International Journal of Future Generation Communication and Networking
Vol. 9, No. 11 (2016), pp. 137-152
http://dx.doi.org/10.14257/ijfgcn.2016.9.11.13

A Survey on Methods for Basic Unit Segmentation in Off-Line Handwritten Text Recognition

Aysadet Abliz, Wujiahemaiti Simayi, Kamil Moydin and Askar Hamdulla

Institute of Information Science and Engineering, Xinjiang University, China
Corresponding Author: askarhamdulla@sina.com

Abstract

Studies on recognizing different kind of handwritten texts have been conducted and achieved great success for some letters. This paper reviews the segmentation techniques on English handwritten recognition, which is one of the most successful one up to date. Also, considering the very much relations between Arabic and Uyghur which we are aiming to get progress on its handwritten recognition technology, references from Arabic handwritten recognition are very much hoped to get. Characteristics of Uyghur handwriting texts and some of the encountered difficulties are described. Then referencing the successful work on English and Arabic basic unit segmentation, this paper tries to give some suggestions for Uyghur basic unit segmentation research.

Keywords*: basic units, segmentation, off-line, handwritten recognition*

1. Background

With the wide application of computers, people prefer to use computers for their daily needs and more convenient applications are being welcomed. For example, the need for recognizing the text documents both in printed and handwritten form makes the Optical Character Recognition (OCR) technology a must. OCR became an important research field of pattern recognition (PR), it plays an important role in the development of pattern recognition.

OCR is the machine replication of human reading and has been the subject of intensive research for more than three decades. OCR can be described as mechanical or electronic conversion of scanned images where images can be handwritten, typewritten or printed text. It is a method of digitizing printed texts so that they can be electronically searched and used in machine processes. It converts the images into machine-encoded text that can be used in machine translation, text-to-speech and text mining.

Character recognition is an important research direction in text recognition. According to different recognition objects and recognition process, it can be divided into following categories as shown in Figure 1.

It is perhaps acceptable to give wider definition for the term 'Character'. Nowadays, the character recognition technology refers the word recognition and the related work. However, there are different kinds of scripts in real application. Some of them are alphabetic, such as Latin based and Arabic based scripts. For theses scripts, character is referred to the basic or the smallest unit to compose a word and sentences at last. Contrary to the alphabetic scripts, symbolic scripts do not clearly show its smallest unit of script. For example, in Chinese a character may be a word which contains full meaning, and sometimes a few characters make a word. But the character that can utter meaning is composed of several strokes. In fact, the strokes are the basic units in Chinese script. So, in symbolic scripts, a character sometimes shows a word, and sometimes refers a sub-content of a word.

ISSN: 2233-7857 IJFGCN

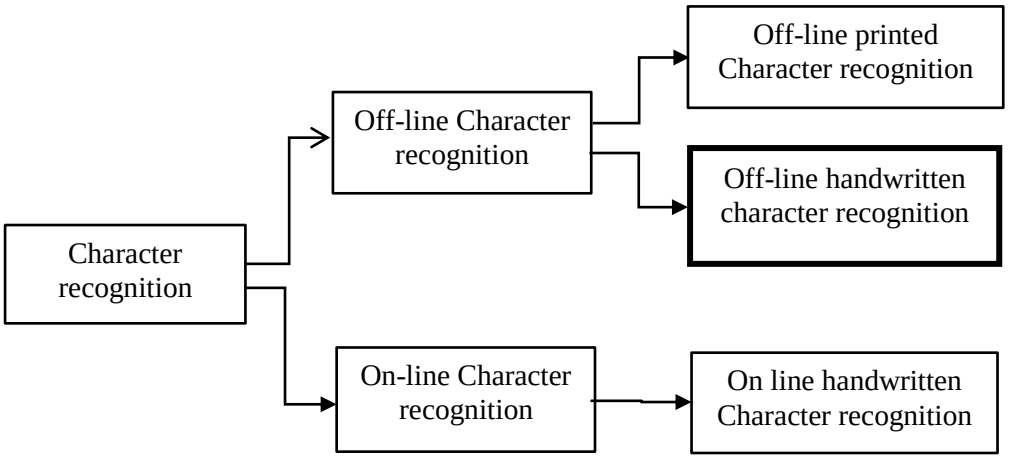


Figure 1. Classification of Character Recognition according to the Research Object

Basic units (word, connected component - CCs, character or letter *etc*.) segmentation is a critical stage in word or character recognition, as well as word spotting. The basic units mentioned in this paper are a word or part of a word in the handwritten text.

The primary assumptions in most of the word segmentation approaches are that: i) the document is already segmented into text lines, ii) each CC belongs to only one word and iii) gaps between words are greater than gaps between consecutive segments which belonging to the same word[28]. Furthermore, due to the irregular spaces between words and variations of writing styles depending on person who writes it, more challenging problems have to be considered in the segmentation of handwritten documents comparing with machine printed documents.

With the upsurge in the research field of OCR, relevant studies have been carried out on Uyghur texts, too. There are very few studies on Uyghur basic unit segmentation. Although there are some achievements on the recognition of printed documents [2]-[6], Offline handwritten recognition still needs more effort and intensive research on it. In order to find better solutions for the problems which encountered in Uyghur handwritten text recognition, we reference some papers from English and Arabic handwritten text recognition, firstly. And then, on the basis of summarizing the ideas and methods from those papers, we try to forward our views and suggestions for the further study of Uyghur handwritten text recognition. The suggestions may provide helpful answers to the questions such as what aspects we should start with and what kind of method is more reasonable and effective for the encountered problems.

This paper is organized as follows: Section 2 introduces the basic steps of offline handwritten text recognition process; Section 3 contains English and Arabic research works review; Section 4 describes the characteristics of handwritten Uyghur texts and comparisons with English and Arabic languages; Section 5 gives a brief summary and some suggestions for Uyghur handwritten text recognition, and the contents of this paper is concluded in section 6.

2. Procedures of Handwritten Text Recognition

In spite of great challenges in off-line handwritten character recognition, many researchers have been attracted to this field. Offline handwriting text recognition process consists of several steps: pre-processing, text line segmentation, basic unit segmentation, feature extraction, and classification *etc*.

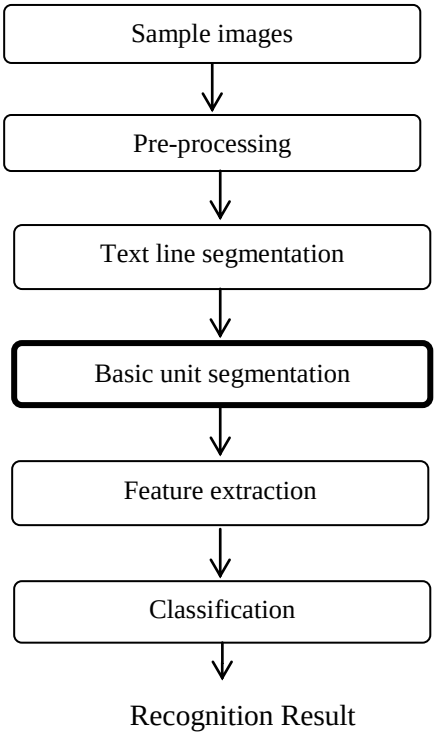


Figure 2. Framework of Offline Handwriting Text Recognition

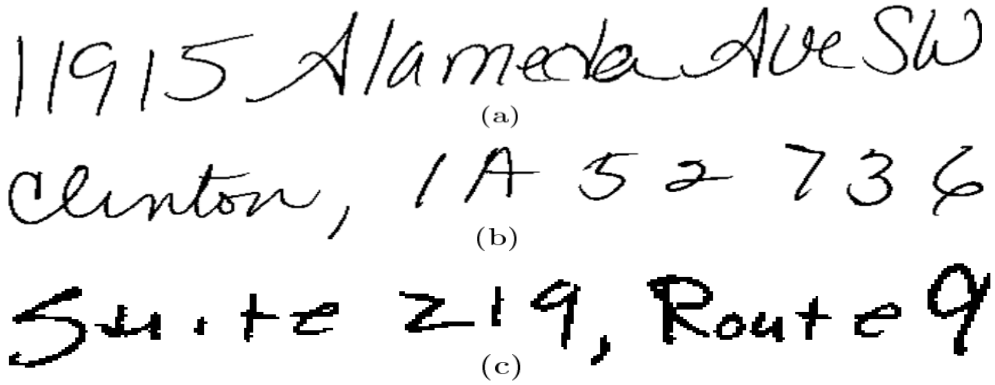
As shown in the diagram, sample collection is the first step for handwriting recognition research. In off-line handwritten text recognition, the writing styles of the samples do not always follow the writing rules, because everyone has his or her own unique writing habits and characteristics. In order to provide conveniences for later processing steps, a preprocessing step which often includes binarization, denoising and normalization techniques *etc*, is required. Segmentation can be defined as the process of dividing or separating an image into smaller sections or useful regions based on some conditions. Segmentation is considered as a core step for any recognition method. Text line segmentation is the beginning of all segmenting work in offline handwritten text recognition. Correct text line segmentation for whole text image and then basic unit segmentation in each line are of great significance in offline handwritten recognition. However, some researchers put forward not doing line segmentation and, instead, propose direct basic unit segmentation and recognition[34]. After the text line segmentation, we have to deal with the basic unit segmentation. The basic unit here refers the units according to different structure characteristics of different scripts. For example, in recognition of English, basic unit is a word or a letter; in Chinese the smallest unit is a Chinese character; in Arabic the smallest unit may be a word or connected component which is easily divided as isolated part; in Uyghur, the basic unit can be a connecting component like Arabic or a word *etc*. The correct segmentation of basic units plays a decisive role in the next step of feature extraction and identification. Finding the unique features of sample is the key to distinguish different samples. The unique features expressed by parameters are the foundation of the recognition. In this step, the classifier selection is very important to get better recognition result. If a document is segmented into basic units then further tasks such as word recognition will be developed.

In offline handwritten text recognition, the segmentation and recognition processes are inseparable and the segmentation is the basis and premise of the recognition. The recognition accuracy depends on the successful segmentation very much. The recognition and segmentation processes are of mutual inspiration and mutual feedback.

3. English and Arabic Research Works Review

3.1. Handwriting Properties of English

An English word is made of one or more English letters. In normal writing of English, there are distances between each letters and between words. Usually, the distance between words is larger than the distance between the letters in a word. Of course, randomness during handwriting, adhesion between words and overlaps are common in handwritten document recognition. Therefore, in English word segmentation, a lot of researchers start from these writing characteristics of English to solve the problems of word segmentation.



(a) shows words that overlap horizontally; (b) shows an inter-character gap (between the digits 2 and 7) that is larger than an inter-word gap (between the character A and the digit 5); (c) shows a text line where many inter-character and inter-word gaps with similar size.

Figure 3. Handwriting Features of English Words

3.2. Related Work on English Word Segmentation

Word segmentation methodology usually go through the three following stages: 1) pre-processing; 2) distance computation and 3) gap classification.[24]

In 1994, a comprehensive discussion on word separation have been carried out by Seni and Cohen *etc*[7]. Eight different distance metrics have been analyzed, and these metrics are used to correct word segmentation results by measuring the number of text lines. A gap classification technique which is based on an iterative procedure over the set of distances and the calculation of a ratio was proposed. Through a study of the drawbacks in previous approaches to gap estimation, in 1995, Seni and Cohen *etc* presented a new technique to estimate inter-component distances, and it got proved to be better than previous method in terms of performance and robustness [8].

In1998, Kazakov and Manandhar described a hybrid approach which is an efficient combination of Genetic algorithms (GA) and Inductive logic programming (ILP) to word segmentation[9].[10] Another methodology that makes use of neural networks was presented by Kim and Govindaraju[10], and obtained an accuracy of 87.36%. In 2008, a similar technique was proposed by Huang and Srihari[20]. The research results was assessed using an unconstrained handwriting database, which contains 50 pages (1026 line, 7562 words images) handwritten documents and obtained word segmentation accuracy of 90.82%.

In 1999, a novel method for segmenting handwritten document images was developed by Manmatha and Srimal[11]. The method analyzed the extent of “blobs” in a scale space representation of the image. The research has been randomly picked from different sections of the George Washington corpus of 6,400 handwritten document images and observed with an average accuracy of around 87 percent. In 2005, Manmatha and Rothfeder presented a scale space approach based on filtering the document images by an

anisotropic Laplacian filter at different scales[18]. This technique was applied on a sample of 100 manuscripts of George Washington and error rate was 17%.

In 2001, a system for recognizing unconstrained English handwritten text based on a large vocabulary was developed by Marti and Bunke. The Convex Hull Distance (CHD) was employed to estimate the gap metric between successive CCs and classified the candidate gaps to "inter" or "intra" words[13]. Testing experiment on 541 text lines, containing 3899 words shows that 95.56% words were correctly segmented.

In 2005, Varga and Bunke presented building a structure tree of the text line and its nodes regarded as possible word candidates[16]. Researches with different gap metrics as well as threshold types showed that the new method produced significant improvements than traditional word extraction methods. In 2005, using unconstrained handwritten carbon copies of PCRs, Nwogu and Gyeonghwan Kim described a valid word segmentation method[17], it was performed for Stroke analyses and extracted image primitives for word detection. At the word boundaries detecting stage, used a heuristics-based approach which was involved gap spacing, height transitions and the average stroke width. Experiments showed the results of 69% correct segmentation.

In 2008,[19] Louloudis *et al*, adopted an SVM-based metric to locate words in each text line[19]. Euclidean distance was viewed as the distance metric and a threshold used between overlapped components. The final word detection rate reached to 91.7%[21]. They extended their work, in[22] to the gap classification stage and developed Gaussian mixture modeling. The SVM-based and Euclidean distance metrics were combined for distance computation[22].

In 2009, a robust evaluation method that was independent in the distance computation and the gap classification stages proposed by Louloudis and Stamatopoulos[24]. In 2009 , two effective techniques for segmenting handwritten documents into text lines and words were presented by Papavassiliou *et al*[26]. Word segmentation stage was based on a gap metric which used the objective function of a soft-margin linear SVM that separates successive CCs.

In 2011, Simistira *etc* presented a technique to enhance the already existing method for handwritten word segmentation by exploiting local special features [28]. In 2015, [32] formulated the word segmentation as a binary quadratic assignment problem and considered correlations between the gaps as well as the likelihoods of individual gaps[32]. In this formulation, all parameters are estimated based on the Structured SVM framework, regardless of writing styles and written languages without user-defined parameters, the proposed method worked well. In 2015, [34]based on Wigner-Ville distribution, Kavallieratou proposed a novel technique for Word segmentation [34]. Training and line segmentation were not required in this technique, but it is adapted to the writing style of the document image. The technique was tested on the subset of ICDAR2013 Handwriting Segmentation Contest and the results were shown as promising.

Table1. Proposed Methods in Previous References

Ref No.	Year	Proposed methods	Data-set Size	Accuracy
[7]	1994	Evaluated eight different distance measures between pairs of connected components for word segmentation.	1084 text lines for test	>90%
[8]	1995	Estimated inter-component distances that was based on the gap between their convex hulls	1084 postal line images	93.17%
[9]	1998	Described the hybrid approach which was efficiently combination of GA and ILP		
[10]	1998	Neural networks	518 images were	85%

			used for training	
[11]	1999	Proposed a noval method that was analyzed the extent of "blobs" in a scale space representation of the image	George Washington corpus of 6,400 handwritten document images	average accuracy of 87%
[13]	2001	Employ the CHD to estimate the gap metric between successive CCs	NG(Not Given)	48%
[18]	2005	Presented a scale space approach based on filtering the document image by an anisotropic Laplacian filter at different scales.	100 randomly documents from the George Washington corpus of handwritten document images	83 %
[16]	2005	Propose to build a structure tree of the text line, whose nodes represent possible word	H_{dev} and H_{test}, of the IAM-Database	>90%
[17]	2005	At the word boundaries detecting stage, used a heuristics-based approach	416 words were tested in 35 text line images	69%
[19]	2008	Adopted to an SVM-based metric to locate words in each text line	tested in the ICDAR2007 handwriting segmentation contest	>90%
[21]		Euclidean distance was viewed as the distance metric and a threshold used between overlapped components		
[22]		Presented use of Gaussian mixture modeling for the gap classification stage and for the distance computation stage combination of two different distance metrics	test set of the ICDAR2007 handwriting segmentation competition	92.3%
[24]	2009	For the distance computation-DC stage implemented 7 different gap metrics, tested 5 different gap classification-GC methodologies	test set of the ICDAR 2007 Handwriting segmentation competition	DC 97.51%, GC 92.9%
[26], [28]	2009 2011	A soft-margin linear SVM was used to separate consecutive CCs.	tested on the bench marking datasets of ICDAR07 handwriting segmentation contest	F Measure 93.01%
[32]	2015	Considers pair-wise correlations between the gaps as well as the likelihoods of individual gaps	ICDAR 2009/2013 handwriting segmentation databases	Average: 92.82%
[34]	2015	Based on Wigner-Ville distribution, proposed a novel technique for Word segmentation,	tested on the subset of ICDAR2013 Handwriting Segmentation Contest	FM:83.8l%

From table 1, it can be seen that the main idea in English basic unic segmentation is measuring the inter-character gap and inter-word gap. After determining the different thresholds, words or basic units are segmented by classification methodologies. For the distance computation stage, gap metrics methods including Bounding Box method, Euclidean method, Minimum run length method, Average run length method, Convex hull method and the hybrid methods with appropriately combining several methods according to the task, are employed. Gap classification methodologies such as global/adaptive threshold, the unsupervised learning techniques such as clustering (SVM) and Gaussian Mixture Model (GMM) and Scale space selection approach supervised-learning techniques such as neural networks are applied for gap classification. Neural network and SVM classifiers are proven to be efficient for segmentation tasks, too. In upcoming research procedure, based on the common methods, we have to pay close attention to new methods and ideas.

3.3. The Characteristics of Arabic Handwritten Text and Related Research

Arabic is one of the main languages in the world with its great influence on the culture and literature of different people. It is spoken by 234 million people and has great influence for the daily writing scripts of many more. The characters of Arabic script and similar ones are still used by a high percentage of the world's population, such as Arabic, Farsi (Persian), and Urdu. Arabic has 28 characters, and Most of the Arabic characters change their shape based on their location within a word. Correspondingly, Arabic characters are written in the beginning, medium, ending or isolated forms. An Arabic word is composed of many sub-words known as PAW. The writing direction of Arabic text is from right to left.

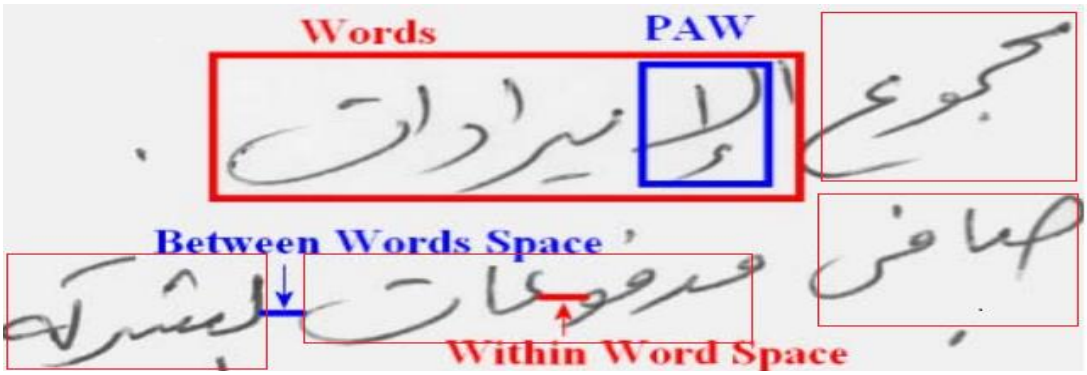


Figure 4. Handwritten Arabic

The Arabic writing style is very different from other languages. For example, English and many other scripts have a "one-letter-after-another" guideline; The smallest unit in Chinese is a stroke and a few strokes constitute a Chinese character that is strictly constrained in printed as well as handwritten documents. But Arabic words follow a mixed and somewhat jumbled writing style that many letters are written before the previous one is finished[40]. So touching and overlapping components and irregularities appears in handwritten Arabic texts so much often.

Quite a few detection approaches have been proposed in literature for Arabic handwritten text recognition. Among them, the projection profile analysis is the most commonly used algorithm[49]. But the projection profile method does not work well on Arabic script, because Arabic text is written with tight and sticky. A lot of the over-lapping letters are missed when projection profile method is used to separate the individual connected components or part of Arabic words (PAW). In Arabic basic unit segmentation, due to the problem of not having clear boundaries of words, extracting

words form text image is not easily realized. Therefore, researchers spotted approaches tend to segment documents into PAWs rather than words, and then find ways to reconstruct the words from the PAWs[40]-[48].

In order to circumvent the problem of word segmentation in Arabic documents, Sarii and Kefali[48] preferred to segment the document into major connected components (PAWs) instead of words[48]. To avoid pre-clustering, Saabni and El-Sana proposed segment the documents into PAWs[47]. Based on shape matching, Moghaddam and Cheriet[46] presented an Arabic word spotting system[46]. Euclidean distance technique and DTW were used for extracting the connected components from the documents and then created their library of PAWs (basic connected components) and clustered it into meta-classes. All three approaches[46]-[48] searched for PAWs rather than words.[42] Aghbari and Brook presented a novel holistic technique for segmenting and classifying HAH manuscripts[42]. The image of HAH manuscript was segmented into words and connected parts. Considering the situation of overlap between the adjacent connected parts of a single word, they developed a str*etc*hing algorithm which is able to reduce the overlap between connected parts and improved the overall results. The accuracy of segmentation improved from 82.11% to 93.16%. Lawgali *etc* proposed an algorithm which was relied heavily on the horizontal and vertical projection method in breaking up words into sub-words and characters [50]. During the research, a lot of overlapping characters were lost and considered as noise so that were removed from the final result. Similarly, Osman used contour analysis to automatic segmentation for the Arabic handwritten text[45]. The horizontal projection was used for line segmentation and vertical projection was applied for word and sub-word segmentation. Due to the overlapping within words, some of sub-words were incorrectly segmented and a lot of characters were lost. 537 randomly selected words were tested, and it was found that only 79.6% of test corpus was correctly segmented. Based on the global binarization of an image at various threshold levels, Khan *et al*. presented independent algorithm for segment sub-words in Arabic words[40]. The presented algorithm was tested on 537 randomly selected words from the AHTID/MW database and showed that 95.3% of the sub-words were correctly segmented and extracted. The presented method has shown considerable improvement over the projection profile method which was commonly used to segment sub-words or PAWs.

Table 2. Proposed Methods in Previous References

Ref No.	Proposed method	Data-set Size	Basic unit segmentation accuracy
[40]	Used multiple threshold levels to segmented Arabic words into PAWs	tested on 537 randomly selected words from the AHTID/MW database	95.3% of the sub-words
[42]	Developed a str*etc*hing algorithm which is able to reduce the overlap between connected parts and improved the overall results.	number of words in the dataset is about 2000	Words: 99.7%, PAWs: 93.16%
[45]	Vertical projection was app for word and sub-word segmentation.	Tested on IFN/ENIT database	Not Given-NG
[46]	Euclidean distance technique and DTW were used for extracting the connected components from the documents	A degraded dataset from Juma Al Majid Center (Dubai), the dataset contains 85	NG

		images(about 160 document pages) , set of 20 pages from the dataset is used in the experiments	
[50]	Proposed an algorithm which was relied heavily on the horizontal and vertical projection method in breaking up words into sub-words and characters	tested using 800 handwritten Arabic words taken from the IFN/ENIT database	NG
[51]	Word extraction is based on an adaptation of gap metrics and clustering algorithm to identify segmentation thresholds as "within word" or "between words" gaps	NG	84.8% correct word extraction

In Arabic basic unit segmentation, is projection profile analysis is commonly used method. But the boundaries between the basic units in handwritten Arabic documents are not very obvious. So a lot of basic units are missed the process of segmentation. Therefore, this method often used pre-segmentation stage. Used multiple threshold and various clustering technique to identify segmentation thresholds which were obtained pre-segmentation stage within word and between words gaps is an ideal segmentation method for the handwritten Arabic basic unit segmentation. Has been used clustering algorithms including Distance Based such as k-means and fuzzy c-means clustering (FCM) Probability Based such as Gaussian mixture model (GMM),Density Based such as density based spatial clustering of applications with noise algorithm (DBSCAN) *etc*. Str*etc*hing algorithm is also effectively reducing overlapped connected parts.

4. About the Handwritten Uyghur Texts and Comparisons with English and Arabic Languages

4.1. The Writing Characteristics of Uyghur

Uyghur is attributed to the Turkish family of Altaic language system. During the long history of evolution and interchange with other people, Uyghur has absorbed many kinds of lexicons from different origins. The modern Uyghur script is an alphabetic script which is based on Arabic and Farsi characters.

Due to the randomness of handwriting, situations including adhesion, overlapping, word spacing and character spacing irregularity are also common in Uyghur Offline handwritten recognition. Figure 5 shows an example of Uyghur handwritten text.

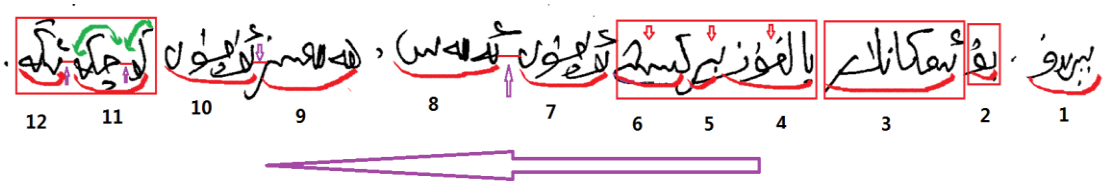


Figure 5. Uyghur Handwritten Text

There are twelve Uyghur words in the picture and we can find following characteristics from this example of handwritten text. It is noticed that we observe the illustration from the point of segmentation purpose.

(1) There is a certain distance between Uyghur words, but because of the randomness of writing, these distances irregularly change with the writing styles of the writers. In Figure 5, the distances between the twelve words can be seen.

(2) The width of each word is not necessarily the same, and the differences among them are obvious. In Figure 5, the width of the word 2 is much greater than of word 3.

(3) There are the situations of overlapping and adhesion between words. There are overlapping and adhesion between words 4 and 5 and overlapping between the words 5 and 6 in Figure 5.

(4) A word is composed of one or more connected segments. In Figure 5, there are three connected segments in word10 and two connected segments in word 11.

(5) There is also a certain distance between the connected segments. In handwriting, this distance is sometimes greater than the distance between the words, and sometimes smaller than the distance between words. In Figure 5, the distance between two connected segments' in word 11 is larger than the distance between the words of 9 and 10, but smaller than the distance between the words of 7 and 8.

(6) A baseline somewhat can be found from the handwritten text in Fig 5. Each word and the baseline for the whole text easily appear inclining phenomenon.

4.2. Comparison

Table 3. Comparison of Uyghur, Arabic and English Text Structure

Attributes Script	Uyghur	Arabic	English
Script kind	Alphabetic	Alphabetic	Alphabetic
Script kind& characters	with 32 characters including 8 vowels and 24 consonants	28 characters [42]	26 characters
Writing direction	From right to left, From top to bottom	From right to left, From top to bottom	From left to right, From top to bottom
Writing forms	Most Characters have 2 or 4 forms, such as isolated, beginning, intermediate and ending forms. Only two vowels have eight writing forms	Most Characters have 2 or 4 forms, such as isolated, beginning, intermediate and ending forms	All characters have two writing forms, such as upper-case and lower-case
Baseline	Text is written in accordance with baseline	Text is written in accordance with the baseline	Text is written in accordance with the baseline
Additional parts	There are 20 characters with one or more additional parts, whose positions are distributed in the upper, middle and lower part of the main body.	There are 16 characters with one or more additional parts, whose positions are distributed in the upper, middle and lower part of the main body.	There are two characters with additional parts, whose positions are appeared only over the character body

Indeed, there are many similar structures between Arabic and Uyghur scripts, and the difference between Uyghur and English scripts structure is relatively large. Therefore, considering segmentation from the structure of the text, English

segmentation methods will probably be less efficient for Uyghur text segmentation. On the contrary, the Arabic handwritten basic unit segmentation methods may be more suitable for Uyghur handwritten basic unit segmentation due to the structural similarities. Anyway, the methods are not necessarily compatible for Uyghur texts, because there are many differences in linguistics such as morphology, syntax and semantics.

5. Summarize and Suggestion

It is acknowledged that languages in the world are different, of course, holding some similarities among them. Scripts for languages are also similar or different from each other. In this paper, we studied Arabic script, which is similar to Uygur, and English, which is different one. Through a brief review of basic unit segmentation methods for English and Arabic handwritten texts, we can conclude that studying the structural and writing characteristics of handwritten text is the starting point of adopting appropriate methods. We can learn from the ideas and methods English and Arabic, and further promote the development of Uyghur handwritten segmentation.

The initial segmentation in Arabic and English handwritten texts is carried out based on the blank space/distance, and by using different distance metrics. The segmented units are obtained using clustering methods appropriate for the unique characteristics of the particular script kind. Of course, recognition results can be returned as a feedback, and neural network approach is used for further training. Similar segmentation approaches may be applied to handwritten Uyghur text segmentation. However, because of the difference between languages, the reviewed methods have no guarantee of full suitability for Uyghur text segmentation. Further and special developed methods for Uyghur handwritten text segmentation are mandatory. Considering the current research situation of the Uyghur handwritten text document segmentation, we put forward some suggestions for further studies as follows:

First, when collecting samples, a close attention to sample origin must be paid in order to ensure the authenticity of the results and convenience for comparative experiments. This involves the collecting handwriting texts from the sample providers with different educational background and age. On this basis, a substantial database for research of Uyghur handwriting images should be established. Because the experimental results are comparable only using the same standard and condition so that qualified and sufficient volume of database helps the development of Uyghur handwriting research go forward in quicker steps.

Secondly, basic units of segmentation lie in the handwritten text line, therefore, the accuracy of text line segmentation has a great impact on the next steps. Preprocessing techniques for adjusting slant of handwriting, adhesion of text lines and mixing between them are need for further perfection.

Thirdly, we believe that due to the same family of languages or scripts obviously have similar structural characteristics. Whether in character segmentation or word segmentation, the methods used in segmentation are closely related to the writing and structural features of the text.

Finally, in segmentation step, although the structure and writing characteristics of the handwritten text is an important study point, we should not point the whole direction to the hard conditions. We can consider the other elements such as the semantic link between basic units, too.

When the segmentation object is an individual character, perhaps there will be little help from English character segmentation to the study of Uyghur character segmentation. Because of English and Uyghur characters can be said to be

completely different in morphological structure. In addition, the linguistic properties of English and Uyghur are not close to each other. So, successful character segmentation approach for English is usually defeated to show good performance for Uyghur handwritten character segmentation. However, the similarities between Arabic and Uyghur make Arabic character segmentation references useful for Uyghur character segmentation.

If the connected segments are taken as the basic unit for segmentation, methods used for the Arabic text such as the baseline characteristics and the blank distance between the connected segments are also suitable for Uyghur handwritten text segmentation. The overlap between Arabic words is more serious and the boundaries between words are not obvious. This phenomenon is very much same in Uyghur handwriting, too. So, the segmentation of connected components is very helpful to extract the whole word in the end. Because of words can be concatenated by other connected components.

When a word is regarded as the basic unit, word segmentation techniques from English are preferred to adopt. Because the widths of Uyghur words are as inconsistent as the width of English words that there are different distances between the letters in each word. There is baseline in Uyghur and in English handwriting, too. As for Arabic, we cannot directly make judgment for the suitability of the Arabic word segmentation methods for Uyghur word segmentation. Because on one hand, Arabic and Uyghur are very much similar in basic characters, but on the other hand, Arabic and Uyghur are far from each other in morphology and syntax.

6. Conclusions

This paper briefly reviewed the relevant research in the basic unit segmentation of handwritten Arabic and English handwritten text recognition. The writing characteristics of Uyghur and the comparison with Arabic and English are discussed. The structural characteristics and difficulties encountered in basic unit segmentation for different scripts are analyzed. Finally, based on the results of these comparisons, we put forward own views and suggestions for further study of Uyghur handwritten text document segmentation. We tried to put what aspects we should start with and what kind of method is more reasonable and effective for these encountered problems. We believe that the suggestions will be important research directions and content for off-line handwritten Uyghur text recognition.

Acknowledgments

This work has been supported by the National Natural Science Foundation of China under grant of 61462080.

References

[1] J. Jian-ming, D. Xiao-qing, P. Liang-rui and W. Hua, “Printed Uyghur Texts Segmentation”, Journal of Chinese Information Processing, **(2005)**, pp. 76-83.
[2] H. Mehmed, “Contour Based Uyghur Character Segment”, The National Language Committee Chinese information society information, **(2007)**.
[3] L. Xiao, Y. Bao-she, C. Qing, R. Zong-yu, Z. Jian-hua, “A Segmentation of Printed Ughur Character Based On Projection Histogram of Pixels”, Computer Technology and Development, **(2012)**, pp. 41-44, 49.
[4] W. Jin-e, Y. Bao-she, L. Xiao, G. C. Mirsali, “An Improved Projection Segmentation Method of Print Uyghur”, Computer Engineering, **(2013)**, pp. 263-266, 271.
[5] L. Ya-nan, C. Xing-wen and Z. Dan, “Improved Segmentation Method of Printed Uyghur Based On Pixels Integral Projection and Connected Domain Search Method”, Journal of Dalian Nationalities University, **(2014)**, pp. 315-318.
[6] Z. Lan, Y. Bao-she and YU Wei, “Segmentation Method of Printed Uygher Based on Drop Fall Algorithm”, Computer Technology and Development, **(2015)**, pp. 107-110, 115.

[7] G. Seni and E. Cohen, "External word segmentation of off-line handwritten text lines", Pattern Recognition, vol. 27, **(1994)**, pp. 41–52
[8] U. Mahadevan and R. C. Nagabushnam, "Gap metrics for word separation in handwritten lines", Document Analysis and Recognition, 1995, Proceedings of the Third International Conference on, Montreal, Que, vol. 1, **(1995)**, pp. 124-127.
[9] D. Kazakov and S. Manandhar, "A hybrid approach to word segmentation", Proceedings of the 8th International Conference on Inductive Logic Programming, vol. 1446, Springer-Verlag, **(1998),** pp. 125–134.
[10] G. Kim and V. Govindaraju, "Handwritten phrase recognition as applied to street name images", Pattern Recognition, vol. 31, **(1998)**, pp. 41-51.
[11] R. Manmatha and N. Srimal, "Scale space technique for word segmentation in handwritten documents", Scale-Space Theories in Computer Vision, **(1999),** pp. 22–33.
[12] J. Park, V. Govindaraju and S. N. Srihari, "Efficient word segmentation driven by unconstrained handwritten phrase recognition", Document Analysis and Recognition, 1999, ICDAR '99, Proceedings of the Fifth International Conference on, **(1999)**, pp. 605-608.
[13] U. V. Marti and H. Bunke, "Text line segmentation and word recognition in a system for general writer independent handwriting recognition", Document Analysis and Recognition, 2001, Proceedings, Sixth International Conference, **(2001)**, pp. 159-163.
[14] S. H. Kim, C. B. Jeong, H. K. Kwag and C. Y. Suen, "Word segmentation of printed text lines based on gap clustering and special symbol detection", Pattern Recognition, 2002, Proceedings 16th International Conference, vol. 2, **(2002)**, pp. 320-323.
[15] M. Feldbach and K. D. Tonnies, "Word segmentation of handwritten dates in historical documents by combining semantic a-priori-knowledge with local features", Document Analysis and Recognition, 2003, Proceedings, Seventh International Conference, vol. 1, **(2003)**, pp. 333-337.
[16] T. Varga and H. Bunke, "Tree structure for word extraction from handwritten text lines", Document Analysis and Recognition, 2005, Proceedings, Eighth International Conference, vol. 1, **(2005)**, pp. 352-356.
[17] I. Nwogu and G. Kim, "Word separation of unconstrained handwritten text lines in PCR forms", Eighth International Conference on Document Analysis and Recognition (ICDAR'05),vol. 2, **(2005)**, pp. 715-719 .
[18] R. Manmatha and J. L. Rothfeder, "A scale space approach for automatically segmenting words from historical handwritten documents", Pattern Analysis and Machine Intelligence, IEEE Transactions, vol. 27, no. 8, **(2005)**, pp. 1212-1225.
[19] T. Stafylakis, V. Papavassiliou, V. Katsouros and G. Carayannis, "Robust text-line and word segmentation for handwritten documents images", 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, **(2008)**, pp. 3393-3396.
[20] C. Huang and S. Srihari, "Word segmentation of off-line handwritten documents", Proceedings Document Recognition and Retrieval (DRR) XV, San Jose, CA, **(2008)**.
[21] G. Louloudis, B. Gatos, I. Pratikakis and C. Halatsis, "Line and Word Segmentation of Handwritten Documents", 1st International Conference on Frontiers in Handwriting Recognition (ICFHR), Montreal, Canada, pp. 247-252.
[22] G. Louloudis, B. Gatos, I. Pratikakis and C. Halatsis, "Text Line and Word Segmentation of Handwritten Documents", to appear on Pattern Recognition Journal, special issue on Handwriting recognition, **(2008)**.
[23] B. Gatos, N. Stamatopoulos and G. Louloudis, "ICDAR 2009 Handwriting Segmentation Contest", Document Analysis and Recognition, 2009, ICDAR '09, 10th International Conference, **(2009)**, pp. 1393-1397.
[24] G. Louloudis, N. Stamatopoulos and B. Gatos, "A Novel Two Stage Evaluation Methodology for Word Segmentation Techniques", Document Analysis and Recognition, 2009, ICDAR '09, 10th International Conference, **(2009)**, pp. 686-690.
[25] X. Du, W. Pan and T. D. Bui, "Text line segmentation in handwritten documents using Mumford–Shah model", Pattern Recognition, **(2009)**.
[26] V. Papavassiliou, T. Stafylakis, V. Katsouros and G. Carayannis, "Handwritten document image segmentation into text lines and words", Pattern Recognition, vol. 43, **(2010)**, pp. 369-377.
[27] B. Gatos, N. Stamatopoulos and G. Louloudis, "ICFHR 2010 Handwriting Segmentation Contest", Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference, **(2010)**, pp. 737-742.
[28] F. Simistira, V. Papavassiliou, T. Stafylakis and V. Katsouros, "Enhancing Handwritten Word Segmentation by Employing Local Spatial Features", Document Analysis and Recognition (ICDAR), 2011 International Conference, **(2011)**, pp.1314-1318.
[29] A. L. Kesidis, B. Gatos, "Efficient Cut-Off Threshold Estimation for Word Spotting Applications", Document Analysis and Recognition (ICDAR), 2011 International Conference, **(2011),** pp. 279-283.
[30] S. Wshah, G. Kumar and V. Govindaraju, "Script Independent Word Spotting in Offline Handwritten Documents Based on Hidden Markov Models", Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference, **(2012)**, pp. 14-19.
[31] N. Stamatopoulos, B. Gatos, G. Louloudis, U. Pal and A. Alaei, "ICDAR 2013 Handwriting

Segmentation Contest", Document Analysis and Recognition (ICDAR), 2013 12th International Conference, **(2013),** pp. 1402-1406.
[32] J. Ryu, H. I. Koo and N. I. Cho, "Word Segmentation Method for Handwritten Documents based on Structured Learning", Signal Processing Letters, IEEE, vol. 22, no. 8, **(2015),** pp. 1161-1165.
[33] A. Ghorbel, J. M. Ogier and N. Vincent, "A segmentation free Word Spotting for handwritten documents", Document Analysis and Recognition (ICDAR), 2015 13th International Conference, **(2015)**, pp. 346-350.
[34] E. Kavallieratou, "Word segmentation using Wigner-Ville distribution", in Document Analysis and Recognition (ICDAR), 2015 13th International Conference, **(2015)**, pp. 701-705.
[35] L. Y. Tseng and C. T. Chuang, "An efficient knowledge-based stroke extraction method for multi-front Chinese characters", Pattern recognition, **(1992)**, pp. 1445-1458.
[36] C. C. Chiang, T. Cheng and S. S. Yu, "An Iterative Rule-Based Character Segmentation Method for Chinese Documents", International Conference on Chinese Computing'96 The Latest Technological Advancement & Applications, **(1996)**, pp. 301-307.
[37] M. Vyas and K. Verma, "A comprehensive survey of handwritten character segmentation", Advanced Communication Control and Computing Technologies (ICACCCT), 2014 International Conference, Ramanathapuram, **(2014)**, pp. 1462-1465.
[38] R. G. Casey and E. Lecolinet, "Strategies in character segmentation: a survey", Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on, Montreal, Que, vol. 2, **(1995)**, pp. 1028-1033.
[39] R. G. Casey and E. Lecolinet, "A survey of methods and strategies in character segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 18, no. 7, **(1996),** pp. 690-706.
[40] F. Khan, A. Bouridane, F. Khelifi, R. Almotaeryi and S. Almaadeed, "Efficient segmentation of sub-words within handwritten arabic words", Control, Decision and Information Technologies (CoDIT), 2014 International Conference on, Metz, **(2014)**, pp. 684-689.
[41] M. Khayyat, L. Lam and C. Y. Suen, "Arabic handwritten word spotting using language models" Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference, Bari, **(2012)**, pp. 43-48.
[42] Z. Al Aghbari and S. Brook, "Word str*etc*hing for effective segmentation and classification of historical Arabic handwritten documents", Research Challenges in Information Science, 2009. RCIS 2009. Third International Conference, Fez, **(2009)**, pp. 217-224.
[43] J. H. A. Khateeb, J. Ren, J. Jiang and S. S. Ipson, "Unconstrained Arabic Handwritten Word Feature Extraction: A Comparative Study", Information Technology: New Generations, 2009. ITNG '09. Sixth International Conference, Las Vegas, NV, **(2009)**, pp. 1655-1656.
[44] L. M. Lorigo and V. Govindaraju, "Offline Arabic handwriting recognition: a survey", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 5, **(2006),** pp. 712-724.
[45] Y. Osman, "Segmentation algorithm for Arabic handwritten text based on contour analysis", Computing, Electrical and Electronics Engineering (ICCEEE), 2013 International Conference, Khartoum, **(2013)**, pp. 447-452.
[46] R. F. Moghaddam and M. Cheriet, "Application of Multi-Level Classifiers and Clustering for Automatic Word Spotting in Historical Document Images", Document Analysis and Recognition, 2009, ICDAR '09, 10th International Conference, Barcelona, **(2009)**, pp. 511-515.
[47] R. Saabni and J. El-Sana, "Keyword searching for Arabic handwritten documents", Proceedings 11th Int. Conf. on Frontiers in Handwriting Recognition, (ICFHR), **(2008)**, pp. 271–277.
[48] T. Sari and A. Kefali, "A search engine for Arabic documents", In Actes du dixieme Colloque Int. Francophone ´ sur l´Ecrit et le Document ´ , pp. 97–102, **(2008)**.
[49] V. Margner and H. Abed, "Guide to OCR for Arabic Scripts", Springer, **(2012)**.
[50] A. Lawgali, A. Bouridane, M. Angelova and Z. Ghassemlooy, "Automatic segmentation for arabic characters in handwriting documents", in Image Processing (ICIP), 2011 18th IEEE International Conference, **(2011)**, pp. 3529–3532.
[51] A. Al-Dmour and F. Fares, "Segmenting Arabic Handwritten Documents into Text lines and Words", International Journal of Advancements in Computing Technology, **(2014)**, pp. 63.

Authors

Aysadet Abliz, she has received her B.E. degree in Electronics from Xinjiang University, China, in 2014. Currently, she is a M.S Student in Signal & Information processing in Xinjiang University. Her research interest is pattern recognition and image information processing.

Wujiahemaiti Simayi, he has received his B.E. and M.S. degree in Electronics, Signal & Information processing from Xinjiang University, China, in 2009 and 2014, respectively. Currently, he is a PhD candidate in Computer Applications in Xinjiang University, and working as research assistant at the Key Laboratory of Intelligent Information Processing, Xinjiang University, China. His research interests include feature extraction and classification techniques, handwritten character recognition, computer simulation for atmospheric pollution dispersion modeling.

Kamil Moydin, he received his B.E. and M.S. degree in radio electronics and computer science from Xinjiang University, China, and Osaka Institute of Technology, Japan in 1983 and 1998, respectively. He has been working as a teacher in School of Information Science and Engineering, Xinjiang University since 1983. He was a visiting scholar in the Osaka Institute of Technology, Japan from 1994 to 1996. In 2002, he got the position of associate professor in Xinjiang University. His research interests include computer network, pattern recognition, and digital image processing.

Askar Hamdulla, he received B.E. in 1996, M.E. in 1999, and Ph.D. in 2003, all in Information Science and Engineering, from University of Electronic Science and Technology of China. In 2010, he was a visiting scholar at Center for Signal and Image Processing, Georgia Institute of Technology, GA, USA. Currently, he is a professor in the School of Software Engineering, Xinjiang University. He has published more than 160 technical papers on speech synthesis, natural language processing and image processing. He is a senior member of CCF and an affiliate member of IEEE.