LINE AND SKEW REMOVAL FROM OFF-LINE CURSIVE HANDWRITTEN WORDS

Amjad Rehman Khan, Dzulkilifli Muhammad and Fajri Kurniawan
Department of Computer Graphics and Multimediam, University Technology Malaysia, Malaysia.

ABSTRACT

This paper presents two new preprocessing techniques for line and skew removal in unconstrained cursive handwritten words inherited at any angle. The former algorithm based on the connected component analysis of black run up to certain length threshold. The detected line is removed by keeping track of junctions to avoid characters shatter. The latter based on mathematical formulation of skew angle by detecting upper and lower boundary points of the skewed word. By hypothesizing a line between two boundary points, angle for skew is estimated. The algorithms are tested on IAM benchmark database. Promising results are reported for the novel techniques.

KEY WORDS: Skew detection; underline removal, preprocessing, character segmentation, word recognition.

INTRODUCTION

In real world applications, documents of prescribed shapes are daily used in public and private institutions. Where the user input is in form of unconstrained cursive handwriting on static surface (Dimauro et al., 1997) such as checks, mail order forms, tax forms, admission forms etc. Handwriting varies in slant and skew as well as characters overwritten on lines of forms documents (Senior A.W and Robinson A.J 1998). Among the others, text overlapping with underlines poses serious recognition problems, particularly when the documents must be filled manually by the writer according to the printed underlines. Furthermore lines can be of different width and length; they can be broken and connected to the handwritten text in many parts. Consequently it is possible that some parts of the text overlap with the underline and therefore that can be deleted during line elimination (Dimauro et al., 1997). The detection and removal of these factors through preprocessing techniques can be helpful to reduce variability and to improve recognition rates. Skew correction is a process which aims at detecting the deviation of the document orientation angle from the horizontal or vertical direction. Skew detection and correction are important preprocessing steps of document layout analysis and OCR approaches (Sarfraz 2007) to make writing style as uniform possible (Pastor, 2004). In this regard, (Watanabe 1997) conducted comparative experiments showing that normalization minimizes the error of recognition.

BACKGROUND

In most of the literature reviewed, preprocessing techniques are described as part of an overall system for handwriting recognition (Senior and Robinson 1998; Bozinovic and Srihari 1999; Madhvanath et al., 1999). Most of the skew estimation techniques can be divided into main classes according to the basic approach they adopt (Jonathan J. H 1998). Some of them are analysis of Projection Profiles (Akiyama and Hagita 1990; Postl 1986; Bloomberg and Kopec 1993; Bloomberg et al 1995; Ishitani 1993; Liolios et al 2002), Principal Component Analysis (Smith 2002; Steinherz et al., 1999; Sarfraz et al., 2005; Steinherz et al., 1999; Okun et al., 1999) and Connected Components Clustering (Ramesh 2006).

The traditional projection profile approach was proposed by (Postl 1986). In this approach, the input document is rotated through a range of angles and a projection profile is calculated at each angle. Features are then extracted from each projection profile to determine the skew angle. This is computationally expensive as it is performed directly on the original document images. Moreover, it is sensitive to the layout of the document image. Another projection profile approach was proposed by (Bloomberg and Kopec 1993) in which the original document image is down-sampled before the projection profile is computed. Therefore, the image data to be processed is reduced and the computational cost is reduced significantly. However, a major weakness is that its detection accuracy is influenced by the document image layout. It often fails on

Scanned with CamScanner

document images with multiple font styles, sizes or the ones that contain a large amount of non-text regions. Projection Profiles are strictly based on lower and upper baselines. If the image is noised then PP based approaches can not bring fruitful results for skew detection and correction (Hull1998 and Postl 1986). (Nicchiotti and Scagliola 1999) employed Generalized Projections (GP) extension of Projection Profile (PP) for skew detection and removal (Nicchiotti and Scagliola 1999). While Generalized Projection approaches is time consuming to implement.

The second class of skew correction is based on component analysis in which most significant eigenvector is calculated which leads to the skew angle of distribution. The problem with this method is that each eigenvector is constructed with support from projections of every point which is expensive in terms of time and that they are least squared estimation techniques and hence fail to account for outliers which are common in images. (Senior and Robinson 1998) described a skew detection technique whereby minima in the lower contour of the word image are first located and a line of best fit was drawn through these points. (DeLauro et al., 1997) used mathematical morphology for removing underlines in handwritten words. Although their system performs well, it does not seem to take into account the possibility of skewed handwriting. (Caesar et al., 1993) fits a straight line through extreme values in the vertical direction to detect reference lines for skew correction. (Blumenstein et al., 2002) introduced new preprocessing techniques for underline and skew removal. Unfortunately for underline removal separate algorithms are developed for different positions of the line present in the word based on stroke thickness. Furthermore skew detection is achieved through elimination of word ascender and descender information that may mislead the entire process.

In this paper, we described new and simple approaches for skew detection and removal. In addition to that, lines present in the word at any position, thickness with any slope are detected and removed without any distortion to actual image by a single algorithm unlike (Blumenstein et al., 2002). The developed techniques are tested on IAM benchmark database (Marti and Bunke 2002) as it is freely available on internet for research community. High accuracy for each technique is reported and discussed. Finally, new techniques come out with successful results even for the failure cases as reported by (Blumenstein et al., 2002) demonstrated in section 3.

The rest of the paper is organized into four sections. Section 2 describes the proposed preprocessing techniques in detail. Section 3 presents the results obtained and lastly conclusions along with future work are presented in Section 5.

PROPOSED PREPROCESSING TECHNIQUES

The proposed-algorithms are tested on word images taken from IAM benchmark database without performing any slant removal or skeleton. The black pixels in the binary images represent the handwriting (foreground pixels).

LINE DETECTION AND REMOVAL

Method of Analysis

The precise analysis is a prerequisite to correct line detection, its removal and junction detection to avoid damaging characters. A junction point is defined as a contact point of characters with line as shown in Figure 1. The connected component (black run) with one pixel up or down is analyzed by keeping record of length and junctions detection. If length is longer than a certain length threshold, we assume it is a line. In the same way, if junction is detected, it is treated as part of character and therefore not removed.

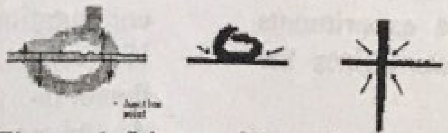

Figure 1. Line and junction points

OVERVIEW OF THE PROPOSED APPROACH

The proposed algorithm consists of line detection, its removal and checking of junction points. In implementation the line is neither a correct straight line nor of uniform thickness. Before removal of detected line, junctions are examined at each pixel.

Proposed algorithm

Let say an images denoted by P where

$P \in \{0,1\}$

$$p_{ij} \in P, \begin{cases} i = 1,2,\ldots,h \\ j = 1,2,\ldots,w \end{cases}$$

hw is height and width of P respectively.

1. Define origin as $O = \{o \in P \mid b = \ \}$.

   Take origin point $o_{ab} \in O$, a & b is left most of P.

2. Define

   $$L = \{l \in \{p_{i,j}\} \mid p_{i,j} = 1, p_{i,j+1} = 1, j = 1,2,..,a, a \leq w\}$$

   Start from o_{ab} trace line (L) to right direction allowing one pixel upward and downward continually.

3. calculate

   $$length(L) = \begin{cases} length(L) + 1, p_{i,j} = 1 \\ length(L), p_{i,j} = 0 \end{cases}$$

4. If $length(L) < \frac{1}{2} w$ then do step 1.

5. If $p_{ij} \in L$ and $\begin{cases} p_{i-1,j} \neq 1 \\ p_{i+4,j} \neq 1 \end{cases}$ then p_{ij} = 0. (changing foreground to background)

Skew detection and removal

The skew detection and removal algorithm is based on mathematical formulation. The proposed algorithm is described as under.

1. Locate and connect lower and upper boundary points as shown in figure 2.
2. Calculate slope of the line L.

$$m = \frac{y_2 - y_1}{x_2 - x_1}$$

1. Calculate angle of slope of line L (skew angle).

   $$\theta = \tan^{-1}(m)$$

2. Take $(x_1 y_1\ )$ as (x_{org}, y_{org})

3. Rotate image through an angle θ at origin (x_{org}, y_{org}).

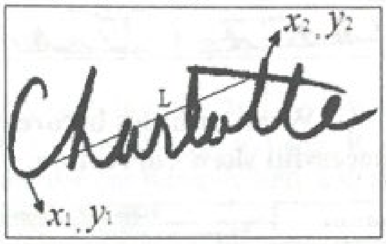

Figure2. Detected lower, upper boundary points of skew word connected by line L.

IMPLEMENTATION AND RESULTS

Implementation and Database

The line and skew detection and removal algorithms were implemented in MATLAB 7.0 and tested on windows XP platform. A number of experiments were conducted to test the proposed techniques by selecting desired images from IAM database (Marti, Bunke 2002). For testing, 1000 words were selected from the test set.

Additionally, for comparison of the proposed techniques with the other researcher (Blumenstein et al., 2002), same images are taken. The proposed approaches performed well even for the failure cases of other researchers. Results are exhibited in figure 3, 4 and 5.

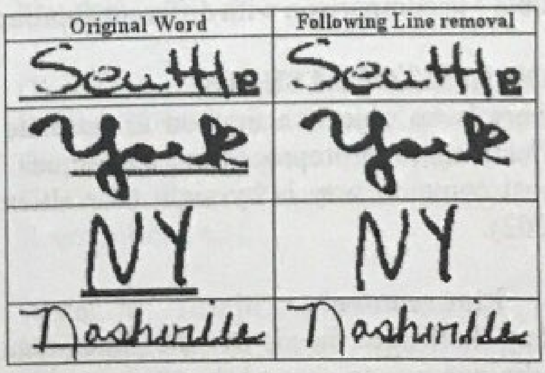

Figure 3. Words samples before and after successful line removal

Scanned with CamScanner

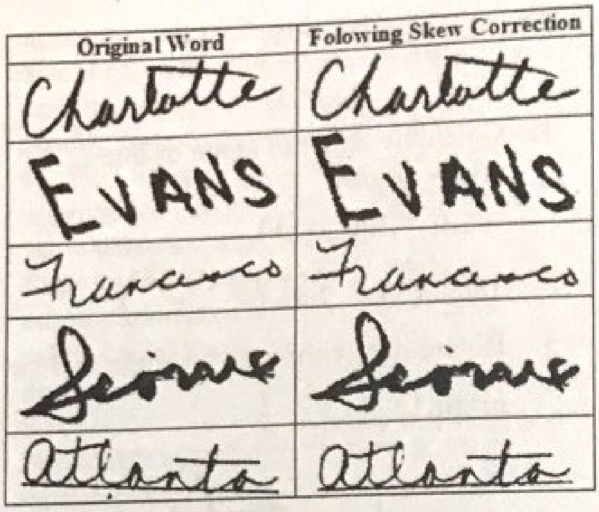

Figure 4. Word samples before and after successful skew correction.

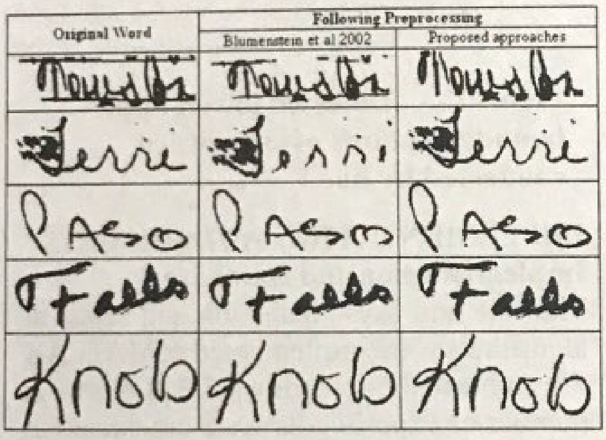

Figure 5. Comparison of results for line removal and skew correction

EXPERIMENTAL RESULTS AND COMPARISON

The performance of each technique was evaluated by visually inspecting the preprocessed word images. We count an error when the result of the preprocessing step is clearly different from what we expect. The percent of errors for each task is presented in Table 1 in comparison with different approaches

ANALYSIS OF RESULTS

There is no objective method to estimate the effectiveness of preprocessing techniques. The most common way is by sight (Kavallieratou 2002).

1. Line removal

The challenge faced by the line removal technique was to accurately retain strokes in a word that are not meant for removal. Therefore before removal of each pixel, its junctions were examined. If foreground pixels were found they were not converted to background.

2. Skew removal

Skew detection and removal was highly successful for almost all cases. The proposed technique performed well, even for the failure cases of (Blumenstein et al., 2002).

Table 1: Comparison of results (%Error) for each preprocessing task with different approaches

Preprocessing Technique	**Skew removal (% Error)**
Generalized Projection (Nicchiott, Scagliola 1999)	2.3
Projection Profile (Liolios et al., 2002)	0.91
Mathematical formulation (Blumenstein et al., 2002)	3.88
Principal component analysis (Sarfraz et al., 2005)	1.48
Connected component analysis (Okun 1999)	0.50
Peak and valley analysis (Sarfraz et al., 2007)	0.91
Proposed Approach	0.62

Preprocessing Technique	**Line removal (% Error)**
Connected components analysis (Yong et al., 1997)	3
Mathematical morphology (Dimauro et al., 1997)	3
Connected components analysis (Blumenstein et al., 2002)	2.84
Proposed Approach	0.75

CONCLUSION

We have presented novel and robust preprocessing techniques to accomplish line and skew removal for unconstrained off-line

Scanned with CamScanner

cursive handwritten words image. Our proposed algorithms relatively save computation time and improve accuracy. Line removal was successful in 99.25 % and word skew was acceptably corrected in 99.38 % of total cases.

In future these preprocessing techniques shall be integrated with entire recognition system to check improvement in recognition rate.

REFERENCES

Akiyama T, Hagita N (1990). Automatic entry system for printed documents. Pattern Recognition. 23(11): 1141 – 1154.

Bloomberg DS, Kopec GE (1993). Method and apparatus for identification and correction of document skew. Xerox Corporation, US Patent 5,187-753.

Bloomberg DS, Kopec GE and Dasari L (1995). Measuring document image skew and orientation. Proc. SPIE 2422, 302–316.

Blumenstein M, Cheng CK and Liu XY (2002). New preprocessing techniques for handwritten word recognition. In proceedings of 2nd IASTED International Conference on visualization, imaging and Image Processing., ACTA. 480-484.

Bozinovic RM, Srihari SN (1989). Off-line cursive script word recognition. Pattern Analysis and Machine Intelligence., 11(1):68-83.

Caesar T, Gloger JM and Mandler E (1993). Preprocessing and feature Extraction for a Handwriting Recognition System. Proceedings of International Conference on Document Analysis and Recognition, 408-411.

Dimauro G, Impedovo S, Pirlo G and Salzo A (1997).Removing underlines from handwritten text: An experimental investigation. Progress in Handwriting Recognition. 497-501.

Hull J. (1998). Document Image Skew Detection: Survey and Annotated Bibliography. Document Analysis Systems II. 40-64.

Ishitani Y (1993). Document skew detection based on local region complexity. In: Proc. 2nd International Conference on Document Analysis and Recognition, Tsukuba Science City, Japan, 49–52.

Jonathan J H (1998). Document Image Skew Detection: Survey and Annotated Bibliography, World Scientific., 40-64.

Kavallieratou E, Fakotakis N and Kokkinakis G (2000).A slant removal algorithm. Pattern Recognition., 33(7):1261-1262

Kim G, Govindaraju V and Srihari S N (1999). Advances in Handwriting Recognition Architecture for handwritten text recognition systems. 163-182

Liolios N, Fakotakis N and Kokkinakis G (2002). On the generalization of the form identification and skew detection problem. Pattern Recognition., 35: 253–264.

Marti U, Bunke H. (2002). The IAM database: An English sentence database for off-line handwriting recognition. International Journal of Document Analysis and Recognition., 15: 65-90.

Madhvanath S, Kleinberg E and Govindaraju V (1999). Holistic verification of handwritten phrases. Pattern Analysis and Machine Intelligence 21: 1344-1356.

Nicchiotti G, Scagliola C (1999). Generalised Projections: a tool for cursive handwriting normalization. Proceedings of 5th International Conference on Document Analysis and Recognition Bangalore India., 729-732.

Okun O, Pietikainen M and Sauvola J (1999). Robust Skew Estimation on Low-Resolution Document Images. 5th International Conference on Document Analysis and Recognition., 621.

Paster, M., Toselli, A., and Vidal, E. (2004). Projection Profile Based Algorithm for Slant Removal. Proceedings of the International Conference on Image analysis and Recognition, 183-190.

Scanned with CamScanner

Postl W (1986). Detection of oblique structures and skew scan in digitised documents. In Proceedings of 8th International Conference on Pattern Recognition., 687-689.

Ramesh DR, Piyush MK and Mahesh DD (2006). Skew Angle Estimation and Correction of Hand Written, Textual and Large areas of Non-Textual Document Images: A Novel Approach. IPCV 2006: 510-515.

Sarfraz M, Zidouri A and Shahab SA (2005). A Novel Approach for Skew Estimation of Document Images in OCR System. Proceedings of the Computer Graphics, Imaging and Vision: New Trends.175-180.

Sarfraz M. Mahmoud SA and Rasheed Z (2007). On Skew Estimation and Correction of Text Computer Graphics, Imaging and Visualization. 308 - 313

Senior AW, Robinson A J (1998). An off-line cursive handwriting recognition system, Pattern Analysis and Machine Intelligence. 20(3): 309-321.

Smith LI (2002) A tutorial on Principal Components Analysis. 26

Steinherz T, Intrator N and Rivlin E (1999). Skew detection via principal components analysis. In Proceedings of the 5th International Conference on Document Analysis and Recognition., 153–156.

Watanabe MM, Hamammoto Y, Yasuda T and Tomita S (1997). Normalization techniques of handwritten numerals for Gabor filters. Proceedings of the International Conference on Document Analysis and Recognition. 1: 303-307.

Yong JY, Kim MK, Yong SB and Kwon YB (1997). Line removal and restoration of handwritten characters on the form documents. -Proceedings of the Fourth International Conference on Document Analysis and Recognition., 1: 128-131

Scanned with CamScanner