Identification of Lampung Script Using K-Neighbor, Manhattan Distance And Population Matrix Algorithm

Identification of Lampung Script Using K-Neighbor, Manhattan Distance And Population Matrix Algorithm

Gladys Ivana Augusta[1], Lukman Hakim[1], Anna Gustina Zainal[2], Hendy Tannady[3*]

[1] Computer Science Department, Universitas Bunda Mulia, Indonesia 14430
[2] Communication Department, Lampung University, Indonesia 35141
[3] Graduate Program of Management, Institut Teknologi dan Bisnis Kalbis, Indonesia 13210

*hendytannady@gmail.com

Abstract. Language is a communication tool that is used as interaction with others. The use of indigenous languages is decreasing and erasing over time. Lampung script is a script that becomes the identity of the province of Lampung. However, only a few native Lampung residents who know the Lampung script and the younger generation often write Lampung in Latin letters because it is considered easier. This study uses Optical Character Recognition to recognize Lampung characters in image images using a smartphone. This study uses the Pixel Population Matrix feature extraction to extract the characteristics of the characters. The distance calculation for each test character and database uses Manhattan Distance and is classified using K-Nearest Neighbor with the value of k is 3. The results of the character recognition process can be translated into Indonesian. Testing this application is done by taking image samples with several conditions. The script image tested is in the form of a screenshot image of printed writing with random fonts, a handwritten photo image, a screenshot image with a combination of several words and a photo image with a smaller size and random slope. The test results according to these conditions obtained an average percentage of success of 91.49%.

Keywords: Lampung Script, Optical Character Recognition, K-Nearest Neighbor, Manhattan Distance, Pixel Population Matrix

1. Introduction

Based on BPS data, the 2010 Population Census states that there are 1,331 ethnic groups in Indonesia [1]. The diversity of ethnicities and nations in Indonesia has made Indonesia a multicultural country that has cultural diversity, one of which is script. However, not all regions in Indonesia have characters, so they need to be preserved and preserved. One of the areas that has script is the Lampung area. Lampung script or so-called Had Lampung is a script that is often used to communicate hereditary and become the identity of the province of Lampung. One of the uses of the Lampung script can be seen in the Lampung Museum where there are 2 inscriptions that read the Lampung script. In addition, the Lampung script is also used on the street name signposts, the Lampung province logo, the Bandar Lampung city logo, and the South Lampung logo.

However, Lampung Province is no longer dominated by indigenous Lampung tribes, but many tribes from outside Lampung Province. Based on ethnicity, according to data from the Lampung government, it turns out that the highest percentage of Javanese is 65.8 percent. Followed by Lampung 12.8 percent, Sundanese 11.36 percent, Minangkabau 3.57 percent, Batak 2.13 percent, Balinese 1.73 percent and other ethnic groups 2.15 percent [2]. This causes only a few native Lampung residents to

know the Lampung script. Unfortunately, the younger generation nowadays rarely use the Lampung script and are accustomed to writing Lampung in Latin because it is considered easier. If this is not corrected, the Lampung script will become extinct.

Based on the above, the use of the Lampung script in the indigenous people of Lampung is rarely used in daily communication. Therefore, the authors feel that there is an opportunity to preserve Lampung script in Indonesia by making an application that implements OCR so that the application can recognize Lampung script using a smartphone. The letter recognition application or often called OCR is an effective solution for the conversion process from printed documents into digital documents [3]. In the OCR application, the feature extraction process to get the characteristics of an object uses the extraction of the Pixel Population Matrix. Then the feature is measured for the closest distance and the longest distance to recognize the character using the distance measurement method, namely the Manhattan Distance method which will then be classified using the K-Nearest Neighbor algorithm. The purpose of this research is to design a Lampung script recognition application to preserve Lampung script, measure the accuracy of the K-Nearest Neighbor classification, measure the Manhattan Distance distance and feature extraction of the Pixel Population Matrix.

2. Methodology

The Lampung script which is called Had Lampung is a form of writing that has a connection with the Pallawa script from South India [4]. Lampung script consists of 20 main letters and 12 sub-letters. The main form of all parent letters is curves or curves [5]. In the Lampung script there are 20 main letters, namely 'ka', 'ga', 'nga', 'pa', 'ba', 'ma', 'ta', 'na', 'ca', 'ja', ' nya ',' yes', 'a', 'la', 'ra', 'sa', 'wa', 'ha', 'gha' [5].

Figure 1. Parent Letters of Lampung Script

Each parent letter of the Lampung script is read with the sound produced from the consonant letter 'a', but not all word fragments end in 'a', but other consonants are also used. To change the sound of a vowel in the parent letter into other vowels such as 'i', 'u', 'e', and 'o', you can use the Lampung script sub-letters. (Optical Character Recognition) OCR is a computer application used to identify images of letters and numbers to be converted into written files [6]. Optical Character Recognition (OCR) in a broad sense is a branch of artificial intelligence and computer vision. This letter recognition system can increase the flexibility or ability and intelligence of computers. An intelligent letter recognition system is very helpful in digitizing information and knowledge, for example in making digital library collections, ancient literary collections, and others [7].

The K-Nearest Neighbor (K-NN) algorithm is a method that uses a supervised algorithm where the results of the new test sample are classified based on the majority of the categories in K-Nearest Neighbor [8]. In pattern recognition, the K-Nearest Neighbor algorithm is a method for classifying objects based on the closest training set in the feature space [9]. The purpose of this algorithm is to classify new objects based on characteristics and training data. The K-Nearest Neighbor algorithm works based on the shortest distance from the test data to the training data to determine its K-Nearest Neighbor value [10]. K-Nearest Neighbor has several advantages, namely that it is resilient to noisy

data templates and is effective when the amount of template data is large. Whereas the weakness of K-Nearest Neighbor is the need to determine the value of the k parameter (the number of closest neighbors), distance-based learning is not clear about what type of distance should be used and which attributes should be used to get the best results, and the computation costs are quite high because It is necessary to calculate the distance from each query instance in the entire training sample [11]. Here are the steps for calculating the K-Nearest Neighbor [12]: 1) Determine the value of k, 2) Calculating the distance of each object to the training data provided using the distance measurement method, 3) Sort the results of the calculation of the similarity value from smallest to largest.

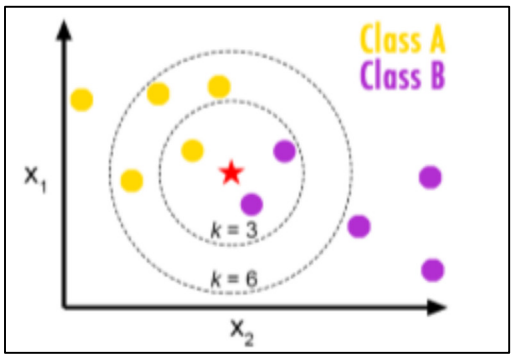

Figure 2. Illustration of the Use of the K Value in the KNN Method

3. Result and Discussion

In the main display in Figure 3 is used to direct the user to select the source of the photo. After that the user needs to crop the image. After the image has been cropped, the user needs to wait until the character recognition process is complete.

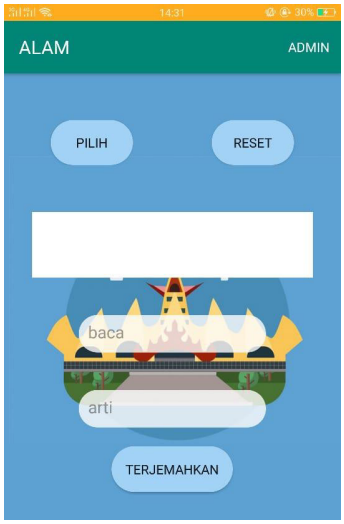

Figure 3. Main Display of Lampung Script Application

After the character recognition process is complete, the image that the user selects will be displayed on the imageView and the results of the character recognition will be displayed on the read edit Text. Users can see the results of the translation of the characters into Indonesian by pressing the 'Translate' button. The results of the script translation process will be displayed in the meaning editText. The display of the results of the introduction and translation can be seen in Figure 4.

Figure 4. The Results of the Introduction of the Lampung Script

Usecase Design Introduction to Lampung Script - In this application there are two actors, namely admin and user. Admin has rights so that they can add new characters and new dictionaries after logging in. Before the user can see how to read characters, the admin needs to add the characters. Meanwhile, users can use the application to read characters through photos without logging in. To read characters, users need to input images that can be done through the camera or smartphone gallery. After the user inputs the image, the system will display how to read the characters and the user can translate them into Indonesian. Storage of Character Features - In the application of the introduction of the Lampung script, there are two processes, namely the process of storing script features and storing the Lampung language dictionary, for the first stage, namely the grayscale process, binaryization to separate characters from the background. The segmentation process is to find the left, right, top, and bottom borders of the characters, followed by feature extraction using the Pixel Population Matrix. After the feature is obtained, it will be added to the database. The Process of Storing the Lampung Language Dictionary - To increase the number of languages or script Lampung, it is necessary to add a dictionary that makes it easier to translate Lampung script into Indonesian.

Manhattan Distance Process - The process of calculating the distance using Manhattan Distance is done by comparing the distance between the features in the test data or the features of the image being tested with the features contained in the database. The distance variable is used to accommodate the distance that will be manipulated for each iteration. In this process, it will be repeated 26 times, according to the number of features of the previously processed image. In the looping process, it will look for the distance by reversing the result of the difference between the image features and the features in the database. The distance variable will hold the sum of all distances in the image. K-Neareast Neighbor - The K-Nearest Neighbor process will use data from the calculated distance calculations in the previous process. In this process, it will begin by sorting the characters based on the smallest distance, then the system will take the 3 characters with the closest distance. After that, the name of the character will be taken for its unique value (no duplication) and stored in the uniqueValues array variable. Next, an iteration process will be performed for each uniqueValue with the condition i <uniqueValue.length () to find the number of times uniqueValue appears. The system will compare whether uniqueValue is the same as each character stored in the character array variable. If uniqueValue is equal to character, then the ith count variable will be incremented. After getting the number of times uniqueValue appears on the script, the system will look for the uniqueValue that appears the most. If uniqueValue is 1, the system will return the closest script name. If unique Value appears more than 1, then the system will return the name of the characters that appear the most.

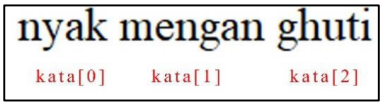

Figure 5. Separation of Words

The Process of Translating Characters - The process of translating how to read characters into Indonesian, the application will require input on how to read characters obtained from the process of recognizing characters. This process is carried out by separating the words in the script reading with the delimiter is a space. The words that have been obtained will be accommodated in the array variable. Furthermore, it will be repeated according to the number of words contained in the array variable, starting from the first word, which is 'nyak'. In the loop, another loop will be performed to find the word in the dictionary data. If the same word is found, the Indonesian word will be taken and displayed in the output. Furthermore, it will be repeated according to the number of words contained in the array variable, starting from the first word, which is 'nyak'. In the loop, another loop will be performed to find the word in the dictionary data. If the same word is found, the Indonesian word will be taken and displayed in the output. Testing The Application of Lampung Script Introduction - Testing the success of the method in this study was carried out by calculating the number of recognized characters. The characters to be tested will be compared with the 160 training data contained in the database. The test was carried out in several conditions using 674 Lampung characters.

4. Conclusion

Based on testing the Lampung script recognition application, namely Pixel Population Matrix feature extraction method, distance measurement using Manhattan Distance and K-Nearest Neighbor classification are able to recognize Lampung script images. The percentage of success from the introduction of Lampung script in Print Text Screenshot with Random Font is 98.433%, Handwritten Photo Image is 78.579%, Print Handwritten Screenshot Image with a Combination of Several Words is 95.351% and Photo Image with Smaller Size and Slope Random is 93.595%. Factors that affect application failure in reading Lampung script are the slope that is too extreme, the length of the image being tested and the difference in writing style for each script, especially in handwritten images.

References

[1] "Badan Pusat Statistik". [Online]. Available: https://www.bps.go.id
[2] S. Hartanto, A. Sugiharto, and S. N. Endah, "Optical Character Recognition Menggunakan Algoritma Template Matching Correlation", J. Masy. Inform., vol. 5, no. 9, 2015.
[3] M. ImrulJubair and P. Banik, "A Simplified Method for Handwritten Character Recognition from Document Image", Int. J. Comput. Appl., vol. 51, no. 14, pp. 50–54, 2012.
[4] I. Mapanga and P. Kadebu, "Database Management Systems : A NoSQL Analysis", Int. J. Mod. Commun. Technol. Res., vol. 1, no. 7, pp. 12–18, 2013.
[5] A. B. M. Moniruzzaman and S. A. Hossain, "NoSQL Database: New Era of Databases for Big data Analytics - Classification, Characteristics and Comparison", Int. J. Database Theory Appl., vol. 6, no. 4, pp. 1–8, 2013.
[6] N. Ameya, P. Anil, and P. Dikshay, "Type of NOSQL databases and its comparison with relational databases", Int. J. Appl. Inf. Syst., vol. 5, no. 4, pp. 16–19, 2013.
[7] J. F. Andry, H. Tannady and F. E. Gunawan, "Purchase Order Information System Using Feature Driven Development Methodology", International Journal of Advanced Trends in Computer Science and Engineering, vol. 9, no. 2, pp. 1107-1112, 2020.
[8] H. Tannady, J. F. Andry, F. E. Gunawan and J. Mayseleste, "Enterprise Architecture Artifacts Enablers for IT Strategy and Business Alignment in Forwarding Services", International Journal of Advanced Trends in Computer Science and Engineering, vol. 9, no. 2, pp. 1465-1472, 2020.
[9] Resdiansyah, J. Darmawan, A. H. Wijaya, L. Hakim and H. Tannady, "Comparing Freeman Chain Code 4 Adjacency Algorithm and LZMA Algorithm in Binary Image Compression", Annual Conference on Science and Technology Research (ACOSTER) 2020, Journal of Physics: Conference Series, 2021.
[10] N. Chatterjee, S. Chakraborty, A. Decosta, and D. A. Nath, "Real-time Communication Application on Android Platform", Int. J. Adv. Res. Comput. Sci. Manag. Stud., vol. 6, no. 4, pp. 74–79, 2018.

[11] F. E. Gunawan, J. F. Andry, H. Tannady and R. Meylovsky, "Designing Enterprise Architecture Using TOGAF Framework in Meteorological, Climatological, and Geophysical Agency", Journal of Theoretical and Applied Information Technology, vol. 97, no. 20, pp. 2376-2385, 2019.
[12] J. F. Andry, H. Tannady and F. Nurprihatin, "*Eliciting Requirements of Order Fulfillment in A Company"*, The 2nd International Conference on Engineering and Applied Sciences 2019 (InCEAS 2019), IOP Conference Series: Materials Science and Engineering, 2019.