Balinese character recognition on mobile application based on tesseract open source OCR engine

I M D R Mudiarta¹, I M D S Atmaja², I K Suharsana³, I W G S Antara⁴, I W P Bharaditya⁵, G A Suandirat⁶, G Indrawan⁷

¹,²,³,⁴,⁵,⁶,⁷Dept. of Computer Science Universitas Pendidikan Ganesha Singaraja, Indonesia

¹ dwiraka@gmail.com, ² mdwi.suta@gmail.com, ³ suharsana@gmail.com, ⁴ suweca@gmail.com, ⁵ bharadityap@gmail.com, ⁶ suandirat@gmail.com, ⁷ gindrawan@undiksha.ac.id

Abstract. Balinese script is a part of Balinese culture which is rarely used today. The Provincial Government of Bali with the Governor Regulation number 80 of 2018 is trying to preserve the Balinese language and script. This study aimed at preserving the Balinese script through a mobile technology approach which is the recent trend with worldwide coverage for supporting ubiquitous learning. This research integrated the Android application to recognize Balinese characters in the form of images into text with Tesseract open source Optical Character Recognition (OCR) engine. The input of this application is a Balinese script image captured by a mobile camera or from a Balinese script image. The application recognized input image into text that can be further processed based on training data available in the application. The new Balinese script training data was created based on eighteen Balinese script’s basic syllables and numbers only. This application can be operated offline with mobile hardware that supports camera functions. The result for testing for 50-word, recognition was 62% obtained in good quality image-based Bali-Simbar font. This application can be further developed to recognize other character repertoire i.e., vowels (Akśara Suara), semi vowels (Arda Suara), additional syllables (Akśara Šwalalita), and sound killers (Pangangge Tengenan).

1. Introduction
Balinese script is one of Balinese culture which is used to write Balinese language. This culture is rarely used in society for daily activities[1]. Because it is rarely used, Balinese script has now begun to be endangered[2].

The Provincial Government of Bali has made a very good breakthrough to prevent the extinction of Balinese script by issuing Bali Governor Regulation Number 80 and Circular Letter No. 3172 of 2019 concerning Protection and Use of Balinese Language, Literacy and Literature and the Implementation of the Balinese Language Month. But the breakthrough alone was felt not enough because it had not touched the wider community to use Balinese script in their daily lives.
This research focuses on preserving the knowledge of reading Balinese script in pictures by combining information technology with Balinese discipline. In this study an OCR application was developed that runs on mobile devices with camera facilities. Input on this application is in the form of images and is processed with Tesseract OCR technology. The resulting output is in Balinese script text which can be copied and used for further purposes.

2. Balinese Script
Balinese script is one of the traditional scripts in Indonesia that evolve on the island of Bali, Indonesia. Balinese script is a sign or symbol used by Balinese people to write Balinese language. Based on its function the Balinese script is divided into four namely Wreastra Script, Swalelita Script, Wijaksara Script, Modre Script. Based on its function, Balinese script has two types of scripts namely Ordinary Balinese Script and Holy Balinese script[3].

Ordinary Balinese Script is a Balinese script which is used to write everyday language, which is divided into the Wreastra script and the Swalelita script. The Holy Balinese script is the Balinese script used for writing things related to religion such as japa mantra, vedas, rerajahan. Based on its form, Balinese script is divided into Basic Balinese Script, Balinese Derivative Script, and Balinese Script with symbol[3]. In this case we used eighteen Balinese script’s basic syllables[4].

The Balinese use decimal system for numbers. There is a simple one-to-one mapping to the Arabic digits.

| Latin | Balinese |
|-------|---------|
| ha    | ḫa     |
| na    | ṇa     |
| ca    | ᱗a     |
| ra    | ḕa     |
| ka    | ᵇa     |
| da    | ᶜa     |
| ta    | ᶘa     |
| sa    | ᵉa     |
| wa    | ᶜa     |

| Latin | Balinese |
|-------|---------|
| la    | ḫa     |
| ma    | ṇa     |
| ga    | ᱗a     |
| ba    | ᵇa     |
| nga   | ᶜa     |
| pa    | ᶘa     |
| ja    | ᶘa     |
| ya    | ᶜa     |
| nya   | ᶘa     |

| Latin | Balinese |
|-------|---------|
| 0     | ᴿ     |
| 1     | ᴸ     |
| 2     | ᴹ     |
| 3     | ᴺ     |
| 4     | ᴻ     |
| 5     | ᴻ     |
| 6     | ᴻ     |
| 7     | ᴻ     |
| 8     | ᴻ     |
| 9     | ᴻ     |

Figure 1. Balinese Basic Syllables and Numeral.
3. Tesseract OCR Engine

The Tesseract OCR engine was originally developed as software in the Hewlett Packard laboratories in Bristol, England and Greeley, Colorado between 1985 and 1994, with some changes made in 1996 to port to Windows, and some migration from C to C++ to 1998 [5]. Software was released as open source in 2005 by Hewlett Packard and the University of Nevada, Las Vegas (UNLV). The development of Tesseract has been sponsored by Google since 2006 [6]. The latest version of this software is Tesseract v4.1.0, which was released on July 7, 2019.

Tesseract is a widely used OCR software that supports a variety of languages, and provides library features [7]. In addition, he also provides training methods for development purposes [5]. So, users can train the tesseract library for the purpose of reading new characters that are more specific.

Tesseract OCR has several main processes [6], namely:

- Adaptive Thresholding: this process converts images into binary images.
- Connected component analysis: in this process extracts character margins. This process is useful for reading images with a dark background and light or white text.
- Find lines and words: in this process segmentation is done per character called Blobs and combines into words to be read.
- The formed blobs are joined together again to form a line of text to be read.
- Recognition is divided into two processes:
  - The first stage focuses on reading letter by letter which is read by the classifier method.
  - The second stage focuses on reading word by word that is not read in the first stage by the same method so that the most possible words are obtained.

Figure 2. Tesseract OCR engine process [8].
4. Method
This study consisted of three main stages which were the process of developing character recognition applications using the Tesseract OCR engine. The first process was to conduct training data that would be used on OCR for character recognition. Training was conducted on basic syllables and Balinese script numeral.

The second process was to build a mobile application that can read Balinese script offline with camera facilities. The last process was testing the applications that have been made with some test data.

5. Results and Discussion

5.1. Build Balinese Training Data for Tesseract OCR

5.1.1. Box Generation. The first step in preparing training data was to form box files. Box files are collections of different characters whose shapes can represent certain characters according to the character set. Making box files in this study using jTessBoxEditor. jTessBoxEditor was a tool to create box files as well as a trainer tool from tesseract developed by third parties. This tool has fully automatic facilities for training data. This program runs on Java Runtime Environment 7 or higher. This tool accepts input in the form of text files containing text or characters that are used as training data. At the time of text input, we must determine the type of letter, font size, prefix of the language we will be training, and the name of the tiff file and file box that will be created. Input text file can be seen in the Figure 3.

![Box file generation from text input.](image)

**Figure 3.** Box file generation from text input.

The Box Editor Tab can display and adjust data box files so that the segmentation of each character is more accurate. The Box Editor tab can be seen in the Figure 4.
5.1.2. Creation of trained data file. The pre-formed box files are the files needed to conduct further training data. The training was processed in the Trainer Tab menu. In this tab, there are file location settings for training data, language prefixes to be trained, and training method options. The training process is shown in Figure 5.

5.2. Mobile Application
This application was developed with Android Studio using the Tesseract API for Android. Image processing and text recognition are performed by the API. The application interface can be seen in the Figure 6.
5.3. Application Testing

In the application testing we used image with single character or number and 50 words from the Bali - Indonesia Dictionary which contained Balinese and Latin scripts published by the Denpasar City Cultural Office in collaboration with the Bali Language, Literacy and Literature Board of Trustees in 2008[9]. The 50 sampling words purely used basic syllables it was limited to eighteen Balinese script’s basic syllables and numbers only. Image for testing create in text editor with white background and black font colour, screenshot saved in jpg image format.

| Character       | Quantity | Correct | Wrong | Accuracy |
|-----------------|----------|---------|-------|----------|
| Single numeral  | 10       | 10      | 0     | 100%     |
| Single character| 18       | 18      | 0     | 100%     |

**Figure 6.** Android mobile OCR application.
Based on the result for testing for 50-word, recognition was 62% obtained in good quality image-based Bali-Simbar font. We remarked many recognition errors, for example:

- Character “da” was recognized as “sa”
- Character “sa” is recognized as “da”
- Character “ka” was recognized as “na”
- Character “ra” was recognized as “ja”

6. Conclusions and Future Work

Many previous works on Tesseract OCR were successful to recognize character of Asian language that was very different from Latin character. Balinese character is the one of Asian character which is not recognized yet with OCR engine. This work was successful to recognize basic syllables of Balinese character.

This OCR mobile application is based on open source Tesseract OCR engine, which get input from capture/insert images entered by the user, processing them into data that can be compared to training data that was previously added to the application. The results of this comparison were used as an output text for the recognition of this Balinese script. In this paper, we conducted the optical character recognition of Balinese Script.

Mobile application we developed based on Tesseract engine could successfully recognize Balinese script with single character and white background in image. Besides, Balinese word with basic syllables recognition was successfully recognized in 62% with 50 words sampling data.

The mobile application was developed in which image input was not only from files but also more useful directly from the camera. For future development, it is expected that the mobile application can recognize more Balinese characters including vowels (Akśara Suara), Akśara Śwalalita, sound killers (Pangangge Tengenan), punctuations, ligatures, and miscellaneous glyph.

References

[1] G Indrawan, I K Paramarta, K Agustini, and Sariyasa, 2018 Latin-to-Balinese script transliteration method on mobile application: A comparison Indones. J. Electr. Eng. Comput. Sci., vol. 10, no. 3, pp. 1331–1342
[2] G. Indrawan, I K Paramarta and K Agustini 2019 A new method of Latin-to-balinese script transliteration based on noto sans balinese font and dictionary data structure ACM Int. Conf. Proceeding Ser. pp 75–79
[3] W R I B Made Suasta, I B Mayun 1996 MODERNISASI DAN PELESTARIAN PERKEMBANGAN METODE DAN TEKNIK PENULISAN AKSARA BALI Departemen Pendidikan dan Kebudayaan Republik Indonesia
[4] I A D P Sari, B Hidayat, and U Sunarya,2015 Pengenalan Aksara Bali Dengan Metode Local Binary Pattern, eProceedings Eng. vol 2 no 2
[5] R Smith 2007 Tesseract OCR Engine Lect. Google Code. Google Inc
[6] R Smith 2007 An overview of the tesseract OCR engine in Proceedings of the International Conference on Document Analysis and Recognition ICDAR
[7] R Smith, D Antonova, and D S. Lee 2009 Adapting the Tesseract open source OCR engine for
multilingual OCR in *ACM International Conference Proceeding Series*

[8] K El G Mohammed and C Amazighe 2015 *Training TESSERACT Tool for Amazigh OCR* no MAY, pp. 172–179

[9] Tim Penyusun 1978 *Kamus Bali-Indonesia Dinas Pengajaran Provinsi Drh. Tingkat I Bali*