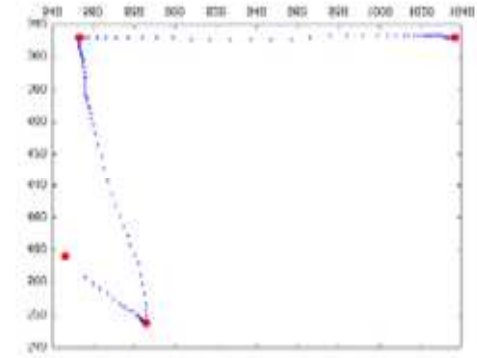
(a)



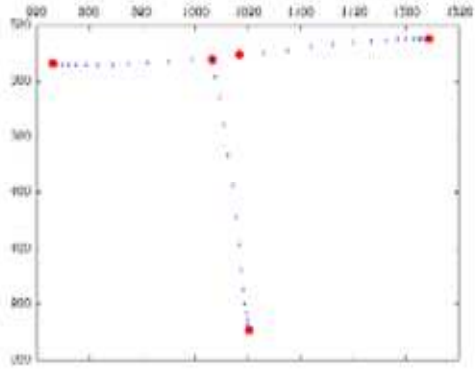
(b)



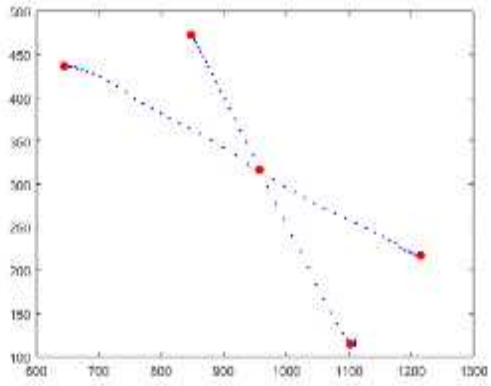
(c)



(d)



(e)



(f)

Fig. 2 Example of line drawings: (a) Bluetooth, (b) fish, (c) Japanese postal mark, (d) root in mathematics, (e) alphabet T, and (f) X.

B. The Clustering Results

In this paper, the shape clustering process is carried out 200 times for each algorithm. It is done to see the convergence of the cluster quality and time complexity by using Equation 17.

$$a = \lim_{n \to \infty} a_n \tag{17}$$

Where a_n is the particular value in nth iteration, and a is the convergence of the particular value[27]. The graph of the cluster quality results for each algorithm using the convergence concept is shown in Figure 3-6.

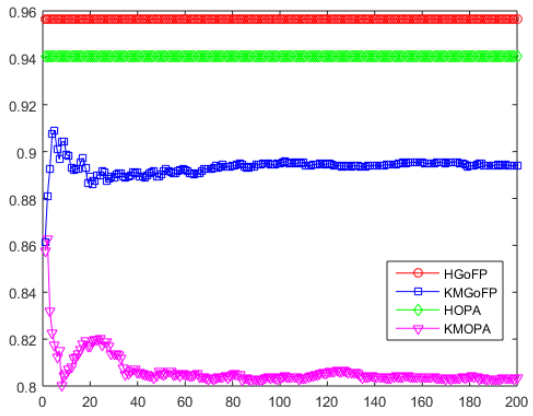


Fig. 3 The convergence of the cluster qualities by using Rand index.

Fig 3 shows the graph of cluster qualities from hierarchical goodness-of-fit of Procrustes (HGoFP), k-means goodness-of-fit of Procrustes (KMGoFP), hierarchical ordinary Procrustes analysis (HOPA), and k-means ordinary Procrustes analysis (KMOPA) by using Rand index as cluster validity. In Rand index, the average values for HGoFP, KMGoFP, HOPA, and KMOPA are 0.956, 0.894, 0.941, and 0.803, respectively. Based on the values, we know that the shape clustering results are satisfactory based on Rand index. We also know that the best results are obtained by the HGoFP algorithm. The second-best algorithm is HOPA, and then KMGoFP and KMOPA are the third and the last, respectively.

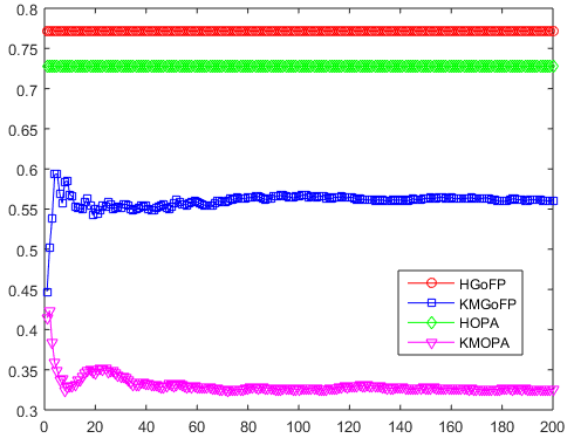


Fig. 4 The convergence of the cluster quality by using the Jaccard index.

Fig 4 shows the cluster qualities based on the Jaccard index. As can be seen, the convergence of the cluster quality from KMOPA is not satisfactory because its value is about 0.300, which is 0.326 precisely. In comparison, other algorithms are good enough because the convergence values are above 0.500. The best algorithm in the Jaccard coefficient is obtained by using HGoFP algorithm, whose cluster quality is 0.772. The second and the third best algorithms are HOPA and KMGoFP, whose values are 0.728 and 0.561, respectively.

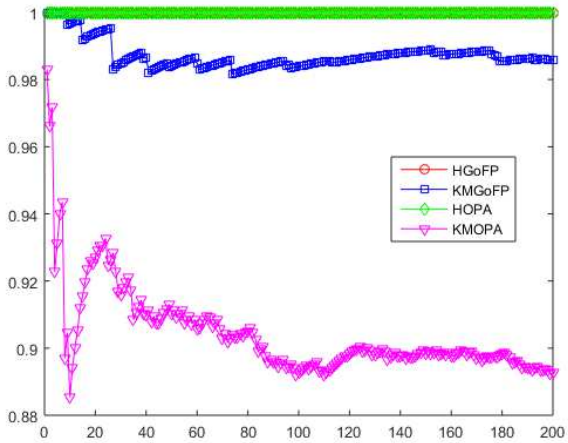


Fig. 5 The convergence of the cluster quality by using the F-measure index.

Fig 5 shows the graph of the F-measure value of each algorithm. The graph shows that each algorithm used gives a satisfactory result where their values are above 0.880. The best algorithms in F-measure are HGoFP and HOPA, which the value is same at about 1. The second-best algorithm is KMGoFP, whose value is 0.986, and then the third-best algorithm is KMOPA, whose value is 0.893.

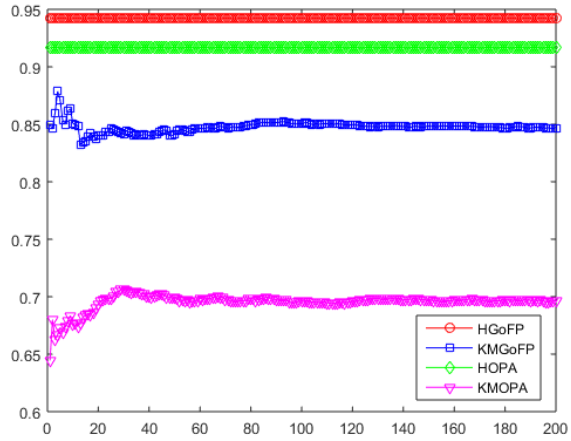


Fig. 6 The convergence of the cluster quality by using a purity index.

The graph in Fig 6 shows that the shape clustering results of each algorithm in Purity validity are good enough because of their Purity value of more than 0.500. The best algorithms in Purity validity are HGoFP, whose value is 0.942. The second, third, and fourth are HOPA, KMGoFP, and KMOPA, whose values are 0.917, 0.847, and 0.697.

Based on the description above, we can conclude that HGoFP is the best algorithm in each cluster validity used. The second-best algorithm is HOPA, then KMGoFP and KMOPA, respectively. The values of HGoFP and HOPA are quite similar in Rand Index, Jaccard coefficient, and Purity. In F-measure, their values are the same. So, the quality of the HGoFP and HOPA is not much different. While the lowest quality is gained by KMOPA generally.

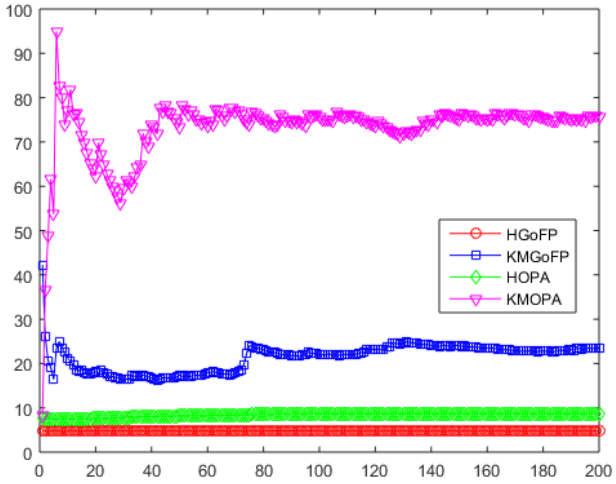


Fig. 7 Graph's visualization of time complexity each algorithm in 200 times iteration

Fig 7 shows the chart of time complexity of each algorithm in 200 times iteration. One of the principal pieces of information from the chart is that the lowest time complexity is achieved by HGoFP whose average of time complexity is 4.733. It means that the HGoFP finishes the shape clustering process faster than other algorithms. The second-lowest time complexity is achieved by HOPA, which the average of time complexity is 8.679. The third and fourth are KMGoFP and

KMOPA. We know that hierarchical shape clustering algorithms are more rapidly than k-means shape clustering based on the results. It can be clarified by the results of HGoFP and KMGoFP or HOPA and KMOPA. We also know that the shape clustering using the GoFP is more rapidly than OPA. It can be clarified by the results of HGoFP and HOPA or KMGoFP and KMOPA.

IV. Conclusion

In the paper, we have discussed shape clustering by using Procrustes analysis. The Procrustes algorithms used in this paper were GoFP, GPA, and OPA. The shape clustering algorithms proposed in this research were HGoFP, KMGoFP, HOPA, and KMOPA. And then, the cluster validities used to evaluate the cluster results were Rand index, Jaccard coefficient, F-measure, and Purity. The clustering process of each algorithm was repeated 200 times to obtain the convergence of each algorithm's clustering quality. This research found that the results of all algorithms used are good enough in Rand index, F-measure, and Purity validities. In Jaccard coefficient, the good clustering results were only from HGoFP, HOPA, and HOPA, whereas the KMOPA algorithm got the low cluster quality. In the time complexity, the HGoFP process is the fastest. Based on the cluster validity used and the time complexity, the algorithms proposed in this paper particularly deserve to be proposed as a new algorithm to cluster the objects in the line drawing dataset. Then, the HGoFP is suggested clustering the objects in the dataset used.

Acknowledgment

The Directorate of Higher Education, the Ministry of Education and Culture of the Republic of Indonesia, funded this research through a 2020 external grant. The authors also are grateful for IT Telkom Purwokerto support.

References

[1] M. Pavithra and R. M. S. Parvathi, "A survey on clustering high dimensional data techniques," *Int. J. Appl. Eng. Res.*, vol. 12, no. 11, pp. 2893–2899, 2017.

[2] R. Ananda, "Analisis Mutu Pendidikan Sekolah Menengah Atas Program Ilmu Alam di Jawa Tengah dengan Algoritme K-Means Terorganisir," *J. Informatics, Inf. Syst. Softw. Eng. Appl.*, vol. 2, no. 1, pp. 65–72, 2019, doi: 10.20895/inista.v2i1.97.

[3] R. Ananda, "Silhouette Density Canopy K-Means for Mapping the Quality of Education Based on the Results of the 2019 National Exam in Banyumas Regency," *Khazanah Inform. J. Ilmu Komput. dan Inform.*, vol. 5, no. 2, pp. 158–168, 2019, doi: 10.23917/khif.v5i2.8375.

[4] R. Ananda, M. Z. Naf'an, A. Beladina, and A. Burhanudin, "Sistem Rekomendasi Pemilihan Peminatan Menggunakan Density Canopy K-Means," vol. 1, no. 1, pp. 19–25, 2017.

[5] R. Adhitama, A. Burhanuddin, and R. Ananda, "Penentuan jumlah cluster ideal SMK di Jawa Tengah dengan Metode X-means clustering dan K-means clustering," *J. Inform. dan Komput.*, vol. 3, no. 1, pp. 1–5, 2020, doi: 10.33387/jiko.

[6] R. Ananda and A. Z. Yamani, "JURNAL RESTI Penentuan Centroid Awal K-means pada proses Clustering Data Evaluasi," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 1, no. 10, pp. 544–550, 2021.

[7] P. Govender and V. Sivakumar, *Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019)*, vol. 11, no. 1. Turkish National Committee for Air Pollution Research and Control, 2020.

[8] J. Lukáč, B. Mihalčová, E. Manová, R. Kozel, Š. Vilamova, and K. Čulková, "The position of the Visegrád countries by clustering methods based on indicator environmental performance index," *Ekol. Bratislava*, vol. 39, no. 1, pp. 16–26, 2020, doi: 10.2478/eko-2020-0002.

[9] U. R. Yelipe, S. Porika, and M. Golla, "An efficient approach for imputation and classification of medical data values using class-based clustering of medical records," *Comput. Electr. Eng.*, vol. 66, pp. 487–504, 2018, doi: 10.1016/j.compeleceng.2017.11.030.

[10] A. A. H. Hassan, W. M. Shah, A. M. Husien, M. S. Talib, A. A. J. Mohammed, and M. F. Iskandar, "Clustering approach in wireless sensor networks based on K-means: Limitations and recommendations," *Int. J. Recent Technol. Eng.*, vol. 7, no. 6, pp. 119–126, 2019.

[11] D. S. Wardiani and N. Merlina, "Implementasi Data Mining Untuk Mengetahui Manfaat Rptra Menggunakan Metode K-Means Clustering," *J. Pilar Nusa Mandiri*, vol. 15, no. 1, pp. 125–132, 2019, doi: 10.33480/pilar.v15i1.403.

[12] I. L. Dryden and K. V. Mardia, *Statistical Shape Analysis with applications in R*, 2nd ed. 2016.

[13] T. Bakhtiar and Siswadi, "Orthogonal procrustes analysis: Its transformation arrangement and minimal distance," *Int. J. Appl. Math. Stat.*, vol. 20, no. M11, pp. 16–24, 2011.

[14] T. Bakhtiar and Siswadi, "On The Symmetrical Property of Procrustes Measure of Distance," vol. 99, no. 3, pp. 315–324, 2015.

[15] Siswadi and T. Bakhtiar, "Goodness-of-fit of biplots via procrustes analysis," *Far East J. Math. Sci.*, vol. 52, no. 2, pp. 191–201, 2011.

[16] Siswadi, T. Bakhtiar, and R. Maharsi, "Procrustes analysis and the goodness-of-fit of biplots: Some thoughts and findings," *Appl. Math. Sci.*, vol. 6, no. 69–72, pp. 3579–3590, 2012.

[17] A. Muslim and T. Bakhtiar, "Variable selection using principal component and procrustes analyses and its application in educational data," *J. Asian Sci. Res.*, vol. 2, no. 12, pp. 856–865, 2012, [Online]. Available: http://www.aessweb.com/pdf-files/856-865.pdf.

[18] R. Ananda, Siswadi, and T. Bakhtiar, "Goodness-of-Fit of the Imputation Data in Biplot Analysis," *Far East J. Math. Sci.*, vol. 103, no. 11, pp. 1839–1849, 2018, doi: 10.17654/ms103111839.

[19] R. Ananda, A. R. Dewi, and N. Nurlaili, "a Comparison of Clustering By Imputation and Special Clustering Algorithms on the Real Incomplete Data," *J. Ilmu Komput. dan Inf.*, vol. 13, no. 2, pp. 65–75, 2020, doi: 10.21609/jiki.v13i2.818.

[20] F. Novika and T. Bakhtiar, "The Use of Biplot Analysis and Euclidean Distance with Procrustes Measure for Outliers Detection," *Int. J. Eng. Manag. Res. Page Number*, no. 1, pp. 194–200, 2018, [Online]. Available: www.ijemr.net.

[21] K. Iwata, *Shape clustering as a type of procrustes analysis*, vol. 11304 LNCS. Springer International Publishing, 2018.

[22] G. Gan and E. A. Valdez, "Data Clustering with Actuarial Applications," *North Am. Actuar. J.*, vol. 24, no. 2, pp. 168–186, 2020, doi: 10.1080/10920277.2019.1575242.

[23] E. Rendón *et al.*, "A comparison of internal and external cluster validation indexes," *Appl. Math. Comput. Eng. - Am. Conf. Appl. Math. Am. 5th WSEAS Int. Conf. Comput. Eng. Appl. CEA'11*, pp. 158–163, 2011.

[24] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On validation techniques," *J. Intell. Inf. Syst.*, vol. 17, no. 2, pp. 107–145, 2001.

[25] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *J. Am. Stat. Assoc.*, vol. 66, no. 336, pp. 846–850, 1971, doi: 10.1080/01621459.1971.10482356.

[26] D. Steinley, "Properties of the Hubert-Arabie adjusted Rand index," *Psychol. Methods*, vol. 9, no. 3, pp. 386–396, 2004, doi: 10.1037/1082-989X.9.3.386.

[27] M. Taboga, *Lectures on Probability Theory and Mathematical Statistics*, 3rd ed. CreateSpace Independent Publishing Platform, 2017.