Clustering OKU Timur Script Images using VGG Feature extraction and K-Means

Liu Toriko[1], Susan Dian Purnamasari[2]*, Yesi Novaria Kunang[3], Ilman Zuhri Yadi[4], Andri[5]
Intelligent Systems Research Group, Faculty of Science Technology [1], [2], [3], [4], [5]
Universitas Bina Darma
Palembang, Sumatera Selatan, Indonesia
liutoriko031102@gmail.com[1], susandian@binadarma.ac.id[2], yesinovariakunang@binadarma.ac.id[3],
ilmanzuhriyadi@binadarma.ac.id[4], andri@binadarma.ac.id[5]

***Abstract*— This study focuses on the utilization of clustering models to group manuscript images from the OKU Timur region based on specific characteristics. OKU Timur is rich in cultural heritage, including a unique writing system known as the OKU Timur script. The development of intelligent systems technology can be employed to recognize the OKU Timur script. For this purpose, a dataset of OKU Timur script is needed, which will later be used for classifying script images. One of the challenges in preparing the dataset is grouping a large number of script image samples according to the number of characters. A proposed solution in this research is to automatically group script images by applying the K-Means algorithm. The dataset comprises 2,280 images, representing 19 characters and 228 variations with different diacritics. Features are extracted using the VGG16 model, which are then clustered with the K-Means algorithm. Clustering performance is evaluated based on the percentage of correctly grouped characters. For 19 groups (character count), the model achieves an accuracy of 82.6%. For 228 groups (variations and diacritics), it correctly groups 48.16% of characters. Despite the challenges, the results demonstrate the model's potential for further refinement. This study's contribution lies in introducing an efficient clustering approach for cultural manuscripts, supporting digital preservation, and advancing automatic recognition of the OKU Timur script. These efforts aim to preserve the script for future generations.**

Keywords*— *OKU Timur Script, Clustering, K-Means, VGG16 Model, Manuscript Images

I. INTRODUCTION

OKU Timur Regency is located in South Sumatra Province and borders Ogan Ilir Regency, Ogan Komering Ulu Regency, South Ogan Komering Ulu Regency, and Lampung Province. This area is rich in historical sites, one of which is Aksara, which refers to letters or symbols that function as symbols of sound (phonemes) [1]. Aksara is also known as a "writing system." Over time, aksara has developed into a visual symbol system that appears on various media, such as paper, stone, trees, wood, or cloth, to convey elements of expression of a language [2]. Currently, the development of information technology is rapid and covers many aspects of life. This progress has led to the availability of extensive and diverse data, covering industries, the economy, science, technology, and various other fields [3]. The OKU Timur Script is an integral part of a unique cultural heritage, reflecting the identity and history of the people in the area.

Like many other ancient writing systems, the OKU Timur script holds significant cultural and historical value [4], [5]. The digital preservation of OKU Timur script is very important for several reasons. First, digitization allows wider access for future generations to learn and understand these characters, which may no longer be actively used in everyday communication [6], [7]. Second, by preserving the script in digital format, we can protect knowledge and culture that may be lost over time, especially as the number of speakers and active users of the script decreases [8].

In an effort to preserve the OKU Timur script, it is necessary to create intelligent applications that can automatically recognize the script, similar to those that researchers have previously developed for other types of scripts. [9], [10]. The initial effort to develop intelligent applications requires a dataset to train the model to recognize variants of the script. A specific challenge in preparing a large-scale dataset is the labor-intensive and error-prone process of manually sorting thousands of example images [11], [12]. To address this, the dataset preparation can be assisted by utilizing clustering algorithms that can automatically group character images into sets based on similarities [13]. In this context, clustering—through techniques such as the K-Means algorithm—plays an important role in the preservation process [14], [15]. Clustering provides a systematic way to organize and group images of OKU Timur script characters. By grouping characters based on visual similarities, we can identify and categorize characters more efficiently.

Clustering is the process of grouping data into various clusters where similar objects are placed in one cluster, while dissimilar objects are placed in different clusters [16]. Each cluster contains data that is as similar as possible to each other, and the degree of similarity is usually measured based on distance. Therefore, each object in one cluster must have similar characteristics, while objects in other clusters must have different characteristics. The goal of this clustering process is to organize the data into several groups, so that similar data is placed in one group while different data is placed in another group [17]. The basis of the clustering concept is to group a number of objects into clusters, with the aim of forming groups that have high similarity among the objects within them and

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i1.2292, Copyright ©2025
Submitted : October 7, 2024, Revised : November 4, 2024, Accepted : November 25, 2024, Published : January 31, 2025

significant differences from the objects in other groups [18].

The effectiveness of the K-Means algorithm in clustering characters, as demonstrated in several previous studies, has inspired researchers to adopt this approach. For instance, the K-Means algorithm has been used to cluster Javanese script images [15], K-Means and spectral clustering for grouping Odia character images [19], clustering handwritten Indic scripts [14], and hierarchical K-Means for clustering Balinese script images [20]. Moreover, the K-Means algorithm is highly popular due to its ability to quickly cluster large amounts of data, including outliers [21], [22]. It also has relatively low computational complexity, making it efficient for application to large datasets [23], [24]. In this study, the researchers use the K-Means algorithm with a feature extraction process utilizing the VGG model.

The novelty of this research lies in the fact that there has been no prior study specifically focusing on the development of a clustering model for OKU Timur script images. The absence of prior research presents a unique challenge in finding suitable information for the development of this model. The development of this clustering model is necessary to make the system more efficient and accurate in grouping OKU Timur script images and to preserve the digital culture of using the OKU Timur script for future generations. Based on the existing issues, the researcher is conducting a study titled "Clustering OKU Timur script images using VGG feature extraction and K-Means"

II. RESEARCH METHODOLOGY

In the development of the clustering model for OKU Timur script images, this can be illustrated in the flowchart in Figure 1. The stages involve data collection, preprocessing, feature extraction, clustering model training, and evaluation of clustering results.

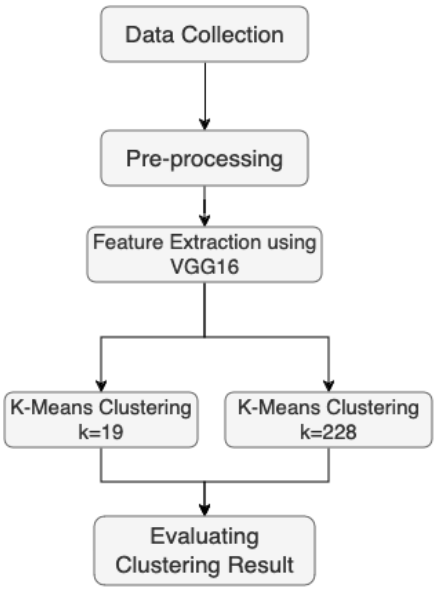

Figure 1. Flowchart

A. Data Collection

The data on the OKU Timur script was obtained from the Traditional Leader and Script Expert of OKU Timur. The data collection was previously conducted by another team from the ISRG (Intelligent System Research Group), and after the data was collected, there were 228 characters of the OKU Timur script. Before performing clustering, several processes must be carried out, such as preprocessing. At this stage, the data collection process for the OKU Timur script has involved local traditional leaders and script experts, covering information about the usage and characteristics of the script. Previously, the ISRG (Intelligent System Research Group) team had collected preliminary data regarding the OKU Timur script, including its use in cultural contexts and technical aspects.

Figure 2. Example filled Questionnaire for the OKU Timur Script

Additional data from the traditional leaders and script experts aims to deepen the existing understanding. The obtained data is then processed into a questionnaire to be distributed to respondents. The image below is an example of the OKU Timur script questionnaire that will be filled out by respondents, consisting of 12 punctuation marks of the OKU Timur script.

After being filled out by the respondents, the questionnaire aims to identify the uniqueness of each respondent's writing. This questionnaire contains 228 characters of the OKU Timur script and was completed by 102 respondents. Figure 2 is an example of the filled questionnaire from the respondent.

B. Preprocessing

In the preprocessing stage conducted on the raw data, the goal is to improve data quality so that clustering can be performed more effectively and yield good groups. The processes involved include downloading and extracting data, then saving, printing the path to the directory, and organizing, checking, and processing the list of image files to prepare them for the K-Means model training. Preprocessing is an important step in data clustering. In this study, 10 sample questionnaire forms were used, with a total of 228 character image types, resulting in a total of 2280 images collected. The number of images taken was then inputted as 2280 images. This process is conducted to prepare the data for use by the model. Several stages are performed during preprocessing, including: data

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i1.2292, Copyright ©2025
Submitted : October 7, 2024, Revised : November 4, 2024, Accepted : November 25, 2024, Published : January 31, 2025

downloading and extraction zip data.

This research begins by downloading data files from an external URL using the 'wget' command. This command allows the system to automatically download files from the internet and save them in a specified location; in this case, the downloaded file is aksara_okut.zip

Next, the file extraction process is carried out using the zipfile module in Python. This module is used to read and extract content from ZIP files. The extraction process involves opening the downloaded file, after which all contents are extracted to a specified directory ('/content/'). After that, to ensure the extraction is successful, the program will display a list of files extracted from that folder.

C. Feature Extraction using VGG16 model

To transform raw image data into a more manageable and informative format, which facilitates effective clustering and analysis, feature extraction is performed. Feature extraction plays a critical role in the process of image clustering because it can reduce dimensions. Raw image data is typically high-dimensional and complex, making it difficult to process directly. Feature extraction reduces the dimensionality by transforming images into a lower-dimensional space while retaining important information. This simplifies the clustering process and can improve clustering performance. Effective feature extraction enhances the quality of clusters formed by algorithms like K-Means [25].

In feature extraction, a pre-trained *VGG16* model trained on the ImageNet dataset will be used. VGG is a convolutional neural network model for image recognition proposed by the Visual Geometry Group at the University of Oxford, where VGG16 refers to a VGG model with 16 weight layers [26]. On the VGG16 model but with the final layer (the classification layer) removed. Instead of utilizing the final fully connected layers for classification, features are extracted from one of the last convolutional layers (typically one of the layers before the global average pooling layer). The output from this layer is a high-level representation of the image features in a lower-dimensional space.

For more details on the feature extraction process, refer to Figure 3. The workflow begins with importing the necessary libraries, specifically Keras, particularly `keras.applications.vgg16`, along with any other required libraries. The next step is importing modules for using the VGG16 model.

```
model = VGG16(weights='imagenet', include_top=False)
```

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i1.2292, Copyright ©2025
Submitted : October 7, 2024, Revised : November 4, 2024, Accepted : November 25, 2024, Published : January 31, 2025

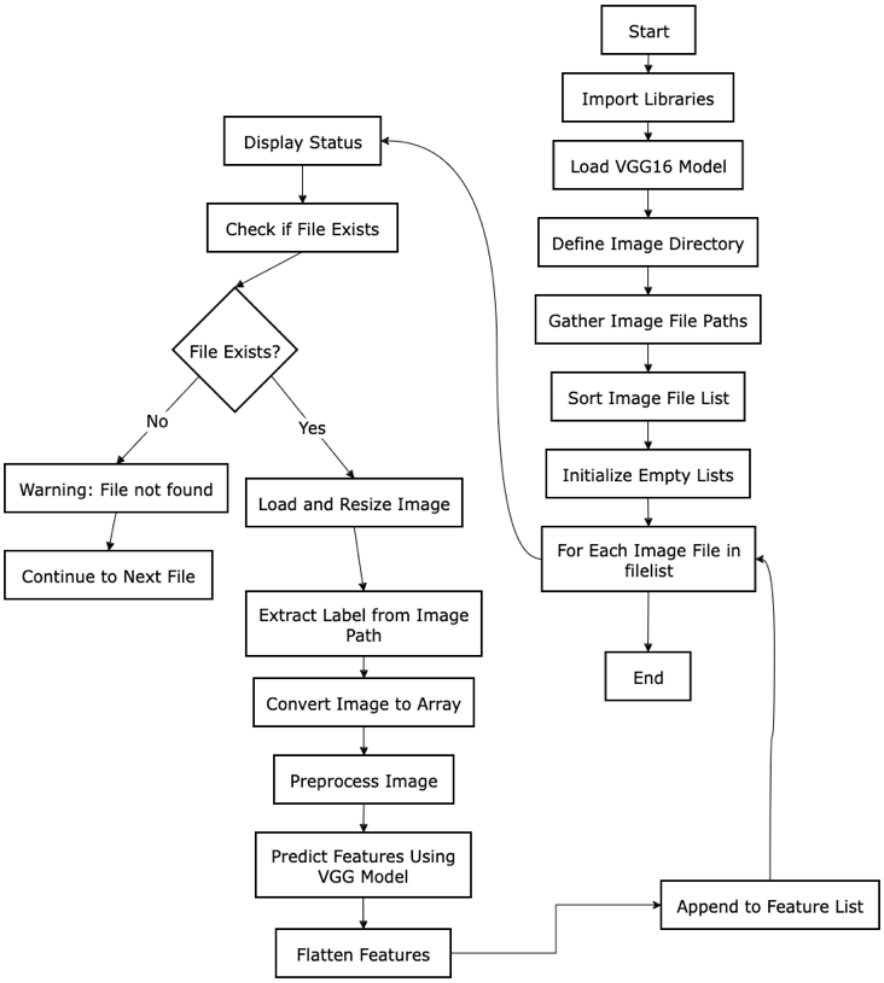

Figure 3. Workflow for feature extraction using the VGG16 model.

At this stage, the task is to create the VGG16 model without the final classification layer, so it can be used for feature extraction from images, such as clustering. In this process, there will be no output displayed because the VGG16 model here serves only as an additional feature to ensure the code runs smoothly without errors.

The next step involves preprocessing all files in the directory to change the file dimensions to match the input dimensions of the VGG16 model. Then, all images have their features extracted and stored in a flattened feature list until all images have been processed. The final result of this feature extraction process produces images with smaller sizes and dimensions corresponding to the feature extraction that was performed.

D. K-Means Clustering

Next, the stage of determining the number of clusters is conducted, as the researchers identified approximately 19 base characters of the OKU Timur script. Each character has 12 variants diacritics, resulting in a total of 228-character variants. Therefore, in the clustering, two types of cluster numbers are used: 19 and 228. These numbers are based on the number of base characters and the variant characters based on the variants.

After the data undergoes feature extraction using the VGG16 model, the clustering process is then performed using K-Means.

```
# Variables
number_clusters = 228

# Clustering
kmeans = KMeans(n_clusters=number_clusters,
random_state=0).fit(np.array(featurelist))
```

The model suitable for clustering the OKU Timur script images is K-Means because it is simple, fast, efficient, and flexible in determining the number of clusters [21], [22]. In the Clustering Model Training, a K-Means model is created with the number of clusters specified by 'number_cluster' and 'random_state=0' for consistent results. The K-Means model is then trained on the image features stored in 'featurelist', which have been converted into a NumPy array. The purpose of this is to group images based on character similarity into the predetermined clusters.

E. Evaluating Clustering Result

To evaluate the clustering results, two approaches are used:

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i1.2292, Copyright ©2025
Submitted : October 7, 2024, Revised : November 4, 2024, Accepted : November 25, 2024, Published : January 31, 2025

Check Cluster Consistency and Calinski-Harabasz Index.
Check Cluster Consistency: Review sample images from each cluster to ensure they are meaningful and consistent [27]. The main objective of this clustering process is to automatically group similar characters, so it is necessary to evaluate the homogeneity of the resulting clusters. To assess this, we calculate how many instances of the same character are correctly grouped together. An analysis is conducted to determine whether an ideal cluster should consist of a single character. If multiple characters are observed within a single cluster, it will be labeled as a cluster with the largest number of character variants or the major cluster. Other characters that fall into this cluster are considered misclassified characters. The final result will show the total number of character images that are accurately clustered compared to the total number of image data overall.

Calinski-Harabasz Index (CHI), also known as the Variance Ratio Criterion, is used to assess the quality of the cluster partitioning generated by the clustering algorithm [28]. This index is calculated by comparing the intra-cluster variance and the inter-cluster variance. The steps to calculate the Calinski-Harabasz Index are presented in equation (1).

$$\left[\text{CHI} = \frac{B_k/(k-1)}{W_k/(n-k)}\right] \quad (1)$$

where:

- (B_k) = *between-cluster variance*
- (W_k) = *within-cluster variance*
- (k) = *Number of clusters*
- (n) = *Total number of objects/data*

This stage involves analyzing the clustering results, observing how the data is grouped, and using the resulting dataset to train deep learning and transfer learning models for classifying the OKU Timur script images.

III. Result and Discussion

Figure 4 shows a visualization of how the image data consisting of 2,280 images undergoes feature extraction using the VGG16 model. The output from the feature extraction process comes from the last convolutional layer, resulting in an output tensor of size (7, 7, 512) for input images sized 224x224. The extracted feature tensors are then grouped using the K-Means method for the number of clusters = 19 and the number of clusters = 228.

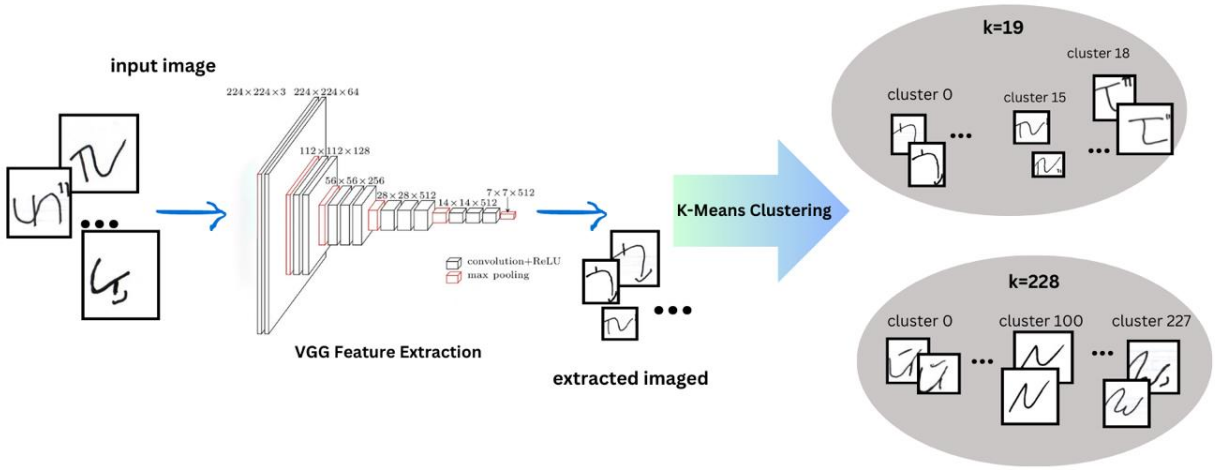

Figure 4. Image Feature extraction process and clustering image results with K-Means

A. Clustering Result

After performing clustering, the next step is to examine the results. The visual representation of images in each cluster provides strong visual confirmation of how these images are grouped by the model. This can help in understanding and validating whether the clustering has successfully separated the characters of the OKU Timur script into the predetermined groups. The clustering results for 19 clusters are presented in Table 1, while the clustering results for the number of classes = 228 are presented in Table 2.

TABLE I. CLUSTERING RESULTS UNTUK NUMBER CLUSTER= 9

Label cluster	Number of images	Mayor Images	Number of Major Images	Percentage of Major Class	Mis-Cluster
0	115	S	115	100%	-
1	182	R	130	71.4%	L, Ny, A
2	111	W	111	100%	-
3	211	J	116	55%	N, W, K
4	117	Ny	117	100%	-
5	109	Ng	109	100%	-
6	68	G	68	100%	-
7	112	L	112	100%	-
8	172	P	92	53.5%	L, R, Y, H
9	106	D	106	100%	-

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i1.2292, Copyright ©2025
Submitted : October 7, 2024, Revised : November 4, 2024, Accepted : November 25, 2024, Published : January 31, 2025

10	121	Y	104	86%	B, P, M
11	121	M	109	90%	P, Ny
12	140	B	113	80.71%	H, L, N, P
13	51	C	51	100%	-
14	138	K	92	66.7%	N, L
15	100	A	92	92%	L, Y
16	70	C	70	100%	-
17	119	H	64	53.8%	N, G, Ng
18	117	T	112	95.7%	D, J, Ng
Total	**2280**		**1883**		

The description of the table 1, 2 are as follows :

- *Class Label is the name of the folder representing the class.*
- *Number of Images is the total number of sample images available in each class category for model testing.*
- *Major refers to the typeface and punctuation in each class.*
- *Number of Major Images is the highest number of images owned by a single class in the directory.*
- *Percentage of Major Class is the proportion of the number of images in the majority class to the total number of images across all classes, expressed as a percentage.*
- *Mis-cluster refers to the grouping error where an item is placed in the wrong group.*

In the process of clustering using the entire dataset, which consists of 2,280 images, no data partitioning is performed. This is because in clustering, the concept of separation between training, validation, and testing, as used in supervised learning models, is not implemented in the same way. Clustering is an unsupervised learning method, where no labels are used to provide context to the data. Therefore, traditional partitioning is not necessary [29].

For cluster consistency with the number of clusters = 19 (Table 1, table 3), it can be seen that the clustering model accurately grouped 1,883 images, which represents the number of major characters in one cluster. Major images represent those that are accurately clustered. Overall, the clustering model achieved an accuracy of 82.6% in grouping the images. Upon closer inspection, there are 3 fundamental OKU Timur characters with significant grouping errors of less than 60%, namely characters J, P, and H. On the other hand, the clustering results with the number of clusters = 228 show a consistency value of 48.16% (table 3) for grouping images of the same character into the same cluster.

TABLE II. CLUSTERING RESULTS UNTUK NUMBER CLUSTER= 228

Label cluster	Num. of images	Mayor Images	Num. of Major Images	%of Major Class	Mis-Cluster
0	26	Bang	16	61.5%	Rang, Ri
1	13	Hu	6	46.2%	Ha
2	5	Ngah	2	40%	Ngai, Ngau
3	6	Rai	2	33.3%	R, Ran
4	19	Wa	6	31.6%	Wi, War
5	21	Yu	10	47.6%	Ya, Yi, Yan
6	25	Ai	15	60%	Ao, An
7	9	Nyan	5	55.6%	Nyu, Nyar
8	5	Ng	3	60%	Ngan, Ngah
9	10	Jau	6	60%	Jan, Jar
10	8	Dan	3	37.5%	Do, Dau
11	9	Di	5	55.6%	Dah, Da
12	2	Car	1	50%	Co
13	11	P	8	72.7%	Pan
14	7	J	5	71.4%	Nyan
15	20	Lau	10	50%	Le, Lo
16	15	Ki	8	53.3%	Kar, Ka
17	34	Gi	9	26.5%	Gah, Gan
18	12	Ngar	7	58.3%	Ngang, Ngi
19	16	Pang	7	43.8%	Par, Pa, Po
20	8	T	4	50%	Ta, Tar, Tang
21	8	H	8	100%	-
22	6	Ro	4	66.7%	Re, Rar
23	13	Car	5	38.5%	Cang, C, Can
24	3	D	2	66.7	Dah
25	20	L	8	40%	La, Lar, Lan
26	14	A	7	50%	Ah, Ar
27	7	Sai	3	42.9%	Sah, Sau
28	3	Ngai	2	66.7%	Ngan
29	9	Ngang	4	44.4%	Ngan, Ngah, Nga
30	2	Ta	1	50%	Tan
31	16	Kai	6	37.5%	Kan, Ka, K, Ke
32	10	Rau	6	60%	Ro, Rah
33	17	Ng	7	41.2%	Ng, Ngau, Ngah
34	9	G	3	33.3%	Gah, Gang, Gu
35	23	Jai	7	30.4%	Jah, J, Jang
36	5	D	3	60%	Da, Dah
37	21	Y	7	33.3%	Ya, Yah, Yan
38	13	Mau	5	38.5%	Ma, Mah
39	4	Ca	2	50%	Cau, Cu
…	…	…	…	…	…
…	…	…	…	…	…
220	19	Wang	6	31.6%	Wa, Wah, War
221	1	An	1	100%	-
222	6	Ja	3	50%	Jan, Jang, Je
223	1	War	1	100%	-
224	10	Nyan	5	50%	Ny, Nyai, Nyang
225	11	Do	6	54.5%	Da, Dar, De
226	10	Hai	6	60%	Han, Hau, Ha
227	6	Nye	3	50%	Nya, Nyah, Nyai
Total	**2280**		**1098**		

B. Evaluation and Discussion

Table 3 shows a summary of the performance metrics from the image clustering process. For the consistency results, it can be observed that the number of character variations and the variation of characters in 228 clusters makes the clustering process challenging. Similarly, for the values of the Calinski-Harabasz Index (CHI), which indicate the quality of the cluster separation produced by the clustering model. For a CHI of 39.85 for 19 clusters, it indicates that with 19 clusters, the model produces a fairly good separation. A higher value suggests that the between-cluster variance is relatively large compared to the within-cluster variance, meaning the formed clusters are well-defined and separated. On the other hand, a CHI of 7.28 for 228 clusters is lower compared to the value for 19 clusters. This may indicate that with more clusters (228), the separation between

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i1.2292, Copyright ©2025
Submitted : October 7, 2024, Revised : November 4, 2024, Accepted : November 25, 2024, Published : January 31, 2025

clusters becomes less clear, or that the clusters become too small, which can lead to redundancy or similarity between clusters. In general, a higher CHI value indicates better cluster quality. In this case, the model with 19 clusters appears to produce better separation compared to the model with 228 clusters

TABLE III. PERFORMANCE METRICS OF THE CLUSTERING RESULTS FOR OKU TIMUR SCRIPT

Number of cluster	Consistency	CHI
19	82.6%	39.85
228	48.16%	7.28

Several factors contribute to the significant errors in the clustering process in the context of clustering OKU Timur scripts, including the similarity in character shapes, particularly diacritics., which makes it difficult for the model to accurately cluster the glyph images (see example in Figure 5). Additionally, another factor is issues with the character's appearance that reduce the clarity of the letters in the writing. Furthermore, the quality of the images produced by the respondents is still lacking.

Figure 5. Examples of variant diacritic in OKU Timur script

Handling data quality is crucial for effective clustering and accurate results, particularly in tasks like clustering scripts or character recognition. To improve image data quality, several strategies can be employed for the development of future research. Data preprocessing techniques such as noise reduction using Gaussian or median filters can enhance clarity, while normalization of pixel values ensures uniform intensity. Resizing images to a fixed dimension is also important for consistency. Image enhancement methods, including contrast adjustment and binarization, can make characters more discernible, and sharpening techniques can enhance edge visibility. Additionally, data augmentation through transformations like rotation and scaling can create variations that help the model generalize better despite poor quality. Automated quality assessments and manual reviews can filter out low-quality images before clustering. Utilizing pre-trained models through transfer learning can adapt to varying image qualities, and ensuring high-quality image capture in future data collection is vital. Finally, outlier removal after initial clustering can mitigate the skewing effects of significantly different images. By implementing these strategies, the impact of poor image quality can be minimized, enhancing the effectiveness of clustering algorithms and leading to more accurate and meaningful results.

The main contribution of this research is the digital preservation of the OKU Timur script and the development of an automatic recognition system that can facilitate access and usage of the script in the future. This research paves the way for further studies in image processing and character recognition, which are crucial in preserving cultural heritage. As a recommendation, future research could explore alternative algorithms such as DBSCAN and Agglomerative Clustering, which may be more effective in handling irregular data distributions. Additionally, the use of optimization methods such as Grid Search or Random Search can help in determining better parameters for the clustering model.

To speed up processing time, the implementation of parallel processing or the use of GPUs can be a concrete solution. By leveraging this technology, it is expected that the time required to train the model and analyze the results can be minimized, allowing for the processing of larger and more complex datasets with greater efficiency. Overall, this research provides a strong foundation for further development in the automatic recognition of local cultural scripts and demonstrates the importance of technology in effectively preserving cultural heritage.

However, the challenges faced in clustering character images are quite complex, including variations in image quality, differences in writing styles, and the complexity of the character forms themselves. These uncertainties and variations can cause difficulties in the clustering process, where characters that appear similar may be expressed incorrectly, and vice versa. Therefore, an effective clustering development model not only helps in organizing data but also contributes to the digital preservation efforts of OKU Timur script. In this way, clustering serves not only as a technical tool but also as a link between technology and cultural preservation, ensuring that this heritage remains alive and accessible to future generations.

IV. CONCLUSION

This research successfully implemented a clustering model to group OKU Timur script images using feature extraction from the VGG16 model and the K-Means algorithm. The test results indicated that the model could group the majority of images with satisfactory accuracy, achieving 82.6% consistency in grouping major characters with 19 clusters. Evaluation metrics such as the Calinski-Harabasz Index provided insights into the quality of the cluster separation produced. However, challenges arose due to the similarity in character shapes and suboptimal image quality, leading to misclassification of some characters. Improvement strategies, such as applying image enhancement techniques and data augmentation, could be implemented to enhance future results. This research makes a significant contribution to the fields of image processing and character recognition, while also opening up opportunities for further studies to develop more effective and accurate clustering methods.

REFERENCES

[1] E. E. Panjaitan and N. Siregar, "THE IMPORTANCE OF LEARNING INDONESIAN LANGUAGE IN PRIMARY SCHOOL," *Ontol. J. PEMBELAJARAN DAN Ilm. Pendidik.*, vol. 2, no. 1, pp. 37–46, 2024.

[2] E. Roza, "Aksara Arab-Melayu di Nusantara dan Sumbangsihnya dalam Pengembangan Khazanah Intelektual," *Tsaqafah*, vol. 13, no. 1, pp. 177–204, 2017.

[3] G. Aceto, V. Persico, and A. Pescapé, "A survey on information and communication technologies for industry 4.0: State-of-the-art, taxonomies, perspectives, and challenges," *IEEE Commun. Surv. Tutor.*, vol. 21, no. 4, pp. 3467–3501, 2019.

[4] A. Drajat, E. W. Harahap, and others, "Rajah dan Spiritualitas Lokal dalam Hukum Islam; Studi Analisis Tafsir Hermeneutik," *Jurisprudensi J. Ilmu Syariah Perundang-Undangan Dan Ekon. Islam*, vol. 16, no. 1, pp. 225–240, 2024.

[5] L. Johanson, "The history of Turkic," in *The Turkic Languages*,

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i1.2292, Copyright ©2025
Submitted : October 7, 2024, Revised : November 4, 2024, Accepted : November 25, 2024, Published : January 31, 2025

Routledge, 2021, pp. 83–120.
[6] C. Agus, S. R. Saktimulya, P. Dwiarso, B. Widodo, S. Rochmiyati, and M. Darmowiyono, "Revitalization of local traditional culture for sustainable development of national character building in Indonesia," *Innov. Tradit. Sustain. Dev.*, pp. 347–369, 2021.
[7] D. Iskandar, S. Hidayat, U. Jamaludin, and S. M. Leksono, "Javanese script digitalization and its utilization as learning media: an etnopedagogical approach," *Int. J. Math. Sci. Educ.*, vol. 1, no. 1, pp. 21–30, 2023.
[8] I. Siregar, "Papuan Tabla Language Preservation Strategy," *LingLit J. Sci. J. Linguist. Lit.*, vol. 3, no. 1, pp. 1–12, 2022.
[9] Y. N. Kunang, I. Z. Yadi, Mahmud, and M. Husin, "A New Deep Learning-Based Mobile Application for Komering Character Recognition," in *2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, Yogyakarta, Indonesia: IEEE, Dec. 2022, pp. 294–299. doi: 10.1109/ISRITI56927.2022.10053072.
[10] T. P. Sari and Y. N. Kunang, "Pengembangan Aplikasi Transliterasi Teks Latin ke Aksara Ulu (Komering) Berbasis Web," *J. Process.*, vol. 18, no. 2, 2023.
[11] S. Huang, H. Wang, Y. Liu, X. Shi, and L. Jin, "OBC306: A large-scale oracle bone character recognition dataset," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, IEEE, 2019, pp. 681–688.
[12] R. Deng *et al.*, "Automatic Identification of Sea Rice Grains in Complex Field Environment Based on Deep Learning," *Agriculture*, vol. 14, no. 7, p. 1135, 2024.
[13] A. E. Ezugwu, A. K. Shukla, M. B. Agbaje, O. N. Oyelade, A. José-García, and J. O. Agushaka, "Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature," *Neural Comput. Appl.*, vol. 33, pp. 6247–6306, 2021.
[14] I. Chatterjee, M. Ghosh, P. K. Singh, R. Sarkar, and M. Nasipuri, "A clustering-based feature selection framework for handwritten Indic script classification," *Expert Syst.*, vol. 36, no. 6, p. e12459, 2019.
[15] A. R. Widiarti, G. R. Prima, and C. K. Adi, "Preliminary research for provision of Javanese script image dataset from Javanese script printed book," in *AIP Conference Proceedings*, AIP Publishing, 2024.
[16] J. Oyelade *et al.*, "Data clustering: Algorithms and its applications," in *2019 19th international conference on computational science and its applications (ICCSA)*, IEEE, 2019, pp. 71–81.
[17] S. Setyaningtyas, B. I. Nugroho, and Z. Arif, "Tinjauan Pustaka Sistematis: Penerapan Data Mining Teknik Clustering Algoritma K-Means," *J. Teknoif Tek. Inform. Inst. Teknol. Padang*, vol. 10, no. 2, pp. 52–61, 2022.
[18] A. Ghosal, A. Nandy, A. K. Das, S. Goswami, and M. Panday, "A short review on different clustering techniques and their applications," *Emerg. Technol. Model. Graph. Proc. IEM Graph 2018*, pp. 69–83, 2020.
[19] S. Panda, M. Nayak, and A. K. Nayak, "Clustering of Odia character images using K-means algorithm and spectral clustering algorithm," in *ICICCT 2019–System Reliability, Quality Control, Safety, Maintenance and Management: Applications to Electrical, Electronics and Computer Science and Engineering*, Springer, 2020, pp. 55–64.
[20] A. R. Widiarti and C. K. Adi, "Clustering Balinese Script Image in Palm Leaf Using Hierarchical K-Means Algorithm," in *International Conference on Innovation in Science and Technology (ICIST 2020)*, Atlantis Press, 2021, pp. 38–42.
[21] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Inf. Sci.*, vol. 622, pp. 178–210, 2023.
[22] N. H. Shrifan, M. F. Akbar, and N. A. M. Isa, "An adaptive outlier removal aided k-means clustering algorithm," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 8, pp. 6365–6376, 2022.
[23] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Inf. Sci.*, vol. 622, pp. 178–210, 2023.
[24] M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means algorithm: A comprehensive survey and performance evaluation," *Electronics*, vol. 9, no. 8, p. 1295, 2020.
[25] S. Sen, P. Chakraborty, S. Das, K. Pandey, and P. Narayana, "Investigation of Clustering Methods for SDSS Galaxy Images through Feature Extraction with VGG-16," in *2024 IEEE Space, Aerospace and Defence Conference (SPACE)*, IEEE, 2024, pp. 660–664.
[26] S. Tammina, "Transfer learning using vgg-16 with deep convolutional neural network for classifying images," *Int. J. Sci. Res. Publ. IJSRP*, vol. 9, no. 10, pp. 143–150, 2019.
[27] Y. Ren *et al.*, "Deep clustering: A comprehensive survey," *IEEE Trans. Neural Netw. Learn. Syst.*, 2024.
[28] I. Ioannou, C. Christophorou, P. Nagaradjane, and V. Vassiliou, "Performance Evaluation of Machine Learning Cluster Metrics for Mobile Network Augmentation," in *2024 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET)*, IEEE, 2024, pp. 1–7.
[29] W. Bao, N. Lianju, and K. Yue, "Integration of unsupervised and supervised machine learning algorithms for credit risk assessment," *Expert Syst. Appl.*, vol. 128, pp. 301–315, 2019.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i1.2292, Copyright ©2025
Submitted : October 7, 2024, Revised : November 4, 2024, Accepted : November 25, 2024, Published : January 31, 2025