International Journal of Information Processing, 8(3), 23-30, 2014
ISSN : 0973-8215
IK International Publishing House Pvt. Ltd., New Delhi, India

Segmentation of Handwritten Text from Underlined Variable Regions in Documents

Suman V Patgar[a], Y H Sharath Kumar [b], Vasudev T[b]

[a]PET Research Foundation, P E S College of Engineering, Mandya, India, 571401,
Contact: sumanpatgar@gmail.com

[b]Maharaja Research Foundation, Maharaja Institute of Technology Mysore, Belawadi, S.R Patna, Mandya, India, 571438

In this paper, we propose a system to identify underlined handwritten region from a given document. The approach initially eliminates the outer boundary surrounding the text of interest and later, the handwritten words are extracted from the remaining contents through proposed line detection algorithm using central moments. The performance of the proposed method is evaluated using five different region based measures. In order to substantiate the efficacy of the proposed model experimentation is conducted based on a dataset over 300 various samples such as conference certificates, birth/death certificates, bank challans, degree certificates, attendance certificates, *etc.*. Experimental results show that the proposed segmentation algorithm achieved best average Dice similarity measure and Measure of overlapping accuracy.

Keywords : Document, Handwritten Text, Segmentation, Underlines, Variable Regions.

1. INTRODUCTION

Many authorities in India trust and consider the photocopied documents submitted by citizens as proof and accept the same as genuine. Few such applications like to open bank account, applying for gas connection, requesting for mobile sim card, the concerned authorities insist photocopy documents like voter id, driving license, ration card, pan card and passport as proof of address, age, photo id etc to be submitted along with the application form. Certain class of people could exploit the trust of such authorities, and indulge in forging/ tampering/ fabricating photocopy document for short / long term benefits unlawfully. Fabricated photocopy is generated normally by making required changes intelligently in the photocopy obtained from an original document and then taking the recursive photocopy from the modified photocopy [1]. It is learned that in majority cases, fabrications are made by replacing a different photograph in place of photograph of an authenticated person; replacing contents in variable regions [2] through cut-and-paste technique from one or more documents; overlaying new content above actual content; adding new content into existing content; removing some content from existing; changing content by overwriting; intellectually changing character in contents.

It is quite evident from the applications listed above, the fabrication could be mainly made in the variable regions [1] of documents. Most of the variable regions contain handwritten text in the documents of the above applications. Figure 1 (a) and (b) show samples of documents having variable regions being encircled. There is a strong need to detect fabrication in the photocopied documents submitted for the applications mentioned above. Since fabrication could be suspected in variable regions, it is quite necessary to identify the variable regions in such documents. Once the variable regions in a document are identified, further investigations could be performed to check for the possibility of fabrication. Thus identification of variable regions in a document is a prerequisite step to detect fabrication in photocopied

Table1: Values of region based measure for segmentation from different class of documents

Samples Type	MOS	MOL	MUS	DSM	ER
Conference Certificates	0.2132	0.7921	0.0723	0.8011	0.1893
Cheques	0.2248	0.7621	0.0623	0.8217	0.1874
Birth/Death certificate	0.2018	0.7789	0.0739	0.8300	0.1985
Attendance certificate	0.2011	0.7945	0.0698	0.8174	0.1996

5. CONCLUSIONS

The implemented method serves as an intelligent system for segmentation of handwritten text in variable region in a document. The method is essentially based on central moments of intensity values of document. It shows an acceptable better segmentation efficiency, which is close to ground truth segmentation. It can be used without a complex hardware setup for detection of variable regions in documents in applications where only photocopy documents are sufficient. Certain cases exhibit mis-segmentation due to appearance of characters having line structure in the components. The method is limited to work on document in which variable regions are underlined. There is much scope to segment variable regions which are not underlined and also the contents are printed instead of handwritten. Further the performance efficiency can be enhanced through exploring other moments and also preparing the document to be free from noise, dirt and background art.

REFERENCES

1. Suman Patgar and Vasudev T. Estimation of Recursive Order Number of a Photocopied Document Through Entropy From Gray Level Co-occurrence Matrix, *ICSEC2013*, pages 313–317, 2013.
2. Vasudev T. Automatic Data Extraction from Pre-Printed Input Data Forms: Some New Approaches, PhD thesis supervised by Dr. G Hemanthakumar, University of Mysore, India, 2007.
3. Gorman L O. The Document Spectrum for Page Layout Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 15:1162–1173, 1993.
4. Nagy G, Seth S and Stoddard S. Document Analysis with an Expert System, *Pattern Recognition Practices II*, 2:149–155, 1984.
5. Ergina Kavallieratou, Stathis Stamatatos and Hera Antonopoulou. Machine-Printed from Handwritten Text Discrimination, *in Proceedings of* 9^{th} *International Workshop on Frontiers in Handwriting Recognition (IWFHR-9 2004)*, 2004.
6. M S Shirdhonkar and Manesh B Kokare. Discrimination between Printed and Handwritten Text in Documents, *IJCA Special Issue on Recent Trends in Image Processing and Pattern Recognition RTIPPR*, 2010.
7. Ranju Mandal, Partha Pratim Roy and Umapada Pal. Signature Segmentation from Machine Printed Documents using Conditional Random Field, *in Proceedings of International Conference on Document Analysis and Recognition*, pages 1170–1174, 2011.
8. Xujun Peng, Srirangaraj Setlur, Venu Govindaraju and Ramachandrula Sitaram. Handwritten Text Separation from Annotated Machine Printed Documents using Markov Random Fields, *IJDAR* 16:1–16, 2013,
9. Suman Patgar and Vasudev T. An Unsupervised Intelligent System to Detect Fabrication in Photocopy Document Using Geometric Moments and Gray Level Co-Occurrence Matrix, *IJCA(0975-8887)*, 74(12), July 2013.
10. Kuo Chin Fan. Marginal Noise Removal of Document Image, *ICDAR 01*, pages 317–321, 2001.
11. Rafael C Gonzales and Richard E Woods. Digital Image Processing, 2nd Edition, *Pearson*

Education Publication, 2002.
12. Arash Asef Nejad and Karim FaezA. Novel Method for Extracting and Recognizing Logos, *International Journal of Electrical and Computer Engineering (IJECE)*, ISSN: 2088-8708, 2(5):577–588, October 2012.
13. Tinku Acharya and Ajoy K Ray. Image Processing Principles and Applications, *A Wiley-Interscience Publication*.
14. Jain A K. Fundamentals of Digital Image Processing, *Prentice Hall, Englewood Cliffs, NJ*, 1998
15. M R Teague. Image Analysis *via* the General Theory of Moments, *Journal of the Optical Society of America*, 70(8):920–930, 1979.
16. P B Mallikarjun and D S Guru. Performance Evaluation of Segmentation and Classification Of Tobacco Seedling Diseases, *International Journal of Machine Intelligence* ISSN: 09752927 and E-ISSN: 09759166, 3(4):204–211, 2011.
17. Elter M, Held C and Wittenberg T. Physics in Medicine and Biology, 55:5299–5315, 2010.

Vasudev T is Professor, in the Department of Computer Applications, Maharaja Institute of Technology, Mysore. He obtained his Bachelor of Science and post graduate diploma in computer programming with two Masters Degrees one in Computer Applications and other one is Computer science and Technology. He was awarded Ph.D in Computer Science from University of Mysore. He is having 30 years of experience in academics and his area of research is Digital Image Processing specifically Document Image Processing.

Suman V Patgar is Research Scholar, P.E.T Research Center Mandya. She obtained her Bachelor of Engineering from Kuvempu University in 1998. She received her Masters degree in Computer Science and Engineering from VTU Belgaum in 2004. She is pursuing doctoral degree with the supervision of Vasudev T under University of Mysore.

Y H Sharath Kumar is Assistant Professor in the Department of Computer Science and Engineering, Maharaja Institute of Technology, Mysore. He obtained his Bachelor of Engineering from VTU Belgaum. He received his Masters degree in Computer Science and Technology from University of Mysore. He is pursuing doctoral degree with the supervision of D S Guru under University of Mysore.