Document Data Extraction & Verification Using Process Automation Techniques

Rupali Umbare
Information & Technology Department
JSPM's Rajarshi Shahu College of Engineering
Pune,India
umbarerupali1@gmail.com

Nishad Raisinghani
Information & Technology Department
JSPM's Rajarshi Shahu College of Engineering
Pune,India
nishd268@gmail.com

Ninad Bhole
Information & Technology Department
JSPM's Rajarshi Shahu College of Engineering
Pune,India
ninad.bhole227@gmail.com

Sarvesh Kulkarni
Information & Technology Department
JSPM's Rajarshi Shahu College of Engineering
Pune,India
sarvesh.kulkarni1606@gmail.com

Sumit Bhattacharya
Information & Technology Department
JSPM's Rajarshi Shahu College of Engineering
Pune,India
sumit.180200@gmail.com

***Abstract*—We've all heard the phrase "time is money," and since then, we've been developing tools and technologies to help us save time and increase our productivity. Humans currently assist in data extraction from different non-textual and unstructured documents such as photographs; however, our research intends to automate this process by utilizing computer vision and a proprietary algorithm to properly extract data and validate it with official sources.**

Keywords—Automation, Extraction, Data, Digitization, Verification

I. Introduction

Documents issued by various institutions, such as our government, are among a citizen's most precious possessions. At various phases of life, these documents aid in proving one's identity, uniqueness, and eligibility. A driver's license, for example, serves as confirmation of identification and eligibility to operate motor vehicles safely on public roads. Although these papers have several advantages, the most significant downside is that they are difficult to digitize and utilize for other purposes. Automation procedures look into and provide a number of user-friendly solutions for this problem.

II. Problem Identification

As described above the problems that are faced with issued documents are properly listed and explained below:

- Lack of ease of digitization.
- Repetitive in nature - the process of extracting useful data from such documents.
- Lack of tools to aid such conversions in the market.
- Time taken to process such documents is very long and may extend upto weeks.
- Requirement of extra man-power to carry out digitization.
- Large amounts of money wasted in hiring data entry specialists.
- Tedious task.

III. Objective

Our project's main objective is to build an approach which is scalable to various other applications by mainly focusing on the process of opening a bank account. We aim to build a user-friendly application which automates this process along with reducing the time and effort required to do so. Our objectives are laid down below in a list:

- A user-friendly UI for consumers to easily create accounts and upload documents.
- A 3-step process optimized for instant bank account creation and verification.
- Automatic data extraction using various methods described below and verification of the data.

IV. Research & Market Analysis

A. Feasibility Analysis

The idea of streamlining the process of opening a bank account is a great one but is it possible to build a system to aid this problem? Was one of the questions that aroused at the beginning. To find answers to this problem we set out to gather data from our friends and family who have applied or already have a bank account. After tallying the results we found out that people were willing to use such a system where there is minimal human interaction and faster results, especially the newer generations who are accustomed to web based portals feel comfortable using one for a bank as well. Apart from this, managers at Xoriant Solutions Pvt. Ltd. showed great interest in the project and would be willing to show our project as a proof of concept to their clients.

B. Existing Systems

Companies like UIpath and Automation Anywhere provide solutions to aid this problem using their proprietary coding solutions which require special software engineers to program and customize the tool for the application which is equivalent to building a system from scratch. No solid solution for all exists in the market at the moment.

Some drawbacks in the existing systems are as follows:

- Automation Anywhere has poor usability and is unable to use OCR properly..
- To set up and maintain efficiently, both technologies require considerable coding knowledge.
- Both technologies concentrate on automating numerous activities rather than digging down and polishing a single solution..
- User interfaces are difficult to understand and use, with a high learning curve.

C. *Technologies & Tools*

After researching techniques and methodologies required to accomplish this project it was found that Python was the programming language of choice as it provided extensive support for image processing, machine learning and data analysis. Apart from that tools like government APIs to verify such documents were also one of the most important parts required.

D. *Documents Involved*

With the advent of Digital India movement in India, the number of documents required to create a bank account have shrunk down to the two most important ones:

- Permanent Account Number (PAN Card).
- Unique Identity (Aadhar Card).

V. Proposed System

Since the start of the development process of this system we have been able to build a basic user interface for easy use and PAN card data extractor. We have listed down the problems we faced during the development of current iteration of the project along with the solutions that we came up with:

A. *Data Extraction*

We initially planned on using Darknet and YOLO machine learning frameworks to extract data from PAN card images:

The Problem : The sheer lack of images of PAN cards one can find on the internet due to strict government policy over data protection. Overfitting of ML models during training to similar kinds of images generated using image augmentation.Low accuracy due to no drastic changes in training images as all PAN cards have similar structure

The Solution : As all PAN cards have a similar structure we can extract the data from the image as whole using Tesseract OCR and extract data with help of line numbers. For eg: A pan card has Holder's name at the top, then fathers name,date of birth and permanent account number so we can correspond the order with line numbers of the extracted text. The fail-safe to this is the data will be verified by the user before verification from the government is done.

B. *Data Verification*

The problem: The problem that we faced with data verification was that the PAN API is only provided for use to institutions which are recognized as registered financial entities by the government for use.

The Solution: A basic form of verification can be done using the public page for citizens to verify PANs when a contract is made. We built a web scraper to insert the data extracted from the image in the form provided and verify it for us, apart from that we are looking to buy access from third party vendors for better verification.

VI. Implementation

During the implementation of the above mentioned solutions we tested and compared the accuracy of both the solutions in order to choose the best possible method. We chose OCR to recognize data and tested two approaches to extract the meaningful data. The approaches were as follows:

- Machine Learning based Data Extraction.
- Custom Line Matching Algorithm based Data Extraction.

A. *OCR Based Text Recognition*

Optical Character Recognition is the identification of printed characters using photoelectric devices and computer software. We used OCR to recognize the text from images of the Aadhar Card and PAN Card.Instead of writing an OCR algorithm from scratch we used the tested and tried Tesseract OCR library to achieve text recognition.

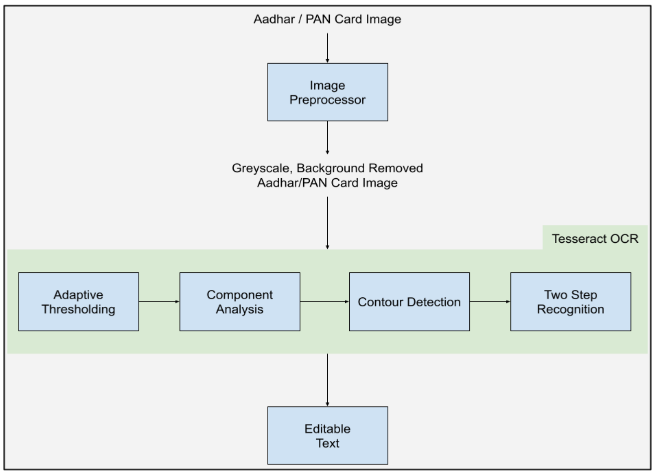


Fig. 1. *Text Recognition Process Flow*

The flow mentioned in Fig. 1. shows how an image is processed and fed to the Tesseract OCR. The end result is text which can be edited ie. it can be manipulated and transformed according to our project's requirements.
For this project's usage the text recognition step occurs after data extraction for the ML based approach and before for the Line Matching algorithm based approach.

B. *ML based data extraction*

Machine Learning based data extraction approach suggests that in a document with help of the bounding boxes, we can identify the positions of important text inside of that document.During training one must provide a large dataset with labels and coordinates of bounding boxes. This helps the ML algorithm to learn and replicate the same on different documents.

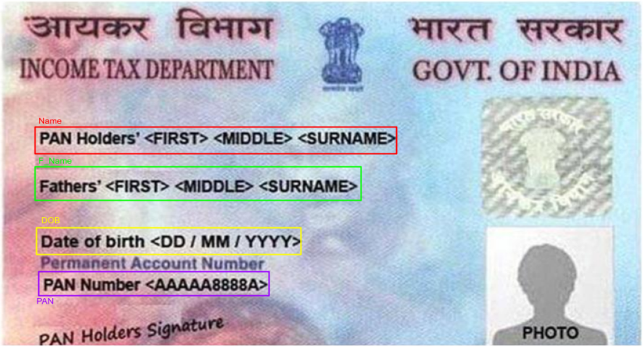


Fig. 2. *Bounding Boxes Around Meaningful Text on PAN Card*

For this approach YOLOv3 was the algorithm of choice, you only look once (YOLO) is a state-of-the-art, real-time object detection system. R-CNN algorithms generally apply a model to various regions of an image, the region with the most score is marked as the detected object. YOLO on the other hand applies the same model to the whole image and attempts to predict various bounding boxes. The bounding boxes with the highest probability are selected and marked as an object [4]. This method increases the speed with similar accuracy to other slower algorithms.These qualities of the algorithm are required in a low latency web application and that is why YOLOv3 was chosen. Steps in extracting the required text:

- Mark and label dataset with bounding boxes coordinates and class names.
- Load pre-trained weights of the coco-dataset to boost performance.
- Load training and testing data and start training.
- Save the trained model.
- Load a new image.
- Using the weights of the new model, predict the positions of the text.
- Using openCV and predicted positions extract image segments from the image.
- Use OCR to recognize the text.Line Matching based data extraction.

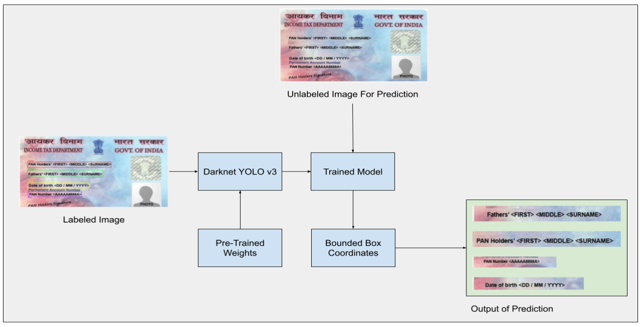

Unlabeled Image For Prediction
Labeled Image
Darknet YOLO v3
Trained Model
Pre-Trained Weights
Bounded Box Coordinates
Fathers' <FIRST> <MIDDLE> <SURNAME>
PAN Holders' <FIRST> <MIDDLE> <SURNAME>
PAN Number <AAAAA8888A>
Date of birth <DD / MM / YYYY>
Output of Prediction

Fig. 3. *Data Extraction Using Yolo V3*

C. Line Matching based data extraction

For the second approach, we observed that the text positions in the above mentioned document are static, i.e. each and every PAN/Aadhar card has the meaningful text printed on approximately the same position.Developing on this observation we wrote a custom text cleaning and text extraction algorithm. This algorithm suggested that if we are able to identify the line number of any meaningful text like date of birth, we can extract other text with help of their line number orders by subtracting or adding to the line number of the already identified text.

This approach works flawlessly for all documents which have a standard format, most government documents have the same formats Steps in extracting required text are:

- Use OCR to get all the text from the image.
- Clean the editable text by removing unnecessary data from the text like empty lines, sentences like Income Tax Department etc.
- Split the cleaned text by a new line (‘\n’) and store it in an array.
- Loop through the array to find heading like Date Of Birth, store the index of the line in a variable (let variable be N).
- With help of this index positions can be estimated as follows Name : N - 4 , Father’s Name: N - 2 , Date of Birth : N + 1, PAN : N + 3

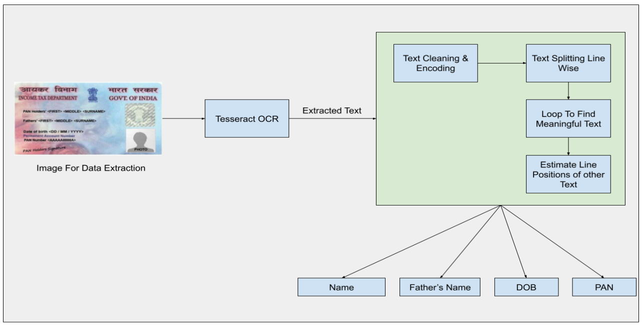
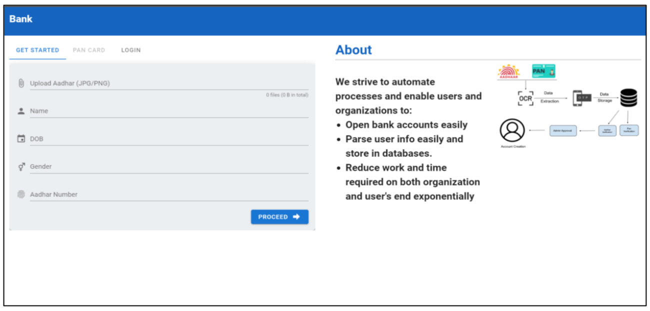


Fig. 4. *Data Extraction Using Line Matching*

D. Data Verification

Data verification is relatively a simpler process, as the government of India provides API’s for the same. These API’s are only available to financial entities i.e. banks, payment gateways, insurance companies. The data extracted is private and sensitive, so to prevent data leaks the government asks for 2.5 million rupees as a security deposit from these entities.In order to demonstrate that the data extracted is proper, we implemented a custom verification API using selenium and open portals that allow civilians to verify the Aadhar/ PAN Card. The selenium webdriver scrapes the website and enters the extracted data on the portal and waits for the result, if the portal shows verified text we consider our documents verified.This process is added to a CRON job so that it runs every 10 minutes of the day, the reason being that sometimes the captcha is not properly solved.

E. System GUI & Results


Bank

GET STARTED PAN CARD LOGIN

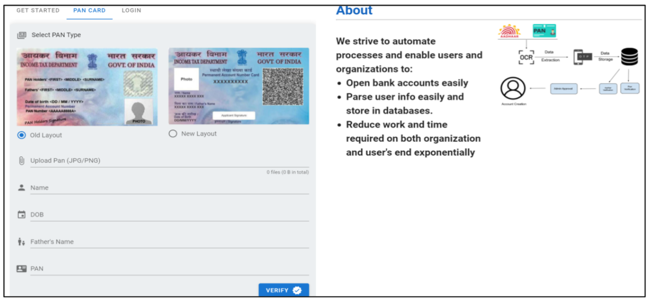
Upload Aadhar (JPG/PNG)
0 files (0 B in total)
Name
DOB
Gender
Aadhar Number
PROCEED

About

We strive to automate processes and enable users and organizations to:

- Open bank accounts easily
- Parse user info easily and store in databases.
- Reduce work and time required on both organization and user's end exponentially

Fig. 5. *Aadhar Data Extraction and Manual Verification Screen.*



About

We strive to automate processes and enable users and organizations to:

- Open bank accounts easily
- Parse user info easily and store in databases.
- Reduce work and time required on both organization and user's end exponentially

Fig. 6. *Aadhar Data Extraction and Manual Verification Screen.*

After implementing both the approaches for data extraction we found out that the ML based approach was inconsistent and had a very high error rate and very low accuracy. The accuracy was around 60% at the 100th epoch. The time taken for the model to load and give results was also large. Data augmentation did not help as there was very little change in the images of the dataset which led to model overfitting. This defeated the purpose of using YOLO as speed and accuracy was its strength.

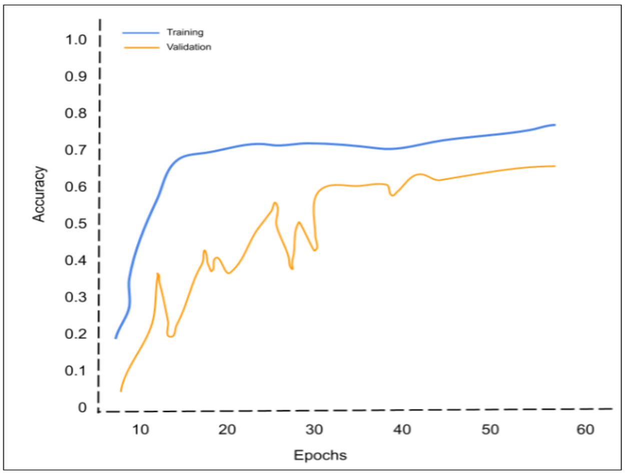


Fig. 7. *ML Accuracy Graph.*

The second approach, Line Matching Algorithm, worked flawlessly for our use-case as it depended on the static positions of the text, the same nature of these documents which led to the failure of the ML approach. We tested the accuracy by running load tests and storing the results in a CSV, out of the 100 iterations this algorithm failed in 17 iterations. After analysis of the 17 iterations, it was observed that when the image has too much noise and is skewed, it leads to the failure of Tesseract OCR which inturn leads to the failure of the line matching algorithm. Taking in consideration that tesseract OCR does not fail, we can get an accuracy greater than 80%.

VII. Conclusion

After rigorous research and development, we conclude that the Line Matching approach for process automation has a better accuracy and provides results which are acceptable.This project intended to reduce human effort while increasing productivity, which was accomplished through quick data extraction and identification as well as a simplified, user-friendly interface. Using this approach, the number of steps necessary to create a bank account can be minimised. allowing banks to concentrate on other issues while saving money.

References

[1] Deloitte-robotic and cognitive automation Deloitte Perspectives on FinancialServices (https://www2.deloitte.com/content/dam/Deloitte/sg/Documents/financial-services/sg-fsi-seminar-2017-robotic-cognitive- automation.pdf)

[2] WorkFusion - Intelligent Automation for Document Processing Learn why Intelligent Automation for document processing is a critical capability for enterprise companies (https://www.workfusion.com/intelligent-automation-for-document-processing/)

[3] Object Detection using YOLOv3 and openCV (https://towardsdatascience.com/object-detection-using-yolov3-and-opencv-19ee0792a420)

[4] Nanonets - Guide to OCR with RPA and document understanding (https://nanonets.com/blog/ocr-with-rpa-and-document-understanding-uipath/)

[5] UI-Path : Document Understanding (https://www.uipath.com/product/document-understanding)

[6] RPA for OCR - Accelirate (https://www.accelirate.com/rpa-intelligent-automation-optical-character-recognition-based-business-processes/)

[7] Angelica Gabasio, "Comparison of optical character recognition (OCR) software." June-2013.

[8] R. Smith, "An Overview of the Tesseract OCR Engine," in Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), 2007, vol. 2, pp. 629–633.

[9] M. Brisinello, R. Grbić, M. Pul, and T. Anđelić, "Improving optical character recognition performance for low quality images," in 2017 International Symposium ELMAR, 2017, pp. 167–171.

[10] M. Shen and H. Lei, "Improving OCR performance with background image elimination," in 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2015, pp. 1566–1570.

[11] Datong Chen, H. Bourlard, and J.-P. Thiran, "Text identification in complex background using SVM," 2001, vol. 2, p. II-621-II-626.

[12] Q. Ye, W. Gao, and Q. Huang, "Automatic text segmentation from complex background," in 2004 International Conference on Image Processing, 2004. ICIP '04, 2004, vol. 5, p. 2905–2908 Vol. 5.

[13] N. Shivananda and P. Nagabhushan, "Separation of Foreground Text from Complex Background in Color Document Images," in 2009 Seventh International Conference on Advances in Pattern Recognition, 2009, pp. 306–309

[14] PROCESS AUTOMATION IN DOCUMENT ANALYSIS : A Survey (Nishad Raisinghani, Ninad Bhole, Sarvesh Kulkarni, Sumit Bhattacharya , Rupali Umbre.