ABSTRAK

Asroni, Ahmad (2024), Implementasi Hirarki Dataset Dalam Membangun Model Language Aksara Bali Menggunakan Framework Tesseract OCR.
Tesis, Ilmu Komputer, Program Pascasarjana Universitas Pendidikan Ganesha.

Tesis ini sudah disetujui dan diperiksa oleh Pembimbing I : Dr. Gede Indrawan, S.T., M.T. dan Pembimbing II : Dr. Luh Joni Erawati Dewi, S.T., M.Pd.

Kata kunci**:** Aksara Bali, *Optical Character Recognition*, *Tesseract OCR*, *Web Scraping*, *Mobile*

Salah satu faktor utama yang menyebabkan penurunan penggunaan Aksara Bali adalah masyarakat Bali kurang tertarik untuk membaca Aksara Bali karena keengganan dalam mempelajari Aksara Bali yang relatif rumit dalam proses pengenalannya. Perkembangan teknologi komputer saat ini telah banyak dimanfaatkan untuk melakukan pengenalan karakter optik atau diistilahkan dengan OCR (*Optical Character Recognition*). Pada penelitian ini dilakukan eksperimen menggunakan Tesseract OCR yaitu salah satu engine OCR terpopuler. Proses eksperimen yang dilakukan terdiri dari beberapa tahapan yaitu pertama melakukan persiapan dataset, kedua melakukan menggunakan metode *Web Scraping* untuk melakukan generate dataset, ketiga tahap training dataset, dan tahapan terakhir adalah melakukan implementasi model language ke dalam aplikasi *mobile*. Hasil penelitian membuktikan bahwa proses generate menggunakan metode *Web Scraping* dataset dapat menjadi pilihan lebih baik jika diperhadapkan dengan training dataset yang memerlukan dataset yang besar dibandingkan dengan beberapa penelitian sebelumnya yang sejenis dalam pengenalan karakter nol-latin. Model language terbaik yang dihasilkan adalah kombinasi hirarki dataset karakter, kata, kalimat dan paragraf (*Combination Hierarchy of Character, Word, Sentence, and Paragraph Datasets*) dengan tingkat *coincidence* sebesar 66.67%. Hirarki dataset tersebut memperoleh tingkat *coincidence* paling tinggi dibandingkan dua jenis hirarki dataset yang lain yaitu kombinasi dataset secara acak (*Random Dataset Combination Hierarchy*) dengan tingkat *coincidence* sebesar 25% dan hirarki dataset per karakter (*Single Character Dataset Combination Hierarchy*) dengan tingkat *coincidence* sebesar 40%. Semakin beragam dan terstruktur hirarki dataset yang digunakan maka akan memberikan peningkatan tingkat *coincidence*. Hasil penelitian menunjukkan bahwa tingkat *coincidence* masih jauh dari optimal, memerlukan perhatian pada karakteristik dataset yang terbatas pada penggunaan synthetic data images.

ABSTRACT

Asroni, Ahmad *(2024), Implementation of Dataset Hierarchy for Building a Balinese Script Language Model Using the Tesseract OCR Framework. Thesis, Computer Science, Postgraduate Program Ganesha University of Education.*

This thesis has been approved and checked by Preceptor I: Dr. Gede Indrawan, S.T., M.T. and preceptor II: Dr. Luh Joni Erawati Dewi, S.T., M.Pd.

Keywords: Balinese Script, Optical Character Recognition, Tesseract OCR, Web Scraping, Mobile

One of the main factors causing the decline in the use of Balinese Script is that Balinese people are less interested in reading Balinese Script due to reluctance in learning Balinese Script which is relatively complicated in the recognition process. The development of computer technology today has been widely used to perform optical character recognition or termed OCR (Optical Character Recognition). In this research, experiments were conducted using Tesseract OCR, which is one of the most popular OCR engines. The experimental process consists of several stages, the first is to prepare the dataset, the second is to use the Web Scraping method to generate the dataset, the third is to train the dataset, and the last stage is to implement the model language into the mobile application. The results prove that the process of generating using the Web Scraping dataset method can be a better choice if faced with a training dataset that requires a large dataset compared to some similar previous studies in zero-latin character recognition. The best language model produced is a hierarchical combination of character, word, sentence, and paragraph datasets (Combination Hierarchy of Character, Word, Sentence, and Paragraph Datasets) with a coincidence rate of 66.67%. The dataset hierarchy obtained the highest coincidence rate compared to the other two types of dataset hierarchies, namely a random dataset combination (Random Dataset Combination Hierarchy) with a coincidence rate of 25% and a single character dataset hierarchy (Single Character Dataset Combination Hierarchy) with a coincidence rate of 40%. The more diverse and structured the dataset hierarchy used, the higher the coincidence rate. The results show that the coincidence rate is still far from optimal, requiring attention to the characteristics of the dataset which is limited to the use of synthetic data images.