International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 02 | Feb 2022 www.irjet.net p-ISSN: 2395-0072

Optical Character Recognition based KYC System

Manasi Bhabal[1], Dhruv Desai[2], Harsh Desai[3], Prof. Sejal D'mello[4]

[1,2,3]B.E. Student, Information Technology Engineering, Atharva College of Engineering, Mumbai, India
[4]Professor, Information Technology Engineering, Atharva College of Engineering, Mumbai, India

---

Abstract - *The aim of this paper is to build a platform using Machine Learning and Optical Character Recognition to make it easier for customers to complete their KYC (Know Your Customer) for banking purposes such as opening a bank account, applying for a policy, and various other business purposes, for the sake of convenience and time savings. This will not only significantly speed up the KYC process, but it will also eliminate errors. Customers can update their KYC in our system by scanning and uploading their AADHAR and PAN cards. The app will use OCR (Optical Character Recognition) to reduce typing errors, verify documents, and thus auto-fill the form without errors, saving time.*

Key Words: **KYC (Know your customer), OCR (Optical Character Recognition), Tesseract, Authentication, Data Extraction.**

1.INTRODUCTION

Know Your Customer (KYC) is one of the most important processes for any businesses and digitalization of this process is need of the hour. Almost everything in the digital era is automated, and information is stored and communicated in digital formats. KYC can also refer to regulated banking practices that are used to verify clients' identities. The traditional method of working for KYC is a bit tedious and takes a long time to complete. The proposed system is a software-based and dependable KYC method that uses the concept of OCR (Optical Character Recognition) to extract text from images, verifies it, and then stores the data in a database. Additionally, auto form filling is implemented in the project.

1.1 PURPOSE

Customers typically fill out forms on paper sheets by hand. Because of human errors, unclear handwriting, and defective writing materials, this might lead to a lot of discrepancy. This could result in a significant amount of paper waste (considering this is a process adopted worldwide). Next, there could be errors produced by the authorities responsible for data sentry, while referring the handwritten form. Even the current digital KYC system requires manual efforts in data entry. Furthermore, because of the manner customers enter their data, there may be inconsistencies in the customer data, generating serious validation issues for the customer whose primary goal was to authenticate their identity in the first place.

Customer identification also aids in the control of financial fraud, the detection of money laundering and suspicious activities, and the scrutiny and monitoring of large cash transactions. To avoid these problems, the Reserve Bank of India (RBI) directed all banks and financial institutions in India to implement a policy framework that requires them to know their customers before opening any accounts. This entails verifying customers' identities and addresses by requesting documents that are accepted as relevant proof. KYC norms require proof of identity and proof of residence as mandatory details. Proof of identity can be a passport, voter's ID card, Permanent Account Number (PAN) card, or driving license, and proof of residence can be a ration card, an electricity or telephone bill, or a letter from the employer or any recognized public authority certifying the address, in addition to proof of identity being used as residence proof if they carry address.

1.2 OBJECTIVES

- To ease the process of KYC for various organizations.
- To create an E-KYC system and make it convenient for users to complete their KYC.
- To extract data from the uploaded documents of users.
- To verify and authenticate the uploaded documents
- To auto fill the user form and complete the process.
- Making a database of the information gathered for each customer for further use.
- Driving down cost of operations by automating the whole process.
- By increasing customer satisfaction in a number of different functional operations across the organization.

2. FEASIBILITY

2.1. Technical Feasibility

The technical resources (hardware and software) required to build the project are the focus of technical feasibility. It also investigates the specifics of how you intend to deliver the product and whether the technical team is capable of

translating the concept into working systems. In our project, we are creating a website with HTML and CSS and storing the extracted data in an SQLite database.

In addition, we use Python libraries to implement the algorithm. Other than a laptop, no additional hardware is required for this project.

2.2. Economic Feasibility

The economic feasibility of a project is determined by the costs incurred in the project's development. HTML and CSS were used to build this site. It is a costless language. This project does not necessitate the use of any additional hardware. As it is purely a software project, we did not have to spend any money on it.

2.3. Operational Feasibility

It evaluates how well a proposed system solves problems and meets the requirements identified during the requirement analysis phase. The primary goal of this project is to make the KYC process as simple as possible for employees while reducing the risks associated with the manual process. This project implements KYC in a single click with form automation, which is superior to traditional KYC methods.

3. PROPOSED SYSTEM

The flow chart in figure 2.1. illustrates how the system works. First, the user registers in the system, after which the home page is displayed. Next, the user crops and uploads images of an Aadhar card and a Pan Card, and the data is extracted using OCR. Following data extraction, each uploaded document is authenticated by a verification process, and a KYC form is automatically filled out and the process is terminated.

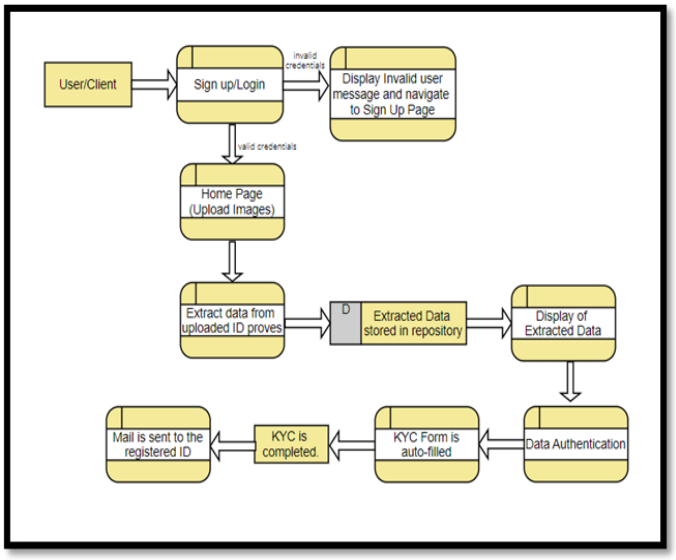

Fig-2.1. Flowchart

4. METHODOLOGY

Data extraction, authentication of ID proofs, and auto-filling KYC forms are the three primary phases in the system. Data from the customer's documents is extracted and mapped, then saved in a database for future auto-form filling. As soon as the customer presents his documents, the system identifies the relevant elements in the image. To auto-fill the form, the extracted data from the photos is checked first, then mapped to the form fields. After the form has been correctly filled, the customer is notified through e-mail about successful completion of KYC.

Data is extracted from images using the Tesseract Python library. The algorithm maps fields on images where data is to be fetched from. Once this data is fetched, the extracted data is displayed to the user to ensure its accuracy.

Furthermore, the retrieved data is checked to confirm that the provided documents are authentic. To obtain the API for verifying Indian Aadhar cards, an official application should be submitted to UIDAI at https://www.uidai.gov.in/. The user clicks the "Fill Form" button after the obtained data has been verified. The KYC form is auto-filled, and a Selenium web-driver is utilized to connect an external form to our system in order to accomplish this.

Tesseract is a text recognition (OCR) engine that is open source and licensed under the Apache 2.0 license. It can be used directly or through an API (for programmers) to extract written text from photos. It helps us with a wide range of languages. Tesseract doesn't come with a graphical user interface, however there are a few on the 3rdParty website. Tesseract can be used with a variety of computer languages and frameworks thanks to wrappers available. It can be used in conjunction with the existing layout analysis to recognize text inside a huge document, or with an external text detector to recognize text from a single text line image.

5. SCOPE

KYC is performed using a few required government documents. This system is not intended for all documents. The KYC is performed by extracting relevant details from the predefined documents. Furthermore, the form is pre-defined, and the extracted details are eventually mapped to the form.

The system can be improved further to be more secure and to perform KYC with more documents. Basic data is gathered, and software generally compares it to lists of individuals known for corruption, sanctioned, suspected of committing a crime, or at high risk of bribery or money laundering.

6. RESULTS

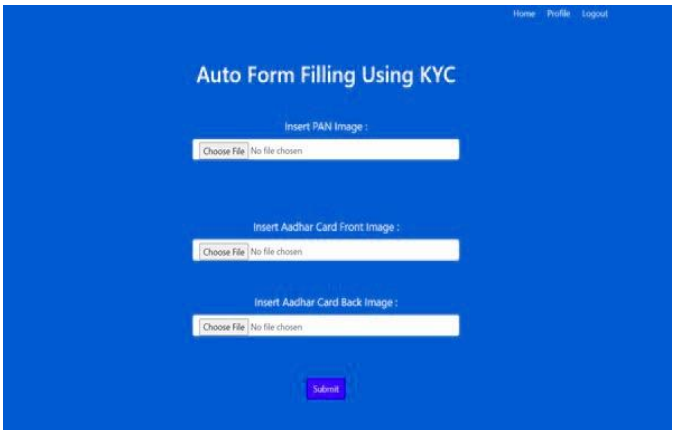

Figure 6.1

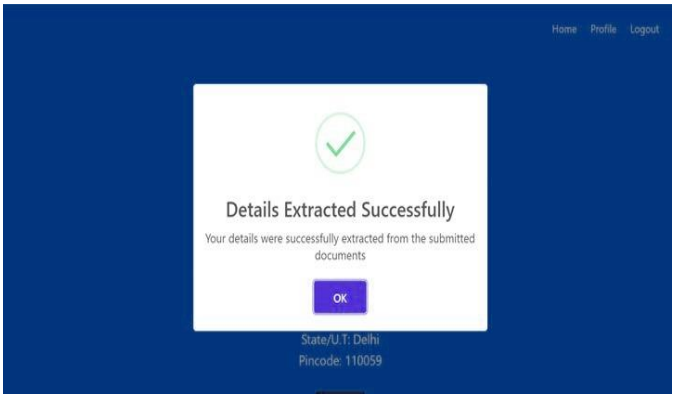

Figure 6.2

7. CONCLUSIONS

The keyboard is the most common way to enter data into a computer. However, this isn't always the best or most efficient solution. The goal of this system is to develop a new KYC system that allows users to complete KYC in a few clicks and includes verification and auto-form filling with user data that is precisely mapped to the form. Although many processes are still carried out on paper, it is clear that automatic data recognition technologies are gaining popularity. During subsequent processing steps, the document is repeatedly copied and changed, resulting in a large number of copies. In some cases, they can help humans, but in others, they are useless. The aim of this system is to successfully enter data into the form by accurately mapping every user detail from the database to the form field using OCR technology without compromising data quality.

8. ACKNOWLEDGEMENT

We would like to express our heartfelt gratitude to our guide, Professor Sejal D'mello, for her motivation and guidance throughout the process. We would also like to thank the faculty of the Information Technology Department for their valuable assistance with our project.

9. REFERENCES

[1] Rishabh Mittal and Anchal Garg, "Text extraction using OCR: A Systematic Review." IEEE (2020)

[2] S. Tomovic, K. Pavlovic, M. Bajceta, "Aligning document layouts extracted with different OCR engines with clustering approach." Science Direct (2020)

[3] Yash Kumar, Gaurav Sharma, Komal, Prof. Audumbar Umbare "E-KYC Mobile Application using Optical Character Recognition". IRJET (2020)

[4] Hejing Wu, Fang Liu, Long Zhao Yabin Shao, "Data Analysis and Crawler Application Implementation Based on Python." IEEE 2020

[5] S. D. Bandari, Ankita Jagtap, Namrata Mane, Swarali Garud "Intelligent Framework for Auto filling Web form using Scanned Documents" IJISRT

[6] A Al Mamun, SR Hasan, MS Bhuiyan, M Shamim Kaisar, Mohammad Abu Yousuf Secure and Transparent KYC for Banking System Using IPFS and Blockchain Technology. IEEE (2020)