

Handwritten Balinese Script Recognition on Palm Leaf Manuscript using Projection Profile and K-Nearest Neighbor

Ni Putu Sutramiani^{a1}, I Wayan Agus Surya Darma^{b2}, Dewa Made Sri Arsaa^{ac3}

^aDepartment of Information Technology, Faculty of Engineering, Universitas Udayana, Bukit Jimbaran, Indonesia, 80361

^bDepartment of Informatics, Faculty of Technology and Informatics, Institut Bisnis dan Teknologi Indonesia, Denpasar, Indonesia, 80225

^cDivision of Electronics and Information Engineering, Jeonbuk National University, Republic of Korea

e-mail: ¹sutramiani@unud.ac.id, ²surya@instiki.ac.id, ³202155383@jbnu.ac.kr

Abstract

This paper presents a simple approach to the handwritten Balinese script characters recognition in palm-leaf lontar manuscripts. The Lontar manuscript is one of the cultural heritages found in Bali. Lontar manuscripts are written using a pengrupak, which is a kind of knife for writing on palm leaves. Roasted candlenut powder is used to give color to the writing; hence the characters appear clear. The research applied the projection profile at the segmentation stage to get the handwritten Balinese script characters in the lontar manuscript. The palm leaf manuscript is acquired from Wariga Palalubangan manuscript. The recognition process is carried out by implementing K-Nearest Neighbor in the recognition process. The recognition was made on the Wianjana script obtained from lontar manuscripts using 720 images consisting of 18 classes as dataset training. The test results showed that the level of recognition accuracy was obtained by 52% in the characters of handwritten Balinese scripts derived from lontar manuscripts and 92% in the characters of handwritten Balinese scripts on paper.

Keywords: Balinese Script, Handwritten Character, K-Nearest Neighbor, Palm-Leaf, Projection Profile

1. Introduction

The *Lontar* manuscript is one of the cultural heritages found in Bali. The Lontar manuscripts are written using a *pengrupak*, which is a kind of knife for writing on palm leaves. To give color to the results of the writing, candlenut is used so that the writing appears clear. The Lontar manuscripts are written with Balinese script characters [1]. The Balinese script used in this study is the *Wianjana* script consisting of 18 characters.

Research on character segmentation remains a challenge, especially the segmentation of handwritten characters. Projection profile method has been implemented for Balinese script in [2]. Handwriting character segmentation of Arabic characters has been carried out using a new vertical segmentation algorithm. The results of the study show higher accuracy with excellent performance by improving segmentation in the case of interlocking characters [3]. Segmentation techniques have also been applied to Sinhala Handwritten Characters by using pixel labeling techniques to segment overlapping characters [4]. The application of character segmentation with a pixel-based approach and bounding boxes has also been done on handwritten characters with a segmentation rate of 94.45% [5]. The results of segmentation are very dependent on the character object used. Handwritten characters tend to have variations in each writing because it depends on the style of the person who wrote. Printed writing is certainly easier to segment because the character shape will always be the same. Research on character segmentation of printed writing has been done on multilingual Indian document images of Latin and Devanagari scripts which result in segmentation rates of up to 98.86% [6].

The type of media used as a manuscript is also a challenge in character recognition. Image quality improvement to reduce noise has been done on lontar manuscript images using Local Adaptive Thresholding [7]. Preprocessing becomes a very important stage to get the characters in an image. Research on preprocessing thinning has been carried out on lontar manuscripts using Zhang-Suen to produce characters with a thickness of one pixel [8]. Recognition of Tamil characters written on manuscripts made from palm leaf media has been carried out using the Canny Edge Detection Algorithm to examine and delete characters from damaged images [9].

Research on Amharic character recognition has been carried out using a combination of features and Support Vector Machine. The paper discusses the combination of various feature extraction techniques and SVM for the introduction of Amhari characters [10]. Related research on character recognition has been carried out on Arabic characters using decision trees and perception codes. The experimental results in this study indicate the level of accuracy of recognition depends on the way of writing Arabic characters [11]. Handwriting character recognition is still a challenge in the field of pattern recognition. Handwriting character recognition has been carried out on Tamil characters using multi-layered feed-forward neural networks with a back-propagation algorithm [12]. The previous handwritten Balinese script recognition was done by applying K-Nearest Neighbor. The data used in the study were Balinese characters written on paper [13]. Various techniques have been used in the handwritten Balinese script recognition, one of which is by dividing the image area into several zones in the feature extraction process, using semantic features and implementing K-Nearest Neighbor in the recognition process [14]–[16]. The KNN also commonly used in other research e.g., baby foot identification [17], [18].

The segmentation process applied the projection profile at the segmentation stage to get the Balinese script characters handwritten in the lontar manuscript. The palm leaf manuscript is acquired from the *Wariga Palalubangan* manuscript, which was written using a screwdriver, which is a kind of knife for writing on the palm leaf. To color the writing on the palm-leaf, candlenuts are used to give black color. The recognition process is carried out by implementing K-Nearest Neighbor in the recognition process. The recognition was made on the *Wianjana* script obtained from lontar manuscripts using 720 images consisting of 18 classes as training data.

2. Research Methodology

This study uses data from the first page in the *Wariga Palalubangan* manuscript. It was written using Balinese characters. The Projection Profile and K-Nearest Neighbor is utilized to segment the palm leaf manuscript and recognize the Balinese characters. Figure 1 shows the proposed methodology.

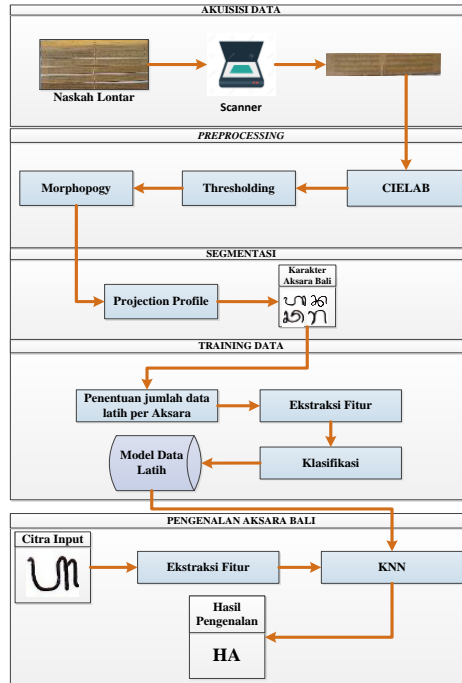


Figure 1. Proposed Methodology

2.1 Data Acquisition

This study uses the *Wariga Palalubangan* manuscript written in Balinese characters. The data acquisition process uses a scanner to get a lontar script image with * jpg format. The data acquisition process in this study which is shown in Figure 2.



Figure 2. Image Acquisition using Scanner

2.2 Pre-processing

The preprocessing stage consists of three processes, determining the color space, thresholding, and morphology. The CielLAB color space is used to determine the pixel position of Balinese scripts. Thresholding is used to separate objects of Balinese script characters with background and morphology to turn Balinese script characters into one-pixel size.

2.3 Segmentation

The projection profile is utilized at the segmentation stage to get each Balinese script characters in the lontar manuscript. Projection profiles make vertical and horizontal projections of the Balinese script characters. The result of the segmentation stage is the character image of the Balinese script which has been segmented according to their respective characters.

2.4 Training Data

The training phase is the stage to train the dataset in machine learning to make predictions based on the data being tested. So that machines can be made to learn according to the dataset being trained. The training process begins with the feature extraction process. In this stage, the training process train the image of Balinese script characters consisting of 18 classes, training on 720 data, each class containing 40 data. The training process refers to previous research [15]. This process is carried out by extracting features in the Balinese script image dataset. The resulting features are then trained to produce a model that is used at the recognition stage.

2.5 The Balinese Script Recognition

The recognition phase is carried out to test the machine that has been built by training the Balinese script dataset. The K-Nearest Neighbor (KNN) is utilized to classify the Balinese script image which is tested to produce the closest neighbor value which is the result of the recognition of the Balinese script character. A comparison of neighboring values is based on a comparison of the value of the Balinese script character image test with the dataset that has been trained.

3. Results and Discussions

3.1. Data Preparation

The data in this study are acquired from the *Wariga Palalubangan* manuscript imagery written using Balinese script characters. Figure 3 shows the *Wariga Palalubangan* Manuscript sample.



Figure 3. The *Wariga Palalubangan* Manuscript sample

Figure 3 shows the *Wariga Palalubangan* Manuscript image that used in the Balinese character recognition system testing. The *Wariga Palalubangan*'s palm leaf imagery was obtained from the Hindu Dharma Negeri Institute of Denpasar library. The image acquisition process uses a scanner with the image dimensions of the acquisition of 2048x191 pixels and 96 dpi. The Balinese character that was the object of introduction in this study is the *Wianjana* script. The *Wianjana* script consists of 18 characters as shown in Figure 4.

ha	na	ca	ra	ka	da	ta	sa	wa
ᮊ	ᮋ	ᮌ	ᮍ	ᮎ	ᮏ	ᮐ	ᮑ	ᮒ
la	ma	ga	ba	nga	pa	ja	ya	nya
ᮓ	ᮔ	ᮕ	ᮖ	ᮗ	ᮘ	ᮙ	ᮚ	ᮛ

Figure 4. *Wianjana* Script

3.2. Preprocessing

Lontar manuscripts made from ental leaves turn out to cause the image generated from the acquisition process to have noise. The preprocessing is needed to reduce the noise contained in the lontar image. It aims to separate the background from the Balinese characters so that they can be detected properly.

The local adaptive thresholding is utilized to produce binary images, which are images that have two gray level values, namely black and white [19]. The process of floating grayscale images to produce binary images, in general based on Equation 1.

$$G(x, y) = \begin{cases} 1 & \text{if } f(x, y) \geq T \\ 0 & \text{if } f(x, y) < T \end{cases} \tag{1}$$

In Equation 1, G (x, y) is a binary image of grayscale f (x, y). T states the threshold value that will affect the quality of the binary image results. The T value can be calculated using the Equations 2-4:

$$T = \frac{\sum_{(x,y)} \sum_{\epsilon \in W} f(x,y)}{N_W} - C \tag{2}$$

$$T = \text{median} \{f(x, y), (x, y) \in W\} \tag{3}$$

$$T = \frac{\max\{f(x,y), (x,y) \in W\} + \min\{f(x,y), (x,y) \in W\}}{2} \tag{4}$$

W states the number of blocks processed; NW is the number of pixels contained in each block W. C is a constant that can be determined freely. Equation 2 is used to calculate T with an average value, equation 3 is used to calculate T with a median value and equation 4 is used to calculate T with the average maximum and minimum pixel values in the window.

The preprocessing on lontar manuscripts is used to improve image quality by reducing noise. The following are the results of preprocessing in the lontar manuscript shown in Figure 5.

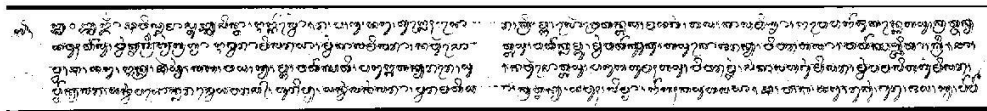


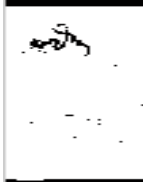

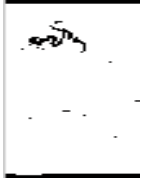
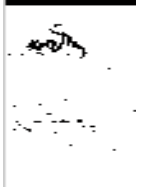
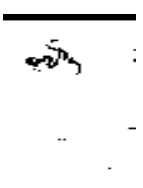


Figure 5. The Preprocessing Results

The thresholding result that shown in overall above do not show any difference in noise in each image. Therefore, the following is table 1 which shows the noise contained in the thresholding image in more detail.

Table 1. Preprocessing Details Based on W and C Value

No	Noise	Value
1		W = 10 C = 0.01
2		W = 100 C = 0.01

No	Noise	Value
3		W= 60 C = 0.05
4		W= 70 C=0.05
5		W= 80 C = 0.05
6		W= 70 C = 0.04
7		W= 70 C = 0.06

3.3. Segmentation

The projection profiles method is utilized to segment the Balinese script character in lontar manuscript. The projection profile make vertical and horizontal projections of the Balinese script characters. Figure 6 shows the horizontal projection profile and Figure 7 shows the vertical projection profile. The result of the segmentation stage is the character image of the Balinese script which has been segmented according to their respective characters.

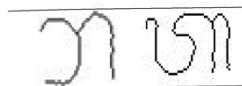


Figure 6. Horizontal Projection Profile



Figure 7. Vertical Projection Profile

3.4. Feature Extraction

The feature extraction process produces six types of features that are used to identify the specific patterns possessed by each Balinese script character. Previous studies of these six types of features were used to recognize handwritten Balinese characters[15]. Figure 8 below shows the features produced in Balinese characters.

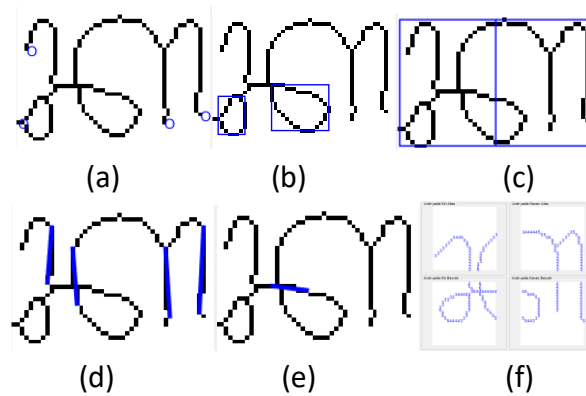


Figure 8 Character Features of Balinese Script: (a) Stop Point, (b) Loop, (c) Length and Width, (d) Vertical Line, (e) Horizontal Line and (f) Direction Feature

The six types of features produced consist of 28 features that will be used in the recognition process. Table 2 shows the 28 features details.

Table 2. Feature Details

Balinese Script	Feature Types	Value
	The number of vertical on the top left	17
	The number of right diagonals on the top left	15
	The number of horizontal on the top left	3
	The number of left diagonals on the top left	3
	The number of vertical on the top right	9
	The number of right diagonals on the top right	17
	The number of horizontal on the top right	7
	The number of left diagonals on the top right	28
	The number of vertical on the bottom left	16
	The number of right diagonals on the bottom left	8
	The number of horizontal on the bottom left	3
	The number of left diagonals on the bottom left	13
	The number of vertical on the bottom right	16
	The number of right diagonals on the bottom right	10
	The number of horizontal on the bottom right	3
	The number of left diagonals on the bottom right	10
	The number of stop points	4
	The number of stop points on the top left zone	1
	The number of stop points on the top right zone	0
	Number of stop points on the bottom left zone	1

Balinese Script	Feature Types	Value
	Number of stop points on the bottom right zone	2
	The number of characters length	2
	the number of characters width	1
	The number of loops	2
	The number of horizontal	1
	The number of vertical	4
	The number of vertical on the left zone	2
	The number of vertical on the right zone	2

























Table 2 shows the detailed features produced in the feature extraction process on the image of Balinese script Ka. There are 28 features produced that contain values on each type of feature. The feature values produced by each character are used for the process of Balinese script character recognition.



























3.5. Recognition

The experiments are conducted using two scenario at the Balinese script character recognition stage. The first experiment was carried out using 50 images of Balinese *Wianjana* script obtained from segmentation in lontar manuscripts. The character image from the segmentation result used is a *Wianjana* character that is successfully segmented exactly one character. Unsuccessful characters are not used in this first experiment. Table 4 shows the sample result of the second experiment. The second experiment was carried out on 50 images of the Balinese script written on paper by 50 writers. The total training data used are 720 images consisting of 18 classes. In each experiment, the implementation of KNN using $K = 3$ [20]

In the first experiment, 50 Balinese characters from lontar manuscript are used as testing data. The Balinese characters are resulted from segmentation process using projection profile method. The experiment shows that the KNN method using $K = 3$ yielded 26 correct recognition, 17 incorrect recognition, and seven failed to recognize. Table 3 shows the sample result of the first experiment























Table 3. Sample Result of the Experiment I

Balinese Script	Recognition Result	Balinese Script	Recognition Result
	Incorrect		Correct
	Incorrect		Correct
	Correct		Failed to Recognize
	Incorrect		Correct
	Incorrect		Incorrect
	Correct		Failed to Recognize
	Correct		Correct
	Correct		Correct
	Correct		Incorrect
	Correct		Incorrect
	Incorrect		Correct
	Failed to Recognize		Correct

Balinese Script	Recognition Result	Balinese Script	Recognition Result
	Correct		Correct
	Incorrect		Incorrect
	Correct		Failed to Recognize
	Incorrect		Correct
	Correct		Incorrect
	Incorrect		Incorrect
	Incorrect		Correct
	Incorrect		Incorrect
	Correct		Correct
	Correct		Failed to Recognize
	Failed to Recognize		Correct
	Correct		Correct
	Failed to Recognize		Correct

In the second experiment, 50 images of the Balinese script written on paper by 50 writers are used as testing data. The experiment shows that the KNN method using K= 3 yielded 46 correct recognition and 4 incorrect recognitions. Table 4 shows the sample result of the second experiment.

Table 4. Sample Result of the Experiment II

Balinese Script	Recognition Result	Balinese Script	Recognition Result
	Correct		Correct
	Incorrect		Correct
	Correct		Correct
	Correct		Correct
	Correct		Correct
	Correct		Incorrect
	Correct		Incorrect
	Correct		Correct
	Correct		Correct
	Correct		Correct
	Correct		Correct

Balinese Script	Recognition Result	Balinese Script	Recognition Result
ꦱ	Correct	ꦱ	Correct
ꦱ	Correct	ꦱ	Correct
ꦱ	Incorrect	ꦱ	Correct
ꦱ	Correct	ꦱ	Correct
ꦱ	Correct	ꦱ	Correct
ꦱ	Correct	ꦱ	Correct
ꦱ	Correct	ꦱ	Correct
ꦱ	Correct	ꦱ	Correct
ꦱ	Correct	ꦱ	Correct
ꦱ	Correct	ꦱ	Correct
ꦱ	Correct	ꦱ	Correct
ꦱ	Correct	ꦱ	Correct

Table 5. Comparison of Recognition Accuracy

	Correct	Incorrect	Failed to Recognize	Accuracy
Experiment I	26	17	7	52%
Experiment II	46	4	0	92%

Table 5 shows the comparison of accuracy between the two tests that have been carried out. The first test results showed 52% accuracy obtained from the *Wianjana* script image in the lontar manuscript. This result was greatly influenced by the Balinese script image contained in the lontar manuscript. The Balinese script is written in a lontar manuscript using a *pengrupak* which is a kind of knife then rubbed with roasted candlenut, so that the Balinese character is black in color. The use of *pengrupak* as a writing tool makes Balinese script writing on lontar manuscripts quite difficult to recognize because there is quite a lot of writing noise.

In the second test, the Balinese script images written by 50 different writers are used as testing data. To find out the comparison of the accuracy of the recognition to the Balinese script written on lontar manuscripts and paper objects. In the second test, the experiment yielded 92% recognition accuracy. A significant difference in accuracy was obtained in the two tests that were carried out.

4. Conclusion

The projection profile and KNN can produce the recognition accuracy of 52% that obtained in the characters of Balinese handwritten scripts derived from lontar manuscripts and 92% from the characters of Balinese handwritten scripts on paper. The result was greatly influenced by the Balinese script image contained in the lontar manuscript. The Balinese script is written on a lontar manuscript using a *pengrupak* which is a kind of knife and then rubbed with roasted candlenut to give color to the writing. The use of *pengrupak* as a writing instrument makes Balinese script writing on the lontar manuscript quite difficult to recognize. Handwritten characters also tend to vary with each writing because it depends on the writer's style.

References

- [1] I. N. Duija, "Keberadaan Aksara Wrésastra Dalam Aksara Bali the Existenace of Wrésastra in Balinese Script," *Kajian Budaya, Institut Hindu Dharma Negeri Denpasar*, vol. 29, no. 51, p. 1, 2017.
- [2] I. W. A. S. Darma and N. P. Sutramiani, "Segmentation of Balinese Script on Lontar Manuscripts using Projection Profile," in *2019 5th International Conference on New Media Studies (CONMEDIA)*, 2019, pp. 212–216. doi: 10.1109/CONMEDIA46929.2019.8981860.
- [3] A. A. A. Ali and M. Suresha, "An Efficient Character Segmentation Algorithm for Recognition of Arabic Handwritten Script," *2019 International Conference on Data Science and Communication, IconDSC 2019*, pp. 1–6, 2019, doi: 10.1109/IconDSC.2019.8817037.
- [4] K. S. A. Walawage and L. Ranathunga, "Segmentation of Overlapping and Touching Sinhala Handwritten Characters," *2018 3rd International Conference on Information Technology Research, ICITR 2018*, pp. 1–6, 2018, doi: 10.1109/ICITR.2018.8736129.
- [5] M. Arun, S. Arivazhagan, and D. Rathina, "Handwritten text segmentation using pixel-based approach," *Proceedings of the International Conference on Trends in Electronics and Informatics, ICOEI 2019*, vol. 2019-April, no. Icoei, pp. 791–796, 2019.
- [6] P. Sahare and S. B. Dhok, "Multilingual Character Segmentation and Recognition Schemes for Indian Document Images," *IEEE Access*, vol. 6, no. ii, pp. 10603–10617, 2018, doi: 10.1109/ACCESS.2018.2795104.
- [7] N. P. Sutramiani, D. Putra, and M. Sudarma, "Local Adaptive Thresholding Pada Preprocessing Citra Lontar Aksara Bali," *Majalah Ilmiah Teknologi Elektro*, vol. 14, no. 1, pp. 27–30, 2015.
- [8] M. Sudarma and N. P. Sutramiani, "The Thinning Zhang-Suen Application Method in the Image of Balinese Scripts on the Papyrus," *International Journal of Computer Applications*, vol. 91, no. 1, pp. 9–13, 2014.
- [9] P. Selvakumar and S. Hari Ganesh, "Tamil Character Recognition Using Canny Edge Detection Algorithm," *Proceedings - 2nd World Congress on Computing and Communication Technologies, WCCCT 2017*, pp. 250–254, 2017, doi: 10.1109/WCCCT.2016.68.
- [10] B. Y. Reta, D. Rana, and G. V. Bhalerao, "Amharic Handwritten Character Recognition Using Combined Features and Support Vector Machine," *Proceedings of the 2nd International Conference on Trends in Electronics and Informatics, ICOEI 2018*, no. Icoei, pp. 265–270, 2018, doi: 10.1109/ICOEI.2018.8553947.
- [11] H. Akouaydi, S. Abdelhedi, S. Njah, M. Zaied, and A. M. Alimi, "Decision trees based on perceptual codes for on-line Arabic character recognition," in *2017 IEEE International Workshop on Arabic Script Analysis and Recognition (ASAR)*, 2017, pp. 153–157. doi: 10.1109/asar.2017.8067778.
- [12] K. Vijayalakshmi, S. Aparna, G. Gopal, and W. Jino Hans, "Handwritten character recognition using diagonal-based feature extraction," *Proceedings of the 2017 International Conference on Wireless Communications, Signal Processing and Networking, WiSPNET 2017*, vol. 2018-Janua, pp. 1178–1181, 2018, doi: 10.1109/WiSPNET.2017.8299949.
- [13] I. W. A. S. Darma and N. K. Ariasih, "Handwritten Balinesse Character Recognition using K-Nearest Neighbor," in *International Conference on Culture Technology*, 2017, pp. 139–144.
- [14] I. W. A. S. Darma, D. Putra, and M. Sudarma, "Ekstraksi Fitur Aksara Bali Menggunakan Metode Zoning," *Majalah Ilmiah Teknologi Elektro*, vol. 14, no. 2, pp. 44–49, 2015.
- [15] I. W. A. S. Darma, "Implementation of Zoning and K-Nearest Neighbors in Character Recognition of Wrésastra Script," *Lontar Komputer: Jurnal Ilmiah Teknologi Informasi*, vol. 10, no. 1, pp. 9–18, 2019, doi: 10.24843/LKJITI.2019.v10.i01.p02.
- [16] M. Sudarma and S. Darma, "The Identification of Balinese Scripts Characters based on Semantic Feature and K Nearest Neighbor," *International Journal of Computer Applications*, vol. 91, no. 1, pp. 14–18, Apr. 2014, doi: 10.5120/15845-4727.
- [17] G. A. Lesmana, I. N. Piarsa, and I. M. S. Putra, "Identification of Baby's Feet Using Principal Component Analysis (PCA) Method Character Extraction with K-Nearest Neighbor (KNN) Classification in Matlab Application," *Jurnal Ilmiah Merpati (Menara*

- Penelitian Akademika Teknologi Informasi*), vol. 9, no. 3, pp. 200–212, May 2021, doi: 10.24843/JIM.2021.V09.I03.P02.
- [18] A. P. Nugraha, I. N. Piarsa, and I. M. S. Putra, “Comparison of Support Vector Machine and K-Nearest Neighbor for Baby Foot Identification based on Image Geometric Characteristics,” *Jurnal Ilmiah Merpati (Menara Penelitian Akademika Teknologi Informasi)*, pp. 84–95, Apr. 2021, doi: 10.24843/JIM.2021.V09.I01.P08.
- [19] D. Putra, *Pengolahan Citra Digital*. Yogyakarta: Andi, 2010.
- [20] I. W. A. S. Darma, “Implementation of Zoning and K-Nearest Neighbor in Character Recognition of Wrésastra Script,” *Lontar Komputer: Jurnal Ilmiah Teknologi Informasi*, vol. 10, no. 1, p. 9, 2019, doi: 10.24843/lkjiti.2019.v10.i01.p02.