CS-591 Project Report

EXTRACTION OF CELL CONTENTS FROM TABLES IN A DOCUMENT IMAGE

Submitted in partial fulfillment of

the requirements for the award of degree of

Bachelor of Technology

In

Computer Science and Engineering

Submitted by

Roll No.	Name of Student
39/CSE/18075/386	Sandeep Saini
39/CSE/18015/325	Ankit Singh
39/CSE/18095/406	Suraj Singh

Under the guidance of

Dr. Sanjoy Pratihar

Department of Computer Science and Engineering

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY

Kalyani, Nadia , West Bengal-741235

Contents:-

Certificate:

This is to certify that the report entitled extraction of cell contents from table in a document image being submitted by Sandeep Saini (Reg No.386), Ankit Kumar Singh (Reg No. 325) and Suraj Singh (Reg No. 406), undergraduate students in the Department of Computer Science and Engineering, Indian Institute of Information Technology, Kalyani , West Bengal- 741235, India, for the award of Bachelor of Technology in Computer Science and Engineering, is an original project work carried by them under my supervision and guidance. This report has fulfilled all the requirements as per the regulations of the Indian Institute of Information Technology, Kalyani and in my opinion, has reached the standards needed for submission. The work and the results presented have not been submitted to any other university or institute for the award of any other degree or diploma.

Dr. Sanjoy Pratihar

Indian Institute of Information Technology

Kalyani, West Bengal-741235

Declaration:

We hereby declare that the work being presented in this project entitled Extraction of cell content from tables in document image submitted to Indian Institute of Information Technology Kalyani in partial fulfillment for the award of the degree of Bachelor of Technology in Computer Science and Engineering during the period from July 2020 to December 2020 under the guidance of Dr. Sanjoy Pratihar, Department of Computer Science and Engineering, Indian Institute of Information Technology Kalyani, West Bengal - 741235, India

Name of the Candidates :-
Sandeep Saini
39/CSE/18075/386

Ankit Kumar Singh
39/CSE/18014/325

Suraj Singh
39/CSE/18095/406

Acknowledgement:-

We would like to express our Gratitude to our supervisor Dr. Sanjoy Pratihar for providing their enormous support , invaluable guidance ,comments and suggestions throughout the course of the project. We would also like to express our genuine gratitude to Department of computer science and engineering, Indian Institute of Information Technology,Kalyani for providing us the platform and opportunity to work on this project and for helping us in every possible way.

Sandeep Saini

39/CSE/18075/386

Ankit Kumar Singh

39/CSE/18014/325

Suraj Singh

39/CSE/18095/406

Introduction:-

In the running world there is growing demand for the software systems to recognize characters in computer system when information is scanned through paper document. As we know that we have number of newspapers and books which are in printed format related to different subjects. These days there is a huge demand in -storing the information available in these paper documents in to a computer storage disk and then later reusing this information by searching process. One simple way to store information in the paper documents in to computer system is to first scan the documents and then store them as IMAGES. But to reuse this information it is very difficult to read the individual contents and searching the contents from these documents line-by-line and word-by-word. The reason for this difficulty is the font characteristics of the characters in paper documents are different to font of the characters in computer system. As a result computer is unable to recognize the characters while reading them. This concept of storing the contents of paper documents in computer storage place and then reading and searching the content is called DOCUMENT PROCESSING. Sometimes in this document processing we need to process the information that is related to languages other than the English in the world. For this document processing we need a software system called OPTICAL CHARACTER RECOGNITION SYSTEM.

Thus our need is to develop character recognition tool system to perform Document Image Analysis which transforms documents in paper format to electronic format. For this process there are various techniques in the world. Among all those techniques we have chosen Optical Character Recognition (OCR) as main fundamental technique to recognize characters. The conversion of paper documents in to electronic format is an on- going task in many of the organizations like banking, insurance sectors, Research and development(R&D) area , Business enterprises, in government institutions , so on.

To effectively use Optical Character Recognition (OCR) tool for character recognition in order to perform Document Image Analysis, we use OpenCV python, easyocr package, GPU to fasten the process.

1.1 Purpose

The main purpose of Optical Character Recognition (OCR) system is to perform Document Image Analysis, document processing of electronic document formats converted from paper formats more effectively and efficiently. This improves the accuracy of recognizing the characters during document processing compared to various existing available character recognition methods. Here OCR technique derives the meaning of the characters, their font properties from their bitmapped images.

- **The primary objective is to speed up the process of character recognition in document processing. As a result the system can process huge number of documents within less time and hence saves the time.**
- **Since our character recognition is based on deep Learning, it aims to recognize multiple heterogeneous characters with different font properties and alignments.**

1.2 Project Scope

The scope of our tool Optical Character Recognition(OCR) is to provide an efficient and enhanced software tool for the users to perform Document Image Analysis, document processing by reading and recognizing the characters in research, academic governmental, industrial, insurance and business organizations that are having large pool of documented scanned images Irrespective of the size of documents and the type of characters in document, the ocr tool is recognizing them, searching them and processing them faster according to the needs of the environment

1.3 Proposed System:

Our proposed system is OCR based on deep learning which is a character recognition system that supports recognition of the characters. The problem of heterogeneous character recognition and supports multiple functionalities to be performed on the document. The multiple functionalities include editing and searching too where the existing system supports only editing of the document.

1.4 Intended Audience :

In this section, we identify the audience who are interested with the product and are involved in the implementation of the product either directly or indirectly As from our project, the OCR system is mainly useful in R&D at various scientific organizations, in governmental institutes, in industrial sector, in insurance sector and in large business organizations, we identify the following as various interested audience in implementing OCR system.

- **The scientists, the research scholars and the research fellows in telecommunication institutions are interested in using OCR system for processing the word document that contains base paper for their research.**
- **The Librarian to manage the information contents of the older books in building virtual digital library requires use of OCR system.**
- **Various sites that vendor e-books have a huge requirement of this OCR system in order to scan all the books in to electronic format and thus make money.**

The Amazon book world is largely using this concept to build their digital libraries.

2.Feasibility Study:

A feasibility study is a high-level capsule version of the entire System analysis and Design Process. The study begins by classifying the problem definition. Feasibility is to determine if it's worth doing. Once an acceptance problem definition has been generated, the analyst develops a logical model of the system. A search for alternatives is analyzed carefully. A search for alternatives analyzed carefully. There are 3 parts in feasibility study.

1. Technical Feasibility

Evaluating the technical feasibility is the trickiest part of a feasibility study. This is because, at this point in time, not too many detailed design of the system, making it difficult to access issues like performance, costs on (on account of the kind of technology to be deployed) etc. A number of issues have to be considered while doing a technical analysis. Understand the different technologies involved in the proposed system before commencing the project we have to be very clear about what are the technologies that are to be required for the development of the new system.

2. Operational feasibility

Proposed project is beneficial only if it can be turned into information systems that will meet the organizations operating requirements. Simply stated, this test of feasibility ask if the system will work when it is developed and installed. To compensate the barriers which appears during implementation.

3.Economic feasibility

Economic feasibility attempts to weigh the costs of developing and implementing a new system, against the benefits that would accrue from having the new system in place. This feasibility study gives the top management the economic justification for the new system. A simple economic analysis which gives the actual comparison of costs and benefits are much more meaningful in this case. In addition, this proves to be a useful point of reference to compare actual costs as the project progresses. There could be various types of intangible benefits on account of automation. These could include increased customer satisfaction, improvement in product quality better decision making timeliness of

information, expediting activities, improved accuracy of operations, better documentation and record keeping, faster retrieval of information.

3. Application of use:

- **Number plates - number plate detection can be used to implement traffic rules, track cars in your taxi service parking, enhance security in public spaces, corporate buildings, malls, etc.**
- **Legal documents - Dealing with different forms of documents - affidavits, judgments, filings, etc. digitizing, databasing and making them searchable.**
- **Table extraction - Automatically detect tables in a document, get text in each cell, column headings for research, data entry, data collection, etc.**
- **Banking - analyzing cheques, reading and updating passbooks, ensuring KYC compliance, analyzing applications for loans, accounts and other services.**
- **Menu digitization - extracting information from menus of different restaurants and putting them into a homogeneous template for food delivery apps like swiggy, zomato, uber eats, etc.**
- **Healthcare - have patients medical records, history of illnesses, diagnoses, medication, etc digitized and made searchable for the convenience of doctors.**
- **Invoices - automating reading bills, invoices and receipts, extracting products, prices, date-time data, company/service name for retail and logistics industry.**

4. Goals and Objective:-

➢ to develop a table extraction tool which detect and decompose table information in a neat document.

➢ the objective of this project is to develop a system that incorporates the knowledge in text after extracting the texts from images.

5. OCR Tool Analysis:-

5.1 Problem Statement-

The problem here is for the OCR (Optical Character Recognition) systems to recognize characters in computer system when information is scanned through paper documents as we know that we have number of newspapers and books which are in printed format related to different subjects. Whenever we scan the documents through the scanner, the documents are stored as images such as jpeg, gif etc., in the computer system. These images cannot be read or edited by the user But to reuse this information it is very difficult to read the individual contents and searching the contents form these documents line-by-line and word-by-word. These days there is a huge demand in "Storing the information available in these paper documents in to a computer storage disk and then later editing or reusing this information by searching process.

The amount of data being collected is drastically increasing day-by-day with lots of applications, tools, and online platforms booming in the present technological era. To handle and access this humongous data productively, it's necessary to develop valuable information extraction tools. One of the sub-areas that's demanding attention in the Information Extraction field is the fetching and accessing of data from tabular forms.

To explain this in a subtle way, imagine we have lots of paperwork and documents where we would be using tables, and using the same, we would like to manipulate data. Conventionally, we can copy them manually (onto a paper) or load them into excel sheets. However, with table extraction, no sooner have we sent tables as pictures to the computer than it extract all the information and stacks them into a neat document. This saves an ample of time and is less erroneous

5.2 OCR implementation:-

To implement this OCR (Optical Character Recognition) all deep Learning part is based on PyTorch . Uses the EasyOcr python package to perform optical character recognition. As well execution of this tool is done through the google colab wih GPU enabled with CUDA toolkit.

Implementation step:-

- **Image processing**

  OCR tool does the certain pre-processing steps (gray scaling and etc.,) within its library and extracts the text..

- **Image display**

  This tool applies the CRAFT (Character Region Awareness for Text Detection) algorithm to detect the text. CRAFT is a scene text detection method to effectively detect text area by exploring each character and affinity between the characters. The recognition model uses CRNN.

- **Bounding boxes**

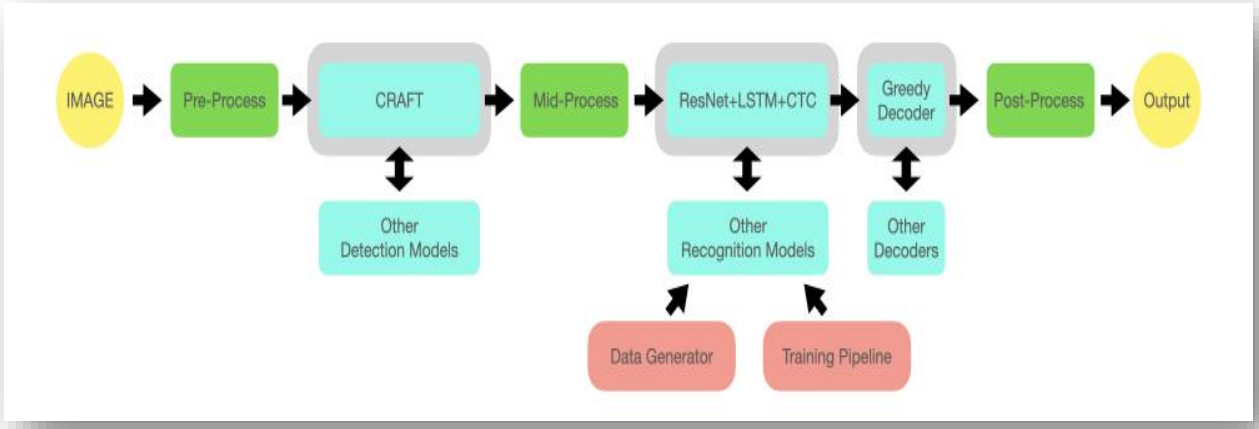
  The sequencing labeling is performed by LSTM and CTC (Connectionist Temporal Classification), here the CTC is meant for labeling the un-segmented sequence data with RNN.

- **Extraction**

  For table content using the pytesserect tool which mainly extract the useful content of tables only.

  Text extraction of bounded boxes through python of bounded box at some coordinates.

5.3 Framework Behind Implementation:-



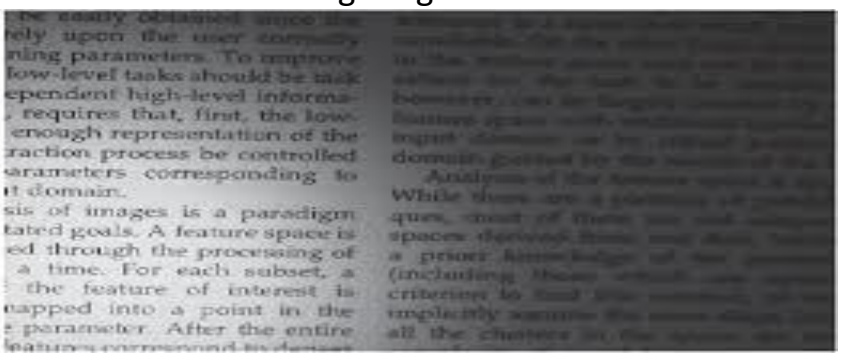
6. Challenges:-

For good quality and high accuracy character recognition, OCR techniques expect high quality or high resolution images with some basic structural properties such as high differentiating text and background. The way images are generated is an important and determining factor in the accuracy and success of OCR, since this often affects the quality of images dramatically. Usually OCR with images produced by scanners gives high accuracy and good performance. In contrast, images produced by cameras usually are not as good of an input as scanned images to be used for OCR due to the environmental or camera related factors.

1. **Scene Complexity:**

In a regular environment, we can see large numbers of man-made objects which are included in camera taken images such as paintings, buildings, and symbols. These objects have comparative structures and appearances to text which makes text recognition very challenging in the processed image. Text itself is
regularly laid out to encourage decipherability. The challenge with scene intricacy is that the surrounding scene makes it hard to segregate text from non-text.

2. Conditions of Uneven Lighting

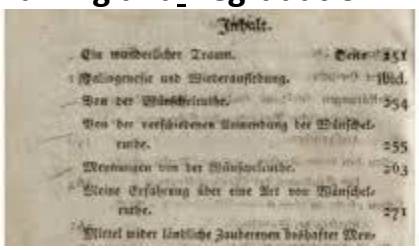


Oftentimes, taking images in natural environments results in uneven lighting and shadows. This poses a challenge for OCR as it degrades the desired characteristics of the image and hence causes less accurate detection, segmentation and recognition results . This condition with uneven lighting is what distinguishes a scanned image form one that is produced with a camera. The lack of such disparities in lighting and shadows makes scanned images preferred over camera images for their better characteristics and quality. Although using an on-camera flash may eliminate such problems with uneven lighting, it introduces new challenges.

3. **Skewness (Rotation)**

For optical character recognition systems, the point of view for the input image that taken from camera of hand-held device or Other gadgets that are used for taking images are not fixed like a scanner input, which skewing of text lines from their unique orientation might be observed. Great degree poor results will be observed when such a skewed image is fed to the OCR classifier.

4. Blurring and Degradation


Inhalt.

Ein wunderlicher Traum. Seite 251
Palingenesie und Wiederaufstehung. Ibid.
Von der Wünschelruthe. 254
Von der verschiedenen Anwendung der Wünschelruthe. 255
Meynungen von der Wünschelruthe. 263
Meine Erfahrung über eine Art von Wünschelruthe. 271
Mittel wider [illegible] Zauberey behafter Men-

Since working over a variety of distances are intended to numerous digital cameras, an important factor is the digital camera's focusing. For the best accuracy of character recognition and character segmentation, character sharpness is required. At large apertures and short distances, uneven focus can be observed when a small point of view changes. For the most part connected with photography, there are two kinds of obscure which is: out of focus obscure and movement obscure . At the point for catching a moving item, when the shade rate of the camera is not sufficiently high, the sensor gets presented to a continually changing scene. Accordingly, blurring will observed in parts in motion.

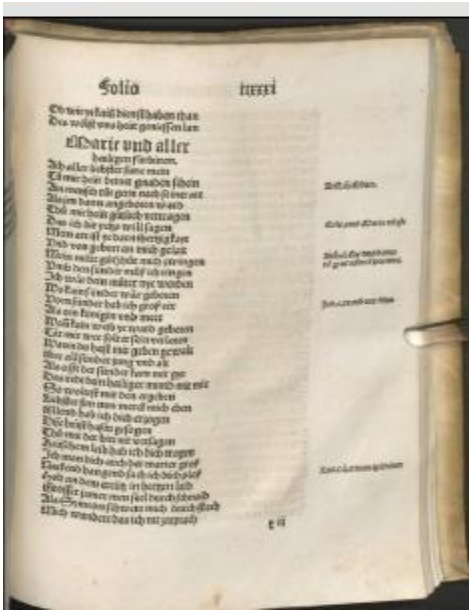
5. Aspect Ratios

Text has different aspect ratios. Text may be brief such as traffic signs, while other text may be much longer, such as video captions. Location, scale and length of text need to be considered with search procedure to detect text, which introduces high computational complexity.

6. Fonts

Italic style and script fonts of characters might overlap each other, making it difficult to perform some of the main OCR processes such as segmentation. Characters of various fonts have large within-class variations and form many pattern sub-spaces, making it difficult to perform accurate recognition when the character class number is large.

7. **WARPING**



Content or text on objects of varying geometries can be another challenge for OCR to be recognize when images of such situation captured by hand-held cameras. A few circumstances may emerge with flatbed scanners, wherein the twisted text observed when the content procured on picture, for example the content towards the binding of an extremely thick book. For convention paper documents, a technique for image dewarping is proposed by Ulges . By expecting the way that content lines are equally separated and parallel to each other,they dewarp pictures.

7.Result and Analysis:

For table content only:

Agent Type	Performance Measure	Environment	Actuators	Sensors
Medical diagnosis system	Healthy patient, reduced costs	Patient, hospital, staff	Display of questions, tests, diagnoses, treatments, referrals	Keyboard entry of symptoms, findings, patient's answers
Satellite image analysis system	Correct image categorization	Downlink from orbiting satellite	Display of scene categorization	Color pixel arrays
Part-picking robot	Percentage of parts in correct bins	Conveyor belt with parts; bins	Jointed arm and hand	Camera, joint angle sensors
Refinery controller	Purity, yield, safety	Refinery, operators	Valves, pumps, heaters, displays	Temperature, pressure, chemical sensors
Interactive English tutor	Student's score on test	Set of students, testing agency	Display of exercises, suggestions, corrections	Keyboard entry

Figure 2.5 Examples of agent types and their PEAS descriptions.

we list the dimensions, then we analyze several task environments to illustrate the ideas. The definitions here are informal; later chapters provide more precise statements and examples of each kind of environment.

FULLY OBSERVABLE
PARTIALLY OBSERVABLE

Fully observable vs. **partially observable**: If an agent's sensors give it access to the complete state of the environment at each point in time, then we say that the task environment is fully observable. A task environment is effectively fully observable if the sensors detect all aspects that are *relevant* to the choice of action; relevance, in turn, depends on the performance measure. Fully observable environments are convenient because the agent need not maintain any internal state to keep track of the world. An environment might be partially observable because of noisy and inaccurate sensors or because parts of the state are simply missing from the sensor data—for example, a vacuum agent with only a local dirt sensor cannot tell whether there is dirt in other squares, and an automated taxi cannot see what other drivers are thinking. If the agent has no sensors at all then the environment is **unobservable**. One might think that in such cases the agent's plight is hopeless, but, as we discuss in Chapter 4, the agent's goals may still be achievable, sometimes with certainty.

UNOBSERVABLE

SINGLE AGENT
MULTIAGENT

Single agent vs. **multiagent**: The distinction between single-agent and multiagent en-

Gray-scale image:

Agent Type	Performance Measure	Environment	Actuators	Sensors
Medical diagnosis system	Healthy patient, reduced costs	Patient, hospital, staff	Display of questions, tests, diagnoses, treatments, referrals	Keyboard entry of symptoms, findings, patient's answers
Satellite image analysis system	Correct image categorization	Downlink from orbiting satellite	Display of scene categorization	Color pixel arrays
Part-picking robot	Percentage of parts in correct bins	Conveyor belt with parts; bins	Jointed arm and hand	Camera, joint angle sensors
Refinery controller	Purity, yield, safety	Refinery, operators	Valves, pumps, heaters, displays	Temperature, pressure, chemical sensors
Interactive English tutor	Student's score on test	Set of students, testing agency	Display of exercises, suggestions, corrections	Keyboard entry

Figure 2.5 Examples of agent types and their PEAS descriptions.

we list the dimensions, then we analyze several task environments to illustrate the ideas. The definitions here are informal; later chapters provide more precise statements and examples of each kind of environment.

FULLY OBSERVABLE
PARTIALLY OBSERVABLE

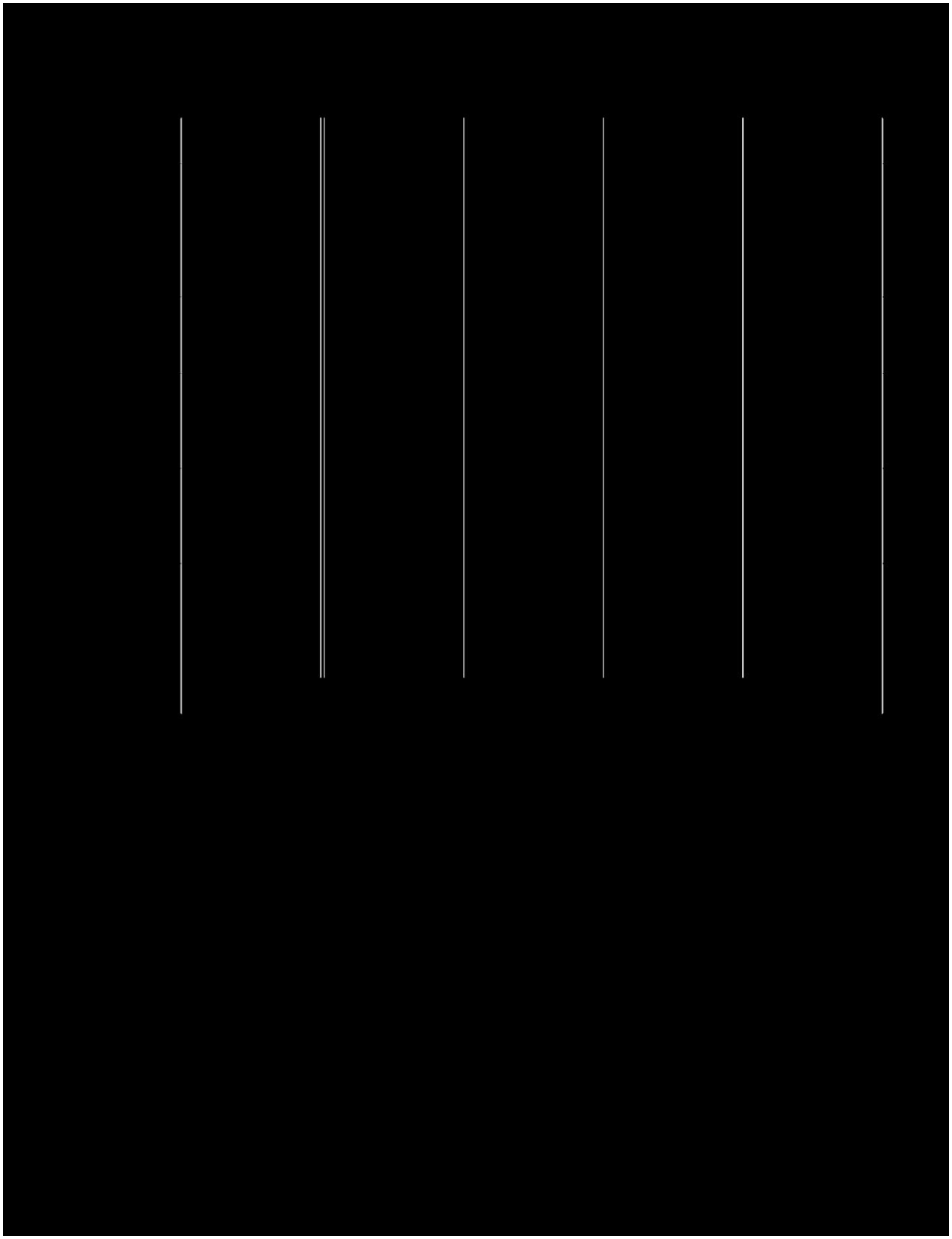
Fully observable vs. **partially observable**: If an agent's sensors give it access to the complete state of the environment at each point in time, then we say that the task environment is fully observable. A task environment is effectively fully observable if the sensors detect all aspects that are *relevant* to the choice of action; relevance, in turn, depends on the performance measure. Fully observable environments are convenient because the agent need not maintain any internal state to keep track of the world. An environment might be partially observable because of noisy and inaccurate sensors or because parts of the state are simply missing from the sensor data—for example, a vacuum agent with only a local dirt sensor cannot tell whether there is dirt in other squares, and an automated taxi cannot see what other drivers are thinking. If the agent has no sensors at all then the environment is **unobservable**. One might think that in such cases the agent's plight is hopeless, but, as we discuss in Chapter 4, the agent's goals may still be achievable, sometimes with certainty.

UNOBSERVABLE

SINGLE AGENT
MULTIAGENT

Single agent vs. **multiagent**: The distinction between single-agent and multiagent en-

Vertical line detection:



Horizontal line detection:

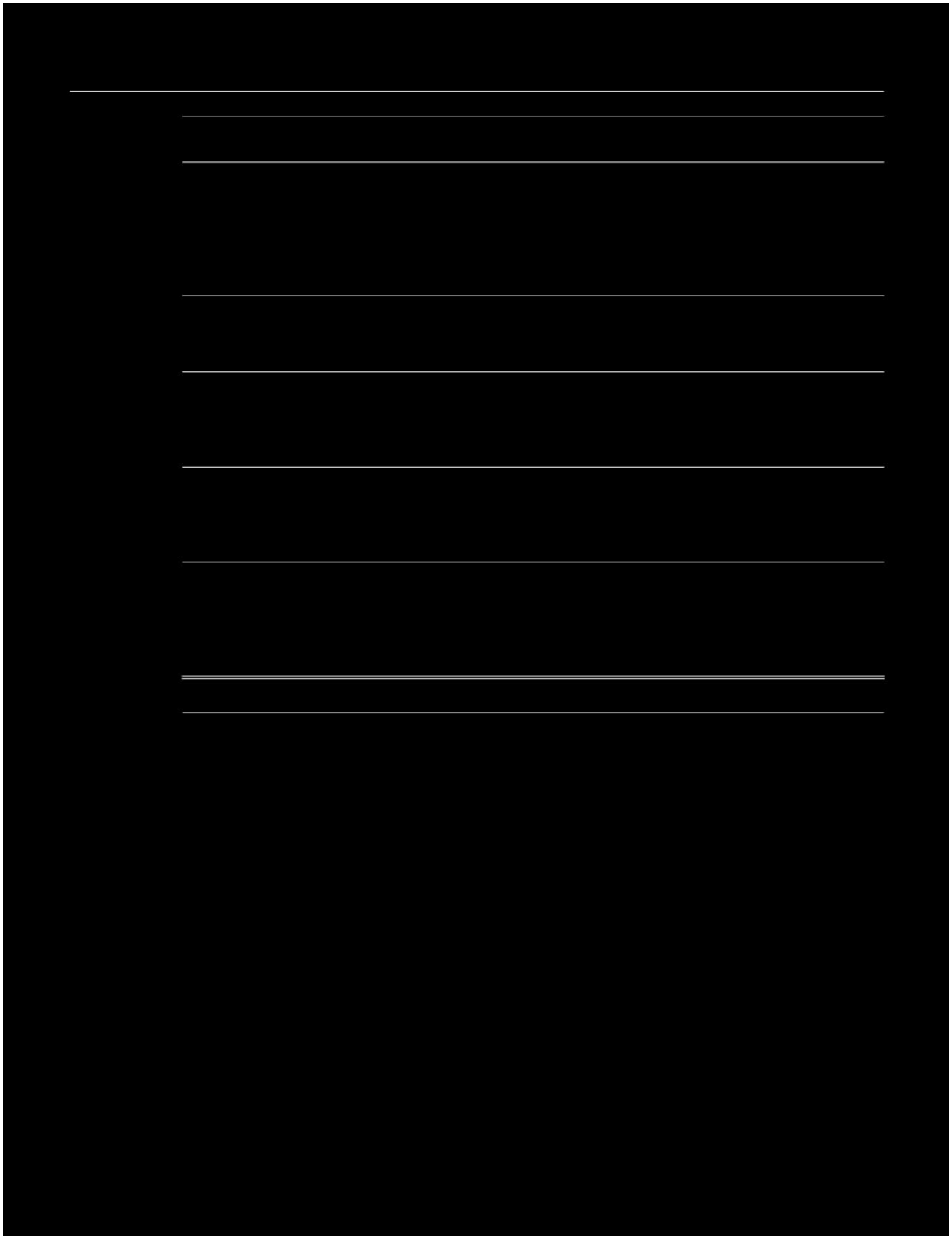


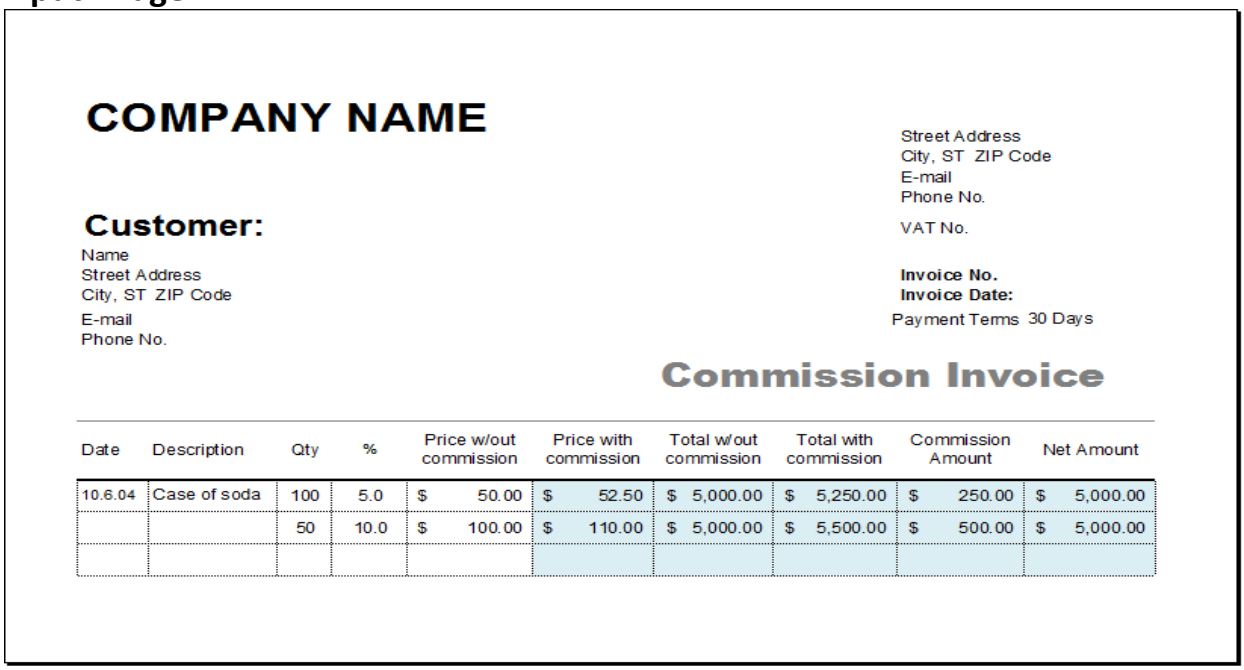
Table detection:

Extracted result in frame:

```
0   Performance\nMeasure\n Healthy patient,\nred...
1                  Correct image\ncategorization\n
2          Percentage of\nparts in correct\nbins\n
3                       Purity, yield,\nsafety\n
4                    Student’s score\non test\n

                                                3  4  \
0   Environment\n Patient, hospital,\nstaff\n
1          Downlink from\norbiting satellite\n
2           Conveyor belt\nwith parts; bins\n
3                     Refinery,\noperators\n
4           Set of students,\ntesting agency\n

                                                   5  \
0   Actuators\n Display of\nquestions, tests,\nd...
1               Display of scene\ncategorization\n
2                         Jointed arm and\nhand\n
3              Valves, pumps,\nheaters, displays\n
4   Display of\nexercises,\nsuggestions,\ncorrect...

                                                   6
0   Sensors\n Keyboard entry\nof symptoms,\nfind...
1                            Color pixel\narrays\n
2                  Camera, joint\nangle sensors\n
3       Temperature,\npressure,\nchemical sensors\n
4                                 Keyboard entry\n
   Unnamed: 0                                      0
0           0    Agent Type\n Medical\ndiagnosis system\n
```

*For Whole Content:

Input image



COMPANY NAME

Street Address
City, ST ZIP Code
E-mail
Phone No.

VAT No.

Customer:

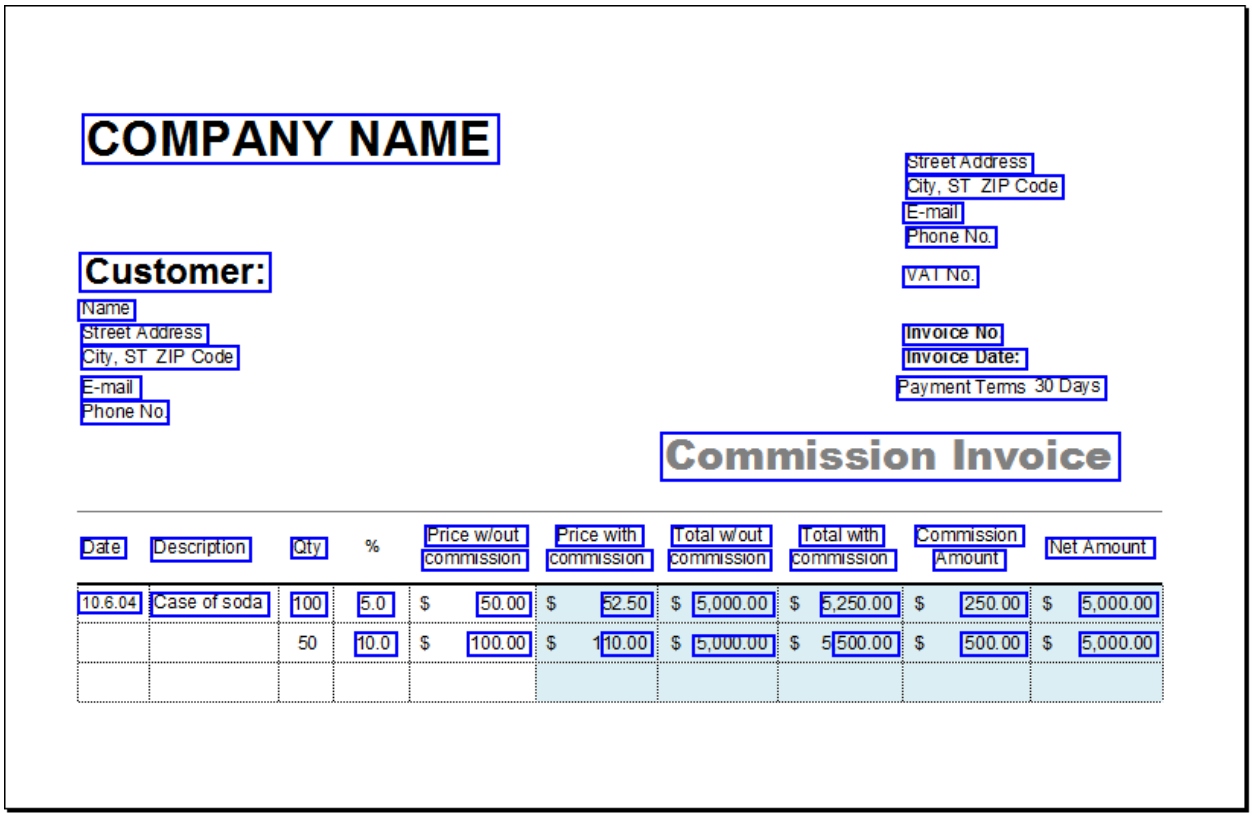
Name
Street Address
City, ST ZIP Code
E-mail
Phone No.

Invoice No.
Invoice Date:
Payment Terms 30 Days

Commission Invoice

Date	Description	Qty	%	Price w/out commission	Price with commission	Total w/out commission	Total with commission	Commission Amount	Net Amount
10.6.04	Case of soda	100	5.0	$ 50.00	$ 52.50	$ 5,000.00	$ 5,250.00	$ 250.00	$ 5,000.00
		50	10.0	$ 100.00	$ 110.00	$ 5,000.00	$ 5,500.00	$ 500.00	$ 5,000.00

Bounded text



COMPANY NAME

Street Address
City, ST ZIP Code
E-mail
Phone No.

VAT No.

Customer:

Name
Street Address
City, ST ZIP Code
E-mail
Phone No.

Invoice No
Invoice Date:
Payment Terms 30 Days

Commission Invoice

Date	Description	Qty	%	Price w/out commission	Price with commission	Total w/out commission	Total with commission	Commission Amount	Net Amount
10.6.04	Case of soda	100	5.0	$ 50.00	$ 52.50	$ 5,000.00	$ 5,250.00	$ 250.00	$ 5,000.00
		50	10.0	$ 100.00	$ 110.00	$ 5,000.00	$ 5,500.00	$ 500.00	$ 5,000.00

Extracted Text

COMPANY NAME

Street Address

Citj,

ST ZIP Code Email

Phone Na. customer: MATIo. Name

Street Address

Invoice Io

Citj, ST ZIP Code

Invoice Date:

E mail

Payment Terms 30 Days

Phone No

Commission Invoice

Price lgut

Price with

Total out

Tatal iith

Commissian

Date

Description

Aty

Net Amount

Commissian

commissian

Commision

Commision

Amaunt

10.6.04

Cost of soda

100

5.0

50.00

52.50

5,000.00

5,250.00

250.00

5,000.00

10.0

1o0.00

0.00

5u0.00

5u0.00

5,000.00

5,uu0.00

8. Future Scope:

- **Analyzing the behaviour on multiple language.**
- **To improve accuracy.**
- **To fasten the OCR speed.**

Analyzing the behaviour on multiple language:

Effort will be concentrated on enabling generic multi-lingual operation such that negligible customization is required for a new language beyond providing a corpus of text. Although change was required to various modules, including physical layout analysis, and linguistic post-processing, no change was required to the character classifier beyond changing a few limits.

To improve accuracy:

OCR tool may face many challenges as mentioned in challenge section to overcome these challenges image processing process will apply to improve the accuracy changes have on low quality image like sharp character borders, high contrasts, well aligned characters and less pixel noise as possible.

To fasten the OCR speed:

To enhance the speed of this tool already running under the GPU machine environment with enabled CUDA toolkit this accessibility makes it easier for specialists in parallel programming to use GPU resources. The challenges facing during pre-processing may take time which resolve through image processing as a result somehow speed got enhanced.

9. References:

- **https://en.wikipedia.org/wiki/Optical_character_recognition**
- **HPE Haven "OCR Document". Archived from the original on April 15, 2016**

- **https://nanonets.com/blog/table-extraction-deep-learning/**
- **http://www.dlib.org/dlib/march09/holley/03holley.html**
- **https://www.explainthatstuff.com/how-ocr-works.html**
- **https://www.researchgate.net/publication/232619447_A_Complete_Optical_Character_Recognition_Methodology_for_Historical_Documents**
- **https://indiaai.gov.in/article/optical-character-recognition-explained**