iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** IX **Month of publication:** September 2022

DOI: https://doi.org/10.22214/ijraset.2022.46520

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue IX Sep 2022- Available at www.ijraset.com

Kannada Text Recognition

Anjali Yogesh Devaraj[1], Omisha N[2], Anup Jain[3], Shobana TS[4]
[1, 2, 3, 4]*Dept. of Information Science and Engineering, BMS College of Engineering*

Abstract: ***The task of automatic handwriting recognition is critical. This can be a difficult subject, and it has gotten a lot of attention in recent years. In the realm of picture grouping, handwritten character recognition is a problem. Handwritten characters are difficult to decipher since various people have distinct handwriting styles. For decades, researchers have been focusing on character identification in Latin handwriting. Kannada has had fewer studies conducted on it. Our "Kannada Text Recognition" research and effort attempts to classify and recognize characters written in Kannada, a south Indian language. The characters are taken from written documents, preprocessed with numpy and OpenCV, and then run through a CNN.***
Keywords: Optical Character Recognition, Convolutional Neural Networks, Segmentation, Image Processing, Handwriting Recognition.

I. INTRODUCTION

One of the computer-related challenges being addressed and investigated is how to detect and classify a scanned picture or document. Our project "Kannada Text Recognition" aims to recognise handwritten Kannada text using optical character recognition (OCR) by uploading an image which contains handwritten text. Several people have been working on character recognition as a subset of pattern recognition issues since the beginning of computer vision. The applications of automated character recognition are broader today than they were a few years ago, thanks to the vast availability of cameras. In Karnataka, reports hand-written in Kannada are the sole form of documentation available in government offices and government hospitals. Trying to type and reproduce these documents would be a tedious and time-consuming task; because these documents are very hard to read and understand by people, especially those who aren't familiar with the language. Thus there comes a necessity for a system that is computerized to act as a bridge between machines and humans. We have proposed a method that would efficiently recognize handwritten Kannada characters. The system uses image pre-processing tools to boost the quality of the image and then use deep learning techniques for performing feature selection.

The Chars74K dataset is used for training. The data set contains various handwritten Kannada alphabets, vowels and consonants consisting of 25 handwritten characters in 657 classes.

II. MOTIVATION AND OBJECTIVES

The majority of documents in government offices are currently hard copies. Being able to digitize this will improve the efficiency of these offices and make data transfer across networks easier.

Because of the vast number of classes created by letters, numerals, kagunitas, and ottaksharas, the complexity is high. As we all know, many individuals in Karnataka these days come from various regions of the country and struggle to read and write Kannada. Even simple duties would be difficult for such persons in government agencies.

In recent years, machine learning, particularly deep learning, has become more significant in artificial intelligence and computer vision. Natural language processing, plant detection, human action recognition, picture segmentation, and image classification have all been studied using learning-based methods. To classify handwritten characters, we used a deep learning technique in this research and project. Convolutional neural networks (CNN) are used to build an architecture from the ground up.

Segmentation, feature extraction, and classification are three machine learning algorithms that are used to recognise characters. A regional language like Kannada has a lot of unusual characters, making it challenging to extract those characters from a scanned document. The proposed system is concerned with the identification and classification of Kannada characters. We have created a system for character classification and recognition in a picture of Kannada text that is primarily handwritten using convolutional neural networks. We began by cleaning up the scanned image by removing noise and then cropping and resizing each character image. We were able to train the CNN for accurate classification into their respective classes thanks to the additional pixels in the image after the pre-processing procedure. Handwritten vowels, consonants, and digits in Kannada are taken into account.

Despite the structural similarities between several simple Kannada characters, the model is able to classify them into their relevant classes.

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue IX Sep 2022- Available at www.ijraset.com

III. LITERATURE SURVEY

In [1] Kannada Handwritten Characters is a challenge in the pattern recognition field because of complicated structure and different handwriting styles. In this research, they propose a technique based on convolutional neural networks and transfer learning for recognising handwritten characters (vowels, consonants, and numbers) in kannada. The model is capable of assigning the character to the appropriate class. After training the created CNN model, the model achieved 86.92 percent validation accuracy.
In [2], two strategies for recognising handwritten Kannada writing are proposed, with great accuracy when compared to prior efforts. The first technique uses the Tesseract tool, while the second employs the Convolution Neural Network (CNN). The Tesseract tool provided 86 percent accuracy, whereas the Convolution Neural Network provided 87 percent accuracy.
In [3], they provide a segmentation-free OCR system that integrates deep learning approaches, synthetic training, data generation, and data augmentation techniques. They represent the interactions between input items using both recurrent and convolutional neural networks. The proposed designs, they argue, outperform leading commercial and open-source engines.
This study [4] employs Machine Learning, Deep Learning, and Transfer Learning techniques for handwriting identification. Data Augmentation has a substantial impact on model performance results. Using the Inception model with both data augmentation and transfer learning, the best recognition rate of 90.27 percent on Dig-MNIST was attained. One major computer-related issue is how to recognise and classify images. In [5], approaches for handwriting recognition are presented. For handwriting recognition, methods such as Convolutional Neural Network Method, Semi Incremental Method, Part-Based Method, and others are utilized. The Convolutional Neural Network approach achieves the maximum accuracy (CNN). Many applications in diverse fields rely on automatic handwriting recognition. Many methods have been used to study the problem of handwriting recognition. Support vector machines (SVMs), K-nearest neighbors (KNNs), neural networks (NNs), and convolutional neural networks are some of these methods (CNNs). The study in [6] was carried out in order to recognise Arabic handwriting. The average test set accuracy was 97 percent, the precision was 96.78 percent, the recall was 96.73 percent, and the F1 score was 96.73 percent. This paper [7] discussed how OCR, Deep Learning, and Segmentation are components of OCR and are used to decipher Telugu text. Different OCR strategies are contrasted, and it is also demonstrated by surveys how OCR evolves as the model that outperforms other machine learning models.

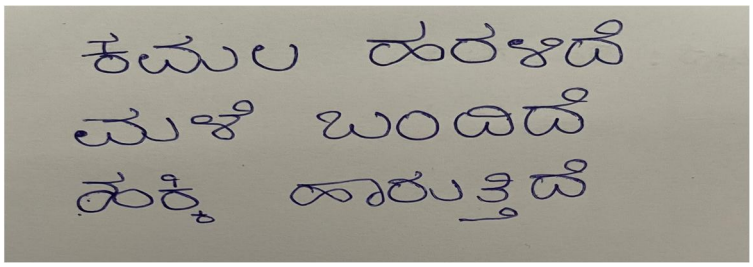
IV. METHODOLOGY

We aim to create a system for reading, classifying, and recognizing handwritten Kannada scanned images.
Four functional modules are included in the system:

- Pre-processing, which converts an image into lower and easier pixel values and suitable sizes for easier manipulation of large page images and image handling procedures, this includes segmentation, denoising, and augmentation;
- Line separation, text line detection, and extracting images from individual lines of text from a scanned document;
- Word segmentation, that includes finding gaps and isolating words from a line of text image, which is followed by character segmentation, that is done using methods contained in OpenCV;
- Character classification and recognition, concerning handwritten text recognition algorithms such as CNN, K-nearest Neighbors, and SVM.

The basic concepts employed in each model and designed for dealing with the uniqueness of each individual's handwriting in its numerous characteristics with the purpose of system robustness and dependability are discussed in detail below.

1) *Data Collection:* We have used the dataset Chars74K. It is a collection of multiple images from various sources. These characters of the dataset are related to over 650 classes and each of those classes had 25 characters that are handwritten, each.
2) *Image Preprocessing:* It plays a very important role in increasing the quality of the image, which is very important in Computer Vision. First, the input is an image as shown below.



©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

The image could be a scanned document and the output obtained would be the characteristics associated with it. When the training of a CNN is carried out with raw images, it might lead to classification. Hence image Pre-processing steps play a very crucial role in deep learning. The steps are converting image to grayscale, then removing noise, followed by contrast normalization, then binarization, and finally segmentation.

a) In pre-processing, first, the grayscale conversion is done. This step involves converting coloured images into grayscale images. A gray-level image generally contains 256 shades of a gray color. That gray color needs to be converted to black and white. It plays an important role in reducing the complexity of the image. This is done by converting the 3D pixel value of the image to its 1D value.

b) The next step is Denoising. Since the input documents are a culmination of various sources, they might be subjected to several noises. Therefore de-noising will further help in reducing the noise from the image. This plays an important role in handwriting recognition from the scanned documents.

c) Next comes the process of Binarization, in which the grayscale image is further converted into a binary image. This is the process in which a grayscale image which consists of 256 shades of gray is converted into black and white, which becomes a binary image.

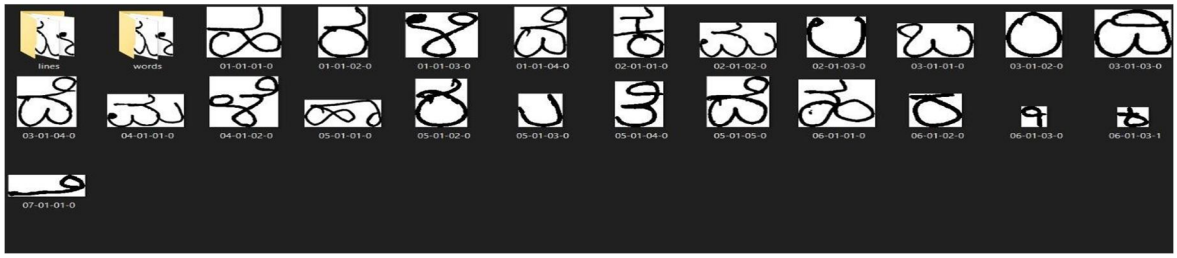
3) Segmentation: This divides an image into regions or objects. Input is first broken down into separate lines and then these lines are separated into individual words which are then segmented into singular characters using various techniques such as line, word, and character segmentation. For analyzing and recognizing the objects in the image, we use contour as it is an efficient tool. We have used the OpenCV contour function to recognize the separate lines, words, and characters in the input image.

There are three steps used in the segmentation process, that is line segmentation, word segmentation, and then character segmentation.

a) Line Segmentation: here images are into line segments. We used the OpenCV findContour method to help us find the contour. After finding the contour, the contour will be extracted and then saved. Then the lines found in the picture will be cropped out from the input image and it is then stored into the line segment folder. To crop the image, we will be using the Bounding Rectangle method which is found in the OpenCV library.

b) Word Segmentation is done later. In Word Segmentation the line segmented part will be further divided into separate words. Every word will be recognized based on the counters as an image. It is then cropped from the original image and stored in a separate folder.

c) Character Segmentation contour is applied for both axes because there is the possibility that an ottakshara might be present. Every character is recognized and then cropped and stored into the character segment folders.



The above image shows how the segmented characters are stored in a separate file.

4) Augmentation: Image augmentation is a technique for modifying existing images in order to generate extra data for the model training process. To put it another way, it is the technique of artificially increasing the available dataset for training a deep learning model.

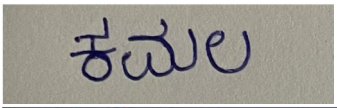


ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue IX Sep 2022- Available at www.ijraset.com

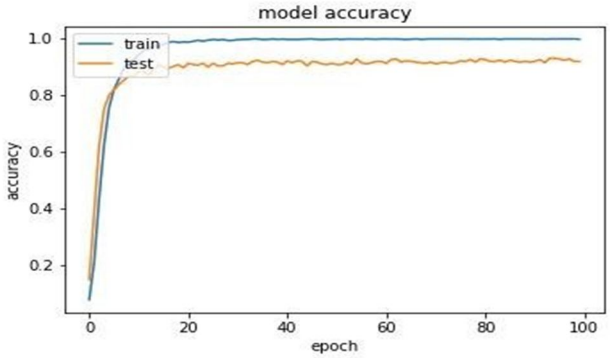
Later, relevant features that are suitable for the classification were selected for which feature extraction CNN was used. That includes convolution layer, max pooling, flattening, and fully connected layers.
Convolutional Neural Networks have major applications in image recognition, specifically when colored images are given as input.

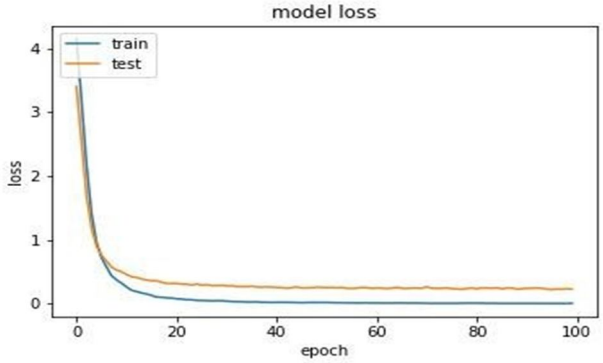
V. RESULTS

The system works with over 90% accuracy for simple Kannada text. The accuracy for character recognition is very high.When ottaksharas and dheergas are introduced as the input the accuracy reduces.









VI. SCOPE

The purpose of this system is to convert already existing Kannada documents into digital format that can be copied and pasted easily. Currently there is no way to send Kannada text other than by sending the image itself or typing it by hand which is why our system is revolutionary. This project can be used extensively in Karnataka government offices and hospitals. All documents dating back several years are in Kannada, and there are no soft copies of these documents. With the help of this project, they can be digitized, boosting the efficiency of such offices. In remote parts of Karnataka where Kannada is the primary source of communication, understanding and using English to text or send messages is impossible. Having an app that can copy paste text would make life a lot easier. Due to the pandemic school children even in villages are forced to buy smartphones to attend classes and tests and teachers need to send material as well. Although the course is taught primarily in Kannada it means the teacher would have to write all the material and send images to the students. Being able to copy and paste the Kannada text will definitely help the teachers simplify this process.
A lot of research is now being conducted in the subject of OCR. This project will help to facilitate future study in this area. Demonstrations of the universal applications of hyperparameter optimization methods could be carried.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

VII. CONCLUSION AND FUTURE ENHANCEMENT

Using CNN (Convolutional Neural Networks) in OCR (Optical Character Recognition) systems has been shown to be a fairly reliable method of converting an image to a text document, which requires the document to be accurately categorized. Due to the lack of precision and increased runtime, template matching proved to be an outmoded technique in this circumstance.

We were able to build an OCR system for Kannada character categorization and recognition with an accuracy of more than 90% at the end of this project. The technology is intended to function solely with printed or handwritten documents containing Kannada characters. Right now the system works well for characters and simple words. Further research and implementation needs to be done to increase the accuracy for complex words and subsequently for paragraphs.

This project could be improved further by developing it in real time so that users can scan and digitize photographs while on the road. A translation tool could be added so that the users could directly translate scanned documents in the same interface.

VIII. ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of this Capstone Project Phase-2 would be incomplete without the mention of the people who made it possible through constant guidance and encouragement. We would like to thank our parents and friends who supported us throughout.

We wish to express sincere thanks to our guide Prof. Shobhana TS, Assistant Professor, Department of Information Science and Engineering for helping us throughout and guiding us from time to time. We wish to express our deepest gratitude and thanks to Prof. Soumya Lakshmi BS, Assistant Professor,

Machine Learning for her help in executing this project and her constant support.

REFERENCES

[1] H. Parikshith, S. Naga Rajath, D. Shwetha, C. Sindhu and P. Ravi, "Handwritten Character Recognition of Kannada Language Using Convolutional Neural Networks and Transfer Learning", 2021.

[2] R. Fernandes and A. P. Rodrigues, "Kannada Handwritten Script Recognition using Machine Learning Techniques," 2019 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER), 2019, pp. 1-6, doi:10.1109/DISCOVER47552.2019.9008097.

[3] Namysl, Marcin, and Iuliu Konya. "Efficient, lexicon-free OCR using deep learning." 2019 International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2019.

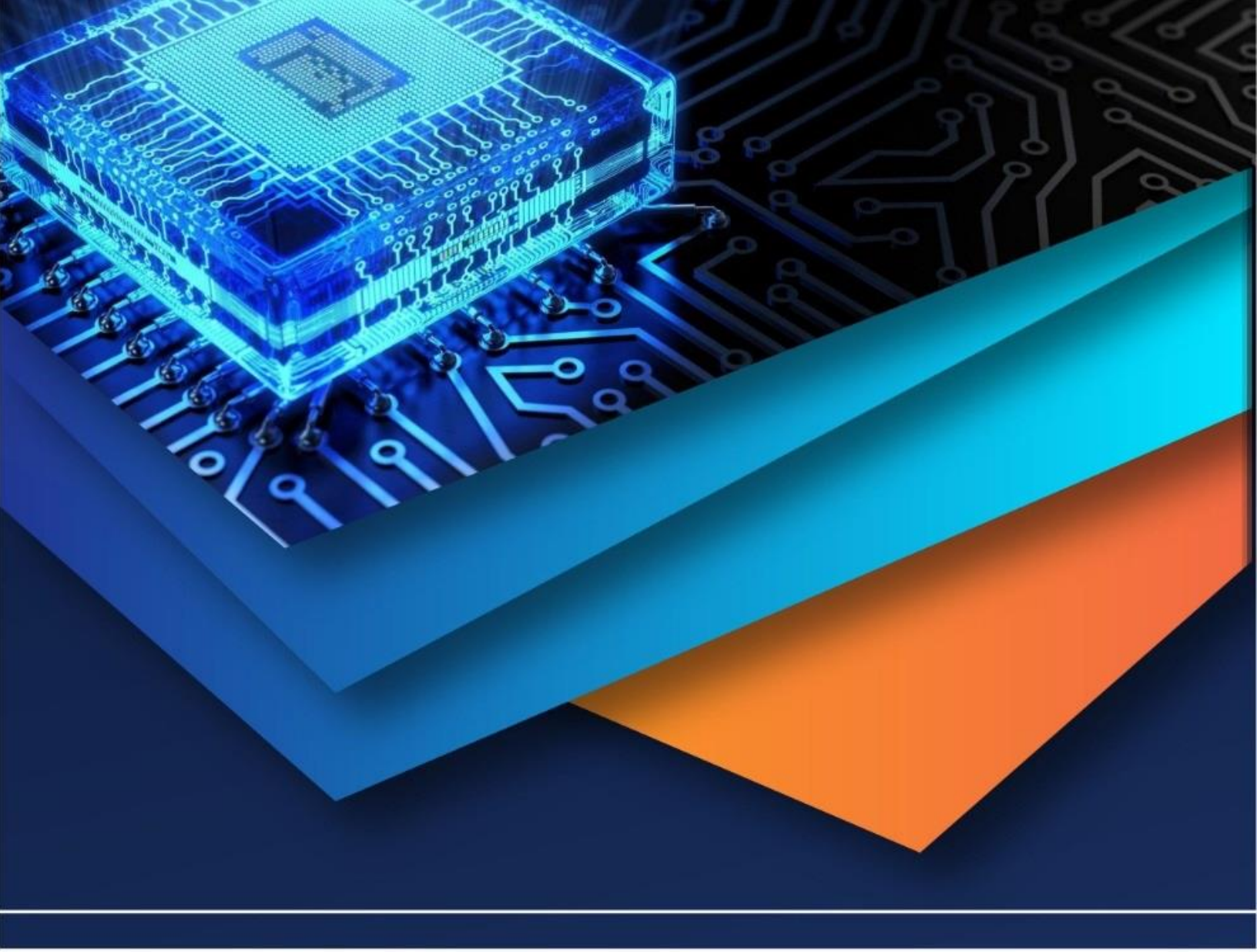
[4] Q. Hu, "Evaluation of Deep Learning Models for Kannada Handwritten Digit Recognition," 2020 International Conference on Computing and Data Science (CDS), 2020, pp. 125-130, doi: 10.1109/CDS49703.2020.00031.

[5] A Review of Various Handwriting Recognition Methods By Salma Shofia Rosyda and Tito Waluyo Purboyo

[6] Altwaijry, N., Al-Turaiki, I. Arabic handwriting recognition system using convolutional neural network. Neural Comput & Applic 33, 2249–2261 (2021).

[7] M. S. Velpuru, P. Chatterjee, G. Tejasree, M. R. Kumar and S. N. Rao, "Comprehensive study of Deep learning-based Telugu OCR," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), 2020, pp. 1166-1172, doi:10.1109/ICSSIT48917.2020.9214087.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)