Imperial Journal of Interdisciplinary Research (IJIR)
Vol-2, Issue-3 , 2016
ISSN : 2454-1362 , http://www.onlinejournal.in

Recognition, Formatting In Image Files By Using Image Processing

Nalin Bhat[1], Avinash Kumar Yadav[2] , Krushna Rajbinde[3]
Prof. Vajid Khan[4]
[1,2&3]Graduate Student, Department of Information Technology Engineering, Genba Sopanrao Moze College of Engineering, Pune
[4]Assistant Proffesor , Department of Information Technology Engineering, Genba Sopanrao Moze College of Engineering, Pune

***Abstract** : The aim of the project 'OCR' is to develop OCR software for offline handwriting recognition. OCR is an Optical character recognition and is the mechanical or electronic translation of images of handwritten or type written. There are so many techniques available in image processing related to text recognition. Techniques were segmentation, image retrieval, accuracy of blurred images. We can read the text in the picture file. In present system the OCR converts the image file of the text in editable digital format. The main objective of our project is to convert the hard copy of the text image file to editable format with additional features like improved accuracy, multiple language recognition (English, Spanish) detection, auto rectification of errors, and scanning of the particular sized images into the editable image file.*

***Keywords**- Optical Character recognition (OCR)*

1. Introduction

We are going to implement the software which will recognize the characters from offline document (in image format).Here we are developing OCR which will recognize machine printed written English, Spanish characters. OCR is an Optical character recognition and is the mechanical or electronic translation of images of typewritten text (usually captured by a scanner or camera) into machine-editable text.Optical Character Recognition (OCR) is a process that translates images of typewritten scanned text into machine-editable text, or pictures of characters into a standard encoding scheme representing them in ASCII or Unicode.

An OCR system enables us to feed a book or a magazine article directly into an electronic computer file, and edit the file using a word processor. Though academic research in the field continues, the focus on OCR has shifted to implementation of proven techniques.

Optical character recognition is currently done by using various optical techniques and digital character recognition (using scanners and computer algorithms).

Very few applications survive that use true optical techniques, the OCR term has now been broadened to include digital image processing as well. OCR systems usually require training (the provision of known samples of each character) to read a specific font. However, this approach is sensitive to the size of the fonts and the font type. Soft computing has been adopted into the process of character recognition for its ability to create input output mapping with good approximation.

2. Literature Survey

A. Adaptive document image binarization. [1]:

"A new method is presented for adaptive document image binarization, where the page is considered as a collection of sub components such as text, background and picture. The problems caused by noise, illumination and many source Type related degradations are addressed. Two new algorithms are applied to determine a local threshold for each pixel.

The performance evaluation of the algorithm utilizes test images with ground-truth evaluation metrics for binarization of textual and synthetic images, and a weight-based ranking procedure for the proposed algorithms were tested with images including different types of document components and degradations. The results were compared with a number of known techniques in the literature.

B. Text Detection and character recognition in scene images with unsupervised learning [2]:

Optical Character Recognition deals with the problem of recognizing optically processed characters. Optical recognition is performed off-line after the writing or printing has been completed, as opposed to on-line recognition where the computer recognizes the characters as they are drawn. Both hand printed and printed Characters may be recognized, but the performance is directly dependent upon the quality of the input documents [2].

C. Combining focus measure operator to predict OCR accuracy [3]:

Mobile document image acquisition is a new trend raising serious issues in business document processing workflows. Such digitization procedure is unreliable, and integrates many distortions which must be detected as soon as possible, on the mobile, to avoid paying data transmission fees, and losing information due to the inability to re-capture later a document with temporary availability. In this context, out-of-focus blur is a major issue: users have no direct control over it, and it seriously Degrades OCR recognition.

In this paper, we concentrate on the estimation of focus quality, to ensure a sufficient eligibility of a document image for OCR processing. We propose two contributions to improve OCR accuracy prediction for mobile captured document images. First, we present 24 focus measures, never tested on document images, which are fast to compute and require no training. Second, we show that a combination of those measures enables state-of-the art performance regarding.

D. Text recognition from Images [4]:

Text recognition in images is a research area which attempts to develop a computer system with the ability to automatically read the text from images. These days there is a huge demand in storing the information available in paper documents format in to a computer storage disk and then later reusing this information by searching process.

One simple way to store information from these paper documents in to computer system is to first scan the documents 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 documents line-by-line and word-by-word. The challenges involved in this the font characteristics of the characters in paper documents and quality of images. Due to these challenges, computer is unable to recognize the characters while reading them. Thus there is a need of character recognition mechanisms to perform Document Image Analysis (DIA) which transforms documents in paper format to electronic format. In this paper we have discussed method for text recognition from images.

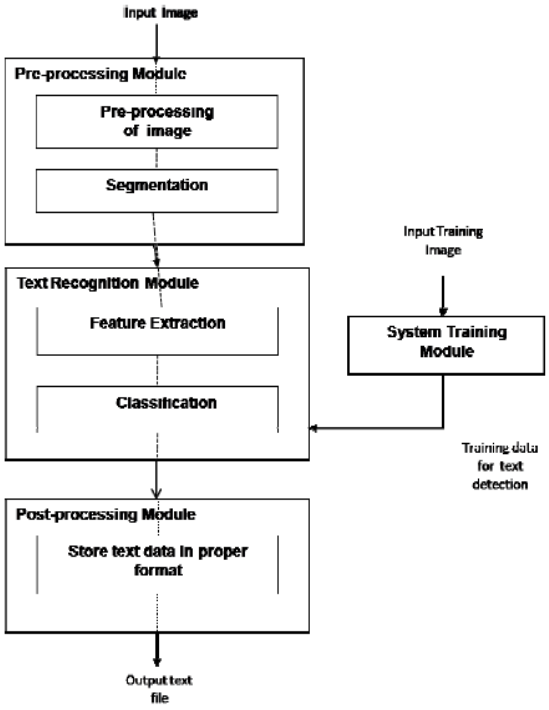

Fig.1 Architecture of text recognition [4].

E. System Training Module [4]:

This module can be used to train the system for text recognition. Before converting the printed documents in to editable and searchable documents, the first and the mandatory step is providing training to the system. Here training in the sense the font followed in the scanned document should be identified by the user. Then the user types all the characters that are required for recognition from the scanned document as an image file. This image file should be provided as an input during the training process.

F. Feature Extraction [4]:

Feature extraction is the process to retrieve the most important data from the raw data. The most important data means that's on the basis of that's the characters can be represented accurately. To store the different features of a character, the different classes

are made. There are many technique used for feature extraction like Principle
Component Analysis (PCA), Linear Discriminate Analysis (LDA), Independent Component Analysis (ICA), Chain Code (CC), zoning, Gradient Based features, Histogram etc.

In this we use matrix feature extraction method. In this
method first we convert the image to binary matrix i.e. black and white image convert to matrix form, text image is converted in to the matrix of 0's and 1's.from this matrix data we was extract text character line by line and word by word by using above segmentation method. After that segmented characters data are normalized and store in fixed dimension [4].

3. Proposed System

A software which will recognize the characters from offline document (in image format).Here we are developing OCR which will recognize machine printed written English, Spanish characters. OCR is an Optical character recognition and is the mechanical or electronic translation of images of typewritten text (usually captured by a scanner) into machine-editable text.

I. Matching phase:

In the matching phase preprocessing of the document is done after taking the image from the scanner or the any other image capturing source with (jpeg,png) as the supporting formats and the image file of a document is segmented and recognized using the neural network

i. Camera Capture Module

With the help of this module various images of the documents are captured in the formats like (.png,.jpeg)

ii. Segmentation

The segmentation is the most important process in text recognition. Segmentation is done to make the separation between the individual characters of an image. Segmentation is one of the most important phases in this project. The performance of this project is depending on segmentation. Segmentation subdivides an image into its constituent regions or objects. Basically in segmentation, we try to extract basic constituent of the script, which are certainly characters. In row segmentation each and every line is separated one by one and in the column segmentation the each and every word is separated.

Algorithm for Line detection from image [4].

Step 1: Start scanning the image horizontally from the topmost left corner row by row.
Step 2: If any black pixel is encountered in a row make the row status as '0'.
Step 3: If no black pixel in encountered in a row while tracing it then mark the row status as '1'.
Step 4: By counting and following the total numbers of
continuous '0' from row status vector number position of lines can be obtained.

Algorithm for Character detection from image [4].

Step 1: Take a single line under consideration.
Step 2: Start scanning the image vertically from the topmost left corner column by column.
Step3: If any black pixel is encountered in a column mark the column status to '0'.
Step 4: If no black pixel in encountered in a column while tracing it then mark the column status as '1'.
Step 5: By counting and following the total numbers of
continuous '0' from column status vector number position of lines can be obtained

iii. Character Recognition using the neural network [5].

Neural networks are typically organized in layers. Layers are made up of a number of interconnected 'nodes' which contain an 'activation function'. Patterns are presented to the network via the 'input layer', which communicates to one or more 'hidden layers' where the actual processing is done via a system of weighted 'connections.

II. Training phase

i. Folder of 62 characters comprising 52 alphabets including upper case and lower case and 10 digits from 0 to 9 is created.

ii. A character is taken and resized to 40 X 40 and boundary lines are made for each character to be trained and the character is cropped.

iii. In down sampling phase

A matrix A is taken
Double [A]=new double (Dw_width*Dw_height)
SampleW=Cw/Dw_width
SampleH=Ch/Dw_height
for (i=0 to SampleW)
for (j=0 to SampleH)
if (pix(i,j) == black)
A[i] = black

IV. Repeat this process from **i** to **iii** for all 62 characters.

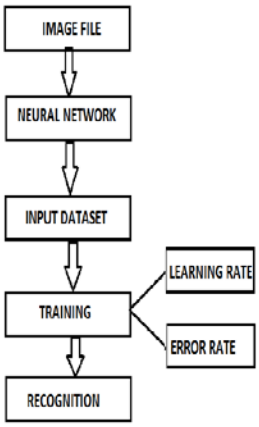

Fig.2 work flow of the OCR system

6. Conclusion

This system can be used by multiple users. We can do this by improving our software for recognizing the printed text document. Also if we can take the stroke information and give it to our system, then it will be possible to recognize even cursive script also. Complete conversion of A4 image document (hard copy) into digital word i.e. Auto detection and rectification of errors.

7. Acknowledgement

Our thanks to our college G.S.M.C.O.E and my department of Information Technology engineering which has provided the support and equipment which we have needed to complete our work. I extend my heartfelt gratitude to my guide, Prof. Vajid Khan Coordinator, Prof. Priyanka More who has supported us throughout our research with their patience and knowledge.

8. References

[1] J. Sauvola, M. PietikaKinen, “Adaptive document image binarization” University of Oulu,1999.

[2] Adam Coates, Blake Carpenter, Carl Case, Sanjeev Satheesh “Text Detection and Character Recognition in Scene Images with Unsupervised Feature Learning”, Stanford University,2010.

[3] Marçal Rusiñol, Joseph Chazalon “Combining Focus Measure Operators to Predict OCR Accuracy in Mobile-Captured Document Images” Université de La Rochelle,2014.

[4] Pratik Madhukar Manwatkar, Shashank H. Yadav “Text Recognition from Images”, YCCE, Nagpur 2015.

[5] https://en.wikipedia.org/wiki/Artificial_neural_network