CHALLENGES IN TEXT RECOGNITIONS OF INDIAN LANGUAGES

A.A. Tayade[1] and R.J. Ramteke[2]
[1]G.S. Science, Arts and Commerce College Khamgaon, MS, India
[2]School of Computer Science, K.B.C. North Maharashtra University, Jalgaon, MS, India
[1]arvindtayade40@gmail.com, [2]rakeshramteke@yahoo.co.in

ABSTRACT

Lots of research work has been carried out on optical character recognition of different languages. Different ancient language languages need to be recognized. The need of recognition of these ancient and some continental languages explore the new knowledge. Huge literature is available on these languages which will gives immense knowledge of science, medicine and civilization. This research paper explore the knowledge about the OCR in different Indian language, formation of database and their challenges in recognition

Keywords: *Character recognition, language database, feature extraction and evaluation*

Introduction

Important research has been performed in recent years in the area of text recognition in Indian languages. There have been several attempts to establish OCRs for acknowledgment in Indian languages such as Bangla, Hindi, Tamil, and Kannada. Since there are several languages and various scripts in India, there are many difficulties in designing state-of-the-art text recognition platforms through all of these applications. Researchers report results on their own datasets, which in most cases are not publicly available. Text recognition of Indian languages is an emerging domain that has only recently gained much-needed traction. There is therefore a marked lack of standard datasets. Indian language consists of several scripts, including only two Devnagari scripts, and Bangla has some significant dataset associated with them. This lack of data sets is a serious concern as it results in sub-par performance in most modern machine learning techniques such as Neural Networks, Long-Term Memory and Support Vector Machines, because most of these modern techniques are highly data-driven.

The wide reach of this domain further compounds the problem. There exist several modalities for each of the languages including scanned papers, born-digital photographs, natural scene images, and text in videos. Often, most datasets in this domain are not freely accessible, and those that do, are fragmented and are individual attempts. Another critical problem is that there is no overarching community-driven effort to monitor and benchmark the various datasets in this domain. This research is carried out to present the character recognition and challenges.

Optical Character Recognition

OCR programme processes a visual image by finding and marking characters such as letters, numbers and symbols. Some OCR software simply exports text, while other programmes may translate characters to editable text directly in the image. Advanced OCR programme can export the size and type of the text as well as the style of the text contained on the page.

The OCR technology can be used to transform a physical copy of a paper to an electronic edition. For example, if you search a multipage document for a digital image, such as a TIFF file, you can load a document into an OCR programme that recognises the text and transforms the document to an editable text file. Some OCR programmers allow you to scan a document and convert it to a word processing document in one stage.

Although OCR technology was originally developed to recognise typed text, it can also be used to recognise and validate handwritten text. For example, postal services such as USPS use OCR software to automatically process letters and address-based parcels. The algorithm checks the scanned information against the current address database to validate the mailing address. The Google Translate software provides the OCR technology that interacts with the camera on your smart phone. It helps you to grab text from records,

magazines, posters, and other items and translate it into a foreign language in real time.

Types of Character Recognition

- **Optical Character Recognition (OCR) –** targets typewritten text, one glyph or character at a time.
- **Optical Word Recognition –** targets typewritten text, one word at a time.
- Intelligent character recognition (ICR) – also targets handwritten print script or cursive text one glyph or character at a time, usually involving machine learning.
- **Intelligent Word Recognition (IWR) –** also targets handwritten print script or cursive text, one word at a time. This is especially useful for languages where glyphs are not separated in cursive script.

OCR is usually an offline method that analyses a static text. There are cloud-based providers that have an OCR API web infrastructure. Handwriting action analysis can be used as input for handwriting recognition. Instead of simply using the outlines of glyphs and words, this method is capable of recording gestures, such as the order in which the fragments are drawn, the path, and the pattern of placing down and raising the pen. This additional knowledge can make the end-to-end method more reliable. The technology is also known as "on-line character recognition" "dynamic character recognition" "real-time character recognition" and "intelligent character recognition"

The name OCR (Optical Character Recognition) has been used in numerous literature contexts, ranging from discrete character recognition to document reading systems. A device that deals with an unregulated document is more properly referred to as a document recognition system. The character classification process after isolation of the character boxes is also referred to as IOCR (Isolated Optical Character Recognition). Both the roles of isolation and recognition of the character boxes face problems in real life circumstances.

In the case of hand-written characters, the difference in the scale and form of the characters, the orientation, the fusions and the fragmentations are more noticeable. Since 1940, many methods have been attempted for restricted/unrestricted print/hand-printed/written acknowledgment of text with little results.

The approach used for the OCR can be narrowly separated into three categories:

a. Statistical Approach
b. Syntactic Approach
c. Hybrid Approach

Available Databases of Indian Scripts

The available work done on Indian scripts has been mostly on small datasets collected in laboratory experiments. However, two large databases for handwritten numerals of the most popular Indian scripts Devnagari and Bangla, was developed by Bhattacharya et al. in 2009 for the first time. After this, ISI Kolkata developed large databases for major Indian scripts which are available to researchers on demand. The various Indian scripts database research centres are tabulated in Table given below:

Table 1. List of Indian scripts database along with their canters

Language Script	Name of research Institute
Devnagari	ISI Kolkata
Bengali	ISI and Jadavpur University, Kolkata
Assamese	IIT Guwahati
Gujarati	M.S. University of Baroda, Vadodara
Kannada	IISc Bangalore and University of Mysore
Malayalam	CDAC, Thiruvananthapuram and Kannur University
Oriya	Utkal University, Bhubaneshwar
Gurmukhi	Thapar Institute of Engineering and Technology, Patiala
Tamil	IIT Madras
Telugu	IIIT and University of Hyderabad
Urdu	CDAC, Pune

Some of these datasets are very large collected from diverse sources while the datasets for Kannada, Tamil, Telugu are small.

Creation of Database

Database is created by using Google input tools. Google IME provides tools that make writing in regional language ease. Transliteration, Input Methods, Keyboard, and Handwriting are the IME.

Database is mainly classified in two types according to their appearance as Numerical database and Alphabetical Database. Alphabets are the combination of both consonant and vowels. There are total 12 vowels and 36 consonant present in Marathi language.

Database is created for all these vowels and consonants. Also database created for numerical values that is from 0,1,2,……9. Each letter of Marathi language is write in Google input tool then copy that letter into Paint. In paint edit letters according to requirements and study it. Letters are edited accordance with various font and different sizes.

For this dissertation total 8 types of Font are used namely Aparjita, Arial, Kokila, Mangal, NIrmala, Nirmala semi light, Sanskrit and Utsaah. Edit letters in each font and resolution done accordingly. This step also followed for creating Numerical database.

Challenges in Evaluation

Many current OCR approaches use two main measurement standards to test text recognition activities. Character error rate is a character metric that is based on the Levenshtein distance, which is the minimal number of single-character editing operations (inserts, deletions, and substitutions) needed to modify the given term to another. Term error rate is a word metric that is often based on the Levenshtein distance with the same character metric as the word metric. That is, a minimum number of single-word operations required to change one text to another.

In the case of the Indian language, CER and WER are unable to accurately represent and evaluate all aspects of the text recognition method in the Indic script.

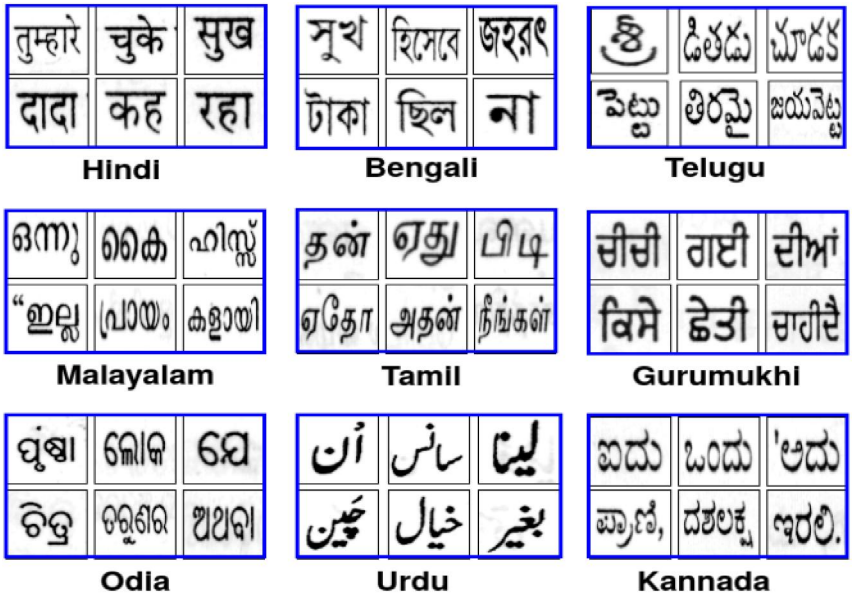

Fig. 1: A visual illustration some Indian scripts.

Algorithms for Document Understanding Segmentation

The problem of document segmentation is that of dividing a document image into a hierarchy of meaningful regions like paragraphs, text lines, words, image regions, etc. These regions are associated with a homogeneity property and the segmentation algorithms. In order to improve the performance of an algorithm over

time using multiple examples or by processing an image multiple times, the algorithm is given appropriate feedback in the form of homogeneity measure of the segmented regions. Feedback mechanisms for learning the parameters could be employed at various levels.

Feature Extraction and Classification

In recent years, there has been a renewed attempt to reformat the classification approaches to the recognition of difficult character sets. It has been found that a multiple classification character recognition schemes have the potential of outperforming individual stand-alone classifiers because of its ability to handle extreme variance in the training and testing samples. The classification process can be carried out in three stages. In the first stag the characters are grouped into three sets depending on their zonal position (upper zone set, middle zone set and lower zone). In the second stage the characters in middle zone set are further distributed into smaller sub-sets by a binary decision tree using a set of robust and font independent features at each node of the tree. In the third stage the nearest neighbor classifier is used and the special features distinguishing the characters in each subset are used. One significant point in this scheme is that in contrast to the conventional single-stage classifiers where each character image is tested against all prototypes, a character image is tested against only certain subsets of classes at each stage. This eliminates unnecessary computations.

Conclusion

As discussed in this research, character recognition is computationally difficulty process, as the hand written characters will be scanned and recognition on the basis of their patterns. The pattern recognition is carried out with the help of a specified algorithm selected for the identification of symbols depending upon the pattern language. This research, provides the different issues Optical Character Recognition with its types, creation of database depending of number of languages selected for recognition and challenges of evaluation.

References

1. Salunkhe, P., Bhaskaran, S., Amudha, J and Gupta, D. (2017). Recognition of multilingual text from signage boards, Sixth International Conference on Advances in Computing Communications and Informatics (ICACCI), pp. 977-982
2. Sinha, R. and Bansal, V. (1995). On devanagari document processing, Systems Man and Cybernetics 1995. IEEE International Conference on Intelligent Systems for the 21st Century, vol. 2, pp. 1621-1626
3. Bansal V. and Sinha, R. (2000). Integrating knowledge sources in devanagari text recognition system, IEEE Transactions on Systems Man and Cybernetics Part A: Systems and Humans, vol. 30, no. 4, pp. 500-505
4. Ashwin T.V. and Sastry, P.S. (2002). A font and size-independent OCR system for printed Kannada documents using support vector machines, Sadhana-academy Proceedings in Engineering Sciences, vol. 27, pp. 35-58
5. Neeba, N., Namboodiri, A., Jawahar C. and Narayanan, P. (2010). Recognition of Malayalam Documents, London, pp. 125-146
6. Bansal V. and Sinha, R.M.K. (2000). Integrating knowledge sources in devanagari text recognition system, vol. 30, pp. 500-505