PROJECT REPORT

ON

HANDWRITTEN MARATHI TEXT OCR

Submitted by

UTKARSH PANDEY (U23AI040)
ARYAN SAWANT (U23AI042)
DEVANG VALA (U23AI044)
JEET TANDEL (U23AI045)
KRISH RATHOD (U23AI049)

DEPARTMENT OF ARTIFICIAL INTELLIGENCE
SARDAR VALLABHBHAI NATIONAL INSTITUTE OF TECHNOLOGY SURAT
395007

Abstract

The task consists of handwritten sentence recognition in the Marathi language using a set of OCR model-based techniques. The processes include preprocessing the input images, segmenting into words and characters, and later getting the prediction from the OCR model. Thus, our solution provides a robust optical character recognition system toward the Marathi language. For the OCR, a convolutional neural network (CNN) was used, testing with a remarkable accuracy of 98.79

Additionally, this project tackles many other challenges such as variability in handwriting, the complexity of the Marathi script, and noise in the input images. We employ techniques of thorough preprocessing, segmentation, and recognition to battle these challenges. The system can be further extended toward other Indian languages and adapted for different handwriting styles, thus adding one more feather in the crown of OCR research.

1 Introduction

1.1 Background and Motivation of the Problem

Optical Character Recognition (OCR) refers to the technology that translates pictures of text—printed, typed, or written by hand—into machine-readable formats. OCR is crucial for digitizing paper documents, data entry automation, facilitating translations, and text-to-speech application support.

OCR has developed over the course of decades—primitively machines in the 1920s to deep learning models of today. Early systems only processed typed English text, but current approaches such as Convolutional Neural Networks (CNNs), CRNNs, and Transformers currently accommodate sophisticated scripts and even full sentence recognition with contextual intelligence.

The project seeks to create an OCR system for handwritten Marathi text. It meets the requirement to digitize regional content, particularly in educational institutions where handwritten content is prevalent. An OCR system of this kind can assist in creating digital study material, enhancing accessibility, and conserving handwritten information.

1.2 Literature Survey or Related Works

Several research have been made in the field of handwritten character recognition for Indian scripts. Below are two research papers that are closely related to this project:

1. **Gujarati Handwritten Character Recognition by Jyoti Pareek et al. (2020):** This study developed an offline OCR system specifically for Gujarati handwritten characters. The authors employed Convolutional Neural Networks (CNN) and Multi-Layer Perceptron (MLP) classifiers, achieving an accuracy of 97.21% using CNN. The paper also highlighted challenges in handwritten character recognition (HCR) compared to printed OCR. However, the study focused only on character-level recognition and did not address word or sentence conversion.
   DOI: 10.1016/j.procs.2020.04.055

2. **Devanagari OCR Using Deep Learning by Acharya, Pant, and Gyawali (2015):** This research proposed a CNN-based OCR system for the Devanagari script. The authors created a large-scale dataset with 46 character classes and incorporated techniques such as dropout and convolutional layers to prevent overfitting. The system achieved high accuracy on character-level recognition, demonstrating the effectiveness of deep learning

for Indian scripts.
DOI: 10.1109/SKIMA.2015.7400041

These related works provide a foundation for exploring handwritten OCR for Indian languages. However, this project extends beyond character-level recognition by focusing on end-to-end handwritten sentence recognition in Marathi, a challenging and less-explored domain.

1.3 Contributions

This project presents a new method for handwritten Marathi sentence recognition and several important contributions. A novel character segmentation method was developed for Devanagari-based scripts involving Shirorekha (the headline). It employed morphological operations to remove the Shirorekha and was done in combination with contour-based segmentations to extract characters while preserving inner character structures. Processing techniques such as gray-scaling, adaptive thresholding, and noise removal techniques have been used to optimize the input images for better segmentation and recognition. After segmentation, all characters were normalized to a fixed resolution of 32x32 pixels and fed into a CNN-based model developed for recognizing 46 Marathi character classes. The model was tested on a balanced dataset, achieving reasonable classification performance. The project also demonstrates end-to-end conversion of handwritten sentence images into recognized Marathi text with fairly consistent results across diverse input samples. This forms the basis for future work like compound character recognition and beginning to understand word-level.

2 Flowchart/ System Diagram

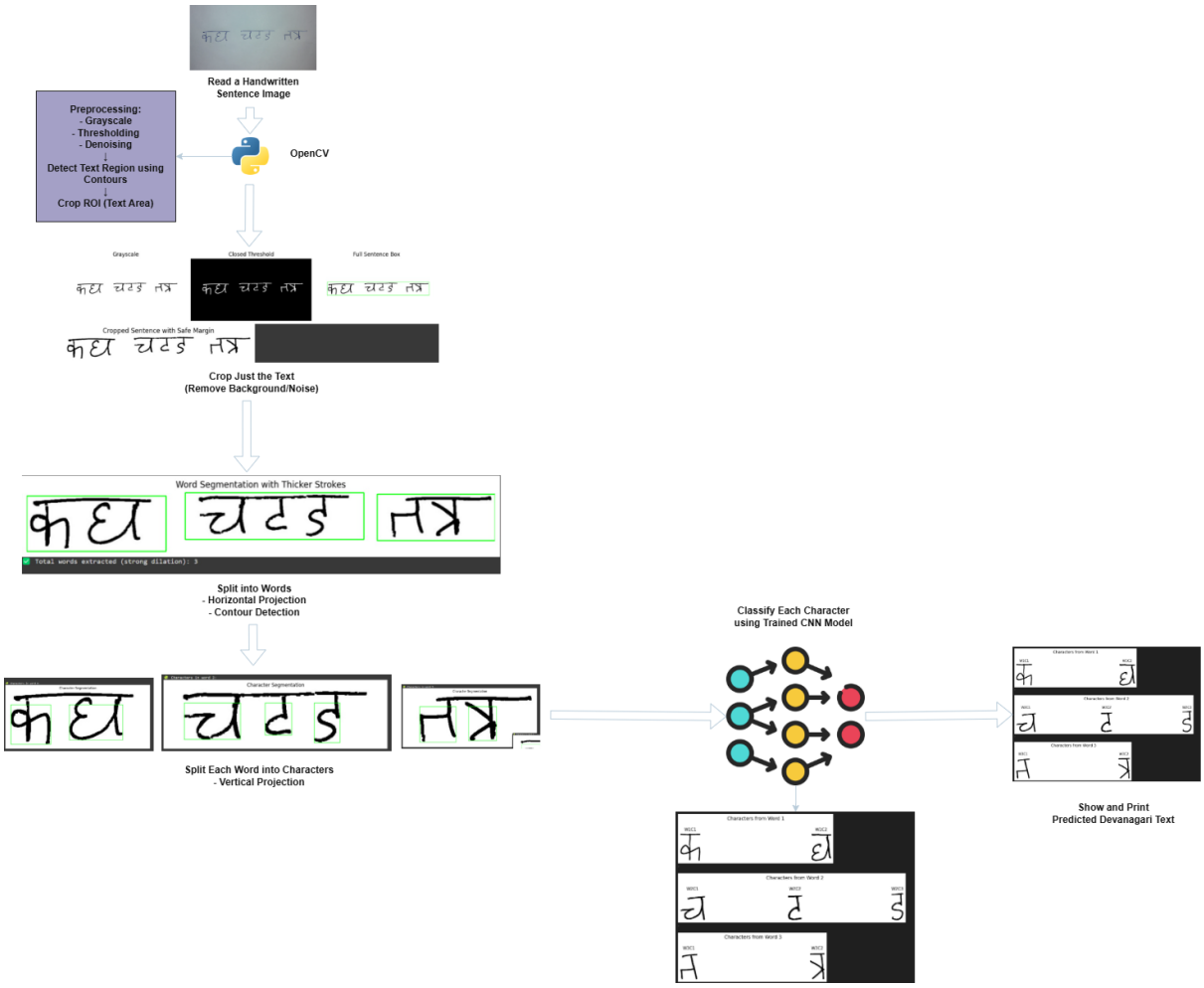


Figure 1: System FlowChart

Description: The very first step in this whole procedure is to input a handwritten Marathi sentence image. The image is then preprocessed, which consists of three very critical steps- grayscale conversion, cropping whitespace, and image enhancement. The grayscale conversion is basically converting the image into shades or tones of gray with the removal of any color information. Cropping removes all of those unnecessary, blank spaces that are lying around the text. Finally image enhancement helps clarify and define the characters.

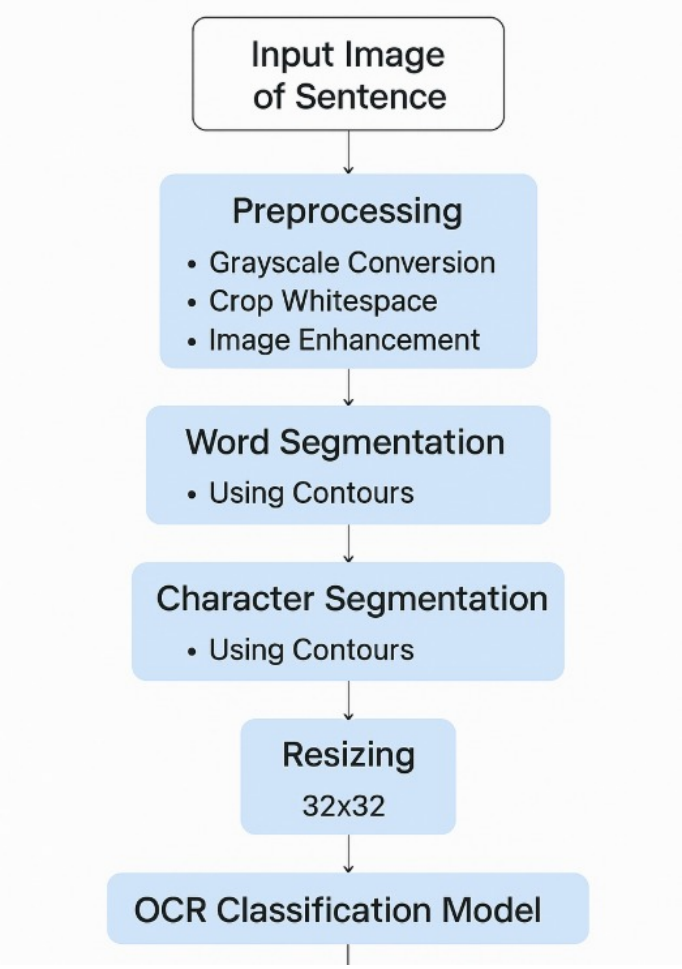
Then comes word segmentation after preprocessing. The segmentation of words is done using contour detection through which the system captures and distinguishes the individual words from the sentence along their boundaries.

Once word extraction is done, the next step is character segmentation. It happens again through contour-based methods with careful consideration to segment individual characters within each word. Shirorekha and such other structures have been removed so as to offer good separation

of characters.

Then, each of the character segments is resized to the standardized 32x32 pixel format, which is essential for feeding into the classification model.

The characters resized are finally given to the OCR Classification Model. The model matches each image to the respective predicted Marathi character and, in the end, sequentially builds up the reconstructed handwritten sentence.



3 Procedure

The methodology involves the following steps:

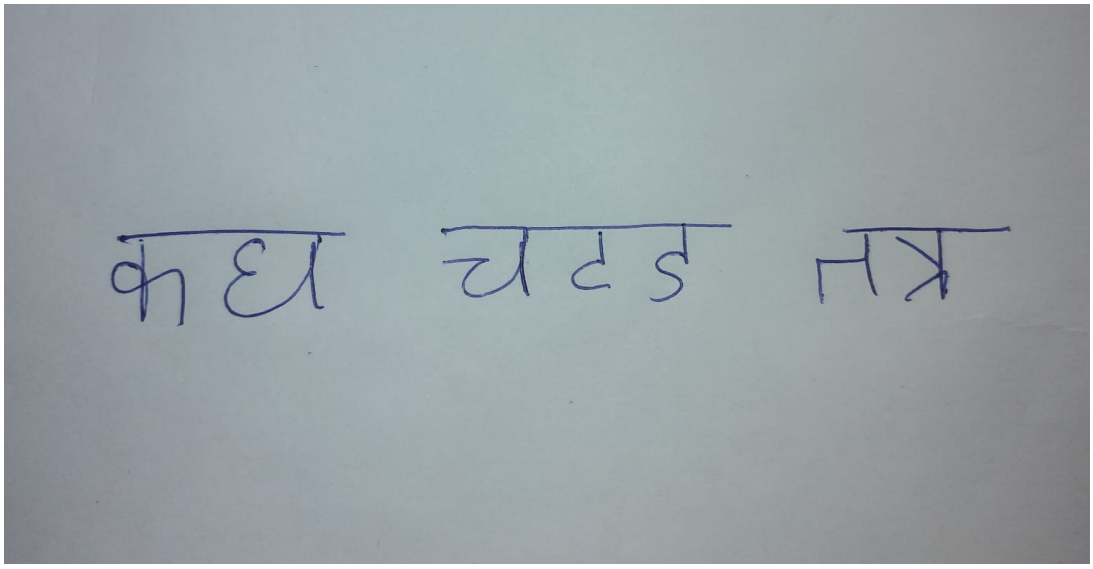
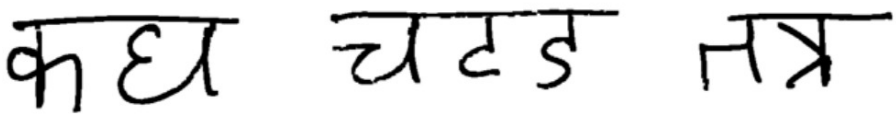


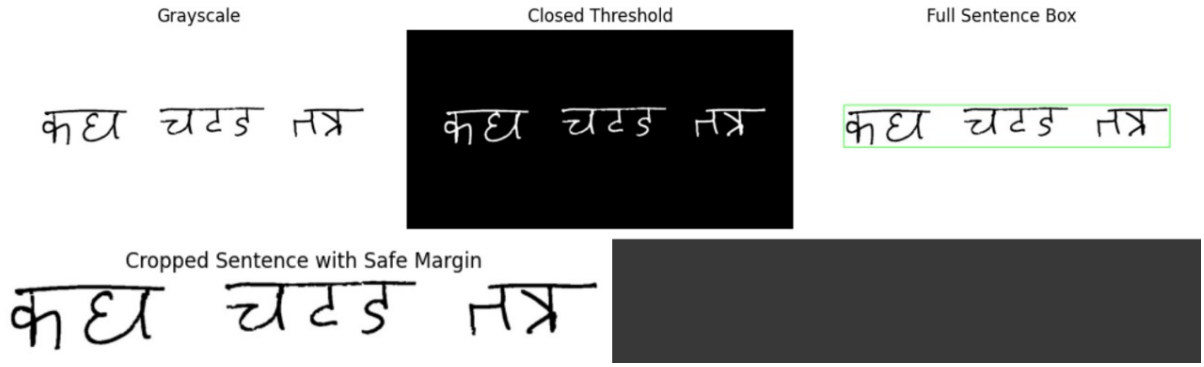
Figure 2: Input Image

3.1 Image Preprocessing



Preprocessing was done to make the handwritten Marathi sentences more legible and organized prior to OCR. The input image was brightened and contrasted, converted to grayscale, and binarized by Otsu's thresholding with binary inversion. A 5×5 dilation filter was applied to make thin or damaged strokes more prominent, and the image inverted to black text on a white surface.

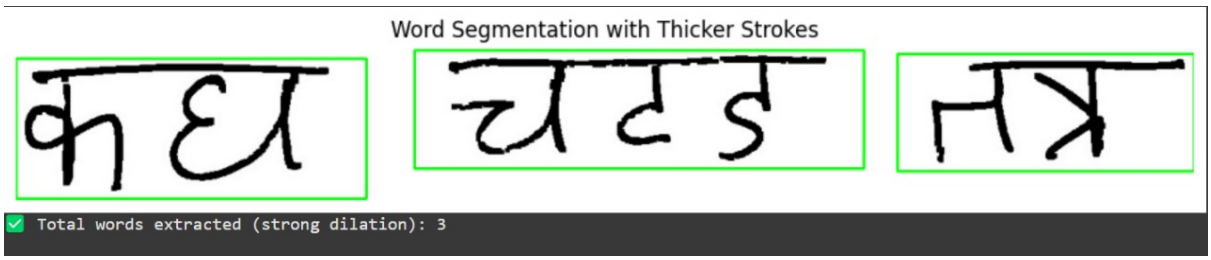
3.2 Background Removal



A image processing pipeline was implemented to properly crop handwritten sentences from document scans. This involved grayscale conversion, Otsu's thresholding followed by binary inversion, and morphological closing to link characters. Following external contour filtering, a bounding box with a safe margin was used to provide clean and intact sentence crops for OCR.

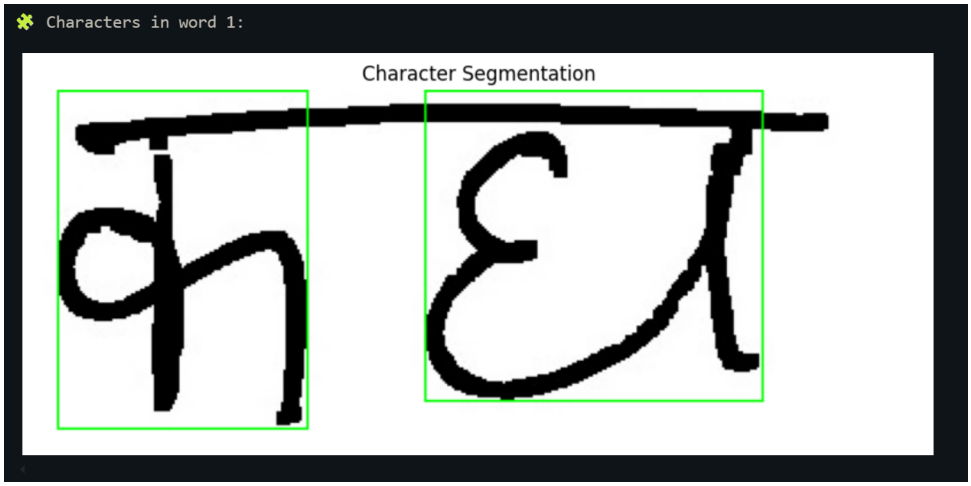
3.3 Segmentation

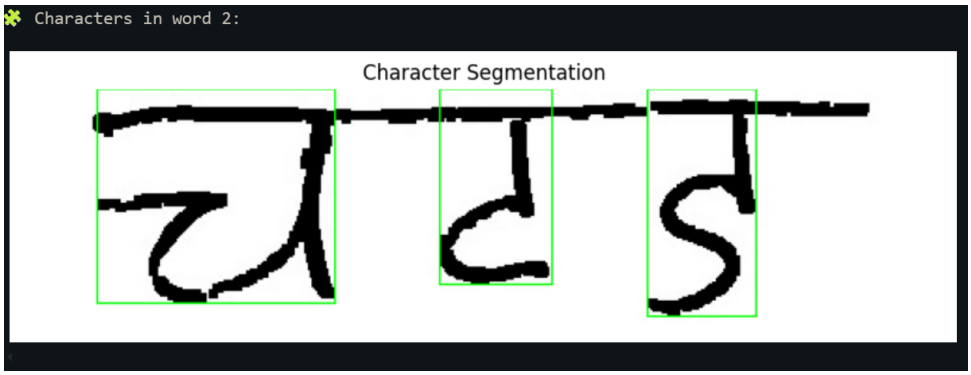
1. Word segmentation is performed similarly using contour detection.

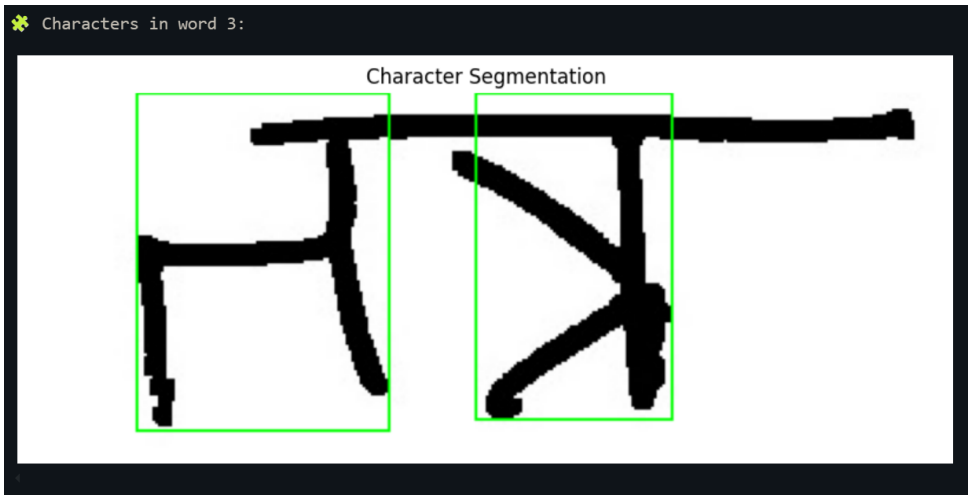


2. similaraly Characters were segmented from words.

For character segmentation, the word image was initially converted into grayscale, and then Gaussian blur and Otsu's thresholding with binary inversion were applied to improve text clarity. The top 1/4th of the image was cropped out to eliminate Shirorekha which was again restored. The external contours of the characters were found using cv2.findContours, and bounding boxes were generated to detect separate characters. The characters were cropped from the original image for processing.

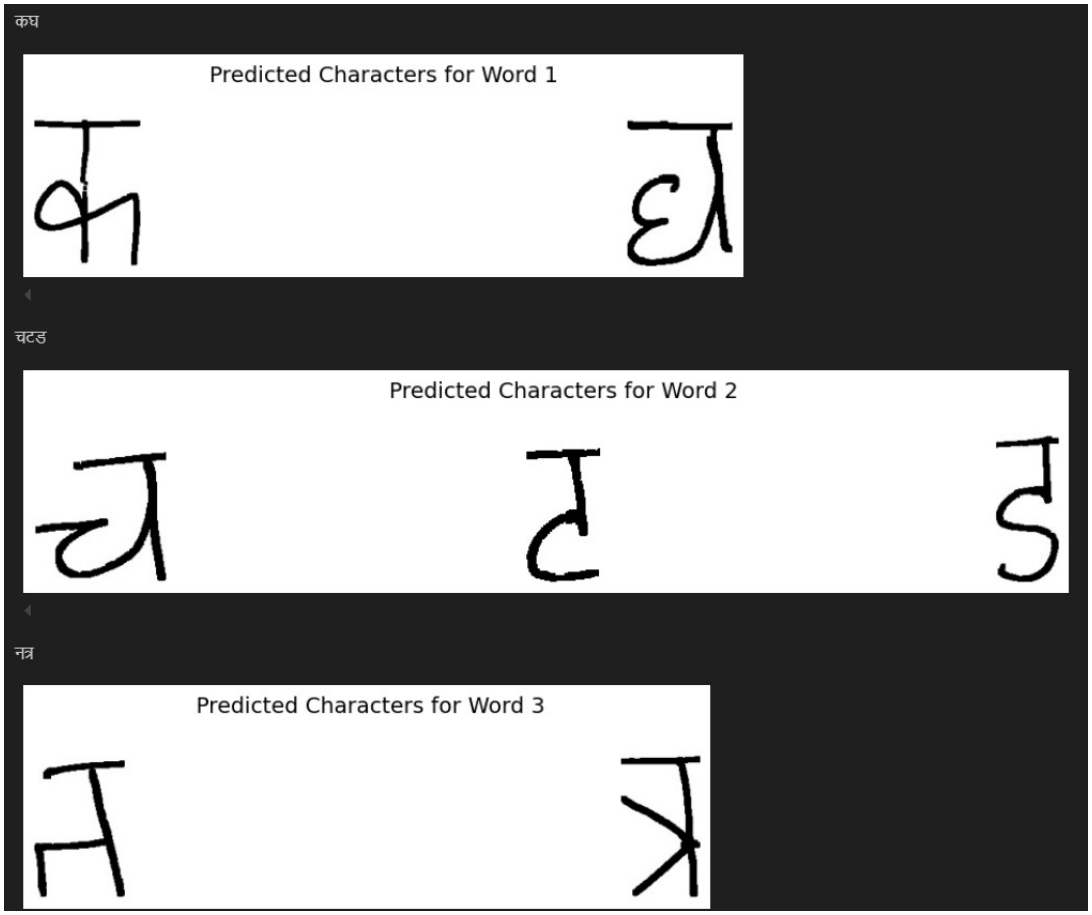

Characters in word 1:
Character Segmentation


Characters in word 2:
Character Segmentation


Characters in word 3:
Character Segmentation

3.4 Character Recognition

A convolutional neural network (CNN) model is trained on Marathi characters to predict the text for each segmented character.



4 Optical Character Recognition Model

The OCR system is based on a sequence Convolutional Neural Network (CNN) architecture for handwritten character recognition. It consists of two layers of convolution with ReLU activation functions, ensuring non-linearity and the ability of the network to recognize complex patterns. The two convolutional layers are followed by a max-pooling layer to downsample the feature maps and decrease computational complexity while retaining dominant features. The flattened output is then passed to a dense layer of 256 neurons and a second ReLU activation for even deeper feature representation. Dropout layers with rates 0.5 and 0.3 are added after flattening and the dense layer to avoid overfitting. The final layer is a softmax activation to generate a probability distribution over multiple character classes. The Adam optimizer with a learning rate of 0.001 and categorical cross-entropy loss are used in training the model, which is suitable for multi-class classification. Accuracy is used as the metric for evaluation during model training.

Table 1: CNN Model Architecture

Layer (Type)	**Output Shape**	**Param #**
Conv2D	(None, 30, 30, 32)	320
MaxPooling2D	(None, 15, 15, 32)	0
Conv2D	(None, 13, 13, 64)	18,496
MaxPooling2D	(None, 6, 6, 64)	0
Flatten	(None, 2304)	0
Dropout	(None, 2304)	0
Dense	(None, 256)	590,080
Dropout	(None, 256)	0
Dense (Output)	(None, 46)	11,822

Activation Functions

The model makes use of the **ReLU (Rectified Linear Unit)** activation function in both its convolutional and hidden dense layers. ReLU serves to introduce non-linearity and facilitate convergence during training by preventing the vanishing gradient problem. Mathematically, it is given by:

$$\text{ReLU}(x) = \max(0, x)$$

For the output layer, **Softmax** activation function is applied to scale the raw output scores (logits) into a probability distribution over several classes. It is particularly suitable for multi-class classification problems. The Softmax function is defined as:

$$\text{Softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

where z_i is the input score for class i, and K is the total number of classes.

Adam Optimizer

Adam optimizer adapts the learning rate based on past gradients and their squared values. It updates model parameters as follows:

$$\theta_t = \theta_{t-1} - \frac{\alpha}{\sqrt{s_t} + \epsilon} v_t$$

Where: - θ_t is the updated parameter, - α is the learning rate, - v_t and s_t are the first and second moment estimates of the gradients, - ϵ prevents division by zero.

Adam helps in faster and more efficient training by dynamically adjusting the learning rate.

Categorical Cross-Entropy Loss

Categorical Cross-Entropy compares the predicted probabilities with the true labels. The loss is calculated as:

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \log(p_i)$$

Where: - y_i is the true label (1 for correct class, 0 for others), - p_i is the predicted probability, - C is the number of classes.

This loss function encourages the model to reduce the error between predicted and true labels, improving accuracy.

Training and Validation Accuracy

The training and validation accuracy are crucial metrics to assess how well the model generalizes. Below is the graph depicting both:

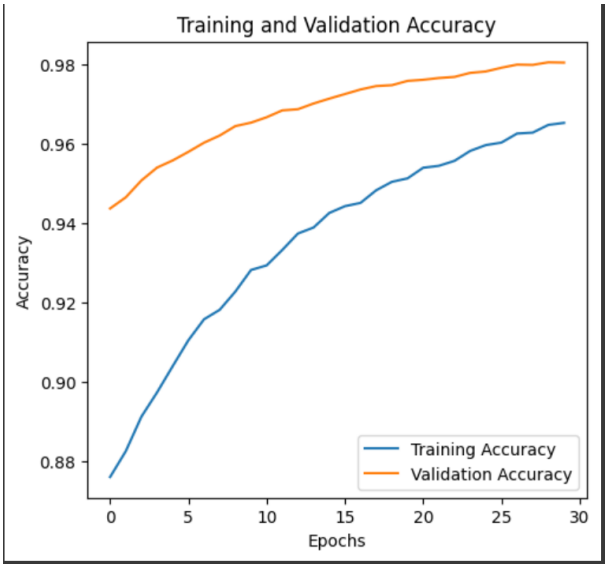


Figure 3: Training and Validation Accuracy over Epochs

In this graph: - The blue line(bottom) represents the training accuracy, showing how the model performs on the training data over epochs. - The yellow line(above) represents the validation accuracy, indicating how the model behaves to unseen data during training.

5 Implementation Environment

The OCR model is trained on a hand-curated dataset of handwritten Marathi characters, consisting of 46 distinct character classes (vowels, consonants, and modifiers). The dataset consists of 2000 samples per class, totaling 92,000 labeled images, allowing for strong training and evaluation.

The dataset is divided into three segments for proper training and testing: 80% (1600 samples/class) is used for training, 10% (200 samples/class) is kept aside for validation purposes to adjust hyperparameters and avoid overfitting, and 10% (200 samples/class) is set aside for actual testing and performance assessment.

Images are in `.png` format, resized to 32×32 pixels to ensure consistency, with each character placed centrally within a 28×28 pixel area and 2-pixel padding on all sides. This ensures consistency and retains significant features during convolution operations.

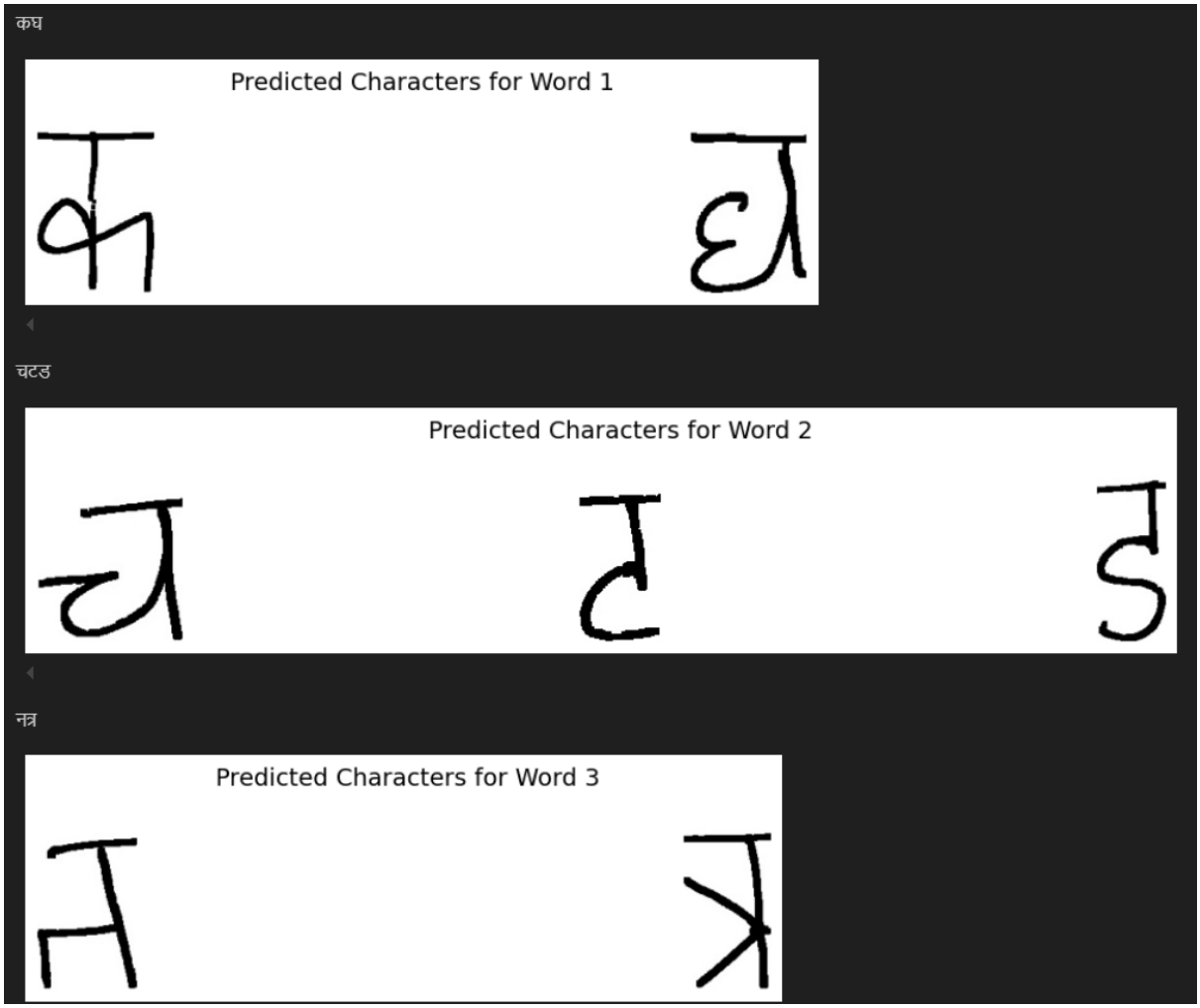
The dataset is obtained from the **UC Irvine Machine Learning Repository**, a reliable and well-established source of high-quality datasets, offering a solid foundation for developing and testing the OCR model.

6 Results

Test accuracy achieved for OCR Model: 98.79

```
Evaluating on Test Set...
144/144 ━━━━━━━━━━━━━━━━━━━━ 7s 51ms/step - accuracy: 0.9857 - loss: 0.0490

Test Loss: 0.0420
Test Accuracy: 0.9879 (98.79%)
```



7 Conclusion and Future Work

7.1 Conclusion

An Optical Character Recognition (OCR) system is implemented in this project for the recognition of handwritten Marathi sentences. It involved preprocessing of the input image, word and character-level segmentation through a contour-based approach, and classification of individual characters by a trained CNN model. Special emphasis was placed on the removal of Shirorekha to improve the accuracy of character segmentation in order to address its inhibitive effect on the aforementioned challenges resulting due to the associated handwriting pattern of the Marathi script. The system performed very well and successfully recovered entire sentences from handwriting input images. The results ascertain the technique capability and suggest that other OCR systems may be established for the low-resource regional languages.

7.2 Future Work

Although the current system produces some excellent results, there are still possibilities for improvement. One would concentrate on instances where characters do overlap or touch one another, fitting them into more sophisticated image segmentation techniques. Similarly, broadening the model's recognition to cover matras and compound characters specific to the Marathi script would make the recognition process even more comprehensive. In addition, those datasets might be supplemented with samples scaled down from a large demographic range such that the model can enhance its generalization across different handwriting styles. Lastly, the analysis task might include end-to-end sequence modeling-based solutions, like CRNNs (Convolutional Recurrent Neural Networks) or Transformer-based solutions, to avoid the task of individual segmentation and obtain improved accuracy in writing whole handwritten sentences.

8 References

1. Acharya, S., & Gyawali, P. (2015). *Devanagari Handwritten Character Dataset* [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5XS53

2. Acharya, S., Pant, A. K., & Gyawali, P. K. (2015). Deep learning based large scale handwritten Devanagari character recognition. In *2015 9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)* (pp. 1–6). IEEE. https://doi.org/10.1109/SKIMA.2015.7400041

3. Pareek, J., Singhania, D., Kumari, R. R., & Purohit, S. (2020). Gujarati Handwritten Character Recognition from Text Images. *Procedia Computer Science*, 171, 514–523. https://doi.org/10.1016/j.procs.2020.04.055