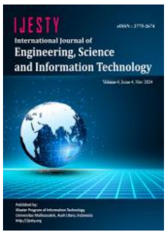
International Journal of Engineering, Science and InformationTechnology
Volume 4, No. 4 (2024) pp. 257-264
ISSN 2775-2674 (online)
Website: http://ijesty.org/index.php/ijesty
DOI: https://doi.org/10.52088/ijesty.v4i4.668
Research Paper, Short Communication, Review, Technical Paper



Pattern Recognition of Chinese Characters Using the Sokal Sneath Four Method

Rasna[1*], Andrian Sah[1], M. Ali Nur Hidayat[2], Jusmawati[1], Mursalim Tonggiroh[1]

[1]Department of Information System, Universitas Yapis Papua, Indonesia
[2]Department of Informatics, Universitas Yapis Papua, Indonesia

**Corresponding author Email: razna.anwar@gmail.com*

The manuscript was received on 20 April 2024, revised on 28 August 2024, and accepted on 18 November 2024, date of publication 5 December 2024

Abstract

Pattern recognition is a discipline that aims to classify or describe objects based on their characteristics, quantitative measurements, or critical properties. Where a pattern is defined as an entity that is initially undefined, it can be identified and named through quantitative analysis. Pattern recognition can be applied to various fields, such as handwriting recognition, face recognition, eye recognition, skin, and image processing. One example of the application of pattern recognition is character recognition in letters used in learning. In this research, the digital image used as input comes from a two-dimensional image obtained through a digital camera. The digital image describes the light intensity in light and dark areas in the form of pixels and provides information about the object's color. To support the process of recognizing alphabet letters, which in this case are specifically Chinese alphabet letters, it will be assisted by using the Sokal Sneath Four Method. This significant mathematical approach helps create a compatible and accurate system for recognizing letter patterns through intensive data training. This method involves a series of steps, including data preprocessing, feature extraction, and classification, to train the system to recognize Chinese characters. The more training given to the system, the higher its accuracy in recognizing letter patterns, especially Chinese alphabet letters. The test results show that this Chinese alphabet letter pattern recognition system has a success rate of 65%, with a failure rate of 35%. Nevertheless, these results show room for further improvement in the algorithms used and the addition of training data to improve system performance and accuracy.

Keywords: *Mandarin Chinese Characters, Pattern Recognition, Sokal Sneath, Digital Image, Characters Recognition.*

1. Introduction

Pattern recognition is the science of classifying or describing something based on quantitative measurements of an object's main features or properties [1]. The pattern is a defined entity that can be identified and named [2]. Pattern recognition can be through handwriting, eyes, face and skin, and image management (Paint). An example of the application of pattern recognition is recognizing characters in letters as learning [3].
This study applies pattern recognition to identify patterns of one type of Chinese letter, Mandarin. Introducing Chinese characters in Chinese language learning, especially in Mandarin letters, is increasingly popular today [4]. Simple use and a high level of letter recognition can increase user interest in learning Chinese. Chinese characters are pretty complex characters when compared to ordinary Roman characters. Primarily, when written by hand, this will result in various forms of writing by each person [5]. A feature extraction process can be used first to recognize patterns from letters. This process is carried out to obtain unique characteristics or features from data. In an image, these features can be pixels in a matrix formed from a digital image. This feature extraction process is implemented in the pre-processing process performed on an image. It is essential to increase the presentation of the successful matching of an object. Among them is resizing the image so that the pixel size of the compared image is similar to the thresholding process used to homogenize

the pixel value of the image and remove existing noise [6]. After the feature extraction process has been carried out, the process of recognizing Mandarin letters using the pattern recognition method is then carried out.

The pattern recognition method used is Sokal Sneath 4 as an application of more straightforward and more complex techniques [7]. In general, the Chinese letter pattern recognition system (Mandarin) using Sokal Sneath 4 consists of several stages, namely image acquisition, greyscale process, segmentation/location using sobel operator edge detection, identification using the Sokal Sneath 4 method, and then output in the form of identification of the Chinese Mandarin letter. Based on these problems, it is necessary to develop a system application that can identify patterns or image managers with this ability level developed using a machine (computer). With the recognition of patterns in the image, we can implement the computer's intelligent capabilities to approach the process of the human brain.

2. Literature Review

2.1. Digital Image Processing

Digital image processing is the digital manipulation and interpretation of images with the help of computers [8]. Image processing aims to:

1. Improving image quality from radiometric and geometric aspects. Radiometric aspects consist of contrast enhancement, image restoration, and color transformation, while geometric aspects consist of rotation, scale, translation, and geometric transformation) [9].
2. Extract information, object description, or object recognition in the image.
3. Selecting optimal feature images for analysis purposes.
4. Performing data compression or reduction for data storage, transmission, and processing time.

The stages of digital image processing are as follows:

1. Image Acquisition
   Image acquisition is the initial stage of obtaining a digital image. Image acquisition aims to determine the required data and select a digital image recording method. This stage involves capturing the object, tools, and imaging preparation [10]. Imaging transforms visible images into digital images (photographs, drawings, paintings, sculptures, landscapes, etc.). Some tools that can be used for imaging are:
   a. Video camera
   b. Digital camera
   c. Conventional camera and analog to digital converter
   d. Scanner
   e. Photo x-ray / infrared ray
   f. Pre-Processing and Segmentation
   This stage is necessary to ensure the smooth running of the following process. Important things that are done at this level include:
   a. Image quality enhancement (contrast, brightness, etc.)
   b. Image Restoration
   c. Transformation (image transformation)
   d. Determining the part of the image to be observed
   Meanwhile, segmentation is a stage that aims to partition the image into essential parts that contain important information [11]. For example, separating objects and backgrounds.
2. Representation and Description
   In this case, representation represents an area as a list of coordinate points in a closed curve, describing its area or perimeter [12]. After defining an area, the following process uses feature selection and extraction to describe the image. Feature selection aims to select quantitative information from existing features that can distinguish object classes well. In contrast, feature extraction seeks to measure the quantitative magnitude of each pixel's features, such as the mean, standard deviation, coefficient of variation, Signal Noise ratio (SNR), and others.
3. Recognition and Interpretation
   The recognition stage aims to label an object whose information the descriptor provides, while the interpretation stage seeks to give meaning to the group of recognized objects.

2.2. Grayscale Image

The initial process that is mainly done in image processing is to convert a color image into a grayscale image; this aims to simplify the color image, which consists of three layers of matrices, namely the R-Layer, G-Layer, and B-Layer so that the following processes are still considered the three-layer process [13]. If each calculation process is carried out on three layers, the same calculation is carried out three times, so the above concept is changed from three-layer matrices to a layer grayscale matrix, which is an image whose gray level is 0-255. A color mode that displays images using 256 shades of gray. Each color is defined as a value between 0 and 255, where 0 is the darkest (black) and 255 is the lightest (white) [14]. To convert a full color (RGB) image into a grayscale image, commonly used methods include:

$$(R + G + B) / 3 \qquad \text{(1)}$$

where :
R: Red color element
G: Green color element
B: Blue color element

2.3. Edge Detection Using the Sobel Operator

The Sobel operator performs edge detection. The advantage of this sobel operator is the ability to reduce noise before calculating edge detection. The process used by the Sobel operator is the process of a convolution that has been determined on the detected image [15]. In the sobel operator, a 3x3 convolution matrix is used, and the arrangement of the pixels around the pixel (x,y) is as follows:

$$\begin{bmatrix} a_0 & a_1 & a_2 \\ a_7 & (x,y) & a_3 \\ a_6 & a_5 & a_4 \end{bmatrix} \qquad \text{………..(2)}$$

The sobel operator is the magnitude of the gradient calculated with :

$$M = \sqrt{s_x^2 + s_y^2} \qquad \text{……....(3)}$$

Where in this case, the partial derivative is calculated by :

$S_x = (a_2 + ca_3 + a_4) – (a_0 + ca_7 + a_6)$ ……………………………………………………………………………………………....(4)

$S_y = (a_0 + ca_1 + a_{22}) – (a_6 + ca_5 + a_4)$ ……………………………………………………………………………………………...(5)

With constant c =2. In mask form, sx and sy can be expressed as follows:

$$S_x \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \text{ dan } S_y \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \qquad \text{…………………………………………………………………………(6)}$$

2.4. Pattern Recognition

Pattern Recognition is a scientific discipline that aims to classify objects or classes [16]. A pattern is a defined entity that can be identified through its features. These features are used to distinguish one pattern from another. A good feature is a feature that has a high distinguishing power so that the grouping of patterns based on their characteristics can be done with high accuracy. An example is as follows.

Table 1. High-accuracy feature

Pattern	Characteristic
Letters	Height, thickness, corner points, line bends
Sound	Amplitude, frequency, pitch, intonation
Signature	Length, complexity, pressure, line curvature
Fingerprints	Curvature, number of lines

The characteristics of a pattern are obtained from the measurement of the test object [17]. Especially for patterns contained in images, the characteristics that can be obtained are derived from information:

1. Spatial: pixel intensity and histogram
2. Edge: direction and strength/pressure, corner points
3. Contours: lines, ellipses, and circles
4. Area/shape: perimeter, area, and center of mass
5. Imagery: thickness, line curvature

2.5. Sokal Sneath 4

This similarity measure is based on the conditional probability of one value of a while the value of the other object is expressed as a predictor calculated from its average value [18]. The formula is as follows:

$$S = \frac{a\ (a+b) + a\ (a+c) + d\ (b+d) + d\ (c+d)}{4} \qquad \text{…………………………………………………………………………………(7)}$$

Where:
a = the frequency of both individuals is 1
b, c = the frequency of one object is 1, and the other object is 0
d = the frequency of both individuals is 0

There are two clustering techniques in cluster analysis: hierarchical and non-hierarchical. The hierarchical technique is used if the number of clusters formed is not known in advance, while the non-hierarchical technique is used if the number of clusters formed has been determined from the beginning [19].

2.6. Mandarin Language

Mandarin Chinese is the official spoken language of China. Mandarin is one of eight significant and minor dialects comprising the Chinese language. Dialects vary regionally in pronouncing the same syllable or speech sound. Chinese writing is based on symbols, called ideographs or characters, rather than sounds. Different dialects pronounce These characters differently, but the meaning is the same [20].

Table 2. Chinese Characters From A to Z

A	诶	Ēi	**N**	艾娜	ài nà
B	比	Bǐ	**O**	哦	Ó
C	西	Xī	**P**	屁	Pì
D	迪	Dí	**Q**	吉吾	jí wú
E	伊	Yī	**R**	艾儿	ài ér
F	艾弗	ài fú	**S**	艾丝	ài sī
G	吉	Jí	**T**	提	Tí
H	艾尺	ài chǐ	**U**	伊吾	yī wú
I	艾	Ài	**V**	维	Wéi
J	杰	Jié	**W**	贝尔维	bèi ěr wéi
K	**开**	Kāi	**X**	艾克斯	Sī
L	艾勒	ài lè	**Y**	吾艾	wú ài
M	艾马	ài mǎ	**Z**	贼德	zéi dé

3. Methods

Several general steps will be carried out in this research, as follows:

1. Create Mandarin writing using paint
   The author makes samples of the Chinese alphabet as a reference in training and testing this application.
2. Literature study and data collection
   Literature study is done by reading, understanding, and looking for references to theories about the Chinese alphabet in various articles, journals, internet media, and books in the library, as well as references from student final assignments related to the Sokal Sneath 4 method.

In carrying out the process to complete this system, the author needs input in the form of images (images), which go through the stages of writing Chinese letter patterns using hands carried out by several students then each image data (image) that has been written will be scanned with a scanner, to train and also test this system later.

The problems faced in building this application are not only due to the large number and varied sentence patterns but also include the issues of checking sentences carried out on processes that are more than one sentence, so techniques are needed to recognize and separate between one sentence and another sentence, by doing some training on all Chinese Letter Patterns, then in the testing process the system will be more accurate in recognizing each Chinese Letter pattern itself.

The output that will be generated after we do the testing process later is after we input a new Chinese letter pattern, and after we test it, the system will automatically recognize the letter pattern that we input, with the output in the form of recognized letters, and how to read each Chinese letter pattern that is recognized.

The system schema illustrates some information about Chinese letter pattern files related to the application to be built. From some of the sampling that has been done, everything will be described in the form of a scheme related to the system linked to one another.

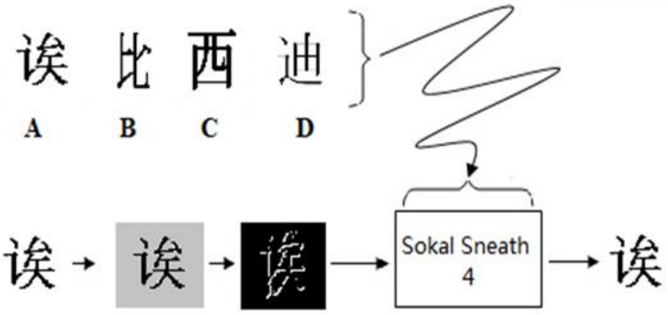


Fig 1. Overall System Workflow

Figure 1 above describes the workings of the overall system scheme, Chinese letter patterns from A to Z. The first Chinese letter pattern is the original image, then the Chinese letter pattern will be grayscaled, and then the edge will be detected. We will test the program one by one in the letter pattern program so that Chinese letters can be recognized.

A flowchart is a section with certain symbols describing the sequence of processes in detail and the relationship between a process (instruction) and other processes in a program. Figure 2 presents the overall work/process of the program after the original image is inputted. The original image is a grayscale process, which then the original color image will turn into a grayscale image, after which the grayscale image will be subjected to an edge detection process using the sobel operator to mark the part that becomes the detail of the image, which will then be continued with the calculation with the sneath four social method to get the value of the calculation which will later become a reference in recognition of the mandarin alphabet pattern. After the sokal sneath 4 value is found, the value and image (image) will be stored in the database. The following is a flow chart of this research:

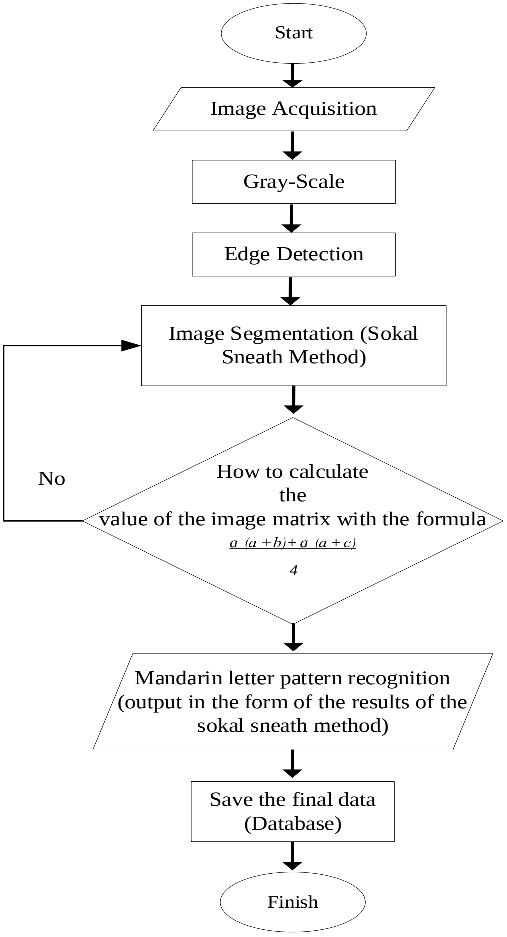

Start
Image Acquisition
Gray-Scale
Edge Detection
Image Segmentation (Sokal Sneath Method)
No
How to calculate the value of the image matrix with the formula
a (a+b)+a (a+c)
4
Mandarin letter pattern recognition (output in the form of the results of the sokal sneath method)
Save the final data (Database)
Finish

Fig 2. Flow Chart

4. Result and Discussions

To perform the Chinese letter pattern recognition process, designing an application to determine the pattern value of an image (image) by training several samples of Chinese letters is necessary. The author intended software using the Borland Delphi 7 programming language in this study. The things discussed include the selection of letter pattern training samples, system training, system testing, and system performance measurement.

System training is conducted to find and know the value of each input letter pattern recognition using the Sokal Sneath 4 method. Then, the letter samples and the value of the search results will be stored in the database, which aims to reference letter pattern recognition later in system testing.

This system test is carried out by inputting new data/samples into the system, which then the system will identify samples based on the data in the database that has been stored during training. Then, the system will bring up the Chinese alphabet from A to Z.

It has been explained through the previous scheme on the system built, and the following will also be described as an implementation of a scheme that shows a process where a line of work performance is as follows:

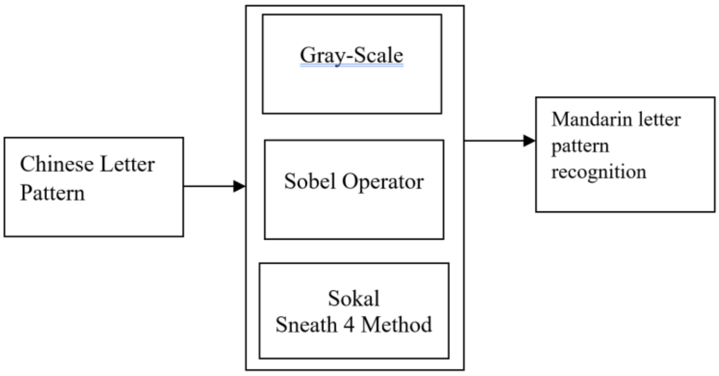
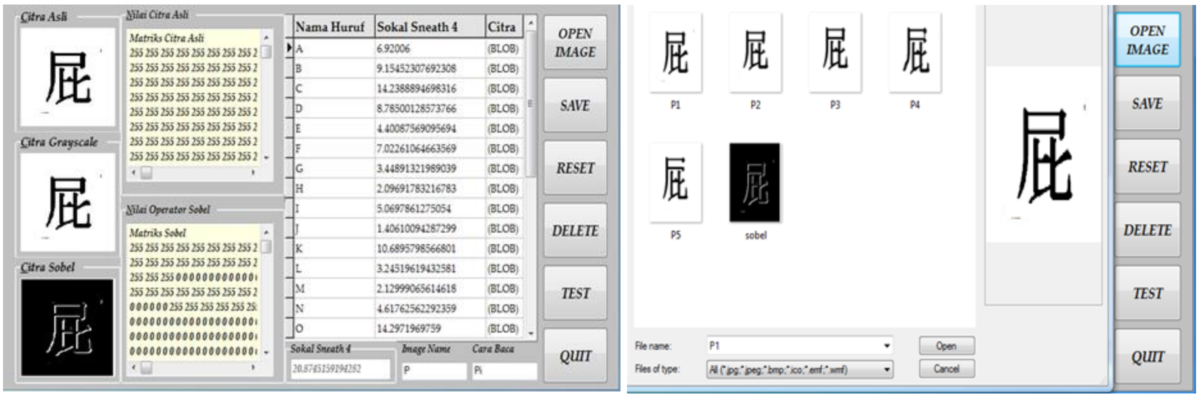


Fig 3. System Schematic Implementation

The training process in this system will be explained in the following image representation:



```
Matriks Citra Asli
255 255 255 255 255 255 255 255 2
255 255 255 255 255 255 255 255 2
255 255 255 255 255 255 255 255 2
255 255 255 255 255 255 255 255 2
255 255 255 255 255 255 255 255 2
255 255 255 255 255 255 255 255 2
255 255 255 255 255 255 255 255 2
255 255 255 255 255 255 255 255 2
```

```
Matriks Sobel
255 255 255 255 255 255 255 255 2
255 255 255 255 255 255 255 255 2
255 255 255 0 0 0 0 0 0 0 0 0 0 0 0 0
255 255 255 255 255 255 255 255 2
0 0 0 0 0 0 255 255 255 255 255 25
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

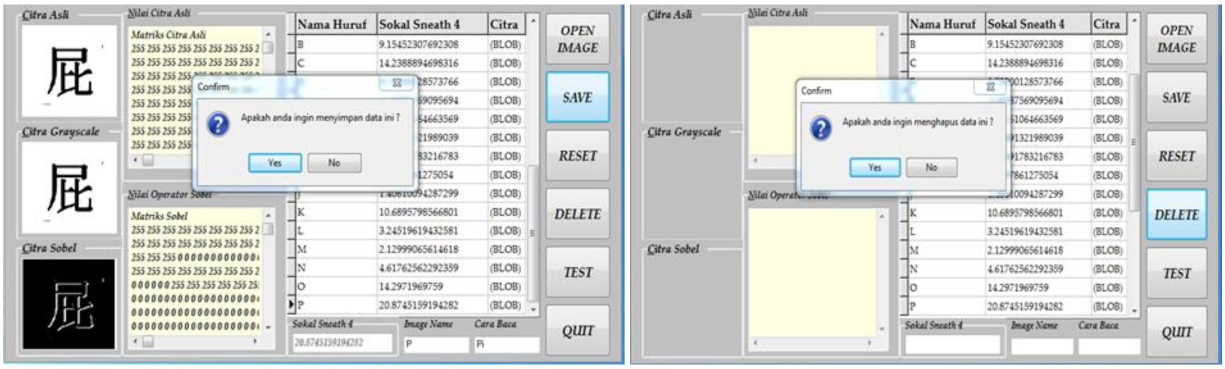
Nama Huruf	Sokal Sneath 4	Citra
A	6.92006	(BLOB)
B	9.15452307692308	(BLOB)
C	14.2388894698316	(BLOB)
D	8.78500128573766	(BLOB)
E	4.40087569095694	(BLOB)
F	7.02261064663569	(BLOB)
G	3.44891321989039	(BLOB)
H	2.09691783216783	(BLOB)
I	5.0697861275054	(BLOB)
J	1.40610094287299	(BLOB)
K	10.6895798566801	(BLOB)
L	3.24519619432581	(BLOB)
M	2.12999065614618	(BLOB)
N	4.61762562292359	(BLOB)
O	14.2971969759	(BLOB)

Sokal Sneath 4: 20.8745159194282
Image Name: P
Cara Baca: Pi

File name: P1
Files of type: All (*.jpg;*.jpeg;*.bmp;*.ico;*.emf;*.wmf)

Fig 4. Training Sample Chinese Letter Pattern Recognition

The picture above is a Training Form where each Chinese Letter pattern from A to Z will be trained so that the final value/result of searching each matrix using the Socal Sneath 4 Method is obtained, which will then be stored in the database to recognize the pattern of each letter pattern during training later. Then, the other picture explains the display form of the system application, which aims to open images in the form of pictures in .bitmap format, where the function of this form is the display of the Chinese alphabet training from A to Z.



Apakah anda ingin menyimpan data ini ?

Apakah anda ingin menghapus data ini ?

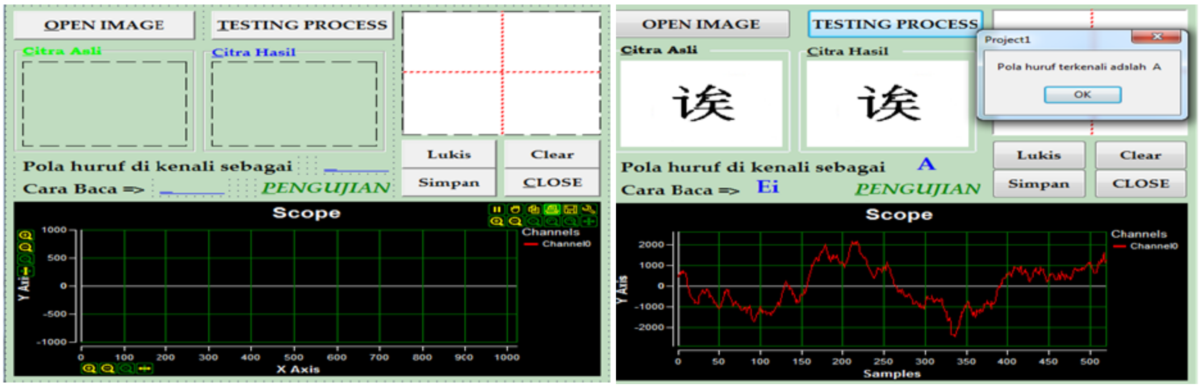
Nama Huruf	Sokal Sneath 4	Citra
B	9.15452307692308	(BLOB)
C	14.2388894698316	(BLOB)
K	10.6895798566801	(BLOB)
L	3.24519619432581	(BLOB)
M	2.12999065614618	(BLOB)
N	4.61762562292359	(BLOB)
O	14.2971969759	(BLOB)
P	20.8745159194282	(BLOB)

Sokal Sneath 4: 20.8745159194282
Image Name: P
Cara Baca: Pi

Fig 5. Save and Delete Data

The picture above shows how the system works in storing data that has been processed from the Original Grayscale Image Image to each letter matrix that has been calculated by the Socal Sneath 4 method, which is in the form of Numeric which is then stored in the database, where the data stored is essential data for the recognition of letter patterns during testing later. Then, another picture displays how the system deletes data stored in the database, where the deleted data is considered unnecessary or there is an error in storing the data.

The Chinese character pattern recognition system was tested after the sample training process. The training was conducted to determine the reference energy the system will identify as a characteristic of a Chinese character. Meanwhile, testing was carried out to compare the letters used as references with the position of the character lines of the letters input later, whether the energy can be recognized as a letter or not. The following shows the results of the Chinese character recognition system testing.



Pola huruf di kenali sebagai

Cara Baca =>

Pola huruf di kenali sebagai A

Cara Baca => Ei

Project1

Pola huruf terkenali adslah A

Fig 6. System Testing Form and Letter A Recognition as an Example

The image above shows the system display on the Testing Form, where the new Chinese letter pattern will be tested in this form. Then, in another image, the test was successful. The original image we input also successfully recognized the letter pattern recognition. After pressing the Testing Button, the system will automatically recognize the letter pattern and its output by reading the letter A sound in the form of the letter A.

5. Conclusion

This study shows that the Sokal Sneath 4 method can be applied to the letter image pattern recognition system, especially Chinese letters from A to Z. More frequent training improves the accuracy of the system in recognizing letter patterns. The test results show a letter recognition success rate of 65% and a failure rate of 35%. Although this method works quite well, the failure rate is still high, indicating the need for further optimization. Factors such as the complexity of Chinese letters or the limited training data may affect these results. The system needs to be improved through more intensive training and parameter adjustment to achieve higher accuracy.

References

[1] A. Lindberg, "Developing theory through integrating human and machine pattern recognition," *J. Assoc. Inf. Syst.*, vol. 21, no. 1, p. 7, 2020.

[2] Z. Nasar, S. W. Jaffry, and M. K. Malik, "Named entity recognition and relation extraction: State-of-the-art," *ACM Comput. Surv.*, vol. 54, no. 1, pp. 1–39, 2021.

[3] S. Bhushan, M. Alshehri, N. Agarwal, I. Keshta, J. Rajpurohit, and A. Abugabah, "A novel approach to face pattern analysis," *Electronics*, vol. 11, no. 3, p. 444, 2022.

[4] S. Ésik, "Teaching Chinese characters to second language learners," *Res. Teach. Chinese As a Foreign Lang.*, vol. 3, pp. 1–22, 2020.

[5] R. N. Morris, "Forensic handwriting identification: fundamental concepts and principles," 2020.

[6] Y. E. Almalki, T. A. Soomro, M. Irfan, S. K. Alduraibi, and A. Ali, "Impact of image enhancement module for analysis of mammogram images for diagnostics of breast cancer," *Sensors*, vol. 22, no. 5, p. 1868, 2022.

[7] G. H. Leazer, R. Montoya, and J. Furner, "Numerical Classification and Complexity: Developing a Classification of Classifications," in *Knowledge Organization at the Interface*, 2020, pp. 217–225.

[8] R. Thakur and R. Rohilla, "Recent advances in digital image manipulation detection techniques: A brief review," *Forensic Sci. Int.*, vol. 312, p. 110311, 2020.

[9] D. Laupheimer, M. H. S. Eddin, and N. Haala, "The importance of radiometric feature quality for semantic mesh segmentation," in *DGPF annual conference, Stuttgart, Germany. Publikationen der DGPF*, 2020, pp. 205–218.

[10] J. A. Richards, J. A. Richards, and others, *Remote sensing digital image analysis*, vol. 5. Springer, 2022.

[11] W. Zhang, J. Pang, K. Chen, and C. C. Loy, "K-net: Towards unified image segmentation," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 10326–10338, 2021.

[12] I. El Bouchairi, A. Elmoataz, and J. Fadili, "Nonlocal perimeters and curvature flows on graphs with applications in image processing and high-dimensional data classification," *SIAM J. Imaging Sci.*, vol. 16, no. 1, pp. 368–392, 2023.

[13] Z. N. Khudhair *et al.*, "Color to grayscale image conversion based on singular value decomposition," *Ieee Access*, vol. 11, pp. 54629–54638, 2023.

[14] N. Rani, S. R. Sharma, and V. Mishra, "Grayscale and colored image encryption model using a novel fused magic cube," *Nonlinear Dyn.*, vol. 108, no. 2, pp. 1773–1796, 2022.

[15] Y. Ma, H. Ma, and P. Chu, "Demonstration of quantum image edge extration enhancement through improved Sobel operator," *Ieee Access*, vol. 8, pp. 210277–210285, 2020.

[16] W. Ali, W. Tian, S. U. Din, D. Iradukunda, and A. A. Khan, "Classical and modern face recognition approaches: a complete review," *Multimed. Tools Appl.*, vol. 80, pp. 4825–4880, 2021.

[17] B. N. Boots and A. Getls, "Point pattern analysis," 2020.

[18] G. H. Leazer, R. Montoya, and J. Furner, "Numerical Classification and Complexity," *at the Interface*, p. 217.

[19] M. Siraj-Ud-Doulah, M. A. Hakim, and M. A. Hamid, "Performance Analysis of Hierarchical and Non-Hierarchical Clustering Techniques".

[20] M. S. Erbaugh, "The acquisition of Mandarin," in *The crosslinguistic study of language acquisition*, Psychology Press, 2022, pp. 373–455.