MVA2005 IAPR Conference on Machine VIsion Applications, May 16-18, 2005 Tsukuba Science City, Japan

8-32

Thresholding, Noise Reduction and Skew correction of Sinhala Handwritten Words

M.L.M Karunanayaka
University of Colombo School of Computing
No.35, Reid Avenue,
Colombo07,Sri Lanka
mlm@ucsc.cmb.ac.lk

C.A Marasinghe
Software Engineering Lab
University of Aizu
Aizu-Wakamatsu
Fukushima, Japan
ashu@u-aizu.ac.jp

N.D Kodikara
University of Colombo School of Computing
No.35, Reid Avenue,
Colombo 07,Sri Lanka
ndk@ucsc.cmb.ac.lk

Abstract

The Sinhala script, which is generally with round characters, is unique among other Brahmi-descended scripts and is used by 70% of the 18 million populations in Sri Lanka. There has been no published research on the cursive unconstrained Sinhala handwriting recognition. This paper proposes vital preprocessing stages, which are categorized under thresholding, noise removal, skew detection and correction algorithm and is useful to improve accuracy of segmentation and recognition of the Sinhala handwritten words.

This paper introduces a novel skew detection method based on least square method and also robust indirect skew correction method of unconstrained cursive Sinhala words. Threshold selection is used in combination with three methods such as QIR, NIR, and analyzing gray level intensity distribution. Median filtering algorithm and connected component analysis method are used to reduce the noise in Sinhala handwritten word images and used vertical projection profile histogram based on validation technique to improve the noise reduction of the image.

In this proposed system over 700 handwritten patterns in Sinhala handwritten real postal addresses in NSF database [3] were tested and the reported accuracy was 97.2% and 500 handwritten patterns, which were written by undergraduate students, were tested and the reported accuracy was 99.6%. Overall accuracy of 98.4% was reported using these two types of handwritten patterns.

1. Introduction

Off-line recognition of handwriting has numerous practical applications in the areas such as mail sorting, banking, census, commerce, etc. There are many techniques available in the computational pattern recognition such as artificial neural networks and statistical approaches such as Hidden Markov Models to recognize handwritten words or isolated characters. By nature, handwriting is very unsteady in shape and quality of tracing but cursive handwriting variability is not only due to writer's style and quality of paper but also due to geometric factors determined by the writing condition. Correction of these factors can be helpful to reduce the variability of handwriting. These corrections are known as preprocessing and it will help to improve the accuracy of segmentation and recognition methods. The crucial landmarks of the preprocessing in handwriting systems are thresholding or binarization, noise removal, skew detection and correction.

Thresholding or binarization methods are used to separate foreground object into the image. Most of the threshold techniques can be classified into two categorizes, that is Global thresholding techniques and Adaptive (or local) thresholding techniques [11]. Global thresholding methods apply one threshold value to the entire image while adaptive (local) thresholding methods apply different threshold values to different regions of the image. The adaptive (local) threshold value is determined by the neighborhood of the pixel to which the thresholding is being applied. In many cases of document processing adaptive (local) thresholding techniques were effectively applied to separate foreground and background of images because these methods can work well with variable illumination, shadows, smears and smudges documents.

In the image processing research noise is described as any unwanted information containing digital image. Digital image acquisition process, which converts an optical image into continuous electronical signal, is the primary source of introducing the noise into the electronic images and noises are introduced into the image during the transmission process. The most common types of noises effected in images are Gaussian noises and impulse noise [4]. Gaussian noises are introduced during acquisition process of the image. Gaussian noise is characterized by each image pixel a value from a zero-mean Gaussian distribution. Impulse noise is characterized by replacing a portion of an image's pixel value with random values, leaving the remainder unchanged. In handwritten words there are three common types of noises. Theses noises are named as background noise, shadow noise and slat and pepper noise.

Skew or slope is the angle between the horizontal direction and the direction of the line on which the writer aligned the word and minimization of this angle($\approx 0^{o}$) or in other words word aligned horizontal direction is known as skew correction of the handwriting words. There are two possible ways of skewing handwriting. The first is, when a document is fed to the optical sensor either mechanically or by a human operation, a few degree of skew is unavoidable. Second is, skewed handwritten words in an original document. Several techniques are used to estimate skew angle in handwriting word images. Such techniques are Chain code [6], projection profile [1],[7],[12], and modified versions of projection profile

such as density distribution[12] ,principle component analysis[10] and generalized projections[7], Hough transform[1],[2],Fourier transform-based methods[2], and nearest neighbor clustering based approach[2], Wigner-Ville distribution[5] and least square [14].Most of skew correction techniques are pixel-oriented. These pixels – oriented techniques are again categorized two different groups that are direct [14] and indirect [1] methods depending on the type of rotational equation used.

Organization of this paper is as follows; Section two of this paper describes the background of the Sinhala handwriting and its styles. The previous research done in Sinhala character recognition is also described in this section. Section three of this paper describes the methods used in preprocessing Sinhala handwriting words. The evaluation of the results of the present work is given in the section four. Finally, the conclusions of the present work is summarized in section five.

2. Background

Sinhala language is used by over 70% of the 18 million populations in Sri Lanka [13]. Being a descendant of a spoken form (Pali) of the root Indic language, Sanskrit, it can be argued that it belongs to the large family of Indo-Aryan languages [8].

Sinhala language is written from left to right pattern and it has curved shape scripts. Characters are written in the three horizontal layers that is the upper layer, the middle layer and the lower layer. Some characters are written across all three layers, some of them are written only in the middle layer and the other sets of characters occupy either upper and middle or middle and lower layers.

There has been a little research carried out on Sinhala handwritten characters [8]. Almost all those research work is focused on identifying regular, well-defined Sinhala handwritten character recognition. There has been no published research done on the recognition of cursive, unconstrained handwritten characters.

3. Methodology

The procedure of preprocessing a Sinhala handwritten word is explained in this section. Section 3.1 describes thresholding techniques used in proposed system. Noise removal techniques are described in section 3.2 and the skew detection and correction methodology of the Sinhala handwriting word is described at the end. Flow of the proposed system is shown in figure 1.

3.1. Thresholding

Thresholding method, which is introduced in this paper, aims to find accurate separation point of foreground and background of the gray level image. This proposed thresholding algorithm is a combination of selected threshold algorithms. These algorithms use three threshold algorithms, which are Native integral ratio technique (NIR)[9], Quadratic integral ratio technique (QIR)[9] and gray level intensity distribution [1] of given image. NIR and QIR techniques are global two stage threshold selection approaches. In the first stage the algorithm divides an image into three subimages: foreground, background and fuzzy subimage where it is hard to determine whether a pixel actually belongs to the foreground or background (figure 2).

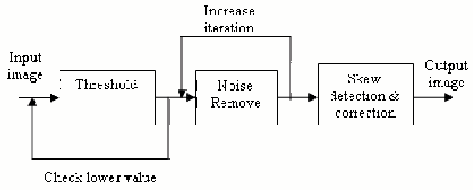


Figure 1. System flow diagram

In the second stage calculate the best separation point between A and C (shown in figure 2) using native integral ration or quadratic integral ratio. Using the intensity distribution histogram calculate the foreground and background peak values and the average value of these two. The propose algorithm is shown in the equations below.

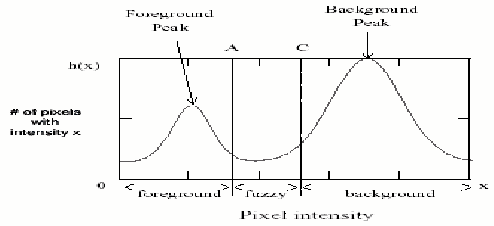


Figure 2. Intensity histogram

NIR threshold = threshold_NIR
QIR threshold = threshold_QIR
AverageMaxPeaks = threshold_MaxPeak
threshold_MaxPeak = (foreground peak + background peak) / 2
Using these three threshold values calculate the average threshold value.
Average threshold value = (threshold_NIR+threshold_QIR+AverageMaxPeaks) / 3

Binarized the image by using the threshold value. Next step of this threshold selection algorithm is the calculation of the vertical projection profile histogram and check whether it is abnormal (shown in figure 3(a) and 3(b)). If the abnormal shape occurs, then choose the lower threshold value next to average threshold value of the above three threshold values and process it until the vertical projection profile histogram is normal. If the vertical projection profile histogram cannot be set to normal, the process then turn off.

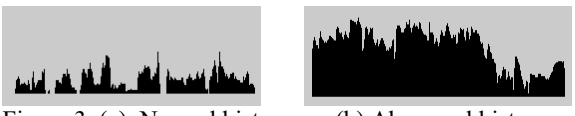


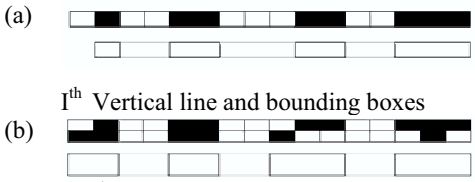
Figure 3. (a). Normal histogram, (b) Abnormal histogram

3.2. Noise removing

Noise removal algorithm in the proposed Sinhala handwritten word is based on the median filtering technique [12]. The kernel of this algorithm is in the shape of " + " and calculate median of five black pixels in the kernel and replace the original gray value of middle pixel with the calculated median value. This process continues with all the pixels in the word image. Example of noisy image is shown in figure 4(a), and after applying the median filtering based noise removing algorithm the image appears as in figure 4(b).

Figure 4. (a)Noise removing kernel and noisy image (b) Noise Removed image

The next procedure in the proposed system is to apply connected component based analysis to remove unwanted objects in the image[1]. Connected components are the rectangular boxes bounding together with connected black pixels. The algorithm used to obtain the connected components is a simple iterative procedure which compares successive vertical scan of an image to determine whether black pixels in any pair of vertical lines are connected together. Bounding rectangles are extended to enclose any groupings of connected black pixels between successive lines; figure 5 shows how it work in this procedure.



(a) I^{th} Vertical line and bounding boxes

(b) $I+1^{th}$ Vertical line and bounding boxes

Figure 5. Connected component analysis process

After calculating the bounding boxes in each connected component measure the width and the volume of bounding boxes. If the volume of the bounding box is less than or equal to 5% of the image volume (i.e. image height multiplied by image width) it is assumed as noise object in the image, and removed. If the bounding box width is greater than 25% of the image width, this bounding box is assumed as underline or any other straight line in the image, and then as it is a noise object it is removed from the image.

3.3.Skew detection and correction

This section describes the proposed novel skew detection algorithm and the robustness skew correction algorithm. Firstly, the Sinhala word image is divided into vertical groups; the width of each group is ten pixels. Then the proposed algorithm starts to search black pixel from bottom to top of the image using kernel 10 x 1. If the kernel meets the black pixel it stops moving and then identify at which height the black pixel is found. To further the process set the coordinates (X,Y) in the founded black pixel using equation 1. This process will be continued to other vertical groups in the image as well. Completely processed image is shown in figure 6(a).

```
kernelwidth = 10;
for( i =0; i<W; i+= kernelwidth){
        X =i+ kernelwidth/2;
}
                                                    (1)
for( j=H-1; j>=0;j--){
        If( I(x,y) == blackpixel){
                Y= H;
        }
}
```

Calculated coordinate is (X,Y) , I(x,y) represent the image. Image height and width are represented as H, W respectively.

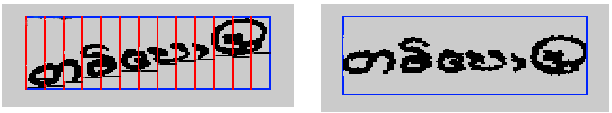


Figure 6. (a)Skew angle calculation, (b) Skew Corrected Image

Most of the vertical groups has these calculated (X,Y) coordinates except in groups which are not included in character stroke. Using these (X,Y) coordinates the best fit line(see figure 7) is calculated. This best fit line is constructed by using least square method, shown is equation 2 and 3[2],[12].

Figure 7. Calculate best fit line

Assume best fit line is, y = A + Bx and A and B of the line is known as intercept and gradient respectively of the line calculated using the equation 2 and 3 respectively.

$$B = \frac{n\sum_{i=1}^{n} X_i Y_i - \left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)} \quad (2)$$

$$A = \frac{\sum_{i=1}^{n} Y_i - b\sum_{i=1}^{n} X_i}{n} \quad (3)$$

Correction of the skew angle of Sinhala word image will be done by using an indirect method [1],[12]. For a pixel $(x',y')^t$ in the word image, the indirect method finds the corresponding pixel $(x, y)^t$ in the original image, and then sets a value of $(x', y')^t$ with that of $(x,y)^t$. The correspondence is computed by applying the inverse rotational matrix to $(x', y')^t$ using equation 4.

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{pmatrix} \cdot \begin{pmatrix} x' \\ y' \end{pmatrix} \qquad (4)$$

Advantage of using indirect rotational method reduce the rounded problem which means that the neighboring connected components are merged and the isolated components are split. Completed skew corrected image is shown in figure 6(b).

4. Results and Evaluation

The proposed methods were applied for the Sinhala handwritten real postal addresses and for some selected words, which were written by different students of the University of Colombo. The Sinhala handwritten database which is available in NSF [3] in Sri Lanka is one of the sources of Sinhala handwriting real postal addresses used to train and test the proposed methods. The advantage of using real postal addresses is, that it covers one of the best samples of the population. In these situations, real postal addresses are the best source of various handwriting samples, which can be used for training and testing procedures. In the proposed system two different kinds of handwritings were tested. Those cover the real postal addresses (RPA) and words written by the students of the same education level (WSEL). Results are shown in table 1. Overall success rate of the proposed method is 98.4%.

Table 1. Results of overall process.

	No of Patterns	Success skew correction	Success(%)
RPA	700	680	97.2%
WESL	500	498	99.6%
Total	1200	1178	98.4%

5. Conclusions

In this paper we have proposed robust approaches for preprocessing of cursive unconstrained Sinhala handwriting. The experimental results show that both the thresholding approach based on NIR, QIR, gray level intensity distribution and the noise removal approach based on median filtering and connected component analysis are accurate. The proposed novel skew detection method and least square method will estimate accurate skew angle of Sinhala word images and the use of indirect rotation method effectively correct skew angle of the cursive words with high accuracy.

The research project will be continued to implement an automated postal address recognition system in the future.

Acknowledgements

The second author's work was supported by grants from the Japan Society for the Promotion of Science (JSPS - P04289).

References

[1] A.Amin, S.Fischer, T.Parkinson and R.Shiu, "Fast algorithm for skew detection", *Proceedings of the IS&T/SPIE Conference on Real-time Imaging*, 1996, pp. 65-76.

[2] Y.Cao and H.Li,"Skew detection and correction in document images based on straight-line fitting", *Pattern Recognition Letters*, 2003, vol. 24, pp. 1871-1879.

[3] H.C. Fernando, N.D Kodikara and Hewavitharana,"A Database of handwritten Text Recognition Research in Sinhala Language", Proceedings of the Seventh International Conference on Document Analysis and Recognition, pp. 1262-1264,2003.

[4] R.Garnett , T. Huegerich, C. Chui and W. He,"A New Framework for Removing Gaussian and Impulse Noises", viwed 6 September 2004, http://www.cs.umsl.edu/~chui/publ/GHCHnoise.pdf.

[5] E.Kavallieratou, N.Fakotakis and G.Kokkinakis,"New Algorithms for Skewing Correction and Slant Removal on Word-Level", *6th IEEE International Conference on Electronics, Circuits and Systems*, 1999, vol. 2, pp. 1159-1162.

[6] G. Kim and V. Govindaraju, "A Lexicon driven approach to handwritten word recognition for real-time applications", *Proceedings of IEEE Transactions on pattern analysis and machine intelligence*,1997,vol 19, no. 4, pp. 366-379.

[7] G. Nicchiotti and C.Scagliola ,"Generalized Projections: A Tool for Cursive Handwriting Normalization", *Fifth International Conference on Document Analysis and Recognition*, Bangalore, India, 1999, pp. 729-732.

[8]. H.L. Premaratne and J. Bigun .Recognition of Printed Sinhala Characters Using Linear Symmetry. Symmetry. *5th Asian Conference on Computer Vision*, pages 23-25, 2002.

[9] Y.Solihin and C,G.Leedham, "Noise and Background Removal from Handwriting Images", *proceeding of the 1997 IASTED International Conference on Intelligent Information Systems*, 1997, pp. 366-370.

[10] T.Steinherz , N.Intrator and E.Rivlin ,"Skew Detection via Principal Components Analysis", *Fifth International Conference on Document Analysis and Recognition*, Bangalore, India, 1999, pp. 153-156.

[11] O.D.Trier and A.K. Jain,"Goal-Directed Evaluaction of Binarization Methods", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1995,vol. 17, no. 12.

[12] A. Vinciarelli and J. Luettin ,"A New Normalization Technique for Cursive Handwritten Words", *Pattern Recognition Letters*,2001,vol. 22,no. 9,pp. 1043-1050.

[13]R.Weerasinghe. A Statistical Translation Approach to Sin hala-Timil Language Translation. *5th International Inform mation Technology Conference*, pages 136 – 141,2003.

[14] C.L.Yu, Y.Y.Tang and C.Y.Suen, "Document skew detection based on the fractal and least square method", *Proceedings of the Third International Conference on Document Analysis and Recognition*, Montreal, Canada, 1995, pp. 1149-1152.