International Journal of Advanced Research in Computer Science RESEARCH PAPER

Available Online at www.ijarcs.info

Text Line Segmentation on Odia Printed Documents

Debabrata Senapati* Master in Engineering, (Knowledge Engineering) P.G. Dept. of Computer Science & Application, Utkal University, Bhubaneswar, India debabratasenapati@gmail.com

Debananda Padhi Master in Engineering,(Knowledge Engineering) P.G. Dept. of Computer Science & Application, Utkal University, Bhubaneswar, India debananda.padhi106@gmail.com Soumya Mishra P.G. Dept. of Computer Science & Application, Utkal University, Bhubaneswar, India soumyalitun@gmail.com

Sasmita Rout Department of Computer Applications ITER, SOA University Bhubaneswar, India rout_mca_sasmita@yahoo.co.in

Abstract: Line segmentation is one of the important steps of OCR system. In this paper we have proposed a robust method for segmentation of individual text lines of Odia printed documents. Both foreground and background information are used for accurate line segmentation. We have tested our method on documents of Odia scripts as well as some multi-script documents and obtained encouraging result. This technique is based on the intensity of pixel values.

Keywords: Optical Character Recognition, Document Analysis, Line segmentation, Odia Documents

I. INTRODUCTION

The goal of Optical Character Recognition (OCR) is automatic reading of optically sensed document text materials to translate human-readable characters to machinereadable codes. Research in OCR is popular for its various applications potentials in banks, library automation postoffices and defense organizations. Other applications involve reading aid for the blind, library automation, language processing and multi-media design.

The OCR technology took a major turn in the middle of 1950s with the development of digital computer and improved scanning devices. For the first time OCR was realized as a data processing approach, with particular applications for the business world. From that perspective, David Shepard, founder of the Intelligent Machine Research Co. can be considered as a pioneer of the development of commercial OCR equipment. Currently, PC-based systems are commercially available to read printed documents of single font with very high accuracy and documents of multiple fonts with reasonable accuracy. In Optical Character Recognition (OCR), the text lines in a document must be segmented properly before recognition.

The organization of the rest of this paper is as follows: In Section 2, we have discussed properties of Odia scripts considered here. Section 3 details proposed approach, Input and Output of algorithm are displayed in Section 4. Finally in section 5, the paper is concluded.

II. PROPERTIES OF ODIA SCRIPT

Odia (ଓଡ଼ିଆ) is the most popular Indian script used in Odisha. Odisha is the ninth largest state by area in India, and the eleventh largest by population. In Odisha (ଓଡ଼ିଶା) Odia is the official and most widely spoken language with 93.33% Odia speakers according to linguistic survey. The complex nature of Odia alphabets consists of 268 symbols (13 vowels, 36 consonants, 10 digits and 210 conjuncts) [1] among which around 90 characters are difficult to segment and recognize because they occupy special size [1].

ଅ ଆଇ ଈ ଉ ଊ ଋ ଌ ଏ ଐଓ ଔଋ

Figure 1.1: 13 Odia Vowels

କ ଖ ଗ ଘ ଙ ଚ ଛ ଜ ଝ ଞ ଟ ୦ ଡ ଢ ଣ ତ ଥ ଦ ଧ ନ ପ ଫ ବ ଭ ମ ଯ ର ଲ ଳ ଶ ଷ ସ ହ ଡ଼ ଢ଼ ୟ

Figure 1.2: 36 Odia consonants

୦୧୨୩୪୫୬୭୮୯

Figure 1.3: 10 Odia digits

କେ ଖଂଇ ଘେଂ ଅଞ୍ଚ ଜ ମ ବ କ ନ ହ ଦ ଦ
୯୬ ୧୯ ଅଟେ ଅଟେ କେଳେଳେଳେଳେଳେ କେଳେ ଅ
ଔ ଘ ଘ ଭ ଓ ଔ ଘ କ୍ଳ ଜ ଳା ବ ୟ ଘ ଷ ଷ
ଟ ଅବ ଟେ ଅକ୍ୟ ସିଥିୟ ଅବ ଅବ

Figure 1.4: Some of the Odia Conjuncts

A text line may be partitioned into three zones. The *upper-zone* denotes the portion above the mean line, the *middle zone* covers the portion of basic (and compound) characters below the mean line and the *lower-zone* is the portion below the baseline. An imaginary line, where most of the uppermost (lowermost) points of characters of a text line lie, is referred as mean line (baseline). Examples of zoning are shown in figure 2 [2].



Figure 2. Different zones of an Odia text line.

From statistical analysis of 4000 touching components, we note that 72% of the touching components touch at the upper part (near the mean line), 11% touch in the lower zone and 17% touch mainly in the lower half of the middle zone. Based on this statistical analysis and the structural shape of the touching patterns [2], we have designed the segmentation scheme.

III. RELATED WORK

Several techniques for text line segmentation are reported in the literature [7, 8, 9, 10, 11, and 13]. These techniques may be categorized into three groups as follows: (i) Projection profile based techniques, (ii) Hough transform based techniques, (iii) Thinning based approach.

As a conventional technique for text line segmentation, global horizontal projection analysis of black pixels has been utilized in [4, 5, 6, and 7]. Partial or piece-wise horizontal projection analysis of black pixels as modified global projection technique is employed by many researchers to segment text pages of different languages [8]. In piecewise horizontal projection technique text-page image is decomposed into vertical stripes. The positions of potential piece-wise separating lines are obtained for each stripe using partial horizontal projection on each stripe. The potential separating lines are then connected to achieve complete separating lines for all respective text lines located in the text page image. Concept of the Hough transform is employed in the field of document analysis in many research areas as skew detection, slant detection, text line segmentation, etc. Based on Hough peaks, text lines can be separated and many pieces of earlier work can be available on it.

Thinning operation also is used by researchers for text line segmentation from documents [12]. In [12], a thinning algorithm followed by post-processing operations is employed in the entire background region to detect the separating borderlines. Since the thinning algorithm is applied on entire background of text page to obtain separation lines, it requires post-processing to remove extra branches.

IV. PROPOSED APPROACH

In this approach we have collected the printed Odia pages from different site, books, magazines and newspapers. The printed document pages are scanned using a flat bed scanner at a resolution of 300 dpi and stored as gray scale or RGB image in jpg, png, or tiff formats. Before proceeding to the actual line segmentation steps, the first step is binarization, and the second step is skew detection and correction. Then the lines can be segmented properly.

A. Binarization:

A binary image is a digital image that has only two possible values for each pixel. Binary images are also called *bi-level* or *two-level*. This means that each pixel is stored as a single bit (0 or 1). A binary image is usually stored in memory as a two dimensional array.

Image binarization converts an image of up to 256 gray levels to a black and white image. Frequently, binarization is used as a pre-processor before OCR. In fact, most OCR packages on the market work only on bi-level (black & white) images. The simplest way to use image binarization is to choose a threshold value, and classify all pixels with values above this threshold as 0, and all other pixels as 1. Global binarization and locally adaptive binarization are two popular types of binarization methods [3]. We have used threshold valueto convert the Odia document to its binary form is given in figure 3.



Figure 3.1: Printed Odia Character

Figure 3.2: Binary Odia Character

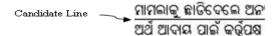
B. Skew estimation and correction:

When a document is fed to the scanner either mechanically or by a human operator to get the digital image a few degree of skew is unavoidable. Skew angle the angle by which the image seems to be deviated from its perceived steady position is the image skew. Skew estimation and correction are important preprocessing steps of line segmentation approaches any documents Skew correction can be achieved by (i) estimating the skew angle, and (ii) rotating the image by the skew angle in the opposite direction [3].

C. Line Segmentation Algorithm:

In this section we used a robust algorithm for segmentation of individual text lines in Odia printed documents. Both foreground and background information is used here for accurate line segmentation. The Algorithm takes input as Odia text printed documents. The printed document is in the form of image file, like jpg, tiff or png formats. It produces the output as text line segment of Odia file. The Algorithm first read the image and then converts it in to gray scaled image.

As we know the region between any two lines is white but the pixels present there are of different intensities. So here we assumed the pixel values ranging from 200 to 255 as white pixels. As the gray scaled images are stored in the memory in the form of matrices and cell values of the matrix varies from 0 to 255, where zero signifies as black and 255 as white. This Algorithm scans through the document and counts the white line between two successive text lines. Choose the candidate line as middle line of two successive text lines, and then replace that with black line.



Original Image

ଲଷ୍ଟି। ୧ ୨। ୧ ୧ (ଏନେନ୍ସି): କଂଗ୍ରେସ ସାଧାରଣ ସଂପାଦକ ତଥା ଆମେଠି ସାଂସଦ ରାକୁକ ରାଜୀକୁ କଡ଼ା ସମାଲୋଚନା କରିଛନ୍ତି ଉଉରପ୍ରଦେଶ ମୁଖ୍ୟମୟା ମାଯାଦତୀ। କଂଗ୍ରେସ ପୁଦରାଇ ରାଜୁକ ରୋଟବ୍ୟାଙ୍କ ରାଜନୀତି ନକରି ବିକାଶମୂଳକ ରାଜନୀତି କରବୁ ଦୋଲି ମାଯା ପରାମର୍ଶ ଦେଇଛନ୍ତି । ରାହୁକ ପୁତି କଥାରେ ରାଜ୍ୟ ସରକାରଙ୍କ ଉପରେ ସେଉଁ ଅସହୋଷ ଓ କ୍ରୋଧ ପ୍ରକଟ କରିଛନ୍ତି, ତାହା ସେ କେନ୍ଦୁ ସରକାରଙ୍କ ବିରୋଧରେ ଦେଖାଇବାକୁ ମାଯା କହିଛନ୍ତି । ରାହୁକ ଜନକକ୍ୟାଣ ପୁସଙ୍ଗ ନେଇ କୌଣସି ଗୁରୁସ୍ ଦେଉନାହାନ୍ତି । ରାଜ୍ୟ ଶାସନ ସଂକ୍ରାଚରେ ଭୁର୍ଭଟକା ତଥ୍ୟ ଦେଇ ଅସହୋଷ ବ୍ୟକ୍ତ କରିବା ପାଂକ ରୋଟବ୍ୟାଙ୍କ ରାଜନୀତି ରହସ୍କୁ ପଦାରେ ପକାଇଛି । ସଦି ପୁକୃତରେ ରାହୁକ ଜନତାକ ପାଇଁ କାଦିବାକୁ ଟାହାନ୍ତି, ତେଦେ ପେଟ୍ରୋଲ ଦରବୃଦ୍ଧି ସହ ଅତ୍ୟାବଶ୍ୟକ ସାମସ୍ଥାର ଆକାଶ ଛୁଆଁ ମୂଳ୍ୟବୃଦ୍ଧି ନିୟରଣ ପାଇଁ କେନ୍ଦୁ ଉପରେ ରାପ ପକାହୁ । ଏହାର ବିକନ୍ତ ଖୋଳବୁ । ଏଥିରେ ଜନତା ଅତିଷ ହୋଇପରିଥିବା ମାୟା କହିଛନ୍ତି । କଂଗ୍ରେସ ନିକର ରାଜଜେତିତ ପାଇଦା ପାଇଁ ସିବିଆର ଓ ବିଏନିକୁ ଅପତ୍ୟବହାର କରିଚାଲିଥିବା ମୁଖ୍ୟମସୀ ଅଭିପୋଗ କରିଛନ୍ତି । ୟୁପି ସରକାରଙ୍କୁ ଛାତି ରାହୁକ କେତ୍ରର ତାଙ୍କ କଂଗ୍ରେସ ଦଳ ପରିଚାଳିତ ସୃପିଏ ସରକାରଙ୍କ ଉପରେ କ୍ରୋଧ ବର୍ଷ କରିବାକୁ ମାଯା ପରମର୍ଶ

Figure 4.1 Printed Odia Document as Input

<u>Algorithm</u>

- Step 1: Read an Odia text document Image
- Step 2: If it is an RGB Image then convert it into gray scale image where the pixel values are from 0 to 255
- Step 3: Store the value of gray scale pixel in a two dimensional array I
- Step 4: For each row of the matrix I do
- Step 5: Count=0
- Step 6: For each row of j^{th} element do
- Step 7: if I(i,j)>200 then
- Step 8: Increment count by one
- Step 9: end if
- Step 10: end for
- Step 11: if count value is equal to the sizeof row then
- Step 12: Increment line by one
- Step 13: Assign bline to zero
- Step 14: else if line>2 then
- Step 15: Increment bline by one
- Step 16: if bline==1 then

Step 17: for each cell of the row do Step 18: Fill the black points at the middle of two Odia line Step 19: end for Step 20: Assign line to zero Step 21: end if Step 22: else Step 23: Increment bline by one Step 24: end if Step 25: end for

Line Segmented Image
ଲକ୍ଷ୍ନୌ, ୧ ୨। ୧ ୧ (ଏଳେନ୍ସି): କଂଗ୍ରେସ ସାଧାରଣ ସଂପାଦକ ତଥା ଆମେଠି ସାଂସଦ ରାହୁକ ଗାହାକୁ
କଟା ସମାଲୋଚନା କରିଛନ୍ତି ଉଢରପୁଦେଶ ମୁଖାମରୀ ମାୟାବତୀ। କଂଗ୍ରେସ ଯୁବରାନ ରାହୁଳ
ରୋଟବ୍ୟାଙ୍କ ରାଜନୀତି ନକରି ବିକାଶମୂଳକ ରାଜନୀତି କରନ୍ତୁ ବୋଲି ମାୟା ପରାମର୍ଶ
ଦେଇଛନ୍ତି । ରାହୁଳ ପୁତି କଥାରେ ରାଜ୍ୟ ସରକାରଙ୍କ ରପରେ ଯେଉଁ ଅସକ୍ରୋଷ ଓ କ୍ରୋଧ ପ୍ରକଟ
କରିଛନ୍ତି, ତାହା ସେ କେହୃ ସରକାରକ ବିରୋଧରେ ଦେଖାଇବାକୁ ମାୟା କହିଛନ୍ତି। ରାହୁକ
ଜନକଳ୍ୟାଣ ପ୍ରବଙ୍କ ନେଇ କୌଣସି ଗୁରୁଦ୍ୱ ଦେଉନାହାନ୍ତି। ରାଜ୍ୟ ଶାସନ ସଂକାରରେ ଭୁଲ୍ଭଟକା
୦ଥ୍ୟ ଦେଇ ଅସରୋଖ ବ୍ୟକ୍ତ କରିବା ତା'ଙ୍କ ଭୋଟବ୍ୟାଙ୍କ ରାଜନୀତି ରହସ୍ୟକୁ ପଦାରେ ପକାଇଛି । ସଦି
ପ୍ରକୃତରେ ରାଜୁଳ ଜନତାଙ୍କ ପାଇଁ କାଦିବାକୁ ଚାହାନ୍ତି, ତେବେ ପେଟ୍ରୋଲ ଦରବୃଦ୍ଧି ସହ
ଅତ୍ୟାଦଶ୍ୟକ ସାମଗ୍ରାର ଆକାଶ ଜୁଆଁ ମୂଲ୍ୟବୃଦ୍ଧି ନିୟନ୍ତଣ ପାଇଁ କେନ୍ଦ୍ର ରପରେ ତାପ
ପକାର୍କ୍ସ ଏହାର ବିକଃ ଖୋଳକୁ। ଏଥିରେ ଜନତା ଅତିଷ ହୋଳପଡିଥିବା ମାୟା କହିଛନ୍ତି।
କଂଗ୍ରେସ ନିଳର ରାଜନୈତିକ ଫାଳଦା ପାଇଁ ସିବିଆଇ ଓ ସିଏନିକୁ ଅପବ୍ୟବହାର କରିଚାଲିଥିବା
ମୁଖ୍ୟମନ୍ତା ଅଭିପୋଗ କରିଛକ୍ତି । ସୂପି ସରକାରଙ୍କୁ ଛାତି ରାହୁଳ କେହୁର ତାଙ୍କ କଂଗ୍ରେସ ଦଳ
ପରିଚାଳିତ ଯୁସିଏ ସରକାରଙ୍କ ଉପରେ କ୍ରୋଧ ବର୍ଷା କରିବାକୁ ମାନ୍ଦା ପରାମର୍ଶ

Figure 4.2 Text Line Segmented Image

V. CONCLUSIONS

In Optical Character Recognition (OCR), the text lines in a document must be segmented properly before recognition. Line segmentation is the preliminary and essential requirement of OCR systems. In this paper we segmented individual Odia text line accurately. The experiment can be further extended to character segmentation and then character recognition through feature extraction.

VI. REFERENCES

- S Mohanty and H N Das Bebartta," A Novel Approach for Bilingual (English - Odia) Script. Identification and Recognition in a Printed Document", International Journal of Image Processing (IJIP), Volume (4): Issue (2)
- [2] N Tripathy and U Pal," Handwriting segmentation of unconstrained Odia text", S⁻adhan⁻a Vol. 31, Part 6, pp. 755–769, December 2006
- [3] Nallapareddy Priyanka, Srikanta Pal, Ranju Mandal," Line and Word Segmentation Approach for Printed Documents", IJCA Special Issue on "Recent Trends in Image Processing and Pattern Recognition", RTIPPR, 2010.
- [4] U. Pal and B.B. Chaudhuri, "Indian script character
- [5] Recognition: A Survey", Pattern Recognition, vol. 37, pp.1887-1899, 2004.
- [6] B. B. Chaudhuri and U. Pal, "A complete printed Bangla OCR system", Pattern Recognition, vol.31, pp.531-549, 1998.
- [7] Vijay Kumar, Pankaj K.Senegar, "Segmentation of Printed Text in Devnagari Script and Gurmukhi Script ", IJCA: International Journal of Computer Applications, Vol.3, pp. 24-29, 2010.

- [8] U. Pal and Sagarika Datta, "Segmentation of Bangla Unconstrained Handwritten Text", Proc. 7th Int. Conf. on Document Analysis and Recognition, pp.1128-1132, 2003.
- [9] K. Wong, R. Casey and F. Wahl "Document Analysis System", IBM j.Res . Dev., 26(6), pp.647-656, 1982.
- [10] F. Hones and J. Litcher, "Layout extraction of mixed mode documents", Machine Vision Application, vol. 7, pp. 237– 246, 1994.
- [11] G. Nagy, S. Seth, and M. Viswanathan, "A prototype document image analysis system for technical journals", Computer, vol. 25, pp. 10-22, 1992.
- [12] H. Yan, "Skew correction of document images using interline cross-correlation", CVGIP: Graph. Models Image Process, vol. 55, pp. 538–543, 1993.
- [13] G. Magy, Twenty years of Document Analysis in PAMI, IEEE Trans. In PAMI, Vol.22, pp. 38-61, 2000.