International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 02 | Feb-2018 www.irjet.net p-ISSN: 2395-0072

Handwritten Text Recognition of Document Form Using Machine Learning

Er. Asadullah Shaikh[1], Shaikh Arsalan[2], Khan Mohammad Altaf[3], Sayyed Athar[4]

[1]*Professor, Department of Computer Engineering, M.H. Saboo Siddik College of Engineering, Mumbai*
[2,3,4] *Department of Computer Engineering, M.H. Saboo Siddik College of Engineering, Mumbai*

Abstract - *An OCR is a wide field of research in recognizing patterns, it is one the most important study and area of interest in computer vision and artificial intelligence .Early versions were very slow and inefficient and they to be trained with images of a single character and worked on one font at a time. Advanced systems can recognize different fonts and have high reliability and accuracy are now popular and common, they support a wide variety of file formats and image formats. Handwriting movement analysis can be used as input to handwriting recognition. This increases the reliability of the end-to-end process more accurate.*

Till now all the solutions available for OCR is for printed text even if there is some system available for handwritten text the accuracy is very low. As most of the NGOs did the surveys to different rural areas where they fill the form in handwritten text after getting the information inserting it into the system is very slow and cumbersome process so we are planned to build a software which will take an image of the form using camera or scanner as an input and try to convert it into text by using image processing and machine learning. As In our system we are particularly focusing on the NGO's survey form, that's why we might achieve high accuracy than the other existing OCR handwritten system.

Key Words: **Optical Character Recognition, Handwritten Text, Machine Learning, Pre-Processing, Segmentation, Recognition.**

1. INTRODUCTION

The problem is to be able to build software which will allow to set up of different types of forms in the system. Based on the form which is set up, the solution needs to be able to recognize English Handwritten alphabets and numbers and store them as information.

There are a large number of OCR software in the industry which allow converting scanned images of documents into searchable text. These work well when the content in the image is in printed form however, this challenge looks at the ability to convert English Handwritten (block capitals) alphabets and number from different kinds of forms and store them.

In the running world, there is growing demand for the software systems to recognize characters in a computer system when information is scanned through paper documents as we know that we have a number of survey forms and other forms which are in the printed format and handwritten block English character. These days there is a huge demand for "storing the information available in these forms into a computer storage disk and then later reusing this information by searching process". One simple way to store information in these paper forms into a computer system is to first scan the forms and then store them as images. But to reuse this information it is very difficult to read the individual contents and searching the contents form these forms line-by-line and word-by-word. The reason for this difficulty is the font characteristics of the characters in the forms are different to a font of the characters in a computer system. As a result, the computer is unable to recognize the characters while reading them. This concept of storing the contents of forms in computer storage place and then reading and searching the content is called Processing the Forms. Sometimes in this processing of some forms we need to process the information that is related to English language and numbers. For this document processing, we need a language and numbers. For this document processing we need a software system called optical character recognition system.

a) In most of NGO's run programs filed workers visit rural areas to gather a large amount of data about people.

b) To ensure that data is atomic and data of people doesn't get mixed up this data is taken in forms..

c) Now, these forms are filled by rural people who don't have good handwriting and later this data has to be inserted in NGO's database.

d) Traditionally this data entry is done manually by the people this process is really cumbersome and time-consuming.

e) This isn't an accurate method because some words may be misinterpreted by the humans who are entering the data and Hence, This will contribute to Errors in data that is collected.

f) We can automate this Process by using optical character recognition which will somehow make this process more reliable and efficient Moreover it will cut down the cost of extra employees needed for the process of Data entry.

2. PROCESS FLOW

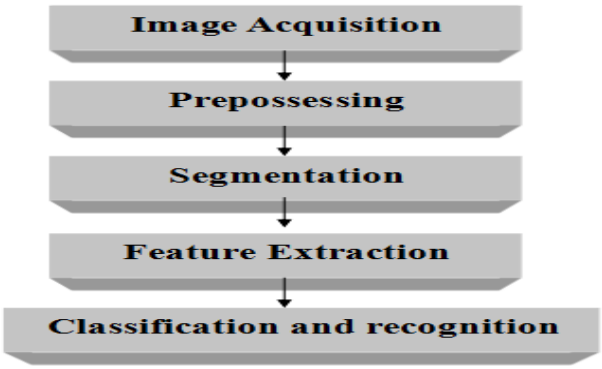

Fig -1: Proposed Flow of Process

3. IMAGE ACQUISITION

We scan the images of the forms using a scanner or any mobile phone camera. After scanning the image is given to pre-processing for the elimination of background noise, and binarization to generate pixels of image in 0s and 1s.

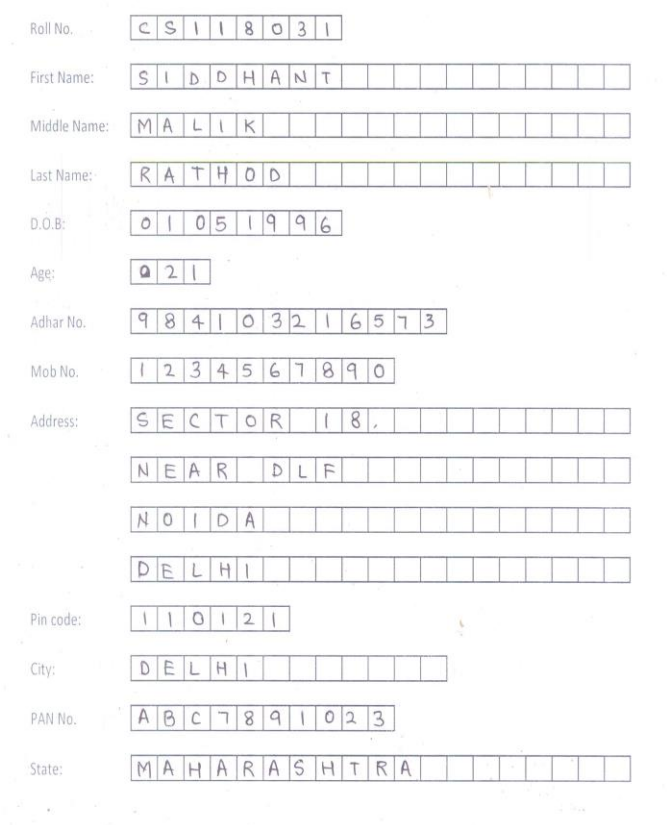

Fig -2: Sample Form

4. PRE-PROCESSING

Preprocessing techniques are needed on color, gray-level or binary document images containing text and/or graphics The steps in pre-processing involve:

Size normalization: Bicubic interpolation is used for the standard sized image.

Binarization: it is the process of converting a grayscale image into a binary image by thresholding. In a binary image, the value of the pixels is either 0 or 1. The background contains white pixels and the foreground contains the black elements. Smoothing: the erosion and dilation smooth the Boundaries of objects.

Opening filter- erosion is followed by dilation. The primary task of an opening

Filter is to remove small object from the foreground and convert them to the background

Closing filter- in the closing filter dilation is followed by an erosion process. It removes small holes in the foreground and changes small regions of background to foreground.

Edge detection: morphological gradient operators are used in edge detection because they enhance the intensity of edges of characters.

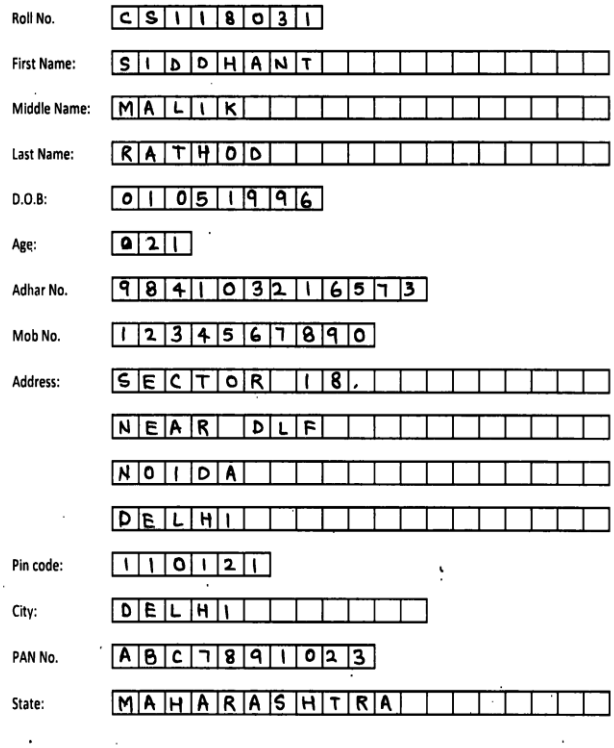

Fig -3: Pre-processed Form

5. SEGMENTATION

The characters are always written in "print fashion", not connected, horizontal histogram profile (for line segmentation), vertical histogram profile (for word segmentation) and connected component analysis is able to handle the character segmentation problem the character

the image is segmented into its sub-components A simple heuristic segmentation algorithm is implemented which scans handwritten words to identify valid segmentation points between characters. The segmentation is based on locating the minima or arcs between letters, common in the handwritten cursive script.

The pre-processed image is divided into lines, words, and characters. An Explanation is given to the below.

Line Segmentation: To separate the text lines, line segmentation is used.

Word Segmentation: - Word segmentation provides the space between words

Character Recognition: It is providing the Spacing between the characters.

We have planned to use neural network and SVM (Support vector machine) to train our network. Basically, the output of both the algorithm gives the approximately same accuracy; still, we will train using both and compare the accuracy.

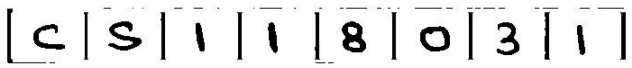

Fig -4: Segmented using contour

6. FEATURE EXTRACTION

Feature Extraction based on Character Geometry: - extracts type of line that forms different character.

A Universe of Discourse-Universe of discourse is defined as the shortest matrix that fits the entire character skeleton.

Zoning-image is divided into equal size of window and feature is done on each window starters-start of window Intersection- have more than one neighbor pixel

Feature Extraction Using Gradient Features: Change in intensity of small neighbor pixel using the Sobel operator.

Gradient vector of each pixel is obtained and gradient image is decomposed in chain code

Crossings and Distances: A popular statistical feature is the number of crossing of a contour by a line segment in a specified direction. The character frame is partitioned into a set of regions in various directions and then features of each region are extracted.

Projections: -Characters can be represented by projecting the pixel gray values onto lines in various directions. This representation creates one-dimensional signal from a two-dimensional image, which can be used to represent the character image

Border Transition Technique (BTT)- In border transition technique it assumes that all the characters are oriented vertically. Each character is divided into four equal quadrants. The scanning and calculation of zero-to-one transition in both vertical and horizontal directions in each division take place.

Graph Matching Method- In a graph matching method it uses a structural feature of the character. It is a useful method to change of font or rotation. In these three features are defined. Here first, an endpoint is connected only one pixel which has information of position. Then a branch point is connected more than three pixels having feature information which is connected the branch.

Fig -5: Extracted each character

7. CLASSIFICATION

We have planned artificial neural network:
A neural network is a powerful data modeling tool that is able to capture and represent complex input/output relationships. The motivation for the development of neural network technology stemmed from the desire to develop an artificial system that could perform "intelligent" tasks similar to those performed by the human brain.

Neural networks resemble the human brain in the following two ways:

1. A neural network acquires knowledge through learning.
2. A neural network's knowledge is stored within inter-neuron connection strengths known as synaptic weights.

To train the network we need to work on each field (name, dob, age etc) of an image individually because the relative position of the fields may change in the form hence we will divide the image into different blocks and try to apply the neural network on each block. So that we can train our network in each field individually. This is the most difficult part of the project. We might face difficulties in dividing image based on a particular field.

Support Vector Machine (SVM):

In machine learning, support vector machines (SVM). Support vector networks are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped

so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

8. CONCLUSION

Thus the Knowledge of Image processing combined along with Machine learning can be used to develop a system which will be useful to identify different types of handwritten text, this implementation can be used to automate process of manual entry , With proper training and large amount of data sets this can be improved further a lot.

ACKNOWLEDGEMENT

We would like to express our gratitude to our Project Guide and teacher Er.Asadullah Shaikh for believing in us that we could do this project and for guiding us to read relevant research papers and previous works.

We would also like to express College Library management because of which we had access to various research papers and had an insight into various concepts because of research papers.

We would like to express our sincere gratitude to the principal of our College Dr. Mohiuddin Ahmed and entire management and the staff of M.H. Saboo Siddik College of Engineering.

We also wish to thank the HOD of Computer Engineering Department Dr. Z.A. Usmani for being helpful.

REFERENCES

[1] Ankit Sharma, Dipti R Chaudhary , "Character Recognition Using Neural Network", International Journal of Engineering Trends and Technology (IJETT) - Volume4Issue4- April 2013

[2] Yuk Yirtg CHUNG, M'an To WONG, "Handwritten Character Recognition by fourier Descriptors And Neural Network", 1997 IEEE TENCON - Speech and Image Technologies for Computing and Telecommunications- conference date-4 Dec 1997

[3] Parag Narendra Achaliya, Sonal Patil, "Handwritten English Character Recognition Based on Artificial Neural Network with Feature Extraction", IJETT ISSN: 2350 – 0808 | September 2014 | Volume 1 | Issue 1 | 124

[4] Anshul Gupta, Manisha Srivastava, Chitralekha Mahanta , "Offline Handwritten Character Recognition Using Neural Network", 2011 International Conference on Computer Applications and Industrial Electronics (ICCAIE 2011)-ISBN 978-1-4577-2059-8/11

[5] Rama Gaur, Dr. V.S. Chouhan, "A Survey on Feature Extraction Techniques for Handwritten Character Recognition", International Journal on Recent and Innovation Trends in Computing and Communication Volume: 5 Issue: 5

[6] Aiquan Yuan, Gang Bai, Po Yang, Yanni Guo, Xinting Zhao, "Handwritten English Word Recognition based on Convolutional Neural Networks", 2012 International Conference on Frontiers in Handwriting Recognition-ISBN 978-0-7695-4774-9/12

[7] Tappert, C. C.; Suen, C. Y.; Wakahara, T. (1990). "The state of the art in online handwriting recognition". IEEE Transactions on Pattern Analysis and Machine Intelligence. 12 (8)

[8] Syed Hassan Tanvir, Tamim Ahmed Khan, Abu Bakar Yamin, "Evaluation of Optical Character Recognition Algorithms and Feature Extraction Techniques", The sixth International conference Innovative Computing Technology (INTECH 2016),-ISBN-978-1-5090-2000-3/16

[9] M. Hanmandlu, K.R.Murali Mohan, Harish Kumar, "Neural based handwritten character recognition" Document Analysis and Recognition, 1999. ICDAR '99. Proceedings of the Fifth International Conference on, 22-22 Sept. 1999, print ISBN- 0-7695-0318-7