OCR for devanagari script

Mitali Singh¹, Pranava Gayathri¹* and Prabhakar Kandukuri¹
¹Department of AI&ML, CBIT, Hyderabad, Telangana, India.

Abstract. The introduction of Optical Character Recognition (OCR) technology revolutionized text digitization, allowing physical documents to be converted into editable and searchable digital representations. This paper goes into the unique challenges and advancements in OCR designed exclusively for the Devanagari script. It helps to preserve cultural heritage by digitizing historic manuscripts and religious writings and making them more accessible. Furthermore, Devanagari OCR has practical uses in administrative activities, data entry, and educational content digitization.

1 Background

Devanagari is a complex script used for languages like Hindi, Marathi, Sanskrit, and others. It includes intricate ligatures, conjunct characters, and a variety of fonts. It consists of 36 Consonants, 10 vowels, 10 digits. OCR faces challenges in accurately recognizing these script features due to its complexity. The main challenge is to recognize conjunct characters. Researchers and developers continue to invest in ongoing research and development to address the evolving challenges of Devanagari OCR, striving for improved accuracy and efficiency.

2 Objectives

1) The primary objective is to accurately recognize and classify individual Devanagari script letters i.e., consonants and numbers
2) Achieve a high classification accuracy to ensure that Devanagari Characters are correctly identified in different applications such as OCR and text analysis

3 Methods

The revised papers have an average accuracy of 94.2%. Even with higher average accuracy, one typical shortcoming is the inability to distinguish between similar-looking letters. Below is a tabulation of a few of the considered papers.

* Corresponding author: pranavagayathriaavuti@gmail.com

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).
<table>
<thead>
<tr>
<th>Ref.No</th>
<th>Paper</th>
<th>Methodology</th>
<th>Accuracy %</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1]</td>
<td>Fuzzy Model Based Recognition of Handwritten Hindi Characters</td>
<td>Feature Extraction: Box Method</td>
<td>91.3</td>
<td>Multiple classifiers are used for 3 different categories</td>
</tr>
<tr>
<td>[2]</td>
<td>Comparative study of Devanagari handwritten character recognition using different features and classifiers</td>
<td>Classification: Mirror Image Learning</td>
<td>95</td>
<td>Increased character set for training</td>
</tr>
<tr>
<td>[3]</td>
<td>Gradient Local Auto-Correlation for handwritten Devanagari character recognition.</td>
<td>GLAC(Gradient Local Auto Correlation)</td>
<td>93.21 &amp; 95</td>
<td>Difficulties in recognising similar words.</td>
</tr>
<tr>
<td>[6]</td>
<td>Recognition of handwritten Devanagari characters using linear discriminant analysis</td>
<td>Feature Extraction: PCA and LDA</td>
<td>94.2</td>
<td>LDA implementation is complex and not so accurate for multiclass classification.</td>
</tr>
</tbody>
</table>

This is the summary of our literature survey. Our objective is to increase the classification accuracy using deep learning methods like using pretrained models, transfer learning etc. Fine-tuning pretrained models on your specific dataset or domain can significantly improve recognition accuracy.
4 Results and Conclusion

In conclusion, the findings of this study demonstrate the effectiveness of Quadratic Support Vector Machine (SVM) as a classification algorithm and Histogram of Oriented Gradients (HOG) as a feature extraction method in the context of our problem domain. With an impressive classification accuracy of 95.9%, our results underscore the robustness and suitability of this combination for the task at hand. In this study, we propose a 96% accurate CNN technique in an effort to improve the current model.

References


