IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 5, May 2024

DOI: 10.17148/IJARCCE.2024.13510

Kannada Handwritten Optical Character Recognition

Mr. Dadapeer[1], Abhishek Alwandi[2], Amanath Rasool[3], Bhuvana[4], K S Mallika Begum[5]

Asst. Prof, Department of Computer Science and Engineering (CSE), Ballari Institute of Technology and Management, Ballari, India[1]

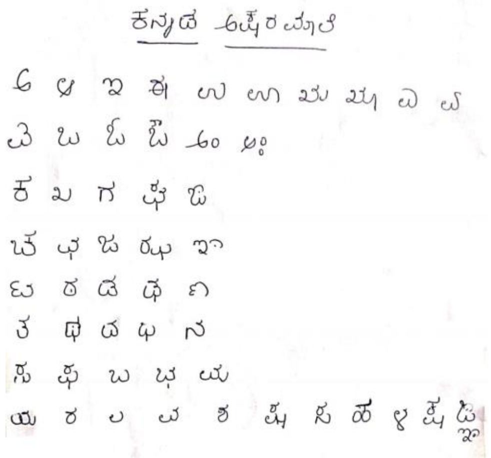
Bachelor of Engineering (CSE), Department of Computer Science and Engineering (CSE),
Ballari Institute of Technology and Management, Ballari, India[2-5]

Abstract: Optical character recognition (OCR) technology place a vital role in converting handwritten text into digital format. This technology explores the development and implementation of OCR system to extract text from various source such as handwritten images and printed images.OCR is one of the challenging topics in the character recognition field. The process begin with the giving an input as image. First the image goes under the pre-processing technique to remove the noise, image obtained is processed to identify the required lines. The identified lines ae extracted using segmentation process. The model is used using Convolutional Neural Network technique.

Keywords: Convolutional neural network, handwritten character recognition.

INTRODUCTION

Now-a-days large amount of characters is stored in images, thus the optical character recognition has obtained popularity, the image processing is mainly about extraction of characters from the image. The process of identifying and recognition of text from image is called optical character recognition (OCR). Handwritten Optical Character Recognition includes four stage of processing those includes Input the image, Pre-processing, Segmentation process, Extraction of individual text(Result). Each stage of result is depending upon the result of previous stage. Kannada language is combination of 52letters, these letters are of 3 categories they are: Swara includes 16letters, Vyanjanas includes 38 letters, and Yogavaka includes 2 letters.



The most difficult challenge is recognizing the character because every individual has their own style of handwriting.

Everytime the handwritten text will not be same as the before. The focus of this work is to recognize the user -uploaded handwritten Kannada image and output as digital text. The accuracy of recognizing the handwriting is higher in case of writer dependent because the text from an individual writer are used to train the model. Accurate text-line segmentation is essential for script identification at line level.Text-line segmentation is very crucial step in OCR, Poor line-segmentation leads to wrong result in recognition. In printed text, line-segmentation is easy but in handwritten text, it is difficult due to problems like overlapping, touching of characters and also due to different writing style of a writer.

DOI: 10.17148/IJARCCE.2024.13510

Image processing module

The purpose of image processing is divided into 5 groups they are:
1. Visualization- Observe the objects that are not visible.
2. Image sharpening and restoration- To create a image.
3. Image retrieval- Seek for the image of interest.

4. Pattern measurement– Measures various objects in animage.
5. Image Recognition– Recognize the objects in an image.

LITERATURE SURVEY

[1].V.Papavauilion, N.Stafylakis, V.Katasouros, G.Carayannis. This paper explains "handwritten document image segmentation into text line and words", here two approaches to extract text lines and words from handwritten document are presented.

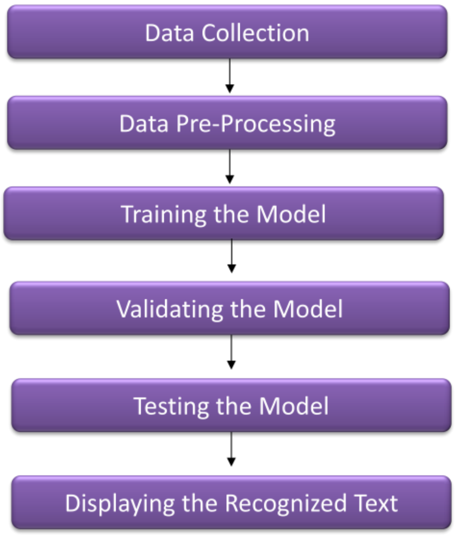
[2]. G. G. Rajput Suryakant B. UmmapurePreethi N. Patil
This paper explains text-line Extraction from Handwritten Document images using Histogram and Connected Component Analysis. An efficient two stage method has been proposed to extract text-lines from handwritten document images. Extraction of text-lines from document images with lines appearing curved (oriented) poses difficulty in segmentation. Connected Component Analysis is used on adjacent lines, in the second stage, wherein on pixel lines of these text lines overlap.

[3].B. Gatos, I. Pratikakis and S.J. Perantonis This paper explains Improved Document Image Binarization by Using a Combination of Multiple Binarization Techniques and Adapted Edge Information. The main novelties of the proposed document image binarization approach consist of (i) combining the binarization results of several state-of-the-art methodologies; (ii) incorporating the edge map of the grey scale image; and (iii) applying efficient image post-processing based on mathematical morphology for the enhancement of the final result.

[4]. K.S. seshkumar, Z.razak et.al clarified with difficulty in separating texts in kannada in text lies in the local distribution of connecting part. Methods based on the topic line detection of baseline and line separation of handwritten characters in kannada.

[5]. K. A. Hamad, Maira Sami and Rizwan Ahmed Khan This paper explains.The OCR system can be used in different practical applications such as number-plate recognition, smart libraries and various other real-time applications. OCR systems use a variety of approaches, but most focus on one letter, phrase, or block of text at a time.Text is copied or read using hardware, while further processing is usually handled by software.

METHODOLOGY



DOI: 10.17148/IJARCCE.2024.13510

- Tesseract is an open source text recognition (OCR) Engine, available under the Apache 2.0 license. It can be used directly, or (for programmers) using an API to extract printed text from images. It supports a wide variety of languages.
- Tesseract is compatible with many programming languages and frameworks
- It can be used with the existing layout analysis to recognize text within a large document, or it can be used in conjunction with an external text detector to recognize text from an image of a single text line.
- Tesseract includes a new neural network subsystem configured as text line recognizer.
- To recognize an image containing a single character, a Convolutional Neural Network (CNN) is used.

The following stages are involved:

Input Document: In this stage we have gathered the images that are required for our experiment. All of these images were collected from old historical places. These images consist of old age Kannada letters or word. The orientation and the size of letter vary in each of the image. The difference in the size and shape of letters, their orientation have a huge impact on identifying the text line. This whole project is identifying the lines in a handwritten document.

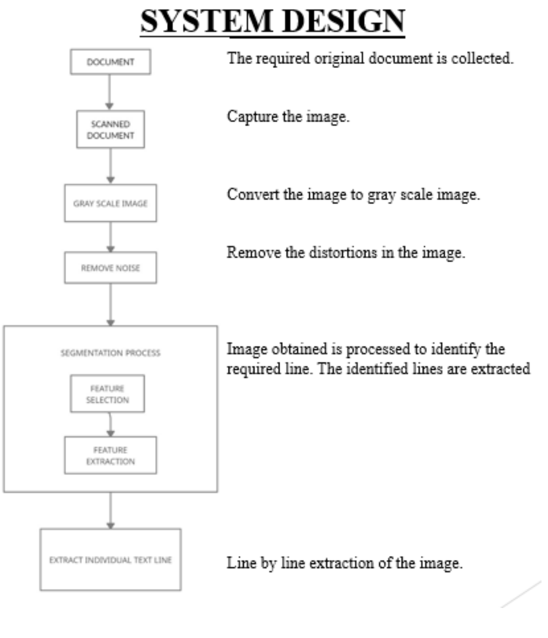
Pre-processing: Old documents tend to have imperfections which has to be removed called noise removal, and then the image is converted to a binary format. The aim of pre-processing is an improvement of the image data that suppresses unwanted distortions or enhances some image features important for further processing. Image pre-processing methods use the considerable redundancy in images. Pre-processing operations in document image analysis transform the input image into an enhanced image more suitable for further analysis.

Segmentation of Line: Segmentation is a process of partitioning image used to representation it in meaningful way. It is the process of extracting object of interest from an image. It subdivides an image into it's constitutes regions or object, which are certainly characters. Segmentation phase is also crucial in contribution to the error due to touching characters, even in good quality documents, some adjacent character touch each other due to inappropriate scanning resolution.

Feature Selection: It is the process where you manually select those features. output in which you are interested in having irrelevant features in our data can decrease the accuracy of the models and make your model learn based on irrelevant features.

Feature Extraction: The input image is used by the feature extraction using CNN network. The extracted feature network are utilized by CNN for classification. Then by using pytesseract we obtained line by line extraction of the handwritten text.

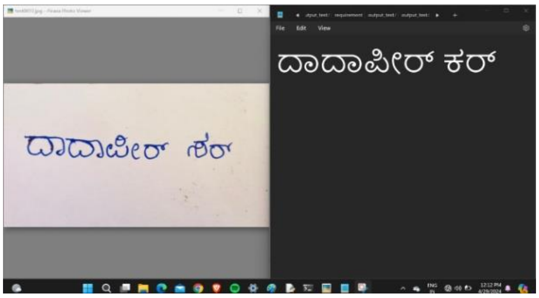
SYSTEM ARCHITECTURE


SYSTEM DESIGN
DOCUMENT
The required original document is collected.
SCANNED DOCUMENT
Capture the image.
GRAY SCALE IMAGE
Convert the image to gray scale image.
REMOVE NOISE
Remove the distortions in the image.
SEGMENTATION PROCESS
FEATURE SELECTION
FEATURE EXTRACTION
Image obtained is processed to identify the required line. The identified lines are extracted
EXTRACT INDIVIDUAL TEXT LINE
Line by line extraction of the image.

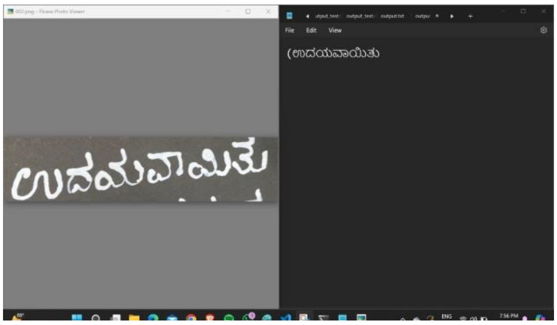
DOI: 10.17148/IJARCCE.2024.13510

- This section explains about the flow of this project. There are number of steps that has to be performed in order to identify the text lines in a Kannada document.
- The first step is data collection, next step is to scan the document and convert it to gray scale image then it undergoes noise removal process in the image.
- The image after noise removal it is subjected to segmentation stage.
- In the last step each of the identified lines are highlighted.

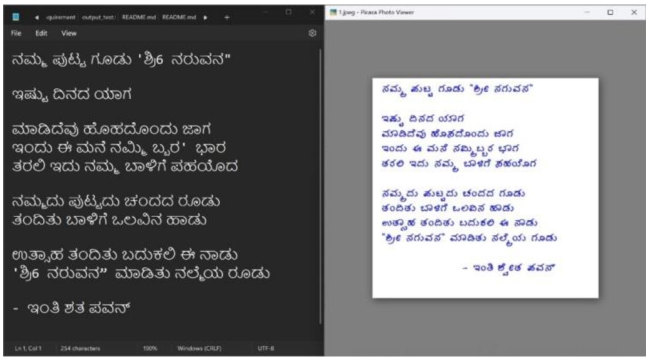
RESULTS



ದಾದಾಪೀರ್ ಕರ್



(ಉದಯವಾಯಿತು



ನಮ್ಮ ಪುಟ್ಟ ಗೂಡು 'ಶ್ರೀ6 ನರುವನ"

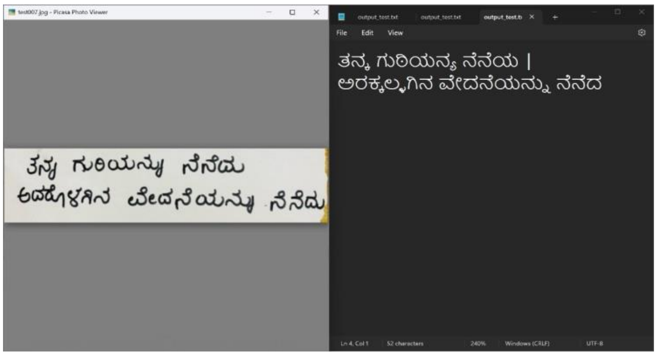
ಇಷ್ಟು ದಿನದ ಯಾಗ

ಮಾಡಿದೆವು ಹೊಹದೊಂದು ಜಾಗ
ಇಂದು ಈ ಮನೆ ನಮ್ಮಿ ಬ್ಬರ' ಭಾರ
ತರಲಿ ಇದು ನಮ್ಮ ಬಾಳಿಗೆ ಪಹಯೊದ

ನಮ್ಮದು ಪುಟ್ಟದು ಚಂದದ ರೂಡು
ತಂದಿತು ಬಾಳಿಗೆ ಒಲವಿನ ಹಾಡು

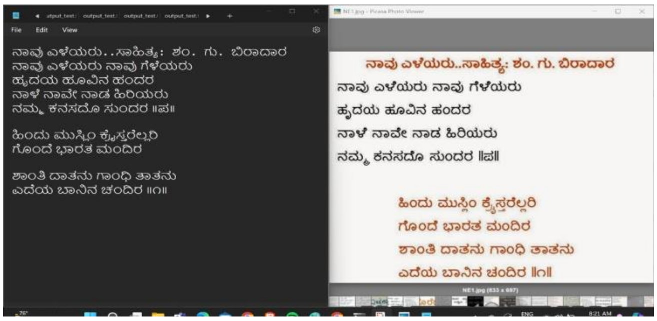
ಉತ್ಸಾಹ ತಂದಿತು ಬದುಕಲಿ ಈ ನಾಡು
'ಶ್ರೀ6 ನರುವನ" ಮಾಡಿತು ನಲ್ಮೆಯ ರೂಡು

- ಇಂತಿ ಶತ ಪವನ್



ತನ್ಕ ಗುರಿಯನ್ಯ ನೆನೆಯ |
ಅರಕ್ಕಲ್ಯಗಿನ ವೇದನೆಯನ್ನು ನೆನೆದ

DOI: 10.17148/IJARCCE.2024.13510



```
ನಾವು ಎಳೆಯರು..ಸಾಹಿತ್ಯ: ಶಂ. ಗು. ಬಿರಾದಾರ
ನಾವು ಎಳೆಯರು ನಾವು ಗೆಳೆಯರು
ಹೃದಯ ಹೂವಿನ ಹಂದರ
ನಾಳೆ ನಾವೇ ನಾಡ ಹಿರಿಯರು
ನಮ್ಮ ಕನಸದೊ ಸುಂದರ ॥ಪ॥

ಹಿಂದು ಮುಸ್ಲಿಂ ಕ್ರೈಸ್ತರೆಲ್ಲರಿ
ಗೊಂದೆ ಭಾರತ ಮಂದಿರ

ಶಾಂತಿ ದಾತನು ಗಾಂಧಿ ತಾತನು
ಎದೆಯ ಬಾನಿನ ಚಂದಿರ ॥೧॥
```

ನಾವು ಎಳೆಯರು..ಸಾಹಿತ್ಯ: ಶಂ. ಗು. ಬಿರಾದಾರ

ನಾವು ಎಳೆಯರು ನಾವು ಗೆಳೆಯರು
ಹೃದಯ ಹೂವಿನ ಹಂದರ
ನಾಳೆ ನಾವೇ ನಾಡ ಹಿರಿಯರು
ನಮ್ಮ ಕನಸದೊ ಸುಂದರ ॥ಪ॥

ಹಿಂದು ಮುಸ್ಲಿಂ ಕ್ರೈಸ್ತರೆಲ್ಲರಿ
ಗೊಂದೆ ಭಾರತ ಮಂದಿರ
ಶಾಂತಿ ದಾತನು ಗಾಂಧಿ ತಾತನು
ಎದೆಯ ಬಾನಿನ ಚಂದಿರ ॥೧॥

CONCLUSION

In conclusion, our Kannada Handwritten OCR project signifies a notable leap forward in optical character recognition, specifically tailored for Kannada. Through rigorous development and testing, we have showcased the efficacy of machine learning techniques in accurately transcribing handwritten Kannada characters into digital text. Beyond technical achievements, our project underscores the imperative of preserving linguistic diversity, thereby contributing to the accessibility of Kannada language resources. While further enhancements in scalability and performance are warranted, our endeavour holds promise for streamlining document digitization and linguistic research workflows. Ultimately, it reflects our commitment to advancing language technology and promoting cultural heritage in the digital age.

REFERENCES

[1] G. G. Rajput Suryakant B. Ummapure Preethi N. Patil, “Text-Line Extraction from Handwritten Document images using Histogram and Connected Component Analysis” ,International Journal of Computer Applications (0975-8887) National conference on Digital Image and Signal Processing, DISP 2015

[2] G. Louloudis, C. Halatsis, “Line And Word Segmentation of Handwritten Documents”, Department of Informatics and Telecommunications, University of Athens, Greece.

[3] B. Gatos , I. Pratikakis , “Line And Word Segmentation of Handwritten Documents”, Computational Intelligence Laboratory, Institute of Informatics and Telecommunications, National Center for Scientific Research “Demokritos”, GR-153 10 Agia Paraskevi, Athens, Greece

[4] B. Gatos, I. Pratikakis and S.J. Perantonis, “Improved Document Image Binarization by Using a Combination of Multiple Binarization Techniques and Adapted Edge Information”, Computational Intelligence Laboratory, Institute of Informatics and Telecommunications, National Research Center "Demokritos". 153 10 Athens, Greece.

[5] T.M.Rath and R. Manmatha, “Word spotting for historical documents”, International Journal on Document Analysis and Recognition (IJDAR), Vol.9, No 2 – 4, pp. 139 – 152 , 2006.

[6] V. Lavrenko, T. M. Rath, R. Manmatha: “Holistic Word Recognition for Handwritten Historical Documents”, Proceedings of the First International Workshop on Document Image Analysis for Libraries (DIAL'04),pp 278-287, 2004.

[7] T. Adamek, N. E. O’Connor, A. F. Smeaton, “Word Matching Using Single-Closed Contours for Indexing Handwritten Historical Documents”, International Journal on Document Analysis and Recognition (IJDAR), special Issue on Analysis of Historical Documents, 2006.

[8] T.M.Rath and R. Manmatha, “Word spotting for historical documents”, International Journal on Document Analysis and Recognition (IJDAR), Vol.9, No 2 – 4, pp. 139 – 152 , 2006.

[9] V. Lavrenko, T. M. Rath, R. Manmatha: “Holistic Word Recognition for Handwritten Historical Documents”, Proceedings of the First International Workshop on Document Image Analysis for Libraries (DIAL'04),pp 278-287, 2004.

[10] T. Adamek, N. E. O’Connor, A. F. Smeaton, “Word Matching Using Single-Closed Contours for Indexing Handwritten Historical Documents”, International Journal on Document Analysis and Recognition (IJDAR), special Issue on Analysis of Historical Documents, 2006.

[11] T.M.Rath and R. Manmatha, “Word spotting for historical documents”, International Journal on Document Analysis and Recognition (IJDAR), Vol.9, No 2 – 4, pp. 139 – 152 , 2006.

[12] V. Lavrenko, T. M. Rath, R. Manmatha: “Holistic Word Recognition for Handwritten Historical Documents”, Proceedings of the First International Workshop on Document Image Analysis for Libraries (DIAL'04),pp 278-

DOI: 10.17148/IJARCCE.2024.13510

287, 2004.

[13] T. Adamek, N. E. O'Connor, A. F. Smeaton, "Word Matching Using Single-Closed Contours for Indexing Handwritten Historical Documents", International Journal on Document Analysis and Recognition (IJDAR), special Issue on Analysis of Historical Documents, 2006.

[14] J. Canny, "A computational approach to edge detection", IEEE Trans. PAMI, 8: 679-698

[15] F.M. Wahl, K.Y. Wong, R.G. Casey, "Block Segmentation and Text Extraction in Mixed Text/Image Documents", Comp. Grap. and Im. Proc., pp. 375-390.