Advances in Engineering Research, volume 208
Proceedings of the International Conference on Innovation in Science and Technology (ICIST 2020)

Clustering Balinese Script Image in Palm Leaf Using Hierarchical K-Means Algorithm

Anastasia Rita Widiarti*, C. Kuntoro Adi
Informatics Department, Faculty of Science and Technology
Sanata Dharma University
Sleman, Indonesia
*rita_widiarti@usd.ac.id, kuntoroadi@usd.ac.id

***Abstract*—This paper proposes a combination approach of clustering, a hierarchical clustering to group similar characters of Balinese Lontar script; followed by k-means clustering as a way to identify the group to find-out the right label for members of the group. Based on the optimal value of the silhouette coefficient, the hierarchical clustering method results in 113 groups of Balinese Lontar scripts. Using these 113 groups to compute the initial centroid of each cluster, k-means clustering process is able to correctly group and label 111 scripts out of 198 Balinese Lontar scripts sample.**

Keywords—clustering and hierarchical clustering approach, Balinese Lontar script, hierarchical K-Means algorithm

I. INTRODUCTION

This paper is a continuation of a preliminary research done by Rita Widiarti in 2019, namely, an "Automatic Segmentation of Balinese script written in Lontar palm leaf". The segmentation of seven Lontar image writing taken from Pustaka Artati Sanata Dharma Library produced 1134 segmented single script of Balinese script with no label, and resulted in 63.76% accuracy.

The question is then, what label is assigned for each character and how? It is interesting to observe that labeling those amounts of segmented characters would be very time consuming, and some experts to verify the labeling results are needed. It would be beneficial if there is a way of automatically cluster the similar character, and then identify the members of the same cluster with same label.

There are methods of clustering such as mini batch k-means, affinity propagation, mean shift, spectral clustering, ward, agglomeration clustering, DB-scan, birch, and Gaussian mixture. The differences of the methods mainly on the objective of the grouping and the types of the data samples. Meanwhile, the performance of clustering algorithms depends mostly on the size of the group, the size of the input data, the randomness of the data, and the existence of noises among the data [1].

K-means clustering approach is a popular, simple and easy to implement. The method requires information of how many clusters beforehand. The success of grouping depends on the ability to guess the number of cluster and the initialization to assign centroid of each cluster. Many researches, therefore, develop methods to identify number of k in k-means clustering [2,3] and way to initialize centroids [4].

The performance of k-means clustering is comparable. Prakosa and Sari [5] examined three approaches (namely, OTSU, FCM and k-means clustering) of segmenting objects for vehicle detection systems. The mean square errors (MSE) values show that k-means offer best segmentation results. Awangga et al. [6] clustered coal data using a combination of k-means and mean-shift methods. K-means is also applied to cluster voice signals [7], however, this approach is less optimal than Gaussian Mixture Models (GMM) with PCA method.

The agglomerative hierarchical clustering (AHC) is another type of clustering of objects based on their similarity. It's a bottom-up approach. Each data observation starts in its own cluster, and similar clusters are merged as to move up the hierarchy. Pandey and Khanna showed that hierarchical clustering is more efficient and effective compared to k-means [8]. Tiayudi and Fitri [9] employed single linkage dissimilarity with global cumulative score standard (SLG) algorithm to improve agglomerative hierarchical clustering approach to analyze student activities in online learning. They successfully expanded average linkage dissimilarity with global cumulative global score standard (ALG) to cluster iris, wine and Wisconsin breast cancer datasets [10].

The success of k-means and hierarchical clustering had inspired researchers to combine both approaches to work some clustering tasks. Arai and Barakbah [11], for example, used k-means hierarchical algorithm to cluster multi-band images. They employed hierarchical clustering as the initialization process before segmenting datasets with k-means clustering. Murthy et al. [12] employed k-means and hierarchical clustering to better search images on web-sites. Yuhefizar et al. [13] used agglomerative hierarchical clustering to determine the initial number of clusters in web-visitor data, and then segmented visitor data using k-means clustering.

This research, therefore, proposes a combination of clustering – that is: a hierarchical clustering to group similar characters of those Balinese script, followed by k-means clustering as a way to identify the group to find-out the right label for members of the group. It is an effort to answer these

following questions: a. how to find out and choose the syllable unit in Balinese Lontar script? b. what kind of preprocessing needed for better clustering? c. what kind of transformation chosen to discover good features (there are variation in size, thickness; and some characters are tilted); d. how many clusters of are in the characters and how good is the clustering results.

The objectives of this paper are, then, a. to find out how many clusters are in the segmented Balinese script. The goodness of clustering results is measured through the silhouette coefficient values; b. to find out the label of each cluster and c. to look for the best feature for clustering.

The research would offer some contributions as follows: a. this clustering and labeling systems would lead to the automatic transliteration of the Balinese palm leaf script or Lontar script to the Roman script, b. this automatic transliteration is important in helping contemporary people to understand texts behind the script; and therefore c. the reservation and dissemination of Balinese Lontar script become possible due to more people have easy access to the texts.

After explaining the method employed in this research, this paper presents its results and discussion, followed by some note as a conclusion.

II. RESEARCH METHODS

Figure 1 presents three processes in the method, namely: data preprocessing, feature extraction and Lontar script clustering.

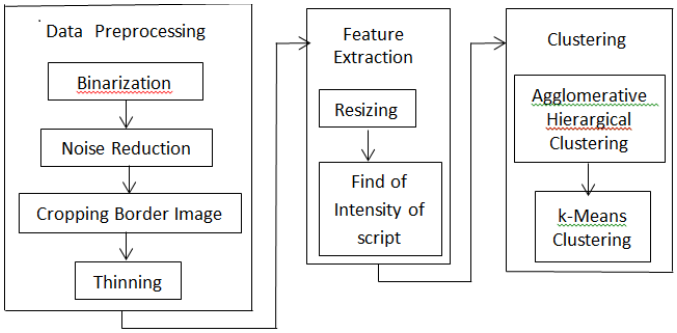


Fig. 1. Proposed of design process.

A. Data Preprocessing

Data observations consist of 1134 characters of Balinese Lontar scripts, segmented from seven Lontar writing. Figure 2 shows one sample of Lontar writing.

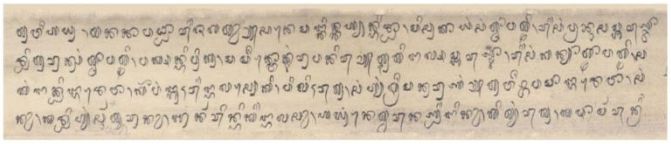


Fig. 2. Lontar script sample.

It may be observed in Figure 3, that some segmentation results may not perfect. Some syllables still contain more than one unit character. Some symbols need to be separated from the main syllable.

Fig. 3. Segmented character of Lontar script.

Theoretically Balinese scripts contains 73 unique syllables, however additional punctuation mark may create more than 73 unique syllables. It is, therefore, before pre-processing step this research separates manually some unit Balinese scripts to give 2079 unique characters for the input of the system.

Data pre-processing involves four steps of converting input image before feature extraction process. In binarization step, data image is modified to black-white image followed by noise reduction step to clean data from salt and pepper noises. Border cropping step isolates the script data from their background. Thinning process then extracts image skeleton of each script.

B. Feature Extraction

Feature extraction is a process of dimensionality reduction of data to smaller size. Feature extraction selects or combines variables into features, and reduce the amount of data to process. At the same time feature extraction process is still accurately maintain the characteristics of the original data set.

This research employed two features of Balinese palm leaf script data, namely the modification of image size and computation of the intensity of character (IoC) for each image data. The first process of feature extraction was, therefore, to resize data images to different sizes of 30x30, 40x40, 50x50, 60x60,70x70, 80x80 pixels. That sizes are chosen as the preferable sizes based on segmentation experiments by Rita Widiarti in 2019. Furthermore, each data image is divided into 3x3, 4x4, 5x5, 6x6, 7x7, and 8x8 areas. This research is then computed the intensity of each area using following equation (1).

$$Feature_{i,j} = \sum_{v=10i+1}^{10(i+1)} \sum_{k=10j+1}^{10(j+1)} M_{y,k} \quad i,j = 0,1,(n-1) \tag{1}$$

with M = size of n x n pixel image.

C. Clustering Process

The clustering process is as follows: for a data set of images $M = \{ m_1, m_2, \ldots, m_N\}$ with N number of features in each image, the clustering process groups M into K clusters $\{ C_1, C_2,\ldots,C_K\}$ with the following conditions [14].

- Each data sample is assigned to a cluster, i.e.

$$\bigcup_{g=1}^{g=K} C_g = M$$

- Each cluster has at least one data sample assigned to it, i.e.:

$$C_g \neq \emptyset, g = 1, \dots, K$$

- Each data sample is assigned to one and only one cluster, i.e.

$$C_g \cap C_h = \emptyset \; where \; g \neq h$$

1) Agglomerative hierarchical clustering: Agglomerative hierarchical clustering is a bottom-up approach. It starts with all data images represents their own cluster. The similar clusters then are merged. The iteration processes continue until all data objects are group in one cluster. There are different ways of measuring cluster similarity, namely, single-link, complete-link and average-link. The linkage criterion determines the distance between sets of data sample as function of the pairwise distances between data samples. This research used single-linkage due to its ability to handle various shape and density [15].

This paper employs silhouette computation as a method to interpret and validate the clustering results, as well as finding out how many clusters are in the data. The silhouette value is a measure of how similar an object is to its own cluster compared to other clusters. The silhouette ranges from -1 to +1, with high value indicates that the data sample is well matched to its own cluster and poorly matched to neighbouring clusters.

The number of cluster (K) is chosen from the best value of silhouette coefficient. The membership of data clusters in each cluster are used to compute the initial centroid of each cluster in the following k-means clustering process.

2) K-mean clustering: K-means clustering is a method to partition data observations into k-clusters. In this case, each observation belongs to a cluster with the nearest mean (cluster centers or cluster centroid). The standard k-means algorithm is as follows:

Given the initial sets of k-centroids c_1, c_2… c_k, the algorithm proceeds by alternating between two steps: assignment step, and update step.

- Assignment step: assign each observation to the cluster with the nearest distance to the cluster centroid.
- Update step: recalculate centroids or means for observation assign to each cluster.

The k-means clustering method has converged when the assignments no longer change the membership of the clusters.

III. RESULTS AND DISCUSSION

A. Preprocessing

As mentioned in the discussion above, the preprocessing steps include binarization, noise reduction, border cropping and thinning processes. Border cropping process of the data Balinese scripts shows that on average they have 47.01 pixel height and 41.86 pixel width as Table 1 shows. Table 1 gives information that the data observation has height range of 21 to 90 pixels and width range of 20 to 75 pixels.

TABLE I. PROPERTY OF DATA SIZE SCRIPTS INPUT

	Average	Maximum	Minimum
Height	47,01	90	21
Weight	41,86	75	20

B. Agglomerative Clustering with Single-linkage Distance

The size of feature inputs are varies following the variation of number of areas chosen from the image data. Figure 4 displays the silhouette coefficient values of different number of clusters from the various features combination. Figure 4 shows that IoC features for script with 3x3 areas results in the highest SC values (as seen in more detail in Table 2).

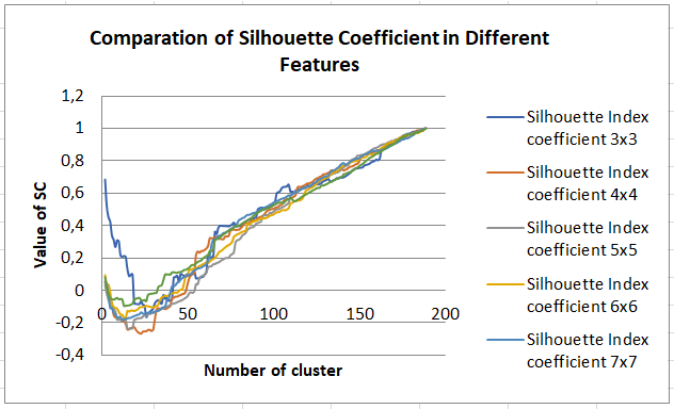


Fig. 4. Silhouette Coefficient (SC) Values of various clustering results.

TABLE II. SILHOUETTE COEFFICIENT (SC) AVERAGE VALUES FOR VARIOUS FEATURE OF WINDOW SIZE

Window size	Average	Standard deviation
3x3	0,473530799	0,332737
4x4	0,424014217	0,399298
5x5	0,401954779	0,405028
6x6	0,424181848	0,37213
7x7	0,440598387	0,381301
8x8	0,450397147	0,336682

As one knows, the silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a SC high value, then the clustering configuration is appropriate.

The question is then: how many clusters are in the data? Figure 5 shows that in the position of cluster number 113 the graph starts to show reasonable SC values. This research does not choose clusters with SC = 1 (when each observation or data sample as a cluster) because it does not give any clue or information of how many observation are in the same cluster.

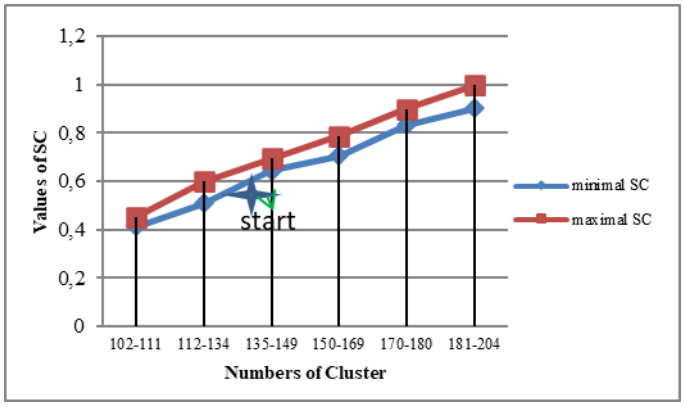
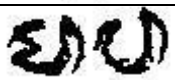
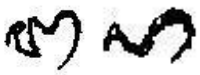
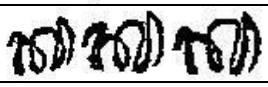
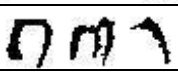


Fig. 5. SC values to discover number of clusters.

C. *K-means Clustering*

This research choose k=113 (the clustering results of hierarchical agglomerative clustering) as the initial number of clusters. The centroid value of each cluster is computed from the average value of features of data samples in each cluster. Table 3 presents some clustering results.

TABLE III. SOME SAMPLES OF K-MEANS CLUSTERING RESULT

Cluster	Number of script	Script image	Number of correct script	Notes
1	1		1	One script – correct grouping
2	1		1	One new script – correct grouping
3	2		2	Two similar scripts – are grouped in the same cluster
4	2		2	Two new scripts – correct grouping
5	2		2	Two new scripts – correct grouping
6	8		4	Two variation of “ma” and “wa”
7	1		1	One new scripts – correct grouping
8	1		1	One new scripts – correct grouping
9	2		1	Two new scripts – one correct grouping
10	3		3	Three new scripts – one correct grouping
11	7		5	Seven new scripts - one correct grouping – 2 scripts false

From the 198 data samples, 111 scripts are correctly clustered. That gives 56.65% of cluster accuracy. It seems that the clustering (and labeling) result is not optimal. However this may content important information to look for better feature to choose and methods to use.

IV. CONCLUSION AND FUTURE WORK

This research proposes a combination approach of clustering, a hierarchical clustering to group similar characters of those Balinese script followed by k-means clustering as a way to identify the group to find-out the right label for members of the group. Based on the optimal value of the silhouette coefficient, the hierarchical clustering method results in 113 groups of Balinese Lontar scripts. Using these 113 groups to compute the initial centroids, k-means clustering process is able to correctly group and label 111 scripts out of 198 Balinese Lontar data sample. As a future work, this research is going to find-out different features, and using

different similarity metrics in agglomerative clustering (single-link, complete-link and average-link) to get optimal results.

ACKNOWLEDGMENTS

This research supporting by a research grant from Sanata Dharma University. We would like to thank David Thanlian Kurniawan and Valentinus Angga Ankrisnar for helpful coding.

REFERENCES

[1] O.M. Abbas, "Comparisons Between Data Clustering Algorithms," IAJIT International Arab Journal of Information Technology, vol. 5, no. 3, pp. 320-325, 2008.

[2] O.J. Oyelade, O.O. Oladipupo, and I.C. Obagbuwa, "Application of k-Means Clustering algorithm for prediction of Students Academic Performance," IJCSIS International Journal of Computer Science and Information Security, vol. 7, no. 1, pp. 292-295, 2010.

[3] T.M. Kodinariya and R.R. Makwana, "Review on determining number of cluster in K-Means Clustering," International Journal of Advance Research in Computer Science and Management Studies, vol. I, pp. 90-95, 2013.

[4] P. Bholowalia and A. Kumar, "EBK-Means: A Clustering Techiniques based on Elbow Method and K-Means," WSN International Journal of Computer Application vol. IX, no. 105, pp. 17-24, 2014.

[5] P.B. Prakoso and Y. Sari, "Vehicle detection using background subtraction and clustering algorithms," Telkomnika vol. 17, pp. 1393-1398, 2019.

[6] R.M. Awangga, S.F. Pane, K. Tunnisa and I.S. Suwardi, "K Means Clustering and Meanshift Analysis for Grouping the Data of Coal Term in Puslitbang tekMIRA," Telkomnika vol. 16, no. 3, pp. 1351-1357, 2018.

[7] S.F. Ahmad and D.K. Singh, "Automatic detection of tree cutting in forests using acoustic properties," Journal of King Saud University - Computer and Information Sciences, 2019.

[8] S. Pandey and P. Khanna, "A hierarchical clustering approach for image datasets," Proc. 9th Int. Con. on Industrial and Information Systems (ICIIS), (Gwalior), pp. 1-6, 2014.

[9] A. Triayudi and I. Fitri, "A new agglomerative hierarchical clustering to model student activity in online learning," Telkomnika vol. 17(3), pp. 1226-1235, 2019.

[10] A. Triayudi and I. Fitri, "Comparison Of Parameter-Free Agglomerative Hierarchical Clustering Methods," ICIC Express Letters ICIC International vol. 12, no. 10, pp. 973–980, 2018.

[11] K. Arai and A. Barakbah, "Hierarchical K-means: An algorithm for centroids initialization for K-means," Reports of the Faculty of Science and Engineering Saga Univ. Saga University vol. 36, no, 1, pp. 25-31, 2007

[12] V.S. Murthy, E. Vamsidhar, P.S. Rao, and G.S.V. Raju, "Application Of Hierarchical And K-Means Techniques In Content Based Image Retrieval," International Journal of Engineering Science and Technology, vol. 2, no. 5, pp. 749-755, 2010.

[13] Y. Yuhefizar, B. Santosa, I.K. Eddy P, and Y.K. Suprapto, "Combination of Cluster Method for Segmentation of Web Visitors," Telkomnika, vol. 11, no. 1, pp. 207-214, 2013

[14] L. Kaufman and P. Rousseeuw, Finding Groups in data: An Introduction to Cluster Analysis, New York: J. Wiley & Son, 1990.

[15] F. Rosa and S. Guillaume, "A hierarchical clustering algorithm and an improvement of the single linkage criterion to deal with noise," Expert Systems with Applications, vol. 128, pp. 96-108, 2019.