International Journal of Information & Computation Technology.
ISSN 0974-2239 Volume 4, Number 6 (2014), pp. 559-564
© International Research Publications House
http://www. irphouse.com

A Review of Various Character Segmentation Techniques for Cursive Handwritten Words Recognition

Amit Choudhary

Maharaja Surajmal Institute, C-4, Janakpuri, New Delhi, INDIA

ABSTRACT

Cursive handwriting recognition is a challenging task for many real world applications such as document authentication, form processing, postal address recognition, reading machines for the blind, bank cheque recognition and interpretation of historical documents. Therefore, in the last few decades the researchers have put enormous effort to develop various techniques for handwriting segmentation and recognition. This review presents the segmentation strategies for automated recognition of off-line unconstrained cursive handwriting from static surfaces. This paper reviews many basic and advanced techniques and also compares the research results of various researchers in the domain of handwritten words segmentation.

Keywords - OCR, Character Segmentation, Character Recognition, Segmentation Techniques

1. INTRODUCTION

The research in the area of Handwriting Recognition has been ongoing for over half a century and the outcomes have been astounding with successful recognition rates for printed characters exceeding 99%, with significant improvements in performance for handwritten cursive character recognition where recognition rates have exceeded the 90% mark (Alginahi, 2010). In the literature (Verma and Blumenstein, 2008), some researchers have obtained very promising results for isolated/segmented numerals and characters using conventional and intelligent techniques. However, the results obtained for the segmentation and recognition of cursive handwritten words have not been satisfactory in comparison (Blumenstein and Verma, 2001; Blumenstein et al., 2003; Vinciarelli et al., 2003; Günter and Bunke, 2005).

The reason for not achieving satisfactory recognition rates is the difficult nature of cursive handwriting and difficulties in the accurate segmentation and recognition of cursive and touching characters (Verma and Blumenstein, 2008). This review reports on the state-of-the-art in handwriting recognition research and methods for segmentation of cursive handwritten words into individual character.

2. CHARACTER SEGMENTATION

Character segmentation is an operation that seeks to decompose an image of a sequence of characters into sub-images of individual symbols (Rehman and Saba, 2012). Several review papers highlighted different issues in cursive script segmentation and acknowledged the segmentation stage as the most difficult step in the process of cursive handwriting recognition (Rehman and Dzulkifli, 2008; Saba et al., 2011).

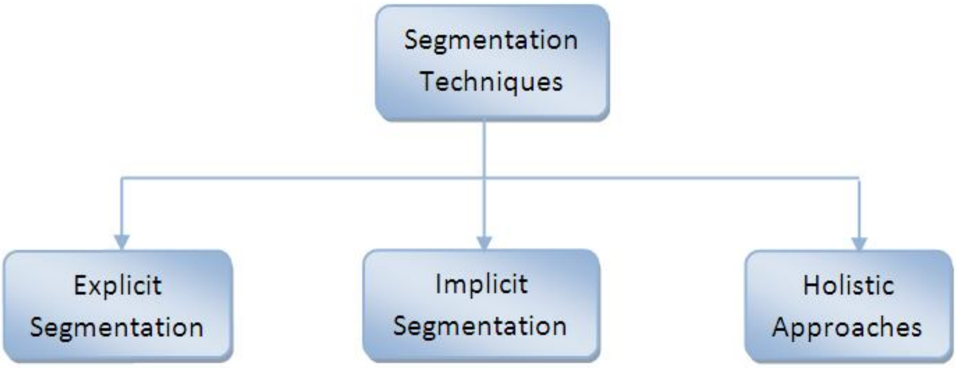

Figure 1. *Categorization of Segmentation Based and Segmentation Free Approaches*

3. SEGMENTATION TECHNIQUES

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 Holistic (Segmentation Free) Approaches as shown in Figure.1.

3.1 Explicit Segmentation

When explicit segmentation (pure segmentation) is adopted for recognition; segmentation becomes the most crucial step of the handwritten word recognition problem. In this classical approach, input word image of sequence of characters is partioned 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 (Saba et al., 2011). 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.

According to (Rehman and Saba, 2012), most of the researchers perform dissection via pre-segmentation. It is used to locate areas in the word containing explicit features that are likely to occur within or between characters in the form of valley such as ligatures. However, it also cuts the characters ‘w’, ‘v’ etc, whose contours contain a valley and therefore, deduce as a ligature.

Some systems investigate ligatures close to the baseline, but such efforts cannot brought fruitful results due to inherited nature of certain characters such as ‘u’, ‘w’, ‘g’ etc that do not contain ligatures close to the baseline. Holt et al. (1992) detect ligatures by locating minima in the upper contour of words, location of holes, contour direction and core region position. Segmentation points are marked if a minima in the upper contour is located, except if the contour component in question formed part of a hole. Similarly, Kimura et al. (1993) propose segmentation–recognition system for handwritten postal words; for segmentation part, they analyze upper contour. According to their investigation, prospective segmentation points are laid in those local minima that are deep enough and are adjacent to local maxima. Finally, segmentation points shift horizontally to the right or left to obtain valid segmented characters.

Ghosh et al. (2004) propose direct segmentation approach in their fully automated off-line handwriting recognition system. The segmentation phase employs many heuristic based set of rules in an iterative procedure and finally a neural network validation system is implemented. Accurate segmentation rate is 83.6%. However, over-segmentation and bad segmentation is considerably high up to 10.8 and 5.4% respectively, whereas, missed segmentation rate is 0.2%.

Samrajya et al. (2006) investigate hypergraph model to segment a cursive handwritten word image into isolated characters. Hypergraph model treats an image as packets of pixels. Authors claim that by recombining these packets of different sizes a given word image can be segmented into characters if at least one of the combinations provided a correct segmentation. However, neither segmentation results are presented for comparison nor the technique seems to yield successful results for horizontal overlapped and touching characters.

Dawoud (2007) introduce iterative cross section sequence graph (ICSSG) for the character segmentation. ICSSG tracks the characters growth at equally spaced thresholds. The iterative thresholding reduces the effect of information loss associated with image binarization. However, the experiments are performed on handwritten digits only.

Lee and Verma (2008) propose a new segmentation algorithm for off-line cursive handwriting recognition. Initially, word images are dissected heuristically based on pixel density between upper and lower baselines. Each segment passed through multiple expert based validation processes to determine valid character boundaries. An average segmentation error up to 5.25% for miss-segmentation, over-segmentation and bad segmentation is reported on 218 test words of CEDAR.

3.2 Implicit Segmentation

Implicit segmentation (recognition based segmentation) based recognition, in which the system searches the image for components that match classes in its alphabet. However, implicit segmentation-based methods are employed as an alternative to integrate segmentation and recognition processes. Accordingly, Hidden Markov

Models (HMM) based approaches are emerged. Actually, this approach is developed for speech recognition where it brought fruitful results (Rabiner, 1989). Therefore, its success diverts researcher's attention to apply HMM in word recognition. The main interest of this category of methods is that they bypass the segmentation problem: No complex "dissection" algorithm has to be built and recognition errors are basically due to failures in classification. The approach has also been called "segmentation-free" recognition.

Cavalin et al. (2006) propose two-stage HMM based method for recognition of strings of characters (words or numerals). In first stage, an implicit segmentation scheme is applied to segment either words or numeral strings and verification performs in the second stage. Accordingly, foreground and background features are combined to compensate the loss in terms of recognition rate during implicit segmentation in previous stage. Word recognition accuracy up to 88.2% is reported on lexicon of size 3,771.

3.3 Holistic Approaches

A holistic (Segmentation Free) process recognizes an entire word as a unit. A major drawback of this class of methods is that their use is usually restricted to a predefined lexicon. Since they do not deal directly with letters but only with words, recognition is necessarily constrained to a specific lexicon of words. This point is especially critical when training on word samples is required. A training stage is thus mandatory to expand or modify the lexicon of possible words. This property makes this kind of method more suitable for applications where the lexicon is statically defined (and not likely to change), like bank cheque recognition. They can also be used for on-line recognition on a personal computer (or notepad), the recognition algorithm being then tuned to the writing of a specific user as well as to the particular vocabulary concerned (Casey and Lecolinet, 1996).

3.4 Hybrid Approaches

The literature is replete with hybrid approaches proposed by a number of researchers to optimize algorithms with linear searching techniques, contextual and lexicon knowledge. Recently, Rehman and Dzulkifli (2008) proposed a new fast segmentation approach for off-line cursive handwritten words with accuracy up to 91.21% on a subset of IAM database. Authors proposed certain rules to analyze ligatures along with knowledge of character shape. The detailed analysis (Blumenstein and Verma, 2001; Rehman and Dzulkifli, 2008) has shown that most existing segmentation algorithms have three major problems: (1) inaccurately cutting characters into parts; (2) missing many segmentation points; and (3) over-segmenting a character many times, which contributes to errors in the word recognition process. Most researchers have evaluated their segmentation accuracy as an overall word recognition performance. Additionally, database and experimental setup is different among the researchers.

5. CONCLUSIONS

In this review paper, a state of the art in off-line cursive script segmentation is presented with the great emphasis on segmentation-based off-line cursive script recognition technique. A critical literature review of existing techniques and comparative study of recent achievements in the area is presented. Novel strategies by the researchers to tackle existing problems in segmentation-based script recognition have also been presented. By the detailed analysis of the literature, it is observed that the research is almost matured in area of numeral recognition however the same accuracy level is not met with alphabets. The problem of cursive character recognition remains very much an open problem. It is mainly due to presence of noisy, broken, multi-stroke, incomplete and ambiguous characters in words. To handle this type of problem new feature extraction/selection techniques and multistage classifiers are desired.

7. REFERENCES

[1] Alginahi, Y. (2010). “Preprocessing Techniques in Character Recognition”, Character Recognition, Minoru Mori (Ed.), ISBN: 978-953-307-105-3, InTechopen Publishers, pp. 1-20.

[2] Blumenstein, M. & Verma, B. (2001). “Analysis of Segmentation Performance on the CEDAR Benchmark Database”, in proceedings of the 6th International Conference on Document Analysis and Recognition, Seattle, USA: IEEE Computer Society Press, pp. 1142-1146.

[3] Blumenstein, M., Verma, B. & Basli, H. (2003). “A Novel Feature Extraction Technique for the Recognition of Segmented Handwritten Characters”, In Proceedings of the 7th International Conference on Document Analysis and Recognition, Edinburgh, UK: IEEE Computer Society Press, pp. 137-141.

[4] Casey, R. G., & Lecolinet, E. (1996). “A survey of Methods and Strategies in Character Segmentation”, IEEE Trans Pattern Anal Mach Intell, 18 (7), pp. 690–706.

[5] Cavalin, P. R., Britto, A. S., Bortolozzi, F., Sabourin, R. & Oliveira, L. S. (2006). “An implicit segmentation based method for recognition of handwritten strings of characters”, in proceedings of ACM symposium on applied computing, pp. 836–840.

[6] Dawoud, A. (2007). “Iterative cross section sequence graph for handwritten character segmentation”, IEEE Trans Image Process, 16(8), pp. 2150–2154.

[7] Ghosh, M., Ghosh, R. & Verma, B. (2004). “A fully automated off-line handwriting recognition system incorporating rule based neural network validated segmentation and hybrid neural network classifier”, Int J Pattern Recognit Artif Intell, 18(7), pp.1267–1283.

[8] Günter, S., Bunke, H. (2005). “Off-line cursive handwriting recognition using multiple classifier systems”. On the influence of vocabulary, ensemble, and training set size, Optics Lasers Eng, 43(3–5), pp. 437–454.

[9] Holt, M., Beglou, M. & Datta, S. (1992). “Slant-independent letter segmentation for off-line cursive script recognition”, in Impedovo S, Simon JC (eds) From pixels to features III, Elsevier, Amsterdam, pp. 41-42.

[10] Kimura, F., Shridhar, M. & Chen, Z. (1993). “Improvements of a Lexicon directed algorithm for recognition of unconstrained handwritten words”, in proceedings of the 2nd international conference on document analysis and recognition, Tsukuba, Japan: IEE Computer Society Press, pp. 18–22.

[11] Lee, H., Verma, B. (2008). “A novel multiple experts and fusion based segmentation algorithm for cursive handwriting recognition”, in proceedings of the international joint conference on neural networks (IJCNN’08), pp. 2994–2999.

[12] Rabiner, L. (1989). “A tutorial on hidden Markov models and selected applications in speech recognition”, Proc IEEE, 77(2), pp. 257–286.

[13] Rehman, A., Dzulkifli, M. (2008). “A simple segmentation approach for unconstrained cursive handwritten words in conjunction with the neural network”, Int J Image Process, 2(3), pp. 29–35.

[14] Rehman, A., Saba, T. (2012). “Off-Line cursive script recognition: current advances, comparisons and remaining problems”, Artif Intell Rev, 37, pp. 261-288.

[15] Saba, T., Sulong, G. & Rehman, A. (2011). “Document image analysis: issues, comparison of methods and remaining problems”, Artificial Intelligence Review, 35, pp. 101-118.

[16] Samrajya, P., Lakshmi, M. & Swaroop. A. (2006). “Segmentation of cursive handwritten words using Hypergraph”, TENCON, IEEE region 10 Conference, pp. 1-4.

[17] Verma, B., Blumenstein, M. (2008). “Pattern Recognition Technologies and Applications: Recent Advances”, Information Science Reference (An Imprint of IGI Global Publications), Hershey, New York, pp. 1-16.

[18] Vinciarelli, A., Bengio, S. & Bunke, H. (2003). “Off-line Recognition of Large Vocabulary Cursive Handwritten Text”, Proceedings of the 7th International Conference on Document Analysis and Recognition, Edinburgh, UK: IEEE Computer Society Press, pp. 1101-1107.