ISBN 978-981-14-1684-2
Proceedings of 2019 the 9th International Workshop on Computer Science and Engineering
(WCSE 2019 SUMMER)
Hong Kong, 15-17 June, 2019, pp. 540-544
doi: 10.18178/wcse.2019.06.080

Chinese Character Translator on Mobile Phone using Optical Character Recognition and Bing Translator API

Andreas Handojo[1, 2+], Anita Nathania Purbowo[1], and Fenny Valentine Budiono[1]

[1]Informatics Department, Faculty of Industrial Technology, Petra Christian University
[2]Mobile Computing Research Group, Petra Christian University

Abstract. Chinese language is one of the international languages whose have users almost 35% of the world's population. Nonetheless the Chinese language has problems in learning how to write and how to read because it is in the form of characters or symbols so that it is more difficult to learn it. Chinese characters that used today is simplified Chinese character with approximately 3000 common characters that daily used. This character / symbol can also be written in Latin alphabet form called hanzi / hanyu pinyin. Some application developers such as Yellow Bridge, Google, Qhanzi, and Bing have provided translator applications from the Chinese characters to the Latin alphabet and vice versa. The application provided is generally still web-based and does not involve the ability to input the shape of a Chinese character in the form of an image, for example image input either from a file or directly from a camera input. This research try to build a Chinese character translator application using Tesseract Optical Character Recognition (OCR) Engine to retrieve the Chinese characters from the image, then translate it using a translator on the Bing API. This application will running on mobile phone. So the user could use image or mobile phone camera as an input. The test results show that the application can operate on various Android devices. OCR Engine has been able to perform the translation function with 74% success rate. The input image could have tolerance angle of approximately 15 degrees.

Keywords: Translator, Chinese character, optical character recognition, mobile phone, Bing API.

1. Introduction

Chinese language is one of the international languages whose have users almost 35% of the world's population. Nonetheless the Chinese language has problems in learning how to write and how to read because it is in the form of characters or symbols so that it is more difficult to learn it. Chinese characters that used today is simplified Chinese character with approximately 3000 common characters that daily used. This character / symbol can also be written in Latin alphabet form called hanzi / hanyu pinyin. Some application developers such as Yellow Bridge [1], Google [2], Qhanzi [3], and Bing [4] have provided translator applications from the Chinese characters to the Latin alphabet and vice versa. The application provided is generally still web-based and does not involve the ability to input the shape of a Chinese character in the form of an image, for example image input either from a file or directly from a camera input.

On the other hand, the use of mobile device technology is increasingly being used by the general public. Where the existing mobile device equipment supports input in the form of images both in the form of image files and directly from the camera. It will be very helpful if there is an application that runs on a mobile device that has the ability to translate Chinese characters that are input directly using the camera or from image files that already exist on the mobile device.

This research try to build a translator application that running on mobile phone. This application has ability to translate Chinese characters into Indonesian language using image as an input. The image could be

+ Corresponding author. Tel.: +62-31-2983456
E-mail address: handojo@petra.ac.id

insert as an image files or directly from the mobile phone camera. The user could take photos of objects (such as writing on paper or road signs) that contain Chinese characters using the mobile phone camera. Then, the system will recognize the Chinese characters by using the Tesseract Optical Character Recognition (OCR) Engine. This mandarin character then sent to Bing Translation API. The application then show the user the translation result. In addition, this application is also has Chinese character play sound feature so the user could hear the pronunciation from the Chinese character that inputted. The user also could save the results of the translation that was entered earlier. The results of the study were tested using several pictures of Chinese characters from the text book, newspaper, board sign. Also tested is the slope factor for taking pictures to determine the level of accuracy that can be obtained.

2. Literature Review

Tesseract is an open-source Optical Character Recognition (OCR) engine. Tesseract was initially developed at Hewlett-Packard between 1985 and 1995, but was never commercially exploited. Tesseract is one of the 3 best engines in the 1995 UNVL Accuracy Test. But at that time, the Tesseract construction had stopped. In 2005, HP transferred the same Tesseract source code to ISRI and released it as open source. In Fig 1 shows the architecture owned by Tesseract. First of all, Tesseract accepts the input results in the form of binary image. Tesseract can do some image processing internally before performing OCR [5].

After that, connected component analysis is performed on the binary image to find out where the component outline is stored. The outline is collected together and forms a BLOB. Blob is compiled into a text line. Text lines are cut into spaces. This character recognition process involves two stages of the process. The first step is to recognize the words in the order. If words can be identified in the first stage, then the words can be forwarded to adaptive classifier as training data to get the classification results. But if the first stage cannot recognize words, a second process stage needs to be done. At this stage, the adaptive classifier has obtained information from the first stage before. There are several similar studies that try to translate words from input images such as Panchal et al. [6], Pai et al. [7], and Manisha & Liyanage [8], but the research carried out has not utilized the existing translation application. This research will use Bing Translation API [4] as translator so this application no need to save translation data. Bing is made by Microsoft, which can be used in applications to do translations. This API can run on REST, SOAP, or JavaScript protocols and return XML or JSON format responses.

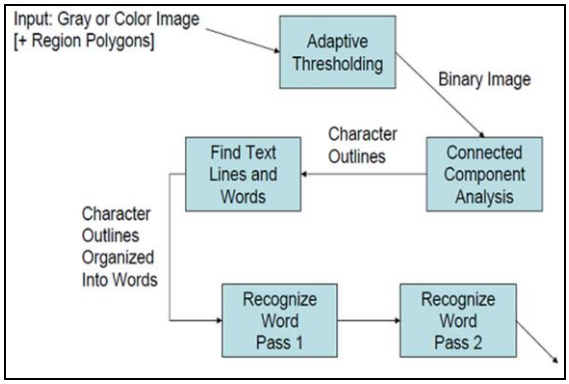


Fig 1: Tesseract Architecture [5].

3. Design and Implementation

The system design for this application could be seen on Fig 2. After install the application user could enter the image of the Chinese character that want to translate. Input can be an image file (from the mobile phone gallery) or directly from mobile phone camera. The application will run the image trough Tesseract OCR to change the image to Chinese character. Then, application will sent the Chinese character to the Bing Translator API to be translated. The translation results then sent back to the mobile phone and show to the user.

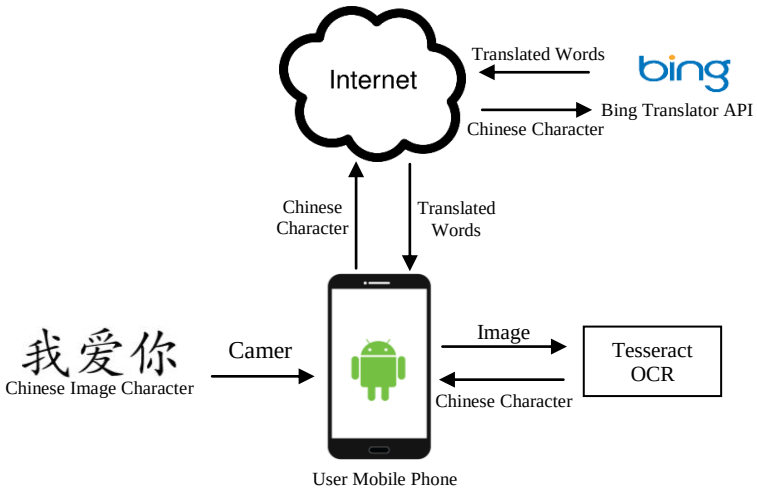


Fig 2: Chinese Character Translator

The application has been tried on several types of mobile phones as can be seen on Figure 3. The user can access the data retrieval menu (can be an image file or directly from a mobile phone camera). For the choice of image files, the application will access the image gallery on the mobile phone. While for the camera option the application will access the camera on the mobile phone (Fig. 3).

Fig 3. Access Mobile Phone Camera

After successfully taking pictures from the mobile phone camera or image gallery, the application will accesses crop image feature that provided by the application (Fig 4). With this feature the user can choose which characters to choose to translate.

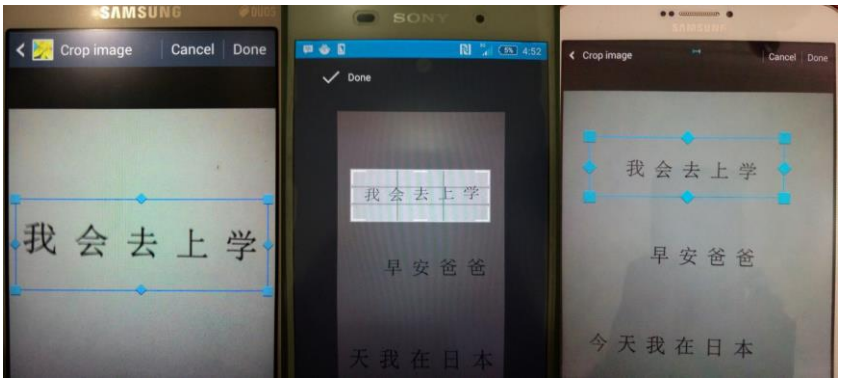


Fig 4: Crop Image Feature

User then could access the translator results menu. The translator result (Fig. 5) will display the images that have been entry, Chinese characters that are recognized, and the text of the translation. Users can also listen to the sounds of the translated characters. The user also could save the translation results by pressing the bookmark button. So the user could see the translator result on another occasion.

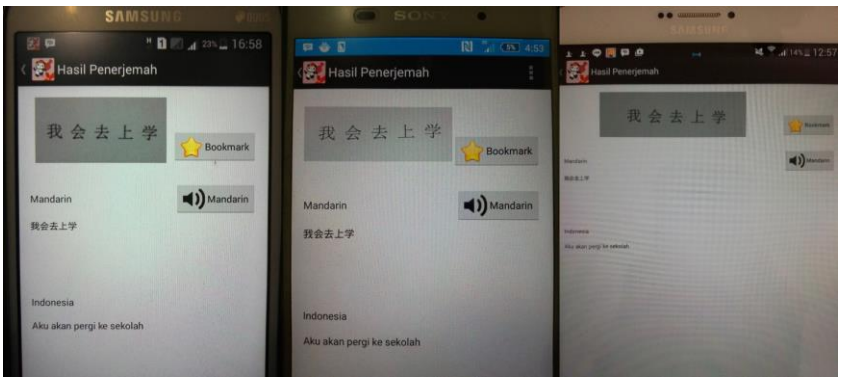
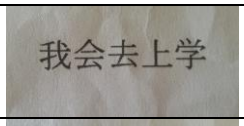
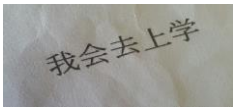
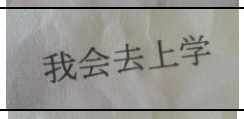
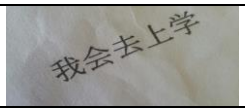
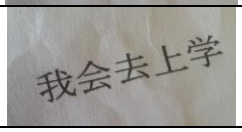


Fig 5: Translate Result

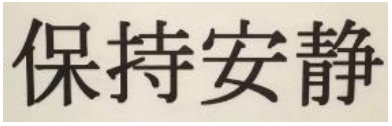
Testing is done by taking photos with different slope angles (samples can be seen in Table 1). The sloping the character is taken, the success rate will decrease. The success rate of the input slope on the Tesseract OCR Engine is approximately 15 degrees.

Table 1: Slope Angle Input Testing

No.	Testing Input	Slop Degree	Success Rate	No.	Testing Input	Slop Degree	Success Rate
1.		0^0	100%	4.		20^0	0%
2.		10^0	100%	5.		25^0	0%
3.		15^0	70%				

Testing is also done by inserting images in various forms (samples can be seen in Table 2) such as road signs, restaurant signs, newspapers, and others. From the tests that carried out, the results show that the application works well for the character shape that has a standard size and shape, while for nonstandard characters and sizes it has a low success rate. From 300 testing, the application success rate is 74%.

Table 2. Image Input Testing

No.	Original Picture	Image Input (Crop Image)	Success
1.			Success
2.			Success
3.			Not Success
4.			Success
5.			Not Success

6.			Success

4. Conclusion

This study focuses to create an application on mobile phone that have ability to translate Chinese character using image input. The inputted image can be entered form an image file (from image gallery on mobile phone) or directly through the mobile phone camera. The image then converted into Chinese character using Tesseract OCR. The Chinese character then sent and translated using Bing translator API. The application then displayed the translation results to the user mobile phone. For testing, this study used Chinese character in various forms of input such as Chinese character that write in paper, road signs, restaurant signs, newspapers, and others. From the tests that carried out, the results show that the application works well for the character shape that has a standard size and shape, while for nonstandard characters and sizes it has a low success rate. From 300 testing, the application success rate is 74%. Therefore, for the future research, this Chinese character image extraction still need to be improve using another method such as Scale Invariant Feature Transform (SIFT) algorithm, Kalman Filtering, deep learning. Based on the testing on degree of slope, the results showed that image input tolerates a degree of slope of 15 degrees. This degree is consider tolerable for a reading purpose.

5. Acknowledgements

This research was supported by Petra Christian University Mobile Computing Research Group.

6. References

[1] Yellow Bridge (n.d.). Yellow Bridge. Retrieved December 17, 2018, from https://www.yellowbridge.com/

[2] Google (n.d.). Google Translate. Retrieved December 17, 2018, from https://translate.google.com/

[3] Qhanzi (n.d.). Input Simplified Chinese characters via handwriting. Retrieved December 17, 2018, from https://www.qhanzi.com/

[4] Bing (n.d.). Bing Translator. Retrieved December 17, 2018, from https://www.bing.com/translator

[5] Tesseract-OCR. (n.d.). *Tesseract-OCR*. Retrieved December 17, 2018, from https://code.google.com/p/tesseract-ocr.

[6] R.B. Panchal, R.G. Sonawane, H. Shaikh, and P.P. Gawli. Design of Text Detection and Translation System for Camera Based Android Smartphone. *International Journal for Scientific Research & Development, Vol. 3, Issue 1*. 2015

[7] K. Pai, G. Pinkyar, M. Pachapur, and V. Mali. Language Translator using Image Processing. *International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 3*, 2016

[8] U.K.D.N. Manisha, and S.R. Liyanage. Sinhala Character Recognition using Tesseract OCR. *3rd International Conference on Advances in Computing and Technology*. 2018.