# *A Novel Approach for Data Retrieval from Identity Cards Using Deep Learning*

Tania Rose Jobi
PG Scholar
Department Of Computer Application
Amal Jyothi College of Engineering
Kanjirappally, India
taniarosejobi@mca.ajce.in

Lisha Varghese
Asst. Professor
Department Of Computer Applications
Amal Jyothi College of Engineering
Kanjirappally, India
lishavarghese@amaljyothi.ac.in

***Abstract*—In today's world and age, there has been an immense demand for digital representation that has fired up a colossal growth in technology and communication. Authentication and authorization play a vital role in digitization. Identifications that are mainly used for verifications are identity cards, licenses, passports and other identification documents. Data from these identity cards are extracted by organizations for valid purposes. This paper aims to propose some methods using various deep-learning tools to gather data from identity cards. The trial results of this research prove the usage of steps such as the Efficient and Accurate Scene Text (EAST) which increases the efficiency.**

***Keywords: Optical Character Recognition (OCR), PyTorch, Easyocr, OpenCV, Language Recognition, Pre-processing***

## I. INTRODUCTION

Nowadays, extracting and thereby understanding textual information manifested in identity cards or any other kinds of documents used for personal identification purposes have become progressively important and popular. Most valid identity cards convey genuine and authentically essential information about the owner of the card. These identity cards are present mostly in terms of color, text layout, patterns followed, and level of importance. Nevertheless, most of these identity cards that are valid, follow the predefined size of the ratio of around 1.58.

In terms of identification, passports are one of the common yet legally and internationally tackled identity proof of identification around the world. In a passport, at the bottom of the main page, an area mainly a barcode area, is found with all the essential fields of information and this area is called the Machine-Readable Zone (MRZ). Machine Readable Zone can be basically found in passports and not in other identity cards such as a driving license or an identity card of a specific organization. Besides, the lamination, emblems, and texture add noise which can make machine reading less accurate. This can affect the data extraction process.

To make the extraction process smooth, we apply some preprocessing steps to remove the noises present such as texture, patterns and symbols. Eventually, this can improve the speed and accuracy of the data being extracted. The data extraction revolves around some preprocessing steps such as increasing the contrast or brightness of the scanned image which can be less likely to be viewable, if not scanned in the presence of proper lighting. Another preprocessing step is noise removal to remove any unwanted images or symbols or emblems in government-related identification documents. The next step is skew alignment, which needs to be done to correct the alignment of a rotated or unaligned ID.

With these preprocessing steps, we do the first phase of data extraction from a valid identity card which includes auto-cropping, locating alignment of the ID card, photo removal and text segmentation.

The second phase aims to represent an exemplary method for data extracted for information or data retrieval. Furthermore, applying some other techniques such as age recognition, and male/female recognition after successful data extraction can prove to be useful for information verification. This paper allows reader to follow up and conclude each module separately in a proper logical manner. Trial results of each module are present towards the end of each module to evaluate the results.

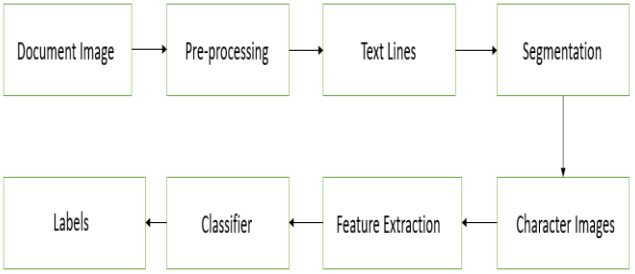


Fig 1. The flowchart of information retrieval from an identity card

## II. LITERATURE SURVEY

Niddal Imam et. al [2] have discussed how Optical Character Recognition systems help in detecting malicious content by classifying using machine learning algorithms.

Ahmed Akkaddo et. al [3] have mentioned the advantages as well as the usage of CN in OCR in the handwritten and printed text.

Daniel Rotman et. al [5] have researched how to improve OCR quality greatly using various preprocessing methods with a masking system which upgrades the OCR quality

Liang Qiao et. al [6] proposed OCR with an open-source tool named DavarOCR that can implement many algorithms.

Moonbin Yim et. al [7] have together proposed a synthetic document image generator through ample analysis for a real-world application.

*DOI: 10.5281/zenodo.7962194*
*ISBN: 978-93-5906-046-0@2023, Dept. of Computer Applications, Amal Jyothi College of Engineering Kanjirappally, Kottayam*

Kimmo Kettunen et. al [8] together has perceived refined OCR quality in use in historical newspaper articles.

Ankit Tiwari [11] has implemented on the OC process done using Tesseract OCR engine.

## III. EXISITING AND PROPOSED SYSTEM

Optical Character Recognition (OCR) is an area that has been a topic of study over the past decades. OCR is a technique of data extraction of a different character from a record picture and additionally provides full alphanumeric recognition of printed or handwritten characters, text numerical, letters, and symbols into a computer processable layout including ASCII, Unicode and so forth. It is deeply used in our day-to-day activities as it is less time-consuming and has a well-determined speed level. The steps in OCR include image acquisition, pre-processing, segmentation, feature extraction, and post-processing. Many methods of OCR are present in data extraction using various tools such as Nanonets, DocParser, Adobe Acrobat and many more.

All modules for the journey of Data Extraction in this paper are as follows:

- Installing wanted OCR and Deep Leaning tools used for data extraction.
- Adjustment contrast or brightness
- Removal of unwanted noise from scanned identity cards
- Gray-Scaling
- Text / Non-Text Segmentation
- Language Recognition
- Box Plotting the extracted data on the ID card

## IV. IMPLEMENTATION

In this system, we are trying to extract textual data from valid and logical identity cards. For this process, the following are the steps required to be done.

### A. Installing tools

The various Optical Character Recognition and Deep Learning tools used in this paper are EasyOCR, PyTorch, and Matplotlib. EasyOCR is basically a python package that allows the scanned image to be converted to text by reading them. PyTorch is a machine-learning as well as a deep-learning framework based on the Torch library. Matplotlib is a plotting library in the Python programming language and its numerical mathematics extension is NumPy. OpenCV is another library of programming functions that supports Machine Learning.

```
In [*]: #install pytorch
        !pip3 install torch torchvision torchaudio

In [*]: #install easyOCR
        !pip install easyocr

In [*]: #install matplotlib
        !pip install matplotlib

In [91]: #importing libraries
         import easyocr
         import cv2
         from matplotlib import pyplot as plt
         import numpy as np
```

Fig 2: Installing Tools and Importing Libraries

### B. Adjustment contrast or brightness

The quality of the source image is a crucial part of data extraction. The quality of the image can in fact determine the accuracy of the extraction to a major level. But if the source image is not clear, the results produced by OCR can most likely include errors. Using a better-quality image makes it easier to differentiate characters from each other, thereby producing high accuracy. We must thereby increase the contrast, density, and brightness before carrying out the process of data extraction.

```
In [110]: #increasing contrast and brightness
          alpha = 2.0 # Contrast control (1.0-3.0)
          beta = 0 # Brightness control (0-100)
          adjusted = cv2.convertScaleAbs(img, alpha=alpha, beta=beta)
```

Fig 3: Increasing contrast or brightness

### C. Noise Removal

Noise Reduction or Removal is an extensive approach to intensify and boost text detection results. Auto-cropping is one of the methods that partially removes unwanted noise thereby cleaning the data or text area. The surface of identity cards can be unclear or dirty due to the prolonged physical touching of the cards by the user. This can accumulate dirt and make the textual area unclear. The background part of the ID card can also be polluted by the scanner or camera lens due to dust or dirt. This causes many conflicts with the following processes described in the next modules.

The process of noise removal done in this module primarily starts by converting the cropped image to grayscale. This strategy is then followed by normalizing for enhancement motive and converting to grayscale again. Eventually, the intensified grayscale is then converted to binary to create a disguise by blurring and finally entwining to create a clean background.

```
In [111]: #noise removal
          im_nr = cv2.fastNlMeansDenoisingColored(img, None, 10, 10, 7, 15)
```

Fig 4: Noise Removal

### D. Text/Non-Text Segmentation

This is one of the most vital modules in data extraction as it can affect the final results in text / non-text segmentation,

*DOI: 10.5281/zenodo.7962194*
*ISBN: 978-93-5906-046-0@2023, Dept. of Computer Applications, Amal Jyothi College of Engineering Kanjirappally, Kottayam*

that is, finding the area containing the textual readable information or the Machine-Readable Zone (MRZ). There are various approaches that are performed to acquire this purpose, such as Maximally Stable External Region (MSER), Stroke Width Transform (SWT), and the most recent one is the Efficient and Accurate Scene Text Detector (EAST) Detector, which is a detector based on deep learning methods.

Here it initially finds high-contrast edges. Then it traverses the image pixel's edge, in the normal direction to obtain another normal edge. This method helps in identifying strokes, which is an element of finite width with two roughly parallel sides. These strokes then play roles as a text area.


Fig 5: Sample ID card for data extraction

```
In [10]: reader = easyocr.Reader(['en'], gpu=False)
         result = reader.readtext(IMAGE_PATH)
         result

         Using CPU. Note: This module is much faster with a GPU.

Out[10]: [([[8, 160], [170, 160], [170, 186], [8, 186]],
           'TANIA ROSE Jobi',
           0.43340534550240095),
          ([[8, 188], [134, 188], [134, 216], [8, 216]],
           '2018-2023',
           0.9998505457030473),
          ([[8, 212], [60, 212], [60, 238], [8, 238]], 'MCA', 0.997291794829096),
          ([[72, 212], [116, 212], [116, 236], [72, 236]], 'INT', 0.932505190372467),
          ([[10, 236], [80, 236], [80, 262], [10, 262]], '11038', 0.9987248405378192)]
```

Fig 6: Data extracted from the ID card

A. Language Recognition

EasyOCR supports almost 70+ languages which encompass languages such as Hindi, Chinese, and Russian. Here, in this research paper, we are extracting data which is written in the language Hindi using Optical Character Recognition.

For this purpose, we have taken an Aadhar Card, a unique identity proof used by the citizens of India. An Aadhar Card has transcripts written in the language, Hindi, which are extracted using EasyOCR.

```
In [12]: reader = easyocr.Reader(['en','hi'], gpu=False)
         result = reader.readtext(IMAGE_PATH)
         result

         Using CPU. Note: This module is much faster with a GPU.
         Downloading recognition model, please wait. This may take several minutes depen
         Progress: |████████████████████████████████████████████████| 100.0% Complete

Out[12]: [([[233, 35], [358, 35], [358, 61], [233, 61]],
           'भारत सरकार',
           0.9641173549342741),
          ([[194, 63], [398, 63], [398, 88], [194, 88]],
           'Government ofIndia',
           0.48849817032224961),
          ([[575, 69], [619, 69], [619, 83], [575, 83]], '3भा', 0.2599998712539673),
          ([[232, 146], [386, 146], [386, 170], [232, 170]],
           'Tania Rose Jobi',
           0.8687603521019104),
```

Fig 7: Hindi transcript extraction

B. Box Plotting

A bounding box is a sort of imaginary rectangle or any shape that is used to outline the extracted data or object as per machine learning or user requirement. They are the main end results of an object detection model. These bounding boxes specify the position and the confidence that tells us the chance of the textual information that is present in that position.

The bounding box is one of the most popular image annotation techniques in deep learning. this method reduces costs and increases the efficiency of annotation. Firstly, the python package, matplotlib must be imported in order to plot in python. To specify the position various parameters are used such as class, width, height, and confidence

```
In [11]: top_left = tuple(result[0][0][0])
         bottom_right = tuple(result[0][0][2])
         text = result[0][1]
         font = cv2.FONT_HERSHEY_SIMPLEX

In [12]: img = cv2.imread(IMAGE_PATH)
         spacer = 100
         for detection in result:
             top_left = tuple(detection[0][0])
             bottom_right = tuple(detection[0][2])
             text = detection[1]
             img = cv2.rectangle(img,top_left,bottom_right,(0,255,0),3)
             img = cv2.putText(img,text,(20,spacer), font, 0.5,(0,255,0),2,cv2.LINE_AA)
             spacer+=15

In [13]: plt.figure(figsize=(10,10))
         plt.imshow(img)
         plt.show()
```


Fig 8: Bounding boxes using box plotting

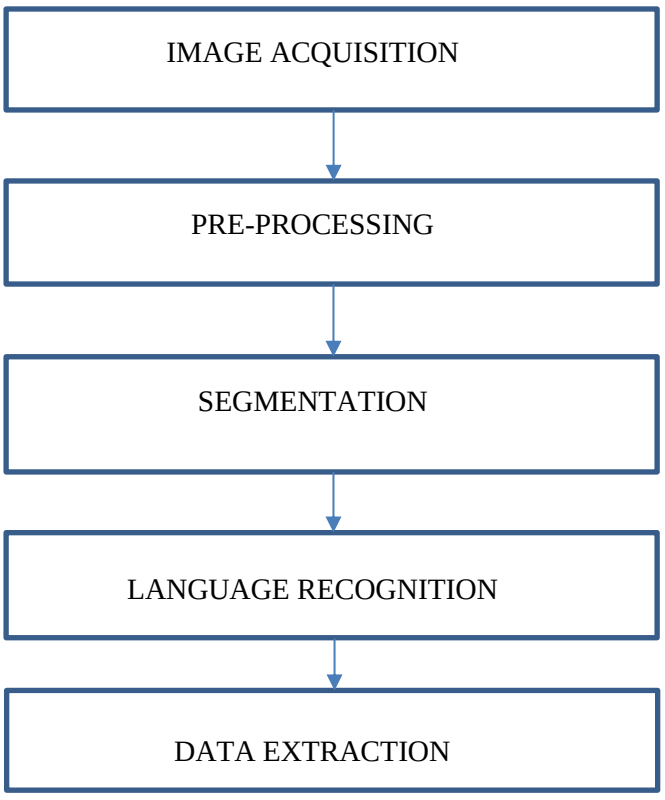


Fig 9: Flowchart of the methods

*DOI: 10.5281/zenodo.7962194*
*ISBN: 978-93-5906-046-0@2023, Dept. of Computer Applications, Amal Jyothi College of Engineering Kanjirappally, Kottayam*

## V. CONCLUSION

This research paper aims to enhance and boost the timing and accuracy of information retrieval from identity cards. It mainly focuses to diminish the waiting time that is required by a customer upon submitting his/her identity card for checking and verification.

Based on this research, various steps from capturing the ID image to Optical Character Recognition include many steps. Many pre-processing steps such as noise removal, gray-scaling, text segmentation, and increasing contrast or brightness are done.

This study attempts to find the most apt solution for text segmentation. EAST is a robust, sturdy and accurate text segmentation method used for this purpose although it has its own drawbacks such as sensitivity towards rotation and improper margins.

Finally, this paper concludes that the methods performed in this research are real-time and offline.

## VI. FUTURE WORKS

Further work and future development following this study can be done by assessing verification methods such as calculating age from date of birth, gender, signature recognition, and nationality.

Another future development bought into this research can be the find the accuracy of the data extracted from the identity cards.

## VII. REFERENCES

[1] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, Jiajun Liang:- "EAST: An Efficient and Accurate Scene Text Detector," Megvii Technology Inc., Beijing, China, 10 July 2017

[2] Niddal Imam, Vassilios Vassailakis, Dimitris Kolovos- "OCR post-correction for detecting adversarial text images", May 2022

[3] Ahmed Alkaddo, Dujan Albaqal- "Implementation of OCR using Convolutional Neural Network*CNN): A Survey", September 2022

[4] K. Chinnasarn, Y. Rangsanseri, and P. Thitimajshima, "Removing salt and-pepper noise in text/graphics images," in IEEE. APCCAS,1998.

[5] Daniel Rotman, Ophir Azulai, Inbar Shapira, Yevgeny Burshtein- "Detection Masking for Improved OCR on Noisy Documents", May 2022

[6] Liang Qiao, Hui Jiang. Ying Chen, Can Li- "DavrOCR: A Toolbox for OCR and Multi-Modal Document Understanding", July 2022

[7] Geewook Kim, Teakgyu Hong, Moonbin Yim, Jeongyeon Nam- "OCR-free Document Understanding Transformer", November 2022

[8] Kimmp Kettunen, Heikki Keskustalo, Sanna Kumpulanen, Tuula Paakkonen- "OCR quality affects perceived usefulness of historical newspaper clippings- a user study", February 2022

[9] J. Wehr, "Card size specification, when does card size matter," Education Library, 2002.

[10] J. E. Bollman, R. L. Rao, D. L. Venable, and R. Eschbach, "Automatic image cropping." Google Patents, 1999.

[11] Anuarg Tiwari- "Data Extraction from images through OCR", August 2022

*DOI: 10.5281/zenodo.7962194*
*ISBN: 978-93-5906-046-0@2023, Dept. of Computer Applications, Amal Jyothi College of Engineering Kanjirappally, Kottayam*