International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 03 | Mar-2018 www.irjet.net p-ISSN: 2395-0072

Optical Character Recognition Using Image Processing

Shyam G. Dafe[1], Shubham S. Chavhan[2]

[1,2]*Department of Electronics & Telecommunication Engineering, Prof. Ram Meghe College of Engineering & Management, Bandera Road, Amravati (444602), Maharashtra, India*

Abstract - *Optical character recognition is the mechanical or electronic conversion of typed, handwritten or printed text into machine encoded text, whether from a scanned document, a photo of document. At the present time, keyboarding remains the most common way of inputting data into computers. This is probably the most time consuming and labor intensive operation. OCR is the machine replication of human reading and has been the subject of intensive research for more than three decades. OCR can be described as Mechanical or electronic conversion of scanned images where images can be handwritten, typewritten or printed text. It is a method of digitizing printed texts so that they can be electronically searched and used in machine processes. It converts the images into machine-encoded text that can be used in machine translation, text-to-speech and text mining. a simple, efficient, and less costly approach to construct OCR for reading any document that has fix font size and style or handwritten style. It is widely used as a form of information entry from printed paper data records, whether passport document ,invoices, bank statements, computerized receipts, business cards, mail ,printouts of static data or any suitable documentation.*

Key Words: **OCR, Speech Synthesis, Speech Conversion, Character Extraction Algorithm, Matlab, Text To Speech**.

1. INTRODUCTION

Machine replication of human functions, like reading, is an ancient dream. However over .The last five decades, machine reading has grown from a dream to reality. Optical character Recognition has become one of the most successful applications of technology in the field of pattern recognition and artificial intelligence. Many commercial systems for performing OCR exist for a variety of applications, although the machines are still not able to Complete with human reading capabilities. Optical character recognition is needed when the information should be readable both too human and to a machine. Both hand written and printed character may be recognized. Optical character recognition is performed off line after the writing or printing has been completed ,as opposed to on line recognition where the computer recognizes the character as they are draawn.OCR method for recognizing documents either printed or handwritten without any knowledge of the font. we discuss different technologies for automatic Identification and establish OCR's position among these techniques. The next chapter gives a brief overview of the historical background and development of character recognition. We also present the different steps, from a methodical point of view, which have been employed in OCR. An account of the wide area of applications for OCR is given in next chapter and in the final chapter we discuss the future of OCR.

1.1 Components of an OCR system

A typical OCR system consists of several components. a common setup is illustrated. The first step in the process is to digitize the analog document using an optical scanner. When the regions containing text are located, each symbol is extracted through a segmentation process. The extracted symbols may then be preprocessed, eliminating noise, to facilitate the extraction of features in the next step. The identity of each symbol is found by comparing the extracted features with descriptions of the symbol classes obtained through a previous learning phase. Finally contextual information is used to reconstruct the words and numbers of the original text.

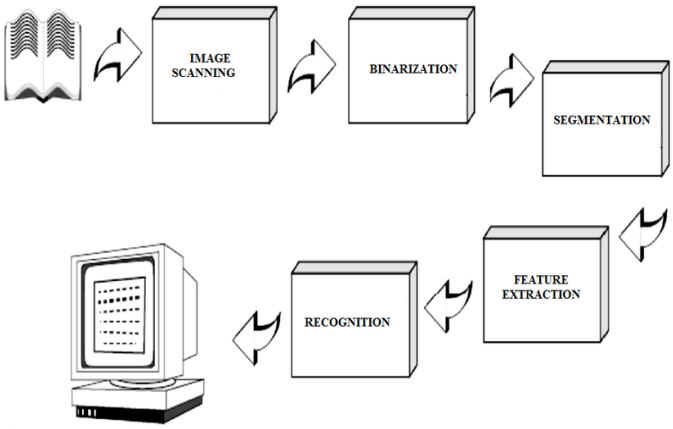

Fig -1: Components of an OCR-system

1.2 Optical scanning

Through the scanning process a digital image of the original document is captured. In OCR optical scanners are used, which generally consist of a transport mechanism plus a sensing device that converts light intensity into gray-levels. Printed documents usually consist of black print on a white background. Hence, when performing OCR, it is common practice to convert the multilevel image into a bi-level image of black and white. Often this process, known as thresholding , is performed on the scanner to save memory space and computational effort. The thresholding process is important as the results of the following recognition is totally dependent of the quality of the level image. Still, the

thresholding performed on the scanner is usually very simple. A fixed threshold is used, where gray-levels below this threshold is said to be black and levels above are said to be white. For a high-contrast document with uniform background, a pre chosen fixed threshold can be sufficient. However, a lot of documents encountered in practice have a rather large range in contrast. In these cases more sophisticated methods for thresholding are required to obtain a good result.

2. Working of OCR

The image resulting from the scanning process may contain a certain amount of noise. Depending on the resolution on the scanner and the success of the applied technique for thresholding, the characters may be smeared or broken. Some of these defects, which may later cause poor recognition rates, can be eliminated by using a preprocessor to smooth the digitized characters. The smoothing implies both filling and thinning. Filling eliminates small breaks, gaps and holes in the digitized characters, while thinning reduces the width of the line.

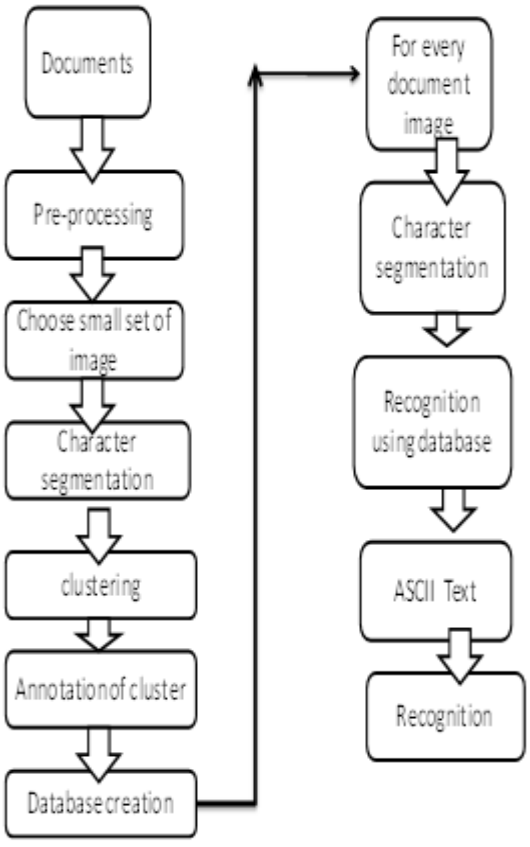

Fig -2 : Flow chart of OCR

Step 1 : Documents

Take the documents in the from of photo of documents or scans the documents.

Step2 : Pre-processing

In the preprocessing Binarization is the starting step of most document image analysis systems and refers to the conversion of the gray-scale image to a binary image. Since historical document collections are most of the times of very low quality, an image enhancement stage is also essential.

Step3 : Segmentation

After document binarization a top-down segmentation approach is applied. First lines of the documents are detected, then words are extracted and finally words are segmented in characters.

Step4 : Text Line Detection

The text line segmentation methodology is based on and consists of three distinct steps. The first step includes, connected component extraction and average character height estimation of the binary image. In the second step, a block - based Hough transform is used for the detection of potential text lines while a third step is used to correct possible splitting, to detect text lines that the previous step did not reveal and, finally, to separate vertically connected characters and assign them to text lines.

Step5: Word and Character Segmentation

Once the text lines have been located, projection profiles based on in order to detect the words are used. Then the following method is adopted in order to separate them into letters. The algorithm is based on the segmentation algorithm described for touching numerals in. The basic idea is that we can find possible segmentation paths linking the feature points on the skeleton of the word and its background.

Step6: Clustering

In our approach the well-known k-Means clustering algorithm is used. An advantage of this algorithm is its computational simplicity. Also, as with all algorithms that use point representatives, k-Means is suitable for recovering compact clusters. Illustrates the k-Means clustering algorithm.

Step7: Annotating Clusters

After feature extraction and clustering algorithm described in Sections of in take place all Characters are grouped into k clusters. Let each cluster represented as $i\ C$ where $i = 1, 2, 3 \ldots k$. The goal here is to turn these clusters into classes. Each class, $'i\ C$ where $i = 1, 2, 3 \ldots l$, has then to be assigned to an ASCII label. Unfortunately, this task cannot be automatic, so a tool has been developed that requires the user's interference. The user is provided with an appropriate tool

to handle clustering errors. Figure 8 is a screenshot of the developed tool. On the left side clusters created by the system are displayed to the user. Since no cluster has an ASCII code, initially all clusters are labeled as '?'. On the right side all the character instances are shown.

Step8: Character Database

At this step, we choose a representative set of images for training and from these images characters are extracted following the segmentation procedure described in Section 3 (.Since class labeling of these characters is not available, no OCR methodology based on supervised learning that requires a training set of labeled patterns can be applied. Thus, our concern is first to "reveal" the organization of characters into "sensible" clusters (groups). Then, these clusters, after performing all required procedures to correct possible errors are labelled and, finally, the character database is created

Step9: Recognition

At this stage every document image that has not participated in the training phase is ready to be converted into a text file. Characters are extracted following the approach described in each character is represented as a feature vector according to and then all characters are classified using the database created in Section character is represented as a feature vector according to cluster and then all characters are classified using the database created in cluster. For this classification problem the Support Vector (SVM) algorithm was used .

3. CONCLUSIONS

Today optical character recognition is most successful for constrained material, that is documents produced under some control. However, in the future it seems that the need for constrained OCR will be decreasing. The reason for this is that control of the production process usually means that the document is produced from material already stored on a computer. Hence, if a computer readable version is already available, this means that data may be exchanged electronically or printed in a more computer readable form, for instance barcodes. The applications for future OCR-systems lie in the recognition of documents where control over the production process is impossible. This may be material where the recipient is cut off from an electronic version and has no control of the production process or older material which at production time could not be generated electronically. This means that future OCR-systems intended for reading printed text must be omnifont. Another important area for OCR is the recognition of manually produced documents.

ACKNOWLEDGEMENT

It's a pleasure and a great blessing of GOD for working on the project named "Optical Character Recognition Using Image Proccesing". Wherein we gained knowledge by working under the able leadership of our Project guide Mr. A.G.Kalbande who helped and supported us in every sphere of our project. who well supported us and provided us with his precious time and support for his valuable advice and help which he provided us throughout the whole duration of this project. Besides that we thank whole of the **EXTC** department for their appreciation and kind support.

REFERENCES

[1] R. G. Casey & K. Y. Wong.Document-Analysis Systems and Techniques.Image Analysisi Applications, eds: R. Kasturi& M. Tivedi, p. 1-36.

[2] Portable Camera-based Assistive Text and Product Label Reading from Hand-held Objects for Blind Persons By, Ying Li Tian .

[3] J-P. Caillot.Review of OCR Techniques. NR-note, BILD/08/087.

[4] R. Bradford & T. Nartker.Error Correlation in Contemporary OCR Systems.Proceedings ICDAR-91, Vol. 2, p. 516-524, 1991.

BIOGRAPHIES:

Shyam G. Dafe pursuing Final year in the year 2018. He is in Prof. Ram Meghe College of Engineering & Management, Bandera Road, Amravati in the department of Electronics & Telecommunication Engineering.

Shubham S. Chavhan pursuing Final year in the year 2018. He is in Prof. Ram Meghe College of Engineering & Management, Bandera Road, Amravati in the department of Electronics & Telecommunication Engineering.