International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Segmentation of Touching Characters in Indian Scripts

B. Hari Kumar[1], N. Sateesh[2]

[1]Assistant Professor, Department of Electronics and Communications Engineering, Wellfare Institute of science Technology & Management Pinagadi Village, Near Pendurthi, Andhra Pradesh, India

[2]Assistant Professor, Department of Electronics and Communications Engineering, Wellfare Institute of science Technology & Management Pinagadi Village, Near Pendurthi, Andhra Pradesh, India

Abstract: ***In a multilingual country like India, a document may contain text words in more than one language. For a multilingual environment, multilingual Optical Character Recognition (OCR) system is needed to read the multilingual documents. So, it is segmentation different language. The objective of this project is to propose visual clues based procedure to segmentation Telugu, Hindi and English text portions by line wise, word wise, character wise of the Indian multilingual document. Segmentation is an important topic in script identification and image processing. Previously to identification script by line wise and word wise segmentation poses now am implementation character wise segmentation it is easy way to identification language by character3 wise segmentation. The objective of segmentation is to segmentation by image line wise, word wise, charter wise. The segmentation is very important role to identification of languages. The world we live in, is getting increasingly interconnected, electronic libraries have become more pervasive and at the same time increasingly automated including the task of presenting a text any language as automatically translated text in any other language. Identification of the language in a document image is of primary importance for selection of a specific OCR system processing multi lingual documents.***

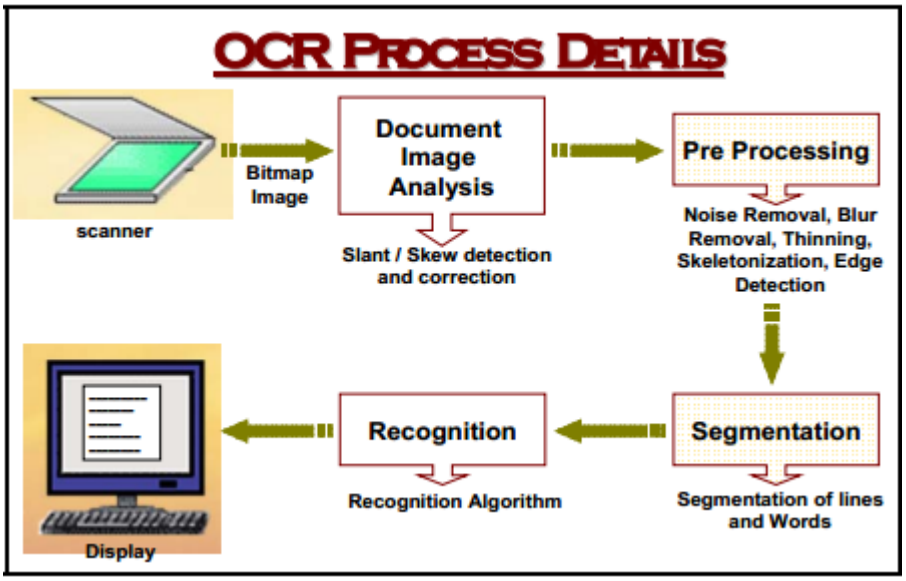
Keywords: Optical Character Recognition tool, Mat lab code, Image processing etc…

1. Introduction

Segmentation of a document images is a very important task for optical character recognition. A lot of research work has been done for character segmentation. In any OCR system segmentation phase is very important step to improve accuracy of printed Indian languages or hand character recognition in OCR heavily depends upon segmentation phase. The term segmentation means subdivides an image into a particular part (like text or graph separation) its constituent region or object. Basically in segmentation techniques that minces we are try to extract a specific part (text line and graph) of the document images [1]. In segmentation we include line based segmentation, word based segmentation and character segmentation. Firstly we are taken line segmentation for any printed document images, because the line segmentation is performed to find number of line in any scanned printed document images and boundaries of each line in any input document images. After completion of line segmentation we apply word based segmentation to perform word wise separation of any scanned printed or handwritten document image[2]. After completion of line and word based segmentation we are processing character segmentation to find the character in any scanned printed Indian languages document or hand written document images.

1.1 Optical Character Recognition (OCR)

What's OCR?
Optical Character Recognition, usually abbreviated to OCR, is the mechanical or electronic translation of scanned images of handwritten, typewritten or printed text into machine-encoded text. The conversion of paper files into electronic document is done by a process called scanning [1]. Not only English but also other foreign languages can be recognized by OCR which makes it universal in application. The two main systems used to perform OCR are "matrix matching" and "feature extraction." Matrix matching is the simpler and the more common, as well as the more limited, of two [2]. All OCR systems include an optical scanner for reading text, and sophisticated software for analyzing images. Most OCR systems use a combination of hardware (specialized circuit boards) and software to recognize characters, although some inexpensive systems do it entirely through software [3]. Advanced OCR systems can read text in large variety of fonts.



Steps involved in OCR processor

Scanning
A flat-bed scanner is usually used at 300dpi which converts the printed material on the page being scanned into a bitmap image.

Document Image Analysis
The bitmap image of the text is analyzed for the presence of skew or slant and consequently these are removed [1]. Quite a lot of printed literature has combinations of text and tables, graphs and other forms of illustrations. It is therefore important that the text area is identified separately from the other images and could be localized and extracted.

Pre-processing
In this phase several processes are applied to the text image like noise and blur removal, binarization, thinning, edge detection and some morphological processes, so as to get an OCR ready image of the text region which is free from noise and blur.

Segmentation
If the whole image consists of text only, the image is first segmented into separate lines of text. These lines are then segmented into words and finally words into individual letters. Once the individual letters are identified, localized and segmented out in a text image it becomes a matter of choice of recognition algorithm to get the text in the image into a text processor.

Recognition
This is the most vital phase in which recognition algorithm is applied to the images present in the text image segmented at the character level. As a result of recognition character code corresponding to its image is returned by the system which is then passed to a word processor to be displayed on the screen where it can be edited, modified and saved in a new file format.

1.2 Applications of OCR

Practical Applications
In recent years, OCR (Optical Character Recognition) technology has been applied throughout the entire spectrum of industries, revolutionizing the document management process [1]. OCR has enabled scanned documents to become more than just image files, turning into fully searchable documents with text content that is recognized by computers. With the help of OCR, people no longer need to manually retype important documents when entering them into electronic databases [2]. Instead, OCR extracts relevant information and enters it automatically. The result is accurate, efficient information processing in less time.

1.2.1 Banking
The uses of OCR vary across different fields. One widely known application is in banking, where OCR is used to process checks without human involvement. A check can be inserted into a machine, the writing on it is scanned instantly, and the correct amount of money is transferred [3]. This technology has nearly been perfected for printed checks, and is fairly accurate for handwritten checks as well, though it occasionally requires manual confirmation. Overall, this reduces wait times in many banks.

1.2.2 Legal
In the legal industry, there has also been a significant movement to digitize paper documents. In order to save space and eliminate the need to sift through boxes of paper files, documents are being scanned and entered into computer databases [1]. OCR further simplifies the process by making documents text-searchable, so that they are easier to locate and work with once in the database. Legal professionals now have fast, easy access to a huge library of documents in electronic format, which they can find simply by typing in a few keywords.

1.2.3 OCR in Other Industries
OCR is widely used in many other fields, including education, finance, and government agencies. OCR has made countless texts available online, saving money for students

and allowing knowledge to be shared [2]. Invoice imaging applications are used in many businesses to keep track of financial records and prevent a backlog of payments from piling up [4]. In government agencies and independent organizations, OCR simplifies data collection and analysis, among other processes. As the technology continues to develop, more and more applications are found for OCR technology, including increased use of handwriting recognition. Furthermore, other technologies related to OCR, such as barcode recognition, are used daily in retail and other industries.

1.3 Benefits of OCR

1.3.1 No More Retyping

If you lose or accidentally erase an important digital file, such as a proposal or invoice, but still have a hard copy, you can easily replace it in your digital filing system by using OCR software to scan the paper original or most recent draft.

1.3.2 Quick Digital Searches

OCR software converts scanned text into a word processing file, giving you the opportunity to search for specific documents using a keyword or phrase. For example, you could effortlessly search hundreds of invoices and locate a specific name or account in moments, without having to thumb through extensive files.

1.3.3 Save Space

Free up storage space by scanning paper documents and hauling the originals off to storage. You can easily turn a filing cabinet worth of information into editable digital files, and create a backup system consisting of a single CD.

1.3.4 Edit Text

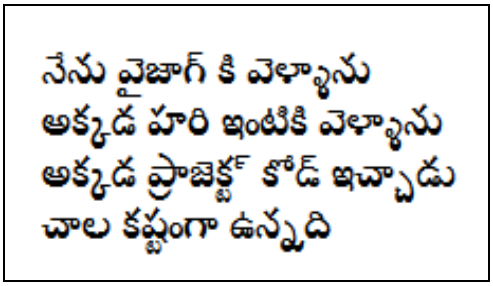
Once you've scanned your document using OCR, you have the option to edit the text within a word processing program of your choice. Scan items that may need to be updated in the future to help expedite the editing process:

- Typed family recipes
- Rental agreements
- Resumes
- Contracts

2. Segmentation Process

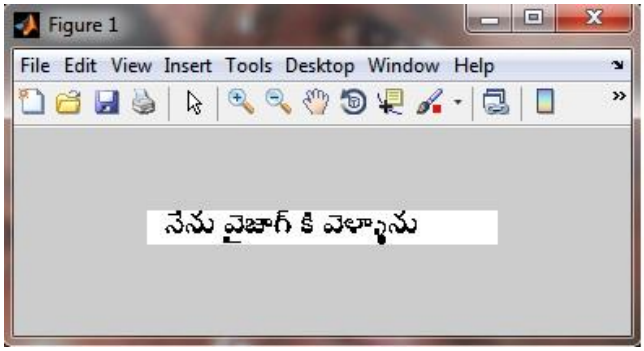
2.1 Line Segmentation

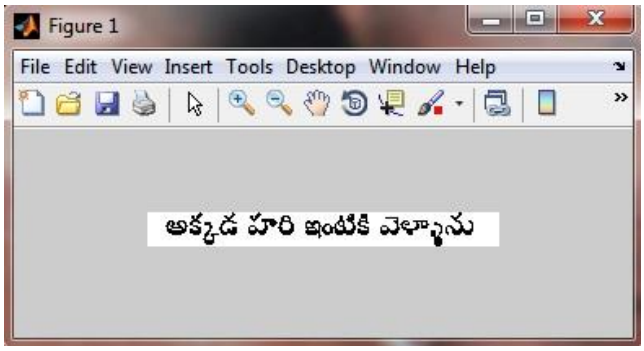
Second step of segmentation process is segmenting the text region into lines, also called as line segmentation. Generally, each text line is separated from the previous and following lines by white spaces. Therefore, the horizontal projection of a document image is the most commonly used technique to extract the lines from the document. If the lines are well separated and not tilted, the horizontal projection will have well separated peaks and valleys. These valleys can be detected easily and used to determine the locations of the line boundaries .shown in below fig2.1

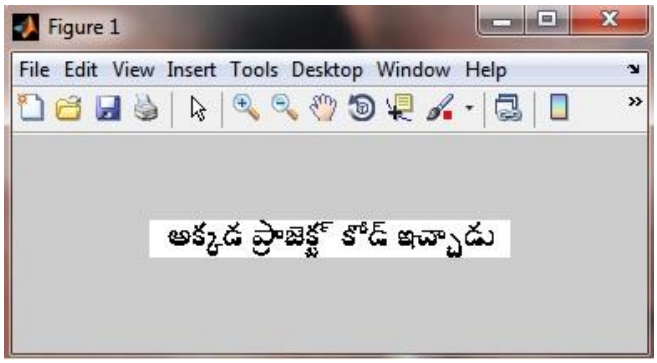


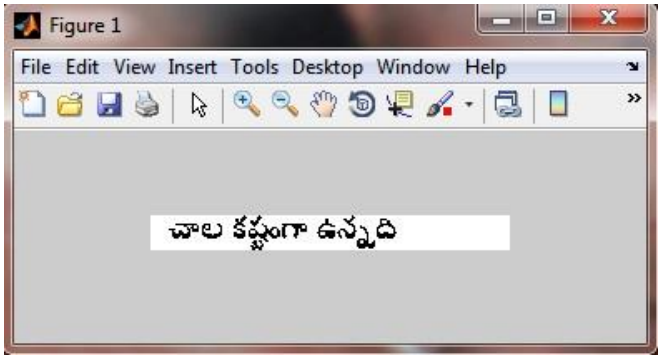
2.1 Input Image

Output Images Of Line Wise Segmentation:



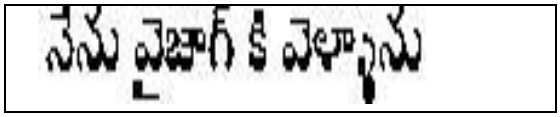






2.2 Word Segmentation

From the extracted text lines, words get separated. Usually, applying vertical projection profile (VPP) and detecting some specific threshold exceeding horizontal gaps, words are separated from a text line. An example is shown in Fig.2.2

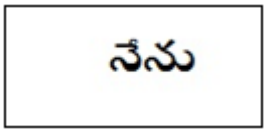


2.2 Input Image

Output Images of Word Wise Segmentation

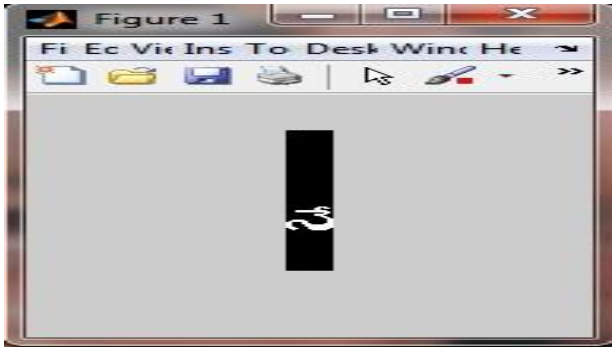
2.3 Character Segmentation

Segmentation of characters from the isolated words is the most challenging part of the script segmentation phase. Since, in computer composed scripts some characters in a container word may partially overlap with one another, it becomes very difficult to isolate those characters properly. Especially the modifiers (both vowels and consonants) most of the time coincide with the modifying characters as shown in Figure 2.3. These kinds of nontrivial combinations of characters make the whole process of character segmentation extremely challenging. Besides, some symbols, like Chandra-Bindu, often come between two consecutive characters in a word; then isolating those becomes a tough job. An example is shown in Figure 2.3.



2.3 Input Image

Output Images Of Character Wise Segmentation:



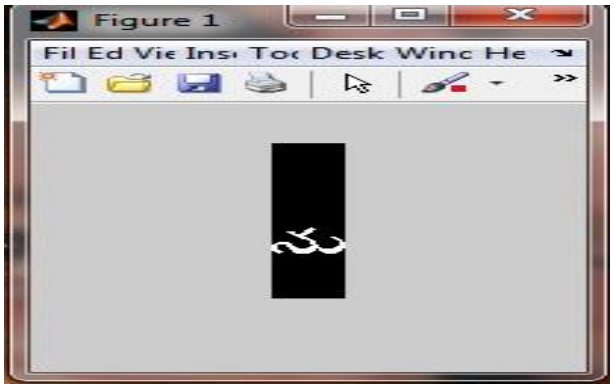


Table 1: Segmentation Table

Sr.No	Line wise	Word wise	Charter
1 (Big text image)	70%	80%	90%
2 (Small text image)	60%	70%	80%

3. Conclusions

In this project we have designed an OCR for the recognition of Latin Printed Document Images. We have conducted experiments to evaluate its performance in which we have got good results on reasonable diverse quality documents. However the performance of the OCR varies with the diversity of the font size and style. For Indian script documented image with sufficient large font the segmentation accuracy is more than 90% and for small font size, the segmentation accuracy declines and will be in the range of 80% to 9

References

[1] S. Mori, C. Y. Suen and K. Yamamoto, "Historical review of OCR research and development", *Proceedings of the IEEE*, Vol. 80(7), pp. 1029-1058, 1992.
[2] U. Pal and B. B. Chaudhuri, "Indian script character recognition: a survey", *Pattern Recognition*, Vol. 37(9), pp. 1887-1899, 2004
[3] R.G.Casey and E. Lecolinet, "A survey of methods and strategies inCharacter segmentation", *IEEE Transactions on PAMI*, Vol. 18(7), pp. 690-706, 1996
[4] C. E. Dunn and P. S. P. Wang, "Character segmentation techniques for handwritten text - a survey", in the Proceedings of 11th ICPR, Vol. 2, pp. 577-580, 1992

References

B. Harikumar completed his M.Tech in Electronics and Communications Engineering in Aurora's Scientific Tech & Research Academy, Hyderabad from 2014. Presently he is working Assistant professor (ECE) Wellfare Institute of science Technology & Management Pinagadi Village, Near Pendurthi, Andhra Pradesh He has 3 years Exp Teaching. His area of interests is ImageProcessing, Optical Character Recognition

N. Sateesh completed his M.tech in "Digital systems signal processing" from Gitam University,Vizag. Presently he is working Assistant professor (ECE) Wellfare Institute of science Technology & Management Pinagadi Village, Near Pendurthi, Andhra Pradesh He has 2 years Exp Teaching. His area of interests is Digital systems& VLSI