International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 5 | ISSN : 2456-3307

Optical Character Recognition - A Review

Dr. S.Vijayarani[*], M.G eetha[#],

[*1]Assistant Professor, Department of Computer Science, Bharathiar University, Coimbatore, Tamil Nadu, India
[#]M.Phil Research Scholar, Department of Computer Science, Bharathiar University, Coimbatore, Tamil Nadu, India

ABSTRACT

Optical Character Recognition (OCR) is the transformation of digital document images into editable text format. Nowadays, most of us downloading the digital documents and e-books from an internet. Normally, most of the downloaded documents are available in the form of images, so it is impossible to edit or to perform search process for extracting any information from these digital documents. Optical Character Recognition is a process of detecting and recognizing the characters from the document images. Documents are converted into digital form with the help of scanners and mobile phones. Scanners scan the documents, whereas mobile phones takes the snapshots of the documents. OCR technique helps to convert handwritten or printed document images into editable form and then we can perform document editing and searching process. The main disadvantage of OCR tool is the accuracy, i.e. it unable to translate the document image accurately into editable form. The OCR does not convert some of the characters, numbers and special symbols in the document images properly. Therefore, there is a need for development of accurate OCR techniques, which can able to perform search and edit processes successfully. This paper studies the fundamental concepts of OCR, significant steps, and their related works.

Keywords : Document Image, Optical Character Recognition (OCR), Segmentation, Feature Extraction.

I. INTRODUCTION

Document Image Processing and Optical Character Recognition (OCR) is most interesting and fascinating field of pattern recognition and human-machine interface for the last few years. Document images have become more popular in digital libraries and digitized organizations. Character recognition is a process of detecting and recognizing character from input and converts into equivalent editing code or ASCII codes [1] [2]. The main goal of the image document analysis and recognition is to identify and extract the text and graphical components from the document images, based on the user requirements. The main two components of document image analysis are, (i) Textual processing and (ii) Graphical processing. Textual processing deals with the text in the document images. Optical Character Recognition is used to recognize both the handwritten and printed document images. Graphical processing deals with graphical images, non-textual line and symbols.

The main aim of OCR is to convert the printed materials into text or word document files that can be easily manipulated and stored. OCR recognizes the document image by dividing the document pages into lines and then further sub-divide into word and then followed by character [3]. Each character is compared with image patterns to predict the same characters. Many open-source OCR tools are available. Some of them are Online OCR, i2OCR, Free Online OCR, Tamil OCR etc. The main drawback of these OCR tools is, unable to convert the document images accurately into an editable format. Some of the characters mismatched with the original one, which reduces the conversion accuracy. Hence, there is a need for development of new tools, which can able to perform the conversion process accurately. The main objective of the paper is study and analyses the OCR and the use of OCR for Tamil Document Images.

This paper is organized as follows; OCR process is discussed in Section 2. Section 3 presents the comprehensive study of existing approaches and conclusion is given in Section 4.

II. OCR Process

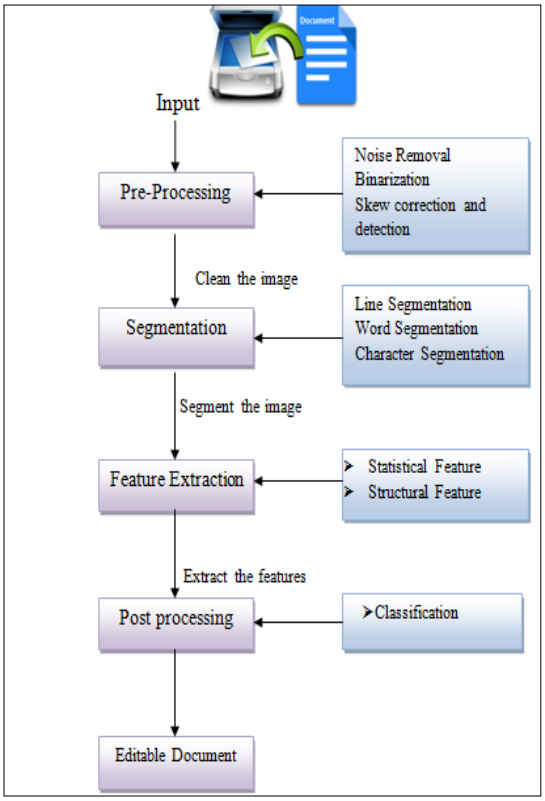


Fig 1. OCR process

OCR Process is given in Figure 1. The process has several components. The important components are (1) Input document image (2) Pre-processing (3) Segmentation (4) Feature Extraction (5) Post Processing and (6) Editable document.

2.1 Input Document image

This paper discussed how OCR handled Tamil handwritten and printed documents. Tamil language is widely spoken in Tamil Nadu. Tamil has the ancient complete literary tradition amongst the Dravidian languages. Tamil is inherited from the Brahmi script. The language Tamil has 12 vowels and 18 consonants. These are combined with each other to yield 216 composite characters and character (aayathaezhuthu). The basic characters of Tamil languages are described in figure 2.

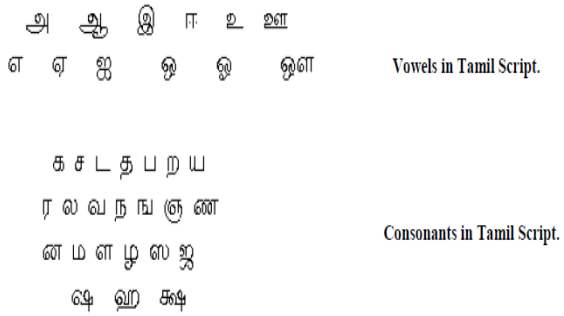


Fig: 2 Tamil Characters

2.2 Pre-processing

Pre-processing is an initial step in performing character recognition. A printed document is scanned and converted into a suitable format for processing the character. It consists of different sub processes for cleaning the document image to make it clear [4]. After the noise removal process, the image is appropriate to carry out the recognition process accurately [5]. Some of the sub processes involved in pre-processing are given below.

- Binarization
- Noise removal
- Skew correction and detection

2.2.1 Binarization

Binarization is a method of converting a grayscale image into a black and white image through the threshold values into 0 and 1 respectively [6]. Image thresholding is a process of separating an image into foreground and background [7][8]. Based on the intensity values the local and global thresholding methods are categorized. They are global and local threshold methods, global method is Otsu and local threshold methods are Niblack, Nick and Savuola for image binarization. This method helps to extract useful information from low quality document images. For example the use of pen quills which was generally used in the historical document images, which is highly responsible for some degradation such as faint, shadowy characters, ink bleeds, large stains. Figure 3 shows the (a) Input image, (b) Otsu, (b) Niblack method, (c) Nick (e) Savuola

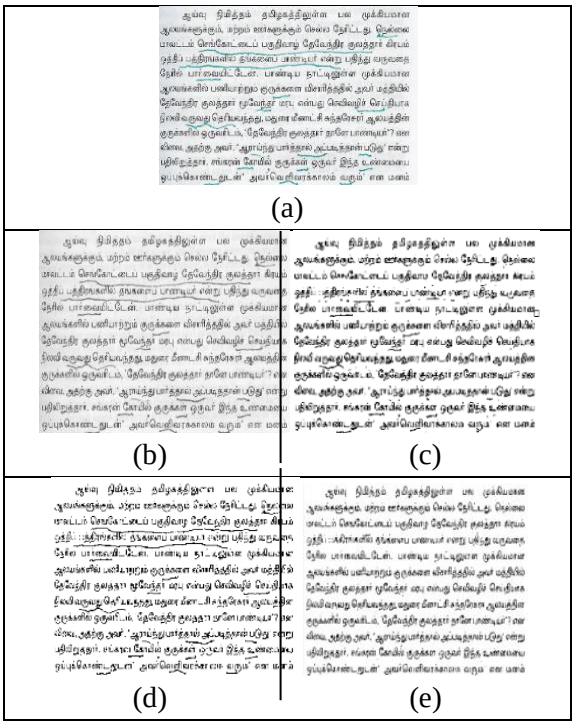


Fig: 3 (a) Input image, (b) Otsu, (b) Niblack method, (c) Nick (e) Savuola

2.2.2 Noise Removal

A noise has a different variation of image intensity and visible as a spot of grains in the image. In the document image, the pixels in the image show different intensity values. Noise removal algorithms are available to remove the visibility of noise by smoothing the image. Some of the noise removal algorithms are Gaussian, average filter and wiener filter. Figure 4 shows the (a) Input image, (b) Median, (c) Average filter, (d) Gaussian filter.

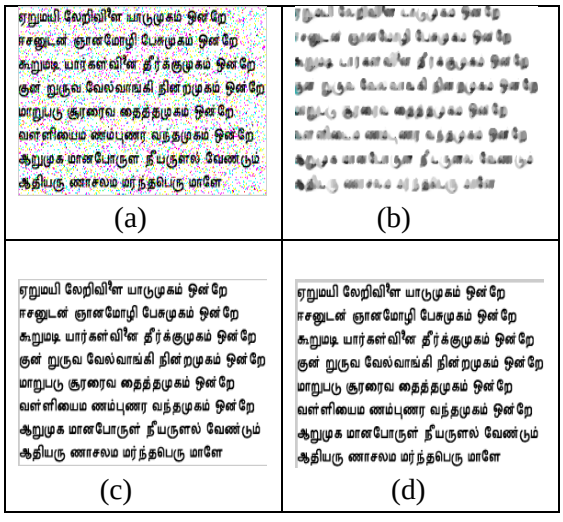


Fig: 4 (a) Input image, (b) Median, (c) Average filter, (d) Gaussian filter

2.2.3 Skew detection and Correction

Document image is fed into an optical scanner manually, a few degrees of tilt or skew are unavoidable. Skew angle is the angle that the text lines can be tilted or skewed either horizontal or vertical direction [9]. Whenever a skew error is detected then, the images have to be de-skewed. The main process of skew correction is to make the correction by rotating the image by an angle equal to skew, but in the opposite direction. Popular skew detection algorithms are Projection Profile Analysis and Hough Transform etc.

2.3 Segmentation

Segmentation is a process of partitioning the document image into groups of pixels. Image segmentation is the process of dividing an image into multiple regions or sets of pixels based on some criterion such as intensity, color, texture etc[10]. There are three main phases of image segmentation.

- Line Segmentation
- Word Segmentation
- Character segmentation

2.3.1 Line segmentation

Line segmentation is the initial step for text based image segmentation. It has horizontal scanning and vertical scanning of the image, the pixels by pixels from left to right and top to bottom [11]. The level of intensity of each pixel is detected or depends on the pixel values, which are grouped into numerous regions.

2.3.2 Word segmentation

Word segmentation is the next step of image segmentation, which includes the vertical scanning of the image. It scans the images through pixel by pixel, row from left to right and top to bottom. The vertical projection profile method is mainly used to segment the word from the text line[12]. The vertical projection profile shows the gap between two different pixel range values. A segmented line is selected and split the text region into the units and then finally to words.

2.3.3 Character segmentation

Character segmentation is the final step for text based image segmentation. It is a similar operation like word segmentation. Character segmentation has a combination of both line and word segmentation phase. For example, the input for character recognition can be either line or word. Figure 5 (a) (b) (c) (d) gives the input image

Line, Word and Character segmentation reults.

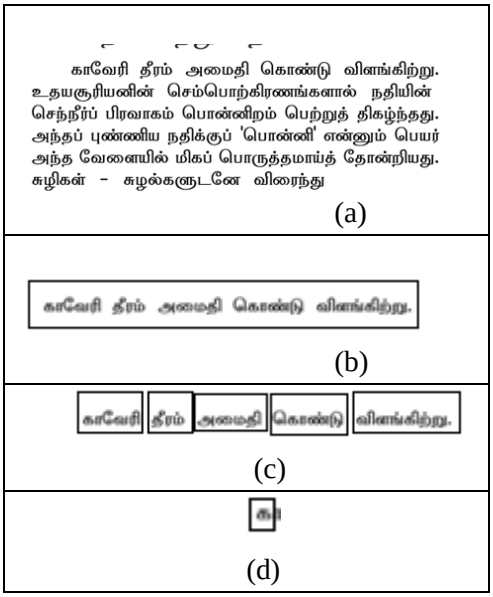


Fig 5. (a) Input image, (b) Line segmentation, (c) Word segmentation, (d) Character segmentation

Feature Extraction

Feature extraction involves extracting meaningful information about an object or a group of object from the document images, features are extracted using character recognition, which helps to identify the different characters or words in document images. It is used to analyze the input images, then recognize characters and words in document image, and then translates character images into character codes [13][14]. Character codes match with the existing template, before that exiting template are stored in a database. The feature extraction methods are broadly grouped into two classes, namely statistical features and structural features.

2.4.1 Structural features

Structural features identify structural features of a character. Structural features are based on topological and geometric properties of the character. Examples of structural features are number of horizontal lines or vertical lines, the direction of the character, number of endpoints, number of cross points, horizontal curves at the top or bottom, etc[15]. This type of representation may also encode some knowledge about the structure of the object or may provide some knowledge as to what sort of components make up that object. Figure 6 gives structural features of the character.

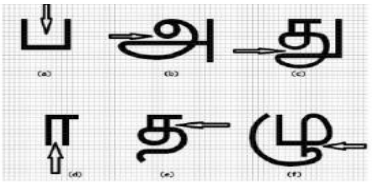
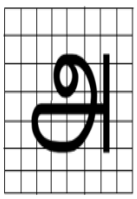


Fig 6. Structural feature of a character

2.4.2 Statistical feature

A statistical feature identifies statistical features of a character. It derived from the statistical distributions of pixels [12]. These features can easily detect as compared to structural features. Statistical features are not affected too much by the noise or distortions as compared to statistical features. A number of techniques are used for statistical feature extraction; some of them are: zoning, projection histograms, crossings and distances, and n-tuples. It identifies the characters by examining their sub feature shapes of the image, sub-vertical and horizontal histograms. Its character repair capability is great for low quality text and newsprint. Figure 7 represents the statistical feature of a character.



No of Loop	1
No of Vertical Line	1
No of Horizontal line	1
Lower arc	0
Upper arc	0
Slant Line	0
No of Curve	1

Fig 7. Statistical feature of a character

2.4 Post Processing

Post processing is required for character recognition, it is an art of detecting, segmenting and identifying characters from an image. Classification techniques are

used for character recognition, it mainly decides the feature space to which the unknown pattern belongs. It is another important component of OCR system. Classification is usually done by comparing the feature vectors equivalent to the input character with the representative of each character class. There are different types of approaches are used to classify the character features in the existing system such as K-Nearest Neighbor, Artificial Neural Network and Support Vector Machine etc.

The conversion of handwritten or printed documents into machine editable forms, so that it can be easily accessed in optical character recognition. It is further divided into off-line character and online character. Offline-Character recognition refers to the process of recognizing words that can be scanned page (as a sheet of paper) are stored digitally in grey scale format [6]. Online character is captured and stored in digital form via different media. The data is generated by users with stylus (pen).Figure 8 shows the different printed and handwritten characters and Offline and Online Characters.

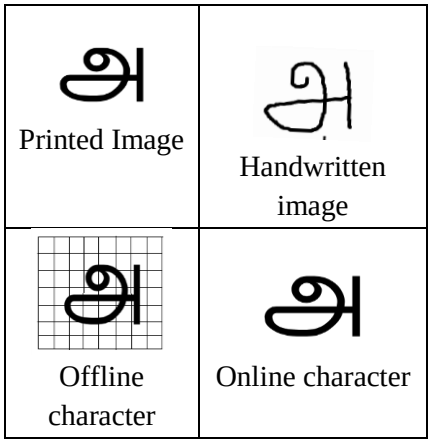


Fig 8. Different types of fonts

III. COMPREHENSIVE STUDY

The table 1 shows the comprehensive study of which has been made on the different OCR's available for character recognition**.**

IV. CONCLUSION

Optical character recognition is one of the important and the challenging research domain in the field of image processing. However, there is no standard solution for the identification of Tamil characters accurately. Various methods have been used for recognition process, but all approaches have given an accurate solution for few character sets. Different types of challenges are recognized for both handwritten and printed document images for abnormal, slanting characters, similar shaped, joined characters, curves and strokes. These challenges are to be explored in future research for getting better results.

V. REFERENCES

[1]. Aparna K G, A G Ramakrishnan,"A Complete Tamil Optical Character Recognition System", International Journal of Machine Learning and Computing, Vol. 2, No. 3, June 2012.

[2]. Seethalakshmi R,Sreeranjani T. R, Balachandar T," Optical character recognition for printed Tamil text using Unicode", November 2005, Volume 6, Issue 11, pp 1297–1305

[3]. AkshayApte and HarshadGado, "Tamil character recognition using structural features" ,2010

[4]. JagadeeshKannan R and PrabhakarR,"An improved Handwritten Tamil Character Recognition System using Octal Graph", Int. J. of Computer Science, ISSN 1549-3636, and Vol 4 (7): 509-516, 2008.

[5]. Jagadeesh Kumar R, Prabhakar R and Suresh R.M, "Off-line Cursive Handwritten Tamil Characters Recognition", International Conference on Security Technology, page(s): 159 – 164, 2008

[6]. "A survey of modern optical character recognition techniques" (DRAFT), February 2004

[7]. Amarjot Singh, KetanBacchuwar, and AkshayBhasin,"A Survey of OCR Applications"International Journal of Machine Learning and Computing, Vol. 2, No. 3, June 2012.

[8]. G.Vamvakas, B.Gatos, N. Stamatopoulos, and S.J.Perantonis, "A Complete Optical Character Recognition Methodology for Historical Documents", 2008IEEEDOI 10.1109/DAS.2008.73

[9]. M. Antony Robert Raj, Dr.S.Abiram,"A Survey on Tamil Handwritten Character Recognition using OCR Techniques" DOI: 10.5121/csit.2012.2213.

[10]. U. Pal and B. B. Choudhuri. A Complete Printed Bangla OCR Systrem. Pattern Recognition. Vol 31. May 1998

[11]. Amarjot Singh, KetanBacchuwar, and AkshayBhasin,"A Survey of OCR Applications"International Journal of Machine Learning and Computing, Vol. 2, No. 3, June 2012.
[12]. RohitVerma, Dr. Jahid Ali, "A-Survey of Feature Extraction and Classification Techniques in OCR Systems", International Journal of Computer Applications & Information Technology Vol. I, Issue III, November 2012 (ISSN: 2278-7720).
[13]. Dr.AmitabhWahi, Mr.Sundaramurthy.S, PoovizhiP,"Handwritten Tamil Character Recognition",2013 Fifth International Conference on Advanced Computing (ICoAC).
[14]. Maya R. Gupta, Nathaniel P. Jacobson, Eric K. Garcia," OCR binarization and image pre-processing for searching historical documents", Pattern Recognition 40 (2007) 389 – 397, 2006.
[15]. RamanathanS.Ponmathavan,N.ValliappanL.Thaneshwaran, Arun.S.Nair, "Optical Character Recognition for English and Tamil Using Support Vector Machines", Advances in Computing, Control, & Telecommunication Technologies, 2009. ACT '09. International Conference on 2009
[16]. 16. V. Ajantha Devi, S Santhosh Baboo," Embedded Optical Character Recognition On Tamil Text Image using Raspberry Pi ", International Journal of Computer Science Trends and Technology (IJCST) – Volume 2 Issue 4, Jul-Aug 2014