IMPLEMENTATION OF SUPPORT VECTOR MACHINE AND ZONE CENTROID ZONE METHODS IN THE INTRODUCTION OF KATAKANA AND HIRAGANA CITRA

Luthfi Nur Adli[1] , Irfan Maliki S.T., M.T[2].

[1,2]Teknik Informatika – Universitas Komputer Indonesia
Jl.Dipatiukur 112-114 Bandung
E-mail : adli.nur11@gmail.com[1],irfanmaliki007@gmail.com[2].

ABSTRACT

Optical Character Recognition (OCR) is the conversion of handwriting into a modified image file, OCR can be used for image recognition of Japanese characters, the methods used are*ArtificialNeuralNetwork [8], Backpropagation, HoughTransform[13],SupportVectorMachine[10]*. From several methods mentioned Support Vector Machine method is widely used because it has advantages in good accuracy. The Support Vector Machine method can be combined with the Zoning-ZoneCentroidZone feature extraction method to produce accurate results in writing recognition, this method was chosen based on previous research. SVM and Zoning classification methods will be more accurate. Therefore, in this study, how is the level of accuracy of handwriting recognition using the Zoning method (ZoneCentroidZone), Support VectorMachine (SVM) in the katakana characters and hiragana.The initial process is to enter the handwritten image into the application. Furthermore, through the image processing stage includes grayscale, thresholding, thining and feature extraction. From the image processing stage, an array of decimal values from character traits is obtained, the decimal array is reprocessed at the classification stage with the SVM method both training and testing. The results of handwriting recognition obtained digital text.The results of testing the handwriting recognition system on Katakana and Hiragana image recognition using the Support Vector machine method obtained the best accuracy of 23.2%. This accuracy is influenced by training parameters, training data and test data used.

Keywords: Artificial Intelligence, Support Vector Machine, Zoning, handwriting image recognition, image processing, SVM.

1. CONTENT OF RESEARCH

Image recognition in letters or often referred to as Optical Character Recognition (OCR) is the conversion of handwriting into a modified image file. The handwriting is then scanned using a scanner to produce an image. OCR techniques are commonly applied to various handwriting, one of which discusses Japanese script writing. Japanese characters are more complicated when compared to the character of Roman characters, besides Japanese characters are more numerous. So it takes a lot of time to understand it.

Japanese is a language that has a different form of language from other foreign languages. The form of language can be observed from pronunciation, vocabulary, grammatical, grammar, ways of disclosure, and the variety of languages used. The variety of Japanese language is strongly influenced by socio-cultural factors such as, region, social class, gender differences, and age [4].

Students and individuals feel the need to learn the letters Hiragana and Katakana, but many difficulties experienced when learning Japanese, among others, are difficult to recognize letters with each other because the letters katakana are many and almost similar, and it is difficult to know the accuracy of the results because of the solid Japanese handwriting each line has its own meaning [3]. So we need a tool to make it easier to recognize hiragana letters and katakana, one of them with the image recognition of Japanese characters.

Many methods can be used for image recognition of Japanese characters, including methods of Artificial Neural Network, Backpropagation, Hough Transform, and Support Vector Machine [17]. Of the several methods mentioned, the Support Vector Machine method is most widely used because it has advantages in accuracy. Support Vector Machine (SVM) is a learning system that uses the hypothetical space of a linear function in a high-feature dimension space developed by Boser, Guyon, Vapnik.The Support Vector Machine (SVM) method in this study is used because it is implemented for various kinds of research efficiently in its use.

Support Vector Machine method has been used in classifying several cases, such as the application of SVM method on the introduction of Sundanese speech, the application of Support Vector Machine method on elementary school accreditation data, signature identification, jawi script recognition, number handwriting recognition and mathematical algorithm based on Algorithm

Support Vector Machine and Hybird, the Support Vector Machine Multiclass (One Vs All) method, this method is used because in the multiclass (svm) comparison study, the best accuracy is 90%.

The Support Vector Machine method can be combined with the Zoning-Zone Centroid Zone feature extraction capable of producing accurate results in writing recognition, this method was chosen because it was based on previous research if the combined SVM and Zoning classification methods would be more accurate.

Based on the explanation that has been discussed, this research is intended to apply the Support Vector Machine and Zone Centroid Zone methods on the image recognition of Katakana and Hiragana.

2. LITERATURE REVIEW

2.1. Image

An citra is a two-dimensional image produced from a two-dimensional analog image which is continuous into a discrete image through a sampling process. Citra is divided into two types, namely analog citra and digital citra. The citra is the output of a data recording system that is optical, analog and digital. Recording citra data is divided into 2, namely Analog Citra and Digital Citra.

2.1.1. AnalogImage

Analog citra are continuous images, such as images on a television monitor, X-ray photos, photos printed on photo paper, paintings, landscapes, CT scans, images recorded on cassette tapes, and so on.

2.1.2. DigitalImage

Digital citra is a two-dimensional image that can be displayed on a computer screen as a set / discrete digital value called pixel / picture elements. In a mathematical review, citra is a continuous function of light intensity in a two-dimensional plane. An citra can be defined as a function of f (x, y) measuring M lines and N columns, with x and y being spatial coordinates, and the amplitude f at the coordinate point (x, y) embedded in the intensity or gray level of the citra at that point. If the value of x, y and the amplitude value of f is finite and is discrete, then it can be said that the citra is a digital citra.

2.1.3. Japanese characters

In general Japanese writing is divided into four types of writing including hiragana, katakana, kanji and romanji.The writing has the function of each as used in foreign language absorption, people's names, conjunctions, and standard words. In this final project, researchers took case studies, namely katakana and hiragana writing.

ア a	イ i	ウ u	エ e	オ o
カ ka	キ ki	ク ku	ケ ke	コ ko
サ sa	シ shi	ス su	セ se	ソ so
タ ta	チ chi	ツ tsu	テ te	ト to
ナ na	ニ ni	ヌ nu	ネ ne	ノ no
ハ ha	ヒ hi	フ fu	ヘ he	ホ ho
マ ma	ミ mi	ム mu	メ me	モ mo
ヤ ya		ユ yu		ヨ yo
ラ ra	リ ri	ル ru	レ re	ロ ro
ワ wa				ヲ (w)o
ン n				

Picture 1 *Katakana*letter

Picture 1. An example of the Katakana letter.

あ a	い i	う u	え e	お o
か ka	き ki	く ku	け ke	こ ko
さ sa	し shi	す su	せ se	そ so
た ta	ち chi	つ tsu	て te	と to
な na	に ni	ぬ nu	ね ne	の no
は ha	ひ hi	ふ fu	へ he	ほ ho
ま ma	み mi	む mu	め me	も mo
や ya		ゆ yu		よ yo
ら ra	り ri	る ru	れ re	ろ ro
わ wa				を wo
				ん n

Picture 2 *Hiragana*letter

Picture2. An example of the Hiragana letter.

2.2. *Preprocessing Digital Image*

2.2.1. *Grayscale*

The grayscale image handles black and white gradations, which produce a gray effect, the color of the image is expressed in intensity. In this case, the intensity ranges from 0 to 255. A value of 0 represents black and a value of 255 denotes white. The following is the color image conversion (RGB) formula to be grayscale.

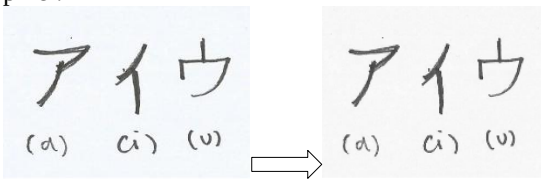
$$I = (0.2989 * R) + (0.5870 * G) + (0.1141 * B) \quad (2.1)$$

Explanation :

I = grayish scale value search function.

R = Shows the red (Red) value of a point per 1 pixel.

G = Shows the green value (Green) of a point per 1 pixel.
B = Shows the blue (Blue) value of a point per 1 pixel.



Picture 3 Examples of Grayscale results

2.2.2. Sauvola Threshold

Sauvola Threshold is the threshold method used in this study. Sauvola Threshold is able to detect writing images well, by calculating the threshold using a vulnerable dynamic value from the standard deviation of the grayscale image. This threshold is included in the local threshold where the threshold value is determined by the subscriber pixel value, the purpose of using this threshold process is to simplify the form of each image.
the following is the formula of the threshold.

$$T(x,y) = m(x,y) * (1 + k * \left(\frac{s(x,y)}{R} - 1\right)) \quad (2.2.1)$$

$$m(x,y) = \frac{\sum_{i=min}^{i=max} \sum_{j=min}^{j=max} img\,(i,j)}{i*j} \quad (2.2.2)$$

$$s(x,y) = \sqrt{\frac{(img\,(i,j) - m(x,y))^2}{(i*j)-1}} \quad (2.2.3)$$

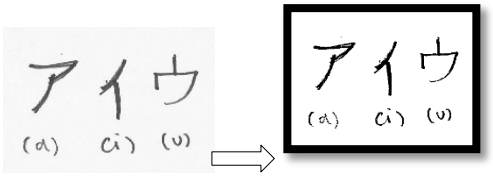
Explanation:
R: maximum value of standard deviation (128 for grayscale image)
k: kernel with a value between 0.2 - 0.5
m: function that produces the average value of a number of pixel images.
s: function that produces deviations from a number of pixel images
T: function that produces a threshold value (threshold)
x: image width coordinate value in formula (i)
y: image height coordinate value in formula (j)
After the threshold value of T (x, y) has been obtained, then enter the equation below.

$$f(x,y) = \begin{cases} 0, img\,(x,y) < T(x,y) \\ 255, img\,(x,y) \geq T(x,y) \end{cases} \quad (2.5)$$

$$f(x,y) = \begin{cases} 1\,; x = 0 \\ 0\,; x = 255 \end{cases} \quad (2.6)$$

Explanation:
f: function that produces a value of 0 or 255.
img: image grayscale value.

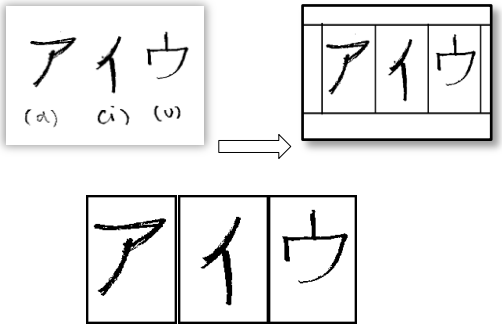
Pixel intensity value which is greater or equal to the threshold value will be converted to 0, therefore 255 values are white. While the pixel intensity value that is less than the threshold value will be converted to 1, therefore the value of 1 is black.



Picture 4 Examples of*Sauvola Threshold*results.

2.2.3. Segmentasi Citra

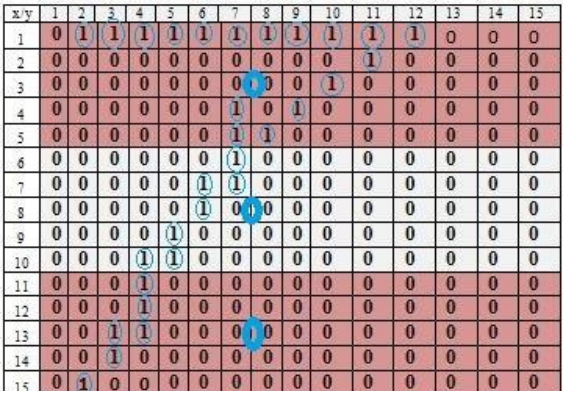
In this process, the input used is a black image resulting from the sauvola threshold and dilated. Next will be cut to get the image of Japanese letters. Cutting is done for each row (horizontal) in the input image first, then cutting each column (vertical) on each image that is cut in line.



Picture 5 Examples of*Segmentation* results.

Feature Zoning Extraction Zoning is one feature extraction method where the core of this method is dividing the image in the same number of zones to identify the characteristics of each letter character which then produces a value to be processed at the classification stage. Here are the steps in Zone Centroid Zone (ZCZ) algorithm.

1. Share the input image into the same number of n sections.



x/y	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0
2	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
3	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
4	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0
5	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
6	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
7	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
8	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
9	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
10	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
11	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
13	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
14	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
15	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0

Picture 6 Example of Matrix Input Image

2.Calculate the distance of the black pixel from the centroid, with the following equation.

$$x_c = \frac{(x_1.p_1 + x_2p_2 + \ldots + x_np_n)}{(p_1 + p_2 + \ldots + p_n)}$$

$$y_c = \frac{(y_1.p_1 + y_2p_2 + \ldots + y_np_n)}{(p_1 + p_2 + \ldots + p_n)}$$

Explanation:

$x_c = Centroid\ koordinat\ x$

$y_c = Centroid\ koordinat\ y$

$x_n = Koordinat\ x\ dari\ pixel\ ke-n$

$y_n = Koordinat\ y\ dari\ pixel\ ke-n$

$p_c = Nilai\ pixel\ ke-n$

Calculation of the distance of each zone can be seen below.

Centroid Zona 1

$$Xc = \frac{2.1 + 3.1 + 4.1 + 5.1 + 6.1 + 7.1 + 7.1 + 7.1 + 8.1 + 8.1 + 9.1 + 9.1 + 10.1 + 10.1 + 11.1 + 11.1 + 12.1}{17} = \frac{129}{17} = 7{,}58 \sim 8$$

$$Yc = \frac{1.1 + 1.1 + 1.1 + 1.1 + 1.1 + 1.1 + 1.1 + 1.1 + 1.1 + 1.1 + 1.1 + 2.1 + 3.1 + 4.1 + 4.1 + 5.1 + 5.1}{17} = \frac{34}{17} = 2$$

Centroid Zona Ke-2

$$Xc = 5{,}71 \sim 6$$

$$Yc = 3{,}14 \sim 3$$

Centroid Zona Ke 3

$$Xc = 3{,}33 \sim 3$$

$$Yc = 3$$

3.Calculate the Distance of Each Pixel in the Zone, with the following equation.

$$d(P, C) = \sqrt{(x_p - x_c)^2 + (y_p - y_c)^2}$$

d = distance between two points
p = Black Point Coordinate
C = Pixel Coordinate
xp = Coordinate x center of gravity
yp = y center of gravity
xc = x center coordinates
yc = y pixel coordinates

Calculation of distances for each Zone can be seen below.

Zona 1 (2,1) : distances= $\sqrt{(2-8)^2 + (1-2)^2} =$ 6.0827625302982197

....

Zona 2 (7,1) : distances= $\sqrt{(7-6)^2 + (1-3)^2} =$ 2.2360679774997897

Zona 2 (5,5) : distances= $\sqrt{(5-6)^2 + (5-3)^2} =$ 2.2360679774997897

....

Zona 3 (4,2) : distances= $\sqrt{(4-3)^2 + (2-3)^2} =$ 1.414213562373095

....

Zona 3 (2,5) : distances= $\sqrt{(2-3)^2 + (5-3)^2} =$ 2.2360679774997897

4. Calculate the average distance between points

$$Zona\ 1 : \text{average distance} = \frac{(\text{distance}(2.1) + \text{distance}(3.1) + \text{distance}(4.1) + \cdots \text{distance}(8.1)}{many\ points\ (17)} = 1.8611116548726$$

$Zona\ 2 :$ average distance $= 1.589855743498$

$Zona\ 3 :$ average distance $= 1.31439158622$

2.3. *Support Vector Machine* (SVM)

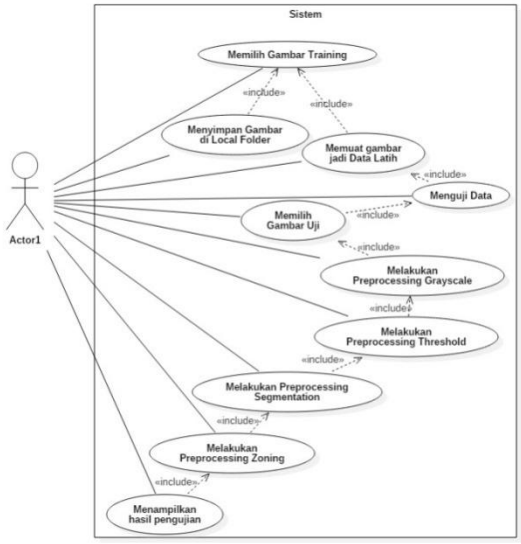
Support Vector Machine is a learning method used for binary classification, the main idea is to find the best Hyperplane that functions as a separator of two classes in input space. This method was introduced by Valdimir Vapnik in 1992 as a superior concept in the field of pattern recognition.

In this research, data representation uses the format of spare data representation with a vector as its input by first inputting +1 or -1 stating the two classes. The numbers both represent the dimensions (row_id) and the third number after the ":" mark the weight of the character.

Tabel 1. Format SVM

Karakter	A	I	U
Zona 1	1.8611116548726	1.1165472688611	1.1511487862616
Zona 2	1.589855743498	1.45898534985	2.898349855574
Zona 3	1.31439158622	1.31456239128	3.35862214391

2.4. Use Case Diagram



Gambar 7*Use case Diagram*

3 CLOSING

3.1 Inference

Based on the results of the testing system for handwriting recognition in the case of the introduction of the image using the method of support vector machine, the greatest accuracy found 23.2%. This accuracy is affected by parameters of training, the training data, test data, and the program that is used.

3.2 Suggestion

For the next research on handwriting recognition using the Support Vector Machine method in the case of image recognition say and hiragana have better accuracy, then the following are suggestions that can be given to future researchers, as follow:

1. It takes the handwriting image segmentation method for separate writing cases.
2. In the image recognition stage can use other methods such as Metaphone, TF-Idf, Cosine Similarity, and Zhu-Takaoka methods.
3. It is recommended to use non-separate or continuous images to make it easier to recognize letters / images because in this case hiragana and katakana images have many letters that have separate parts in 1 (one) letter.

REFERENCES

[1] Krishnapriya, M, dkk 2015, *Optical Character Recognition For Handwritten Text*, Vol.1, Issue.2, International Journal of Contemporary Research in Computer Science and Technology, India.

[2] Javad Sadri,ChingY suen, Tien D Bui,"*Application Of Support Vector Machines For Recognition Of HandWritten Arabic/Persian Digits".*

[3] Danasasmita, Wawan (2009). *Metodologi Pembelajaran Bahasa Jepang*. Bandung:Rizki Press.Buku Bahasa Jepang untuk SMA.

[4] Drs. Sudjianto, M.Hum. (2010). Metodologi Pembelajaran Keterampilan Berbahasa Jepang. Bekasi: Kesaint Blanc.

[5] Eva Tuba,Milan Tuba,Dana Simian."*HandWritten Digit Recognition By Support Vector Machine Optimized By Bat Algoritma".*

[6] Dragomir R. Radev, "*Centroid-based summarization of multiple documents : sentence extraction, utility-based evaluation, and user studies*".

[7] N.Amalia, "pengenalan pola huruf hijaiyah menggunakan SVM," Universitas Sumatera Utara, Medan, 2016

[8] M. I. a. Musthofa, "Sundanese Script Recognition Using Zoning Feature Extraction and Support Vector Machine Classfication," Department Of Computer Science, Vol. 21, no. Bogor Agricultural University, Bogor, p, l, 2013.

[9] Devi Rima. SS, Diana Lady .SS. "*Respon Mahasiswa Terhadap Pembelajaran Bahasa Jepang Di Universitas Andalas:Studi Tentang Penerimaan Huruf Hiragana dan Katakana*".

[10] Muhammad Naeem Ayyaz,dkk. "*HandWritten Character Recognition Using Multiclass SVM Classfication With Hybrid Feature Extraction."*

[11] Muhammad Ferderico S. *Implementasi Metode Support Vector Machine (SVM) dan Zoning pada pengolahan citra untuk pengenalan citra tulisan tangan*

[12] R. P. Rajashekararadhya SV, "Novel Zone Based Feature Extraction Algorithm For Handwritten Numeral Recognition Of Four Indian Scripts," *Digital Technology Journal*, vol. 2, pp. 41-51, 2009.

[13] R. Sianipar, Teori dan Implementasi Java, Bandung: Informatika, 2013.

[14] R. A. a. M.Shalahudin, Rekayasa Perangkat Lunak 'Terstruktur dan Berorientasi objek', Bandung: Informatika, Bandung, 2014.

[15] N. Amelia, "pengenalan pola huruf hijaiyah menggunakan SVM," Universitas Sumatera Utara, Medan, 2016.

[16] H. B., Rekayasa Sistem Berorientasi Objek, Bandung: Informatika, Bandung, 2004.

[17] Krisantus, Tutorial SVM Berbahasa indonesia, Yogyakarta: Krisantus, 2003.

[18] I. J. a. W. M. Muhammad Naeem Ayyaz, "Handwritten character recognition using multclass SVM classification with Hybrid Feature Extraction," *J.Engg & Appl*, vol. 10, no. J.Engg & Appl, pp. -, Januari 2012.