**OCR RESEARCH**

**Submitted by**

**ROLL NO: EC39**

UTSAV .P. MUNGRA(19ECUOS049)

SEMESTER VIII

****

Department of Electronics & Communication

Engineering Faculty of Technology.

Dharmsinh Desai University, Nadiad.

**RESEARCH PAPERS BASED ON OCR**

1. Digitizing Historical Balance Sheet Data: A Practitioner’s Guide

**Abstract:**

To successfully digitize micro-data by augmenting OCR engines with pre and post processing. OCR application have higher error rates, however complementary OCR with other methods can increase success rate. This paper is about show casing this methods and why they are useful. We apply them against two large balance sheet and introduce quipucamayoc(python package)

JEL Classification: C81, C88, N80

Keywords: OCR, Data Extraction, Balance Sheets

**Introduction:**

Optical character recognition (OCR) is a powerful tool that allows researchers to unlock historical data.

we introduce [quipucamayoc,](https://github.com/sergiocorreia/quipucamayoc) an open-source Python package containing our methods in a unified framework.We argue that commercial products should be integrated and combined in a way that serves the researcher’s purpose. We suggest following a data extraction pipeline”that has the following steps:

* First, the original image files are preprocessed (de-warped, contrast adjustments, etc.).
* Second, the commercially available OCR and layout recognition techniques are applied.
* Third, the data are extracted and validated by leveraging relationships that must hold in the data such as accounting identities.
* Last,a crucial step is a human review in which the researcher herself validates some of the data creating a “ground truth.”

Such validated data allows researchers to then test and improve the accuracy of the digitization pipeline by serving as a benchmark against which the digitization output can be compared to construct accuracy metrics. In turn, these metrics allow for more advanced optimization of the parameters used throughout the digitization process.

Original Image

•

PDF

•

PNG

•

...

Data

“Ground Truth”

Image

Processing

OCR Engine

•

Engine 1

•

Engine 2

•

...

Layout Engine

•

Engine 1

•

Engine 2

•

...

Data Extrac-

tion and

Validation

Human

Review

Parameter Tuning

Fig 1. Data extraction Pipeline

**Future Scope:**

quipucamayoc is an actively maintained and developed package, and as such, is open to improvements from both the authors as well as future users of the package. We believe potential users will benefit from the existing tools and “batteries included” design; and for large-scale projects, we encourage them to extend the set of pre- and post-processing tools to improve the quality of their output.

**Summary:**

For large-scale datasets, we suggest that researchers’ time is more valuable when applied to the elements specific to their datasets, instead of focusing on directly managing the OCR tools and the scaffolding connecting the different parts of the OCR pipeline. By avoiding the more repetitive and cumbersome aspects of the OCR aspects, researchers are more able to devote resources to developing metrics for identifying errors in the data, either by constructing ground truths via human reviews or by exploiting characteristics of the data at hand. In the case of balance sheet records, we can exploit accounting identities to allow for straightforward error detection. In contrast, security price data may be less well suited, as they contain fewer constraints that we can validate against. Moreover, accuracy metrics allow researchers to easily test and tune the different components of the digitization pipeline, so instead of being relegated to the latter stages of the digitization, they can be used from the beginning to build a more accurate pipeline.

1. Automatic License Plate Recognition Using Optical Character

Recognition

**Abstract:**

This study aims to recognize vehicle number plates using an Android smartphone. Using a camera on a handphone to scan the plate image, and it is processed to find out the vehicle owner's information from the numbered plate being scanned. This study also uses Optical Character Recognition in processing images scanned by the camera where the steps taken are image acquisition, pre-processing, segmentation, normalization, feature extraction and recognition. The results show that the camera on an Android smartphone can be used to read and display information on number plate owners and also show that vehicle plate recognition using Optical Character Recognition based on the tests conducted shows the percentage of successful character recognition on motorized vehicle plates by 75% while the characters are equal to 97.36%.

**Introduction:**

Automatic plate reading is useful for everyday life. ALPR is a new technology in the transportation system automation world.

This technology utilizes image processing to identify vehicles from the license plate image, however variations on the license plate and the surrounding environment such as font size, font type, font color, number plate location, and differences in intensity due to lights or environment can cause problems during recognition license plate.

In the research conducted by Michael, the method used in image conversion is Optical Character Recognition. Tests carried out resulted in an accuracy of 78.57% if the condition of the number plate was in good condition and for the condition of the number plate in a bad condition it resulted in an accuracy of 57.41% . Referring to further research conducted by Manshur, in 2018, stated that the position of the camera will determine the accuracy of the final result . Budianto, in his research he compared the level of accuracy of various methods on automatic license plate recognition and various factors that influence it, and the results of his research stated that the Neural Network Backpropagation algorithm produced an accuracy rate of 96 percent and there were many factors that influenced this accuracy, one of which was the camera quality and also the distance between the camera and the vehicle.

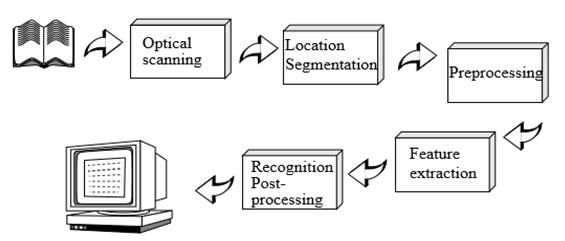


Fig 2. Block diagram of OCR

**Table 1**. Test Results in this table are displayed 10 test results from 100 tests.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No. | **Citra Flat** | **No. Plat** | **j. experiment** | **Introduction**  **Results** |
|  |  | E 3996 R | 3 | E 3996 R |
|  |  | E 5142 TW | 3 | 142 TU |
|  |  | E 4993 TI | 3 | E 4993 H |
|  |  | E 3115 SN | 3 | E 3115 SN |
|  |  | E 4564 SZ | 3 | E 4564 S |
|  |  | E 1245 RG | 3 | E 1245 RG |
|  |  | E 6295 QL | 3 | E 6295 OL |
|  |  | E 5361 TI | 3 | E 5361 TI |
|  |  | E 2956 RP | 3 | E 2956 RP |

E 2393 TT

3

E 2393 TT



**Results:**

After testing carried out on 100 number plate samples, it can be concluded with the percentage of success in the recognition process from 100 number plates, 75 number plates can be recognized, namely 75%. Character recognition accuracy with 722 total characters and 703 number of successful character recognition, amounting to 97.36%.

**Conclusion:**

In this study, the program was able to recognize license plates using Optical Character Recognition (OCR) techniques. The application of the OCR method on the android operating system results in good accuracy in number plate readings, where from the test results are obtained 75% accuracy, where the accuracy for the reading of each character is 97.36 %. The system is able to display vehicle owner data that is successfully recognise correctly. Recognition failures are due to the resemblance of characters to each other, such as B and 8, 2 and Z, 1and I, D and 0, A and 4. Imperfections in character segmentation can cause characters to be truncated that can change the perception of character recognition.

1. HENet: Forcing a Network to Think More for Font Recognition

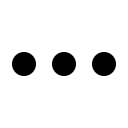
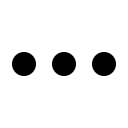
**Abstract:**

This paper proposes a novel font recognizer with a pluggable module solving the font recognition task. The pluggable module hides the most discriminative accessible features and forces the network to consider other complicated features to solve the hard examples of similar fonts, called HE Block. Compared with the available public font recognition systems, our proposed method does not require any interactions at the inference stage. Extensive experiments demonstrate that HENet achieves encouraging performance, including on character-level dataset Explor all and word-level dataset AdobeVFR.

**Introduction:**

The proposed method provides a solution without tedious pre-processing processes and interactions for font images, which still achieves a satisfactory performance.

With the rapid development of deep learning algorithms, great success has been achieved on many computer vision tasks(e.g., image classification, object detection, and optical character recognition). Font recognition can be regarded as a special task of image classification, which takes a raw image as input, and subsequently learns its class-specific feature representation through a CNN-based network. Finally, the result can be predicted with the class-specific featurerepresentation through the classifier layer. The current research on the use of CNNs for deep learning in computer vision has got many encouraging results, including AlexNet [5] , ResNet [3] , VGGNet [8] , GoogleNet [10] and others. However, directly using these networks on font recognition definitely can not get a satisfactory performance. Some task-specific methods are needed for the task of font recognition.



**"CascadeScriptLTStd"**

**Class:**

Font System

Font System

User Interactions

**Public Font System**

**"CascadeScriptLTStd"**

**Class:**



HENet

**Our Proposed Method**

Fig3. HENet block diagram

**Results:**

Evaluated our algorithm on the AdobeVFR dataset. The experiment demonstrates that HE Block improves top-1 performance on synthetic validation set as well as the real-world test set by 0.61% and 1.14%. Although The real-world test set is extraordinarily challenging and the training set contains no real-world data, our method still brings considerable improvement on word-level.

**CONCLUSION:**

In our paper, we proposed a method to get better performance on font recognition whose categories are very similar. Our proposed HE Block suppresses the prominently accessible feature so that the whole network can think more and learn more complicated stroke details. What’s more, the utilization of HE Block is pluggable during the training phase and does not bring any extra computation cost during inference time. We conduct related experiments with different backbone networks, proving that our method works on different backbones and surpasses the state-of-the-art font recognition models. In our future work, we will make a large-scale real-world font dataset containing both Chinese and Latin characters as a public benchmark for font recognition. Then, we will further explore the popular transformer-based model to obtain global and local strokes information.

1. Smart Reader Glass for Blind and Visually Impaired People

**Abstract:**

Visually impaired people fail to read the text with existing technology. The proposed project targeted to design a spectacle with a camera by which the blind visually impaired people can read whatever they want to read based on contemporary OCR technique and text-to-speech (TTS) engines. This proposed smart reader will read any kind of documents like books, magazines and mobiles. People can access this novel technology with blindness and limited vision. Theearlier version of the proposed project was developed successfully with mobile readerwhich had certain drawbacks such as high cost due to the need of android mobile, not

user friendly and improper focusing. To overcome these disadvantages, a spectacle type

reader with camera is proposed in this project, which will be cost effective and more efficient.

Keywords: Micro camera, Python, Raspberry pi, Smart reader, Text-to-speech

**Introduction:**

This Project is aimed to design assistive smart glasses in wearable design format for the blind and visually impaired people. This smart spectacle with micro camera setup proposed in this project is designed to support reading printable version of any books, documents, mobile texts by converting text to audio,which can be heard by microphones or speakers. This portable and economical smartspectacle is programmed with raspberry pi module and the image processing techniquehelps in recognizing and extracting the text from the image [4, 5]. Finally, the extracted text is converted into speech and can be heard by blind and visually impaired people. Thefinal hardware model is tested with two test samples, the first one is with book page and another one is mobile document. The designed smart glass converted both test samples into a right audio format. This project is very affordable to all category of people and will be more useful.

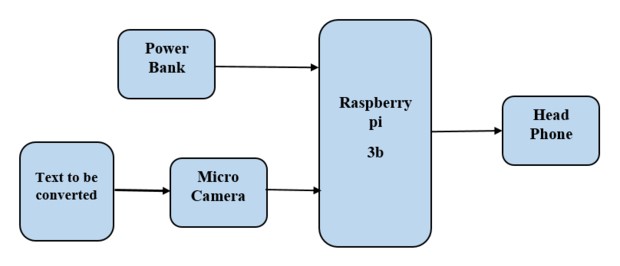


Fig 4. Block diagram of smart reader glass

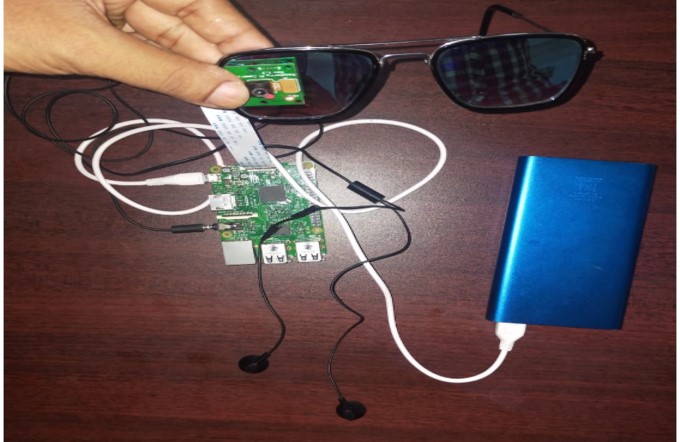
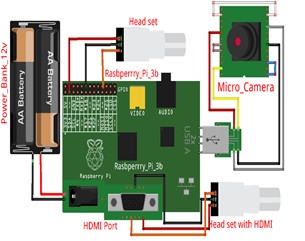


Fig 5. Circuit diagram of smart reader glass

**Results:**

The audio output can be heard using normal microphones used for mobile phone. The rigid assembling of hardware components ensure the lifetime of the proposed smart glass. Really, this smart reading spectacles will be an optimal choice for visually impaired peoples.

**Summary and Conclusion:**

The proposed project - Smart reader glass for blind and visually impaired people mainly aimed to assist the blind and visually impaired people to read the documents in the form of text book, mobile screen or computer screen. Even though some of the glasses are available in the market they are too costly and not affordable to common people. Our proposed project is designed with simple components which can be placed in shirt pocket. Hence this product is user friendly and will be available at very low price and compact size. The audio output can be heard using normal microphones used for mobile phone. The rigid assembling of hardware components ensure the lifetime of the proposed smart glass. Really, this smart reading spectacles will be an optimal choice for visually impaired peoples.

1. Sophisticated and modernized library running system with OCR algorithm using IoT

**Abstract:**

This research on library control system operates on the basis of IoT and OCR algorithm rules and its training.A closed-circuit television (CCTV) watched mechanism is created to control the book issuing and returning phenomenon via tag studying system in the library. In this proposed word text file converted into an audio file and it is being played and the contents of the book can be heard via the headset, which is useful for blind people. Now a days OCR widely focused in machine processes such as machine transformation,TTS extraction and text data mining. Using OCR to scan the damaged book in the library converted into pdf format the book gets new life and sharing the contents to multiple readers. In this paper aims to implement IoT based library management system to maintaining books in digital format.

**Introduction:**

This system is more ineffective, and they spent more time for searching books. Sometimes the students are unaware whether the books are existing or not. To overcome this issue to implement the proposed system to provides the automatic solution. In the library every book is inserted with RFID tag and RFID readers [1]-[6].

Manually managing library systems are vulnerable human error and slow process it difficult to store massive amounts of data or records efficiently, the library staff spends maximum of time on mechanical, clerical tasks rather than liaising with library visitors. Issuing a book is a slow process also books are more theft prone.

Combining the RFID with text recognition software to offers effective methods to maintaining the library system. The proposed system is could be helpful to overcoming these challenges.

Optical character recognition (OCR) is a mechanism for converting the scanned documents or printed texts, handwritten data into computer recognizable universal character encoding (UNICODE) format or machine-readable text over an optical mechanism. It captures the text and reading capabilities of a human eye. It widely used in data entry for digitizing the documents like passport, bills, bank reports, mail, static data, and printed data.

By using this technique, the documents are digitally modified, stored compactly, searching process is quick and fast. OCR is used in machine translation processes such as text-to-speech recognition, key data, text mining and cognitive computing. In recent days most of the academics focus their research in artificial intelligence, pattern recognition, and computer vision.

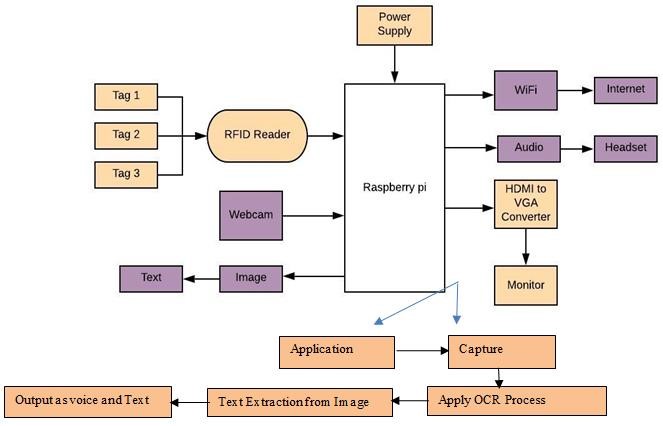


Fig 6. Block diagram of proposed system

# Conclusion:

Here updated version of library management system has been analyzed and compared with the regulatory of the existing one, at the whole, outcome of this experiment shows that the proposed one is better compared to the earlier technology with more advancement in the field of library management that pay the way to develop the advanced library management system which inculcates the OCR implementation very wisely. The book issuing and returning system created is highly efficient and secured added with two important concept like conversion of pdf file into highly visible text file and it also has been converter to audio file where it has been widely used for blind people in the society which improves there eager in the library usage the above technique has been verified using coding and outputs.

## **OCR & Handwriting Datasets for Machine Learning**

[NIST Database](https://catalog.data.gov/dataset/nist-handprinted-forms-and-characters-nist-special-database-19): The US National Institute of Science publishes handwriting from 3600 writers, including more than 800,000 character images.

[MNIST Database](http://yann.lecun.com/exdb/mnist/): A subset of the original NIST data, has a training set of 60,000 examples of handwritten digits.

[Devangri Characters](http://www.iapr-tc11.org/mediawiki/index.php?title=Devanagari_Character_Dataset): A dataset of handwritten Devangari characters, composed of 1800 samples from 36 character classes obtained by 25 native writers.

[Mathematics Expressions](http://www.iapr-tc11.org/mediawiki/index.php?title=CROHME:_Competition_on_Recognition_of_Online_Handwritten_Mathematical_Expressions): More than 10,000 expressions, including more than 101 mathematical symbols.

[Chinese Characters](http://www.iapr-tc11.org/mediawiki/index.php?title=Harbin_Institute_of_Technology_Opening_Recognition_Corpus_for_Chinese_Characters_(HIT-OR3C)): A dataset of handwritten Chinese characters containing 909,818 images that corresponds to about 10 news articles.

[Arabic Printed Text](http://diuf.unifr.ch/diva/APTI/): Contains a lexicon of 113,284 words, and uses 10 Arabic fonts.

[Document database](http://www.iapr-tc11.org/mediawiki/index.php?title=IAM_Online_Document_Database_(IAMonDo-database)): Contains 941 online handwritten documents by 189 writers, and covers lists, tables, formulas, diagrams and drawings.

[Iam On-line Handwriting](http://www.fki.inf.unibe.ch/databases/iam-on-line-handwriting-database): Contains forms of handwritten English text acquired on a whiteboard, and includes more than 1700 entries.

[Street View Text](http://www.iapr-tc11.org/mediawiki/index.php?title=The_Street_View_Text_Dataset): The Street View Text dataset was harvested from Google Street View, and mostly deals with outdoor street level signs and boards.

[Street View House Numbers](http://www.iapr-tc11.org/mediawiki/index.php?title=The_Street_View_House_Numbers_(SVHN)_Dataset): Contains 73257 digits of house street numbers, taken from Google Street View.

[Natural Environment OCR](http://www.iapr-tc11.org/mediawiki/index.php?title=NEOCR:_Natural_Environment_OCR_Dataset): A dataset that contains 659 real world images with 5238 annotations of text.

[Scene Text](http://www.iapr-tc11.org/mediawiki/index.php?title=KAIST_Scene_Text_Database): Contains 3000 images captured in different environments, including outdoors and indoors scenes under different lighting conditions (clear day, night, strong artificial lights, etc).

[Text Detection](http://www.iapr-tc11.org/mediawiki/index.php?title=MSRA_Text_Detection_500_Database_(MSRA-TD500)): Contains 500 natural images, which are taken using a pocket camera. The indoor images are mainly signs, doorplates and caution plates while the outdoor images are mostly guide boards and billboards.

[Stanford OCR](http://ai.stanford.edu/~btaskar/ocr/): Contains handwritten words dataset collected by MIT Spoken Language Systems Group, published by Stanford.