

**DIGITALIZATION: DEVELOPMENT OF A PDF SCRAPING PROGRAM TO
AUTOMATE TASKS FROM OPERATIONS TEAMS IN MAERSK CUSTOMS
SERVICES**

By

Theophilus Ariel Gee
12002017

BACHELOR'S DEGREE
in

INFORMATION TECHNOLOGY WITH
ARTIFICIAL INTELLIGENCE AND DATA SCIENCE
FACULTY OF ENGINEERING AND INFORMATION TECHNOLOGY



SWISS GERMAN UNIVERSITY
The Prominence Tower
Jalan Jalur Sutera Barat No. 15, Alam Sutera
Tangerang, Banten 15143 - Indonesia

Revision after the Thesis Defense on 4 July 2024

STATEMENT BY THE AUTHOR

I hereby declare that this submission is my own work and to the best of my knowledge, it contains no material previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any other degree or diploma at any educational institution, except where due acknowledgement is made in the thesis.



Theophilus Ariel Gee

Student

17.07.2024
Date

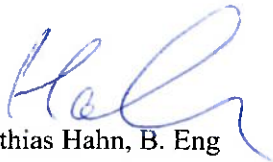
Approved by:



Dipl.-Inf. I Eng Kho

Thesis Advisor

17.07.2024
Date



Matthias Hahn, B. Eng

Thesis Co-Advisor

17.07.2024
Date

Dr. Maulahikmah Galinium S.Kom, M.Sc

Dean

17.07.2024
Date

Theophilus Ariel Gee

ABSTRACT

DIGITALIZATION: DEVELOPMENT OF A PDF SCRAPING PROGRAM TO AUTOMATE TASKS FROM OPERATIONS TEAMS IN MAERSK CUSTOMS SERVICES

By

Theophilus Ariel Gee
Dipl.-Inf. I Eng Kho, Advisor
Matthias Hahn, B. Eng, Co-Advisor

SWISS GERMAN UNIVERSITY
MAERSK CUSTOMS SERVICES

It is undeniable that PDF invoice documents are extensively utilized across various stages of the business process, regardless of the industry. In the context of a customs service company, invoices serve as a crucial information medium that will be passed through different organizations. The amount of time required to extract information from PDF invoices hinders the business operations, especially when the volume is large. Therefore this thesis aims to offer a solution by implementing a machine learning approach for scraping data out of PDF invoices, utilizing the object detection model YOLOv8, in conjunction with a novel AutoTrainer program to help streamline model training and training data preparation. The divide and conquer strategy is introduced, a crucial process in reducing noise and increasing accuracy in detection. Additionally, this thesis provides a comparison between an existing rule based approach and the machine learning approach. The result proves that the machine learning approach with object detection is superior to the rule based approach. Although there is a challenge to reach perfect accuracy, the reusability and flexibility it offers overcomes its limitations, especially when compared to the rule based approach.

Keywords: Document Analysis and Recognition, PDF invoice scraping, Table Detection, YOLO Object Detection, Robotic Process Automation.

© Copyright 2024
by Theophilus Ariel Gee
All rights reserved

DEDICATION

I dedicate this work for future research in document intelligence and to all parties to whom it may benefit.

ACKNOWLEDGEMENTS

I would like to thank the host company for this thesis, Maersk Customs Services, for the opportunity to be able to conduct this research until completion. I would also like to express my gratitude towards my digitalization manager at Maersk Customs Services, Matthias Hahn and my advisor from Swiss German University, Dipl.-Inf. I Eng Kho for their continuous support and guidance.

Thank you,

Theophilus Ariel Gee

TABLE OF CONTENTS

	Page
STATEMENT BY THE AUTHOR.....	2
ABSTRACT.....	3
DEDICATION.....	5
ACKNOWLEDGEMENTS.....	6
TABLE OF CONTENTS.....	7
LIST OF FIGURES	9
LIST OF TABLES.....	11
CHAPTER 1 – INTRODUCTION	12
1.1 Background.....	12
1.2 Research Problems.....	17
1.3 Research Objectives.....	17
1.4 Research Questions.....	18
1.5 Research Hypothesis.....	18
1.6 Significance of Study.....	19
1.7 Research Scope	19
1.8 Research Limitation	19
CHAPTER 2 - LITERATURE REVIEW	20
2.1 Theoretical Perspectives	20
2.1 Related Works.....	21
2.1.1 PDFDataExtractor: A Tool for Reading Scientific Text and Interpreting Metadata from the Typeset Literature in the Portable Document Format	21
2.1.2 Variable Indexing Method in Rule Documents for Ship Design Using Extraction of Portable Document Format Elements	22
2.1.3 Image-Based Table Recognition: Data, Model, and Evaluation.....	23
2.1.4 Information Extraction from Semi and Unstructured Data Sources: A Systematic Literature Review	25
2.1.5 Current Status and Performance Analysis of Table Recognition in Document Images With Deep Neural Networks	26
2.2 Comparison Table	28
CHAPTER 3 – RESEARCH METHODS	33

3.1 Analysis.....	33
3.2 Design	34
3.2.1 Process Overview	34
3.2.2 Training Data Preparation Pipeline	35
3.3 Implementation	40
3.3.1 PDF Into PNG	40
3.3.2 Training Data Annotation	41
3.3.3 Divide and Conquer Strategy.....	43
3.3.4 OCR Process.....	46
3.3.5 Auto Trainer.....	46
3.3.6 GOV.json.....	48
3.3.7 Rule Based Approach	55
CHAPTER 4 – RESULTS AND DISCUSSIONS.....	58
4.1 First Phase: Root Detection Model	58
4.2 Second Phase: Seller And Client Data Sub Detection Model.....	61
4.3 Third Phase: Items Table Sub Detection Model	64
4.4 Fourth Phase: Item Row Sub Detection Model	67
4.5 Fifth Phase: Summary Table Sub Detection Model	68
4.6 AutoTrainer Complete Detection Result.....	70
CHAPTER 5 – CONCLUSIONS AND RECOMMENDATIONS	74
5.1 Conclusions.....	74
5.2 Recommendations.....	75

LIST OF FIGURES

Figures	Page
Figure 1. 1 Customs Clearance Process Overview	13
Figure 1. 2 Deeper look at declarant's job.....	14
Figure 1. 3 RPA Support in customs clearance process	16
Figure 3. 1 Different Table Understanding Methodologies Taken from Hashmi et al.	33
Figure 3. 2 Process Overview Diagram	34
Figure 3. 3 Data Preparation Process Diagram	35
Figure 3. 4 Pre-Determined Labels Visualizer.....	37
Figure 3. 5 Sample PDF 1.....	38
Figure 3. 6 Sample PDF 2.....	39
Figure 3. 7 Code Snippet of PNG Transformer	40
Figure 3. 8 Label-Studio Dashboard	41
Figure 3. 9 Label-Studio Annotation Canvas.....	41
Figure 3. 10 Label-Studio Annotation Export in YOLO Format.....	42
Figure 3. 11 Bounding Box Coordinates in YOLO Format.....	42
Figure 3. 12 Example Detection Result	44
Figure 3. 13 Example Sub Detection Result	45
Figure 3. 14 Code Snippet for OCR.....	46
Figure 3. 15 Auto Trainer Mode: Training	46
Figure 3. 16 Auto Trainer Mode: Detection	47
Figure 3. 17 Example of Generated Out.json	48
Figure 3. 18 Annotation With 6 Labels for Training Root Detection Model	49
Figure 3. 19 GOV.json for training phase #1.....	50
Figure 3. 20 GOV.json for training phase #2.....	51
Figure 3. 21 Implementing rule based strategy on the sample PDF invoice	55
Figure 4. 1 Annotation data for root detection model visualizer	59
Figure 4. 2 Root detection model: confusion matrix	59
Figure 4. 3 Root detection model: detection result	60
Figure 4. 4 Seller data sub detection model: confusion matrix.....	62
Figure 4. 5 Seller data sub detection model: detection results.....	62
Figure 4. 6 Client data sub detection model: confusion matrix	63
Figure 4. 7 Client data sub detection model: detection results	63
Figure 4. 8 Items table sub detection model: acceptable detection result.....	64
Figure 4. 9 Items table sub detection model: confusion matrix	65
Figure 4. 10 Items table sub detection model: overlapping detection of 2nd row	65
Figure 4. 11 Items table sub detection model: missing detection on 3rd row and cropped detection on 6th row	66

Figure 4. 12 Item row sub detection model: detection results	67
Figure 4. 13 Item row sub detection model: confusion matrix	67
Figure 4. 14 Summary table sub detection model: detection results	68
Figure 4. 15 Summary table sub detection model: confusion matrix	69
Figure 4. 16 Example PDF invoice	71
Figure 4. 17 AutoTrainer complete detection results.....	72

LIST OF TABLES

Table	Page
Table 2. 1 Related work comparison table.....	28
Table 3. 1 Comparison Table of Different Object Detection Methodologies Taken from Hashmi et al.....	33
Table 3. 2 Auto Trainer training_opts Tags.....	52
Table 3. 3 Comprehensive List of Possible Tags in Detection Objects.....	53
Table 4. 1 Root Detection Model Training Report.....	61
Table 4. 2 Seller Data Sub Detection Model Training Report.....	62
Table 4. 3 Client Data Sub Detection Model Training Report	63
Table 4. 4 Items Table Sub Detection Model Training Report.....	64
Table 4. 5 Error detections table out of 420 rows.....	66
Table 4. 6 Items Row Sub Detection Model Training Report	68
Table 4. 7 Summary Table Sub Detection Model Training Report	69
Table 4. 8 Runtime benchmark	73
Table 5. 1 Approach strengths and weaknesses comparison table	74