Investigation on OCR Systems for Odia

Swayam Parida

Stanford | ENGINEERING
Computer Science

Background

- Though state-of-the-art OCR systems today boast > 95% accuracy, they are only available for widely spoken global languages.
- Languages such as Odia, a vernacular spoken widely in eastern India, are yet to benefit from high quality OCR systems.

Objectives/Aims

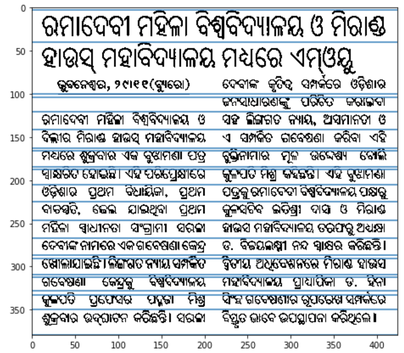
- Implementing the two major components of an OCR pipeline: *segmentation* and *classification*.
- Identifying the challenges unique to Odia text extraction.

Methods

- Segmentation was performed using histogram projection and choosing the rows and columns consisting only of background pixels.
- Feature extraction: Histogram of foreground pixels on vertical and horizontal scan lines, Histogram of background pixels encountered before first foreground pixel.
- Classification: Logistic Regression.

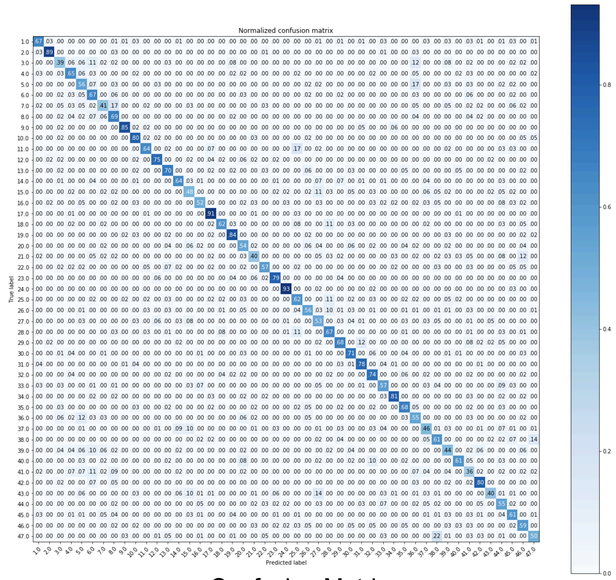
Results & Analysis

- The line segmentation algorithm accurately segmented each individual line of text.
- Over-segmentation occurred sometimes when diacritic marks that should have been considered as part of the same line of text were segmented into individual lines.
- Character segmentation failed occasionally when binarization on low resolution images would leave pixels connected without a column of background pixels.
- Character segmentation was also unreliable when diacritics extended into the separator column of background pixels.



Line Segmentation Result

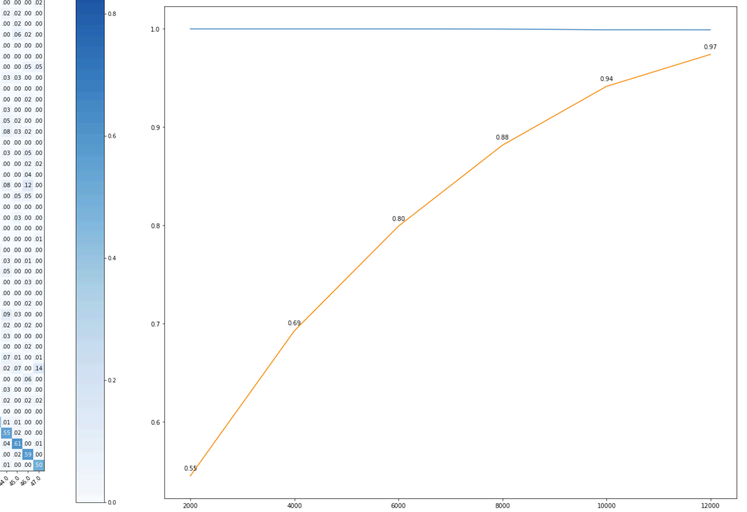
Character Segmentation Result



Confusion Matrix

- Macro-precision: 97%
- Macro-recall: 98%
- Macro-F1: 97%

Some characters are similar enough that when badly written and viewed in isolation, they are indiscernible by humans. Hence, it is necessary to incorporate contextual information such as n-grams to correctly classify these characters.



Improvements in Training Accuracy (Blue) and Testing Accuracy (Orange) wrt number of training examples

- Training accuracy: 99.92%
- Validation accuracy: 97.17%

Since both training and testing accuracy are high, it implies that the model has low bias and low variance. However, these results are achieved on preprocessed images. The efficacy of the classifier is yet to be evaluated on less optimal images.

Limitations

- The presence of diacritics in Odia, i.e. alphabet modifiers analogous to accents in Latin-script languages, posed challenges to neat segmentation.
- The only publicly available dataset comprises of base characters without diacritics.
- Accommodating diacritics would significantly increase the label space.
- No n-gram datasets are publicly available to improve classification accuracy.

Conclusions

- The most significant challenge facing Odia OCR systems is the lack of comprehensive standardized datasets.
- High classification accuracy on individual characters can be achieved with existing feature extractors and classification algorithms.

Future Directions

- Build a n-gram dataset from Odia corpus to enhance classification.
- Create dataset with images of entire words and use CNNs for learning. This mitigates errors caused by incorrect segmentation.

Contact Information: Swayam Parida | 562 Mayfield Ave, Stanford, CA 94305 | swayamp@stanford.edu.