Comparative Study on Text Segmentation Techniques

Dr. Bharat C. Patel[1] and Dr. Jagin M. Patel[2]

[1]*Associate professor, Smt. Tanuben & Dr. Manubhai Trivedi college of Information science, Surat, Gujarat,India.*

[2]*Assistant Professor, M. K. Institute of Computer studies, Bharuch, Gujarat, India.*

E-mail: [1] patelbharat99@yahoo.co.in *, [2] jagin_2k@yahoo.com

Abstract

Text segmentation, whether printed, handwritten or cursive, is one of the most complicated phases in any OCR. The accuracy of recognition will be heavily reliant on good segmentation. Image segmentation is a crucial component of image analysis and the field of computer vision. Researchers have developed several techniques for segmentation, each of which is used for different types of segmented objects. At present no any universal method is available for image segmentation. Existing image segmentation techniques are not capable to deal with images of any types. This survey looked at a variety of image segmentation techniques, evaluated them, and discussed the issues that came up as a result of using them.

Keywords*: Image segmentation, implicit segmentation, explicit segmentation, holistic approach, region-based segmentation.*

1. Introduction

Image segmentation is the technique of subdividing an image into small regions in order to analyze them separately and get detailed information. An image is divided into different components as per the features of pixels to recognize objects or boundaries to simplify an image and can be examining more efficiently. Image segmentation is one of the intermediate steps in image processing. It plays an important role in various applications like text segmentation, medical applications, remote sensing, aerial imaging, pattern recognition, and many more. There are many image segmentation methods are available that can be applied to a particular type of image to analyze the interesting segment of the image. There is no universally accepted image segmentation method can be available for all types of images and a particular type of image. Image segmentation is used to distinguish different objects in the image.

2. Segmentation

There is the various meaning of segmentation depending on context. Segmentation means to separate text area from the background [1], lines from a paragraph, words from lines, characters from the word, etc. So, during the text segmentation process, a document is segmented into its logical sub-components, such as text and graphics, lines in a paragraph, words in a line, and characters in words.

Script recognition relies greatly on segmentation. Correct segmentation is a base for precise script recognition [2]. The selection of the segmentation approach influences the technique employed in subsequent processing steps, such as feature extraction and recognition of characters. Different segmentation approaches may create different characters. As a result of this issue, the classifier's performance suffers.

A major difficulty in analyzing segmentation is how to categorize methods. For example, Tappert et al. [3] categorized into “external” vs. “internal” segmentation, based on whether recognition is necessary in the procedure, Dunn and Wang [4] named as “straight segmentation” and “segmentation-recognition”, while [5] categorized image segmentation into two main categories: layer-based segmentation and block-based segmentation. Rehman and Saba [2] talk about the explicit and implicit segmentation as well as the issues they cause in English cursive writing.
Apart from this, several segmentation approaches exist, which can be divided into two categories according to the techniques used [6,7,8,9] namely analytical and holistic. The analytical approach is classified into explicit segmentation [10, 11] and implicit segmentation [12, 13].

2.1. Implicit Segmentation

The implicit approach is also known as recognition-based segmentation or straight segmentation because in this approach characters are segmented and recognized at the same time. The implicit segmentation method does not require the words to be segmented into characters prior to recognition. In this approach, the system looks for components in the image that match classes in its alphabet.

The maim benefit of implicit segmentation is that they overcome the problem of segmentation. They do not require any complex "dissection" algorithms, and recognition errors are mainly caused by poor classifications. This form of segmentation is typically constructed with rules that aim to identify all of the segmentation points for the character.
The majority of implicit segmentation methods are language-independent. Their accuracy is relay on classification success. They are utilized to get around the difficulties of cursive script segmentation. Apart from this, they provide all temporary segments and allow the recognizer/classifier to select the most appropriate segmentation. As a result, implicit segmentation-based recognition methods need a significant amount of training data.

While using fewer segments cuts down on computation time, it also increases the problem of under-segmentation. Furthermore, when there is overlapping between adjacent characters problem increase because it is necessary to recognize all possible combinations of valid characters rather than just valid characters. On the flip side, the uses of a large number of segments decreases under segmentation issue and hence decrease the number of ligatures. But, it creates more segments which increases the computational time as well as over-segmentation issues. Furthermore, portions of characters are recognized as valid characters, a problem known as class overlapping.

Naz et. al. [6] suggested implicit segmentation of printed Urdu text lines, which is in the Nasta'liq writing style. They used Multidimensional Long Short-Term Memory (MDLSTM) Recurrent Neural Networks.

Koteswara and Negi [9] proposed Telugu printed text recognition model based on HMM. This model calculates statistical features intensity and derivative of intensity by means of sliding window technique on the training set of word images.

In order to generate segmentation cuts, the researcher normally considers a set of heuristics and information from the background [14,15,16], the foreground [17,18,19], or a combination of these [20,21]. One of the key drawbacks of most of these algorithms is the huge number of cuts that must be evaluated, and the number of heuristics that must be set. Vellasques et al. [22] proposed a method for reducing the number of segmentation cuts.

To avoid explicit segmentation as well as the complication of setting various heuristics, various authors proposed implicit segmentation to recognize strings of digits [23,24]. The major disadvantage is the high sensitivity to slanted images.

Cavalin et. Al. [25] presented work for Recognition of Handwritten Strings of Characters using implicit segmentation. Strings of Characters may be words or numerals. To solve the difficulty of finding the optimum balance between segmentation and recognition, the authors proposed a two-stage HMM-based technique. The first step is divided into three modules: (i) pre-processing (PP), (ii) foreground feature extraction (FFE), and (iii) segmentation recognition (SR). The isolated digits are used to train the character HMMs. Foreground and background features are combined to boost the recognition rate during the second stage, which is the verification stage.

Radwan and khalil [26] developed a method that is based on the multichannel neural network for character segmentation. This method finds out the characteristics of a segmentation window and calculates the probability of the current window to a segmentation area. Rosenbeg and Nakum [27] SIFT algorithm extract local feature of character for classification. Each word scanned using growing window sizes, which is results in the setting of segmentation points where the classifier received the highest confidence.

2.2. Explicit Segmentation

It is Pure Segmentation. Here, segments are identified based on “character-like” attributes. This process cuts an image into meaning full sub-images or components, it is therefore also known as dissection segmentation [2]. If explicit segmentation is employed for word segmentation, Words are explicitly split or cut into characters, which are subsequently categorized or recognized separately. Due to the complexity of finding optimal word hypotheses, this strategy is more expensive [28].

In explicit segmentation, word is separated into smaller independent units/characters. These units may be ligatures, characters, or part of the character(strokes), depending on a set of a given hypothesis, attributes, or rules which are utilized to decide the validity of the segmentation points [8,29]. A few of the common dissection methods used in OCR systems are white space and pitch discovery, connected component processing, projection analysis, etc

Explicit segmentation methods are based on: projection, Contour Tracing, Upper Distance Function, Skeletonization, Morphological Operations, and Template Matching.

(a) A Segmentation Methods Based On Projection: The purpose of this method is to simplify the segmentation of text by reducing two-dimensional information into one dimension. It performs better with printed documents and poorly with handwritten text. Projection profile methods are normally used for lines, words, and characters segmentation. They are computationally easy and get good results for simple font types. There are horizontal and vertical projection profiles methods. The horizontal and the vertical projection profile method are used for lines segmentation and words or characters segmentation respectively.
When the vertical projection is used for cursive text, it may result in (i) over-segmentation (ii) Under-segmentation. Over-segmentation is occurred because characters are formed with the help of several parts whereas under-segmentation is occurred because of neighbouring characters are overlie.

(b) Segmentation Methods Based on the Upper Distance Function: The upper distance function is a particular type of vertical projection. It is set of the highest points in each column [30].

(c) Segmentation Methods Based On Contour Tracing: It is possible to accomplish segmentation by tracing the outer contour of a word [31].
The contour tracing methods examine the structural shape of characters as they are scanned, avoiding the difficulties caused by the thinning process. Conversely, in many situations, the contour must first be smoothed.

These methods give a clear narrative of the shape or outline of the characters which can help with the under segmentation issues produced by characters overlapping [32]. Moreover, it reduces baseline extraction errors by eliminating the need to adjust the baselines multiple times. On the other hand, this type of segmentation suffers from over-segmentation issues due to the presence of noise and characters brakes.

(d) Segmentation Methods Based On Template Matching: Template matching methods [32, 33] discover the character's probable cutting points depending on the sliding window as well as predefined character templates. It all starts with establishing the baseline. Then, by sliding the templates over the text, it looks for matches between the templates and the text-image [29,32,34].

The template matching approach works well for printed text with simple fonts. Though, it depends on the variation in size of the characters. When using more font types and styles, the performance suffers as the number of predefined segments grows. Checking all predefined templates is computationally expensive.

(e) Segmentation Methods Based On Morphological Operations: In handwriting, in any language, the majority of characters are linked by horizontal lines. Therefore, Morphological operations such as closing and opening are significant for segments of the word.

(f) Segmentation Methods Based On Skeletonization-based: The character skeleton stores vital information about a character, which is used to recognize the character. The methods based on Skeletonization [35,36,37,38,39] utilize a variety of morphological operations such as opening, closing, and thinning. It recognizes character or word elements based on features like curvatures, boundaries, skeletons, angles, etc that define region shape information.

2.3. Holistic Approaches

To keep away from the hardships of the segmentation stage, researchers emerged with an alternative approach known as a holistic or global approach, in which recognition is generally achieved over the entire representation of words and where no attempt is made to recognize characters individually. Hence this approach is known as the segmentation-free approach [40]. It recognizes an entire word as a unit.

One major disadvantage of this type of algorithm is that its application is usually limited to a predefined lexicon. Because this approach deals with words rather than letters, recognition is limited to a given lexicon of words. This method is best suited to applications where the lexicon is static and unlikely to change, such as the recognition of bank checks.

Region-Based segmentation

The most important objective in this method is that the segmentation is performed in particular image based on similarities and dissimilarities. It looks for similarities in adjacent pixels which possess the same attributes and are grouped into unique regions. The regions formed by region based method have the following features.

- The summation of all the regions is equal to the whole image.
- Each region is contiguous and connected
- A pixel belongs to a single region only; hence there is no overlap of pixels.
- Each region must satisfy some uniformity condition
- Two adjoining regions do not have anything in common.

There are four different approaches to carry out region-based segmentation such as: region growing, splitting, merging and split & merge.

(I) Region Growing: In region growing, the approach is to start with set of seed points and the region is growing by aggregation of those neighboring pixels having similar gray level, texture, color, shape properties. When edges are difficult to detect in noisy image then region growing based techniques gives better result than the edge-based techniques.

(II) Region Splitting: Primarily the complete image is considered to be a single region. During region splitting, the original image subdivides into the regions such that each region assurance a condition of uniformity.

(III) Region merging: It works opposite of region splitting. Region merging process start with small region and the region having similar characteristics are collected to obtained region having similar gray level and variance.

(IV) Split and merge: Both split and merge processes are carried out parallel to get the desired region.

Region based segmentation is useful in the applications such as in medical images to find the tumors, veins etc, in satellite images to find target object, in surveillance images to find the people, video summarizing and may more.

3. Application areas of image segmentation

(i) Medical applications: Image segmentation is useful to diagnosis of various diseases that are related to heart, brain, knee, spine, prostate, pelvis, and blood vessel and pathology localization.

(ii) Remote sensing:** In remote sensing, it detects energy that is reflected by the earth and gets the data from it. For example, take an example of the satellite which captures the data from the surface of the earth and then it can be analyzed in many ways like which area is green or where is the water present also we can also check in which areas population is increased. The remote sensing of space-based is responsible for monitoring the changes in our environment like deforestations (cutting down the trees), ecosystem degradation, changes in the forest carbon stocks, and carbon recycling, among others. These models are prepared to capture the dynamics of changes to and impacts on global ecology.

(iii) Aerial imaging: It involves the analysis of images/photographs which are taken through an aircraft or flying object. For example, capturing an image using a drone without going anywhere or any location and getting the important details in very short time duration. It is useful in recognition tasks like face recognition, fingerprint recognition, iris recognition, and many more.

(iv) Machine vision and Robotics: Many robotic machines work on digital image processing. Robots are capable to find their ways using image processing technique such as finding the obstacle on the path and line follower robot.

(v) Pattern recognition: This field is useful for the study of image processing. Pattern recognition is also combined with artificial intelligence to implement different application such as computer-aided diagnosis, natural language processing and images recognition. Currently, image processing is used for pattern recognition.

4. Future direction and conclusions

In this paper, a variety of segmentation strategies for printed/handwritten/cursive word/character are discussed. By a thorough review of the literature, it is observed that there is no global segmentation method. The Holistic approach is better suited for applications area where the lexicon is defined statically. The explicit segmentation strategy requires more computation than the implicit segmentation strategy, but it produces slightly better results.

A brief survey of segmentation strategies was conducted in this paper, which may assist researchers in the field of OCR in discovering new ideas and providing new solutions to the challenges of printed/handwritten/cursive text segmentation.

References:

[1] *J. M. Patel and A. A. Desai, "Gujarati Text Localization, Extraction and Binarization from Images", International Journal of Computer Sciences and Engineering vol. 6, no. 8 (2018), pp. 714-724.*

[2] *A. Rehman and T. Saba, "Off-Line cursive script recognition: current advances, comparisons and remaining problems", Artif Intell Rev, vol. 37, no. ,4 (2012), pp. 261- 288.*

[3] *CC Tappert, CY Suen and T. Wakahara, "The state of the art in on-line handwriting recognition", IEEE Trans Pattern Anal Mach Intell, vol. 12, no. 8, (1990), pp. 787–793.*

[4] *CE Dunn and PSP Wang, "Character segmenting techniques for handwritten text—a survey", In Proceedings of 11th international conference on pattern recognition, vol. 2, (1992), pp 577–591.*

[5] *Nida M. Zaitoun and Musbah J. Aqel, "Survey on Image Segmentation Techniques", Procedia Computer Science, volume 65,(2015), pp. 797-806.*

[6] *Naz Saeeda, Arif Iqbal Umar, Riaz Ahmed, Muhammad Imran Razzak, Sheikh Faisal Rashid and Faisal Shafait, "Urdu Nasta'liq text recognition using implicit segmentation based on multi-dimensional long short term memory neural networks", SpringerPlus, vol. 5, no. 1, (2016), pp. 1-16.*

[7] *Naz Saeeda, Arif I. Umar, Syed H. Shirazi, Saad B. Ahmed, Muhammad I. Razzak and Imran Siddiqi. "Segmentation techniques for recognition of Arabic-like scripts: A comprehensive survey", Education and Information Technologies, vol. 21, no. 5, (2016), pp. 1225-1241.*

[8] *A. Choudhary, "A review of various character segmentation techniques for cursive handwritten words recognition", International Journal of Information & Computation Technology, Vol. 4, No. 6, (2014), pp.559-564.*

[9] *Rao D. Koteswara and Atul Negi, "An implicit segmentation approach for Telugu text recognition based on hidden Markov models.", In Advances in Signal Processing and Intelligent Recognition Systems, vol. 425, (2015), pp. 633-644.*

[10] *El-Yacoubi A, Gilloux M, Sabourin R and Suen CY, "An HMM-based Approach for on-line unconstrained handwritten word modeling and recognition", IEEE Trans Pattern Anal Mach Intell, Vol. 21, no. 8, (1999), pp. 752–760.*

[11] *N. Arica and Yarman-Vural FT, "Optical character recognition for cursive handwriting", IEEE Trans Pattern Anal Mach Intell, vol. 24, no 6, (2002), pp. 801–813.*

[12] *M. Gillies, "Cursive word recognition using hidden markov models", In Proc fifth US postal service advanced technology conference, (1992), pp. 557–562.*

[13] *W. Cho, SW Lee and JH Kim, "Modeling and recognition of cursive words with hidden Markov models", Pattern Recognit, vol. 28, no. 12, (1995), pp.1941–1953.*

[14] *M. Cheriet, Y.S. Huang and C.Y. Suen, "Background region based algorithm for the segmentation of connected digits", In Proceedings of the 11th International Conference on Pattern Recognition, vol. 2, (1992), pp. 619–622.*

[15] Lu, Z., Chi, Z., Siu, W., Shi, P., "A background-thinning-based approach for separating and recognizing connected handwritten digit strings", Pattern Recognit, vol. 32, no. 6, (1999), pp. 921–933.

[16] U. Pal, A. Belaid and C. Choisy, "Touching numeral segmentation using water reservoir concept", Pattern Recognition Letter, vol. 24, no. 1-3, (2003), pp. 261– 272.

[17] K. K. Kim, J. H. Kim and C. Y. Suen, "Segmentation-based recognition of handwritten touching pairs of digits using structural features", Pattern Recognit. Lett, vol. 23, no. 1, (2002), pp. 13–21.

[18] E. Lethelier, M. Leroux and M. Gilloux, "An automatic reading system for handwritten numeral amounts on french checks", In Proceedings of 3rd International Conference on Document Analysis and Recognition, (1995), pp. 92–97.

[19] J. Sadri, C. Y. Suen and T. D. Bui, "A genetic framework using contextual knowledge for segmentation and recognition of handwritten numeral strings", Pattern Recognition, vol. 40, no. 3, (2007), pp. 898–919.

[20] Y. K.Chen and J. F. Wang, "Segmentation of single- or multiple-touching handwritten numeral string using background and foreground analysis", IEEE Trans. Pattern Anal. Mach. Intell, vol. 22, no. 11, (2000), pp. 1304– 1317.

[21] L. S. Oliveira, R. Sabourin, F. Bortolozzi, and C. Y. Suen, "Automatic recognition of handwritten numerical strings: a recognition and verification strategy", IEEE Trans. Pattern Anal. Mach. Intell, vol. 24, no. 11, (2002), pp. 1438–1454.

[22] E. Vellasques, L. S. Oliveira, A.S. Jr. Britto, A. Koerich and R. Sabourin, "Filtering segmentation cuts for digit string recognition", Pattern Recognition, vol. 41, no. 10, (2008), pp. 3044–3053.

[23] A. S. Britto, R. Sabourin, F. Bortolozzi and C. Y. Suen, "The recognition of handwritten numeral strings using a two-stage HMM-based method", IJDAR, vol. 5, (2003), pp. 102–117.

[24] S. Procter and A. J. Elms, "The recognition of handwritten digit strings of unknown length using hidden markov models", In Proceedings of 14th International Conference on Pattern Recognition, (1998), pp. 1515– 1517.

[25] P. R. Cavalin, Alceu de Souza Britto Jr, Flavio Bortolozzi, Robert Sabourin and Luiz E. Soares Oliveira, "An Implicit Segmentation based Method for Recognition of Handwritten Strings of Characters", Proceedings of the 2006 ACM symposium on Applied computing - SAC '06, (2006), pp. 836-840.

[26] M. A. Radwan and M. I. A. H. Khalil, "Predictive segmentation using multichannel neural networks in arabic ocr system", In Artificial Neural Networks in Pattern Recognition, Springer International Publishing, Cham, (2016), pp. 233–245.

[27] Andrey Rosenberg, and Nachum Dershowitz, "Using SIFT descriptors for OCR of printed Arabic", Tel Aviv University, (2012).

[28] A. Cheung, M. Bennamoun and N. W. Bergmann, "An Arabic Optical Character Recognition System using Recognition-Based Segmentation", Pattern Recognition, vol. 34, no. 2, (2001), pp. 215-233.

[29] A. Lawgali, "A survey on arabic character recognition", International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 8, no. 2, (2015), pp. 401–426.

[30] Ali Hussein Saleh, "Technology Diffusion and Adoption: Global Complexity, Global Innovation", Technology & Engineering, Idea Group,U.S. (2013).

[31] M. Omidyeganeh, K. Nayebi, R. Azmi, and A. Javadtalab, "A new segmentation technique for multi font farsi/arabic texts", In Proceedings. (ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2. (2005).

[32] R. Saabni, "Efficient recognition of machine printed arabic text using partial segmentation and hausdorff distance", In 2014 6th International Conference of Soft Computing and Pattern Recognition (SoCPaR), (2014), pp. 284–289.

[33] Y. M. Alginahi, "A survey on arabic character segmentation", Int. J. Document Analysis and Recognition, vol. 16, no. 2, 2013, pp. 105–126.

[34] Y. Zhang, Z. Q. Zha and L. F. Bai, "A license plate character segmentation method based on character contour and template matching", In Applied Mechanics and Materials, vol. 333, (2013), pp. 974–979.

[35] S. Naz, A. I. Umar, S. H. Shirazi, S. B. Ahmed, M. I. Razzak and I. Siddiqi, "Segmentation techniques for recognition of arabic-like scripts: A comprehensive survey", Education and Information Technologies, vol. 21, no. 5, (2016), pp. 1225–1241.

[36] M. M. Altuwaijri and M. A. Bayoumi, "A thinning algorithm for arabic characters using art2 neural network", IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing, vol. 45, no. 2, (1998), pp. 260–264.

[37] D. Motawa, A. Amin and R. Sabourin, "Segmentation of arabic cursive script", Proceedings of the Fourth International Conference on Document Analysis and Recognition, vol. 97, (1997), pp. 625–628.

[38] B. Timsari and H. Fahimi, "Morphological approach to character recognition in machine printed persian words", In Proceeding of SPIE. Document Recognition III, (1996).

[39] F. U. Qomariyah, and W. F. Mahmudy, "The segmentation of printed arabic characters based on interest point", Journal of Telecommunication, Electronic and Computer Engineering, vol. 9,no. 2-8, (2017), pp. 19–24.

[40] J. Ahmad, "Optical character recognition system for arabic text using cursive multi-directional approach", Journal of Computer Science, vol. 3, no. 7, (2007), pp. 549–555.

[41] B. Al-Badr and R. Haralick, "A Segmentation-free Approach to Text Recognition with Application to Arabic Text", International Journal on Document Analysis and Recognition (IJDAR), vol. 1, no. 3, (1998), pp. 147-166.