Abstract

Optical Character Recognition (OCR) is a technology used to recognize and convert printed or handwritten text into digital text that can be processed and analysed by computers. While OCR technology has been developed for many languages, OCR specifically for Chhattisgarhi, an Indo-Aryan dialect spoken in the central-eastern part of India, may be less common.

In this paper, we propose the development of OCR technology for Chhattisgarhi. This involves building a dataset of scanned Chhattisgarhi text images, preprocessing the images to remove noise and enhance contrast, and training an OCR model using machine learning algorithms to recognize and convert Chhattisgarhi text images into digital text. The accuracy of the OCR model will then be evaluated on a separate dataset of scanned text images.

The development of OCR technology for Chhattisgarhi will require specialised knowledge of the Chhattisgarhi script, including its unique characters, punctuation, and grammar rules. However, with the increasing demand for regional language technologies, the development of OCR technology for Chhattisgarhi and other regional languages is essential and can greatly benefit the wider community.

1.Introduction

Chhattisgarhi is an Indo-Aryan dialect spoken in the central-eastern part of India, primarily in the state of Chhattisgarh. It has a rich cultural heritage and is widely used in literature, music, and movies. The Chhattisgarhi script is the writing system used to write the Chhattisgarhi dialect. The Chhattisgarhi script is a variation of the Devanagari script,

which is also used to write other Indian languages such as Hindi and Marathi.

The Chhattisgarhi script is used in various forms of written communication, including literature, journalism, and official documents.

Despite the widespread use of the Chhattisgarhi dialect, the script is less well-known outside of the Chhattisgarh region of India. However, with the increasing demand for regional language or dialect technologies, the development of OCR technology and other language technologies for Chhattisgarhi is essential and can greatly benefit the wider community.

1.1 What is Optical Character Recognition (OCR) ?

Optical Character Recognition (OCR) is a technology that enables the conversion of scanned images, PDF files, or handwritten documents into digital text that can be edited, searched, and analysed by computers. OCR software scans the document image and uses pattern recognition algorithms to identify the characters in the image and convert them into text.

OCR technology has a wide range of applications in various industries, including finance, healthcare, legal, and education. It is used to digitise paper documents, automate data entry, and enable text analysis. OCR software can recognize text in a wide range of fonts, sizes, and styles, and can also recognize handwritten text with varying degrees of accuracy.

OCR technology is often used in conjunction with other technologies, such as natural language processing (NLP), to analyse and extract information from large volumes of text. OCR can also be used to generate subtitles for videos and images, or to recognize licence plates and other types of alphanumeric characters.

OCR technology has made it possible to digitise vast amounts of information that were previously only available in hard-copy form,

making it easier to access and analyse this information. However, OCR technology is not always perfect and can still make mistakes, especially when dealing with handwritten or poorly scanned documents. Therefore, it is important to verify and proofread the output of OCR technology to ensure accuracy.

1.2 Introduction to Chhattisgarhi Script

Chhattisgarhi script dates back to the 12th century when the Nagari script was used to write the Chhattisgarhi dialect. The Chhattisgarhi dialect, which is an Indo-Aryan dialect, was developed over time by incorporating words and expressions from various regional languages.

During the 17th century, a distinct form of Chhattisgarhi dialect emerged, which had its own unique vocabulary and grammar. It was during this time that the Chhattisgarhi script began to evolve and take shape as a separate script from Nagari.

In the early 20th century, the Chhattisgarhi script was standardised and given its current form. The script contains 52 letters, including vowels and consonants, and is written from left to right.

1.3 Literature Review

Chhattisgarhi literature is a rich and diverse body of work that reflects the cultural heritage and linguistic identity of the Chhattisgarh region in India. The literature includes poetry, fiction, non-fiction, drama, and folk literature, and dates back to the 16th century.

The earliest known work of Chhattisgarhi literature is the 'Dhola Maru', a folk epic composed in the 16th century. Other important works from the early period of Chhattisgarhi literature include the 'Bhaktamala' and the 'Chandayana'.

In the 19th and early 20th centuries, the Chhattisgarhi dialect and literature underwent a significant transformation as the region came

under British rule. Many writers began to experiment with new forms of literary expression, and literary magazines and journals were established to promote Chhattisgarhi literature.

In the post-independence period, Chhattisgarhi literature continued to flourish, with writers producing works that explored themes of social justice, identity, and regionalism.

In recent years, there has been a renewed interest in Chhattisgarhi literature, with efforts to promote the language and its literature through cultural events, literary festivals, and the establishment of Chhattisgarhi language academies. The emergence of digital media has also provided a platform for Chhattisgarhi writers to reach a wider audience.

Overall, Chhattisgarhi literature reflects the rich cultural heritage and linguistic diversity of the Chhattisgarh region, and is an important part of India's literary tradition.

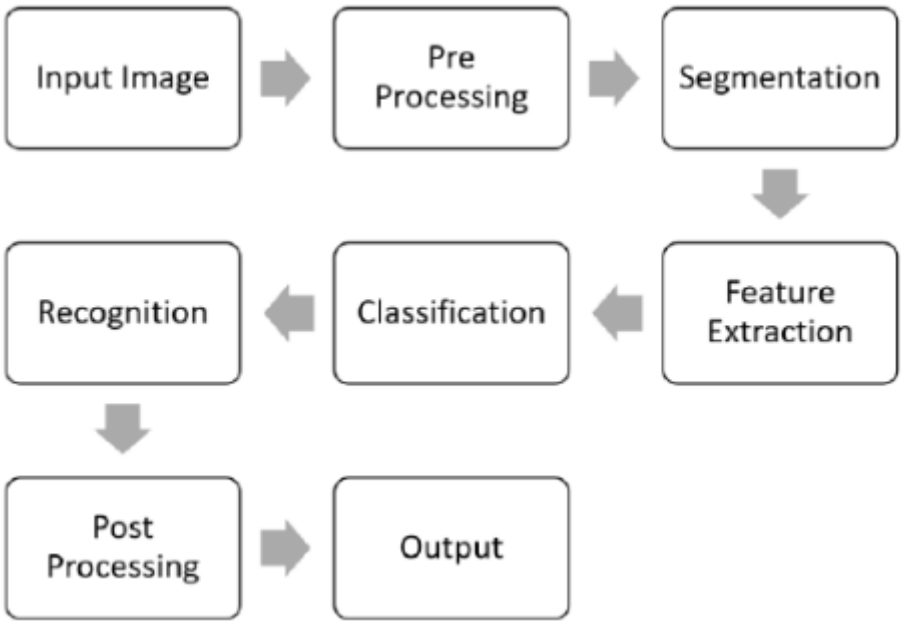
Chhattisgarhi script is the same as the other regional languages like Bangla, Devanagari, Kannada, Tamil, Telugu etc when it comes to the comparative ratio of OCR success rates.

KNN, Decision Tree, Binary Tree Classifier etc approaches required some manual work to be done on almost every step which can be automated up to some extent with the use of the existing OCR engines which are really effective when it comes to OCR operations.

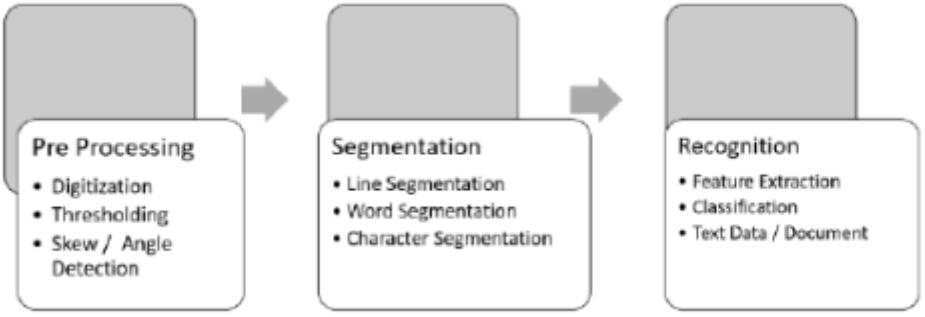
Tesseract is one of them which is used by many researchers worldwide to perform OCR operations on different national and international language‘s scripts because that can be trained as similar as we train any core neural network based mechanism as it‘s already a part of it in many OCR engines. To train Tesseract OCR engine one can gain the information on the official repositories which are managed by Google Inc.

2. Steps of Chhattisgarhi Character Recognition

After the literature review, we explored that each OCR research operation may have the following common steps which may differ in implementation but all the steps are necessary to carry out the basic OCR on any language‘s character recognition. The major steps which are as followed: Input of Image/ Document (know as Image Acquisition or Digitization too), Pre-Processing, Segmentation, Feature Extraction, Classification, Recognition, Post-Processing, Output.



Basic Step of OCR



Major Step to perform OCR

2.1 Preprocessing

Pre-processing is the initial stage of any OCR in which pre-processing work is performed to acquire images using cameras or scanners and after that, some more processing has been carried out to make those images appropriate for further operations. Basically pre-processing step is divided into following sub-steps:

A. Digitization

This is the very first phase of pre-processing step in which the physical or digital document is transformed into images/files using scanners, high-quality cameras or other techniques to make it capable to be processed by a computer system. This process is also called an image acquisition.

B. Thresholding

After the digitization now it comes to thresholding where the image is processed and gets converted into a binary image which is consist of two values (0) or (1) respectively for each pixel value. Thresholding can be done by two methods namely global thresholding and adaptive thresholding

C. Skew / Angle Detection

Skew detection or angle detection is performed to recognize skewed images angle if any of image wasn‘t taken or scanned properly, to make it correct there are several skew correction algorithms can be applied.

D. Noise removal

This involves removing unwanted elements such as dots, lines, or smudges from the input image. This can help to improve the accuracy of character recognition by reducing interference from irrelevant features

E. Smoothing

This involves applying a filter to the input image to reduce noise and smooth out the edges of the characters. This can help to improve the accuracy of character recognition by reducing the effects of pixelation.

F. Binarization

This involves converting the input image to a binary image by setting a threshold value that separates the foreground (text) from the background. This can help to enhance the contrast of the text and remove noise from the image.

Overall, preprocessing is an essential step in Chhattisgarhi OCR that can significantly improve the accuracy of character recognition. The specific techniques used will depend on the characteristics of the input images and the requirements of the OCR system

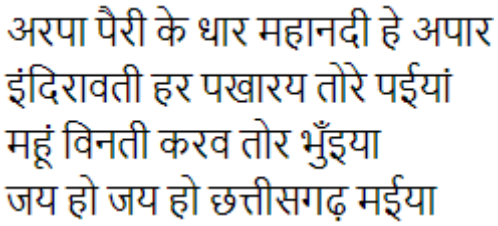
2.2 Segmentation

Segmentation is the most adaptive method acquired by almost every OCR as it split the image into different levels respectively like line, word and characters segments so it makes easier to recognize the characters. As pre-processing, segmentation is also having separate steps as following :

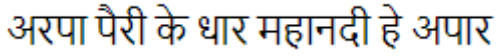
A. Line Segmentation

A horizontal scanning method is applied for segmenting the text paragraphs into lines. While performing the segmentation to

extract the lines from the text blocks, it performs horizontal scanning starting from the top of the scanned document till it locates the last row containing all white pixels, before a black pixel row is encountered. It continues the scanning further, till it

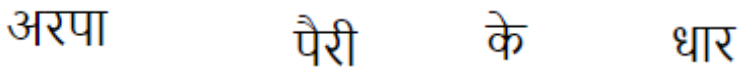


locates the first row containing all white pixels, just after the end of last row of black pixels. This determines a line, and is eventually extracted. This whole process is repeated on the entire text page to segment all the text lines present in that particular page/paragraph



B. Word Segmentation

Word segmentation is as similar to the line segmentation but it works on the segmented lines which are further segmented into individual words. After segmenting the lines it segments the individual words embedded in each line. To per-form this operation a vertical scanning method is applied. The vertical scanning is applied to the width of the line only.



(a) (b) (c) (d)

C. Character Segmentation

Once the words are segmented from text lines using the method described here, a further segmentation process is applied to achieve the individual characters out of the segmented words. Before segmenting words at character level, the header line or shirorekha is identified and re - moved. This process is done by finding the rows with the maximum number of black pixels in a

word. Here we applied a heuristic approach to locate the shirorekha as sometimes, not all the rows of shirorekha contain same number of black pixels. In this paper we have taken black pixel difference counts among rows of shirorekha, though it is not the same value for all other scripts. This difference count for Hindi was decided after an exhaustive analysis of rows of many shirorekha. After locating the shirorekha, it is removed, i.e. converted into white pixels. Once the shirorekha is properly removed, the word is divided into three horizontal zones known as upper, middle and lower zones. Individual characters are sepa-rated from each zone by applying vertical scanning. The vertical scanning is performed for the width of zones and length of the word only. As stated earlier, only modifiers are present in upper and lower zones. So before performing vertical scanning in these zones, it is checked whether any modifier exists or not. This is required because in several cases words may contain only middle zone or middle and upper zone or middle and lower zone only.

There are several algorithms and approaches available for various levels of segmentation which can be studied and applied as per the requirement. Like: Region- growing algorithm and much more.

2.3 Recognition

Recognition is considered as the most important phase as in this step we perform various methods and techniques to recognize the script‘s characters.

A. Feature Extraction

Feature extraction is a process which is carried out in the various phases of segmentation. In feature extraction, each of the characters is processed through the specific technique used for feature extraction to train and test further. Some of the

well-known feature extraction techniques are Template Matching, Zoning, Transformations etc.

B. Classification

After the feature extraction classification and recognition take place in which already extracted features are used along with the various methods and classifiers to classify and recognize each character. Some popular methods and classifiers are Nearest Neighbour (NN), Euclidean Distance, Neural Network, Support Vector Machine (SVM), etc.

C. Text Data / Document

This step can be considered as the last step where we retrieve our classified data in direct text format or in a document file. Along with this step various post-processing step like error identification, correction and grouping can be performed as well for better and quality results.

3. Challenges in Chhattisgarhi OCR

Chhattisgarhi script‘s complexity is higher as compared to the other regional languages so it‘s quite more challenging to perform the Chhattisgarhi OCR and get a more effective result.

Some of the challenges which often affects the accuracy and decrease the resulting quality are as follows:

Lack of standardisation:

There is no standardised version of the Chhattisgarhi script, which makes it difficult to create an accurate OCR system. The

script has undergone several modifications over the years, and there are regional variations in the way it is written.

Limited availability of training data:

The lack of digital text in the Chhattisgarhi language and limited availability of training data makes it difficult to train machine learning models for OCR. This is a major obstacle in the development of accurate OCR systems for the Chhattisgarhi dialect.

Complex script:

The Chhattisgarhi script is complex and has a large number of ligatures (combined characters). These ligatures can make it difficult for OCR systems to recognize individual characters, leading to errors in the recognition process.

Limited research:

There is limited research on OCR for Chhattisgarhi dialect, which makes it difficult to build on existing knowledge and techniques in the field.

These are some of the issues which makes Chhattisgarhi OCR much more difficult.

4. Conclusion

Optical Character Recognition (OCR) for Chhattisgarhi dialect presents a unique challenge due to the lack of standardisation in the Chhattisgarhi script and limited availability of training data for machine learning models. However, with advancements in OCR technology and the growing interest in promoting regional languages, there is potential for the development of accurate OCR systems for Chhattisgarhi language in the future. This would greatly benefit Chhattisgarhi speakers by enabling digital access to a vast amount of historical and contemporary texts in their native language.

REFERENCES :

1. “Chhattisgarhi Language - wikipedia” Online, Available : https://en.wikipedia.org/wiki/Chhattisgarhi_language
2. “Language of India - wikipedia” Online, Available : https://en.wikipedia.org/wiki/Category:Languages_of_India
3. “Tesseract OCR Github” Online, Available : https://github.com/tesseract-ocr
4. Milind Kumar Audichya, Prof. Dr. Jatinderkumar R. saini “An Overview of Optical Character Recognition for Gujarati Typed and Handwritten Character”
5. Kiran R. Dahake, S. R. Suralkar, S.P. Ramteke “Optical Character Recognition for Marathi Text”
6. Divakar Yadav, Sonia Sanchez-Cuadrado, Jorge Morato “Optical Character Recognition for Hindi Language Using a Neural Network Approach”
7. R. Jagadeesh Kannan, R. Prabhakar “A Comparative Study of Optical Character Recognition for Tamil Script”
8. Rijuka Pathak, Somesh Dewangan “Natural Language Chhattisgarhi: A Literature Survey”