OPTICAL CHARACTER RECOGNITION FOR BUSINESS CARDS

Archana Chaudhari[1], Praveen Pol, Varun Manwatkar, Amit Mohite, Krushna More, Shruti Mane and Kuber Markad

Department of Instrumentation and Control Engineering, Vishwakarma Institute of Technology, Pune, India

Abstract

The paper focuses on the application that can extract the data from the business card and present it to the user. Further, the user can also manipulate the data according to his perspective. Business cards have been used deliberately nowadays in many industries and companies. Also, the information can be further used to get the contacts. The technology that has been used to extract this information from the image is OCR which is Optical Character Recognition. Further, the NER from NLP is used for tagging the data and predicting it. The paper further tells how one can access the data from the business card from a web application.

Keywords:

Optical Character Recognition, Natural Language Processing, Named Entity Recognition, text recognition, business card.

1. INTRODUCTION

1. 1 BACKGROUND

Text recognition is a growing technology nowadays in the case of data analysis and various other works. Identifying texts from various documents, getting the data from them, and manipulating them is easy. But if the text is in the form of an image, then many systems fail to even detect the document. OCR that optical character recognition helps here. If there are several hard copies of textual images in the job and one must analyse them then OCR also helps in extracting the text from those images. There can be several times when soft copies of images are non-editable.

In all the above cases or situations, OCR works best. It is basically a tool that can quickly convert files of images into readable and editable text versions. The advantages of using OCR to extract the data are fewer human efforts, easy edits, prevention of human errors, etc. Time and money can be saved by reducing the paperwork with the help of OCR. In business, the main communication medium is giving business cards to each other. In this way, companies and industries also communicate with each other. In this case, if anyone wants to manipulate the data from the business card or want to even store it for future use then that can be possible with the help of OCR. As the name itself suggests, it only works on images consisting of texts. It detects the image, converts its parts into text, processes it and stores it for further use.

1.2 MOTIVATION, OBJECTIVE and CONTRIBUTION

There have been so many models of OCR even for business card recognition. But some fails in accuracy and some has not been deployed yet. The purpose of deploying the model is that it will be easy for the user to deal with the data. Developing a good model and converting into a simple web application can result in a good project.

As we know that a traditional business card consists of the name of the person, organization, contact info and the e mail id. The objective of the project is to extract the name, organization, contact and the e-mail from the business card. These are called as entities. The objective includes choosing a perfect method or technology to train the model, getting the best results out of it, etc.

The model thus created has a good accuracy. Also there has not been any OCR system that has been deployed on web. The uniqueness of the project is that it allows you to upload any type of business card which can be of any layout. The model itself corrects the orientations of the business cards and thus sends it for the further execution.

Applications such as banking, insurance papers, mortgage applications, and loan documents widely use this OCR technology. This paper focuses mainly on the application that is business cards. A system has been made in which an image of any business card is given by the user and that image undergoes several steps of the model to predict the data and to generate a final editable excel sheet or file.

2. RELATED WORK

Business Card recognition technique has been developed a lot in the past years. New methods based on character segmentation has been developed. There are some models in which linguistic approach has been also developed that can verify and modify the results. [1] In the recent years, for, IOS users, business card reader applications have also been created that are based on Tesseract. As the resolution of the phone camera has been increased it has become very easy to upload an image with a correct resolution and get the data from it. [2]

There have been several projects and research done on the same. Optical Character Recognition on images with

colourful backgrounds is one of the interesting works done by Matteo and Ratko in 2018. They have presented a pre-processing method to improve Tesseract OCR performance on images with colourful backgrounds. The method consists of segmentation steps and an image classification step. The problem with the colourful background of the images is resolved by them with 95.5% accuracy. [3]

There have been some situations where Hindi text mustextract which became difficult because of the Devnagari script. This work is done by Veena Bhansal and R.M.K. Sinha in 2016. They tested the Devnagari Document Reading System on various printed documents and they got a performance of approximately 93%-character level. [4]

Later some similar kinds of projects have been done such as a bilingual OCR system for Hindi-Telugu documents. [5] For Marathi documents, some systems have been developed as well in which handwritten Marathi compound character recognition is possible. [6]

Some OCR-based video text recognition technologies have been developed as well. The edge analysis method is used to extract the superimposed text in the video. [7] A system has been developed for Optical Character Recognition of the scientific equations by Sagar and Lakhan in 2022. This is the latest system that they have developed and had also made an application for the same. The application allows the user to take a snip of their present screen. After processing that image, the latex code of that equation can be obtained. [8]

There have been several review papers made as well some of which compare the performances of various OCR models and some of them also studied the technologies that are required for an OCR model or system to be made. Studies found that neural networks work best in the case of OCR execution. [9]

The advancements in OCR for Indian scripts have been achieved rapidly. Different optical character recognition systems have been developed to read and manipulate Indian Scripts. These systems have been studied together in a research paper. It focuses on the advancements in this field. [10] While dealing with the OCR model for extraction of the data, several errors may generate. These errors can affect the Natural Language Processing model. Sentence boundary detection, tokenization, and part-of-speech tagging are some of the parts of the model where these errors may generate. All these are analysed and signified by DanielLopresti. [11]

There is a lot in the field of OCR and its hardware implementation. The system works best for the documents. A new approach to developing hardware has been developed. The research work in this field also includes the comparison between the Simulink model of the system and its hardware using FPGA. [12] Handwritten and printed text recognition is possible using tesseract-OCR.The system thus developed has good accuracy on printed and typed text. Training the dataset of the tesseract model at a higher level can increase the accuracy of the model. [13]

There are so many techniques in developing an OCR model. Improving them leads to a good rate of accuracy. Summarization of these technologies is important to get a proper idea of the model. Different aspects and various issues revolving in this field have been summarized in the review paper.[14]

In the field of handwritten optical character recognition, there has been a lot in the literature reviews. A lot of SLRs have been made in order to classify the methods that come under OCR. Handwritten documents are difficult to process and manipulate. The OCR is the best platform to proceed with them.

We all know that in the field of science there has been a lot of new inventions increasing day by day. This includes making a review of the research that has been done till now. It becomes so difficult to go through the pages of the loads of the documents. Here, OCR does the best job to extract the information from the documents and make the work easy.

The whole setup gives an accuracy of nearly 70 per cent with business cards. The system differs from other systems in development and handling. A fully developed web application has been made to make the system user-friendly. Detecting the data correctly, extracting it and making it available for the user for manipulation are the several benefits of this system. Further, the model can detect many kinds of layouts and structures as there can be many depending on image dimensions. The system is also able to detect documents other than business cards. The whole setup generates an editable excel sheet which can be further used for the manipulation of the data.

There has been a great evolution of Optical Character Recognition fora long back. It was first invented by Emanuel Goldberg. He had developed a machine which was able to read the characters and convert them into telegraph. Retrieving the records such as financial, etc. was possible then. Since that time this OCR technology had grown up and is growing rapidly with its accuracy.

3.FLOWCHART

The following figure shows the process or flow of the proposed model or system. [Fig.1]

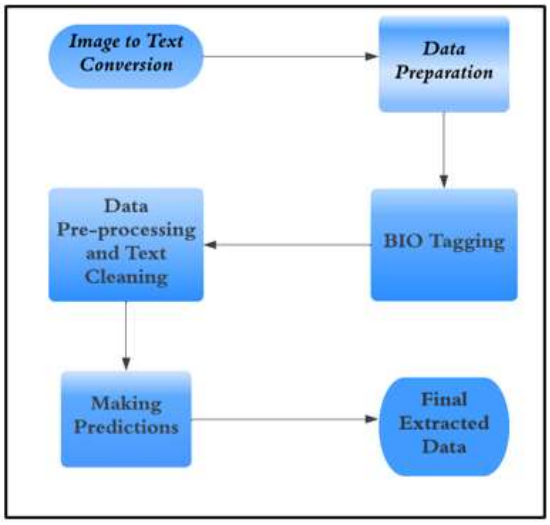

Fig. 1 Flow of the proposed model

The above flowchart shows the flow of the proposed system or model. Conversion of the image into text is the first and basic thing the model starts with. The data thus got from the first step is then used for the data preparation process which includes basic data segregation. BIO tagging is done on the data that has been provided by the second step. Cleaning of the data is significantly important as it can contain white spaces, etc. Making predictions based on the text that has been cleaned is the last step ending with the final data that is provided by the system in an excel sheet. The objective of making the flowchart is to understand the process precisely. Because of this, the ideology got precise and also to present the model it helps a lot. All these processes have been discussed briefly in the methodology below.

4.PROPOSED METHODOLOGY

The traditional OCR model works on the basic NLP that us Natural Language Processing. It includes two stages i.e., image processing and extraction of the data from the physical form of the document. Image processing includes uploading the image of the physical document into the model, processing of the image, etc. The extraction of the data can be done using NLP which lets the model extract whatever data is present on the physical document or to be specific the business cards. The model discussed in this paper works parallelly but the data processing steps are somewhat different. The model is discussed below. It is divided into several steps i.e., extracting text from image, text cleaning, model training, making predictions, etc.

4.1 MODEL ARCHITECTURE:

4.1.1DATA COLLECTION:

For a machine learning model to be made, the data set has to be collected to train and to test the model. Also, the more data for training you have the more the accuracy can be made. Various images of the business cards have been found out for the training and the testing purpose. Traditionally, a business card consists of the contact information, name of the organization and about. Accessing this data in the form of text is a great task.

The model has been divided into training and testing tests. For this, the business card data has been collected of various kinds. There are about 300 images of diverse business cards collected from various websites. These scanned copies of the business cards with unique names are loaded into the Jupyter Notebook. Either OpenCV or PIL i.e., Python Imaging Library can be used to load the images into the Jupyter notebook. PIL can read the image and the object and it can also look into the type of the same.

4.1.2EXTRACTION OF THE TEXT FROM IMAGE

To get the text from an image it must be first converted into text. This is where the role of an OCR model comes into picture. Optical Character Recognition is used to convert the desired image into text. Not only image it is also applicable for all the physical documents. The desired documents must go under the scanning process. The data can be extracted then. There are methods to convert an image into text. One of them includes uploading a file to google drive and opening it with the google docs. The image file can be converted but the format might differ. The fonts, bolds and italics may be retained in some of the documents and in some they may not. OCR knocks out all these limitations and successfully converts the desired file into a text.

In the proposed model, the image is converted into text using the PyTesseract package from python. PyTesseract is an OCR-based tool or engine used in python. It supports both the OpenCV and the PIL Library. Anyone with the PyTesseract can be used to extract the text from the image. Text can be converted into string and can be extracted using this package. It contains some steps which include finding the text, classifying it. From finding the pages and layouts to detecting the words from the lines, it works best. Fig. 2 shows the architecture for the same. [Fig.2]

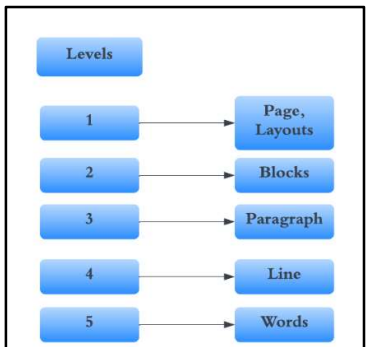

Fig. 2 Hierarchy of PyTesseract

There are some levels in the architecture of the PyTesseract package i.e., level 1 to level 5. Starting from the level 1, it detects the pages of the uploaded document as there can be many. For example, if there are two pages then the package has to work for both the pages. But as the business card

contains only one page then the package will work only on one page. Further the text is recognized as the package undergoes the block level, paragraph level, line level and letter level. The letters are then passed to the machine learning model or deep learning model and can be then further used for the data cleaning and so on. The classification model can then be able to detect of the letter is a number or an alphabet or any special character.

4.1.3CLEANING OF THE TEXT:

Data preparation traditionally means getting the data ready to perform the data processing on it. Drawing the bounding boxes on top of the images and all the levels can be done in order to get the cleaned text. According to these bounding boxes each word can be stored with a unique image id and then can be extracted into a csv file for the ease of handling.

Labelling of the data is important as there can be so many words of each and every image and cannot be classified. For this, BIO tagging is done to tag the words. The model cannot distinguish this data, so we need to pre-process the data in a specific format. This can be done by accommodating the starting and ending position and also the size of the particular word.

Further, the cleaning of the text can be done by removing all the white spaces and special characters such as, %, !,&. As @ is used in the e-mail address and also the / can be used while mentioning the two different contact numbers, these cannot be removed. Spacy is used to train the NER (Named Entity Recognition) model. Spacy v1.3 is used for training.

In the proposed model, NER, that is Named Entity Recognition model, is trained in Spacy. So, the final cleaned data has to be converted into Spacy format. The purpose of using Spacy to train NER model is that it provides the accurate analysis of the NLP library. As there are multiple images of the business cards, so after the BIO tagging, all the texts and words from one image are grouped together according to their image ids. Here, B and I tags are considered. The positions for each and every word can be defined.

E.g., If there is a phone number then it can have (2,10, B-PHONE) which means it starts at position 2, ends at position 10 and this is the beginning of the phone number. Further the training and testing sets can be created. And all the functions can be called together to make the predictions.

4.2 IMPLEMENTATION:

The whole setup has been done in Jupyter. The technologies used here are computer vision for reading the images and NLP that is Natural Language processing. Environment used is Python. Firstly, all the required packages have been installed. E.g., PyTesseract, tesseract OCR. Spacy is also installed which is used later for the training purpose.

The image must undergo some steps. The image is first scanned. The image is then refined as it can be of any dimension and layout. It is converted into a specified standard format. Binarization is then done in which the text is recognized and aligned in a specific order. Identifying the characters is the next step. It matches the pixels of each scanned letter with the existing database of fonts. Later it ensures accuracy by reducing errors. Lastly, it provides a readable and editable text file, in a form such as an excel sheet, to the user. With this file, the user can also manipulate the data.

The setup for the model is as follows:

Collecting various business card data→ Extracting the text from them→ Cleaning it→ Labelling it manually→ passing it to the NER model to get the results.

First the data has been collected from various websites. Around 300 images of various business cards having different blocks and layouts have been obtained and stored by naming them with a unique image id. The scanned copies of the business cards are stored in a folder.

Now, to start with the implementation and making of a model, the image is to be loaded into the notebook. This is done using PIL. It read the image and object. Using PyTesseract the data is extracted and is converted into a string. As discussed above, various levels and respective words have been then extracted and identified.

Level 5 defines words. Level 2 defines the block number. As shown below, a table is created for the easy dealing of the data. There are levels from 1 to 5. Each of which indicating layout, block, paragraph, line and word respectively. [Fig.3] E.g., HONDA is the word and thus indicated by level 5. It is present in page 1, block 1, paragraph 1, line 1 and word 1. Its height, width and position on the entire business card is also defined using pixels. Bounding box around all the levels are drawn. [Fig.3]

	level	page_num	block_num	par_num	line_num	word_num	left	top	width	height	conf	text
0	1	1	0	0	0	0	0	0	720	401	-1	
1	2	1	1	0	0	0	36	29	207	25	-1	
2	3	1	1	1	0	0	36	29	207	25	-1	
3	4	1	1	1	1	0	36	29	207	25	-1	
4	5	1	1	1	1	1	36	29	207	25	95	HONDA
5	2	1	2	0	0	0	237	166	246	63	-1	
6	3	1	2	1	0	0	237	166	246	63	-1	
7	4	1	2	1	1	0	237	166	246	16	-1	
8	5	1	2	1	1	1	237	166	112	16	96	DONNIE
9	5	1	2	1	1	2	364	166	119	16	96	HANSEN
10	4	1	2	1	2	0	245	194	231	12	-1	
11	5	1	2	1	2	1	245	195	108	11	96	MOTOCROSS

Fig.3 Table showing PyTesseract Frames

In the proposed model, till now the rough data of words have been collected from all the images. These words have been stored as data frames in an excel sheet having columns as image ids and text. All the words are separately stored with the proper image ids. Even the white spaces have been given a unique image id and are stored. Figure 4 shows the data frames.

	A	B
1	id	text
2	000.jpeg	
3	000.jpeg	.
4	000.jpeg	040-4852
5	000.jpeg	8881,
6	000.jpeg	90309
7	000.jpeg	52549
8	000.jpeg	Fi

Fig. 4 Data frames in Excel Sheet

The texts including the special characters, numbers and words have been separated and stored with their corresponding image ids.[Fig.4] The data i.e., texts, numbers, e-mails have been stored and made ready for the data processing. This manipulated excel sheet is then further passed for data pre-processing.

Tagging is done using all the words in capital just to get it classified from the real entities. BIO Tagging is done. [Fig.5]Here, B means beginning, I mean inside and O mean outside. B prefix before a tag indicates the beginning of the word. I indicate that the particular part is inside the specified chunk. O indicates that the it is an empty chunk or there is no entity present. E.g., B-ORG means the name of the organization starts from this position. The tagging has been done manually as there is no any other way of doing that. Each and every word of each and every entity has been tagged manually. Figure 5 shows the BIO tagging.

A	B	C
id	text	tag
000.jpeg		O
000.jpeg	.	O
000.jpeg	040-4852	B-PHONE
000.jpeg	8881,	I-PHONE
000.jpeg	90309	B-PHONE
000.jpeg	52549	I-PHONE
000.jpeg	Fi	O
000.jpeg	/laurelsoverse	O
000.jpeg	‰ÛÏ@:	O
000.jpeg	LAURELS	B-ORG
000.jpeg	OVERSEAS	I-ORG

Fig.5 BIO Tagging

Excluding the @ and /, all the white spaces and the null values have been removed using the predefined functions. The final cleaned data is then extracted. Using the official documentation of Spacy, the NER model is trained. Further, for the training and testing purpose, two datasets have been created with the ratio 90:10. 240 cards are for the training purpose and the remaining are for the testing purpose. For the ease of the handling in Jupyter Notebook, these files have been converted into pickle files.

4.3 MAKING PREDICTIONS:

To identify the name, number etc. in a text, NER is used from Spacy. NER which is Named Entity Recognition helps to identify the key elements in a text easily. It detects the named entities and classifies them into a set of predefined categories. A new data frame is added which is a token in which a value is stored which is associated with the tags.

According to the tag names, the bounding boxes have been drawn around each word separately. Similar tags have been then grouped together. Each value associated with the tags has been then parsed to categorize the names, emails, and phone numbers. [Fig.5] The beginning, end and inside tags for the same key have been then found out and joined together. The same data has been then stored in the excel sheet.

Accuracy of the model.

```
========================= Training pipeline =========================
ℹ Pipeline: ['tok2vec', 'ner']
ℹ Initial learn rate: 0.001
E    #       LOSS TOK2VEC  LOSS NER  ENTS_F  ENTS_P  ENTS_R  SCORE
  1     200         814.43   5478.80   45.06   53.57   38.89    0.45
  3     400         898.72   3326.67   61.16   69.16   54.81    0.61
  5     600        1880.83   1910.93   59.78   59.56   60.00    0.60
  8     800         562.15   1299.74   60.95   62.75   59.26    0.61
 12    1000         549.38    921.39   64.79   70.93   59.63    0.65
 17    1200         574.73    681.16   61.41   69.81   54.81    0.61
 23    1400         738.35    617.56   61.86   69.77   55.56    0.62
 30    1600         560.75    506.99   61.54   64.00   59.26    0.62
 39    1800         819.27    581.80   64.27   69.70   59.63    0.64
 50    2000         681.04    504.68   63.60   63.14   64.07    0.64
 63    2200         784.13    513.58   63.01   66.80   59.63    0.63
 80    2400         729.80    568.98   62.28   67.53   57.78    0.62
 96    2600         631.30    483.66   66.67   72.81   61.48    0.67
113    2800         678.31    456.91   62.91   66.12   60.00    0.63
130    3000         760.78    477.66   61.89   69.27   55.93    0.62
146    3200         647.88    451.74   63.10   62.87   63.33    0.63
163    3400         633.86    413.86   62.75   66.67   59.26    0.63
180    3600         630.00    416.55   60.38   61.54   59.26    0.60
196    3800         585.14    403.63   64.16   68.94   60.00    0.64
213    4000         369.96    357.21   63.58   69.60   58.52    0.64
230    4200         890.56    435.80   64.94   70.26   60.37    0.65
✔ Saved pipeline to output directory
output\model-last
```

Fig.6 Accuracy score

The model has a good accuracy of nearly 65 percent. This can be improved by two methods. First one includes the feeding of the data. As the Spacy is the BERT architecture model, the better accuracy can be achieved by feeding more and more data into the model. Second method is to change the data preparation format. But this is not a guaranteed method. As we know that the organization names are in Capital letters, also the name of the person starts with a capital letter, so if we let this known to our model then the accuracy can be improved by nearly 3 to 4 percent.

A web application is created to make the process simple and user-friendly. The user has to just upload the file after adjusting it with a proper layout and on the next page user will get all the predictions in the form of a table.

5. RESULTS AND DISCUSSIONS

Figure 6 shows the predictions made by the model. [Fig.6]

```
In [5]: entities
Out[5]: {'NAME': ['Dr T S Reddy'],
         'ORG': ['Lea Associates South Asia Pvt Ltd'],
         'DES': ['Senior Consultant'],
         'PHONE': ['91', '66747135', '9182230'],
         'EMAIL': ['tsr@lasaindia.com', 'limmappagari@ymait.com'],
         'WEB': ['www.lasaindia.com']}
```

Fig.6 Predictions

To get the results, the user must go through several steps. First, the path in which the image is stored must be given by the user into the model. Then that respective image must undergo the steps that have been discussed earlier like data pre-processing, data cleaning, etc. The figure below shows the final data obtained in an excel sheet. [Fig.7]

A	B	C	D	E	F
NAME	ORG	DES	PHONE	EMAIL	WEB
['Thathine	['Life Insu	['Insuranc	['8099948	['lictsrikan	[]

Fig.7The Final Data

All the data sought till now is then obtained in an excel sheet. [Fig.6] The excel sheet can be named by the user's choice. It contains all the extracted data from the business card. The excel sheet formed at the last can be then used to manipulate the data if one wants. The following figures show the created web application and its web pages. [Fig.8 -Fig.10]

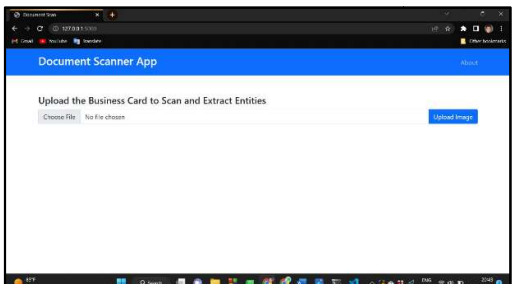

Fig.8 The Web Application

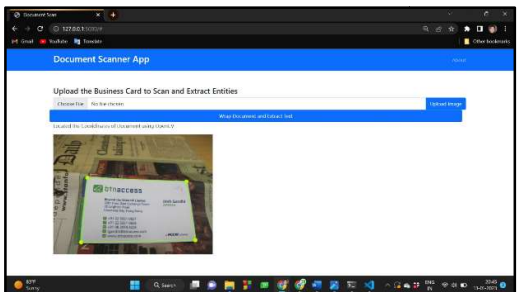

Fig.9 Image Layout Adjustment

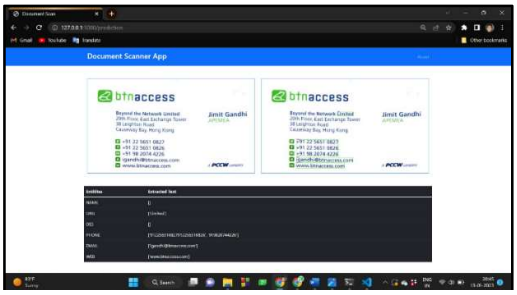

Fig.10 Final Extracted Data

There are several web pages in the application. The first one is the file upload page where the user must upload the image of the business card. [Fig.8]After this the model will detect the image and it will be shown on the very next web page.The next step is to adjust the image of the business card if it is not fitting the layout. [Fig.9] The user can easily make changes there. On the third page, the extracted data from the business card will be detected and will be displayed on the screen. [Fig.10] A simple GUI is made just to make the process user-friendly and precise.

6. CONCLUSION

As discussed till now, the OCR system has been developed to extract the data from the business cards. It converts the image of the text into a readable text format. The extracted text can be then used for several purposes. The developed system has several advantages. The image given by the user can be of any layout or any dimension. This model can detect that layout and then convert them into its standard format and then move it to the actual system.

There are several elements on the business card such as logos, etc. These elements cannot be detected by the system as it only extracts data from the image containing text. The system developed till now can be useful only for business cards. The system can detect the other document types and can also undergo the operations but the accuracy in case of them is very low. The predicted data may or may not be correct. While in the case of business cards, the accuracy of the model is nearly 70 per cent. This accuracy can be increased if one can add more and more images to the datasets.

There have been several applications, of several kinds of systems, such as in banks - to extract the text from various cheques and documents, etc. The proposed system can be modified further in case of developing and handling. A web application is also created to make the system more user-friendly.

REFERENCES

[1] An OCR System for Business Cards, Hisashi Saiga, Yasuhisa Nakamura, Yoshihiro Kitamura, Toshiaki Morita, IEEE,2011.

[2] A Business Card Reader Application for iOS devices based on Tesseract, Bello Dangiwa, Smitha S Kumar, ResearchGate, 2018

[3] Optical Character Recognition on images with colourful background, Matteo Brisinello, Ratko Grbic, Dejan Stefanovic, IEEE, 2018.

[4] A complete OCR for Printed Hindi Text in Devnagari Script, Veena Bansal, R.M.K. Sinha, IEEE, 2001.

[5] A Bilingual OCR for Hindi-Telugu Documents and its 5pplications, C. V. Jawahar, M. N. S. S. K. Pavan Kumar, S. S. Ravi Kiran, IEEE, 2003.

[6] Handwritten (Marathi) Compound Character Recognition, Minakshi Bhandare, Annuradha Kakade, IEEE, 2015.

[7] Research on Video Text Recognition Technology Based on OCR, Ding Jie, Zhao Guotao, Xu Fang, 10th International Conference on Measuring Technology and Mechatronics Automation, 2018

[8] Optical Character Recognition (OCR) For Scientific Equation, Sagar Tete, Lakhan Ghayade, Archit Datarkar, Kunal Jaiswal, Yogesh Dahake, International Research Journal of Modernization in Engineering Technology and Science, 2022

[9] Optical Character Recognition – A Review, Sushant Chandra, Saurav Sisodia, Preeti Gupta, IRJET, 2020.

[10] An Overview of OCR Research in Indian Scripts, B. Anuradha Srinivas1, Arun Agarwal2 & C. Raghavendra Rao2, International Journal of Computer Sciences and Engineering Systems, 2018.

[11] Optical Character Recognition Errors and Their Effects on Natural Language Processing, Daniel Loprest, International Journal on Document Analysis and Recognition, 2009

[12]Design of an OCR System and its Hardware Implementation, Gulfeshan Parween1 , Dr. Satadal Saha2, International Journal for Research in Applied Science & Engineering Technology (IJRASET), 2021.

[13]Handwritten and Printed Text Recognition Using Tesseract-OCR, Jay Amrutbhai Patel, International Journal of Creative Research Thoughts,2021.

[14] Improving Optical Character Recognition Techniques, International Journal of Engineering and Technology,2018.