Abstract

In a multi script environment, majority of the documents may contain text information printed in more than one script/language forms. For automatic processing of such documents through Optical Character Recognition (OCR), it is necessary to identify different script regions of the document. In this context, this paper proposes to develop a model to identify and separate text words of Kannada, Hindi and English scripts from a printed tri-lingual document. The proposed method is trained to learn thoroughly the distinct features of each script and uses the simple voting technique for classification.Experimentation conducted involved 1500 text words for learning and 1200 text words for testing. Extensive experimentation has been carried out on both manually created data set and scanned data set. The average success rate is found to be 1/4.
99% for manually created data set and 98.5% for data set constructed from scanned document images.

References

- S. L. Wood, X. Yao, K. Krishnamurthy and L. Dang, "Language identification for

Index Terms

Computer Science Pattern Recognition

Keywords

Multi-lingual Document Processing Script Identification Feature Extraction Binary Tree Classifier