Abstract

Optical character recognition is a very challenging area. Many works have been done and still being done for many languages across the world. For many Indian languages too good amount of work has been done. However, Gujarati is a language for which hardly any work can be found. Gujarati has a rich literary heritage, and therefore it is important to preserve it for the next generation. In this paper an attempt has been done to segmenting out the words and characters from old typewritten Gujarati documents. Here an algorithm is presented which makes use of global threshold for converting scan RGB documents to blank and white documents. Noise removal has also been applied. Here Radon transform is utilized for skew detection. The novel concept of using Radon transform is presented here in this work. Here Radon transform is used for segmenting documents into lines and then vertical profiles has been used for further segmentation of lines in characters. At last this segmentation algorithm is also tested for the documents typewritten in Hindi. The algorithm presented here gives very good results.

References

- C. Bhagvati, T. Ravi, SM. Kumar, A. Negi, On developing high accuracy ocr system for telugu and other indian scripts, Proceedings of the Language Engineering Conference (LEC’02), 18-23
- B.M. Sagar, Shobha G., R. Kumar, OCR for kannada text to machine editable format using database approach, WSEAS Transactions on Computers, 6(7).2008, 766-769
- Sheetalkumari, Shreeranjani, Balachandar, A Singh, M. singh, R Ratan, S. Kumar, Optical character recognition for printed tamil text using unicode, Journal of Zhejiang University SCI, 64(11), (2005), 1297-1305

Index Terms

Computer Science
Signal Processing
Keywords
Segmentation; Radon transform; skew correction; digitization; noise removal