International Journal of Research Studies in Computer Science and Engineering (IJRSCSE)
Volume 2, Issue 3, March 2015, PP 41-45
ISSN 2349-4840 (Print) & ISSN 2349-4859 (Online)
www.arcjournals.org

A Hierarchical Pre-processing Model for Offline Handwritten Document Images

Ch. N. Manisha[1#], E. Sreenivasa Reddy[2*], Y.K. Sundara Krishna[3#]

[#]Krishna University, Machilipatnam, Andhra Pradesh, India
[1]ch.n.manisha@gmail.com
[3]yksk2010@gmail.com
[*]Acharya Nagarjuna University, Guntur, Andhra Pradesh, India
[2]esreddy67@gmail.com

Abstract: *This paper proposed a Hierarchical preprocessing model for offline handwritten based digital document images. This hierarchical model has level 0 to level 5. i.e., Noise Removal to Segmentation totally six levels in this hierarchical preprocessing model. This paper describes how various preprocessing methods will apply as in sequence for preprocessing of offline handwritten based digital document images for recognizing accurate characters and numerals from offline handwritten based digital document images.*

Keywords*: Noise, Skew, Slant, Skeletonization, Segmentation.*

1. INTRODUCTION

To convert physical document into digital format is called digital document images. To increase the lifetime of the document, most physical documents are converting into digital document images. To recognize the characters in digital document is a more complicated task. Before recognizing the characters to apply preprocessing methods to the digital document is a very important task. Since the errors occur by scanning the document or taking the photograph of the document or quality of the document or improper writing style, the noise will added to the digital document images.

Depend on types of characters made on offline documents, the digital document images are two types. One is offline printed based digital document images and another is offline handwritten based digital document images. Compare to preprocessing of offline printed based digital document images, preprocessing of offline handwritten based digital document images are toughest task.

To recognize offline handwritten characters is difficult task compare to online handwritten characters. Since online handwritten characters have temporary information like pen strokes, pen movements and pen pressure, etc., therefore less preprocessing methods are used for online handwriting recognition. Offline handwritten characters have an only spatial structure available.

Therefore offline handwritten characters require more and proper preprocessing methods than online handwritten characters.

The Hierarchical model is one of the tree-structure in [14]. If the hierarchical model has n levels, then it can be indexed as 0 to n-1 levels. Level 0 consider as the root node for hierarchical models. We have lot of preprocessing methods are available, but It is necessary to understand step by step how to apply preprocessing methods to offline handwritten based digital document images is important and which type of method is suited for it. Because a proper pre-processed digital document images will only give accurate recognized characters. We proposed a hierarchical preprocessing model for how to apply each and every step of preprocessing methods to offline handwritten based digital document images as in sequence.

2. LITERATURE REVIEW

Different preprocessing methods are used in recognition of offline handwritten based digital document images. But the main problem is very less papers gave the information on how to they implemented methods for preprocessing. Kanika Bansal et.al. [1] proposed K-Algorithm for removing noise from handwritten document images. Angle of document estimated by vertical projections in [2]. Skew angle corrected based on piecewise covering by parallelograms in [3]. Based on eigen-point document skew was corrected in[4].

Handwritten signature slant correction described in [5], slant characters corrected by shear operation in[6]. Using dynamic programming Uchida, S. et. al. [7] corrected the slant of the characters.

Handwritten overlapped characters are segmented in[8]. Munish Kumar et.al. [9] used water reservoir method to segment touched characters of gurmukhi

script. Veena Banasal et.al [10] segmented touching characters in Devanagari.

Segmentation of Telugu touched characters described in [11,12,13]. Telugu overlapped characters were segmented by drop fall algorithm in[11], by split profile algorithm in[12], segmented using minimum area bounding boxes in[13].

3. PROPOSED MODEL

We proposed six levels hierarchal preprocessing model for preprocessing offline handwritten based digital document images. The level numbers indexed as Level 0 to Level 5. Level 0 is a root node of hierarchical preprocessing model.

Level 0: Noise Removal

Level 1: Binarization

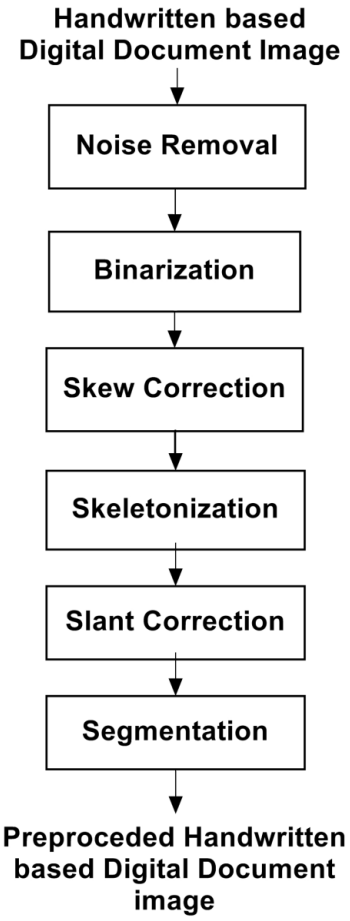

Fig1. *Proposed Hierarchical Pre-processing Model*

Level 2: Skew Correction

Level-3: Skeletonization

Level-4: Slant Correction

Level-5: Segmentation

Fig. 1 shows the proposed hierarchical preprocessing model for offline handwritten based digital document images.

Level-0: Noise Removal

Improper scanning of document or malfunction of camera or low quality of the documents , the digital document added by noise. It disturbance the recognition of the characters and numerals. Common noise are Gaussian noise and salt and pepper noise etc. Using different filters to removing noise on the image. Fig. 2 Shows noised image. Fig. 3 Shows noise removed image. We have Linear smoothing filters and Non-linear smoothing filters are available to remove the noise from the image.

Fig2. *Noisy Image*

Fig3. *After Removal of Noise*

Level-1: Binarization

Binarization is simply extract the foreground data from background data. Digital Document means scanning or photography of the document. Normally digital document in the form of RGB format. For recognition of character means analyse the structure of the character. No need to analyse the colour of the character. The RGB image contains a large format of data that is each byte range is from 0 to 255. But we need to analysis only character structure. Otsu method is one of the method, which is most people used for Binarization. A colour image can be converted to gray scale image. Then gray scale can be converted to binary image. In binary image 0 represents space and 1 represents the part of the character. Therefore with binary numbers it is easy compute and recognizes the characters. To selecting Threshold is very important in Binarization. To convert to Binarization we may use Global Thresholding or Local Thresholding. Fig. 4 shows RGB format of Digital Document Image and Fig 5 shows After Binarization digital document image.

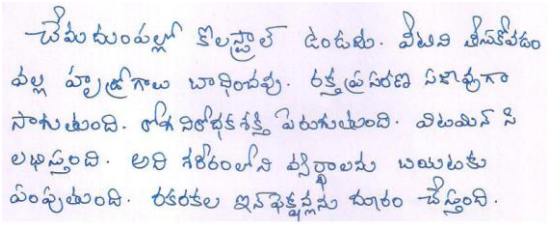

Fig4. *RGB format of Digital Document Image*

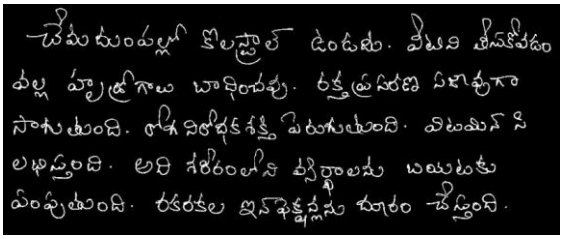

Fig5. *After Binarization digital document image*

Level-2: Skew Correction

Skew correction another important step in preprocessing. Due to the improper scanning angle the document, angle of the document will changed. Normally we have lot of skew correction algorithms are available for entire document skew correction. Most popular skew correction algorithms based on Hough transform method Fig. 6 shows the Telugu handwritten based digital document image before skew Correction. Fig. 7 shows After Skew Correction Telugu handwritten based digital document image. Many of the algorithms implemented for document skew correction based on horizontal projection profiles and vertical projection profiles.

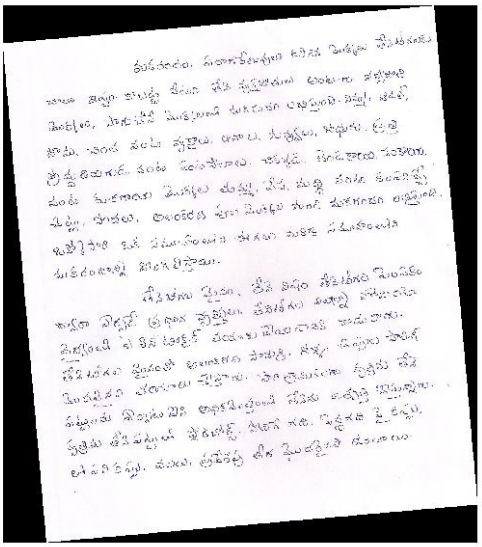

Fig6. *Before Skew Correction*

Another skew correction needed that is Text baseline skew correction. Some of languages have characters in the form of unconnected components. There for Text base line skew correction is not common to all the languages. Depend on language of the characters, the method of text base line skew correction method will change. Some of the skew correction algorithms depend on Morphology operations, some are depends on projection Profiles, some algorithms depends on Nearest-neighbour approach and some algorithms depends on Hough transform etc.

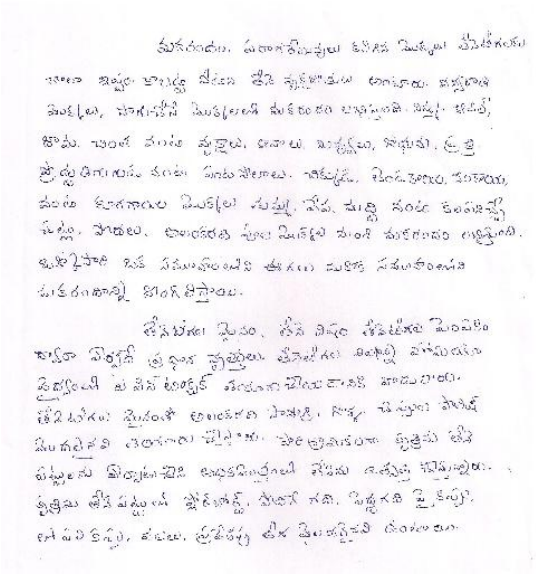

Fig7. *After Skew correction*

Fig8. *Before Skeletonization*

Fig9. *After Skeletonization*

Level-3: Skeletonization

To understand the structure of the character we need more detailed structure. Therefore we recognize the skeleton of the character by thinning process. Fig. 8 shows the Telugu handwritten characters and Fig 9 shows after Skeletonization Telugu characters. 8-neighborhood of the pixel commonly used method for thinning of the image.

Level-4: Slant Correction

After Skeletonization next level is slant correction. Skew correction and slant correction have some similarities. Some of people wrote the characters fall on left or right direction. The slanted character like italic font in word processing. Therefore slant is correction needed for this type of characters. Fig. 10 Shows before Slant Correction handwritten characters. Fig. 11 Shows after Slant correction handwritten characters.

Fig10. *Before Slant Correction*

Fig11. *After Slant Correction*

Fig12. *Overlapped Handwritten Telugu Characters*

Fig13. *After Segmentation Handwritten Telugu Characters*

Level-5: Segmentation

Important and a very complicated level is segmentation. Segmentation process depends on language of the characters. When we want to segment text into characters we need to idea about the language. Major problem in segmentation is overlapped text lines or overlapped text characters. Fig shows the overlapped Telugu based handwritten characters. Fig shows the after segmentation Telugu based handwritten characters. Many of the authors concentrated on printed touched characters. Complex segmentation occurs while the characters were handwritten characters. A proper segmentation method will properly segment the characters even the characters are overlapped or broken. Fig 12 shows overlapped Handwritten Telugu characters and Fig 13 show Handwritten Telugu characters after segmentation.

4. CONCLUSION

In this paper we explained a systematic hierarchical preprocessing model for step by step applying various preprocessing methods to offline handwritten based digital document images. We proposed six levels in hierarchical preprocessing model. However some of the levels have different methods or algorithms need to apply depend on language of characters. We concluded that skew correction for text line, slant correction and segmentation levels are important levels and these applied algorithms may vary depend on language of the characters.

REFERENCES

[1] Kanika Bansal and Rajiv Kumar. "K-Algorithm: A Modified Technique for Noise Removal in Handwritten Documents." *arXiv preprint arXiv:1306.1462* (2013).

[2] A. Papandreou and B. Gatos. "A Novel Skew Detection Technique Based on Vertical Projections", *2011 International Conference on Document Analysis and Recognition*, 2011, pp. 384-388.

[3] Chien-Hsing Chou, Shih-Yu Chu and Fu Chang. "Estimation of skew angles for scanned documents based on piecewise covering by parallelograms", Pattern Recognition, Vol. 40, No. 2, 2007, pp. 443-455.

[4] Yang Cao and Heng Li. "Skew detection and correction in document images based on straight-line fitting". *Pattern Recognition Letters 24*, No. 12, pp. 1871-1879.

[5] L. B. Mahanta and Alpana Deka. "Skew and Slant Angles of Handwritten Signature", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 1 Issue. 9, 2013, pp. 2030-2034.

[6] Changming Sun and Deyi Si. "Skew and Slant Correction for Document Images Using Gradient Direction", 4th International Conf. on Document Analysis and Recognition, Germany, 1997, pp. 142-146.

[7] Uchida, S., Taira, E., & Sakoe, H. Nonuniform slant correction using dynamic programming. In *Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference on* IEEE. 2001, pp. 434-438.

[8] Fajri Kurniawan , Mohd Shafry Mohd Rahim, Daut Daman, Amjad Rehman, Dzulkifli Mohamad, and Siti Mariyam Shamsuddin. "Region-based touched character segmentation in handwritten words." *Int J Innov Comput Inf Control* Vol. 7, no. 6, 2011, pp.3107-3120.

[9] Munish Kumar, M. K. Jindal, and R. K. Sharma. "Segmentation of Isolated and Touching Characters in Offline Handwritten Gurmukhi Script Recognition." *International Journal of Information Technology and Computer Science*, Vol. 6, No. 2, 2014, pp. 58-63.

[10] Veena Bansal and R. M. K. Sinha. "Segmentation of touching characters in Devanagari". In *Proceedings CVGIP, Delhi*, 1998, pp. 371-376.

[11] Srinivasa Rao A. V., Mary Junitha M., Shankara Bhaskara Rao G. and Subba Rao A. V.

"Segmentation of Touching Telugu Characters under Noisy Environment", Journal *of Emerging Trends in Computing and Information Sciences*, Vol. 5, No. 9, 2014, pp. 698-702.

[12] L.Pratap Reddy, T.Ranga Babu, N.Venkata Rao and B.Raveendra Babu. "Touching Syllable Segmentation using Split Profile Algorithm", *International Journal of Computer Science Issues*, Vol. 7, Issue 3, No 9, 2010, pp. 17-26.

[13] J. Bharathi and P. Chandrasekar Reddy. "Segmentation of Touching Conjunct Consonants in Telugu using Minimum Area Bounding Boxes", *International Journal of Soft Computing and Engineering*, Vol. 3, Issue. 3, 2013, pp. 260-264.

[14] Wikipedia contributors. (2015) Hierarchical database model on Wikipedia. [Online]. Available: http://en.wikipedia.org/w/index.php?title=Hierarchical_database_model&oldid=648001695