Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 17 (2013) 88 – 95

Procedia Computer Science

Information Technology and Quantitative Management (ITQM2013)

A New Character Segmentation Approach for Off-Line Cursive Handwritten Words

Amit Choudhary[a,*], Rahul Rishi[b], Savita Ahlawat[c]

[a]Maharaja Surajmal Institute, New Delhi, India
[b]*UIET, Maharshi Dayanand University, Rohtak, India*
[c]*Maharaja Surajmal Institute of Technology, New Delhi, India*

Abstract

Character Segmentation is the most crucial step for any OCR (Optical Character Recognition) System. The selection of segmentation algorithm being used is the key factor in deciding the accuracy of OCR system. If there is a good segmentation of characters, the recognition accuracy will also be high. Segmentation of words into characters becomes very difficult due to the cursive and unconstrained nature of the handwritten script. This paper proposes a new vertical segmentation algorithm in which the segmentation points are located after thinning the word image to get the stroke width of a single pixel. The knowledge of shape and geometry of English characters is used in the segmentation process to detect ligatures. The proposed segmentation approach is tested on a local benchmark database and high segmentation accuracy is found to be achieved.

Selection and peer-review under responsibility of the organizers of the 2013 International Conference on Information Technology and Quantitative Management

Keywords: Character Segmentation; OCR; Word Recognition; Pre-processing Techniques.

1. Introduction

The importance of the piece of paper cannot be ignored towards enhancing the people's memory and in facilitating communication between people. People generally store important information by writing on paper for retrieving at a later stage. Paper is still a convenient and feasible way of storing data in the form of handwritten script or printed text. A large amount of historical data is also written on papers. In today's life also, a large amount of data is written on papers in Banking System, Postal Department and Insurance companies etc. Many researchers are attempting to simulate intelligent behavior and mimic the human brain's

* Corresponding author. Tel.: +91-991-133-5069.
E-mail address: amit.choudhary69@gmail.com.

Selection and peer-review under responsibility of the organizers of the 2013 International Conference on Information Technology and Quantitative Management
doi:10.1016/j.procs.2013.05.013

ability to read and recognize the handwritten or printed characters from the paper surface so that the computer can understand this script and process the data [1]. The automated processing of handwritten material optimizes the data processing speed as compared to manual processing and delivers very high recognition accuracy in least cost.

1.1. Historical Background

In the literature, for achieving high recognition accuracy, several segmentation techniques are proposed that can be broadly classified into three categories, namely Explicit Segmentation (Pure Segmentation), Implicit Segmentation (Recognition Based Segmentation) and Segmentation Free (Holistic) Approaches [2].

The segmentation approach proposed in this paper lies in the category of Explicit Segmentation or Pure Segmentation. When classical approach is adopted for recognition, segmentation becomes the most crucial step of the handwritten word recognition problem. In classical approach, input word image of sequence of characters is portioned into sub images of individual characters, which are then classified. The process of cutting up the word images into classifiable character sub images is termed as dissection. Many researchers in the literature adopted this dissection based segmentation techniques [3, 4, 5]. These techniques are used to find all the interconnections between character images (also called ligatures) and cut the word image through all the detected ligatures.

The key letters present in the cursive script provided the clue for character segmentation [6]. To extract key letters, face-up and face-down valleys along with open loop regions in cursive script are detected. Finally, fine segmentation points were determined by background analysis. Segmentation technique [7] on handwritten postal words was applied. A set of heuristic rules based on geometrical, topological and English character features was used. Segmentation accuracy of 85.7% reported on 50 real-world address images. The contour features for segmentation purpose in the complete word recognition system were investigated [8]. The top word recognition rate was 91.7% with lexicon size of 50 words. Segmentation of handwritten cursive words based on natural segmentation points and ligatures was hypothesized [9]. Accordingly, natural segmentation points were analyzed using histogram projection taken from five different angles and ligature candidates obtained from morphological operations of opening and closing. To search best structuring elements to determine ligature in the set of training words, genetic algorithm was employed.

1.2. Various Challenges during Segmentation

Off-line handwritten word segmentation is a subject of much attention due to the presence of many difficulties such as:

- There can be variation in shapes and writing styles of different writers.
- Cursive nature of handwriting i.e. two or more characters in a word can be written connected to each other.
- Characters can have more than one shape according to their position inside the word image.
- Words may be written by a pen having ink of different colors.
- Some characters (e.g. ‘u’ and ‘v’) in a handwritten word image can have similar contours.
- Some characters can give the illusion of presence of two similar characters e.g. ‘w’ can be segmented into two ‘v’ and ’v’ characters.

1.3. Motivation and Contribution

English language is used by a much higher percentage of the world's population. People will be benefited all over the globe if an automation system is designed for off-line handwriting recognition. The off-line handwriting recognition system enables the automatic reading and processing of a large amount of data printed or handwritten in English language. Although such automated systems for recognizing off-line handwriting already exist, the scope of further improvement is always there.

A new vertical segmentation technique is developed to enhance the over-segmentation of the handwritten word image by thinning the word image to a single pixel width. The objective of the proposed approach is to over-segment the handwritten word image sufficient number of times to ensure that all possible character boundaries have been dissected. Another technique is also developed to merge more than one successive potential segmentation points present between any two characters into a single segmentation point to enhance the segmentation performance.

2. Techniques and Methodologies

The selection of a segmentation technique depends on the nature of the script to be segmented. The proposed segmentation approach in this work is developed for segmenting touched handwritten characters from the words written in English language.

2.1. Handwritten Words Local Database

For conducting the segmentation experiment by the proposed segmentation technique, handwriting samples from 10 different people (age 15-50 years) has been gathered. Some of these samples were written on white paper and others on a colored or a noisy background. Exactly 200 words has been selected randomly from these handwriting samples containing all shapes of English characters written by those persons. Some of the word image samples from the collected database are shown in Fig 1.

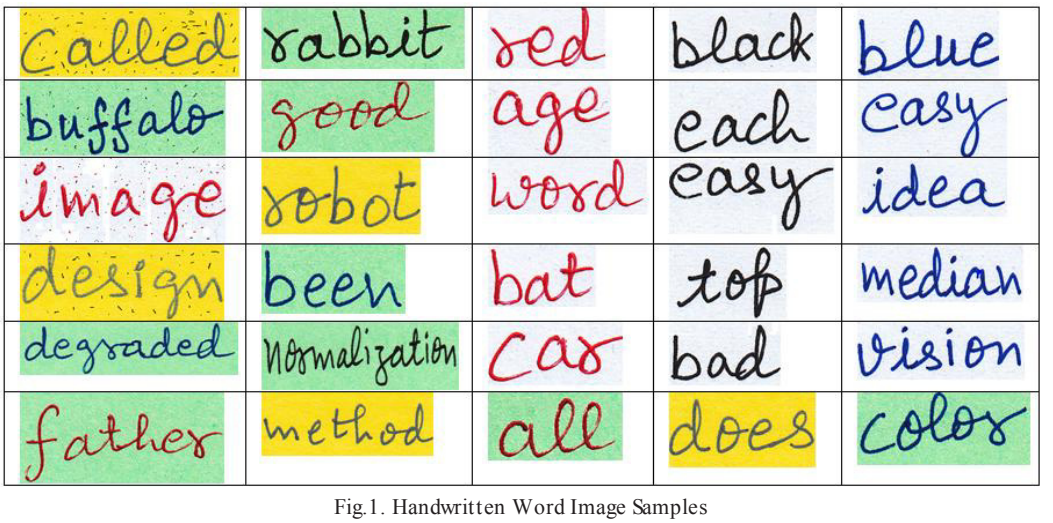

Fig.1. Handwritten Word Image Samples

2.2. Word Image Acquisition

In image acquisition, the word images have been acquired through a scanner or a digital camera. The input word images were saved in JPEG or BMP formats for further processing. Three such handwritten word images from the local database are shown in Fig 2(a).

2.3. Word Image Pre-processing

The aim of pre-processing is to eliminate the inconsistency that is inherent in cursive handwritten words. The handwriting samples may be written on a noisy or coloured background and also the quality of the word images may be degraded due to the noise that is introduced in the process of scanning or capturing the word images. It is necessary to remove the background noise to improve the quality of the word images to be used in the segmentation experiment. The outcome of the pre-processing techniques, which has been employed in an attempt to increase the performance of the segmentation process, is shown in Fig 2(b-f).

2.3.1. Thresholding

In this phase of preprocessing, the RGB images in BMP format as shown in Fig 2(a) has been converted to grayscale format as shown in Fig 2(b). This step is necessary so as to overcome the problems that may arise due to the use of pens of different colors and different intensities on various noisy and colored backgrounds. These grayscale images have been converted in binary matrix format. The resultant binary images has values of 0 each for all the foreground black pixels and 1 each for all the background white pixels as shown in Fig 2(c).

2.3.2. Thinning and Skeletonization

A large amount of variability may be present among the handwritten words because writers can use different type of pens of unequal stroke width. The thinning process delivers all the words used in the proposed experiment, a uniform stroke width of one-pixel. Such word images after thinning are shown in Fig 2(d).

2.3.3. Noise Removal

Noise (small dots or foreground components) may be introduced easily into an image while scanning the handwritten word image during image acquisition. The resultant images (without noise dots), while retaining the portions of the characters are shown in Fig 2(e).

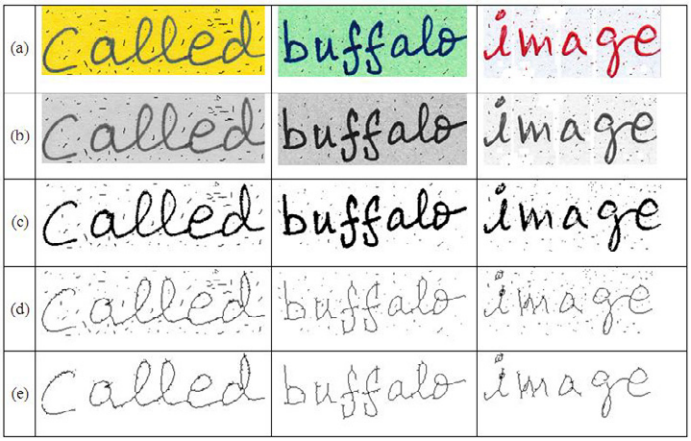

Fig.2. Word Image Pre-Processing (a) Input Scanned Word Images; (b) Word Images after Gray Scale Intensity Threshold; (c) Word Images in Binary Format; (d) Word Images after Thinning; (e) Word Images after Noise Removal.

3. Segmentation Technique

Many segmentation techniques have been developed by the researchers in the recent years. These techniques are basically script dependent and may not work well if applied for segmentation of words written in any other script. For example, the technique developed for segmenting touched characters in Roman script may not work well to segment touched characters of a word written in Arabic or Chinese script.

3.1. Overview

There are two types of characters in English language. First type of characters are called Closed Characters and contain a loop or a semi-loop such as 'a', 'b', 'c', 'd', 'e', 'g', 'o', 'p', 's' etc. Second type of characters are termed as Open Characters and are without a loop or a semi-loop e.g. 'u', 'v', 'w', 'm', 'n', 'i' etc. In case of Open Characters, it is very difficult to differentiate between ligatures and characters because of the cursive nature of handwriting. In case of cursive handwritten words, a ligature is a link (small foreground component) which is present between two successive characters to join them. Two consecutive 'i' characters may give an illusion of the presence of a character 'u' and vise versa. Two consecutive characters 'n' and 'i' may look like 'm'. Also, 'w' may look like the presence of two consecutive characters 'u' and 'i'. To overcome such type of challenges in the domain of cursive handwriting segmentation and recognition, a new segmentation approach is developed which is based on the analysis of the character's geometric features, such as, the shape of the character to identify the characters and the ligatures.

3.2. Methodology

After the preprocessing of the input handwritten word image, the height and width of the word image is calculated for the analysis of the ligatures in an accurate manner [2]. The word image is scanned vertically, from top to bottom, column wise and the number of foreground pixels in the inverted word image are counted in each column. The positions of all these columns are saved for which the sum of foreground black pixels is either 0 or 1. All these identified columns are termed as PSC (Potential Segmentation Columns) as shown in Fig. 3 (d).

3.3. Over-segmentation Problem

Many consecutive PSC are present in various groups in the whole word image where sum of foreground pixels are 0 or 1. This situation can be termed as over-segmentation as shown in Fig. 3(d).

The over-segmentation problem is occurred in three cases. First, when the two consecutive characters in the word image are not touching each other and the sum of foreground pixels of the columns in this area are 0. Second, when the two consecutive characters in the word image are connected by a ligature and the sum of foreground pixels in these columns crossing this ligature are 1. Third, when the characters are Open Characters like 'u', 'w', 'm', 'n' etc. without having any loop or semi-loop. A ligature is present within all of these Open characters. Due to the presence of this ligature-within-character, the characters of such type in the word image are still over-segmented and the sum of foreground pixels in these columns is also 1.

3.4. Solution of Over-Segmentation Problem

When there is a clear vertical space between two consecutive characters in a word image, the problem of over-segmentation is eliminated by taking average of all the PSC present in that area as the sum of foreground

pixels for all these PSC columns is 0.

When there is a ligature between two consecutive characters or there is a ligature-within-character (Open Characters such as 'w', 'm' etc.), the over-segmentation is eliminated to a great extent by taking average of those PSC which are at a distance less than a particular value (threshold) and by merging them into a single SC (Segmentation Column). The threshold value is the minimum distance along the width of the word image between consecutive PSCs and is so chosen that its value must be less than the width of the thinnest character possible (e.g. 'i', 'l') in a word image. By experimenting several times, the value of threshold is set to a value 7. This means that all those PSCs which are separated by a distance of 7 pixels or less by another PSC will be merged to a single SC.

4. Implementation

In the first step, preprocessed word images after thinning are taken as input images to be segmented into characters as shown in Fig 3(a). To minimize the computation complexity, the input word images are inverted to make them suitable for further processing. By complementing the input binary images, white pixels become the foreground pixels and the black pixels become the background pixels as shown in Fig 3(b). Hence, it becomes easier to count the foreground white pixels (represented by 1) in each column of the word images. These images are now converted from binary format to a RGB format as shown in Fig 3(c). Now, it becomes computationally easier to display the PSC (Potential Segmentation Columns) in different color other than black and white. All PSCs over-segmenting the word image are printed in red color as shown in Fig 3(d). It is clear from Fig 3(d) that each column in the word images, for which the sum of foreground white pixels is 0 or 1, is a PSC and vertically cuts the word image. All PSCs, which are at a distance less than a threshold value (7 pixels) from each other, are merged into a single column termed as Segmentation Column (SC) as shown in Fig 3(e). The final segmented word images are obtained by changing the black background of the image with a white background as shown in Fig 3(f).

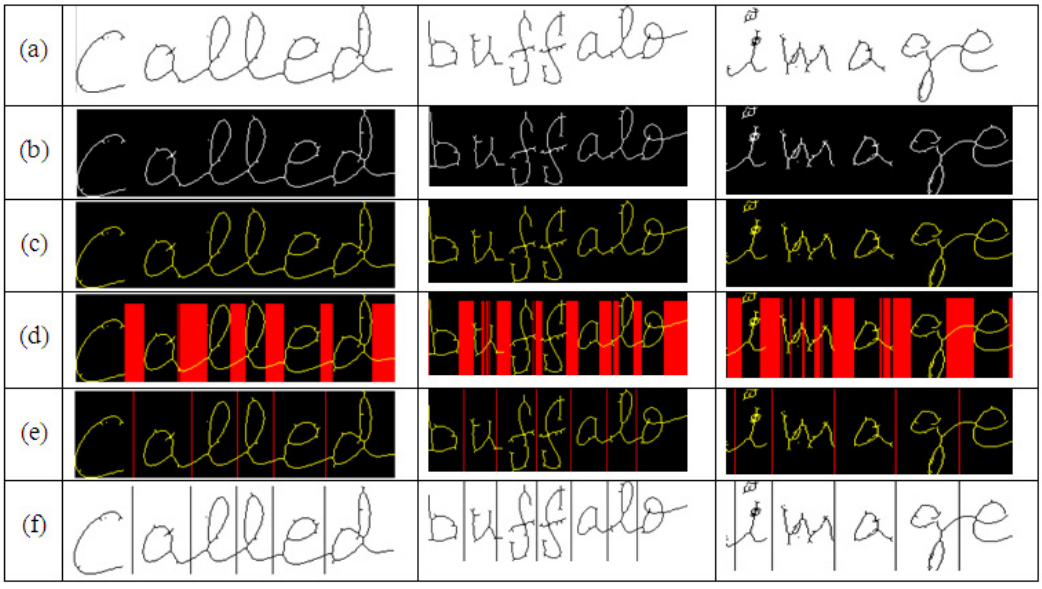

Fig.3. Word Image Segmentation (a) Pre-processed Word Images; (b) Inverted Binary Images; (c) RGB Images; (d) Over-segmentation in Images; (e) Image after removing Over-segmentations; (f) Final Segmented Output Word Images

5. Result Analysis

For evaluation of the proposed segmentation approach, 200 handwritten word samples were selected from the handwriting samples of 10 different writers. The performance of the proposed segmentation approach was judged on the basis of segmentation errors of the three types, namely, Number of Over Segmentations, Number of Missed Segmentations and Number of Bad Segmentations and is shown in Table 1.

Table 1. Segmentation Result of the Proposed Approach

Total Number of Handwritten Words	Total Correctly Segmented Words (%)	Total Incorrectly Segmented Words (%)	Number of Words with various Segmentation Errors		
			Over-Segmented	Miss-Segmented	Bad-Segmented
200	167 (83.5%)	33 (16.5%)	14	5	24

Out of 33 incorrectly segmented words, some words were over-segmented in one place as well as bad-segmented in some other place. While putting the results in Table 1, such types of words were counted in both of the error categories i.e. counted in Over-Segmented as well as Bad-Segmented. Similarly, in some words, the correct segmentation point was missed and shifted to some other place resulting in bad-segmentation. Some word images which are over-segmented or miss-segmented or bad-segmented are shown in Fig 4.

Fig.4. Word Images Showing all type of Segmentation Errors

It is very difficult to compare the segmentation results achieved by the proposed approach with the segmentation results of some other researchers because different researchers used different databases of handwritten words and reported the segmentation results under various constraints such as some researchers assumed the absence of noise, some researchers collected the handwriting samples from different number of writers and so on. Although some researchers [10, 11] used various benchmark databases e.g. CEDAR or IAM for their experiment but they used different number of words from the benchmark database. As the character segmentation in word images is done before the character recognition phase, most of the researchers mentioned only the recognition results and not the segmentation results.

6. Conclusions and Future Scope

The proposed segmentation approach guaranteed correct segmentation when characters in a word image were not touching each other. This approach also delivered excellent results in case of segmentation of ligatures present between consecutive Closed Characters. It also minimized the problem of over-segmentation that appeared during segmentation of Open Characters. The presence of Over-Segmentation in case of Open Characters was due to the presence of ligatures-within-characters. In case of Open Characters, ligature-within-characters were sometimes appeared as they are ligature between two consecutive characters and were over-

segmented by the proposed technique. Although the segmentation accuracy of 83.5% obtained was very favourably, yet there is always a scope of improvement.

In future, better pre-processing techniques will be used specially the thinning technique. The skew and slat correction module will be incorporated in the future work. This module was not applied here in this work because it was assumed that all the handwritten word samples are free from slant and skew. Some intelligent technique to validate the correct segmentation points (e.g. a neural-based validation) will be applied to further improve the segmentation accuracy in future work.

Acknowledgements

The authors acknowledge their sincere thanks to the management and the director of Maharaja Surajmal Institute, C-4, Janakpuri, New Delhi, India, for providing infrastructure and financial assistance to carry out this research. The excellent cooperation of the fellow colleagues and library staff is highly appreciated. The authors are grateful to the anonymous reviewers for their comments and suggestions.

References

[1] Peng Y., Kou G., Shi Y. and Chen ZA. Descriptive Framework for the Field of Data Mining and Knowledge Discovery , International Journal of Information Technology & Decision Making, 2008, Vol. 7, Issue: 4, Page 639 – 682.
[2] Rehman A, Mohamad D, Sulong G. Implicit Vs Explicit based Script Segmentation and Recognition : A Performance Comparison on Benchmark Database, Int. J. Open Problems Compt. Math., 2009, Vol. 2, No. 3, 352-364.
[3] Rehman A, Saba T . Performance analysis of character segmentation approach for cursive script recognition on benchmark database, Digital Signal Processing, 2011, (21) 486-490.
[4] Saba T , Rehman A, Sulong G. Cursive Script Segmentation with Neural Confidence, Int. J. Innovative Computing, Information and Control, 2011, Vol 7, No. 8, 4955-4964.
[5] Al Hamad HA, Zitar RA. Development of an efficient neural-based technique for Arabic handwriting recognition, Pattern Recognition, 2010, (43) 2773-2798.
[6] Cheriet M. Reading cursive script by parts. In: Proceedings of the 3rd international workshop on frontiers in handwriting recognition, Buffalo, 1993, 25–27 May, pp 403–408.
[7] Han K, Sethi IK. Off-line cursive handwriting segmentation. In: Proceedings of the 3rd international conference on documents analysis and recognition, 1995, pp 894–897.
[8] Yamada H, Nakano Y. Cursive handwritten word recognition using multiple segmentation determined by contour analysis. IEICE Trans Inf Syst, 1996, E79-D:464–470.
[9] Veloso LR, Sousa RP, De, Carvalho JM. Morphological cursive word segmentation. In: Symposium on computer graphics and image processing, 2000, XIII, Brazilian, pp 337–342.
[10] Marti U, Bunke H. The IAM database: An English sentence database for off-line handwriting recognition, Int. J. Doc. Anal. Recognit, 2002, (15): 65–90.
[11] Hull JJ. A database for handwritten text recognition, IEEE Trans. Pattern Anal. Mach. Intell.; 1994, (16):550–554.