IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

OCR FOR HINDI LANGUAGE

[1]Assistant Professor, [2]Student, [3] Student, [4] Student, [5] Student
[1]Department of Information Technology,
[1]Finolex Academy Of Management And Technology, Ratnagiri, Maharashtra, India

Abstract : With over 300 million speakers, Hindi is India's most spoken language. The Optical Character Recognition (OCR) systems designed for the Hindi language have a very low recognition rate due to the lack of character separation in Hindi texts compared to English texts. An Artificial Neural Network (ANN)-based OCR for printed Hindi text written in Devanagari script is proposed in this paper to increase its efficiency. One of the significant purposes behind the unfortunate acknowledgment rate is blunder in character division. The fact that the scanned documents contain touching characters makes the process of segmentation even more difficult. As a result, designing an efficient method for character segmentation presents a significant challenge. A general OCR consists of preprocessing, character segmentation, feature extraction, classification, and recognition at the end.
The paper looks at the preprocessing tasks of converting grayscaled images to binary images, rectifying images, and segmenting the text of the document into paragraphs, lines, words, and then basic symbols. The neural classifier recognizes the fundamental symbols that were obtained as the fundamental unit through the segmentation process.
In this work, three element extraction procedures : histogram of projection in light of mean distance, histogram of projection in light of pixel worth, and vertical zero intersection, have been utilized to work on the pace of acknowledgment. Even distorted characters and symbols can have their features extracted using these powerful feature extraction methods. A back-propagation neural network with two hidden layers is used to build the neural classifier.
For printed Hindi texts, the classifier is trained and tested. It is possible to achieve a performance with a correct recognition rate of roughly 90%.

Keywords—OCR, Pre-processing, Segmentation, Feature Vector, Classification, Artificial Neural Network (ANN)

- **INTRODUCTION**

The technology known as optical character recognition (OCR) makes it possible to automatically recognise text in images, scanned documents, and other kinds of visual media. It is crucial for the digitization and preservation of cultural and historical documents written in a variety of languages. India's official language is Hindi, which is one of the world's most spoken languages. Characters are written in a continuous script in the Hindi script, which has a unique structure because there are no clear lines between words and characters. OCR systems face significant challenges due to the script's complexity.

To overcome these difficulties, scientists have created Hindi OCR models that utilize various methods like division, highlight extraction, and character acknowledgment. By digitizing and preserving important Hindi-language documents, these models have made them more accessible to a wider audience. Digital archives and libraries have also been made possible by the development of Hindi OCR models, making it easier for researchers to study the language and its culture.

Overall, the rich cultural heritage of the Hindi language has been digitized and preserved thanks to Hindi OCR models. In addition, they have provided new opportunities for education and research in the Hindi language and literature.

- **TYPES OF OCR :**

Basically, there are three types of OCR. They are briefly covered below:

1. Offline Text Written by Hand :

Offline Handwritten Text is text that has been written on paper with a pen or pencil by a person and afterwards the record has been digitised.

2. Online Handwritten Text :

found Online handwritten text is text created using a number of digital devices directly on a digital platform. The result is a list of x-y coordinates with the pen's position and other details like writing pressure and speed.

3. Machine-Printed Text :

Machine-printed texts are often delivered by offset techniques and are seen as being in printed records. Using optical character recognition, several sorts of documents, including PDF files and photos, can be scanned and turned into editable files.The OCR system is used for the following purposes:

1. Processing Bank cheese
2. Documenting library materials into digital format.
3. Storing documents in digital form, searching text and extracting data.

- **About Devanagari Script**

Devanagari script is an abugida script used to write several languages including Hindi, Marathi, Nepali, Sanskrit, and others. For OCR (optical character recognition) of Hindi language text, it is essential to have a model trained on the Devanagari script.

To train an OCR model for Hindi language text, the first step would be to collect a large and diverse dataset of Hindi language text written in Devanagari script. This dataset would be used to train a machine learning model, such as a convolutional neural network (CNN) or a recurrent neural network (RNN), to recognize and classify individual characters and words in the text.In addition to the dataset, it would be necessary to pre-process the images of the text to remove noise, correct skew, and enhance contrast, among other techniques. Once the model is trained, it can be used to recognize and transcribe Hindi language text written in Devanagari script from images or scanned documents.

It is important to note that OCR models trained on Devanagari script can also be used for other languages that use the same script, such as Marathi and Nepali. However, if the text is written in a different script, such as Urdu or Tamil, a different OCR model would need to be trained specifically for that script.

- **CHARACTERISTICS OF OCR**

Due to the complexity of the Hindi script, an OCR Hindi model has some unique properties that set it apart from OCR models for other languages. The following are some essential properties of an OCR Hindi model:

1. Hindi script support: The OCR Hindi model should be able to recognise and decipher the intricate Hindi script. The many ligatures and diacritical markings used in the Hindi script are included in this.

2. Accuracy: The OCR Hindi model ought to be capable of correctly identifying both printed and handwritten Hindi text. For the system to be dependable and practical for tasks like document digitization, this is essential.

3. Language models: In order to recognize the context of the text being recognized, the OCR Hindi model needs to have language models. This is crucial for handwritten text recognition since a character's context can be crucial in determining how it should be interpreted.

4. Text segmentation: The OCR Hindi model needs to be capable of dividing the text into separate characters, words, and lines. For the system to recognize text effectively, primarily when characters are written near together or overlap, it is crucial to do this.

5. Pre-processing: To increase the precision of text recognition, the OCR Hindi model should use pre-processing methods including noise reduction, contrast enhancement, and skew correction.

6. Machine learning algorithms: To increase the precision of text recognition, the OCR Hindi model should include machine learning techniques such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs).

7. Training data: A sizable and varied dataset of Hindi text, including both printed and handwritten text, should be used to train the OCR Hindi model. This will make it easier for the model to recognize text from many sources and with different levels of complexity.

8. Post-processing: To increase the precision of recognized text, the OCR Hindi model should use post-processing techniques such as error correction and normalization.

- **Devanagari Script from OCR Point of View:**

Devanagari is a two-dimensional script composed logically of its constituent symbols. The alphabet is used to write it. In Devanagari, letters are written evenly from left to right without upper- or lowercase. It comprises 36 simple consonants and 12 vowels. Additional Devanagari constituent symbols include the set of vowel modifiers known as Matra, which are positioned to the left, right, above, or below a character or conjunct, and pure consonants (sometimes called half letters), which, when joined with other consonants, generate conjuncts. Some of the Hindi vowels and consonants are shown in the figure.Figures 2(a) and 2(d) alone. The vowels are used in two different ways in Hindi and English: (i) They have the ability to produce their own sounds; Devanagari uses the vowels shown in Figure 2(a) for this purpose. (ii) They are employed to alter the sound of a consonant; for this purpose, the relevant modifier from Fig. 2(b) is suitably linked to the consonants. Some of the Hindi modifiers attached to the first consonant letter "ka" are seen in Fig.

अ आ इ ई उ ऊ ए ऐ ओ औ अं अः अँ ऋ

[a] Some of the Vowels

[b]Modifier Symbols

क का कि की कु कू

[c] The modifier symbols attached to the consonants,

क्य ख्य च्य ज्य त्य त्य

[d] Some sample conjuncts

Fig. 01. Characters and Symbols of Devanagari Script

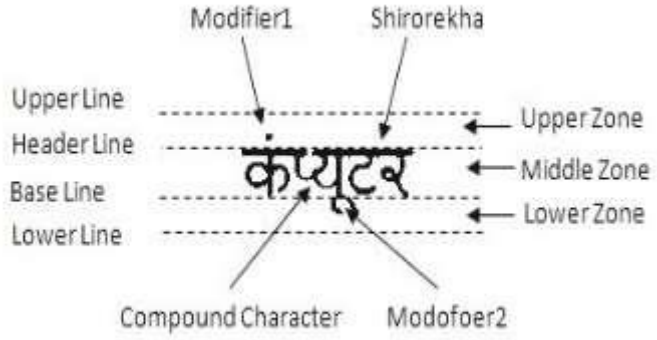

Fig: 02 Three Zones of Devanagari Script

2(c). The figure's visual inspection reveals that some modifier symbols are located next to the consonants (core modifier), above the consonants (top modifier), and below the consonants (lower modifier). There is a top modifier and a core modifier in some of the modifiers; The top modifier is above the core modifier, and the core modifier is placed before or next to the consonant. For nearly every consonant, there is a pure form known as a half character in the Devanagari script. A consonant in unadulterated structure generally contacts the following person happening in the content and consequently yielding conjuncts, contacting characters or melded characters. Fig. 2(f) shows a portion of the conjuncts framed by composing unadulterated structure consonants followed by consonants. The header line, or shirorekha, is a horizontal line that is drawn at the top of every character in a word. Three horizontal strips or zones can be used to represent Devanagari words: a top strip, bottom strip, and core strip The header line divides the middle and upper zones. Fig. 3 shows a word that comprises of three characters, one top modifier, and one lower modifier. The three strips and the header line have been checked.

- **Related Work:**

In the paragraphs that follow, we'll go over a brief synopsis of previous studies on OCR for Hindi. It was mentioned in the preceding section that Devanagari is the script for Hindi, India's national language. We also talked about how, in addition to Hindi, it can be written in Sanskrit, Marathi, Nepali, and many other Indian languages. Devanagari is gotten from antiquated Brahmi script through different changes. OCR for the Devanagari script has received significant research attention over the past forty years [4-5], [8-16], and [26-31]. However, there is currently no complete OCR for Devanagari that can be used in a noisy environment. Errors in character segmentation are a major contributor to an OCR system's low recognition rate [4]. Another significant obstacle encountered when developing an efficient character segmentation method is the presence of touching characters in the scanned documents. A novel method for identifying and segmenting touching characters is described in [4]. The method depends on fluffy multi factorial examination. In order to effectively select feasible cut columns for segmenting the touching characters, a predictive algorithm is developed. Devanagari and Bangla-language printed documents have been subjected to the proposed approach.

The method for character division treats characters contacting each other as a solitary character and prompts the disappointment in the person acknowledgment stage. Faxed reports, copies, old books, papers, and so on., contain an impressive number of contacting characters. A module for programmed division of contacting characters is fundamental for effective OCR of such Archives.

- **PROPOSED ARCHITETURE**

1. Input Image: The input to the OCR system is an image containing Hindi text.

2. Pre-processing: The input image undergoes pre-processing, which involves operations like noise reduction, contrast enhancement, and skew correction. This step prepares the image for segmentation and text recognition.

3. Segmentation: In this step, the pre-processed image is segmented into individual characters, words, and lines. This is important for accurate recognition of text, especially in cases where characters are written closely together or overlap.

4. Feature Extraction: The segmented text is then passed through a feature extraction module, which extracts features from the text to be used in the recognition process. This module can use techniques such as histogram of oriented gradients (HOG) or convolutional neural networks (CNNs) to extract features.

5. Character Recognition: The feature vectors generated from the previous step are then passed through a character recognition module that classifies the characters into their corresponding Hindi script.

6. Word and Line Recognition: Once the characters have been recognized, the OCR model then proceeds to recognize the words and lines of text. This can be done using a language model that takes into account the context of the recognized characters.

7. Post-processing: The recognized text then undergoes post-processing, which involves techniques such as error correction and normalization. This step aims to improve the accuracy of recognized text.

8. Output: The final output of the OCR model is the recognized text, which can be saved in a digital format for further processing or analysis.This is a high-level description of the flow diagram for a Hindi OCR model. The exact implementation and details of the different steps can vary depending on the specific OCR model and the techniques used.

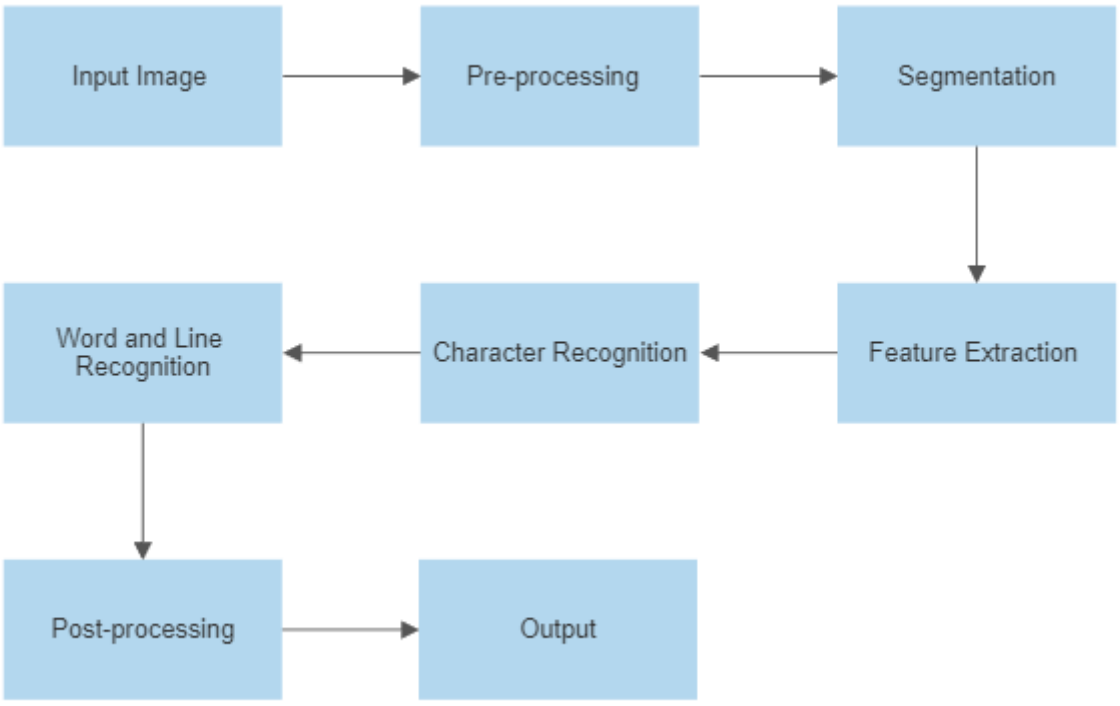

Fig : 03 OCR Flow diagram

- **CONCLUSION**

Over the years, the Hindi OCR (Optical Character Recognition) system has advanced significantly. With advances in technology and machine learning, OCR systems for the Hindi language have improved in their ability to recognise and translate printed or handwritten text with greater accuracy and efficiency.

In conclusion, Hindi OCR systems have enormous potential for use in a variety of industries, including education, business, and research. However, there is still room for improvement in terms of precision and effectiveness, particularly for the identification of handwritten text. Future OCR systems for the Hindi language are likely to be even better as a result of ongoing research and development in this area.

- **REFERENCES**

[1] Mori, S. et. al.: Historical Review of OCR Research and Development. Proceeding IEEE, Vol.80,No.7, 1992, pp. 1029-1058.

[2] Chaudhari, A.A., Ahmad, E.A. S., Hossain, S., Rahman, C.M.: OCR of Bangla Character Using Neural Network: A better approach. 2nd International Conference on Electrical Engineering (ICEE 2002.),khuln, Bangladesh, 2002.

[3] Mahmud, J.U.; Raihan, M.F., Rahman, C.M.: A complete OCR System for Continuous Bengali Char- acters. TENCON 2003, IEEE conference on convergent Technologies for Asia-Pacific Region Vol.4,2003, pp.1372-1376.

[4] Garain, U., Chaudhuri, B. B.: Segmentation of Touching Character in Printed Devanagari and Bangla Script Using Fuzzy Multifactorial Analysis. IEEE Transaction on System, Man and Cybernetics -PartC: Applications and Reviews, Vol.32, No.4, 2002, pp.449-459.

[5] Jawahar, C.V., Pavan Kumar, M.N.S.S.K., Ravi Kiran, S.S.: A Bilingual OCR for Hindi-Telugu Documents and its Application. Document Analysis and Recognition. IEEE Proceedings Seventh In- ternational Conference on, Vol. 1, 2003, pp.408-412.

[6] Lakshmi, C.V., Patvardhan, C.: A High Accuracy OCR System for printed Telugu text. Conference on Convergent Technology for Asia-pacific Region Vol.2, 2003, pp.725-729.

[7] Ashwin, T.V., Sastry, P.S.. A Font and size independent OCR for printed Kannad documents usingsupport vector machines.