International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 04 | Apr 2019 www.irjet.net p-ISSN: 2395-0072

Offline Transcription using AI

Priyanka Patil[1], Babli Sah[2], Ruchita Shah[3], Jyoti Deone[4]

[1,2,3,4]*(Dept. Of Information Technology, Ramrao Adik Institute of Technology, Navi Mumbai, Maharashtra)*

Abstract - *Regardless the abundance of technological writing tools, people as well as many organizations still take their notes traditionally: with pen and paper. In this paper, offlinehandwritten character recognition and image processing is applied to read the usual detail forms,wherein we have used aadhaar link form. The scanned image of this form is being pre-processed using morphological operations to remove noise and localize the boxes. For offline handwritten character recognition we have developed a tool which is based on an open source tesseract optical character recognition engine. It successfully reads the whole form containing handwritten English capital letters and digits which gives more than 80% accuracy. The extracted text is post-processed to get required information to store it in excel sheet with information. It involves the automatic conversion of input image form into document which areeditable and usable within computer to store link to database and other text processing applications.*

Key Words*: *Handwritten character recognition, Image Processing, Recurrent Neural Network, Tesseract engine.

1. INTRODUCTION

Machine reading is a difficult task. It is one of the most engrossing and challenging field of pattern recognition among the researchers. Handwritten character recognition principally entails optical character recognition from an image or video. It is broadly divided into two ways offline and online. In an on-line method, handwritten character is characterized by structure -or- shape-based representation of a stroke on touch pad using special pen. Off-line character recognition involves scanning a form or document written sometime in the past. Offline handwritten character recognition is a very problematic research area because writing styles may vary from one user to another [1]. The proposed system is automation of the manual paper processing at low cost. Many organizations use forms with boxes. These forms are hand filled by different users, one character or digit per box. Thus, the problem we define here is to read the user filled data with the system to automate the manual reading of the form. The system itself would read all the handwritten information and extract more amount of information in less amount of time, thus reducing the cost of manual data processing. This can be used by Aadhar link centers, NGO's.

Our approach to reach the goal will include image processing, computer vision and tesseract engine. Image processing is a technique to change over an image into digital shape and play out a few operations on it, so as to get an improved image or to separate some helpful information from it. It is a sort of signal administration in which input is image, similar to video edge or photo and yield might be image or qualities related with that image. Image processing is used in almost every domain, pattern recognition, object recognition, security, etc. Image recognition also known as computer vision is a technical field that deals with searching the ways to automate all the job that a human visual system can do. But, challenges are with handwriting text as it's hard to store and access physical records in an effective way, look through them efficiently and to share them with others. Hence, a great deal of critical knowledge gets lost or does not get inspected in view of the way that reports never get transformed to digital format. So in this manner we chose to handle this issue, since we believe the significantly greater ease of management of digital text compared to composed content will help individuals all the more viably get to, pursuit, share, and examine their records, while as yet permitting them to utilize their favored composition strategy.

1.1 LITERATURE REVIEW

Today various research on image based character recognition has been proposed with different algorithms in different applications. One of which is proposed by Fabian Tschopp in Nov 2016,who enhanced this by presenting a three layer convolutional neural network (CNN) for efficient pixelwise classification of images[2]. The most punctual Artificial Neural Network models were examined in the mid1940s as models of biological neural networks. It wasn't until the thought of error correction and back propagation algorithms that the ANN wound up pervasive in processing essentially for its learning ability [3].Numerous Traditional OMR work with a committed scanner gadget that reflects a light emission onto the structure paper. The contrasting reflectivity at predefined positions on a page is then used detected these marked areas because they reflect less light than the blank areas of the paper. With printed or cursive composition, specifically, the product is as of now unfit to render these sorts of reports which are machine-coherent. The driving force behind handwritten text classification was for digit classification for postal mail. OCR innovation has for quite some time been utilized by the United States Postal

Service, among different associations, to peruse addresses on mail[4].

Algorithms of computer vision able to read the string of handwritten digits developed by Yann la cun, using neural network. These algorithms have been incorporated into a system that reads the handwritten-digits in the U.S. mail[5]. Most of the researches are working on accuracy of recognition with the help of neural network. Earlier KNN classifier and Hidden markov model was used to identify and classify characters [6]. Neural Networks have been successfully applied to pattern recognition, association and classification. In previous studies, ANNs have proven to be excellent recognizers of printed that is machine type characters and handwritten digits (0~9) [7]. RNN's were initially created in the 1980's, but can only show their real potential since a few years, because of the increase in available computational power, and to handle the huge amounts of data that we create nowadays. LSTM, long-short term memory was invented in the late 1990's, because of their internal memory, RNN's are able to remember important things about the input they received, that enables the model to be very precise in predicting about every new challenge[8]. This is the reason why we have chosen to use this algorithm to read the form characters. Enhancing OCR programming in handwriting recognition fits on perfect for Tesseract open source OCR engine taking to the another level of accuracy and speed.

2. PROPOSED WORK

Development process is divided in following steps:

- Collection of handwritten data sets.
- Localization
- Character Recognition
- Formatting
- Database (Excel sheet)

System architecture of any project gives the complete insight of the project. It alludes to the high-level structure of a software and the control of making such logical and systematic structures of frameworks. System Architecture of this project is as in Fig-1.

Input Material

Input material includes a hard copy of forms (with squares in it) which are being filled manually by different people with different handwriting. We need to provide this form to the system for digitizing purpose i.e. to convert them into machine readable format. Here we receive the scanned document in the form of image so that it will be easy to read the document and perform operations. The crucial part of the project is this, localizing the boxes on the scanned sheet. A box detection function applied over the preprocessed image to detect each box based on the predefined kernel length which can be updated as per size of the box.

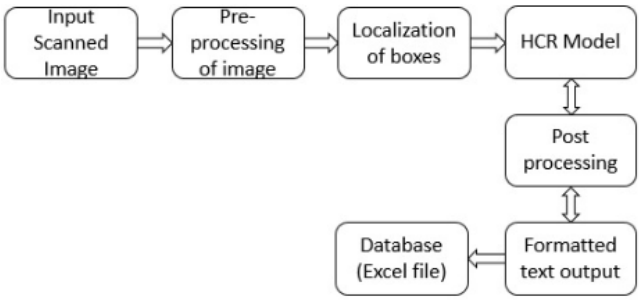


Fig-1: System Architecture

HCR Model:

Today researches are ongoing in neural network for improving the accuracy of the model whether by rigorous training with huge data set or by improving the algorithm. Recurrent Neural Networks (RNN) is a powerful type of neural networks because they are the only ones with an internal memory. Tesseract OCR engine with LSTM is a prominent method which uses RNN in its architecture. Thus this HCR model is able to read the handwritten as well as the printed text of visible length size of minimum 20 pixels.

Post-processing:

In post-processing, the output or the data obtained is understood and analyzed for its accuracy. Accuracy will depend on the number of correctly identified letters with respect to total number of letters. Further the mapping of the content is done that is, data at its specific title of data into the excel sheet. Now, data processing techniques are applied to convert the raw data into meaningful format. Formatted text is the output of post processing in editable format and stored in database.

2.1. IMPLEMENTATION

The proposed system is automation of the manual paper processing at low cost. Many organizations use forms with boxes for example Fig-2. These forms are hand filled by different users, one character or digit per box. Thus, the problem we define here is to read the user filled data with the system to automate the manual reading of the form. The system itself would read all the handwritten information and extract more amount of information in less amount of time, thus reducing the cost of manual data processing. Our approach to reach the goal will include image processing and neural network.

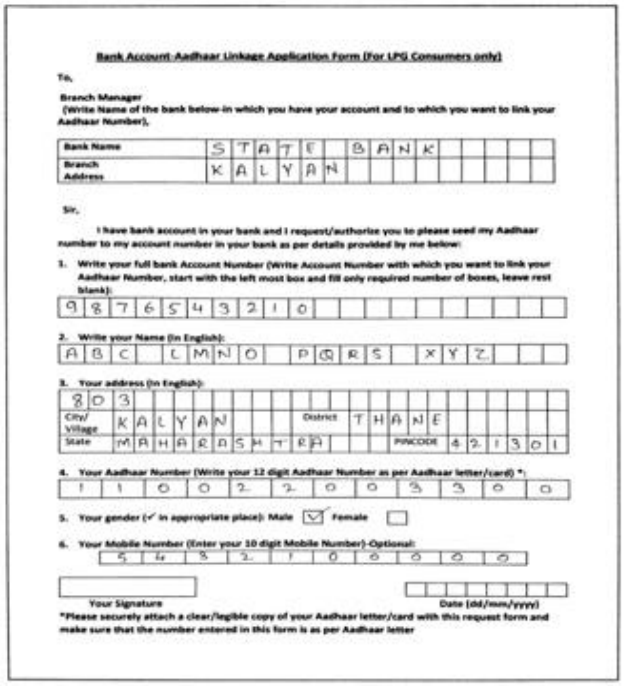


Fig-2: Input Image

Image pre-processing: Image processing is a technique to change over an image into digital shape and play out a few operations on it, so as to get an improved image or to separate some helpful information from it. It is a sort of signal administration in which input is image, similar to video edge or photo and yield might be image or qualities related with that image. Image processing is used in almost every domain, pattern recognition, object recognition, security, etc. Image recognition also known as computer vision is a technical field that deals with searching the ways to automate all the job that a human visual system can do. As already implemented Tensor Flow by Google, Deep Face by Facebook and many more.

Morphological operations: Image enhancement entails removal of different types of noise in the input image. There is a need to boost the quality of image for proper recognition of information present in it. The steps followed are described in the Fig-3.

Input Image: The input image is required to be scanned through digital camera or any other suitable digital input device in a readable quality not necessarily high quality as it undergoes filtering while pre-processing stage Fig-2. The system supports different file formats such as JPEG, PNG and PDF.

Thresholding: So as to decrease storage necessities and to expand processing speed, it is often desirable to represent grey scale or color images as binary pictures by picking some threshold value for everything over that value is set to 1 and everything beneath is set to 0. In this we threshold the image to separate foreground and the background. It gives clarity of pixel in black and white, an acceptable image for further processing.

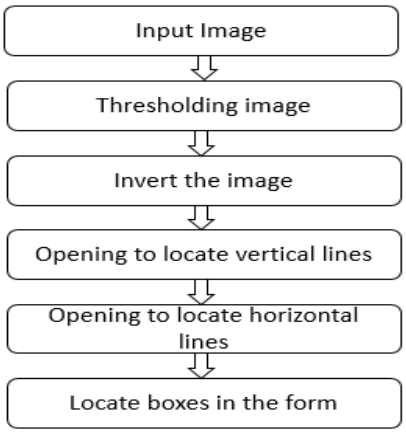
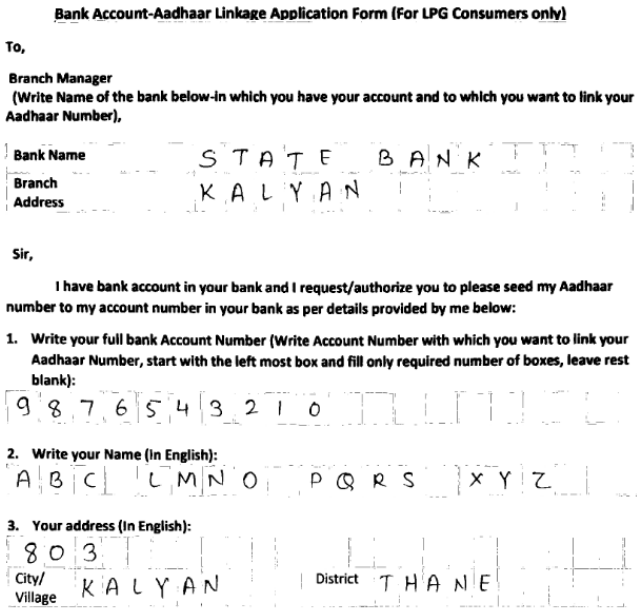


Fig-3: Localization of boxes

Opening: Opening comprises of erosion followed by dilation. a structuring element is defined as per the required box size and can be varied independently. A kernel size is set to the minimum height and width which tends to find the horizontal and vertical lines as per structuring element. Thus the boxes are localized and then they are removed from the original image and further recognition techniques can be applied. Output of the above process is shown in the Fig.4, where boxes in form are successfully detected with their respective co-ordinates. The co-ordinate values are based on the width, height, left and right position respective of the given image.

Handwritten Character Recognition: Handwriting recognition is the most explored area which has led to many inventions. However, a complete handwritten recognition system should also handle formatting and editing of data. In the past few years, Deep Learning and neural network based methods have surpassed traditional machine learning techniques by a vast scope in terms of accuracy in many areas of Computer Vision. Handwriting recognition is one of the high-flying examples. So, there was time before when Tesseract too had a Deep Learning based recognition engine. To recognize an image containing a single character, a traditional convolutional neural network can be used successfully for predictions. But they lack while implementing over a string of characters or a sequence of characters, such problems can be solved using RNNs and LSTM. They are widely considered to solve the sequence prediction problem, because of their property of selectively remembering and forgetting patterns for long durations of time. LSTM has explicitly introduced a memory unit called cell into the network. This single unit makes decision by considering the current input, previous output and previous memory. And it generates a new output and alters its memory.


Bank Account-Aadhaar Linkage Application Form (For LPG Consumers only)

To,

Branch Manager
(Write Name of the bank below-in which you have your account and to which you want to link your Aadhaar Number),

Bank Name STATE BANK
Branch Address KALYAN

Sir,

I have bank account in your bank and I request/authorize you to please seed my Aadhaar number to my account number in your bank as per details provided by me below:

1. Write your full bank Account Number (Write Account Number with which you want to link your Aadhaar Number, start with the left most box and fill only required number of boxes, leave rest blank):
9 8 7 6 5 4 3 2 1 0

2. Write your Name (In English):
A B C L M N O P Q R S X Y Z

3. Your address (In English):
8 0 3
City/ Village KALYAN District THANE

Fig-4: Output of localization process

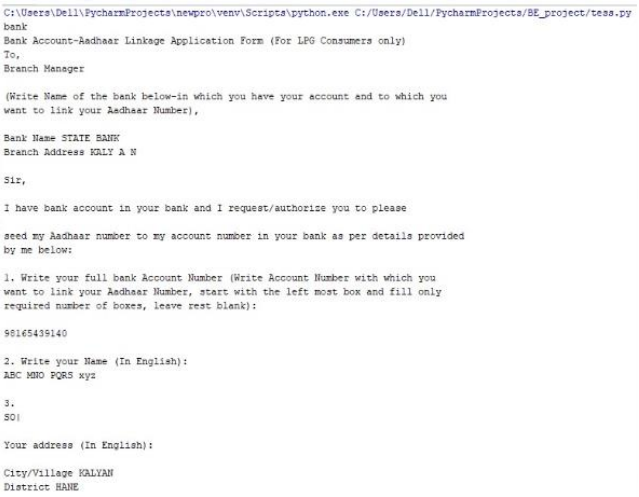
In this paper, Tesseract engine with the recurrent neural network (RNN) is used to recognize the digits and characters in the image. The Tesseract OCR engine library is implemented with the same. Technologies used for development are python programming language and pyCharm IDE. Tesseract library of python called pytesseract is integrated with for character recognition. We have fed in 10 such forms shown above, these forms are hand filled by different people. The results are pretty accurate for bold characters. The faded ones are incorrectly detected as shown in the Fig-4. We have calculated the accuracy of the recognition for one form as

$$\text{Accuracy } \% = \frac{\textit{no. of correctly recognized characters}}{\textit{total no. of handwritten characters}} \times 100$$

Taking the example form there were total 78 handwritten character out of which 66 were correctly recognized, thus giving the accuracy of 84.61%. Similarly we obtain efficiency of this tool over the sample handwritten datasets.

2.3. EXPERIMENTAL RESULT

Input was given in the form of images. We applied the tool over the English alphabet along with the digits. The language contains 26 letters as forms have general rules of filling into capitals only. Along with this it include recognition of digits from 0-9, for fields like contact number, etc. The algorithm of box detection applied on the input image to localize the boxes. The output of this process results in the elimination of the box leaving the letters into them beside along with the co-ordinates of the boxes, in image format shown in fig.4. Now we are ready to apply the algorithm over the resultant image which has only characters. The model reads each character in the document and prints in the console window, Fig-5. The final formatted data is obtained after some file operations over the output and extract only useful data for example the name, address, aadhaar number, etc. This information is wrote into an excel file where each row has the sequential data of one form each row, Fig-6. Thus, there can be multiple reads and then stored in database.


```
C:\Users\Dell\PycharmProjects\newpro\venv\Scripts\python.exe C:/Users/Dell/PycharmProjects/BE_project/tess.py
bank
Bank Account-Aadhaar Linkage Application Form (For LPG Consumers only)
To,
Branch Manager

(Write Name of the bank below-in which you have your account and to which you
went to link your Aadhaar Number),

Bank Name STATE BANK
Branch Address KALY A N

Sir,

I have bank account in your bank and I request/authorize you to please

seed my Aadhaar number to my account number in your bank as per details provided
by me below:

1. Write your full bank Account Number (Write Account Number with which you
went to link your Aadhaar Number, start with the left most box and fill only
required number of boxes, leave rest blank):

98165439140

2. Write your Name (In English):
ABC MNO PQRS xyz

3.
SO!

Your address (In English):

City/Village KALYAN
District HANE
```

Fig-5: HCR Output

	A	B	C	D
1	Bank Name	Branch Address	Account Number	Name
2	PunjabNational	abc	0123456789	xyz
3	STATEBANKOFINDIA	XYZ	5439824453	ABC

Fig-6: Data Table

3. CONCLUSIONS AND FUTURE WORK

The system automates handwritten paper processing which will play an important role towards digitization of various systems across the world. We have combined the efforts of researchers to develop a required product which is able to detect contours of square present in the dummy form and the individual letters within them along with the machine typed text. The traditional tools are used to read printed text whereas handwritten texts are identified by neural network techniques.

The aim is to read machine printed text along with some handwritten text (in squares) in range of English alphabet (A a-Z z) and numbers(0-9). So the future scope to this project are engross different regional languages within the system, automation for reading new/different types of forms, system powerful enough to process low quality image.

REFERENCES

[1] Ayush Purohit, Shardul Singh Chauhan: "A Literature Survey on Handwritten Character Recognition",Centre for Information Technology, University of Petroleum and Energy Studies Dehradun, India,2016.

[2] Fabian Tschopp:"Efficient Convolutional Neural Networks for Pixelwise Classification on Heterogeneous Hardware Systems", Department of Computer Science, ETH Zurich, 2015.

[3] Tristan Wright:"Handwriting Recognition with Artificial Neural Networks and OpenCV",CS488-Senior Capstone, 2012.

[4] KazJaszczak: "Optical character recognition:A backbone for postal and mailing service application".

[5] Yann le cun:"Reading handwritten digits: A Zip-Code recognition system", AT&T laboratories, Holmdel.

[6] ElieKrevat, Elliot Cuzzillo:"Improving Off-line Handwritten Character Recognition with Hidden Markov Models", Department of Computer Science Carnegie-Mellon University.

[7] SavithaAttigeri:"Neural Network based Handwritten Character Recognition system", M.Tech, Department of Computer Science and Engineering STJIT,Ranebennur, March 2018.

[8] Seong-Whan Lee, Young- Jaon Kim :"A new type of recurrent neural network for handwritten character recognition", Department of Computer Science Korea University 1, 5-ka, Anam-dong, Seongbuk-ku Seoul 136-701, Korea.