(IJID) International Journal on Informatics for Development,
Vol. 4, No. 2, 2015, Pp. 8-11

Japanese Letter Pattern Recognition Application with Tesseract Engine

Akhmad Imam Fahrizal
Informatics Department
Islamic State University (UIN) of Sunan Kalijaga
Yogyakarta, Indonesia

Ahmad Subhan Yazid
Informatics Department
Islamic State University (UIN) of Sunan Kalijaga
Yogyakarta, Indonesia
yazid.anfalah@gmail.com

Shofwatul ‘Uyun
Informatics Department
Islamic State University (UIN) of Sunan Kalijaga
Yogyakarta, Indonesia
shofwatul.uyun@uin-suka.ac.id

***Abstract*— Digital image processing is a field that is being cultivated by many researchers at this time because it is interesting to apply to various activities, both analysis and production activities. One branch of the digital image is pattern recognition. This study uses Tesseract as a tool to recognize patterns from Hiragana letters. This study was conducted to find out how much Tesseract was able to recognize a Japanese text and handwritten text. This study uses 1 image as training data containing 74 Hiragana letters which are processed through training for each letter. This study has several testing criteria based on font size and resolution to find the best results in pattern recognition. This pattern recognition system is able to do data training and recognize 74 Hiragana letters using the Tesseract Engine. The system can also recognize images with the best success percentage of 98.24% with an image resolution of 200dpi (dots per inch) at size 18. This system can also recognize handwritten images with the best percentage of success of 90% with 200dpi image resolution.**

Keywords- Hiragana; Pattern Recognition System; Tesseract.

I. INTRODUCTION

Foreign languages have become popular with people, such as English, Japanese, German, Mandarin, French, and others. The number of foreign language enthusiasts makes software developers compelled to create applications that support the learning of these languages. Many of them are applications of artificial intelligence.

One of the fields studied in artificial intelligence is artificial neural networks. This artificial neural network has many advantages, including allowing computers to conduct training: accepting sets of inputs and setting targets. Input can be done with various devices including a mouse, keyboard, joystick, and others. In its development, using a device input such as a keyboard is no longer efficient. Technology development is directly proportional to the design of practical and efficient technological features.

On the other hand, there are several languages whose writing letters and characters are different from the alphabet in general. These languages tend to use symbols to convey information. This is better known as calligraphy. Various kinds of characters or letters including calligraphy, including Mandarin, Japanese, Arabic, and many more. Japanese has three types of letters, namely Hiragana, Katakana, and Kanji. However, the most common and easiest to learn is Hiragana letters, while Katakana is usually used for foreign absorption words, and kanji is usually quite difficult to learn.

The input will be more easily recognized by the computer when the characters or letters entered are digital. For printed or handwritten inputs, certain handling needs to be done. One method that is commonly used to transform non-digital data into digital is through a pattern recognition system with objects in the form of images.

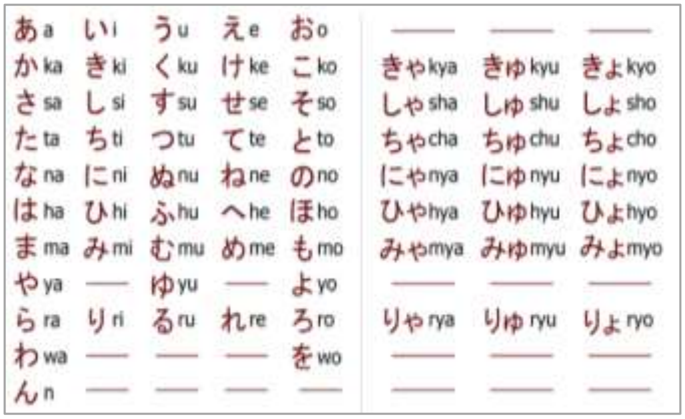
The image is a continuous function of the intensity of light in a two-dimensional field. When an optical device that records light reflection is a digital machine, for example, a digital camera, the resulting image is a digital image [1]. The recognition of digital images can be done using the Tesseract engine. Tesseract is one of the most accurate OCR sources. Combined with the Leptonica Image Processing library, Tesseract can read various image formats and convert them to text in more than 60 languages [2].

This study aims to determine how much the accuracy of Tesseract in recognizing Japanese handwriting and Japanese text as well as knowing the best percentage of Tesseract to recognize Japanese handwriting and Japanese text handwriting based on predetermined parameters. Following is a brief description about Hiragana.

II. HIRAGANA

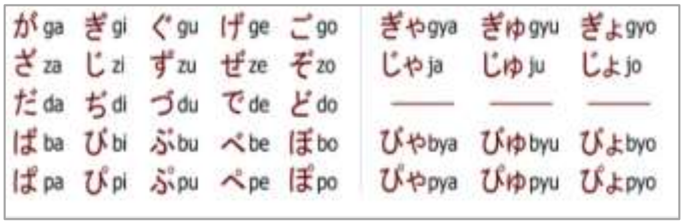
Hiragana (ひ ら が な) is a way of writing Japanese and represents the designation of syllables. In the past, it was also recognized as onna de (女 手) or 'women's writing' because it is commonly used by women [3]. Men at that time wrote using the writings of Kanji and Katakana. Hiragana was used extensively in the 10th century.

Hiragana letters are divided into two, namely the original letter and Hiragana letter development [4]. In Figure 1, the left-hand side shows the original letters and the right panel shows Hiragana development. While Figure 2 shows the development of a combination of Hiragana letters [5].



あ a	い i	う u	え e	お o	—	—	—
か ka	き ki	く ku	け ke	こ ko	きゃ kya	きゅ kyu	きょ kyo
さ sa	し si	す su	せ se	そ so	しゃ sha	しゅ shu	しょ sho
た ta	ち ti	つ tu	て te	と to	ちゃ cha	ちゅ chu	ちょ cho
な na	に ni	ぬ nu	ね ne	の no	にゃ nya	にゅ nyu	にょ nyo
は ha	ひ hi	ふ hu	へ he	ほ ho	ひゃ hya	ひゅ hyu	ひょ hyo
ま ma	み mi	む mu	め me	も mo	みゃ mya	みゅ myu	みょ myo
や ya	—	ゆ yu	—	よ yo	—	—	—
ら ra	り ri	る ru	れ re	ろ ro	りゃ rya	りゅ ryu	りょ ryo
わ wa	—	—	—	を wo	—	—	—
ん n	—	—	—	—	—	—	—

Figure 1. Vowels, consonants and their development



が ga	ぎ gi	ぐ gu	げ ge	ご go	ぎゃ gya	ぎゅ gyu	ぎょ gyo
ざ za	じ zi	ず zu	ぜ ze	ぞ zo	じゃ ja	じゅ ju	じょ jo
だ da	ぢ di	づ du	で de	ど do	—	—	—
ば ba	び bi	ぶ bu	べ be	ぼ bo	びゃ bya	びゅ byu	びょ byo
ぱ pa	ぴ pi	ぷ pu	ぺ pe	ぽ po	ぴゃ pya	ぴゅ pyu	ぴょ pyo

Figure 2. Development of a combination of Hiragana letters

III. METHODS

This study is begun with a preliminary study, which examines the variables used and relates to tesseract. The data in this study were obtained through literature studies of books, papers, and data on the internet.

The development of this system refers to the pattern recognition method which in general consists of the following stages: Making image training (providing sample images), training data, and testing systems. The system design flow chart can be seen in Figure 3.

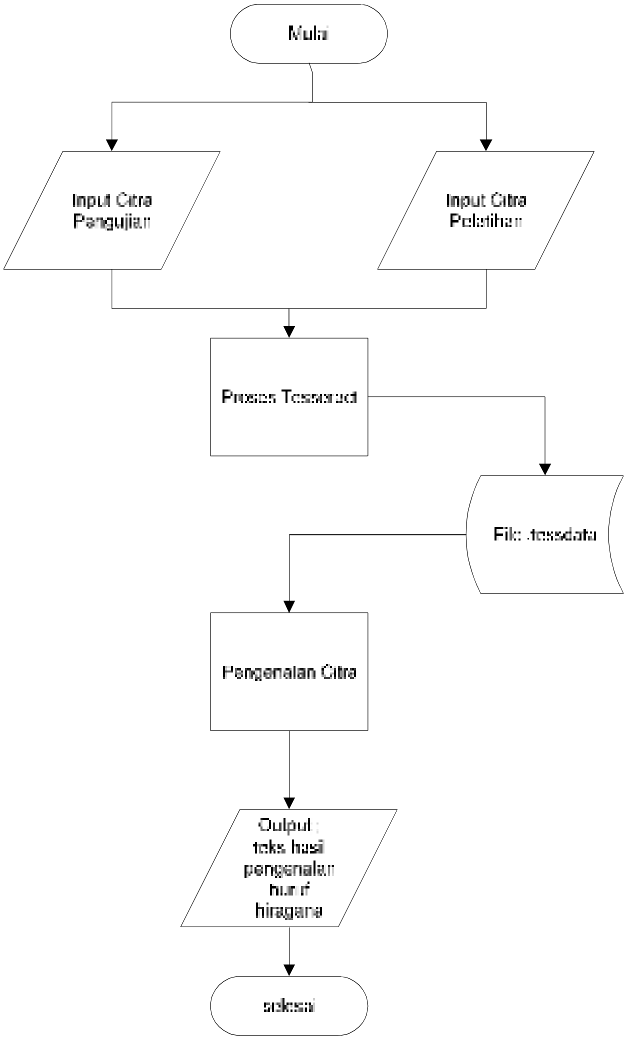


Figure 3. System flow chart

IV. RESULT AND DISCUSSION

The Japanese letter pattern recognition system with the Tesseract Engine briefly has two stages: drawing, training, and letter recognition stages.

A. *Letter Drawing*

This stage is the initial stage, where the researcher draws an image to be used in the training. Images made in the format of the tagged image format (.tif) with the contents of the image are 71 Hiragana letters (Figure 4).

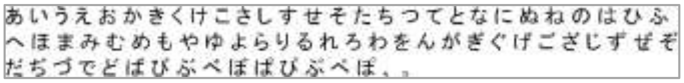
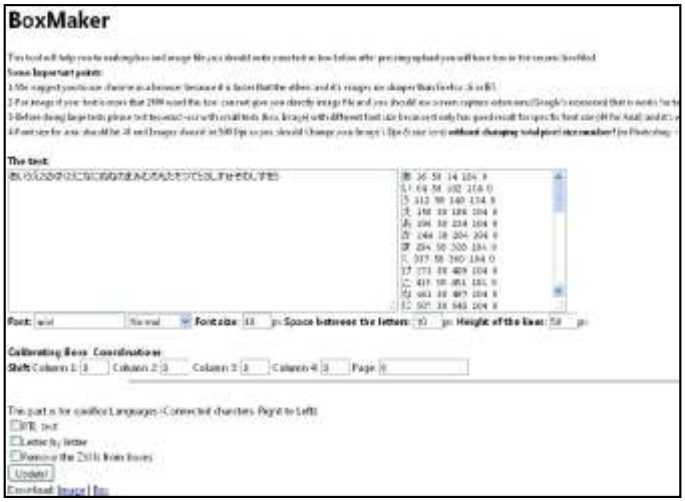

あいうえおかきくけこさしすせそたちつてとなにぬねのはひふ
へほまみむめもやゆよらりるれろわをんがぎぐげござじずぜぞ
だぢづでどばびぶべぼぱぴぷぺぽ、。

Figure 4. Hiragana letters as training data

B. *Training*

The training stage is the stage for training Hiragana letters so that they can be recognized by tesseract using the basis of the training is 71 Hiragana letters (Figure 4). At this stage, first is the creation of Box Files. The researcher utilizes online media, namely BoxMaker, which functions to make images in the .png format while creating its own Box Files (Figure 5).


BoxMaker

Some Important points:

The text:

Font: Font size: px Space between the letters: px Height of the line: px

Calibrating Box Coordinations

Shift Column 1: Column 2: Column 3: Column 4: Page

□RTL text
□Letter by letter
□Remove the ZWJs from boxes

Update!

Download: Image | Box

Figure 5. The interface of Box Maker

After getting the .box extension files, the thing to do is to manually correct each letter to improve the accuracy of its introduction. In this case, the researcher uses the JtessBoxEditor to edit and correct the letters one by one (Figure 6).

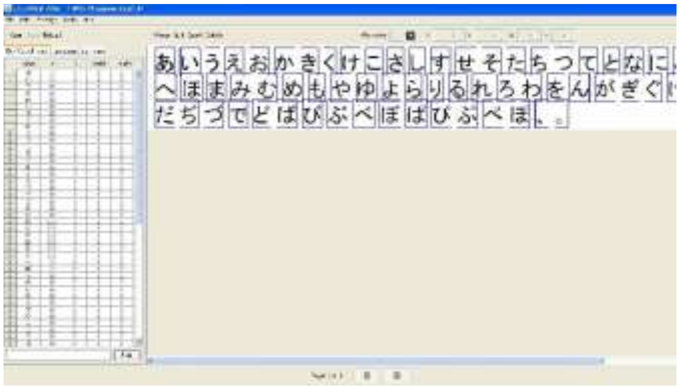


Figure 6. The interface of JtessBoxEditor

The left panel of figure 06 is a panel to correct letter errors that occur when creating a file box. This panel also functions to see what letters are recognized by the image in the right panel.

The next stage is the Box Training stage. This stage is carried out training for the file box, by running the command below and the file with the extension *.tr.

```
tesseract
Lang.Hiragana.exp0.tif
Lang.Hiragana.exp0.box nobatch
box.train
```

The next step is charset. This stage aims to find out the set of possible characters that can be produced. At this stage, tesseract will take the charset from the box files. At this stage the command entered is:

```
unicharset_extractor
Lang.Hiragana.exp0.box
```

The next step is the font properties. At this stage, the usual text (.txt) file is created, which will then delete the .txt format. The purpose of this file is to provide letter style information that will appear when the letter has been recognized. Fill in the file with the format:

```
Hiragana 0 0 0 0 0
```

After the character features of all training, pages have been extracted, the next stage is clustering which aims to create a prototype. Character shape features can be grouped using shape cluster, mftraining, and contraining. All of that will produce Inttemp, pffmtable and normproto files.

The next step is renaming by adding a trained language prefix. In this case, the researcher added the "high" prefix for all files. The last training phase is the merging stage. This stage is the stage to combine all the files resulting from previous training with commands:

```
Combine_tessdata hgn.
```

C. *Letter Recognition*

This stage is the stage where the researcher uses the Tesseract Engine to recognize letters based on the input image of the user (Figure 7).

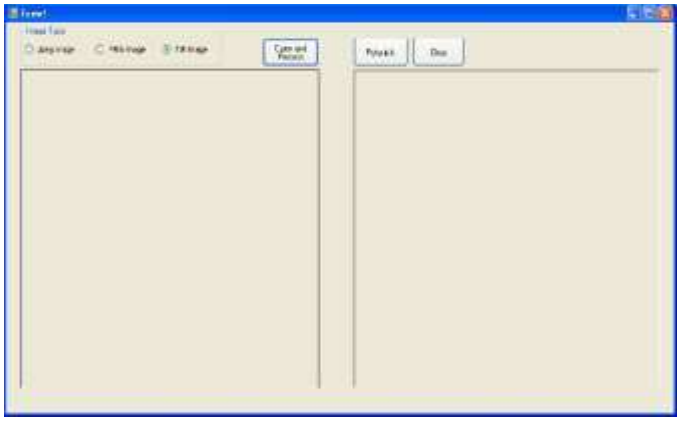


Figure 7. System interface display

Function of the “Open and Process” button is to browse folders when the user wants to select the image to be recognized. When the user has selected the image, the system will immediately process and display the results in the text box on the right side. While function of the “Petunjuk” button is to help by giving instructions to users about how to use this system.

In order for this system to run smoothly, there are two system requirements that must be fulfilled by the user: (1) Tesseract 3.02 must be installed on the user's computer. (2) The hgn.traineddata file must be copied to the Tesseract-Ocr/tessdata folder.

D. *Testing*

The testing stage uses 2 image bases and 10 Hiragana letter handwritten images which will be divided into 56 tests based on image resolution and font size. Table 1 shows the best results from testing.

TABLE I. SYSTEM TEST RESULTS

File Name	Resolution	Font Size	Total Letters	Total Letter Errors	Total Words	Total Word Errors
Img_0004	600dpi	11	73	8	17	4
Img_0006	200dpi	18	73	2	17	2
Img_0008	600dpi	18	73	2	17	2
Img_0011	300dpi	11	279	7	61	5
Img_0013	100dpi	18	279	3	61	3
Img_0030	200dpi	-	20	2	-	-

V. CONCLUSION

Pattern recognition for Hiragana letter handwriting developed with Tesseract Engine can recognize handwriting with a success percentage of 52.90% and the percentage of success for recognizing Hiragana text is 91.75%. Tesseract can recognize handwriting properly in the state of 200dpi resolution with a success percentage of 90%, while for Hiragana text in size 18 and resolution 200dpi with a percentage of success of 98.24%.

REFERENCES

[1] A. R. Kardian, “Pengantar Pengolahan Citra,” in *Pengolahan Citra Digital*, Jakarta, 2012.

[2] R. Smith, “Tesseract OCR engine What it is, where it came from,where it is going,” 2007. .

[3] E. and C. L. Tsujita, *Mahir Bahasa Jepang dalam Sepekan*. Jakarta: Kesaint Blanc, 2005.

[4] A. Hasnan, “Pengantar Belajar Bahasa Jepang,” 2009. .

[5] S. Center, “Hiragana,” 2006.