Role of Skeletonization Techniques for Recognition of Odia Off-Line Characters

Sagarika Mishra, Chandra Sekhar Panda

Abstract*: *Many character recognition methods use a thinning stage which facilitates shape analysis and stroke identification. There are several thinning algorithms reported in the literature of OCR. Odia, one of the Indian languages, has structurally different scripts than other Indian languages. Unlike in other Indian scripts, there has not been any special effort by OCR researchers to develop recognition techniques of this unique structural character set. In this paper, some major thinning algorithms are examined with Odia character set and show that these algorithms exhibit some sort of deficiencies when applied to Odia characters and the vital features of the character is not retained in the process. We propose a new parallel thinning technique that preserves all important features of the script.

Keywords* : *Odia Scripts, Thinning, Segmentation, OCR.

I. INTRODUCTION

Character recognition is becoming extremely important in digital preservation of cultural heritage. Several recognition techniques for printed characters of different Indian scripts have been reported in literature The importance of OCR for Indian languages cannot be underestimated as massive volumes of print-media resources representing the cultural and historical heritage of India are available in several Indian languages (scripts). *Odia* is one of the scheduled languages of India. It is spoken by over 30 million people residing in the state of Odisha and its neighboring states. *Odia* language has a rich heritage and *Odia* script is found in stone engravings, copper plates, and palm-leaf manuscripts. Like Devanagari scripts and Bangla scripts, Odia script derives its origin from Brahmi scripts. But unlike other scripts, *Odia* scripts are mostly circular in nature and do not have top horizontal line. This is possibly due to influence of Dravidian scripts as well as to facilitate palmleaf writing. But unlike many Dravidian scripts, the circular shapes of *Odia* scripts have bottomward openings.

Revised Manuscript Received on November 25, 2020.
Manuscript Received On October 06, 2020

***** Correspondence Author
Sagarika Mishra, Department of Computer Science, SUIIT, Sambalpur University.
Dr Chandra Sekhar Panda, Associate Professor and Department of Computer Science & Applicationm Sambalpur University.

The vowels, depending on their position in a word, take different shapes, called *modifiers* or, *matra*. Vowel modifiers, *matra*, when attached to a consonant depending on the vowel, are placed to the left, right (or on both sides), top or bottom of the consonant [1]. Normally, *matras* do not disturb the shape of

the basic characters to which they are attached. Moreover, two or more basic characters combine to form a new complex shaped character, called a *compound character*.

The curve and circular nature of the script, the absence of the upper line connecting the characters in a word are some of the features that demand a separate technique for recognition of such characters. The basic character-set, modifiers (*matras*) and numerals are shown in Figure 1.

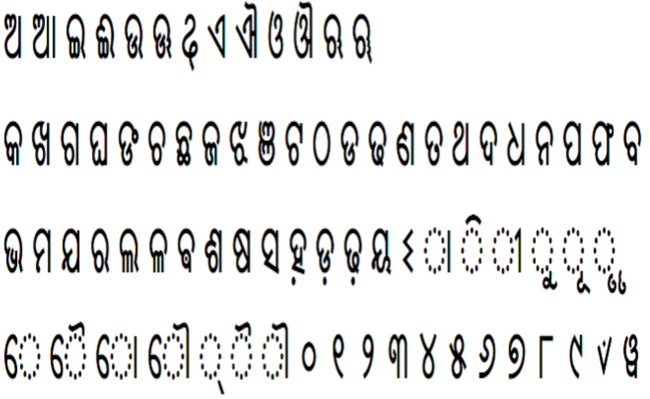


Figure-1(Odia Character Set)

Mask-based iterative thinning algorithms delete successive layers of pixels on the boundary of the pattern until only a skeleton remains. The deletion (or retention) of a particular pixel depends on satisfaction of certain condition specified by the thinning-mask which captures the local neighborhood distribution of black pixels. These masks aim at identifying boundary pixels that are likely to be crucial in preserving geometric and topological properties[2]. According to the way they examine pixels, these algorithms can be classified as sequential or parallel. In a parallel algorithm, the decision of deletion of a pixel depends purely on data of previous iteration and not on any data of the current iteration. Thus, pixels are examined independently in a parallel manner in each iteration. On the other hand, in sequential algorithm pixels are examined in a fixed sequence [3].

The paper is organized as follows. In Section II, suitability of various thinning algorithms for skeletonization of Odia characters is examined. In Section III, a new thinning algorithm is proposed and results are described.

II. EXISTING METHODOLOGY

Many thinning algorithms have been developed in the past few years for general application as well as for applications in OCR. A well-known algorithm, proposed by Zhang and Suen [4], iteratively run two sub iterations. ZS algorithm makes neighbour connectivity and successfully removes the boundary noise points as shown in figure-3 (a1) .But as shown in figure-3 (a2) slanting line does not have 1-pixel width line.. Hol et al. [5] propose an improvement which requires examining larger mask but gets over the complexity of having two subiterations.

Chen and Hsu[6] propose an improvement of original Zhang and Suen algorithm and use a look up table to speed up the process. Lu and Wang (LW) algorithm [7] is a modified version of ZS algorithm where the range of values for B(p) is [3, 6] instead of [2, 6]. As shown in figure-3(b1), LW algorithm preserves the vertical and horizontal lines. But the junction points and end points are not preserved properly as shown in figure-3 (b2).

ETA algorithm[8] improves the ZS algorithm by adding some additional post processing condition after completion of all iteration. ETA algorithm preserves the structure of the image and can remove the contour noise points as shown in figure-3 (c1). But as shown in figure-3 (c2) slanting lines have 2-pixel width line. Datta and Parui [9] propose a thinning algorithm that ensures unit-pixel thick skeleton but has four sub iterations. In ZW algorithm the shape of curve and lines are better maintained as shown in figure-3 (d1). But as shown in figure-3 (d2), it cannot preserve the end points and junction points. It also does not remove contour noise points. Robust algorithm preserve the neighbor connectivity as shown in figure-3 (e1) . But it is also very sensitive to contour noise and cannot preserve the end points, junction points as shown in figure-3 (e2). BM-99[10] algorithm maintains shape of curve as shown in figure-15 (a). But as shown in figure-3(f1), it cannot make 1-pixel width line. Also does not remove the contour noise points shown in figure-3(f2).

This paper is concerned with thinning of printed Odia characters. It is observed that most of the existing thinning algorithms do not perform well for Odia characters. We examine some major algorithms in this investigation and list the shortcomings of each of these algorithms. A new algorithm is presented here that exploits the basic features of the language and performs better than any of the earlier algorithms.

III. PROPOSED METHODOLOGY

It is assumed, in this paper, that input to any thinning process is a binary image of an isolated character. A pixel *p* is examined for deletion if it is a black pixel with value 1.The neighbor pixels in its 3×3 window are labeled as shown in Table-I. The pixels $x_2, x_3, ..., x_9$ are 8-neighbours of *p* and are collectively denoted by *N(p).* The number of black pixels in *N(p)* is denoted by *B(p).* A(p) is the number of (0-1)patterns in the ordered set $x_2, x_3 ..., x_9, x_2$

TABLE- I: 3 X 3 WINDOW MASK

x_9	x_2	x_3
x	p	x_4
x_7	x_6	x_5

In the proposed algorithm 3 steps are done:

- Two sub iterations of ZS algorithm
- A new mask is added for neighbour connectivity.
- Vertical stroke is added

ZS algorithm [11] works in two sub iterations as follows.

1st Sub-Iteration:

The pixel p is deletable if it satisfies the following conditions:

- $2 \leq B(p) \leq 6$
- $A(p) = 1$
- $x_2 \times x_4 \times x_6 = 0$
- $x_4 \times x_6 \times x_8 = 0$

2nd Sub-Iteration:

The pixel p is deletable if it satisfies the following conditions:

- $2 \leq B(p) \leq 6$
- $A(p) = 1$
- $x_2 \times x_4 \times x_8 = 0$
- $x_2 \times x_6 \times x_8 = 0$

Here an extra restoring mask is added for preserving the neighbor connectivity shown in Figure – 2.2. After completion of all iteration a post processing condition is checked to get the final skeleton image. No previous algorithms are able to give a thin vertical line in odia character hence an extra vertical stroke is added to those characters that have vertical lines.

TABLE- II: EXTRA RESTORING MASK

		0	0	0
		0	0	0
	1	P	0	0
0	0	1	1	
0	0	0		

Pseudo code of the proposed methodology as follows..

```
Repeat
        Set a flag to false
        For a set of black pixel X
        counter = 1
        IF counter = 1
          Repeat
              Y = set of pixel in X that satisfies conditions of sub-iteration1
              X = X \ Y
                  Until Y = ϕ
          Repeat
              Z = set of pixel in X that satisfies conditions of sub-iteration2
              X = X \ Z
          Until Z = ϕ
        counter = counter + 1
        ELSE
          Repeat
                        W = set of pixel in X which match any one of the thinning mask but no restoring mask
                        X = X \ W
          Until W = ϕ
  counter = counter + 1
```

```
        ENDIF
        Set the flag to True
        Repeat
          $V$ = set of pixels that satisfies post processing conditions
          $X = X \setminus V$
        Until $V = \phi$
Until flag is not set
IF vertical line is present
        Draw vertical stroke
ENDIF
```

IV. RESULT AND DISCUSSION

To evaluate the effectiveness of the proposed algorithm, the comparison and experimental results are described here. In figure-1 [a1,a2,b1,b2,c1,c2,d1,d2,e1,e2,f1,f2] the result of the existing characters is compared with ZS,LW, ETA, ZW ROBUST and ZW algorithms. Our dataset consists of isolated characters of 2000 characters of several fonts and sizes. Documents are scanned at 300dpi with high-resolution scanner. Characters are isolated by connected component method. These are size-normalized to 4848. In the present experiments, only the alphabets and compound characters are taken leaving aside numerals and modifiers. k is 2% of image size and hence k =1.

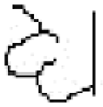
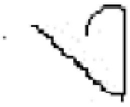
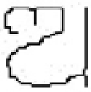
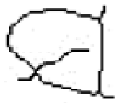
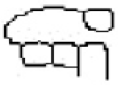
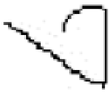
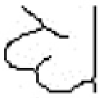
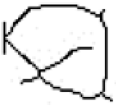
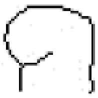
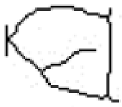
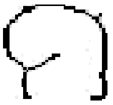
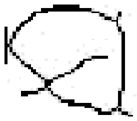
Algorithms	Successful Thinning	Unsuccessful Thinning
ZS	 (a1)	 (a2)
LW	 (b1)	 (b2)
ETA	 (c1)	 (c2)
ZW	 (d1)	 (d2)
ROBUST	 (e1)	 (e2)
BM99	 (f1)	 (f2)

Figure-I (Results of various thinning Algorithm)

In Figure-II, our proposed algorithm's experimental results are shown including its original characters. It is found that most characters able to preserve 1 pixel width line. Topological structure of the characters is well preserved including vertical stroke, slanting line and also elliptical shape of the Odia characters.

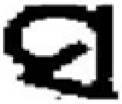
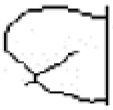
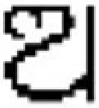
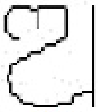
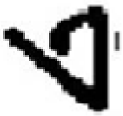
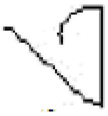
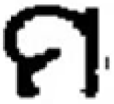
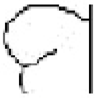
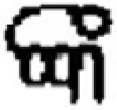
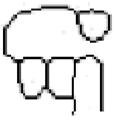
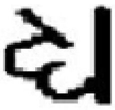
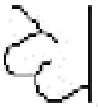
Original Characters	Proposed Thinning Characters
	
	
	
	
	
	

Figure-II (Results of Proposed Thinning Algorithm)

From our extensive experimentation, we observe that BM99, LW and ZW are not able to remove contour noise. BM99, LW and GH92 yield skeletons of 2-pixel width. H tends to remove tips of linear strokes. Robust algorithm often fails to preserve shape and topology. ETA does not preserve vertical or diagonal lines and sensitive to contour noise points. LW does not preserve junction or end points. Robust algorithm is also sensitive to contour noise.

V. CONCLUSION

The present work is concerned with investigating existing thinning algorithms for their suitability to a specific script, namely *Odia* script. It is observed that none of the existing algorithms is very suitable and has some shortcoming or other. By combining some steps of different algorithms, a new algorithm is proposed that not only preserves the desirable features like shape, topology, connectivity etc but also retains basic strokes.

This is very useful for subsequent recognition process. A new concept of stroke correction is also introduced here. We propose to integrate this with our recognizer in future and extend the concept of stroke correction to many other basic strokes.

REFERENCES

1. Pushpalata, P.,& Babita , M. (2015). A Survey on Odia Character Recognition. *International Journal of Emerging Science and Engineering*, *8*(2), 2319–6378.
2. Pal, U., Wakabayashi, T., & Kimura, F. (2007). A system for off-line oriya handwritten character recognition using curvature feature. *Proceedings - 10th International Conference on Information Technology, ICIT 2007*, 227–229.
3. Chaudhury, B., Pal, U.,&Mitra,M. (2002). Automatic Recognition of Printed Oriya Script.*RESEARCH GATE*, *27*(January), 23–34.
4. Zhang TY, Suen CY (1984) A fast parallel algorithm for thinning digital patterns. *CACM* 27:236-239
5. Zhang YY, Wang PSP (1988) A modified parallel thinning algorithm. *CPR* 1023-1025
6. Hall RW (1989) Fast parallel thinning algorithm: parallel speed and connectivity preservation. *CACM* 32:124-131
7. Lu HE, Wang PSP (1985) An improved fast parallel algorithm for thinning digital pattern. *IEEE CVPR'85* 364-367
8. Kwon JS, Gi JW, Kang EK (2001) An enhanced thinning algorithm using parallel processing. *ICIP* 3:752-755
9. Datta A, Parui SK (1994) A robust parallel thinning algorithm for binary images. Pattern Recognition 27:1181:1192K. Elissa, "Title of paper if known," unpublished.
10. Bernard TM, Manzanera A (1999) Improved low complexity fully parallel thinning algorithm. *ICIAP* :367-374

AUTHORS PROFILE

Sagarika Mishra is perusing her Ph.D in the Department of Computer Science, SUIIT, Sambalpur University. She has a teaching experience of more then 10 years and currently working as an academic consultant (Comp Sc) in Odisha State Open University(OSOU). Her research area includes Character Recognition, Computer Vision, OCR, Image Processing, Natural Language Processing.

Dr Chandra Sekhar Panda is serving as an Associate Professor and HOD in the Department of Computer Science & Applicationm Sambalpur University.He has a vast teaching & research experience of more than 20 years. There are more then 50 students have completed their M.Phil dissertation under his guidance. His research area includes Image Processing, Segmentation, Feature Extraction, Computer Vision, Natural Language Processing, Video Motion Analysis, Neural Networks etc.