PROJECT REPORT

Text detection and text extraction on images with Tesseract OCR

WILLIAM KAMDESU SAMOSIR
16.K1.0032

Faculty of Computer Science
Soegijapranata Catholic University
2021

APPROVAL AND RATIFICATION PAGE

HALAMAN PENGESAHAN

Judul Tugas Akhir:	:	Text detection and text extraction on images with Tesseract OCR
Diajukan oleh	:	William Kamdesu Samosir
NIM	:	16.K1.0032
Tanggal disetujui	:	21 Januari 2021
Telah setujui oleh		
Pembimbing	:	R. Setiawan Aji Nugroho S.T., MCompIT., Ph.D
Penguji 1	:	Rosita Herawati S.T., M.I.T.
Penguji 2	:	R. Setiawan Aji Nugroho S.T., MCompIT., Ph.D
Penguji 3	:	Y.b. Dwi Setianto
Penguji 4	:	Hironimus Leong S.Kom., M.Kom.
Penguji 5	:	Yonathan Purbo Santosa S.Kom., M.Sc
Ketua Program Studi	:	Rosita Herawati S.T., M.I.T.
Dekan	:	R. Setiawan Aji Nugroho S.T., MCompIT., Ph.D

Halaman ini merupakan halaman yang sah dan dapat diverifikasi melalui alamat di bawah ini.

sintak.unika.ac.id/skripsi/verifikasi/?id=16.K1.0032

STATEMENT OF ORIGINALITY

I, the undersigned:

Name :WILLIAM KAMDESU SAMOSIR

ID : 16.K1.0032

Certify that this project was made by myself and not copy or plagiarize from other people, except that in writing expressed to the other article. If it is proven that this project was plagiarizes or copy the other, I am ready to accept a sanction.

Semarang, January, 21, 2021

METERAI TEMPEL
D344FAFF493339931
6000
ENAM RIBU RUPIAH

METERAI TEMPEL
5F4B8ADC432179367
3000
TIGA RIBU RUPIAH

WILLIAM KAMDESU SAMOSIR
16.K1.0032

APPROVAL PAGE FOR PUBLICATION OF SCIENTIFIC PAPERS FOR ACADEMIC INTEREST

The undersigned below :

Name : WILLIAM KAMDESU SAMOSIR

Undergraduate Program : TECHNICAL INFORMATION

Faculty : COMPUTER SCIENCE

Type of work : SKRIPSI

Approved to give Non-Exclusive Royalty Free Right to Soegijapranata Catholic University Semarang for scientific work entitled **"Text detection and text extraction on images with Tesseract OCR"** along with the existing tools (if needed). With this Non-Exclusive Royalty Free Right Soegijapranata Catholic University has the right to store, transfer data / format, manage in the form of a database, maintain, and publish this final project as long as I keep my name as a writer / creator and as a Copyright owner.

This statement I made in truth.

Semarang, January 21, 2021

Sincerely

WILLIAM KAMDESU SAMOSIR

ACKNOWLEDGEMENTS

First and foremost, thank God Almighty for His blessings, inclusion and love from the beginning and the end of this final project. The final project is intended as one of the requirements to take the Bachelor of Computer Science exam in the Informatics Engineering Study Program at Soegijapranata Catholic University of Semarang.

In the preparation of this final project, the author received assistance from various parties. Therefore, on this occasion the author would like to express his thanks to:

1. Mr. Gondo Wiyanto and Mrs. Susi for all the assistance given to college.

2. Father, Mother, Sister and grandmother who always provide prayer support and motivation.

3. Families as co-workers of San'en Semarang who have been kind enough to provide time, thought and energy assistance in the work.

4. R. Setiawan Aji Nugroho S.T., McompIT., Ph.D for his kindness as a supervisor and lecturer who has provided guidance and input to the author, so that the final project can be completed properly.

5. Kevin, Julius, Ryan and Bagas friends and colleagues for motivational assistance in the construction and completion of the final project.

6.And other parties which the author cannot mention one by one and have providedsupport and assistance provided to the author during the preparation of this final project.

Semarang, January 21, 2021

WILLIAM KAMDESU SAMOSIR

ABSTRACT

In this day, documents is very important. Documents can be in the form of archives in notes or in the form of typed files, for documents in the form of notes, its is usually in print out or handwriting. For documents in printed or handwritten form, they usually have difficulty in the storage process because documents documents records can be damage, for example such as faded print ind and easily torn print paper. In modern time this can be overcome by using OCR(Optical Character Recognition), which is image processing that can detect text and text exctraction into a documents file format that can be edited and stored on a computer device for easier storage of documents text.

OCR (optical character recogniton) is an image processing that can detect text and text extraction. through OCR (optical character recognition), the text of the document will be processed using the LSTM (long short term memory) algorithm to perform text detection and text extraction. LSTM (long short term memory) image will be pre-processed using tresholding which will help the process of detecting text. then the image will be processed in convolutional which will turn the image into a matrix, then the batch normalization process is carried out to add stability to the neural network (CNN). After that using Leaky Relu (Leaky Rectified Liniear Unit) is a type of activation function based on a ReLU, but it has a small slope for negative values instead of a flat slope as layer function , max pooling layer as the output or the final result of the detection. The image detected by the text character will be extracted into a document format in the form of a .txt file which is ready to be processed and stored.

Based on the final results of OCR (optical character recognition) using the LSTM (long short term memory) algorithm, it has a satisfactory level of accuracy for text detection, while the process speed in recognizing character letters is good enough. The detected language recognition still has limitations due to the written character of the language

Keyword: OCR, LSTM, CNN, Text detection, Text exctraction

TABLE OF CONTENTS

ILLUSTRATION INDEX

INDEX OF TABLES