Journal of Applied Engineering and Technological Science
Vol 4(1) 2022 : 58-63

IMPLEMENTATION OF OCR (OPTICAL CHARACTER RECOGNITION) USING TESSERACT IN DETECTING CHARACTER IN QUOTES TEXT IMAGES

Ikha Novie Tri Lestari[1*], Dadang Iskandar Mulyana[2]
STIKOM Cipta Karya Informatika[1,2]
ikhanovie21@gmail.com

Received : 10 August 2022, Revised: 01 September 2022, Accepted : 02 September 2022
*Corresponding Author

ABSTRACT
The development of technology in Indonesia is currently increasingly advanced in people's lives and cannot be avoided. The use of Artificial Intelligence in helping humans in dealing with problems is growing. Humans can take advantage of computer/smartphone media in today's technological era. One of its uses is Optical Character Recognition. This research is motivated by the problem where the running system requires development in terms of technology to detect characters in the quote text image, because the previous system still performs manual input. Optical Character Recognition has been widely used to extract characters contained in digital image media. The ability of OCR methods and techniques is very dependent on the normalization process as an initial process before entering into the next stages such as segmentation and identification. The image normalization process aims to obtain a better input image so that the segmentation and identification process can produce optimal accuracy. To get maximum results, it takes several pre-processing stages on the image to be used. To achieve this, it is necessary to perform Optical Character Recognition which can be done using Tesseract-OCR. The OCR program that was created was successfully used to scan or scan a quote text image if the document was lost or damaged, and it could save time for creating, processing and typing documents.
Keywords: *Optical Character Recognition, Tesseract, Quotes*

1. Introduction

The development of a very practical era makes people more inclined to find a fast way to do something. In modern times, technology has become a part of everyday human life. Technology has been able to help humans in carrying out their daily activities. Along with these developments, mobile applications are also growing rapidly and are in great demand by various groups of people because of their ease of use, and their nature that can be used anywhere (Andreas et al., 2020; Hamzah et al., 2022; Sintia et al., 2021).

The development of technology in Indonesia is currently increasingly advanced in people's lives and cannot be avoided. The use of Artificial Intelligence in helping humans in dealing with problems is growing (Trilaksono et al., 2008). Humans can take advantage of computer/smartphone media in today's technological era. One of the uses is Optical Character Recognition (Bagus et al., 2017; Li, et al., 2021).

This research is motivated by a problem where a system that requires development in terms of technology to detect characters in image images, one example is the detection of text in quote images, because the previous system still input text manually. Optical Character Recognition has been widely used to extract characters contained in digital image media (Qashlim et al., 2022; Pino, et al., 2021; Nazaruddin, 2022).

The ability to OCR methods and techniques is very early in the normalization process as a process before entering the next stages such as segmentation and surprise. The image normalization process aims to obtain a better input image so that the segmentation process can produce optimal accuracy (Siregar, 2019; Phoenix, et al., 2021).

To get maximum results, several pre-processing stages are needed on the image to be used. To achieve this, it is necessary to perform Optical Character Recognition which can be done using Tesseract-OCR (Aprilianto Susanto & Richard Beeh, 2015). The OCR program that has been created has been successfully used to scan or an image of a text quote if the document is lost or damaged, and can save time in creating, creating and typing documents (Firdaus et al., 2021;

Thorat, et al., 2022). Therefore, the authors are interested in conducting research related to the OCR program to detect text characters in image quotes.

2. Research Methods

The method in implementing this research consists of the following stages:

1. Pre-Processing
   This pre-processing stage aims to get a single character from a scanned text in a good and clean condition so as to facilitate the recognition process. According to Bieniecki et al, preprocessing begins with normalizing the condition of the text by eliminating noise such as dots and correction of text image orientation, binaryization, and segmentation. This stage if done correctly will increase the character recognition accuracy ratio(Mamuriyah & Jacky, 2021).
2. Feature Extraction Stage
   The purpose of feature extraction is to find the attributes of the most important character patterns and different from other characters so that they can be classified. The role of humans is to determine and select features that enable an efficient and effective recognition process. The question is then what can be used as a feature for a set of letters or alphabets in a particular language writing system (Marizal, 2022).
3. Introduction Process
   When character patterns have been mapped into vector values, the next problem is how to group characters that have the same or almost the same vector values. This problem is solved by classification. This series of vector value classification processes is known as the character recognition process. Thus the character recognition process deals with and is at the level of character representation.
4. Post-Process
   Every OCR system built with the most sophisticated algorithms often makes mistakes, in the sense that not all characters read are converted to their equivalent characters. For this reason, the post-process character matching stage is carried out to increase the accuracy of character recognition. This post-processing system is also known as the correction process because this module is tasked with correcting errors that are often made at the word level.

3. Results and Discussions

In previous research, it was stated that OCR is a method for converting handwriting in digital form so that it can be updated, and consists of three main processes including pre-processing, recognition and post-processing. OCR also serves to distinguish one character from another in a digital image (Sanjaya et al., 2019). Another study states that OCR is a process of converting images into text that can be updated on a computer, where the text and numbers in the image cannot be changed because the character consists of an arrangement of pixel dots that form the image display of text and numbers. OCR is classified into two types, namely offline recognition and online recognition (Sandhika, 2014). In the offline OCR type, the test image can be scanned from a document, while online OCR is consecutive points represented as a function of time and there is a sequence of the image patterns (Fauzan & Wibowo, 2021).

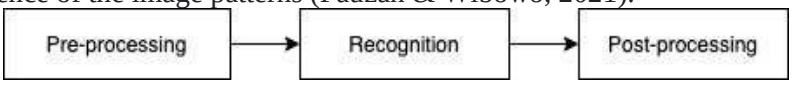

Fig 1. Optical Character Recognition

a. Pre-processing: this stage serves to improve accuracy in character recognition which includes binaryization, noise reduction, normalization of ratio and scale, and image cropping.
b. Character recognition: there are four main types of OCR algorithms, namely Template Matching, Statistical Approaches, Structural Analysis, and Neural Networks.
c. Post-processing: the stage of processing data from character recognition results for further processing.

One approach to the OCR method in detecting characters can be done using Template Matching (TM) by calculating the smallest error value to determine the level of character match from the input image (Ashar et al., 2020).

$$\min e = \sum (Ix,y - Tx,y)\ 2$$

Where the value of I is the pixel value of the image that will be matched with the pixel value of the template (T). The match between the input image and one of the template images is calculated based on the smallest error value.

Tesseract OCR is a library used to detect characters in binary images based on pixel distribution analysis and training (Sahertian et al., 2020). How Tesseract OCR works can be seen in the following image:

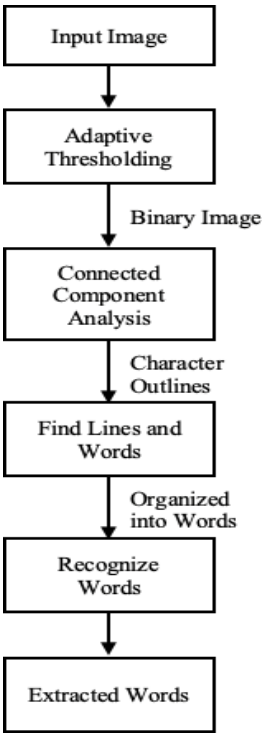

Fig 2. How Tesseract Engine Works

$$(A, B) = 1\ m{\cdot}n \sum \sum ([i,] - B[i,j])\ m{-}1\ 2$$

Tesseract considers the input image to be a binary image, the first step is component analysis that is connected to the stored component outline or template. At this stage all outlines are collected arranged in the form of blobs. Blobs are arranged in the form of text lines, where the areas and lines are analyzed and corrected into a proportional text form. Lines of text are broken down into words based on the type of character spacing using definite spaces and fuzzy spaces (Kevin Wiguna et al., 2019).

The recognition stage is then continued in a stage known as adaptive recognition using letter shape recognition with a high level of confidence (first pass). Furthermore, the remaining characters in the previous stage will be recognized better at the next stage (second pass). Tesseract uses libraries to improve accuracy at the character segmentation stage.

Fig 3. Image dataset to be extracted

```
!pip install Pillow==9.0.0
image_path_in_colab="/content/gdrive/MyDrive/quotes_ikha.jpeg"
extract = pytesseract.image_to_string(Image.open(image_path_in_colab))
print(extract)
#type(extract)!
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: Pillow==9.0.0 in /usr/local/lib/python3.7/dist-packages (9.0.0)
Oprah Winfrey

Tantangan adalah hadiah yang
memaksa kita untuk mencari
pusat gravitasi baru. Jangan
melawan mereka. Temukan saja
cara baru untuk berdiri
```

Fig 4. Image extraction process using Tesseract OCR

```
extract=extract.split('\n')

extract
```

```
['Oprah Winfrey',
 '',
 'Tantangan adalah hadiah yang',
 'memaksa kita untuk mencari',
 'pusat gravitasi baru. Jangan',
 'melawan mereka. Temukan saja',
 'cara baru untuk berdiri',
 '\x0c']
```

Fig 5. The Process of split a string by newline

```
l=list()
for i in extract:
  if i!='' and i!=' ':
    l.append(i)
l
```

```
['Oprah Winfrey',
 'Tantangan adalah hadiah yang',
 'memaksa kita untuk mencari',
 'pusat gravitasi baru. Jangan',
 'melawan mereka. Temukan saja',
 'cara baru untuk berdiri',
 '\x0c']
```

Fig 6. The Process of inputting each line of a sentence into a variable list

4. Conclusion

Based on the description in the previous chapter, the authors can conclude that: The OCR (Optical Character Recognition) program can be used to detect characters in the quote text image.

To get maximum results, it takes several pre-processing stages on the image to be used. To achieve this, it is necessary to perform Optical Character Recognition which can be done using Tesseract-OCR. The image normalization process aims to obtain a better input image so that the segmentation and identification process can produce optimal accuracy. The OCR program created was successfully used to scan or scan a quote text image. The OCR (Optical Character Recognition) program can save time on document creation, processing and typing.

References

Andreas, Y., Gunadi, K., & Purbowo, A. N. (2020). Implementasi Tesseract OCR untuk Pembuatan Aplikasi Pengenalan Nota pada Android. *Jurnal Infra*, *8*(1), 312-317.

Ashar, M. K., Setyawan, G. E., & Setiawan, E. (2020). Navigasi Robot Beroda Berdasarkan Pengenalan Teks untuk Melakukan Pergerakan Menggunakan Metode Optical Character Recognition (OCR). *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *4*(4), 1150–1159.

Fauzan, M. R., & Wibowo, A. P. W. (2021). Pendeteksian Plat Nomor Kendaraan Menggunakan Algoritma You Only Look Once V3 Dan Tesseract. *Jurnal Ilmiah Teknologi Infomasi Terapan*, *8*(1), 57–62. https://doi.org/10.33197/jitter.vol8.iss1.2021.718

Firdaus, A., Syamsu Kurnia, M., Shafera, T., Firdaus, W. I., Teknik, J., Politeknik, K., & Sriwijaya -Palembang, N. (2021). Implementasi Optical Character Recognition (OCR) Pada Masa Pandemi Covid-19 *1. In *Jurnal JUPITER*. 13(2).

Hamzah, M. L., Rahmadhani, R. F., & Purwati, A. A. (2022). An Integration of Webqual 4.0, Importance Performance Analysis and Customer Satisfaction Index on E-Campus. *Journal of System and Management Sciences, 12*(3), 25-50. https://doi.org/10.33168/JSMS.2022.0302

Kevin Wiguna, A., Suciati, N., & Khotimah, W. N. (2019). Aplikasi Penerjemah Gambar Teks Berbahasa Inggris Menggunakan Teknologi Realitas Tertambah pada Perangkat Berbasis Android. *Jurnal Teknik ITS*, *8*(1). https://doi.org/10.12962/j23373539.v8i1.40070

Li, M., Lv, T., Cui, L., Lu, Y., Florencio, D., Zhang, C., ... & Wei, F. (2021). Trocr: Transformer-based optical character recognition with pre-trained models. *arXiv preprint arXiv:2109.10282*.

Mamuriyah, N., & Jacky, J. (2021). Perancangan dan Pembuatan Alat untuk Mendeteksi Teks Hangul dan Inggris pada Menu Makanan Menggunakan metode OCR (Optical Character Recognition). *Telcomatics*, *6*(1), 1–10. https://doi.org/10.37253/telcomatics.v6i1.5054

Marizal, M. (2022). Classification of The Risk of Comorbid Covid-19 Patient at Bengkalis Hospital Using Bayesian Binary Logistics Regression . *Journal of Applied Engineering and Technological Science (JAETS), 3*(2), 168–177. https://doi.org/10.37385/jaets.v3i2.812

Nafsin, M., Qashlim, A., & Khairat, U. (2022, May). Sistem Informasi Data Siswa Berbasis Ocr (Optical Character Recognition) Pada Smk Bina Harapan. In *Journal Peqguruang: Conference Series*. 4(1), 412-417.

Nazaruddin, N. (2022). Implementation of Quality Improvements to Minimize Critical to Quality Variations in Polyurethane Liquid Injection Processes. *Journal of Applied Engineering and Technological Science (JAETS), 3*(2), 139–148. https://doi.org/10.37385/jaets.v3i2.771

Okta, M. D. U., Aulia, S., & Burhanuddin, B. (2021). Pengenalan Pola Berbasis OCR untuk Pengambilan Data Bursa Saham. *Jurnal Rekayasa Elektrika*, *17*(2), 100–106. https://doi.org/10.17529/jre.v17i2.19656

Phoenix, P., Sudaryono, R., & Suhartono, D. (2021). Classifying promotion images using optical character recognition and Naïve Bayes classifier. *Procedia Computer Science*, 179, 498-506.

Pino, R., Mendoza, R., & Sambayan, R. (2021). Optical character recognition system for Baybayin scripts using support vector machine. *PeerJ Computer Science*, 7, e360.

Sahertian, J., Khotmuniza, M. I., & Helilintar, R. (2020). Sistem Parkir Menggunakan Ocr (Optical Character Recognition) Plat Nomer Dan Iot (Internet of Things). *Joutica*, *5*(2), 363. https://doi.org/10.30736/jti.v5i2.443

Sandhika, R. (2014). *KITAB FIQIH SAFINAH AN-NAJA.*

Sanjaya, E., Prasetiadi, A., & SAPUTRA, W. A. (2019). Klasifikasi Analisis Sentimen Pada Gambar Meme Politik Dengan Library Tesseract Dan Algoritme Support vector machine. *Journal of Informatics, Information System, Software Engineering and Applications (INISTA)*, *2*(1), 56–64. https://doi.org/10.20895/inista.v2i1.96

Sintia, S., Defit, S., & Nurcahyo, G. W. (2021). Product Codefication Accuracy With Cosine Similarity And Weighted Term Frequency And Inverse Document Frequency (TF-IDF) . *Journal of Applied Engineering and Technological Science (JAETS)*, *2*(2), 62–69. https://doi.org/10.37385/jaets.v2i2.210

Siregar, R. (2019). *Implementasi OTSU Thresholding pada Optical Character Recognition Menggunakan Engine Tesseract*.

Susanto, F. A., & Beeh, Y. R. (2015). Pemanfaatan Teknologi Optical Character Recognition (OCR) Untuk Mengenali Alfabet Yunani Berbasis Android. *Artikel Ilmiah Teknologi Informasi. Universitas Kristen SatyaWacana. Salatiga*.

Thorat, C., Bhat, A., Sawant, P., Bartakke, I., & Shirsath, S. (2022). A Detailed Review on Text Extraction Using Optical Character Recognition. *ICT Analysis and Applications*, 719-728.

Widja, I. B. P. (2017). Rancangan Binarisasi Citra dan Pengenalan Karakter Teks Dengan Raspberry Pi. *E-Proceedings KNS&I STIKOM Bali*, 766-771..

Trilaksono, M. (2008). Implementasi Optical Character Recognition (Ocr) Dengan Pendekatan Metode Struktur Menggunakan Ekstraksi Ciri Vektor Dan Region IT Telkom.