STUDY OF VARIOUS CHARACTER SEGMENTATION TECHNIQUES FOR HANDWRITTEN OFF-LINE CURSIVE WORDS: A REVIEW

[1]AMANDEEP KAUR, [2]SEEMA BAGHLA, [3]SUNIL KUMAR

[1]M.Tech. Student (Computer Engg.), [2]Assistant Professor (Computer Engg.), [3]Assistant Professor
Yadavindra College of Engineering, Punjabi Univ. Guru Kashi Campus, Talwandi Sabo, Bathinda, Punjab, India
E-mail: agill.amandeep333@gmail.com, bgarg_seema238@yahoo.co.in

Abstract- Segmentation of cursive handwriting is the challenging step of Optical Character Recognition (OCR).The recognition accuracy will highly depend on the good segmentation. Segmentation of cursive handwriting is very difficult. The segmentation can be done on the basis of zoning, a line segment of text, a word segment from line and character segment from word. This can be done by the use of horizontal, vertical method. This paper reviews many basic and advanced techniques of handwritten word segmentation.

Index Terms- Character segmentation, character segmentation techniques, optical character recognition.

I. INTRODUCTION

Optical character recognition is a program that translates a scanned image of a document into a text document that can be edited. Segmentation of cursive handwriting is very difficult [2]. Character segmentation is an operation to decompose an image into the sub-image of individual symbols [5].

There are mainly three phases of a character recognition system, namely preprocessing, segmentation and recognition. Preprocessing aims to produce data that are easy for the OCR system to work accurately. It reduces noise and distortion, removes skewness and performs skeletonising of the image, thereby simplifying the processing the rest of the stages [4]. The next stage is segmenting the document into its sub-components. It separates the different logical parts, like text from graphics, line of a paragraph, and character of a word [7].

II. PHASES OF OCR

Phases of OCR can be listed in the form of following flowchart shown in fig. 1.

A.Image acquisition

This involves scanning a document and storing it as an image. Their solution (number of dots per inch, dpi) determines the rate of process [4].

B.Pre-processing

Process of representing the scanned image of further processing. Preprocessing aims to produce data that are easy for the OCR system to operate accurately.

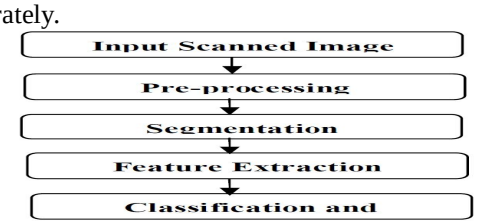

Fig.1: Steps in OCR

It reduces noise and distortion, removes skewness and performs skeletonising of the image, thereby simplifying the processing the rest of the stages [4].

C.Segmentation

After the preprocessing stage, a 'clean' document is obtained. The next stage is segmentation. In this stage, segmenting the document into its sub-components. It separates the different logical parts, like text from graphics, line of a paragraph, and characters of a word [7]. Segmentation is an important phase of OCR, because it can reach in separation of words, lines or characters directly affect the recognition rate of the script [4]. In fact correct recognition based on correct segmentation [5].

D.Feature Extraction

A set of rules stored on OCR engine comparing against character's shape and its features that distinguishes each character identify a character. The main part of the recognition system design is the selection of a stable representative set of features. It is the most consequential issue in the designing issues involved in building an OCR system.

E. Classification

The main decision making stage of an OCR system is classification. Classification uses the features extracted in the feature extraction stage to identify the text segment.

F. Post-processing

It is the final stage, post-processing refining the decisions taken by the previous stage, improves the recognition and recognizes words using context. It is ultimately responsible for outputting the best solution and is often implemented as a set of techniques that rely on character frequencies, lexicons, and other context information.

III. SEGMENTATION

The preprocessing stage yields a clean document. The

sufficient amount of shape information, high compression and low noise on normalized image is obtained. The next stage after preprocessing is segmentation. Segmentation is the process of segmenting the whole document into sub components [14]. Segmentation is of two types, external and internal segmentation. While external segmentation is the isolation of the sentences, paragraphs, and other such writing units. Internal segmentation is the isolation of the characters and letters [1].

G. Segmentation processes

Segmentation processes, including following processes:

- Line segmentation

Line segmentation is the process in which from the image, we extract only lines or differentiate the lines. Horizontal projection of a document image is most commonly used to extract the lines from the document. The horizontal projection will have separated peaks and valleys for the lines that are well separated and are not tiled, which serve as the separators of the text lines. These valleys are easily detected and used to determine the location of boundaries between the lines. Word segmentation is the process in which from the line segmentation, we extract only words. As we know that there is a distance between one word another word, this concept is used for word segmentation.

- Word segmentation

Word segmentation is a process of dividing a string into its component words. Word splitting is the process of parsing concatenated text to infer where word breaks exist. By using vertical projection profile, one can get column sums. By looking for minima in horizontal projection profile of the page, we can separate the lines and then separate words by looking at minima in the vertical projection profile of a single line. By using the valleys in the vertical projection of a line image, one can extract words from a line and also extracting individual characters from the word [2].

- Character segmentation

In character segmentation, we extract only characters from word. Character segmentation is a difficult step of OCR systems as it extracts meaningful regions for analysis. This step decomposes the images into classifiable units called character [2]. A poor segmentation process leads to incorrect recognition or rejection segmentation process carried after out only after the preprocessing of the image. According to Casey and Lecolinet.

IV. CHARACTER SEGMENTATION

Character segmentation is the most crucial step for any OCR system because the characters are the smallest unit of any language script. After segmentation of the character features can be ext, cannot be recognized accurately by feature extraction algorithm [11]. Segmentation of character is quite easy in case of printed documents as compared to the handwritten documents. Vertical projection is used for character segmentation.

A.Character segmentation techniques

Several segmentation techniques can be broadly classified into following three categories:

- Explicit segmentation

In the explicit segmentation, the input word image of a sequence of characters is portioned into sub images of individual characters, which are then classified. This process is termed as a dissection [2].

Vertical segmentation approach lies in the category of explicit segmentation. In this approach, after the preprocessing of the input handwritten word image, the word image is scanned from top to bottom [4]. The positions of all these columns are saved for which the sum of foreground black pixels is either 0 or 1. These columns are known as PSC (Potential segmentation columns) [4]. In the word images, each column for which the sum of foreground pixels is 0 or 1 is a PSC. By using PSC, vertically cuts the word image as shown in fig. 3. All the PSC, for which distance is less than a threshold value, are integrating into a single column.

Pre-processing algorithm & character segmentation algorithm:

This algorithm segments the characters from the words to form broken character which can be segmented into characters. This algorithm consists of these following steps:

Step 1. Collection of various samples of cursive handwritten script from standard database IAM handwriting database version 3 is used as input for the proposed system.

Step 2. Conversions of collecting documents of PNG format and then convert them into binary format.

Step 3. Conversion of binary images into 0's and 1's so that edge detection is applied to it.

Step 4. After finding edges, the histogram profile projection technique is worked out.

Step 5. After applying histogram profile projection technique, we gets out cut points of a character at a distance of 20 pixels.

Fig. 2: word image samples

Fig. 3: word image samples after segmentation

- Implicit segmentation

Implicit segmentation is also called recognition based segmentation. In this approach segmentation and recognition of characters are achieved at the same time [5]. In this, the system searches the image for the components that image classes in its alphabet [2]. Implicit segmentation approach is to split words into segments that should be characters, and then pass each segment to a classifier. If the classification results are not satisfactory, call segmentation once more with the feedback information about rejecting the previous result [8].

The implicit segmentation approach provides all the tentative segments and let recognizer to decide best segmentation hypothesis. However, there is a tradeoff in selecting the number of segments for a word. Less number of segments is the base of efficient computation but wide character cannot be covered in the hypothesis. Whereas, large number of segments, it is computationally expensive [5].

To overcome problems, the implicit segmentation approach is a two stage process. In the first stage, heuristic rule based segmentation scheme is applied to segment the words. In the second stage, verification is performed [2]. Fine segmentation was finalized based on recognition. In this approach, other segmented characters were segmented into maximum two halves [5].

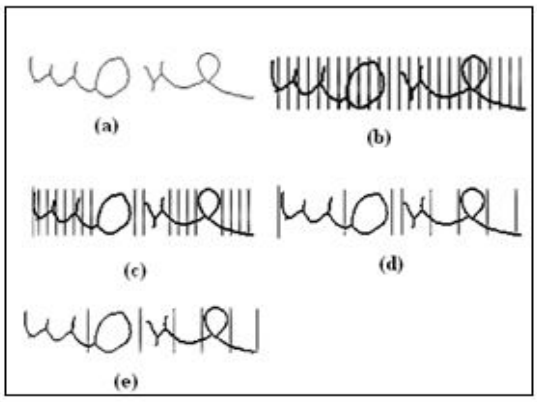

Fig. 4: a sequence of processing results of stage1 [5]

- Holistic approaches

Holistic segmentation approach is also known as a segmentation free approach. By using holistic approach, one can extract the entire word as a unit from a string. This approach directly concern with words, not letters. Use of the holistic approach is limited to a predefined lexicon. An application for which the lexicon is statically defined, holistic approach is used, like bank cheque recognition [2].

White space and pitch approach lies in the category of holistic segmentation technique. Vertical white space serves to separate successive characters. In billing application, design of a document specifically designed for OCR, additional spacing is bulit into the fonts used. The basis provided for estimating segmentation points by pitch, or number of characters per unit horizontal distance. The sequence of segmentation points obtained for a given line of print should be approximately equally spaced at the distance corresponding to the pitch. White space and pitch was heavily dependent on the quality of the input images.

V. SOME STUDIES ON CHARACTER SEGMENTATION TECHNIQUES

Several segmentation techniques are proposed for achieving high recognition accuracy. Recognition accuracy depends on the correct segmentation.

Rehman et al. (2009) represented the comparison between implicit and explicit based segmentation techniques for off-line cursive handwriting recognition. Implicit segmentation technique achieved segmentation and recognition at the same time. Implicit segmentation based recognition removed the class overlapping problem. The explicit segmentation based approach was computationally complex than the implicit segmentation approach, but the explicit segmentation technique achieved better results than implicit segmentation technique.

Lee et al. (2009) described the novel binary segmentation with neural validation. The segmentation and validation approach contains over-segmentation based suspicious segmentation point generator, binary segmentation and neural validation modules. The chain-failure successfully reduced by binary segmentation was investigated.

Brodowska (2012) discussed about the most common method of character recognition was initially over segmentation. Holistic approach, classical approach, recognition-based segmentation approaches and mixed approach were discussed. Selection of a particular technique should be dependent on the task, primarily on the kind alphabet the character strings were built over.

Nath and Rastogi (2012) discussed the various stages of optical character recognition (OCR). Explicit and implicit segmentation approaches were discussed.

Classifier with different features was also discussed.

Choudhary et al. (2013) discussed the character segmentation approach that was based on selecting the objects or regions that meet the criterion of having an area greater than some threshold value. For the segmentation of untouched characters, this character extraction technique was applied. These words of varying length were written on a noisy background.

Choudhary et al. (2013) proposed a new vertical segmentation technique in which segmentation points were located after thinning the word image to get the stroke width of a single pixel. This technique enhanced the over-segmentation of handwritten word image. Proposed approach over-segment the handwritten word image sufficient number of times to ensure that all possible character boundaries had been dissected. The proposed technique minimized the problem of over segmentation that appeared during segmentation of open characters.

Phukan and Borah (2014) discussed about what was the character recognition system and the various stages included in the process. External and internal segmentation were also discussed. Moments could be used as a pattern for feature extraction. There were four broad categories for recognizing the character viz. template matching, statistical technique, structural technique and neural networks.

Choudhary (2014) discussed the segmentation strategies for automated recognition of off-line unconstrained cursive handwriting from static surfaces. This paper reviewed the explicit segmentation, implicit segmentation and holistic segmentation techniques and also compared the research results of various researchers in the domain of handwritten word segmentation.

Pal et al. (2003) purposed the segmentation technique of water reservoir for touching numerals for free handwritten. In this concept the water was poured from the top or from the bottom of the numerals. The water was stored in the cavity and we will differentiate the numerals. The accuracy obtained is 94.8%.

Dongre and Mankar (2011) the bounded box method for segmentation of lines, words and characters were used. The histogram obtained by the pixels using present method. The main work was on segmenting lines; words to individual characters of Devanagari script. Some character were connected to various places caused the problem in character segmentation. The accuracy of line, word and character segmentation was about 91%.

Rani and Kumar (2013) proposed the problem occur during the character segmentation. The major problems of character segmentation were described in it. Handwritten characters were not of fixed size caused the problem occurred were broken character, overlapped, touching, skewed and of irregular intensity.

Tapkir and Shelke (2012) described about the segmentation based on profiles. The segmentation was by line, word and then character segmentation. OCR gave wrong result in character recognition due to the error in character segmentation. It also described the components of OCR system; the work was on Devanagari script. There was the 100% experimental result of line segmentation and about 98% accuracy of character segmentation. For feature extraction the use of Euclidean Minimum Distance Classifier that gave around 92.77% result.

Kumar and Singh (2010) discussed the detection and segmentation of handwritten document, to segment the document they used the scanned image and segment into lines, word and finally to characters. For the segmentation the concept of flexible window was used, this window could be adjusting the size according to the need of the document.

Table I: Result of various techniques

Author	Method	Result
Amit Choudhary,Rahul Rishi and Savita Ahlawat[4]	Vertical Segmentation Approach	83.5%
Amjad Rehman, Dzulkifli Mohamad and Ghazali Sulong[5]	Implicit Based Segmentation	79.23
	Explicit Based Segmentation	80.91
Hong Lee, Brijesh Verma[7]	Binary Segmentation with Neural validation	61.4%
Rajiv Kumar, Amardeep Singh[9]	Segmentation with Flexible Window	89.89%
U.Pal, A.Belaid and Ch.Choisy[10]	Water Reservoir	94.8%
Vinaya.S.Tapkir, Sushma.D.Shelke[12]	Segmentation Based on Projection Profile	92.77%
Vikas J Dongre, Vijay H Mankar	Histogram Based Character Segmentation Approach	55%

CONCLUDING REMARK FROM LITERATURE

In this paper three segmentation based approaches for cursive handwriting recognition are presented. By the detailed analysis of the literature, it is observed holistic approach is more suitable for applications where the lexicon is statically defined. Explicit segmentation based approach was computationally complex than implicit segmentation, but gives slightly better results than less complex implicit based segmentation approach.

REFERENCES

[1] A. Phukan, M. Borah, "A survey paper on character recognition focusing on offline character recognition," International Journal of Computer Engineering and Applications, vol. 6, pp. 51-60, 2014.

[2] A. Choudhary, "A review of various character segmentation techniques for cursive handwritten words recognition," International journal of Information and Computation Technology, vol. 4, pp. 559-564, 2014.

[3] A. Choudhary, R. Rishi, and S. Ahlawat, "A new approach to detect and extract characters from off-line printed images and text," Information Technology and Quantitative Management, pp. 434-440, 2013.

[4] A. Choudhary, R. Rishi, and S. Ahlawat, "A new character segmentation approach for off-line cursive handwritten words," Information Technology and Quantitative Management, pp. 88-95, 2013.

[5] A. Rehman, D. Mohamad, and G. Sulong "Implicit Vs Explicit based script segmentation and recognition: A performance comparison on benchmark database," International Journal Open Problems Compt. Math., vol. 2, pp. 352-364, 2009.

[6] H. Lee, and B. Verma, "Binary segmentation with neural validation for cursive handwriting recognition," Proceedings of International Joint Conference on Neural Networks, pp. 1730-1735, 2009.

[7] M. Brodowska, "Oversegmentation methods for character segmentation in off-line cursive handwritten word recognition," Schedae Informaticae, vol. 20, pp. 44-65, 2012.

[8] R. kumar, and A. Singh, "Detection and segmentation of handwritten text in Gurmukhi script using flexible windowing," International journal of Computer Theory and Engineering, vol. 2, pp. 329-332, 2010.

[9] U. Pal, A. Belaid, and Ch. Choisy, "Touching numeral segmentation using water reservoir concept," Elsevier Science B.V., pp. 261-272, 2002.

[10] V. Rani, and P. Kumar, "Problems of character segmentation in handwritten text documents written in Devnagari script," International Journal of advanced Research in Computer Engineering and Technology, vol. 2, pp. 1026-1029, 2013.

[11] V. S. Tapkir, and S. D. ShelkeRishi, "OCR for handwritten Marathi script," International Journal of Scientific and Engineering Research, vol. 3, pp. 1-6, 2012.

[12] V. J. Dongre, and V. H. Mankar, "Devnagari document segmentation using histogram approach," International journal of computer science, Engineering and Information Technology, vol. 1, pp. 46-53, 2011.

[13] R. K. Nath, and M. Rastogi, "Improving various off-line technique used for handwritten character recognition," International journal of Computer Applications, vol. 48, pp. 11-17, 2012.

★★★