ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Character Segmentation For Telugu Image Document

M.Shalini, MCA, M.Phil, (Ph.D)
Head, Department of Computer Science,
Kasturba Gandhi Degree & PG College For Women,
West Marredpally Secunderabad - Telangana
shalu_praveenkumar@yahoo.co.in

Dr. B Indira Reddy, MCA, Ph.D
Associate Professor, Department of Computer Science,
Kasturba Gandhi Degree & PG College For Women,
West Marredpally Secunderabad - Telangana
indira.baddam@gmail.com

Abstract

Content line division is one of the significant part of record picture investigation. Content line division is important to recognize all content areas in the record picture. In this paper we propose a calculation taking into account numerous histogram projections utilizing morphological administrators to concentrate elements of the picture. Level projection is performed on the content picture, and afterward line portions are recognized by the tops in the flat projection. Limit is connected to separate the content picture into sections. False lines are disposed of utilizing another edge. Vertical histogram projections are utilized for the line portions and disintegrated into words utilizing edge and further deteriorated to characters. This methodology gives best execution in light of the test results, for example, Detection rate DR (98%) and Recognition Accuracy RA (98%).

Keywords—optical character acknowledgment, division, histogram projection, telugu scripts.

Introduction

Content line division is a key preprocessing stage for acknowledgment in numerous Optical Character Recognition (OCR) frameworks. Division of content line is an indispensable stride in light of the fact that erroneously portioned content lines result in blunders amid acknowledgment stage. Division of the written by hand archive is still a standout amongst the most concerned testing issues. A few strategies for content line division are accounted for in the writing for dividing Indian script records. These strategies incorporate projection profile (white space examination) [1], voronoi and docstrum [2], diagram cut, associated parts based. Division is not exact with these techniques. Jawahar [3] proposed the chart cut technique that requires from the earlier data about the script structure to cut. Rajasekharan proposed a strategy in light of projection technique for Kannada script record division [4]. As an ordinary system for content line division, worldwide flat projection examination of dark pixels has been used in [5, 6, 7, 8]. Fractional or piece-wise flat projection investigation of dark pixels as adjusted worldwide projection method is utilized by numerous analysts to fragment content pages of various dialects [9, 10, 11]. In piecewise level projection procedure content page picture is deteriorated into vertical strips. The positions of potential piece-wise isolating lines are taken for every strip utilizing fractional flat projection on every stripe. The potential isolating lines are then associated with vertical strips to accomplish complete isolating lines for all individual content lines situated in the content page picture.

In this paper a vigorous technique for division of reports into lines and words and the proposed strategy depends on the altered histogram as the Telugu script is extremely intricate. For precise line division Foreground and foundation data is additionally utilized. This technique take considerations of dispensing with false lines and recouping the loss of content in covered content lines

Characteristics of Telugu Script

Telugu is the most mainstream South Indian talked script based dialect. The Telugu character set contains 16 vowels, 36 consonants, vowel (maatras) and consonant modifiers (vaththus). These characters are

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

consolidated to speak to a few as often as possible utilized syllables (evaluated somewhere around 5000 and 10000) in the dialect [12, 13, 14]. We allude to these essential orthographic units as glyphs (single associated segment representation). These characters will have variable size. (i.e. width and stature). In Latin based scripts the greater part of the characters have same size with the exception of few characters. Division of such characters is troublesome when contrasted and Latin based scripts like English. The figure 1 indicates test Telugu basic and compound character pictures.

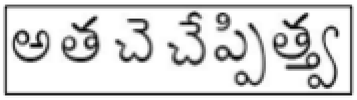


Figure 1 : Examples for simple and compound characters

Proposed Approach

Here we propose another procedure which naturally recognize and section the content line areas of transcribed reports. Figure 2 demonstrates the essential strides in our proposed calculation.

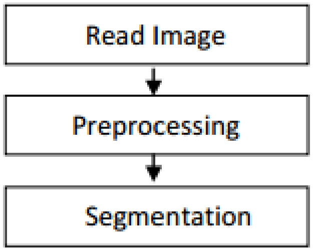


Figure 2 : Shows the basic steps in segmentation algorithm

Preprocessing the crude information is subjected to various preparatory handling ventures to make it usable in the phases of character examination. Pre-handling means to deliver information that are simple for division precisely.

The main objectives of pre-processing include:

- Binarization
- Noise reduction
- Skeletonization/Normalization
- Skew correction

We have utilized paired picture for our work and to change over the first dark level report pictures into double picture, we have connected the calculation because of Otsu [15]. At that point commotion evacuated, skew remedied yield picture from the pre-handling stage is given as contribution to the Segmentation stage. For Noise evacuation we utilize morphological administrators. Figure 3 demonstrates ventures in Noise expulsion.

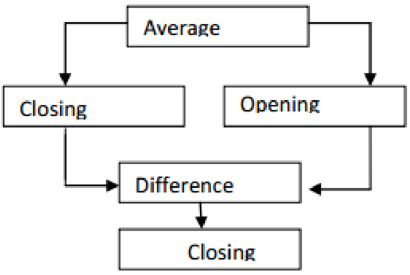


Figure 3 : Steps in Noise Removal

$$E(I(x,y) = \frac{1}{mn} \sum_{i=-\frac{m}{2}}^{i=\frac{m}{2}} \sum_{j=-\frac{n}{2}}^{j=\frac{n}{2}} I(x+i, y+j) \quad S(i,j) \qquad (1)$$

$$I(x,y).S = (I(x,y) \oplus S) \odot S \qquad (2)$$

$$I(x,y) \circ S = (I(x,y) \odot S) \oplus S \qquad (3)$$

$$D(x,y) = I1(x,y) - I2(x,y) \qquad (4)$$

$$T(I(x,y)) = \begin{cases} 255, if \ I(x,y) > T \\ 0 \ otherwise \end{cases} \qquad (5)$$

In Equations 1-5 I(x, y) means the dark level estimation of the pixel situated at position (x, y). S is the basic component of size m x n where m and n are odd qualities bigger than zero. Here we utilize an organizing component of 3x3 in numerical morphological substance. Shutting operation is performed to turn the outskirt of the subsequent picture more minimized and closer. Standardization gives an enormous lessening in information size, diminishing concentrates the shape data of the characters. The report then must skew remedied which is the contribution for Segmentation.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Segmentation:

Once the pre-handling is finished then the histogram projections in y heading are gotten keeping in mind the end goal to perform Line division and after that x histogram projections for words and character division.

$$Profile(y) = \sum f(x,y)$$

Line segmentation

It is the process of identifying lines in a given image. Steps for the line Segmentation is as follows.

1. Scan the preprocessed image horizontally and find the number of ON pixels in each row.
2. Plot the histogram in y direction for the ON pixel count for the image.
3. Scan the histogram projection to find first ON pixel count with zero and remember that y coordinate as y1.
4. Continue scanning the histogram projection then we would find lots of ON pixel counts to be nonzero since the characters would have started.
5. Finally we get the first ON pixel count as zero and remember that y coordinate as y2.
6. Scan the image from y1 to y2 rows for the segmented line.
7. Clear y1 and y2.
8. Repeat the above steps till the end of the histogram.

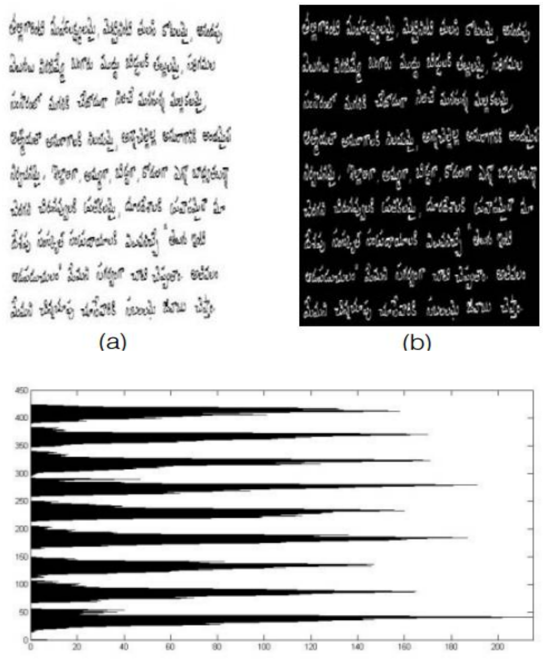
Word Segmentation:

As every word is isolated with a separation, we utilize that idea for word division. Filter the fragmented line picture vertically for word division. Ventures for the line Segmentation is as per the following:

1. Scan the segmented line image vertically and find the number of ON pixels in each column.
2. Plot the histogram in x direction for the ON pixel count for the image.
3. Scan the histogram projection to find first ON pixel count with zero and remember that x coordinate as x1.
4. Keep examining the histogram projection then we would discover loads of ON pixel numbers to be nonzero since the characters would have begun.
5. Finally we get the first ON pixel count as zero and remember that x coordinate as x2.
6. Scan the image from x1 to x2 columns and get the segment word.
7. Clear x1 and x2.
8. Repeat the above steps till the end of the vertical histogram.

Character Segmentation:

Rehash the same calculation characterized in 3.2.2 for fragmenting the word into characters. In step 1 give the info the divided word picture and in step 3 utilize a character isolating separation (as 2) in light of the histogram. In the wake of finishing step 8 we will have the fragmented characters.

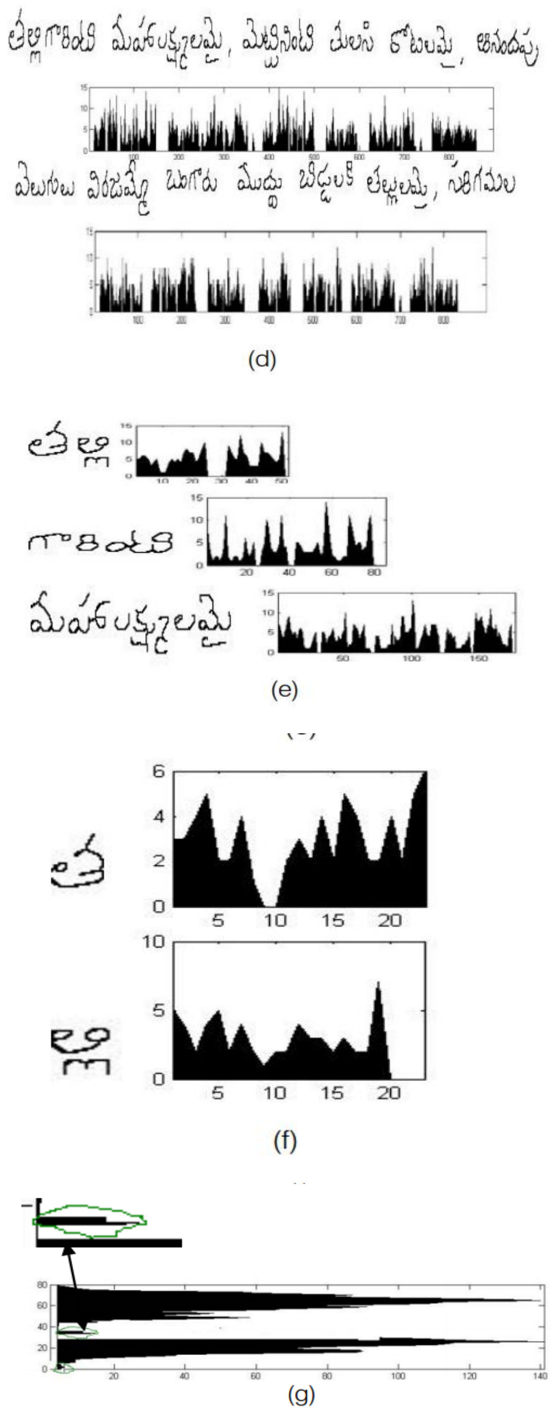


(a)

(b)

(c)

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal



(d)

(e)

(f)

(g)

Figure 4 : Moderate stages: (an) Input mage, (b) Pre-prepared stride, (c) Y histogram projection (d)Text line detachment with level histogram projections, (e) x histogram projections for portioned words.(f) X histogram projection for divided characters. (g)False line.

False Line Exclusion

This strategy tries to bar conceivable commotions near the content line areas. Once the conceivable content lines are divided by expelling a balance from the histogram, we figure the normal tallness of these line districts to bar false lines that may be recognized. Figure 4.g a little crest in the histogram appeared in green, in the event that this area has enough stature it can be mistaken for a content line portion by the calculation. The condition beneath gives the normal tallness of the lines found in a histogram:

$$\sum (ymax - ymin)/Nr$$

Where Ymax is the maximum tallness of the content line locale and Ymin is the start of content area and Nr is the aggregate no of line districts.

The lines with tallness beneath a pre-decided edge are expelled. The estimation of this edge is corresponding to the normal stature of the content lines in the entire picture.

False Word Exclusion :

We will locate the normal tallness of the word in x heading and the word not fulfilling the decided edge will be dealt with as false word.

Performance Evaluation:

The execution is assessed by checking the tally of number of matches between the fragmented substances with that of elements in the ground truth [16].A Match Score table is made where the pixels of the sections and the ground truth are agree. Give I a chance to be the arrangement of all picture focuses, Gj the arrangement of all focuses inside the j ground truth area, Si the arrangement of all focuses inside the i portioned area, T(s) a capacity that numbers the components of set s. Coordinating consequences of the j ground truth area and the i fragment area:

$$\text{Match Score(i,j)} = \frac{T(Gj \cap Si \cap I)}{T(Gj \cup Si \cup I)}$$

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

A coordinated match is utilized if the coordinating score is equivalent to or over the evaluator's acknowledgment edge Ta. On the off chance that G is the tally ground-truth components, S is the tally of result components, and o2o is the quantity of coordinated matches, we ascertain the discovery rate (DR) and acknowledgment exactness (RA) as takes after:

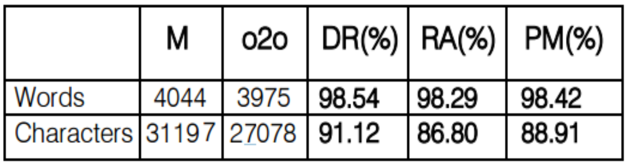
$$DR = \frac{o2o}{G}, \quad RA = \frac{o2o}{S}$$

DR and RA is used to extract the performance metric which is

$$PM = \frac{2DR.RA}{DA+RA}$$

Results and Discussion:

The calculation is actualized in MATLAB. The calculation is tried with a few archive pictures. Test results are appeared in Figure 4. From the examination the proposed technique is quick and dependable to notwithstanding for manually written records which have non-covered lines. The line division precision with DR is 99% and RA is 98% for good quality archives. The constraint of this strategy is that it brought about division blunders for touching characters.



	M	o2o	DR(%)	RA(%)	PM(%)
Words	4044	3975	98.54	98.29	98.42
Characters	31197	27078	91.12	86.80	88.91

Figure 4

Conclusion and Future Work:

In this test, the proposed calculation is tried with a few record pictures. Despite the fact that this calculation gives powerful results it couldn't precisely fragment the covered lines. A heuristic calculation should be considered if there should arise an occurrence of covering lines and words to recoup the misfortune content.

References

1. C. V Lakshmi, C. Patvardhan. (2004): An optical character recognition system for printed Telugu text, Pattern Analysis & Applications, Volume 7, pp. 190-204.

2. Agarwal, David Doermann. (2009): Voronoi++: A Dynamic Page Segmentation approach based on Voronoi and Docstrum features, 10th International Conference, ICDAR.

3. K.S. Sesh Kumar, A. M. Namboodiri, C.V. Jawahar. (2006): Learning Segmentation of Documents with Complex Scripts, Fifth Indian Conference on Computer Vision, Graphics and Image Processing, Madurai, India, LNCS 4338, pp.749-760.

4. B.M. Sagar, DR. G. Shoba, DR. P. Ramakanth Kumar. (2008): Character Segmentation algorithms for kannada optical character Recognition, Proceedings of the 2008 International Conference on Wavelet Analysis and Pattern Recognition.

5. U. Pal and B.B. Chaudhuri, “Indian script character 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 SagarikaDatta, "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. Likforman-Sulem, L., Zahour, A. and Taconet, B., “Text line Segmentation of Historical Documents: a Survey”, International Journal on Document Analysis and Recognition, Springer, Vol. 9, Issue 2, pp.123-138, 2007.

11. U. Pal and P. P. Roy, “Multi-oriented and curved text lines extraction from Indian documents”, IEEE Trans. On Systems, Man and Cybernetics- Part B, vol. 34, pp.1676-1684, 2004.

12. U. Pal, B.B. Chaudhuri. (2004): Indian script character recognition: a survey, Pattern Recognition, 37, 1887 – 1899.

13. B. Anuradhaand, ArunAgarwal and C. RaghavendraRao. (2008): An Overview of OCR Research in Indian Scripts, IJCSES, Vol.2, No.2.

14. U. Pal and B.B. Chaudhuri, “Indian script character recognition: A Survey”, Pattern Recognition, vol. 37, pp. 1887-1899, 2004.

15. N. Otsu. (1979): A threshold selection method from gray-level histograms, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, VOL. SMC-9, NO.

16. I. Phillips, A. Chhabra, "Empirical Performance Evaluation of Graphics Recognition Systems", IEEE Trans. of Patt. Analysis and Machine Intell., Vol. 21, No. 9, September 1999, pp. 849-870.