IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-11, Nov 2019 E-ISSN: 2347-2693

Aadhar & Driving License Information Extraction System

Kothagattu Surya Teja[1*], Kunam Siri Chandana[2], R. Srinivas[3], B. Prasanthi[4]

[1,2,3,4]Dept. of Computer Science, Mahatma Gandhi Institute of Technology, Hyderabad, India

[*]*Corresponding Author ksteja99@gmail.com, Tel.: +91-9603370905*

DOI: https://doi.org/10.26438/ijcse/v7i11.173176 | Available online at: www.ijcseonline.org

Accepted: 07/Nov/2019, Published: 30/Nov/2019

***Abstract*—** OCR based Aadhar & driving license info Extraction System may be a time period embedded system that mechanically acknowledges the kind of document whether or not it's a Aadhar or driving license and extract the out there info from it. There square measure several applications starting from advanced security systems to common official work. OCR primarily based Aadhar & driving license info Extraction System has advanced characteristics because of various effects like totally different pattern in numerous Aadhar Card, totally different Spacing in text etc. Most of the OCR primarily based info Extraction System square measure designed mistreatment proprietary tools like MATLAB that takes a protracted method and time and conjointly will have many limitations and conjointly they're unable to sight pattern and can't extract the desired info severally. this concept presents an efficient technique of implementing OCR primarily based Aadhar & driving license info Extraction System mistreatment Free software package together with Python and therefore the Open pc Vision Library.

***Keywords*:** Software Requirement, Python, OpenCV, Tesseract.

I. INTRODUCTION

Aadhar and driver's license area unit getting used to unambiguously establish someone and is additionally used as address proof, proof of family etc. OCR primarily based Aadhar & driver's license data Extraction System will plays a vital role in several applications like storing worker knowledge, knowledge security, extracting knowledge from Brobdingnagian image info etc. OCR primarily based Aadhar & driver's license data Extraction System is convenient and price economical because it is machine-controlled.

II. EXISTING SYSTEM

The scientific world is deploying analysis in intelligent transportation systems that have a big impact on people's lives. OCR primarily based Aadhar & license info Extraction System may be a laptop vision technology to extract the name, address, ID variety etc. from pictures. it's Associate in Nursing embedded system that has various applications and challenges. OCR primarily based info Extraction System are enforced exploitation proprietary technologies like MATLAB that are pricey and not economical. they solely extract the alphamerical characters and don't reason information the info the information in correct format as they cannot discover that data is what as a result of typically Aadhar and license don't follow a strict pattern. This closed approach conjointly prevents any analysis and development of the system at a reasonable worth from several free sources.

III. PROPOSED SYSTEM

In India, basically, there are two kinds of most used ID Proof, Aadhar Card and Driving License. Our proposed system is to show that free and open source technologies are matured enough for scientific computing domains. Also, we have implemented noise reduction techniques to increase more accuracy in detection of information. Moreover, we have implemented several custom-made algorithms for recognition of type of document, data coming from document, category of data etc. Our system not just extract the data, but utilizing the algorithms it can automatically detect which text is name, which text is address, which text is DOB etc. Further all these data is stored in databases and can be searched with no hassle.

The system works satisfactorily for wide variations in illumination conditions and different types of Aadhar Cards and Driving License commonly found in India. It is definitely a better alternative to the existing proprietary systems, even though there are known restrictions with high resolution to detect and extract information using OpenCV and python which is most easy to understand and make changes.

IV. SOFTWARE REQUIREMENTS

Python:

Python is associate degree understood, high-level, general purpose artificial language. Created by Guido van Rossum and initial free in 1991, Python's style philosophy emphasizes

code readability through use of great whitespaces. Its language constructs and object-oriented approach aims to assist programmers write clear, logical code for little and large-scale comes.

Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, as well as procedural, object-oriented, and purposeful programming. Python is usually represented as a "batteries included" language thanks to its comprehensive customary library.

OpenCV:
OpenCV was started at Intel in 1999 by metropolis Bradsky, and therefore the 1st unharness came enter 2000. Vadim Pisarevsky joined metropolis Bradsky to manage Intel's Russian computer code OpenCV team. In 2005, OpenCV was used on Stanley, the vehicle that won the 2005 Defense Advanced Research Projects Agency Grand Challenge. Later, its active development continued underneath the support of Willow Garage with metropolis Bradsky and Vadim Pisarevsky leading the project. OpenCV currently supports a mess of algorithms associated with pc Vision and Machine Learning and is increasing day by day.

OpenCV supports a good form of programming languages like C++, Python, Java, etc., and is obtainable on totally different platforms as well as Windows, Linux, OS X, Android, and iOS. Interfaces for high-speed GPU operations supported CUDA and OpenCL are underneath active development.

OpenCV-Python is that the Python API for OpenCV, combining the simplest qualities of the OpenCV C++ API and therefore the Python language.

Tesseract:
Tesseract was originally developed at Hewlett-Packard Laboratories city and at Hewlett-Packard Co, Greeley Colorado between 1985 and 1994, with some a lot of changes created in 1996 to port to Windows, and a few C++sizing in 1998. In 2005 Tesseract was open sourced by power unit. Since 2006 it's developed by Google

OpenCV-Python is that the Python API for OpenCV, combining the most effective qualities of the OpenCV C++ API and also the Python language.

V. ALGORITHM

There square measure 2 basic forms of core OCR rule, which can manufacture smart result.

Matrix matching involves examination a picture to a keep glyptography on a pixel-by-pixel basis; it's additionally referred to as "pattern matching", "pattern recognition", or "image correlation". This depends on the input glyptography being properly isolated from the remainder of the image, and on the keep glyptography being in an exceedingly similar font and at an equivalent scale. this system works best with typed text and doesn't work well once new fonts square measure encountered. this can be the technique the first physical photocell-based OCR enforced, rather directly.

Feature extraction decomposes glyphs into "features" like lines, closed loops, line direction, and line intersections. The extraction options reduce the spatiality of the illustration and makes the popularity method computationally economical. These options square measure compared with associate degree abstract vector-like illustration of a personality, which could scale back to 1 or a lot of glyptography prototypes. General techniques of feature detection in laptop vision square measure applicable to the present form of OCR, that is often seen in "intelligent" handwriting recognition and so most up-to-date OCR software system. Nearest neighbor classifiers like the k-nearest neighbors rule square measure wont to compare image options with keep glyptography options and select the closest match.

Tesseract use a two-pass approach to character recognition. The second pass is thought as "adaptive recognition" and uses the letter shapes recognized with high confidence on the primary pass to acknowledge higher the remaining letters on the second pass. this can be advantageous for uncommon fonts or low-quality scans wherever the font is distorted (e.g. blurred or faded).

In our project, once extraction of all the characters we have a tendency to square measure applying our custom created pattern recognition to classify the sort of document and additionally classify the data in numerous classes accurately like Name, DOB, Address etc. This info are kept in SQLite databases

VI. DESIGN METHODOLOGY

Figure 1 represents the design flow chart of this OCR based Aadhar and driving license information extraction software.

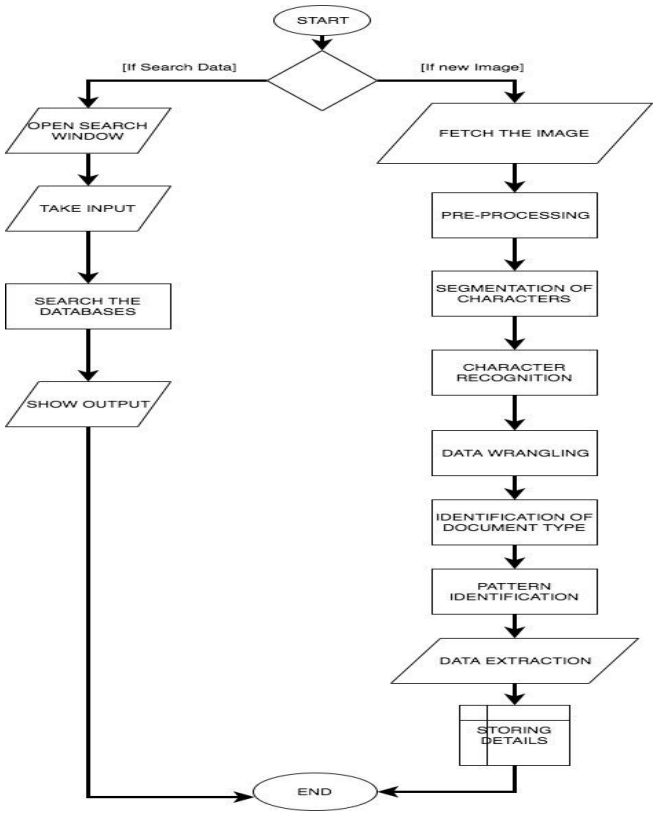

Figure 1. Flow chart diagram of Aadhar & Driving License Information Extraction System

VII. TESTING AND RESULTS

The interface of this system looks like figure 3 and figure 4.

STEP 1:-
Multiple input images are selected at once on clicking 'select file'. For example, figure 2 and 3

Figure 2. input file

Figure 3. input file

Step 2:-
To upload all the images, click on 'Upload'. As shown in figure 4

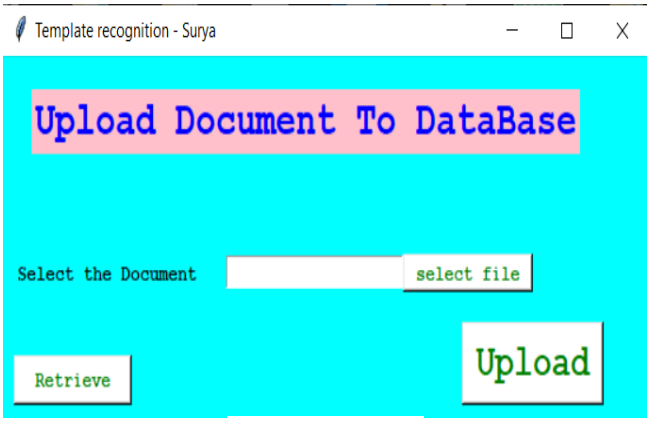

Figure 4. interface

Step 3:-
To retrieve data from the images, click on 'Retrieve'. As shown in figure 5

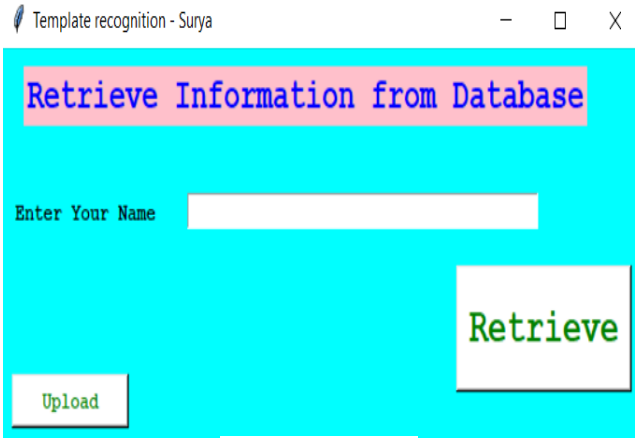

Figure 5. Interface

Step 4:-
On searching a text that is on the image will fetch you the data available. As shown in figure 6 and 7

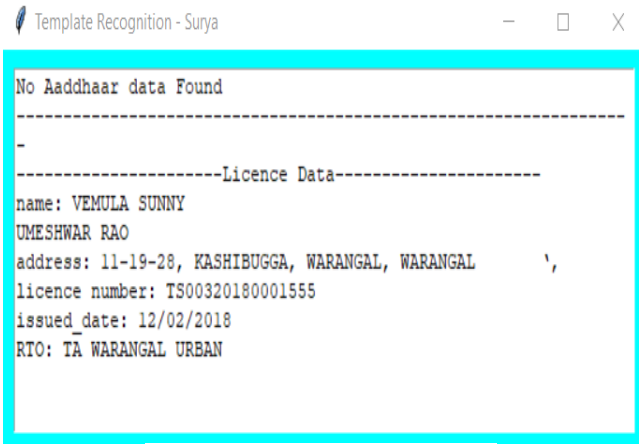

Figure 5. Output after Retrieving

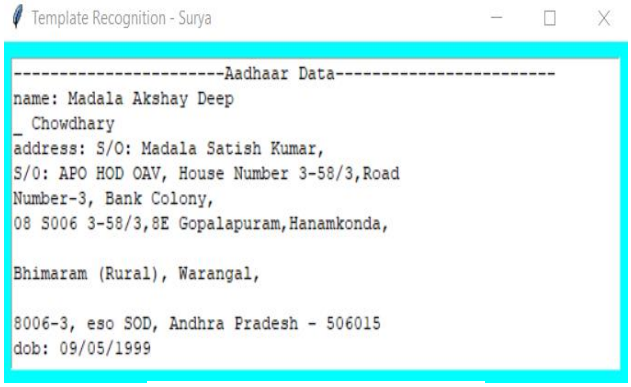

Figure 6. output after retrieving

VIII. FUTURE SCOPE

Aadhar & Driving License Information Extraction System can store all the extracted information according to fields that are present on Aadhar and driving license. Similarly, this can be used to fill forms in institutions in which forms filled manually from the hard copy.

IX. CONCLUSION

Aadhar & Driving License Information Extraction System will give the accurate information available in the images that are used as inputs for this software. The data extracted from this software is stored in SQLite database where the data is stored as two different tables one contains Aadhar and other contain license details in it.

The data that appears is not just the data obtained after OCR. After performing OCR, the data is transferred to database. And the data id fetched from here.

REFERENCES

[1] Rafi, Ali, Faraz, Athaul, "OCR Engine to extract Food-items and Prices from Receipts Images via Pattern matching and heuristics approach", SMIU, 1st International Conference on computing and related technologies, 2017 [Souvik Das "The Development of a Microcontroller Based Low Cost Heart Rate Counter for Health Care Systems" International Journal of Engineering Trends and Technology- Volume4Issue2- 2013.

[2] Chaki, Nabendu, Soharab Hossain Shaikh, and Khalid Saeed. "A comprehensive survey on image binarization techniques." In Exploring Image Binarization Techniques, pp. 5-15. Springer India, 2014.

[3] Zhang, Mi, Anand Joshi, Ritesh Kadmawala, Karthik Dantu, Sameera Poduri, and Gaurav S. Sukhatme. "OCRdroid: A Framework to Digitize Text Using Mobile Phones." In MobiCASE, pp. 273-292. 2009.

[4] Brisinello, Matteo, Ratko Grbić, Matija Pul, and Tihomir Anđelić. "Improving Optical Character Recognition Performance for Low Quality Images." In 59th International Symposium ELMAR-2017. 2017.

[5] ZHAO, Yan, Yue CHEN, and Shi-gang WANG. "Corrected fast SIFT image stitching method by combining projection error." Optics and Precision Engineering 6 (2017): 029.

Authors Profile

Mr. Kothagattu Surya Teja, Student of Bachelors in Engineering at Mahatma Gandhi Institute of Technology, Hyderabad. He worked on projects with IT giants like Tech Mahindra in Jun 2019 on Robotic Process Automation. His places of interest in researches are IOT, Artificial Intelligence, Machine Learning, Web technologies. He worked on projects with Innoprime technologies in June 2018 on web technologies and Andriod application development .

Ms. Kunam Siri Chandana, Student of Bachelors in Engineering at Mahatma Gandhi Institute of Technology, Hyderabad. She worked on projects with IT giants like Tech Mahindra in Jun 2019 on Robotic Process Automation. Her places of interest in researches are IOT, Artificial Intelligence, Machine Learning, Web technologies.

Mr.R.Srinivas, Senior Assistant Professor at Mahatma Gandhi Institute of Technaology, Hyderabad. He is working as assistant professor from the past 16 years.He was involved in many projects which got published in national and international journals. He has a PhD in Digital Image Processing and Computer Vision.

Ms.B.Prasanthi, Senior Assistant Professor at at Mahatma Gandhi Institute of Technaology, Hyderabad. She is working as assistant professor from the past 15 years. She was involved in many projects which got published in national and international journals and also in many International conferences. He has a PhD in Image Processing.

