DECLARATION

I **BODAPATI SOHAN CHIDVILAS and VENKATA NAGA SAI RAKESH KAMISETTY** hereby declare that the Project Report entitled **DIGITIZATION OF DATA FROM INVOICE USING OCR** done by me under the guidance of **Dr. S.Revathy M.E., Ph.D.,** (Internal) at **cSoft Technologies** (Company name and address) is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in Computer Science and Engineering.

DATE: 09-11-2021

BODAPTI SOHAN CHIDVILAS
VENKATA NAGA SAI RAKESH KAMISETTY

PLACE: CHENNAI

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

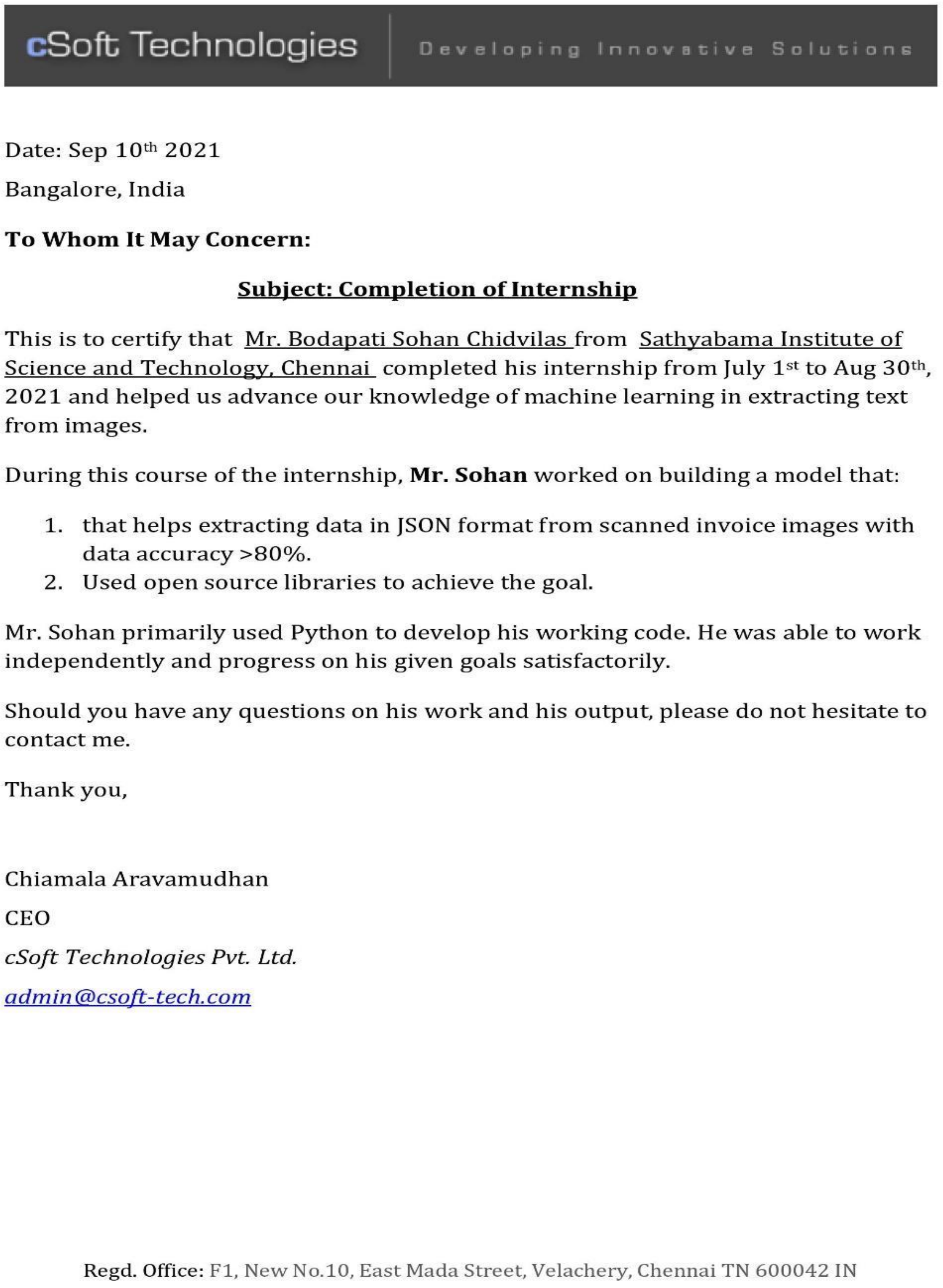
I am pleased to acknowledge my sincere thanks to **Board of Management** of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T.Sasikala M.E., Ph.D.**, **Dean**, School of Computing , **Dr.S.Vigneshwari M.E., Ph.D., and Dr.L.Lakshmanan M.E., Ph.D.,** Heads of the Department of Computer Science and Engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr. S.Revathy M.E., Ph.D.,** for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

TRAINING CERTIFICATE



cSoft Technologies | Developing Innovative Solutions

Date: Sep 10th 2021

Bangalore, India

To Whom It May Concern:

Subject: Completion of Internship

This is to certify that Mr. Bodapati Sohan Chidvilas from Sathyabama Institute of Science and Technology, Chennai completed his internship from July 1st to Aug 30th, 2021 and helped us advance our knowledge of machine learning in extracting text from images.

During this course of the internship, **Mr. Sohan** worked on building a model that:

1. that helps extracting data in JSON format from scanned invoice images with data accuracy >80%.
2. Used open source libraries to achieve the goal.

Mr. Sohan primarily used Python to develop his working code. He was able to work independently and progress on his given goals satisfactorily.

Should you have any questions on his work and his output, please do not hesitate to contact me.

Thank you,

Chiamala Aravamudhan

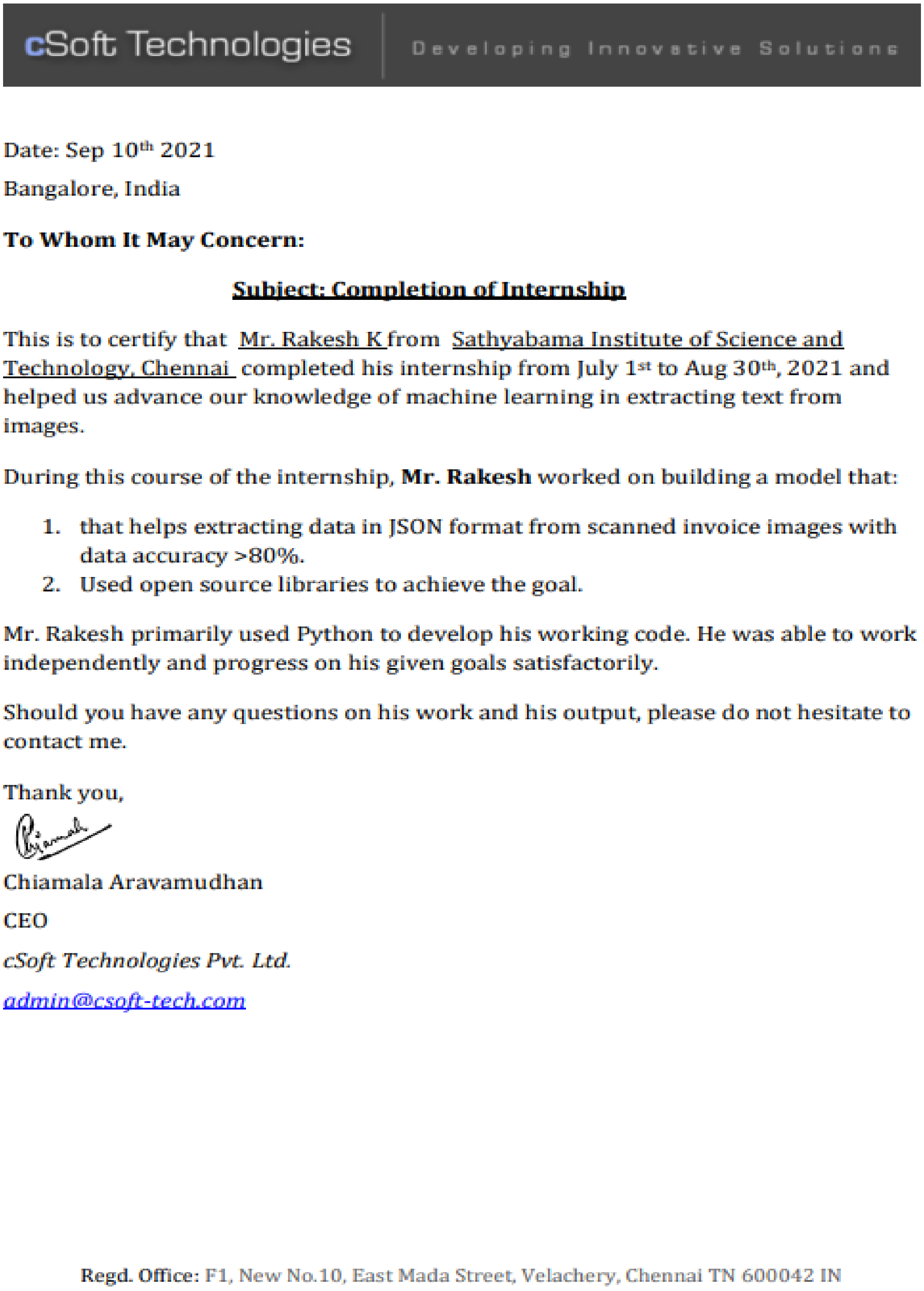
CEO

cSoft Technologies Pvt. Ltd.

admin@csoft-tech.com

Regd. Office: F1, New No.10, East Mada Street, Velachery, Chennai TN 600042 IN

TRAINING CERTIFICATE



cSoft Technologies | Developing Innovative Solutions

Date: Sep 10th 2021

Bangalore, India

To Whom It May Concern:

Subject: Completion of Internship

This is to certify that Mr. Rakesh K from Sathyabama Institute of Science and Technology, Chennai completed his internship from July 1st to Aug 30th, 2021 and helped us advance our knowledge of machine learning in extracting text from images.

During this course of the internship, **Mr. Rakesh** worked on building a model that:

1. that helps extracting data in JSON format from scanned invoice images with data accuracy >80%.
2. Used open source libraries to achieve the goal.

Mr. Rakesh primarily used Python to develop his working code. He was able to work independently and progress on his given goals satisfactorily.

Should you have any questions on his work and his output, please do not hesitate to contact me.

Thank you,

Chiamala Aravamudhan
CEO
cSoft Technologies Pvt. Ltd.
admin@csoft-tech.com

Regd. Office: F1, New No.10, East Mada Street, Velachery, Chennai TN 600042 IN

ABSTRACT

Optical Character Recognition (OCR) is a predominant aspect to transmute scanned images and other visuals into text. Computer vision technology is extrapolated onto the system to enhance the text inside the digitized image. This preliminary provisional setup holds the invoice's information and converts it into JSON and CSV configurations. This model can be helpful in divination based on knowledge engineering and qualitative analysis in the nearing future. The existing system contains data extraction and nothing more. In a paramount manner, image pre-processing techniques like black and white, inverted, noise removal, grayscale, thick font, and canny are applied to escalate the quality of the picture. With the enhanced image, more OpenCV procedures are carried through. In the very next step, three different OCRs are used: Keras OCR, Easy OCR, and Tesseract OCR, out of which Tesseract OCR gives the precise result. After the initial steps, the undesirable symbols (/t, /n) are cleared to get the escalated text as an output. Eventually, a unique work that is highly accurate in giving JSON and CSV formats is developed.

Impact statement— In our protrude, a front-end android app is developed which takes input from the user and stores the output onto the database. The JSON and CSV files can be viewed through an app by the end.

ABBREVATIONS

OCR – Optical Character Recognition
JSON - JavaScript Object Notation
CSV – Comma-separated values

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Computer vision drew attention by swaying as a data- reliant stratified feature in extraction methods.

Visualization technology has been imposed to decipher an image to make the machine understand. Optical Character Recognition (OCR) automatically extracts characters from the image and recognizes text quickly using an existential database.

OCR is a meticulous technology that comes up with legible recognition of inscribed or in-written characters from images which will be further digitized in our apparatus. Various procedures have been in use already. Despite this, the existing OCRs cannot convert the text into the desired form that the end-user needs. In this current era, OCR has been the most dominant technology. OCR can be used in an enchanting number of ways apart from just extracting the text. They are shown in a different dimension here.

Among the OCRs around globe, the least preferred is Keras OCR as it goes with line segmentation. The other one is Easy OCR, a parasite of spaces. Finally, Tesseract OCR is the best open-source choice as it can be corelated with python libraries called pytesseract. Tesseract OCR extracts the text based on the invoice format. The exact explanation of how tesseract OCR extracts the text form the image is inscribed in section V under phase 4. The primary Python library used in our structure is OpenCV, which helps the machine find objects in the image and make OCR work efficiently. In this adorable framework, pdf or an image (JPG, JPEG, PNG) is taken as input from the android app. If it is a pdf, it will be converted into a picture, then pre-processing techniques (Black and white, no noise, Grayscale, thick font) have been used to amplify the text. Textual content is a conduit where details are confronted with a machine orderly to give a valid result. Multiple approaches are there to extract text in many different ways with an OCR to get the most accurate result. It has to be validated with a set of pre-trained images to get an efficient output. Then the noise should be filtered from the photo to make the above statement work. Below are the few processing methodologies for an image to be intensified under OpenCV.

Thresholding is a form where the image will be segmented to understand the image better. Several procedures have been applied like spatial to correspond with the pixels and further computerize it to black and white for highlighting the words to bring back the highest quality. The pivot OpenCV methodology also sharpens the image by blurring the borders to make the essential fields stand out. Threshing also includes smoothening where it evens rough side to blend the text. The text got from the OCR may not be error-free. So, regular expressions have been used to clean the printed characters further. The string format must be converted into a list by splitting it as a ratio. In the concluded part of our setup, the cleaned text is converted into JSON and CSV formats for better comprehension.

Here JavaScript Object Notation (JSON) is a format that will return the object from the back-end server and edit cookies. Apart from the primary use, the web developers mostly use it to deploy output onto the web page. Mainly, Key pair values are generated and commonly known as dictionaries in python. CSV is a format where a comma separates the values, and a tabular column is created, which returns as an excel sheet. The basic idea of developing an app is to make it uncomplicated. A rudimentary java file picker has been evolved to residue the complexity.

1.1 PROBLEM STATEMENT

It is observed that the performance of the computer vision degrades drastically under the course of action. The results are also compared with existing computer vision fooling approaches to evaluate the accuracy drop. We propose a primary state-of-the-art performance using the solution in terms of the computer robustness under OCR is observed in the experiments. No OCR in the world which can detect only the specific contents with more than 80 percent accuracy which can convert it into JSON or CSV.

1.2 PROJECT JUSTIFICATION

The necessity of this protrude is extracting the relevant data instead of unnecessary matter. For instance, take medical bills, when we need only the tabular contents then there is no OCR that can detect the tabular columns separately, that to with 80 % accuracy and returning the output as JSON as well as csv. The main advantage of JSON and CSV is that the end user need not enter contents as they will directly return the particulars. Our methodology proves to be highly accurate while tested on a variety of input images of bills and invoices. This course of action achieves an increase in accuracy by 80%. The proposed approach can be used to improve the robustness of Computer Vision.

CHAPTER 2

LITERATURE SURVEY

2.1 OCR text extraction

Abstract:

This research tries to find out a methodology through which any data from the daily-use printed bills and invoices can be extracted. The data from these bills or invoices can be used extensively later on - such as machine learning or statistical analysis. This research focuses on extraction of final bill-amount, itinerary, date and similar data from bills and invoices as they encapsulate an ample amount of information about the users purchases, likes or dislikes etc. Optical Character Recognition (OCR) technology is a system that provides a full alphanumeric recognition of printed or handwritten characters from images. Initially, OpenCV has been used to detect the bill or invoice from the image and filter out the unnecessary noise from the image. Then intermediate image is passed for further processing using Tesseract OCR engine, which is an optical character recognition engine. Tesseract intends to apply Text Segmentation in order to extract written text in various fonts and languages. Our methodology proves to be highly accurate while tested on a variety of input images of bills and invoices.

2.2 Mixed-Initiative Approach to Extract Data from Pictures of Medical Invoice

Abstract:

Extracting data from pictures of medical records is a common task in the insurance industry as the patients often send their medical invoices taken by smartphone cameras. However, the overall process is still challenging to be fully automated because of low image quality and variation of templates that exist in the status quo. In this paper, we propose a mixed-initiative pipeline for extracting data from pictures of medical invoices, where deep-learning-based automatic prediction models and task-specific heuristics work together under the mediation of a user. In the user study with 12 participants, we confirmed our mixed-initiative approach can supplement the drawbacks of a fully automated approach within an acceptable completion time. We further discuss the findings, limitations, and future works for designing a mixed-initiative system to extract data from pictures of a complicated table.

2.3 OCR for Data Retrieval: An analysis and Machine Learning Application model for NGO social volunteering

Abstract:

With the increase in amount of information being made available in digital format, information retrieval is a challenging task. Currently there exists a gap between organizations, volunteers and NGOs for volunteering work. There has been an upsurge of NGOs, non-profit events and corresponding independent volunteers or organizations willing to interconnect especially during these pandemic times. There is a need to fill this gap and connect the stakeholders minimizing the emergency response times. This paper proposes a novel design and implementation of an OCR based application for Automated NGO connect using machine learning. Phases implemented include image de-noising, binarization, data extraction and data conversion. The framework integrates deep learning-based Tesseract OCR with image processing module and Data visualization module. The proposed model can be extended to other application domains as well for research purposes.

2.4 Comparative Analysis of Text Extraction from Color Images using Tesseract and OpenCV

Abstract:

Image-based Text Extraction has a growing requirement in today's generation. Students, doctors, and engineers generate a lot of images every day. It is very important to extract text from these images in a simple yet effective manner. We can obtain useful information by testing these images. We aim is to summarize the visual information and retrieve its content. The Optical Recognition System involves several algorithms that fulfill this purpose. Text Extraction involves a lot of processes from text detection, localization, segmentation and, text recognition. Tesseract is the most optimized OCR Engine build by HP Labs and owned by Google. Text Detection involves the recognition of text from desired input images. Text Localization involves identifying the position of text on the images. Tesseract works pretty well on the light-colored background but unable to recognize text on darker shades. We have tried to apply various image processing techniques. This method will allow us to recognize text from most types of background. We propose to provide methods for easy text extraction. Track bar allows the user to adjust various parameters to extract a required text from an Image. This method is gaining huge importance in years to come. For Automation, we can use a set of image processing techniques such as edge detection, filtering and, blurring for better results. A series of these steps will enable us to extract text from images efficiently. This experiment compares the optimized result by two methods for efficient Text Extraction.

2.5 Analysis of Image Classification for Text Extraction from Bills and Invoices

Abstract:

Optical Character Recognition (OCR) technology offers a complete alphanumeric recognition of printed or handwritten characters from pictures such as scanned bills and invoices. Intelligent extraction and storage of text in structured document serves document analytics. The current research attempts to find a methodology through which any information from the printed invoice can be extricated. The intermediate image is passed over using an OCR engine for further processing. Segmentation extracts written text in various fonts and languages. Image classification helps in making a decision based on the classification results. This paper surveys these techniques and compares them in terms of metrics, algorithm and results.

2.6 Text Orientation Detection Based on Multi Neural Network

Abstract:

Optical character recognition (OCR) is an important research area in the field of pattern recognition, such as Vehicle License Plate Recognition. With it, we can extract textual information from the images to facilitate digital processing. However, most existing systems are designed to detect or recognize horizontal (or near-horizontal) texts and can't be applied to recognize texts of varying orientations. Text Orientation Detection is an important but challenging task. It can be used as a pre-processing for previous researches, and allows them to be adapted to more complex situations. Once we get angle of the text, we can use a series of transformations to get horizontal text. Although this problem is a multiclass classification, the results using common multiclass classification methods are not ideal. Our algorithm is inspired by the human behavior of recognizing texts. In this paper, we propose a new algorithm containing three neural networks to detect text orientation. The first is for capturing the abstract information of text image, and trained on multi-orientation, synthetic texts. Then the second is for evaluating the correctness of meaning of texts and trained on horizontal texts. The output of these two neural networks serves as the input to the last. In this way, the last neural network can obtain information about the image of the text as well as its meaning,