IMPLEMENTATION OF K-NEAREST NEIGHBOR AND PRINCIPAL COMPONENT ANALYSIS ALGORITHMS ON HANDWRITING RECOGNITION IN ESSAY ANSWERING

Yuyus Bahtiar Gustani[1], Galih Hermawan[2]

[1,2] Teknik Informatika – Universitas Komputer Indonesia
Jl. Dipatiukur 112-114 Bandung
Email : yuyus_rsbi2@yahoo.co.id[1], galih.hermawan@email.unikom.ac.id[2]

ABSTRACT

Optical Character Recognition (OCR) is a tool that can be used for translating characters digital image into text format. [1]. This research are using principal component analysis and k-nearest neighbor methods. The purpose of PCA method is too reduce big dimension from observed variables intrigue dimension that smaller from independent variables, it is used to describe smaller data [2]. K-NN is classification method that based on nearest neighbor class. The image that used for this research are training image and test image. Training image will be handwriting image that contains characters from a-z, A-Z and 0-9. Test image will be handwriting essay answers. Step by step of this research are preprocessing that include grayscaling, thresholding, dilation, segmented eruotion, and resizing. After that extraction process that using PCA method. For introduction step is by using extraction of feature from K-NN method result.. after getting the result from identification, next to do is assessing process essay by measuring similarity level with the answer sheet using rabin karp method. Based on the testing result that have been with five K values combination parameter. The accuracy of the method is 77,79 percent with K=5.

Keyword: Handwriting Introduction, Principal Component Analysis, Feature Extraction, K-Nearest Neighbors, Digital Image Processing

1. INTRODUCTION

Optical Character Recognition (OCR) is a tool that can be used for translating characters digital image into text format. [1]. OCR has been widely used in its application such as the introduction of license plates, handwriting Java letters, Japanese letters, and Balinese script, and many other applications of OCR use. In the image processing, to get good results in the introduction, then used two methods namely methods for feature extraction and methods for recognition.

The characteristic extraction method in this study will use Principal Components Analysis (PCA). The purpose of PCA method is too reduce big dimension from observed variables intrigue dimension that smaller from independent variables, it is used to describe smaller data [2]. PCA can be used as a feature extraction on handwriting recognition [3]. The use of PCA in recognizing patterns depends on the dimensions and number of zones [3]. While the learning method used in this research is K-Nearest Neighbor (K-NN) method. K-NN is a classification method based on the class of the nearest neighbor. K-NN can be used for handwriting of arabic numbers (indian) [4]. The best accuracy value of using K-NN method is 86% with k = 1 [4].

Handwriting recognition is one form of computer ability to recognize patterns. Research in the field of handwriting recognition has evolved over a considerable period of time. The study was conducted because more and more use of handwriting models in everyday life, such as identification of important documents, evidence of validation in the banking world, and other things involving handwriting in it.

Based on research on handwriting recognition that has been done found obstacles in the process of recognition, this is because some letters of the alphabet have similarities form with other letters [5], as well as on the application of new handwriting recognition implemented in the case in general. In this research, handwriting recognition will be implemented in the case of essay assessment. Research on the assessment of essays has also been done but is limited to digital text only [6]. So in this research the researcher will make prototype of K-Nearest Neighbor algorithm implementation and Principal Components Analysis on handwriting recognition in essay assessment. Then for accuracy, will be calculated by counting the number of handwriting characters that can be recognized.

2. RESEARCH CONTENTS

The problem in this research is how to recognize handwriting using K-NN and PCA method. In previous research handwriting recognition has been used to recognize letters of the alphabet, but there are obstacles in the process of recognition, this is because some letters of the alphabet have similarity in shape with other letters as well as on the application of new handwriting handwriting is implemented in the case

in general. In this study the PCA method will be used to generate the characteristic of handwriting images which will then be classified using K-NN. The results of the writing introduction will be used for the essay assessment process. Accuracy of handwriting recognition obtained will affect the results of the essay assessment, so please note the accuracy of the introduction.

The system to be built includes handwriting recognition process divided into two stages, where the training and testing phase. At the training stage there is a preprocessing process, and feature extraction with PCA. Furthermore, at the testing stage there is preprocessing, feature extraction with PCA, K-NN classification, and Essay Evaluation. While the input data used is divided into two data ie training data in the form of images with extension. Jpg which contain handwriting collection of characters A-Z, a-z, and 0-9. Test input data is handwritten image with .jpg extension containing essay answers. Here is the general architecture of the system to be built.

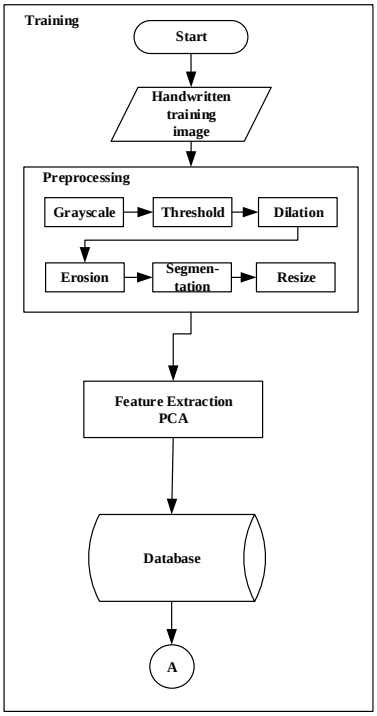


Figure 1 Training Stage

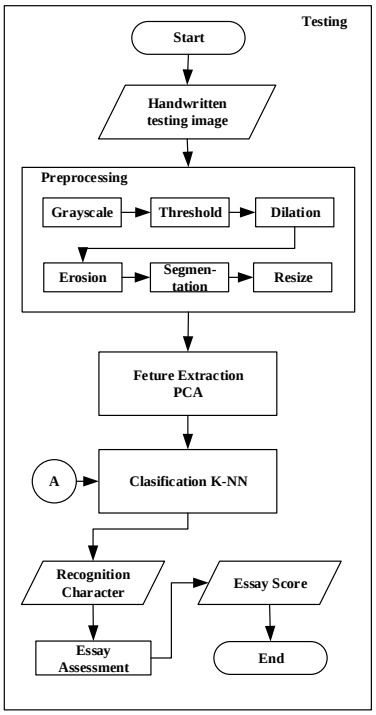


Figure 2 Testing Stage

2.1. Input Data

Data input used in this research, that is handwritten image with image format .jpg. Imagery can be obtained through scan. Handwritten images used are written using a pen in HVS paper with a white background. In addition, the color of the image is also not too dark for the application to be more leverage character recognition of each handwriting.

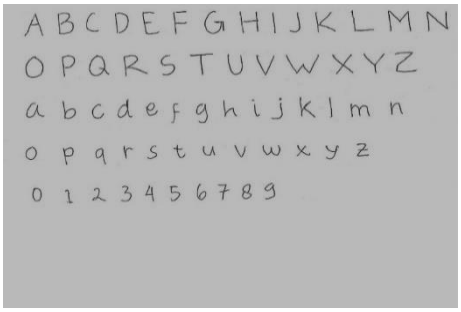


Figure 3 Input Data

2.2. Prepocessing

Digital image processing means processing two-dimensional images via a digital computer. According to Efford (2000), image processing is a generic term for various techniques that exist to manipulate and modify images in various ways [7]. In the process of preprocessing the data input image will be processed in several stages of grayscale, thresholding, dilation, segmentation and resize. This process is done so that the data input image can meet

certain criteria before the process of feature extraction and classification.

A. Grayscale

The grayscale image handles the gradations of black and white, which of course produce a gray color effect. Color images expressed with intensity. In this case, the intensity ranges from 0 to 255. The value 0 is black and the value 255 is white [8]. Changing the color format to grayscale serves to shrink the color range from 0 to 255. This process will make it easier when wanting to thresholding the image into black and white image. Here is the formula used in the grayscale process.

$$I = (0.2989*R)+(0.587*G)+(0.1141*B) \quad (1)$$

Explanation:

R = The red value component (Red) of a pixel point
G = The green value component (Green) from a pixel point
B = The blue value (Blue) component of a pixel point

Therefore, these colors are given a higher weight to ensure that the relative intensity balance in the resulting grayscale image is similar to that of the RGB color image [9].

B. Thresholding

Thresholding method used in this research is using Sauvola Threshold method. Sauvola threshold is a threshold method capable of providing enhancement by calculating the threshold using the dynamic value range of the standard devaluation value of grayscale image [10]. Sauvola is a category of local threshold method, where the threshold value of each pixel depends on the number of neighbors used. Here is the formula used in the thresholding process.

$$T(x,y) = m(x,y) * (1 + k * (\frac{s(x,y)}{R} - 1)) \quad (2)$$

$$m(x,y) = \frac{\sum_{i=min}^{i=max} \sum_{j=min}^{j=max} img(i,j)}{i*j} \quad (3)$$

$$s(x,y) = \sqrt{\frac{\sum_{i=min}^{i=max} \sum_{j=min}^{j=max} (img(i,j)-m(x,y))^2}{(i*j)-1}} \quad (4)$$

Explanation:

R = Maximum value of standard deviation (128 for grayscale image)
k = Kernel with a value between 0.2 - 0.5
m = Function that produces the average value of a number of image pixels.
s = The function that generates the standard deviation value of some image pixels
T = Function that produces threshold value (threshold)
x = Coordinate value of image width
y = High image coordinate value
i = Value of neighboring pixel size
j = Value of neighboring pixel size

After the threshold value T (x, y) has been obtained, then enter into the equation below.

$$f(x,y) = \begin{cases} 0, & \text{img}(x,y) < \text{T}(x,y) \\ 255, & \text{img}(x,y) \geq \text{T}(x,y) \end{cases} \quad (5)$$

Explanation:

f = Function that results in 0 or 255 values
img = Grayscale image value

C. Dilation

The dilation process works to fix the pixels on broken images by comparing each input image pixel with the SE center value.

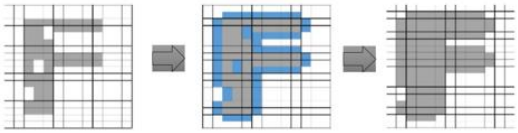


Figure 4 Dilation Example

D. Erosion

The process of erosion is the reverse of the dilation process. If in the dilation process produces a wider object then in the process of erosion will produce a narrowed object (smaller). The erosion operation can be expressed as follows [8].

$$D(A,B)=A \ominus B= \{x:B_x \cap A \neq \emptyset\} \quad (6)$$

Explanation:

A = Input image
B = Structuring Elements (SE)
E = Functions that produce erosion images

E. Segmentation

Segmentation is a technique for dividing an image into regions where each region has similar attributes [7].



Figure 5 Segmentation Example

F. Resize

Resizing the image means changing the size of the digital image in pixels. Sometimes its size changes to smaller than its original size and exists times otherwise. Image Resizing process can affect the image size larger or smaller and also the image quality becomes better or worse. Resize/Normalization with size 16x16 [11].



Figure 6 Resize Example

G. Principal Component Analysis (PCA) Algorithm

PCA is a powerful technique for extracting the structure of a set of data with considerable dimensions. Problem in PCA is finding eigenvalue and eigenvector. PCA is an orthogonal (perpendicular) transformation of the coordinate system in which we describe the data. Our new coordinates represent data called principal component or PC. These coordinates are selected where the variance of the data reaches the maximum. Often only with some principal component is sufficient to explain the structure of the original data. If the data in the original dimension is difficult to represent via graphics, then with two PCs or one PC can be drawn through graphs [12]. Here is the methodology of PCA [13]:

Each 2-D character image is depicted as a 1-D vector by combining each column (or row) into long thin vectors. Thus, the resulting vector must be present as:

$$\bar{x}^i = [x_1^i \dots x_N^i]^T \quad (7)$$

Explanation:
$\bar{x}$ = A set of 1D character imagery
T = Transpose
X = 1D character image
N = Lots of characters

The average image is a column vector where each entry is the average of all the corresponding training image pixels. The previous theories can be expressed in the following phrase:

$$\bar{x}^i = x^i - m \quad (8)$$

Explanation:
$\bar{\bar{x}}^i$ = 1D character image
m = Average column matrix of characters

where

$$m = \frac{1}{p}\sum_{i=1}^{p} x^i \quad (9)$$

Explanation:
m = Average column matrix of character imagery
P = Lots of data
X = The 1D i character image
I = Lots of characters

Center data:

$$\bar{X} = [\bar{x}^1|\bar{x}^2| \dots |\bar{x}^p|] \quad (10)$$

Explanation:
$\bar{x}$ = A set of 1D character imagery
p = Lots of data
x = 1D character image

Covariance matrix

$$\Omega = \bar{X}\bar{X}^T \quad (11)$$

Explanation:
Ω = Covariance matrix
$\bar{X}$ = A set of 1D character imagery
T = Transpose

$$det(C - \lambda) = 0 \quad (12)$$

Explanation:
C = Covariance matrix
λ = A set of 1D character imagery

Sort order of eigenvectors:

$$V = [V_1|V_2| \dots |V_p|] \quad (13)$$

Explanation:
V = Eigenvector Matrix
p = Lots of data

Image projection:

$$\tilde{x}^i = V^T\bar{x}^i \quad (14)$$

Explanation:
$\tilde{x}$ = Training projection matrix
V = Eigenvector matrix
T = Transpose

$\bar{x}$ = 1D Image Image

Identify new images:

$$\bar{y}^i = y^i - m \quad (15)$$

Explanation:
$\tilde{y}$ = New projection matrix
$\bar{y}$ = The new 1D image image data center
V = Eigenvector matrix
T = Transpose

H. K-Nearest Neighbor (K-NN) Algorithm

This technique belongs to the nonparametric classification group. This technique is very simple and easy to implement. Similar to clustering techniques, grouping of new data based on the new data distance to some of the nearest neighbors. In this case the number of nearest data / neighbors is determined by the user expressed by k. Suppose set k = 5, then every data testing is calculated distance to training data and select 5 training data closest to data testing. Once you know these 5 data, check the output or label respectively. Then determine which output is the most frequent [12].

To calculate the distance between two points x and y Euclidean distance can be used as follows:

$$d(x,y) = \| x - y \|^2 = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (16)$$

Explanation:
$d(x,y)$ = The distance between two points
x_i = Train data of i
y_i = Test data to i
i = Numbers count i = (1,2,3, … , n)

I. Algoritma Rabin Karp

Rabin-karp algorithm is a string matching algorithm that uses hash functions as a comparison between the search string (m) and substring in text (n). The Rabin-karp algorithm is based on the fact that if two strings are equal then the hash value value must be the same. But there are two problems that arise from this, the first problem is that there are so many different strings, this problem can be solved by assigning multiple strings with the same hash value. The second problem is not necessarily a string that has the same hash value matching to overcome it for each string that is assigned to do the Brute-Force string matching. The key to the efficient Rabin-Karp algorithm is in its hash value selection. One effective way is to treat each substring as a number with a particular base [14]. The steps performed in the evaluation of essay answers are the process of tokenizing and case folding, filtering, Rabin-Karp algorithm (K-gram, hashing, similarity).

- **Tokenizing and Case Folding**

In the tokenizing and case folding stage the process of splitting the text into words, then change all the words that have been dialed into lowercase. Here is an example of Tokenizing and Case Folding.

Tabel 1 Tokenizing and Case Folding

Text	Memori adalah sistem yang sangat berstruktur, yang menyebabkan organisme sanggup merekam fakta tentang dunia dan menggunakan pengetahuannya untuk membimbing perilakunya
Result of tokenizing and case folding	memori | adalah | sistem | yang | sangat | berstruktur | yang | menyebabkan | organisme | sanggup | merekam | fakta | tentang | dunia | dan | menggunakan | pengetahuannya | untuk | membimbing | perilakunya |

- **Filtering**

After going through the process of tokenizing and case folding, then done the process of filtering to eliminate words that belong to stoplist / stopword. The words 'adalah', 'yang', 'sangat', 'tentang', 'dan', 'menggunakan', and 'untuk', are omitted so the result is as follows.

Tabel 2 Filtering

Text	memori | adalah | sistem | yang | sangat | berstruktur | yang | menyebabkan | organisme | sanggup | merekam | fakta | tentang | dunia | dan | menggunakan | pengetahuannya | untuk | membimbing | perilakunya |

The result of filtering	memori | sistem | berstruktur | menyebabkan organisme | sanggup | merekam | fakta | dunia | pengetahuannya | membimbing | perilakunya

- **K-gram**

In this process, text that has been through tokenizing process and case folding and filtering is done splitting as much as K letter continuously. In this study using K = 4 then the results of K-gramnya as follows:

Tabel 3 K-gram

Text	memori | sistem | berstruktur | menyebabkan organisme | sanggup | merekam | fakta | dunia | pengetahuannya | membimbing | perilakunya
K-gram results	memo | emor | mori | oris | risi | isis | sist | iste | stem | temb | embe | mber | bers | erst | rstu | stuk | tukt | uktu | ktur | turm | urme | rmen | meny | enye | nyeb | yeba | ebab | babk | abka | bkan | kano | anor | norg | orga | rgan | gani | anis | nism | isme | smes | mesa | esan | sang | angg | nggu | ggup | gupm | upme | pmer | mere | erek | reka | ekam | kamf | amfa | mfak | fakt | akta | ktad | tadu | adun | duni | unia | niap | iape | apen | peng | engt | ngta | gtah | tahu | ahua | huan | uann | anny | nnya | nyam | yame | amem | memb | embi | mbim | bimb | imbi | mbin | bing | ingp | ngpe | gper | peri | eril | rila
	| ilak | laku | akun | kuny | unya |

- **Hashing**

Hashing is a way to transform a string into a unique value of a fixed length that serves as a string marker. Here is the equation to get the hashing value for each K-gram.

$$H = C_1 * a^{(k-1)} + C_2 * a^{(k-2)} + \ldots + C_3 * a^{(0)} \quad (17)$$

Explanation:

H = The hashing value
C_k = The ascii value of each letter in K-grams
A = Constant / base value
K = Lots of characters

- **Similarity**

Similarity is used to calculate the level of similarity of a text, which is compared to the result value of the hashing process. To measure the percentage of similarity using the Dice's Similarity Coefficient method where the same hash value of two different abstract text documents is computed and the numbers are stored in the form of arrays. Here is the similarity equation of Dice's Similarity Coefficient method.

$$S = \frac{2 * C}{A + B} \quad (18)$$

Explanation:

S = The value of similarity between the two texts
A = Number of K-grams in text 1
B = Number of K-grams in text 2
C = The sum of the same K-grams of the comparable text

The research undertaken focuses on handwriting recognition obtained from PCA extraction results. The resulting feature extraction is 13 feature extraction and will be classified using the K-NN method. In training training process, training is done by extracting the sequential character image from A-Z, a-z, 0-9 for eigenvalue and principal component (PC). After obtaining the eigenvalue and principal component (PC) of each character image, the eigenvalue and principal component (PC) will be stored in the database to be reused in the testing process. Data used 13 handwriting data training image and 5 handwriting data test image.

2.3. Testing Accuracy

In the test the accuracy will be calculated the value of precision or suitability of the application results handwriting recognition essay answers. Testing is done after training data has been done training process.

The parameter used in this research is the K value of the K-NN classification method used in the testing phase. This test is performed on 806 data sets. Here are experiments performed with different parameters on image testing.

Table 4 Test Result

No	K	Accuracy (%)
1	1	74.83
2	2	74.83
3	3	76.62
4	4	77.55
5	5	77.79

Based on the results of accuracy testing of 5 test data that has been done, evaluation of accuracy results obtained on handwriting essay recognition are as follows:

a. Accuracy results that have the highest percentage of the test to-1 using K = 5 with an accuracy of 77.79% of handwriting test data.
b. Accuracy results that have the lowest percentage of the 5th test using K = 1 and K = 2 with an accuracy of 74.83% against handwriting test data.

The higher the K value the higher the accuracy. The value of K greatly influences the K-NN classification. And the system can provide the value of essay assessment of the answers using the algorithm Rabin Karp by way of membenadingkan similarity of text between the introduction of answers with answer answer.

3. CLOSING

In this section will explain the conclusions and suggestions of the research.

3.1. Conclusion

In this study has produced a handwriting recognition system for essay reviewers using KNN and PCA methods. The introduction made to 62 types of output consisting of the letters A-Z, a-z and numbers 0-9.

Based on test results that have been done, testing of handwriting recognition system as follows:

1. Handwriting recognition system using K-NN and PCA method obtained the best accuracy of 77.79% with the value of K = 5. This accuracy is influenced by PCA Builder generated in the training process, the amount of training data and test data used.
2. The handwriting recognition system can provide an essay assessment value.

3.2. Suggestion

The system that is currently built can still be developed again in accordance with the needs of users in getting better accuracy results. Suggestions for further development on the system that can be taken into consideration are:

1. Using segmentation method that can be used for handwritten and handwriting handwritten image segmentation.
2. Using better feature extraction and classification methods to get the best accuracy.
3. Using better essay scoring methods on the essay assessment process.
4. Pay attention to the meaning in the sentence in order to provide an accurate assessment of the answer.

BIBLIOGRAPHY

[1] R. S. Bahri dan I. Maliki, “Perbandingan Algoritma Template Matching dan Feature Extraction Pada Optical Character Recognition,” *Jurnal Komputer dan Informatika (KOMPUTA)*, vol. I, no. 1, hal. 29-35, 2012.

[2] M. F. Miftah, “Pengenalan Karakter Huruf Tulisan Tangan Menggunakan Metode Principal Components Analysis,” dalam *Tugas Akhir Jurusan Teknik Informatika Fakultas Ilmu Komputer*, Bandung, Universitas Dian Nuswantoro, 2013.

[3] D. Puspitaningrum, D. K. Sari dan B. Susilo, “Dampak Reduksi Sampel Menggunakan Principal Components Analysis (PCA) Pada Pelatihan Jaringan Syaraf Tiruan Terawasi (Studi Kasus : Pengenalan Angka Tulisan Tangan),” *Jurnal Pseudecode*, vol. II, no. 1, hal. 83-89, 2014.

[4] R. Akbar dan E. A. Sarwoko, “Studi Analisis Pengenalan Pola Tulisan Tangan Angka Arabic (Indian) Menggunakan Metode K-Nearest Neighborsdan Connected Components Labeling,” *Dinamika Rekayasa*, vol. XII, no. 2, hal. 45-51, 2016.

[5] I. A. Ersyaputra, "Analisis Dan Implementasi Pengenalan Citra Tulisan Tangan Alfabet Menggunakan Metode Modified Direction Feature dan Klasifikasi Jaringan Saraf Tiruan-Learning Vector Quantization," Ilmu Komputasi, Fakultas Teknik Informatika, Universitas Telkom, Bandung, 2013.

[6] S. Hamza, M. Sarosa dan P. Budi Santoso, “Sistem Koreksi Soal Essay Otomatis Dengan Menggunakan Metode Rabin Karp,” *Jurnal EECCIS*, vol. 7, no. 2, hal. 153-158, 2013.

[7] D. Putra, Pengolahan Citra Digital, Yogyakarta: ANDI, 2010.

[8] A. Kadir dan A. Susanto, Teori dan Aplikasi Pengolahan Citra, Yogyakarta: ANDI, 2013.

[9] C. Solomon and T. Breckon, Fundamental of Digital Image Processing, West Sussex: John Wiley & Sons, Ltd, 2011.

[10] K. Khurshid, I. Siddiqi, C. Faure and N. Vincent, "Comparison of Niblack Inspired Binarization Methods for Ancient Document," in *Document Recognition and Retrieval XVI*, San Jose, 2009.

[11] R. Lesmana, "Pengenalan Karakter Tulisan Tangan Menggunakan Ekstraksi Fitur PCA & LDA," *Jurnal Teknik Elektro Universitas Kristen Petra*, 2003.

[12] B. Santosa, Data Mining: Teknik Pemanfaatan Data untuk Keperluan Bisnis, Yogyakarta: Graha Ilmu, 2007.

[13] A. Khan and H. Farooq, "Principal Component Analysis-Linear Discriminant Analysis Feature Extractor for Pattern Recognition," *IJCSI International Journal of Computer Science Issues*, vol. VIII, no. 2, hal. 267-270, 2011.

[14] H. Noprisson, B. Susilo dan E. , “Implementasi Algoritma Rabin-Karp Untuk Menentukan Keterkaitan Antar Publikasi Penelitian Dosen Tahun 2013,” *Jurnal Teknologi Informasi*, vol. 10, no. 1, hal. 110-127, 2014.