Print ISSN: 2321-6379
Online ISSN: 2321-595X
DOI: 10.17354/ijssl/2017/90

Original Article

Improvement in Persian Text Words Segmentation Using Morphology Operations

Mohammad Sadeq Navabi, Erfan Golzar, Seyyed Mohammad Razavi

Department of Electrical and Computer Engineering, University of Birjand, Birjand, Iran

Abstract

Text word segmentation is one of the most important issues in the OCR systems. Research in this area has been started for several decades with good results achieved in the field of English words. However, satisfactory results have not been achieved for Persian and Arabic because there are points, vertical overlapped broken and continuous lines in words and research is still under way in this regard. The application of these operations causes improved performance of OCR systems and the increased accuracy of word recognition in the text. The current study used morphology operations in text words segmentation. The paper aims to find span of each word and words segmentation by putting them in separate boxes to increase the accuracy of Persian words segmentation and reduce error due to the segmentation of subwords and letters instead of the words. Dilation operation is used in this article. The system test was applied using three typed text images including 713 words. The recognition rates of 96 percent, 94.6 percent, and 95.27 percent were obtained, respectively. The general reason behind the emergence of errors in all three tests is non-compliance in typing space between words.

Keywords: Segmentation, Morphology operations, Persian text, OCR, Improving the recognition, Subword and letter

INTRODUCTION

The emergence of new science and technology has made human society face with different forms of information. The level of development of a society can be evaluated by the produced amount of information and knowledge. Information is increasingly produced in various forms with varying degrees of complexity. Therefore, there is an increased need for information processing systems [1]. The application of OCR systems did not lead to satisfactory results for Persian and Arabic languages because there are points, vertical overlapped broken and continuous lines in words. The current paper, using morphology operations, studied the accuracy of segmentation in Persian and Arabic languages.

OCR term refers to techniques to recognize text in scanned images and converts them to editable areas. In recognition of offline text the systems input is the scanned image of the text, but in recognition of online text the systems input is the pen trajectory points. In this case, a man's relation with the computer is usually a pen and a digitizing tablet. Figure 1 shows the input method in both modes [2].

Persian alphabet recognition is not a long history. First published official reports of efforts done in this regard date back to the early years of the 1980s [3]. There have been few studies on the recognition of the Persian alphabet despite the relative learning use of alphabet among different nations of Asia. Because there are fundamental differences between the wording of Persian words and Latin words, such as a attached word and the change in the relative form in different position of the letters in a word of Persian, it is not possible to directly apply the conventional methods for recognizing English letters on Persian words.

Literature Review

Zand [4] proposed an algorithm to separate letters based on the Artificial Neural Network (ANN) classification method [5]. He also described the horizontal and vertical histogram method for separating text lines [6]. In this method if the difference between the two black rows pixels is greater than the specified threshold, a new line of text can be found.

Access this article online

www.ijss-sn.com

Month of Submission : 00-0000
Month of Peer Review : 00-0000
Month of Acceptance : 00-0000
Month of Publishing : 00-0000

*Corresponding Author: Mohammad Sadeq Navabi, Department of Electrical and Computer Engineering, University of Birjand, Birjand, Iran. E- mail: mnavabi@birjand.ac.ir

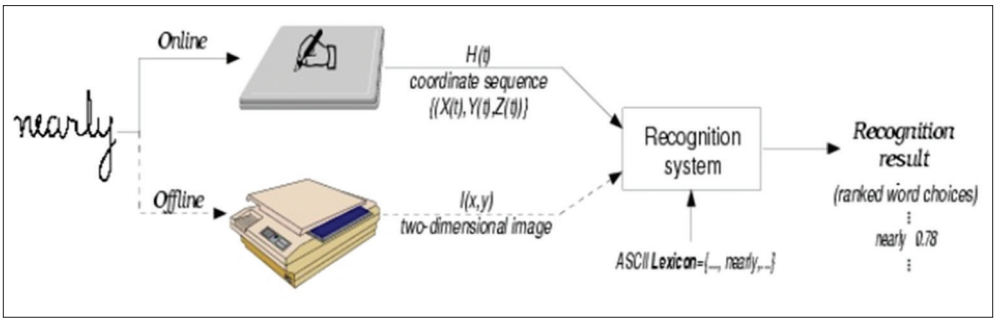


Figure 1: Text offline and online recognition [3]

Frank [7] described the horizontal and vertical histogram in English Handwritten Character Recognition.

Shahraki [8], introduced a method to separate text lines using morphology operations. The biggest problems in Persian lines of text separation using morphology operations are letters sticking to each other through the skewed parts of two letters "ک", and "گ", the letters point's noise. So elements with these features are ignored.

Jelodar [9] designed an algorithm that does not lose its efficiency by changing the font, also suitable for digits. Morphology operations are very powerful and their only disadvantage is sensitivity to the noise. To reduce the sensitivity, the skeleton image is used instead of thinned image. AlKhateeb [10] proposes a new way to find baseline in handwritten text using horizontal histogram. The result of this technique was tested on 200 images with 85% correct recognition and 15% incorrect recognition due to the scriber change in spaces between words and subwords.

Al-Dmour [11] used horizontal histogram to separate handwritten Arabic text to line and word. Clustering and threshold methods have been used to separate the word from the text by calculating the interval s between the words. Mamatha [12] studied the differences between Kannada, Telugu, Assamese in the words of Hindi language and the reasons for continuing the investigation in the field of OCR and refers to some features of Hindi language compared to other languages in the world, including a lot of overlap among these reasons and proposes morphology operation and using the histogram in segmentation of the word and subword of the text. The result led to the segmentation of lines of text with 94% accuracy, with the results of words and letters segmentation of 82% and 73% accuracy, respectively. Karmakar[13] proposed a way to separate lines of text in Hindi language. In this method, the white pixels can be found in each row of the input image. The threshold for the number of Consecutive Whiterow (CWR) is calculated as the interval between two lines. The current paper used ten different images as input with the result of 100% accuracy of line recognition as well as100% accuracy of word from line recognition.

Alipour [14] described the general trend of OCR process. In this article, the process includes binary phase, filtering, smoothing, and thinning. To test this, a 100 typed page Persian text with 10 different fonts and sizes were used. Segmentation was done with the accuracy of 98%. The point is that segmentation accuracy increases by increasing the font size. Fiirojj [15] presented a new algorithm for segmentation of Arabic words using bounding box regions for Arabic letters; finally, the interval between the bounding box regions for Arabic letters to fill bounding box regions for Arabic words. This method ignored segmentation of lines that is done in most methods of segmentation. Mariam et al., 2016, [16], in their article as spelling correction model for OCR errors in Arabic, investigated to show that the Multi-character model is better than the single character one, and showed that using the classification and alignment of characters as well as their model the accuracy of 94% in recognizing the correct spelling of 502167 words.

METHOD

The design of an automated segmentation of words of text in rectangular boxes with maximum precision and speed is the main objectives of this article. To achieve this, the design of appropriate algorithms to recognize the font size and perform some pre-processing schemes will be put on the agenda. K-mean clustering method and analytical methods for image processing is proposed to recognize the font size. The following perspectives are proposed for the purpose of improving the accuracy and speed of words segmentation.

In this view, the aim is maximizing segmentation of the word precision.

This method should cause the minimum error in a large database. Moreover, the extracted features must be

independent of changes in scale and font size. In this view, the focus will be on increasing speed, and in addition to maintaining the highest segmentation precision, speed is the maximum. Human visual system is the best pattern in this field as well as similar fields of image processing, the system with the best resolution and the highest speed. The interval of horizontal space between neighboring bounding box in Arabic words is measured as the X axis and is called F Interval. The interval between two Arabic words is the interval between the last letter of the first word with the first letter of the second word.The interval between the Arabic word is more than the interval between the Arabic letters. The current paper aims to identify the interval between Arabic words and the interval between the Arabic letters and bridge the gap between letters without words filling the spaces between the words. F interval can have different values. Recognize different parts of the Arabic image are complicated, there is a way to solve this problem that is to transform grayscale image to binary image so that each pixel is limited the range of one and zero. Morphological analysis is a method for processing Arabic images based on form [18, 17]. Morphological operators convert the original Arabic image into another Arabic image through interactions with other Arabic image that have a fixed size (which this known as structuring element). Morphological operators can simplify information, keep the characters forms and ignore irrelevant information. Morphological operators are used for different purposes among them are edge detection, segmentation, enhance photos and graphics computations. Two main morphological operators are dilation and erosion. In dilation, the objects are expanded and are used to fill small holes. In contrast, in erosion objects become small. The operators can be improved through the accurate calculation and selection of appropriate structural characteristics to accurately estimate the amount of dilation and or erosion of objects. The process of dilation is conducted by the structural characteristics B located on the Arabic image A and slip it over the Arabic image like the case of a convolution.

If the origin of the structuring element corresponds to white pixel in the Arabic image, there will be no motion to modify next pixel. If the origin of the structuring element corresponds to black pixels the entire image region by the structural characteristics is covered in black. In general, expansion leads to the spread image, and as a result fills small holes in the background. If the unidirectional dilation is horizontal (left to right), the suitable dilation width can be controlled and this can be achieved only through distinctive single row.

The proposed structural characteristics of dilation are expressed in Equation 2:

P=1,2,3,4,5…etc. Thus
For P=1, ME1=[0 1 1]
For P=2, ME2=[0 0 1 1 1]
For P=3, ME3=[0 0 0 1 1 1 1]
For P=4, ME4=[0 0 0 0 1 1 1 1 1]
For P=5, ME5=[0 0 0 0 0 1 1 1 1 1 1]

Equation 1. The used structure features

The above equation, represents a P zero and represents a P + 1. Structural characteristic matrix is of 1 × (2P + 1).

Bounding box sticking during dilation occurs by counting the bounding boxes before and after dilation. Sticking occur when an F interval is filled with reduced number of bounding boxes of the unit.The proposed algorithm application gradually integrates bounding boxes and the final Arabic image is a series of horizontal segments with no F interval. Figure 2 and Figure 3 depict bounding boxes of 7 Arabic script lines before and after dilation using the proposed algorithm, respectively.

The left to right dilation of the target image causes the integration of all between the letters intervals while the intervals between the words remain intact.

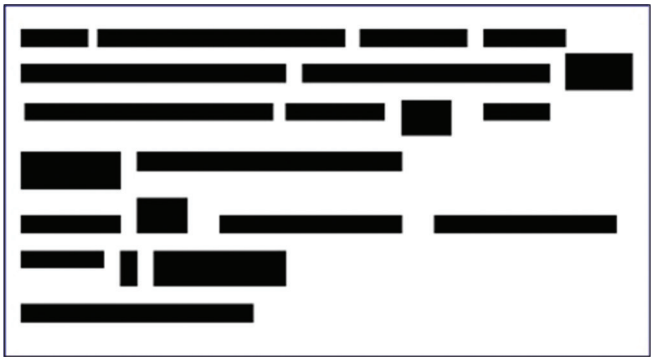


Figure 2. Bounding boxes of 7 Arabic script lines before dilation

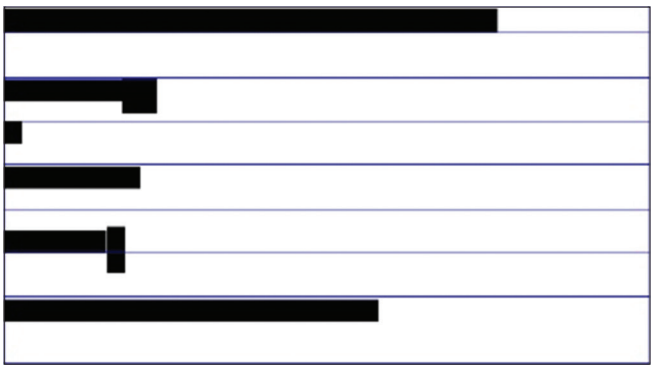


Figure 3. Bounding boxes of 7 Arabic script lines after full F dilation

The algorithm implementation

- First, the primary image is entered into the algorithm and the RGB image is converted into GRAY. Then the image edge is detected with an appropriate threshold using canny method.
- Second, using Radon command, the image is exposed to a 90 ° angle ray to detect lines. Then the returned values are plotted by the vector R. Horizontal histogram obtained from radiation using the radon can be seen in Figure 4. The histogram clarifies available lines in text.
- Third, to find the font size is important. First, the horizontal histogram ascenders should be obtained and then calculated the average interval between the ascenders. The average is the relative font size. This stage ascenders are depicted in Figure 7. The original entered text constitutes 18 lines, with one line the page number and the other line the Persian date "Esfand 1345". The first 7 lines are with a fixed interval, and then appears a gap and then 10 lines with the same interval and then the page number.

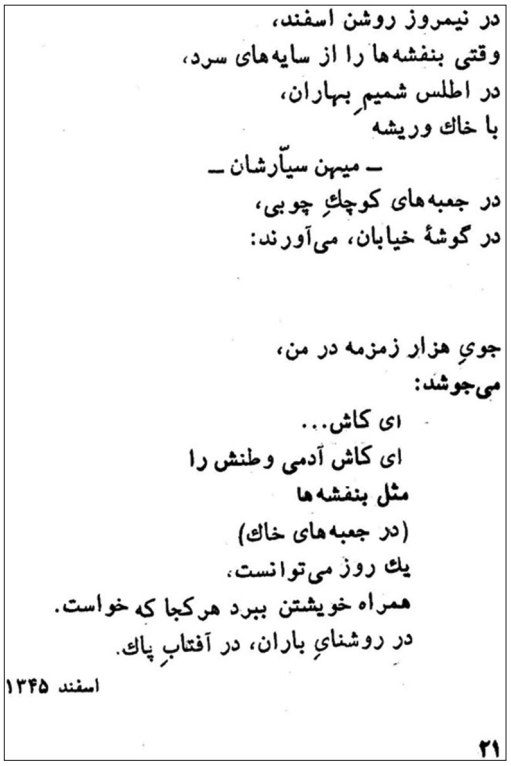
A total of 18 lines and the gap between a number of first lines and the second line segments are marked in the histogram in Figure 6.

Figure 7 shows that there are points in histogram ascenders that we do not care for considered relative peaks. These points are known as error.

Figure 4. The entered image after radon ray exposure to detect lines

These errors generally occur at low points and below 3×. You should always be looking for the absolute peaks in the histogram. Each of the red peaks, indicating a line and precise diagnosis of the red peaks have a direct impact on the quality and accuracy of the algorithm. K-mean method is used to resolve these errors.

Fourth, we look for the lines intervals as a citation to detect font size. In Figure 7, if the difference between the absolute peak and an absolute minimum is greater than 3×, the algorithm enters into the K-mean clustering. The red points are clustered in two clusters. The larger average Y coordinates cluster considered as the main points and the average difference between X s in the cluster is considered as a measure to detect the font size. Cluster with the less


در نیمروز روشن اسفند،
وقتی بنفشه‌ها را از سایه‌های سرد،
در اطلس شمیمِ بهاران،
با خاك وریشه
ــ میهن سیّارشان ــ
در جمبه‌های کوچكِ چوبی،
در گوشهٔ خیابان، می‌آورند:

جویِ هزار زمزمه در من،
می‌جوشد:
ای کاش...
ای کاش آدمی وطنش را
مثل بنفشه‌ها
(در جمبه‌های خاك)
یك روز می‌توانست،
همراه خویشتن ببرد هرکجا که خواست.
در روشنایِ باران، در آفتابِ پاك.
اسفند ۱۳۴۵

۲۱

Figure 5. The original entered image

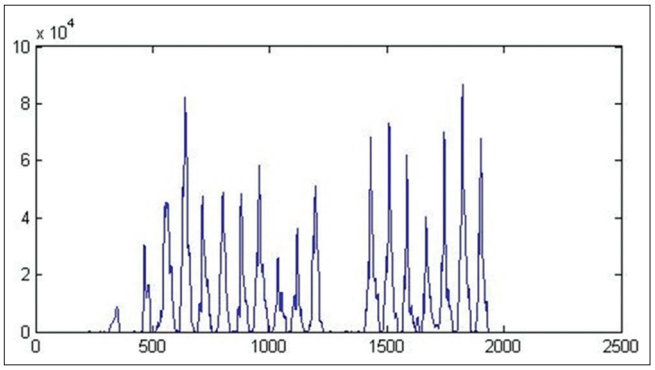


Figure 6. Entered image horizontal histogram

than the average Ys is considered error because most errors usually occur in low Y.

If the difference between the absolute peak and the absolute minimum is less than 3×, the algorithm does not enter into the clustering and in fact the algorithm does not recognize an error and the average difference between the total points available is as criterion for determining the font size.

In this figure, the red cluster is eliminated as error. Of course, all points in the red cluster are not error and a point at a interval of 0 to 20 indicates a page number in the original text (number 21) that is mistakenly considered error and eliminated due to being in red cluster, however, this deletion has no effect on the final result as a result of the large number of lines. Average blue points difference is considered a criterion to detect font size.

Fifth (final), here the number obtained in the fourth stage is rounded upward. The number obtained is a criterion for placement in the structure feature. Then, dilation operator implemented on the edge detected image, taken into account the first stage. Figure 10 shows the results obtained up to this point:

Next, the hole is filled and the number of stuck together points are counted and placed inside the bounding boxes. The end result of the algorithm is shown in Figure 9.

The Results of The Implementation of The Proposed Algorithm

In order to test the system, the proposed algorithm was tested on database contains 5000 words and an average of more than 98% accuracy was obtained as a result of utilizing the segmentation algorithm. Here we examine a few experiments and expression errors causes, if any.

First Test

Figure 12 is selected from the database containing 200 words. The algorithms text output and the end result of segmentation is shown in Fig. 11.

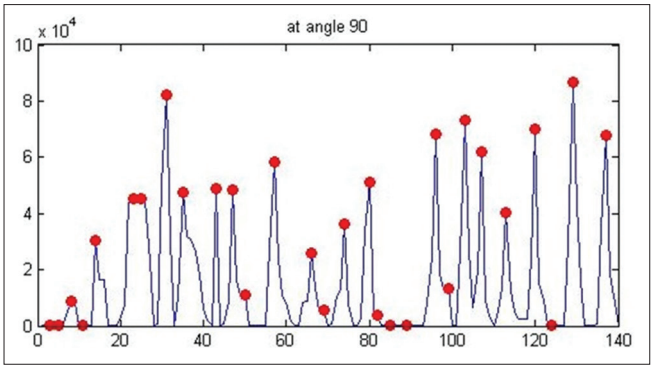


Figure 7. Detecting peaks in the histogram horizontal

There are 8 errors including:,"احت‌الحلقوم" "باگوشهایم" "می‌گفتم:خداحافظ", "می‌گفتم:سلام", "خل", "راحت‌الحلقوم:نوعی", "شیرینی از نشاسته"", "و شکر, with two main causes. The first are typing errors of not considering the words intervals including: "راحت‌الحلقوم", "باگوشهایم", "می‌گفتم:خداحافظ","خل "می‌گفتم:سلام"". The second are as the result of the changes in font size including: "راحت‌الحلقوم:نوعی", "شیرینی "از نشاسته", "و شکر".

In the end of each line, the punctuations such as question marks and dots together with the last word are considered

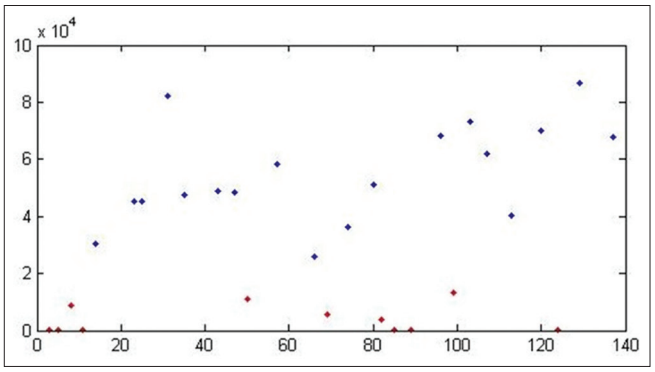


Figure 8. Absolute peaks segmentation of errors using K-mean

Figure 9. Word segmentation end results

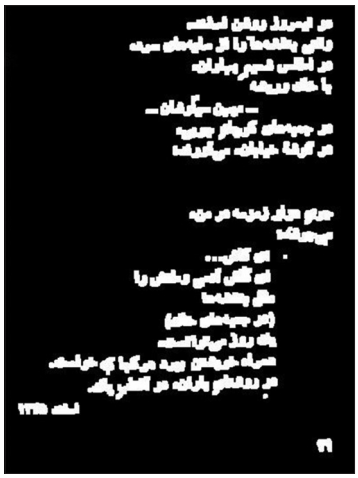


Figure 10. Dilation operator implemented on the edge detected image

Figure 11. Word segmentation end results of first test

Figure 12. Input text image

Figure 13. Input text image

as a part of the word that the error is due to typing error in not paying attention to spaces between the words and punctuations. This text segmentation quality is 96% and all errors result from non-compliance of the interval between words and signs.

Second Test

Figure 13 is selected from the database containing 280 words.

Figure 14 depicts the algorithm output and word segmentation end results.

There are 15 errors including: “”,“حرکت”, “مترقی ترین”, عبارتنداز”, “عاشوراو”,” فراموش شده”, “وعاشورا”,” ستم،و”, “ترک”,” قبرستان”, “وانتظار”,” حیات‌بخش”, “راخواهند”,” “ومانیزروح”, “وحیات”,” جهان‌بینی”, caused by typing errors of not considering the words intervals. In case of standard intervals between the words, the segmentation will be implemented with 100% accuracy. Here, the segmentation was implemented by 94.6% accuracy and all the errors caused by non-compliance of the interval between words.

Third Test

Figure 16 is selected from the database containing 233 words. Figure 15 depicts the algorithm output and word segmentation end results.

There are 11 errors including: "", "بی‌پرده گشته‌چون"، " ولیکن‌صحبت"، "دهدچون‌من‌زبان‌دست"، "مراخواهداجل‌دست"، "وزبان‌بست"، "مگرهوشنگ"، "کرده‌باشم اگراولاد"، "گرت‌ازعلم‌ودانش‌توشه‌ای‌هست"، "ازاین‌خرمن‌تراهم‌خوشه‌ای‌هست" caused by typing errors of not considering the words intervals. Here, the segmentation was implemented by 95.27% accuracy and all the errors caused by non-compliance of the interval between words.

CONCLUSION

The current paper proposes a simple method for segmentation of the word of the text. The algorithm is simple but enjoys high efficiency. The current paper used morphology operations with different amount of dilations per image based on the font size. The current study used a database contains 10,000 words. All images are scanned at 300 dpi quality. First the primary RGB image is converted into GRAY. Then the image is edge detected. Then, the image is exposed to ray to detect lines. Later on, the font size should be detected that is done by the average ascenders points in horizontal histogram. This interval is a criterion to regulate dilation of algorithm.

Three typed text image containing 713 words were used to test the proposed system.The first test examined a text image selected from the database containing 200 words with the system recognition of 96% accuracy. The errors are of two main causes, the first are typing errors of not considering the words intervals including: "", "راحت‌الحلقوم"، "باگوشهایم"، "می‌گفتم:خداحافظ"، "می‌گفتم:سلام"، "خل". The second are as the result of the changes in font size including: "و شکر"، "شیرینی از نشاسته"، "راحت‌الحلقوم:نوعی".

Figure 14. word segmentation end results in second test

Figure 15. Word segmentation end results of third test

Figure 16. Input text image

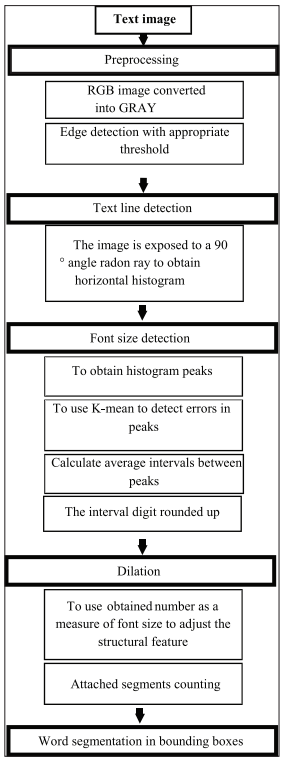


Figure 17: The general trend of words of text segmentation algorithm performance

The second test examined a text image selected from the database containing 280 words with the system recognition of 94.6% accuracy and the same errors as that of the first test.

The third test examined an image of relatively long poem selected from the database containing 233 words with the system recognition of 95.27% accuracy. All the errors caused by non-compliance of the interval between words in typing. In case of standard intervals between the words, the segmentation will be implemented with 100% accuracy.

Recommendations

Given the results, the following is recommended for further studies:

- All the errors caused by non-compliance of the interval between words and signs in typing. In case of standard intervals the errors are prevented.
- The database used in this study is 300 dpi scan quality, the higher scanning quality increases segmentation quality with less errors.
- There are sometimes digits, and signs in addition to letters and the word requiring paying attention to the intervals between the digits and words typed into the database.
- There are sometimes English or Arabic digits and letters in a text in addition to letters and words in Persian, therefore compliance with the intervals inserted in other languages in typing may contribute to more algorithm evolution.

There might be some overlapped, words attachments to each other or to signs, in low quality image scanning or lower than 300dpi scan because of the noise. The algorithm needs a part of the initial processing on the image in order to neutralize the negative impact of listed factors.

REFERENCES

Azmi Reza, "Recognition of printed texts in Persian," PhD thesis, Supervisor: Kabir Ehsanallah, Tarbiat Modares University, Faculty of Engineering, Tarbiat Modarres University, 1999.

Razavi Seyyed Mohammad, "Recognition of handwritten online Persian", PhD thesis, Supervisor: Kabir Ehsanallah, Faculty of Engineering, Tarbiat Modares University, 2006.

Timsari , Bijan, "Recognition of Persian typed words' letters using morphology", Supervisor: Hamid Fahimi, Faculty of Electrical Engineering, University of Technology, 1992.

Mohsen Zand, AhmadrezaNaghshNilchi, and S. AmirhassanMonadjemi," Recognition-based Segmentation in Persian Character Recognition".

R. J. Schalkol, "Pattern Recognition: Statistical, Structural and Neural Network", Wiley, New York, 1992.

B. Timsari, "Character recognition in typed Persian words", a morphological approach, M.S. thesis, Isfahan Univ. of Tech., Iran (1992).

Frank de Zeeuw "Slant Correction using Histograms" Bachelor's Thesis in Artificial Intelligence Supervised by Axel Brink &Tijn van der Zant, July 12, 2006.

Abdollah Amirkhani-Shahraki, Amir EbrahimiGhahnavieh, SeyyedAbdollahMirmahdavi" A Morphological Approach to Persian Handwritten Text Line Segmentation" 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation.

M. Salmani Jelodar, M.J. Fadaeieslam, N. Mozayani, M. Fazeli" A Persian OCR System using Morphological Operators" World Academy of Science, Engineering and Technology International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:1, No:4, 2007.

Jawad H AlKhateeb, Jianmin Jiang, JinchangRen, and Stan S Ipson" Component-based Segmentation of Words from Handwritten Arabic Text" World Academy of Science, Engineering and Technology International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:2, No:5, 2008.

Ayman Al-Dmour1 and Fares Fraij2" Segmenting Arabic Handwritten Documents into Text lines and Words" Segmenting Arabic Handwritten Documents into Text lines and Words Ayman Al-Dmour, Fares Fraij.

Mamatha H R, Srikantamurthy K" Morphological Operations and Projection Profiles based Segmentation of Handwritten Kannada Document" International Journal of Applied Information Systems (IJAIS) – ISSN: 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 4– No.5, October 2012 – www.ijais.org.

PriyankaKarmakar, BiswajitNayak, NilamaniBhoi" Line and Word Segmentation of a Printed Text Document" PriyankaKarmakar et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (1), 2014, 157-160.

Mir Mohammad Alipour ,Department of Computer Engineering, University of Bonab, Bonab 5551761167, Iran" A New Approach to Segmentation of Persian Cursive Script based on Adjustment the Fragments" International Journal of Computer Applications (0975 – 8887) Volume 64– No.11, February 2013.

FiirojjParwejj,,Ph..D ,Department of Computer Science ,Jazan University, Jazan, Kingdom of Saudi Arabia" A Perceptive Method for Arabic Word Segmentation using Bounding Boxes by Morphological Dilation" International Journal of Computer Applications (0975 – 8887) Volume 71– No.1, June 2013.

Luc Vincent, "Morphological grayscale reconstruction in image analysis:

Applications and efficient algorithms," IEEE Trans. on Image Processing, 2(2):176-201, 1993.

H. Hadwiger:"Vorlesungen über Inhalt, Oberfläche und Isoperimetrie", Springer Verlag, (1957).

Muhammad M, ELGhazaly T, Ezzat M, Gheith M. A Spell Correction Model for OCR Errors for Arabic Text. InInternational Conference on Advanced Intelligent Systems and Informatics 2016 Oct 24 (pp. 124-136). Springer International Publishing.

How to cite this article: Navabi MS, Golzar E, Razavi SM. Improvement in Persian Text Words Segmentation Using Morphology Operations. Int J Sci Stud 2017;5(5):653-661.

Source of Support: Nil, **Conflict of Interest:** None declared.