In this paper an attempt is made to retrieve the text region alone from a degraded text document. For doing that, four different filters are used for noise removal in the text document. Later document binarization is done using thresholding. Three different thresholding techniques are implemented for foreground-background separation. Then candidate region is selected and features are extracted. The features are then fed to an SVM to classify text and non-text regions. The proposed approach is implemented and tested on various hand written and machine printed degraded text documents.

References

- Yahia S. Halabi, Zaid SA, FarisHamdan, Khaled Haj Yousef, Modeling
- T Kasar, J Kumar and A G Ramakrishnan, Font and Background Color Independent Text Binarization
- K. S. Sim *, C. P. Tso, Y. Y. Tan, Recursive sub-image histogram equalization applied to gray scale images

Index Terms

Computer Science

Keywords

Image Filtering  Wavelet Decomposition  Feature Extraction  Document Binarization