Paper Title:
A complete printed Bangla OCR system

Paper Link:
https://www.sciencedirect.com/science/article/abs/pii/S0031320397000782

1 Summary

1.1 Motivation

The motivation behind this endeavor was rooted in the absence of Optical Character Recognition (OCR) systems tailored for Indian scripts, with a specific focus on the intricacies of the complex Bangla script. The overarching goal was to pioneer the development of a comprehensive OCR system specifically designed for printed Bangla documents.

1.2 Contribution

This project marks a significant milestone as the first-ever OCR system crafted exclusively for the Bangla script. The contributions lie in the application of robust techniques, including skew correction, zone separation, and a tree classifier for character recognition. These innovations collectively elevate the accuracy and reliability of the OCR system.

1.3 Methodology

The methodology employed a multi-faceted approach, encompassing flatbed scanning, skew correction, zone separation, character segmentation, a tree classifier for basic characters, and template matching for compound characters. Notably, a dictionary-based error correction mechanism was integrated to enhance the precision of the OCR system.

1.4 Conclusion

The culmination of our efforts resulted in a highly accurate OCR system, achieving a remarkable 95.5% accuracy at the word level and an impressive 99.1% accuracy at the character level, particularly on clear printed Bangla documents.

2 Limitations

2.1 First Limitation

While our OCR system demonstrated exceptional performance on clear printed documents, its evaluation on noisy documents remains a subject for further exploration. We acknowledge the need for extensive testing in diverse document scenarios to comprehensively understand its limitations.

2.2 Second Limitation

Another limitation pertains to the font variations considered during evaluation. The focus was primarily on the popular Linotype font, raising questions about the system's adaptability to other font styles. Further research is essential to address this limitation and enhance the system's versatility.

3 Synthesis

In synthesis, the techniques developed in this project extend beyond the Bangla script. Skew correction, zone separation, and the tree classifier exhibit potential applicability to other Indian scripts, such as Devnagari. Furthermore, the dictionary-based error correction mechanism proves valuable for inflectional languages. This pioneering work not only addresses the OCR needs of the Bangla script but also lays a foundation as a model for the development of OCR systems in various other scripts.