



PROJECT REPORT
IMAGE PRE PROCESSING FOR TESSERACT OCR

GOEI, STEVEN CHRISTIAN SUGIHARTO
17.K1.0016

Faculty of Computer Science
Soegijapranata Catholic University
2021



HALAMAN PENGESAHAN

Judul Tugas Akhir : IMAGE PRE PROCESSING FOR TESSERACT OCR

Diajukan oleh : Goei, Steven Christian S

NIM : 17.K1.0016

Tanggal disetujui : 11 November 2021

Telah setuju oleh

Pembimbing : Hironimus Leong S.Kom., M.Kom.

Penguji 1 : R. Setiawan Aji Nugroho S.T., MCompIT., Ph.D

Penguji 2 : Hironimus Leong S.Kom., M.Kom.

Penguji 3 : Rosita Herawati S.T., M.I.T.

Penguji 4 : Y.b. Dwi Setianto S.T., M.Cs.

Penguji 5 : Yulianto Tejo Putranto S.T., M.T.

Penguji 6 : Yonathan Purbo Santosa S.Kom., M.Sc

Ketua Program Studi : Rosita Herawati S.T., M.I.T.

Dekan : Dr. Bernardinus Harnadi S.T., M.T.

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

sintak.unika.ac.id/skripsi/verifikasi/?id=17.K1.0016

DECLARATION OF AUTHORSHIP

I, the undersigned:

Name : GOEI, STEVEN CHRISTIAN SUGIHARTO

ID : 17.K1.0016

declare that this work, titled " IMAGE PRE PROCESSING FOR TESSERACT OCR", and the work presented in it is my own. I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at Soegijapranata Catholic University
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
3. Where I have consulted the published work of others, this is always clearly attributed.
4. Where I have quoted from the work of others, the source is always given.
5. Except for such quotations, this work is entirely my own work.
6. I have acknowledged all main sources of help.
7. Where the work is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Semarang, November, 11, 2021



Goei, Steven Christian Sugiharto

17.K1.0016

APPROVAL PAGE FOR PUBLICATION OF SCIENTIFIC PAPERS FOR ACADEMIC INTEREST

The undersigned below :

Name : Goei ,Steven Christian Sugiharto

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 “IMAGE PRE PROCESSING FOR TESSERACT OCR” along with the existing tools (if needed). With this Non- Exclusive Royalty Free Right Soegijapranata Catholic University has the right store, transfer data / format, manage in the form of database, maintain and publis 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, November, 11, 2021

Sincerely



Goei,Steven Christian Sugiharto

ACKNOWLEDGMENT

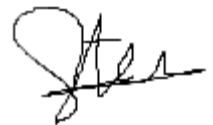
First of all, thank you to the Lord Jesus Christ for His blessing, so that I can finish my final project successfully. The final project is a requirement to take a the Bachelor of Computer Science Exam in the Informatic Engineering Study Program at Soegijapranata Catholic University Semarang.

In the preparation and masking of this final project, I was always supported and encouraged by people around me, special thanks to :

1. Father, Mother and my family who always support me in working on final project.
2. Hironimus Leong S.Kom., M.Kom. as a Supervising Lecturer who has kindly provided guidance and always be patient with me so that my final project can be completed.
3. My Girlfriend, Elke Melvinda who always beside me and support me when I get tired and start giving up on this final project.
4. All my friend who can't help solve the problem, but are always there when it's up and down also always turn the mood again to work on this final project, you're the mvp boys,glhf.
5. And other person that I cant mention one by one who pray for me and support me while working this project.

Semarang, November, 11, 2021

Sincerely



Goei,Steven Christian Sugiharto

ABSTRACT

Sometimes processing text data or numbers in images, it makes us difficult to process the data. Ocr is software that converts text in image format or image files into text format that can be read and edited by computer applications, but sometimes there are also some that can't be detected

And in my opinion through this pre processing will help the process of refinement or accuracy of this conversion process to a more accurate one, I use grayscale, then the image will go through the opening process where the image will be eroded first and then dilated, why don't I use the closing process, because what I want to detect here is text so that the results if using dilation will look worse than opening because it makes the writing close.

I tried to use all pre-processing processes to find out which accuracy value was the best, where I compared the erosion, dilation, opening and closing processes. where the result is that dilation has the lowest value with 34% and the highest opening with 59% and that makes me use opening, I also compare that converters that go through pre-processing are higher than those that only use tesseract by comparison when using tesseract only get 43% while pre-processing is 59% more accurate.

Keyword: tesseract ocr, convert text, image processing, image conversion

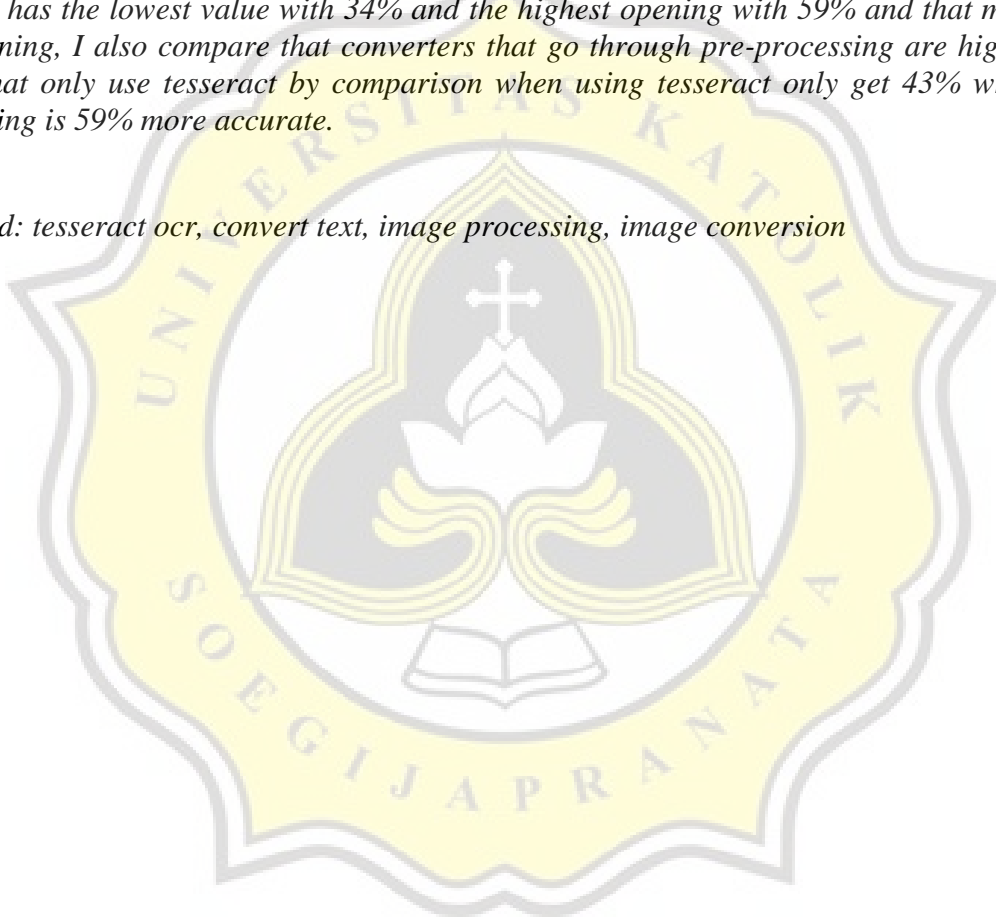
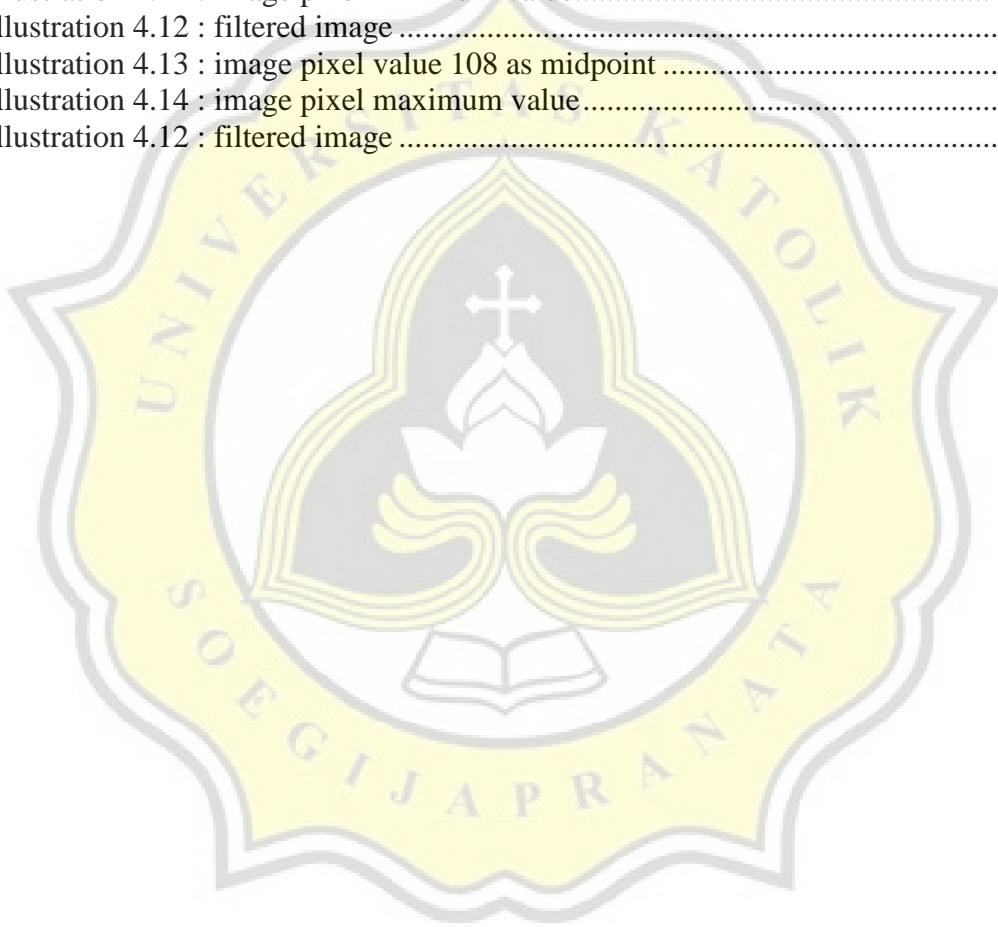


TABLE OF CONTENTS

COVER.....	i
APPROVAL AND RATIFICATION PAGE	ii
DECLARATION OF AUTHORSHIP.....	Error! Bookmark not defined.
ACKNOWLEDGMENT	iv
ABSTRACT (Abstract Title).....	vi
TABLE OF CONTENTS	vii
LIST OF FIGURE.....	1
LIST OF TABLE.....	1
CHAPTER 1 INTRODUCTION.....	3
1.1. Background.....	3
1.2. Problem Formulation.....	3
1.3. Scope	4
1.4. Objective.....	4
CHAPTER 2 LITERATURE STUDY.....	5
CHAPTER 3 RESEARCH METHODOLOGY	9
CHAPTER 4 ANALYSIS AND DESIGN	11
4.1. Analysis	11
4.2. Design.....	12
CHAPTER 5 IMPLEMENTATION AND RESULTS.....	23
5.1. Implementation.....	23
5.2. Testing	25
CHAPTER 6 CONCLUSION	29
REFERENCES	29
APPENDIX	a

LIST OF FIGURE

Illustration 4.1: Flowchart	10
Illustration 4.2: Original Image	11
Illustration 4.3: Value from 260x60 to 269x69pixels.....	11
Illustration 4.4: Value of the grayscaled image	11
Illustration 4.5: Grayscale image.....	12
Illustration 4.6: example location of yellow square	12
Illustration 4.7: example locationof red square	12
Illustration 4.8: example location of green box	12
Illustration 4.9: image pixel value 108is midpoint.....	12
Illustration 4.10 : image pixel value 172 as midpoint	13
Illustration 4.11 : image pixel minimum value.....	14
Illustration 4.12 : filtered image	14
Illustration 4.13 : image pixel value 108 as midpoint	14
Illustration 4.14 : image pixel maximum value.....	14
Illustration 4.12 : filtered image	14



LIST OF TABLE

Table 5.1: Table result of dataset image text	24
Table 5.2: Table result of dataset handwritten text	25
Table 5.3: Table result of dataset notimage text	25
Table 5.4: Table result of dataset photo image	25
Table 5.5: Table result of dataset mix image	25
Table 5.6: Measures of pre processing	26
Table 5.7: measures of tesseract	26

