ISSN 2714-9714

The Effect of Rescaling on the Performance of Recognition with Arabic Characters Using Tesseract OCR Based on Long Short Term Memory

Timur Gagah Prawiro [1*], Arifatul Khasanah [1]

[1]Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Semarang, Indonesia
*Corresponding author: timurgagahprawiro@students.unnes.ac.id

ARTICLE INFO

Article history
Received 13 August 2020
Revised 22 September 2020
Accepted 14 October 2020

Keywords
Tesseract OCR
Arabic Handwritten
Long Short-Term Memory

ABSTRACT

The development of the ability to recognize handwritten character images is one of the branches of science that includes pattern recognition and image processing using Optical Character Recognition (OCR) technology. The performance achieved in the case of Arabic characters is not optimal because of it is cursive nature and relatively severe difficulty. Tesseract OCR Engine is a popular OCR framework that is open source and accurate in character recognition development. The Tesseract OCR Engine works well with images that are 300 dpi (dots per inch). This study focuses on rescaling analysis on recognizing handwritten Arabic characters using Tesseract OCR Engine based Long Short-Term Memory, with scaling sizes 90%, 80%, 70%, and 60% of the source image size. And effect performance on the recognized characters were measured with character accuracy as a method of success. This study used 70 images from publicly available IFN / ENIT image samples.

1 Introduction

The development of information and communication technology has an impact on everyday life, where information technology has made an activity/activity easier, including manual work that will be minimized and changed as much as possible by applying computers (Ryan & Hanafiah, 2015). One of the developments in information and communication technology is computers that are trained to recognize images of handwritten human characters and are converted into ASCII characters so they can be recognized by computers (Mohammad, Anarase, Shingote, & Ghanwat, 2014).

The development of the ability to recognize human handwritten character images is one of the branches of science that includes pattern recognition and image processing using Optical Character Recognition (OCR) technology. Optical Character Recognition (OCR) is the conversion of an image containing a character or a set of characters into a set of characters recognized by a computer machine (Hidayatullah, 2017). However, in the context of handwritten character recognition, some difficulties may occur due to ambiguity, distraction, and large variations in writing style or even the similarity between character entities (Pradeep, Srinivasan, & Himavathi, 2011). In fact, the performance achieved in other characters' case has not been optimal, such as Arabic characters with relatively high cursive and difficulty characteristics (Lamghari, Charaf, & Raghay, 2016).

The Arabic character consists of 28 letters and is written from right to left in cursive. Arabic letters (Alphabet) are used to write various languages such as Persian, Urdu, and Jawi. Each Arabic letter has two or four forms depending on its position in the text (AlKhateeb, 2015). In recognizing handwritten Arabic characters in images, a standard database is needed to determine a good comparison. Some of the research results on recognizing handwritten Arabic characters in images produce high accuracy because the database used is small and not a standard database (AlKhateeb, 2010). An example of a standard database for images of handwritten Arabic characters is the IFN / ENIT database (Pechwitz, Maddouri, Märgner, Ellouze, & Amiri, 2002).

One of the popular frameworks for developing character recognition is Tesseract OCR. Tesseract OCR Engine is a popular OCR framework that is open source and accurate (Abandah, Jamour, & Qaralleh, 2014). Tesseract OCR can be improved with image processing techniques to get more accurate output, such as scaling/rescaling images. The Tesseract OCR Engine works well on images with a size of 300 dpi (dots per inch). This research was applied to 70 images from a sample IFN / ENIT database and focuses on analyzing the effect of rescaling on the performance of character recognition of handwritten Arabic characters using Long Short Term Memory based Tesseract OCR. Figure 1 shows a few examples of the image IFN/ENIT database.

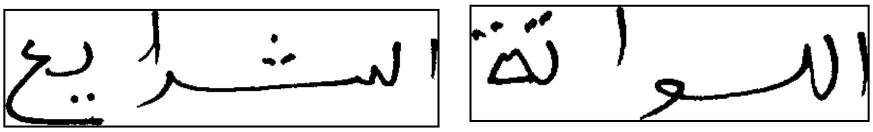


Figure 1. Example image IFN/ENIT database

2 Related Work

Abandah *et al.* (2014) introduce the Arabic handwritten character recognition system in 2014. This development was carried out based on the previous development done by Abandah using the Tesseract OCR Engine. They used segmentation, feature extraction, and the Recurrent Neural Network (RNN) by modifying the Tesseract OCR Engine, which resulted in high accuracy recognition with the smallest error label is 5.5%.

In 2016, Kef *et al.* (2016) developed Arabic handwriting character recognition based on structural characteristics and fuzzy classifier. The proposed method is based on explicit segmentation. The first process of pre-processing is thinning, contour tracing, and connected components detection. Then, the character's features are extracted using invariant pseudo-Zernike moments. Classification using fuzzy ARTMAP neural network and tested on the IFN / ENIT database with an accuracy of 93.8%. Elleuch *et al.* (2016) using an integrated approach between the two types of classifications of Convolutional Neural Network (CNN) and Support Vector Machine (SVM) on the IFN / ENIT and HACDB databases, where the failure in classification was 7.05% and 5.83%.

3 Methodology

The rescaling method applied in this research is based on the percentage scale; the percentage of rescaling that is applied is rescaling 90% of the image size, rescaling 80% of the image size, rescaling 70% of the image size, and rescaling 60% of the image size. Using the OpenCV library is an open-source computer vision library (Pulli, Baksheev, Kornyakov, & Eruhimov, 2012). One of the rescaling functions in OpenCV Imgproc.resize with the INTER_NEAREST scaling type in OpenCV. The following functions are used:

Imgproc.resize(src, dst, size, 0, 0, Imgproc.INTER_NEAREST)

Where	src	: Input image
	dst	: Output image
	size	: Image scaling value
	Imgproc.INTER_NEAREST	: Scaling type

After the rescaling process, the OCR process uses Tesseract OCR based on LSTM, using the tessdata_best data training model, and the page segmentation method "Raw line. Treat the image as a single text line, bypassing hacks that are Tesseract-specific" the 13th option on Tesseract OCR. The measurement of the success of a recognized character is measured using character accuracy. According to Rice *et al*. (1993), character accuracy can be formulated as follows:

$$CA = \frac{(n-(\#error))}{n} \tag{1}$$

Whereas n : The total number of characters
$\#error$: The number of characters not / failed to be recognized.

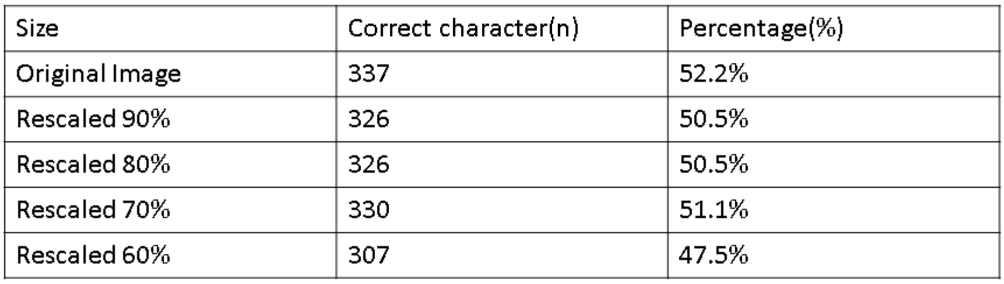
This character accuracy calculates the error value on the OCR. To calculate the OCR performance in percentage, formula 2 is used:

$$(100 - CA) * 100\% \tag{2}$$

Whereas CA : Character Accuracy

4 Result and Conclusion

After the image is rescaled and processed by Tesseract OCR, then the text results are reviewed per character with character accuracy. The original image is also used to compare the results, the total character is 645 characters, and the results are shown in Table 1.



Size	Correct character(n)	Percentage(%)
Original Image	337	52.2%
Rescaled 90%	326	50.5%
Rescaled 80%	326	50.5%
Rescaled 70%	330	51.1%
Rescaled 60%	307	47.5%

Table 1. OCR performance result

According to the results, the original image still gives high accuracy, and the difference in any other rescaled image is 2-4%.

5 Conclusion

Tesseract OCR performs about 50% on handwriting Arabic IFN/ENIT image. Some solution to increase Tesseract OCR's performance is using custom segmentation or retrain tessdata_best as training model of Tesseract OCR.

References

Abandah, G. A., Jamour, F. T., & Qaralleh, E. A. (2014). Recognizing Handwritten Arabic Words using Grapheme Segmentation and Recurrent Neural Networks. *International Journal on Document Analysis and Recognition (IJDAR)*, *17*(3), 275-291. doi:10.1007/s10032-014-0218-7

AlKhateeb, J. H. (2010). *Word-based Offline Handwritten Arabic Classification and Recognition. Design of Automatic Recognition System for Large Vocabulary Offline Handwritten Arabic Words using Machine Learning Approaches* (Doctoral dissertation). University of Bradford, West yorkshire.

AlKhateeb, J. H. (2015). A Database for Arabic Handwritten Character Recognition. *Procedia Computer Science*, *65*, 556-561. doi:10.1016/j.procs.2015.09.130

Elleuch, M., Maalej, R., & Kherallah, M. (2016). A New Design Based-SVM of the CNN Classifier Architecture with Dropout for Offline Arabic Handwritten Recognition. *Procedia Computer Science, 80*, 1712-1723. doi:10.1016/j.procs.2016.05.512

Hidayatullah, P. (2017). *Pengolah Citra Digital Teori dan Aplikasi [Digital Image Processing Theory and Applications]*. Bandung: Penerbit Informatika.

Kef, M., Chergui, L., & Chikhi, S. (2016). A Novel Fuzzy Approach for Handwritten Arabic Character Recognition. *Pattern Analysis and Applications*, *19*(4), 1041-1056. doi: 10.1007/s10044-015-0500-4

Khorsheed, M. S. (2002). Offline Arabic Character Recognition: A Review. *Pattern Analysis & Applications*, *5*(1), 31-45. doi:10.1007/s100440200004

Lamghari, N., Charaf, M. E. H., & Raghay, S. (2016). Template Matching for Recognition of Handwritten Arabic Characters using Structural Characteristics and Freeman Code. *International Journal of Computer Network and Information Security*, 14(12), 31-40. Retrieved from https://www.researchgate.net/publication/312087970_Template_Matching_for_Recognition_of_Handwritten_Arabic_Characters_Using_Structural_Characteristics_and_Freeman_Code

Mohammad, F., Anarase, J., Shingote, M., & Ghanwat, P. (2014). Optical Character Recognition Implementation using Pattern Matching. *International Journal of Computer Science and Information Technologies*, *5*(2), 2088-2090. Retrieved from https://ijcsit.com/docs/Volume%205/vol5issue02/ijcsit20140502254.pdf

Pechwitz, M., Maddouri, S. S., Märgner, V., Ellouze, N., & Amiri, H., (2002, October). IFN/ENIT-Database of Handwritten Arabic Words. In *Proceeding of CIFED*, 2, 127-136. Citeseer. Retrieved from https://www.semanticscholar.org/paper/IFN%2FENIT%3A-database-of-handwritten-arabic-words-Pechwitz-Maddouri/732c94369298f0d3df48ed4035703e56aaf39892

Pradeep, J., Srinivasan, E., & Himavathi, S. (2011, April). Diagonal Based Feature Extraction for Handwritten Character Recognition System using Neural Network. In *2011 3rd International Conference on Electronics Computer Technology*, 364-368. IEEE. doi:10.1109/ICECTECH.2011.5941921

Pulli, K., Baksheev, A., Kornyakov, K., & Eruhimov, V. (2012). Real-Time Computer Vision with OpenCV. *Communications of the ACM*, *55*(6), 61-69. doi:10.1145/2184319.2184337

Rice, S.V., Kanai, J., & Nartker, T. A. (1993). An Evaluation of OCR Accuracy. *Information Science Research Institute, 1993 Annual Research Report*, *9*, 20. Retrieved from https://www.researchgate.net/publication/243780291_An_Evaluation_of_OCR_Accuracy

Roza, E. (2017). Aksara Arab-Melayu di Nusantara dan Sumbangsihnya dalam Pengembangan Khazanah Intelektual [Arabic-Malay Script in Nusantara and its Contribution to Development of Intellectual Property]. *TSAQAFAH*, *13*(1), 177-204. doi:10.21111/tsaqafah.v13i1.982

Ryan, M., & Hanafiah, N. (2015). An Examination of Character Recognition on ID Card using Template Matching Approach. *Procedia Computer Science*, *59*, 520-529. doi:10.1016/j.procs.2015.07.534

Supriana, I., & Nasution, A. (2013). Arabic Character Recognition System Development. *Procedia Technology*, *11*, 334-341. doi:10.1016/j.protcy.2013.12.199