DATA EXTRACTION

A PROJECT REPORT

Submitted by

Gediya Miteshkumar G.

Amin Sanket K.

Bhadaja Pavankumar V.

Shah Raj R.

In fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

in

Computer Engineering

LDRP Institute of Technology and Research, Gandhinagar

Kadi Sarva Vishwavidyalaya

Sept, 2020

LDRP INSTITUTE OF TECHNOLOGY AND RESEARCH GANDHINAGAR

CE-IT Department

CERTIFICATE

This is to certify that the Project Work entitled **“Data Extraction”** has been carried out by **Name of Student (#Enronllment No)** under my guidance in fulfilment of the degree of Bachelor of Engineering in Computer department Semester-7 of Kadi Sarva Vishwavidyalaya University during the academic year 2020-21.

Dr. Bela Shremali
Internal Guide
LDRP-ITR

Dr. Shivangi Surati
Head of the Department
LDRP-ITR

ACKNOWLEDGEMENT

With immense pleasure we would like to present this report on our topic “Data Extraction”. We are thankful to all that have helped us a lot for successful completion of our project and providing us courage for completing the work.

We are thankful to our Head of the Department **Dr. Shivangi Surati**, our project mentor **Dr. Bela Shrimali** for providing encouragement, constant support and guidance which was of a great help to complete this project successfully.

We thank our Dearest parent, who encourages us to extend our reach with their help and support, we have been able to complete this work. We are also thankful to all my friends who directly or indirectly have been helpful in some or the other way.

With regards,
Shah Raj
Bhadaja Pavan
Amin Sanket
Mitesh Gedia

ABSTRACT

Optical Character Recognition (OCR) is the technology for identification of characters with utmost accuracy possible by employing suitable preprocessing, processing and post processing refinements. Practical application of OCR is very wide ranging from day-to-day need to scientific research purposes. One very crucial application is to Automate Digitalization of Business cards. Since Business cards comes in different fonts and sizes, and most importantly with different lighting conditions, applying OCR can be done after careful processing. Avoiding noise from source image is one of the most crucial step in any image-processing process and it has major weightage in the accuracy of further step and thus indirectly has a huge contribution for our final outcome. Further proper noise cancellation in our source image can reduce number of future steps required to attain good accuracy and also avoid problem of iterating our sample over cycles to avoid better contrast or to distinguish text in our source lying in a noisy matrix. Digitalising business cards aims at classification of the text extracted from our source image in the hard copy of the business card directly into respected following classified fields so that it becomes a lot easier to proceed any desired function ones aims to do with that business card in this online era.

TABLE OF CONTENTS:

Acknowledgement 1

Abstract 2

Table of Contents 3

List of Figures 4

List of Tables 5

1 Introduction **6**

1.1 Introduction 7

1.2 Aims and Objective of the work 7

1.3 Brief Literature Review 7

1.4 Problem definition 8

1.5 Plan of their work 8

2 Technology and Literature Review **9**

3 System Requirements Study **11**

3.1 User Characteristics 12

3.2 Hardware and Software Requirements 12

3.3 Assumptions and Dependencies 13

4 System Diagrams **-**

5 Data Dictionary **14**

6 Result, Discussion and Conclusion **16**

7 References **18**

1. Introduction

1.1 Introduction

1.2 Aims and Objectives of the work

1.3 Brief Literature Review

1.4 Problem Definition

1.5 Plan of Work

1.1 INTRODUCTION:

Digital image processing refers to use of computerized algorithms to perform image processing on digital images. Optical character recognition or OCR(Tesseract algorithm) is a form of information entry for business cards, e-mails, pan cards, Id cards, which scans a document in written form or printed form and retrieving the text out of it. The idea of Business card Digitalization has been evolved from Automatic License Plate Recognition. Digitizing the Business Cards is a real challenge. They come in different formats and fonts. Text extraction in different lighting conditions is very difficult. A formal structure for the business card reader which was innovated is reported. A Boundary Detection method is proposed called Biggest Contour method and Hough lines Transformation method, image extraction and segmentation technique based on a statistical method called Connected Component Method. The task is to detect the boundary of the card by eliminating the background. The image was subdivided into an array of smaller blocks, over which gray thresholding is used to compute local thresholds. These thresholds are then stored in another array.

1.2 Objectives:

- To provide an easy user interface to input the object image.
- User should be able to upload the image.
- System should be able to pre-process the given input to suppress the background.
- System should detect text regions present in the image.
- System should retrieve the text present in the image and display them to the user.
- The system should provide accuracy.
- To separate the text into different categories like Company name, email, website etc.
- To store separated data in the database and then can be easily retrievable.
- To digitalize the physical business cards.

1.3 Brief Literature Review:

Data Gathering is necessary step before preceding to perform any next step. Data extraction from available data on internet is very important and necessary task. Nowadays data can be extracted from image, pdf, word files. We can extract text data from image using OCR technology. This literature focuses on digitizing business card by extracting text data using tesseract engine and parsing the text to get information from it like Name, Company name, Email, Contact, Website, Address. System does not need the access to camera as it takes pre-captured images from user as input. Image will go to preprocessing step in which it will be perspective warped and thresholded to binarize. This binarized image is provided to the tesseract engine. It will return the text extracted from image. Now the text is passed to the parser. The parser have different functions to determine and identify the details like Name, email, contacts. Once text has been parsed, parsed information is passed back to user where user can verify and correct the parsed information. Once verified user can submit it and information will be stored in database.

1.4 Problem definition:

Digitizing business card means extracting text data from the business card and parse into meaningful categories then store the information as contact or in database.

1.5 Plan of work:

- Deciding on which ocr technology to use.
- Getting optimal pre-processed image for ocr engine by method which will be decided by trial and error of different pre-processing method,
- Creating parser to extract information suitable for requirement.

2. Technology and Literature Review

2.1 Tesseract –OCR

2.2 Django

2. Technologies

Tesseract-OCR

Tesseract is an open-source OCR engine developed by HP Labs and currently maintained by Google. It is the leading open-source OCR engine and closely follows commercial OCR products in terms of accuracy.

The Tesseract algorithm assumes its input is a binary image and works in two steps: preprocessing and recognition. Tesseract preprocessing involves finding the outline of the text, identifying the text lines, and separating each character in a word. The recognition step passes the words it identifies to an adaptive classifier that it then uses for the subsequent text in the document. Thus, it should perform better on text further down in the page. This recognition step is run on the entire document for a second pass, since the classifier did not have as much training data for the earlier words in the document.

Django

Django is an open-source python web framework used for rapid development, pragmatic, maintainable, clean design, and secure websites. A web application framework is a toolkit of all components needed for application development.

The main goal of the Django framework is to allow developers to focus on components of the application that are new instead of spending time on already developed components. Django is fully featured than many other frameworks on the market. It takes care of a lot of hassles involved in web development; enables users to focus on developing components needed for their application.

3. System Requirement Study

3.1 User Characteristics

3.2 Hardware and Software Requirements

3.3 Functional & Non Functional Requirements

3.4 Assumptions and Dependencies

3. System Requirements Study

3.1 User Characteristics

Analyzing user characteristic is an important aspect of any project. It allows us to clearly define and focus on who the end users are for the project. Also it allows checking the progress of the project to ensure that we are still developing the system for the end users. The user must have following characteristics.

- User should know the basics of the internet.
- User must register before using the web application.

3.2 Hardware and Software Requirements

Hardware Requirements:

- Internet Connection
- Device supporting the browser.
- 512MB ram

Software Requirements:

- OS: Windows, Linux, MacOs, Android, ios
- Web Browser: Any HTML compliant browser
- Database: SQLlite

3.3 Assumptions and Dependencies

Assumptions:

- User is the person having enough knowledge for the accessing the website.
- We will provide a user friendly interface so that any user can easily navigate through the system, but he/she should be capable of providing valid input.
- The server used for data storing is always secured.

Dependencies:

- The system is dependent upon the user’s valid input. If user inputs wrong file then system will not give expected output and may now work properly.
- This system depends on the server and internet as all the information is collected and then stored in the server through secure internet connection.
- The Business card borders should be clearly visible in the image with the text.

5. Data Dictionary

5. Data Dictionary

Cards:

Column_Name	Data_type	Size	Constraints
Id	int	5	Primary
Name	varchar	25	
Company Name	varchar	25	
Email	varchar	50	
Website	varchar	50	
Contact1	varchar	12	
Contact2	varchar	12	
City	varchar	10	
State	varchar	10	
Pincode	varchar	6	
Address	varchar	50	

6. Result, Discussion and Conclusion

6. Result, Discussion and Conclusion:

❖ Result:

System can take input image and provide digital text and parse it into relevant information which is stored in database.

❖ Discussion:

This system is not 100% accurate. Image should meet some conditions to get good result from OCR engine. OCR engine itself is not perfect. It's quality highly depends on quality of image provided which makes preprocessing step very important. Parser is not perfect for identifying information. It has some guesswork.

❖ Conclusion:

Using different technologies we can extract the data from the physical business card and turn it into digital contact or store the information in database.

7. References

7. References:

- https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3358293
- https://stacks.stanford.edu/file/druid:np318ty6250/Sharma_Fujii_Automatic_Contact_Importer.pdf
- https://pypi.org/project/pytesseract/
- https://opencv.org/
- https://github.com/kumar-shridhar/Business-Card-Detector
- https://github.com/thucdx/business_card_detection
- https://www.danvk.org/2015/01/07/finding-blocks-of-text-in-an-image-using-python-opencv-and-numpy.html