International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-3, February 2020

International, National and Local Languages OCR Segmentation of Running Hand Scripts

B.Hari Kumar, P.Chitra

Abstract*: Script Segmentation of running hand scripts is a complicated job because of different hand writing styles and complex formational quality. Segmentation of moving hand script in Indian language is a difficult assignment .The incident of a title, crossway nature in the mid section & half nature makes the segmentation process is more difficult.

Sometimes, the import space and noises make line shatter a hard task. Without disconnecting the nudge characters, it will be complicated to recognize the character; hence shatter is needed for the moving texts in a word. So, the technique, according to that first step will be pre-processing of a term, then we can identify the joint points, form the bounding boxes around all perpendicular & parallel lines of the script , finally splint the nudge nature based on their height and width. For non-touching printed scripts and running hand scripts, it gives us 95% and 90%result respectively.

This technique fails for some hand written characters because of obligation that means cause will not match always to a lot of variation in the different handwriting scripts i.e., character like occurs problem as space in the middle of parts in scripts that can be resolve by identification system because of this system considers initial part as non-bar character and cut after that, as the future exertion is a concern, in this go these different characters as input to the character identification system. In this method have resolute control of a word which will once more helpful for the identification system to perceive bottom modifiers and higher modifiers. This running hand scripts technique can be applicable to other Indic languages hand written scripts.

Keywords: OCR, Preprocessing, Segmentation, Mat lab code, Image processing;

I. INTRODUCTION

Segmentation of text images is a very essential course for OCR. A hug research work had done for OCR segmentation, for every OCR process segmentation part is a major step to modify the appropriate and accuracy of a written manual. Character recognition in OCR hugely depends upon the segmentation part the phrase segmentation means a subdivision of a virtual picture into a specific part insularly in segmentation methodology that means to identify the abstract of a specific part of a document virtual image [1]. In segmentation, it interludes the line, word, character-based segmentation. Basically the segmentation of any handwritten manual documentation pictures as the lines of segmentation is dramatized to find a number of lines of any form handwritten text images and the barriers of each and every line in any input document pictures [2]. After analyzing the segmentation, generally, it applies word-based segmentation to perform verbatim wise separation of any handwritten document images at lost all the word-based segmentation it is processing by character segmentation to analyze the character in any scanned, printed Indian vernacular languages manuals and images. In order to address the resolutions, it formulates the word segmentation, resolution as in an encoding quadric assignment resolution that are Considered as pair wise ratifications between the Edges as well as hoods of individual gaps [3].

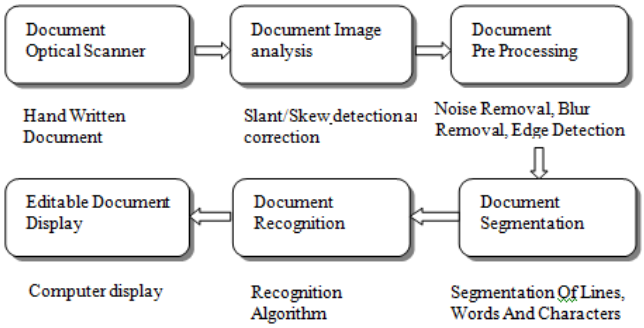
Revised Manuscript Received on February 05, 2020.

* Correspondence Author

B.Hari Kumar*, professor (ECE) Wellfare Institute of science Technology & Management Pinagadi Village, Near Pendurthi, Andhra Pradesh

P.Chitra, Professor Department of ECE, Sathyabama Institute of Science and Technology

II. BLOCK DIAGRAM OF OCR



A. Submission

OCR, usually known as Optical Character Recognition, is used to convert scanned documents (.jpg .png) into editable text in (word pad, notepad) document. The changing of hard copy documents into the computerized script is done by a technique called scanning [1]. It does not only implement international language (English) it can be applied to other Indian languages such as national language (Hindi) and local language (Telugu) can be recognized by OCR. Optical Character Recognition is used to perform two main systems are "matrix match up" and "feature exacting" Matrix matchup is the easy and the more usual, as well as the more finite, of two pictures [2]. All OCR formation includes an optical scanner for perusal script and having software to look over images. The OCR Embedded systems use to accept characters, while some low price systems do it totally across software [3]. Approach Optical Character Recognition systems can read the message in a vast diversity of scripts.

III. OCR PROCESSOR STEPS

Image Scanning

The first step of OCR is Scanner a light scanner is generally used to changing of hard copy documents into softcopy such as a (300 dpi) documents [1].

Image inspection

The second step of OCR is image analysis it is used to detect if is there any skew detection and correction of the scripts [1]. Totally plenty of handwritten documents has been taking here and analyzed those soft copies if is there any slant or skew needed that can be modified here. It is crucial that the text area is recognized separately from the other scripts and could be enclosed and takes out [2].

Image Pre-processing

The third step of OCR is image pre-processing it is used in several processes such as remove the noise of the images, blur removal (blur images can be changed into clear images), binate, scattering, boundary detection and some morphological processes [3], so Optical Character Recognition is getting a prepared image of the script region which is free from noise and blur.

Image Segmentation

The next step of OCR is image segmentation it is used to perform the images into a segmented image which means that the image can be divided into parts If the entire image is a text image, the text image is first segmented into different lines of text this is called line segmentation. Next these lines are then breaking into words it is called word segmentation and finally words into single character it is called character segmentation. The segmentation task is a very important task for image recognition [4]. The segmentation was done properly then the image recognition part is easy.

Image Recognition

The last and most important step of OCR is Image Recognition it is useful to recognize the image. There is line-wise script recognition, word-wise script recognition, and character-wise script recognition. This is the most important stage in which the identification algorithm is applied to the images present in the text image beaked at the character extent. As a result of identification character code compatible to its image is returned in note pad or word pad on the computer screen. That image can be changed into an editable format that can be we can modify also and saved in a different file format [5].

IV. OCR USES

OCR is used to scan file documents or images and convert them into editable text documents.

The OCR system is used for the following purposes:

- Data Entry (E.g. Check, passport, receipt)
- Legal Billing (E.g. Any government document)
- Save space (E.g. Free up storage space)
- Editable text (E.g. Resume ,contracts)
- Automatic number plate recognition
- Speech Recognition
- Electronic images of printed documents (E.g. Google Books)

V. TYPES OF HANDWRITTEN OCR

Offline Manuscript Document Type

In this type of text document can be assembled by a human by running scripts with help of pen or pencil on a white paper then scanned the paper to digitalized them is called Offline manuscript document type [6].

Online manuscript document type

In this type of online manuscript document type, the text is directly on a digital platform using computers. The output is an order of X; Y correlated that convey pen place as well as other information such as pressure and speed of running hand [6].

VI. SEGMENTATION PROCES

In optical character recognition, the script Segmentation images is a very crucial task. Plenty of analyses work has been done for OCR segmentation. In any OCR system segmentation stage is the main step to enhance the closeness of hand character identification in OCR absolutely depends upon the segmentation stage. The meaning of segmentation has split an image into a specific part (like words or histogram separation) its component area or object. Basically in segmentation techniques that grind difficult to take out a fixed part (words and histogram) of the images [1]. In segmentation, there are different types of line based segmentation,word-based segmentation, and character-based segmentation. The line segmentation for any handwritten document which is executed to find a number of lines in any scanned handwritten document images and the border 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 handwritten paperwork [7]. After completion of the line and word-based segmentation, we are processing character segmentation to find the character in any scanned printed international, national and local language handwritten paperwork [8].

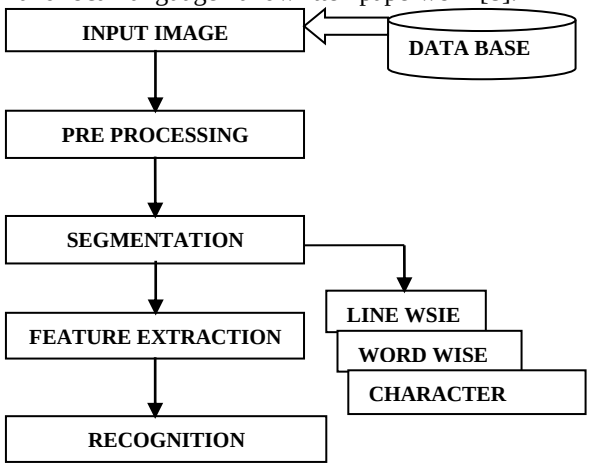


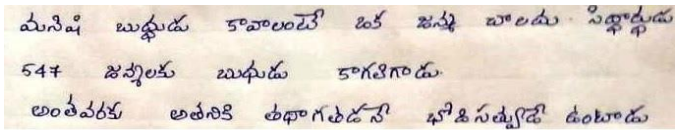
Fig 1 Flow Diagram of Segmentation process

Input image is taken from database (database is the folder which consists of several images, which is underground for the process) after getting input image from the database, pre-processing is carried out, the preprocessing is nothing but getting the binary descriptor for the processor in the binary descriptor there is two values which is nothing but 0's ad 1's the 0' area represents the back region and 1's region are represents the white region after the processing the segmentation is carried out here line-wise, word-wise, character-wise segmentation done. Next feature extraction is carried out in the feature extinction we have to apply LBP feature extraction (local binary pattern) after the recognition is carried out in the recognition the printed documents converted into a text document [4]

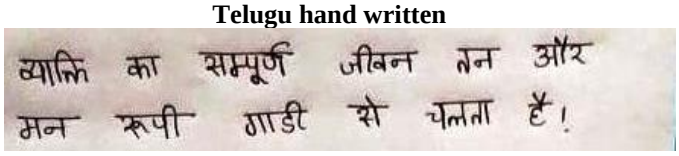
Line segmentation

The First step of the segmentation process is segmenting text part into line part; it is also called 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 frequently used technique to extract the lines from the paper[11]. If the lines are well split and not bend, the vertical estimation will have well-separated heights and gaps. These gaps can be detected simply and used to decide the locations of the line margins. Shown in below fig2.1

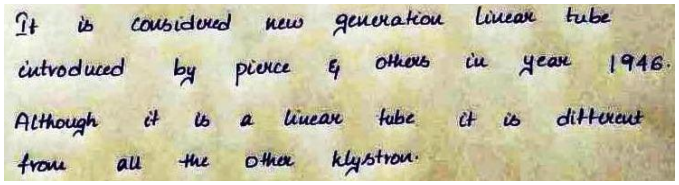
Input images



Telugu hand written

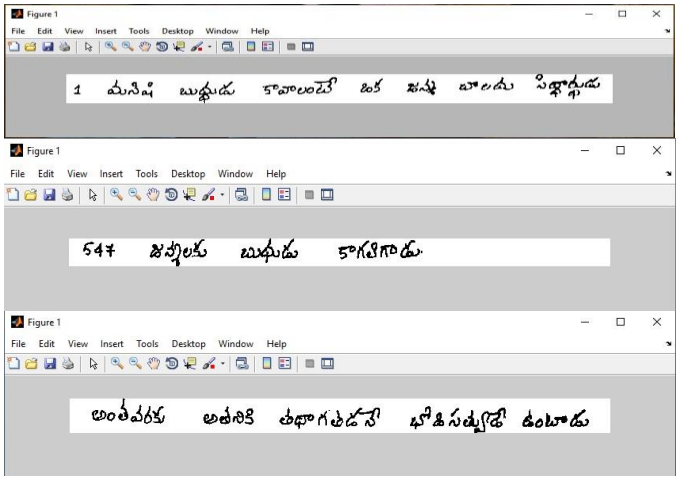


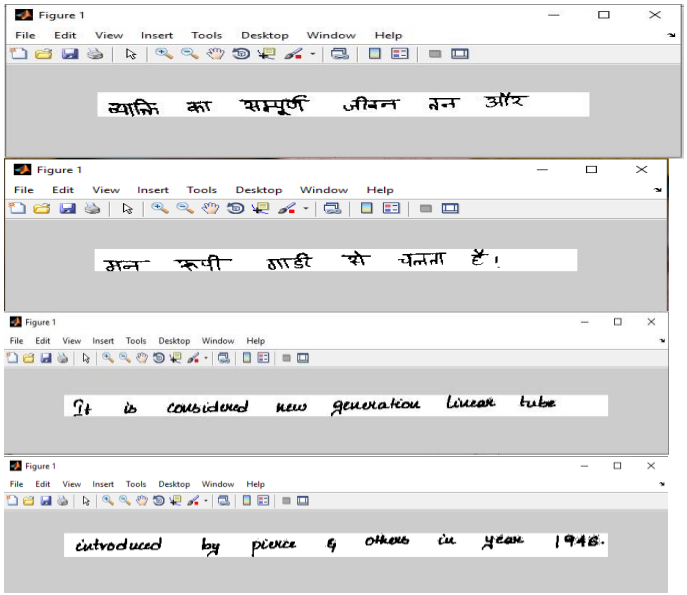
Hindi hand written



English hand written

Output images:

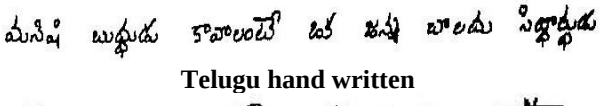




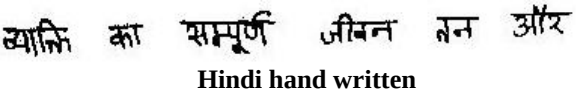
Word Segmentation

The second step of segmentation is word segmentation in this the input images are taken from takeout text lines, by using vertical projection profile (VPP) lines can be segmented into words get.[11]Usually, applying some particular portal considerable vertical gaps, words are split from a text line. An example is shown in Fig.2.2

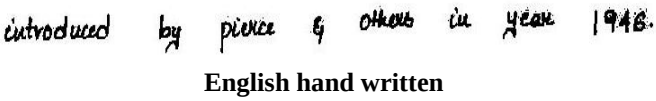
Input images



Telugu hand written

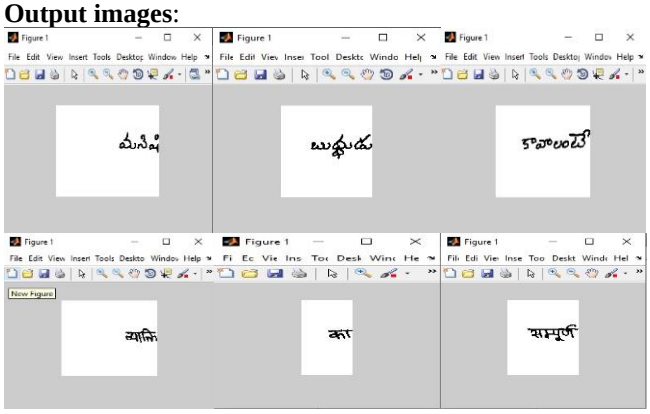


Hindi hand written



English hand written

Output images:



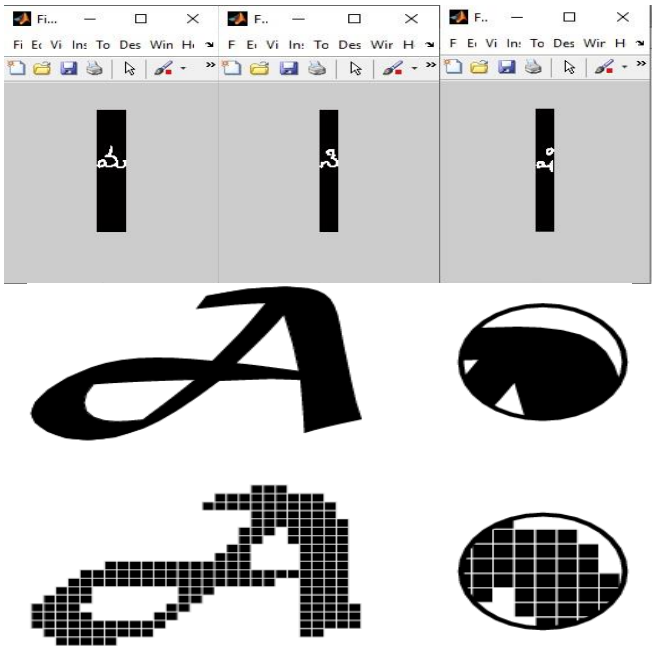
Character Segmentation

The final step of Segmentation is character segmentation; it is the most grueling part of the script segmentation stage. Here the word segmented image is taken as input and the output is in that word each and every character comes individually.

But it is very difficult to segmentation in character-wise in computer collected handwritten scripts some characters in a holder word may moderately overlap with one some other; it becomes very hard to separate those characters properly [5]. This character segmentation is helpful for OCR final step recognition based o individual characters and the structural feature of that character's image recognition can be done. An example is shown in Figure 2.3.

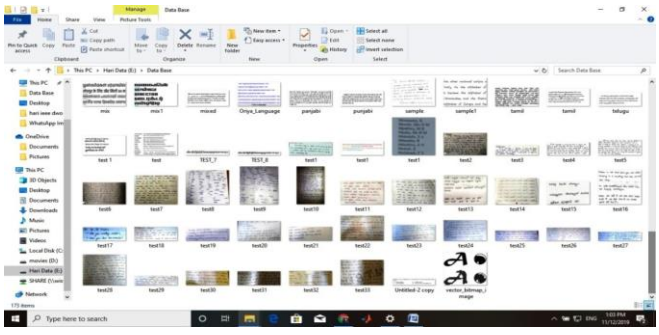
Input image

Output images:



Bit Map BMP Image

Bitmap (BMP) is an image file format that can be used to create and store computer graphics. A bitmap file displays a small dots in a pattern that, when viewed from afar, creates an overall image. A bitmap image is a grid made of rows and columns where a specific cell is given a value that fills it in or leaves it blank, thus creating an image out of the data.



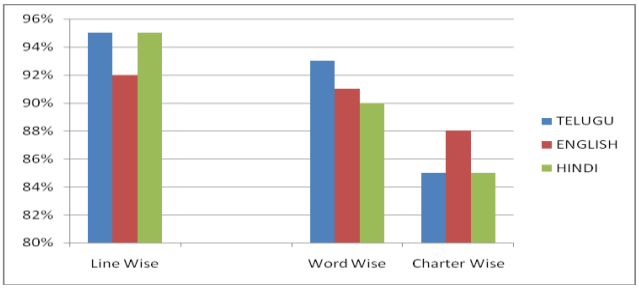
E:\Data Base of all Indian languages

VII. RESULTS

The research was done on OCR (optical character recognition) in which segmentation on handwritten had been observed, the main theme of segmentation on handwritten is to identify or reconcile the language in which local, national, international language scripts are been worked under research. In every language according to MAT lab, line-wise word wise, character-wise segmentation of different handwritten languages.[by above-stated script highest accuracy of the handwritten script is increased]To abstract the segmentation a database maintained in which local, national, international (Telugu, Hindi, English) language handwritten are kept 1000 handwritten scripts are maintained in the above stated data base. While for segmentation from each language two hand written samples are collected and by simulating in MAT lab, the apt and accuracy results are incurred are shown in the given table.

Type Of Hand Written Language	Method	Line Wise	Word Wise	Charter Wise
Telugu	Vertical segmentation method	95%	93%	85%
English	Segmentation based on projection profile	92%	91%	88%
Hindi	Vertical segmentation method	95%	90%	85%

Model Graph



VIII. CONCLUSIONS&FUTURE SCOPE

It's a tough task to reconcile the handwritten scripts by using this partition algorithm language. Segmentation accuracy was raised by doing his process in OCR, recognition of the handwritten languages are easily traced. For future perspective, this experimentation is highly useful for the industrial, commercial, judicial purposes, public sectors, and private sectors. It can be applicable for other Indian vernaculars regarding segmentation on handwritten scripts which hence fruit full results. In this project, the method used to develop an OCR for the identification of running hand Document Images. Conducted tests to evaluate its production in which got good results on good quality documents. Here tested international English language and national language Hindi and local language Telugu tested, with different font sizes and styles. For, the segmentation accuracy is more than 90% and for small font size, the segmentation accuracy declines and will be in the range of 90% to 95%. Apply this method for all other languages running hand documents also.

REFERENCES

1. M.Arun, S.Arivazhagan, D.Rathina, "Handwritten Text Segmentation Using Pixel Based Approach" Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019) IEEE Xplore Part Number: CFP19J32-ART; ISBN: 978-1-5386-9439-8
2. Huo Liulei, Kamil Moydin, Abdusalam Dawut, Askar Hamdulla "The Algorithms For Segmentation Of Text-Lines In Handwriting Images" 2018 3rd International Conference on Smart City and Systems Engineering (ICSCSE)

3. Ge Peng, PengFei Yu, HaiYan Li, HongSong Li, XuDong Zhu "A Character Segmentation Algorithm for the Palm Leaf Manuscripts" 2017 2nd IEEE International conference intelligence and applications Adam GHORBEU, lean-Marc OGIER A segmentation free Word Spotting for handwritten documents 2015 13th International Conference on Document Analysis and Recognition (ICDAR)
4. Shuchi Kapoor and Vivek Verma "Fragmentation Of Handwritten Touching Characters In Devanagari Script" (IJITMC) Vol. 2, No. 1, February 2014
5. Muhammad M. Mehdi, Aqsa Riaz "Optimized Word Segmentation for the Word Based Cursive Handwriting Recognition" 2013 European Modelling Symposium
6. N.Vishwanath, S.Somasundaram,A.R. Jariya Begum,N. Krishnan Nallaperumal A Novel 2-Row Indian LP Character Segmentation Algorithm based on a Hybrid Approach 978-1-4673-1344-5/12/$31.00 ©2012 IEEE
7. Femin P D, Krishna priya K S, Dr. Vince Paul, "Handwritten Text Extraction Using HOG Feature and SVM", IJIRSET, Vol. 6, Issue 4, April (2017).
8. Han X C , Yao H , Zhong G Q , Handwritten text line segmentation by spectral clustering[C]/ /Proceedings of the SPIE 10225 , 8th International Conference on Graphic and Image Processing,Tokyo,Japan: SPIE,2017: #102251A. [DOI: 10. 1117/12.2266982]
9. Sébastien Eskenazi,Petra Gomez-Krämer,Jean-Marc Ogier. A comprehensive survey of mostly textual document segmentation algorithms since 2008[J]. Pattern Recognition,2017,64
10. P.Chitra,B.Sheela Rani,B.Venkatraman,Baldev Raj Evaluation of the Signal to Noise in Different Radiographic Methods and in Standard Digitizer Indian Journal of Computer Science and Engineering (IJCSE) ISSN : 0976-5166 Vol. 2 No. 5 Oct-Nov 2011

AUTHORS PROFILE

B.Harikumar pursing Part Time PhD Scholar, Department of ECE, Sathyabama Institute of Science and Technology (Deemed to be University) 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 5 years Exp Teaching. His area of interests is Image Processing.

Dr P.Chitra professor Department of ECE, Sathyabama Institute of Science and Technology (Deemed to be University) she had received her doctorate in Sathyabama Univeristy in September, 2014, She had more than 13 years of teaching experience and She had taught variety of courses for PG and UG. She had published more than 25 papers in reputed journals and conferences.