Arabic OCR Evaluation Tool

Mansoor A Alghamdi
Department of Computer Science
Community Collage
Tabuk University
Tabuk, Saudi Arabia
malghamdi@ut.edu.sa

Ibrahim S Alkhazi
College of Computers & Information Technology
Tabuk University
Tabuk, Saudi Arabia
i.alkhazi@ut.edu.sa

William J. Teahan
School of Computer Science
Bangor University
United Kingdom
w.j.teahan@bangor.ac.uk

***Abstract*—Performance evaluation of Optical Character Recognition (OCR) systems is an essential task for OCR systems development. However, studies in Arabic OCR suffer from the lack of proper performance evaluation metrics and the availability of evaluation tools. Although the literature provides typical performance metrics, such as character accuracy and word accuracy for OCR performance evaluation, these metrics are not sufficient for evaluating Arabic OCR. This paper presents an open source automated software tool with various metrics for the evaluation of Arabic OCR performance. The developed tool is available for OCR researchers, thus it can be applied for ranking different OCR algorithms.**

Keywords—performance evaluation; Arabic; OCR; Arabic OCR; evaluation metrics.

I. Introduction

Optical Character Recognition (OCR) is a process of converting a machine-printed or handwritten text image into a digital computer format that can be editable. OCR technology is considered as a challenging research area in the field of pattern recognition and artificial intelligence. Many studies have been proposed for Latin and non-Latin scripts. However, development of OCR applications for Latin script is easier than that of Arabic scripts because of Arabic writing characteristics [1]. Fig. 1 illustrates some characteristics of the printed Arabic script that contribute to the challenges in Arabic OCR evolution. Compared to printed Latin script, Arabic script is written cursively from right to left, and contains dots and diacritics. Also, a character of Arabic script may have four dissimilar shapes in relation to its location in an Arabic word.

Furthermore, when comparing the performance of Arabic OCR with the Latin script OCR, researchers call attention to the need for more research solutions for Arabic OCR [2], [3]. Consequently, evaluating the performance of OCR is required [4] in order to develop better OCR algorithms. In other words, having quantitative measures will result in monitoring progress in the OCR development field. Moreover, comparing the performance of OCR systems will contribute to providing a scientific explanation for the system performance, identifying open areas, and analysing the weakness and strengths of the OCR systems, as stated by [5] and [6].

In respect to performance evaluation of Arabic OCR systems, several researchers [7], [8], [9] report the accuracy performance of their systems in the absence of specifying the method of measuring the performance. Hence, researchers in Arabic OCR may not be able to assure the comparability of the performance of Arabic OCR systems. Besides, some studies, such as [10], [11], evaluate Arabic OCR systems only in terms of character accuracy, which is the standard performance metric. However, despite this it is claimed that the accuracy rates of Arabic OCR systems are comparable. Thus, different accuracy metrics to classify the Arabic OCR are required [12].

On the other hand, to obtain statistically meaningful results when evaluating the performance of OCR systems, a great number of text images are needed to be tested [4]. Although tools exist for other languages (e.g. English [13], [14]), none exists for Arabic. Therefore, a programmed test tool is needed to automate the evaluation test. Additionally, conducting the performance evaluation under an automated tool will result in eliminating human error, improve speed, precision and reduce repetition [12].

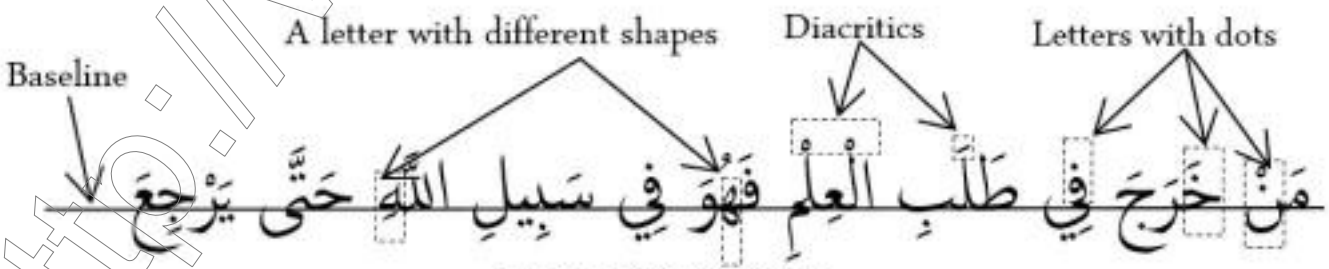

Figure 1. Arabic script characteristics

There are several experimental studies on Arabic OCR evaluation [15], [5] and [16]. Still, only two studies—[12], [16]—indicate that an automated tool was utilised to evaluate the performance of Arabic OCR systems. Unfortunately, the utilised tool is not available for researchers and is limited in accuracy metrics.

This paper first presents an overview of performance evaluation for OCR systems. Then, a set of objective accuracy metrics for Arabic OCR evaluation is discussed. Also, the developed tool for the evaluation of Arabic OCR performance is described.

II. AN OVERVIEW OF OCR EVALUATION

Evaluation of OCR systems can be classified into two types: black-box evaluation and white-box evaluation [15]. In the former, an entire OCR engine is treated as an indivisible unit, so the submodules of the OCR system are not known to the evaluator [15]. Further, the evaluation of this type is concerned with the output rather than how the output is produced, whereas with the latter evaluation, the submodules of the OCR system must be accessed by the evaluator in order to evaluate each submodule [5]. Thus, it is possible to evaluate different proposed OCR systems by relying on the white-box evaluation method, since it does not require having access to the submodules of OCR systems.

Evaluating the performance of OCR systems involves comparing an observed variable, which is the output text of OCR, with a reference variable, which is the original text, called ground truth, under controlled conditions [17], [18].

III. OCR ACCURACY EVALUATION

A general measure of OCR systems, when the ground truth text is available, is to determine the differences between the OCR output text and the ground truth text [18]. Since the total cost of correcting errors in the output of OCR system is an important factor of an OCR performance system, the most meaningful measurement of an OCR system is accuracy [17]. A general definition of the term "accuracy" of OCR systems is the number of items (characters or words) that have been recognised correctly on a text image and normalised by the total number of items in the ground truth text [5].

In order to assess accuracy, the number of correct, inserted, deleted and substituted symbols should be taken into consideration [19]. The accuracy of OCR engines can be determined by the *edit distance* between the OCR output text and the ground truth text. The authors in [18] clarify that the edit distance between the output text of OCR and the ground truth text can be defined as the minimum number of operations required to convert the output text into the ground truth text. The edit operations are: insertion, deletion and substitution.

For example, the edit distance between the two strings, the ground truth string (جامعة بالجور) and the OCR-generated string, (جامعة بياتجور) is 2. The required operations to transform the OCR-generated string into the correct one are: (1) substitute the underlined letter in the OCR output (ـت); for (ـة) and (2) delete the underlined letter (ـا). This number, which is the minimum number of single character edit operations, is known as the Levenshtein distance [20]. Regarding OCR evaluation studies, it is common to determine the edit distance in different levels: word level and character level. However, character level accuracy is useful for predicting improvements in OCR systems in which OCR developers are interested, whereas word level accuracy is useful for analysing the ease of human readability, as [5] emphasise. Word accuracy is outside the scope of this of this paper and is left for future work.

Owing to the inadequate quality of OCR systems, accuracy metrics that provide interesting measures of the OCR performance system have been demanded by many studies [17], [21], [13] and [22]. In particular, for Arabic OCR systems, accuracy rates are comparable, and thus some researchers call attention to the need for different performance accuracy metrics to classify Arabic OCR systems [16]. The next section will provide unique Arabic accuracy metrics that will be implemented in the tool.

IV. ARABIC ACCURACY METRICS

The literature suggests several possible performance measures for evaluating OCR, such as character accuracy, accuracy by character class and marked character accuracy [23], [22] and [12]. However, because of Arabic script characteristics, it is not sufficient to assess Arabic OCR performance by relying on such metrics. Thus, the authors propose different accuracy metrics for evaluation of Arabic OCR engines as discussed below.

A. Character accuracy

Accuracy in an OCR-generated text in respect to the ground truth text is computed by Levenshtein edit distance; that is, the minimum number of primitive operations that are required to correct the OCR output text to be matched with the ground truth text. These operations are substitution, deletion and insertion. Then, character accuracy is determined by:

$$\frac{m-e}{m}\times 100 \qquad (1)$$

where e is the edit distance, and m is the number of characters in the reference text.

B. Character accuracy based on character class

Considering Arabic script, some characters have features that allow them to be classified into a particular class. Consequently, it will be valuable to analyse the accuracy of each class. To determine the accuracy of this metric, Arabic characters have been classified into various classes as below:

1) Character position (form shape) class

As illustrated earlier, an Arabic letter may have various shapes depending on its position in a word, whether it is an isolated letter, an initial letter, a middle letter or a terminal letter. In order to analyse the accuracy of this class, the ground truth Arabic characters are categorised into four classes: isolated, initial, middle, and end, as shown in Table I. Then, the

tool will compute the accuracy of each class by using equation (1).

TABLE I CHARACTER POSITION CLASS

Isolated	Initial	Middle	End
ا	ا	ـا	ـا
ب	بـ	ـبـ	ـب
ت	تـ	ـتـ	ـت
ث	ثـ	ـثـ	ـث
ج	جـ	ـجـ	ـج
ح	حـ	ـحـ	ـح
خ	خـ	ـخـ	ـخ
د	د	ـد	ـد
ذ	ذ	ـذ	ـذ
ر	ر	ـر	ـر
ز	ز	ـز	ـز
س	سـ	ـسـ	ـس
ش	شـ	ـشـ	ـش
ص	صـ	ـصـ	ـص
ض	ضـ	ـضـ	ـض
ط	طـ	ـطـ	ـط
ظ	ظـ	ـظـ	ـظ
ع	عـ	ـعـ	ـع
غ	غـ	ـغـ	ـغ
ف	فـ	ـفـ	ـف
ق	قـ	ـقـ	ـق
ك	كـ	ـكـ	ـك
ل	لـ	ـلـ	ـل
م	مـ	ـمـ	ـم
ن	نـ	ـنـ	ـن
ه	هـ	ـهـ	ـه
و	و	ـو	ـو
ي	يـ	ـيـ	ـي

2) Dot character class

Arabic OCR systems have faced challenges in recognising characters that contain dots, since some characters have the same shape and they can be distinguished only by the number of dots, as illustrated in Fig. 1. Thus, in order to evaluate the accuracy of this class, Arabic characters can be categorised into four classes: one dot, two dots, three dots and non-dot characters, as displayed in Table II. Each class accuracy is given by equation (1).

TABLE II DOT CHARACTER CLASS

One dot	Two dots	Three dots	Non-dot
ب ج خ ذ ز ض ظ غ ف ن	ت ق ي ة	ث ش	ا ح د ر س ص ط ع ل م ه و ك ى

3) Zigzag character class

Arabic script can be written with Hamza, which has the zigzag shape. It would be interesting to measure an Arabic OCR in terms of characters that contain the zigzag shape. Consequently, this tool analyses the accuracy of the zigzag character class that is shown in Fig. 2. The zigzag character accuracy is also computed by equation (1).

4) Dot Character based on baseline class

Dots of Arabic characters can be placed either above or below the baseline. Owing to the significance of the baseline in Arabic OCR, it would be valuable to compute the accuracy of characters that are most related to the baseline. Regarding the baseline, Arabic dot characters will be divided into two groups: above baseline and below baseline, as shown in Table III. Each group accuracy is expressed by equation (1).

TABLE III BASELINE CHARACTER CLASS

Above baseline	Below baseline
ق ن ف غ ظ ض ش ز ذ خ ث ت	ي ج ب

5) Loop Character class:

When considering the characteristics of Arabic script, several letters consist of a loop shape. Such a feature may be an obstacle in Arabic OCR development. Thus, it would be desirable to assess the accuracy of characters that have this feature. In order to assess the accuracy of this class, loop characters are defined according to how they are presented in Fig. 3. The accuracy of loop characters are also determined by equation (1).

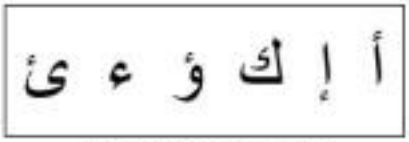

Figure 2. Zigzag characters

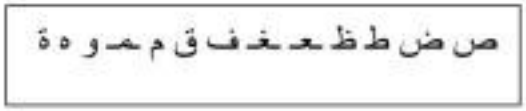

Figure 3. Loop characters

Figure 4. Diacritic marks

C. Diacritic group accuracy

A major feature of Arabic script is the appearance of diacritical marks, which influences the accuracy of Arabic OCR systems. Fig. 4 shows Arabic diacritics, which are analysed by the evaluation tool to provide the accuracy of Arabic OCR in terms of diacritics by equation (1).

D. Digit group accuracy

It is critical for developers to assess the recognition of numbers by Arabic OCR applications. In respect to Arabic script, both Arabic and English numerals may appear in any Arabic text. Thus, the evaluation considers all numerals in computing the accuracy of digits. Digit accuracy is provided by equation (1).

E. Punctuation group accuracy

Several punctuation symbols resemble Arabic characters. For example, the full-stop punctuation (.) is quite identical to the Arabic Zero number (٠). This illustration confirms that the punctuation group has a vital effect on Arabic OCR accuracy. As a result, determining the accuracy of this group will be measured with the presented tool by equation (1).

V. Arabic OCR evaluation tool

The Arabic OCR evaluation tool, which is programmed in Java, allows for comparison of an OCR-generated text file with a ground truth text file. The difference between the two text files is computed in terms of the minimum cost of converting the OCR output text to the ground truth. For this purpose, Levenshtein edit distance algorithm in [20] has been adopted.

The Levenshtein method calculates the edit distance between two strings where edit distance is the minimum number of insertions, substitutions or deletions that are necessary to convert one string into the other. For a comprehensive explanation of the Levenshtein edit distance algorithm refer to [20].

By using the software tool, an evaluator can quantitatively measure the performance of Arabic OCR systems according to various Arabic accuracy metrics. A Graphical User Interface (GUI) has been implemented to facilitate evaluators to select the OCR output text file and the ground truth text file, as demonstrated in Fig. 5. Also, as can be seen in Fig. 5, after clicking on the *Evaluate* button, the GUI displays statistics of all the accuracy metrics discussed.

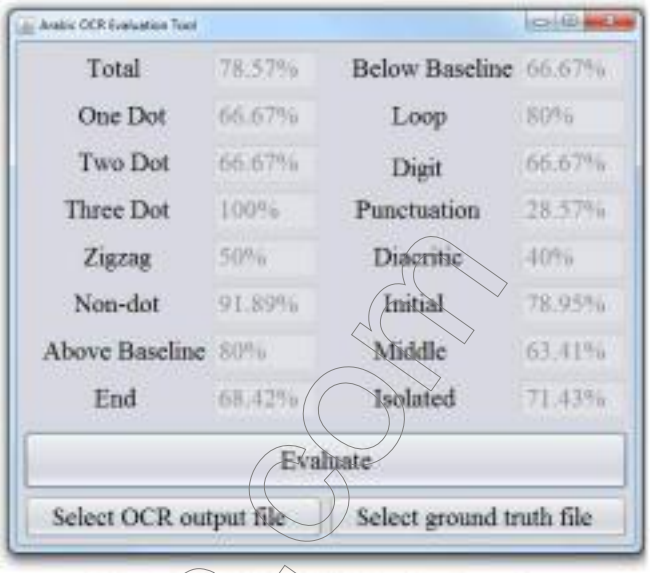

Figure 5. GUI of the Arabic OCR evaluation tool

The Arabic OCR evaluation tool is provided freely as open-source software at the following GitHub address: https://github.com/NLPBangor/OCREvaluationTool.

In order to illustrate how the tool works, an example is provided here. Fig. 6 shows the text image that contains Arabic text with digits, diacritics and punctuation marks. The open source OCR Tesseract engine[1], which supports Arabic, was run to convert the Arabic text image to an editable text. Fig. 7 displays the corresponding OCR-generated text. To evaluate the accuracy of the utilised OCR application, the tool was run on the OCR output text and the ground truth text. The result of the evaluation is presented in Fig. 5.

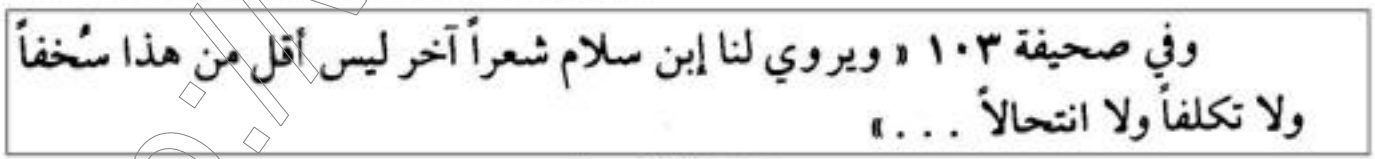

Figure 6. Arabic text image

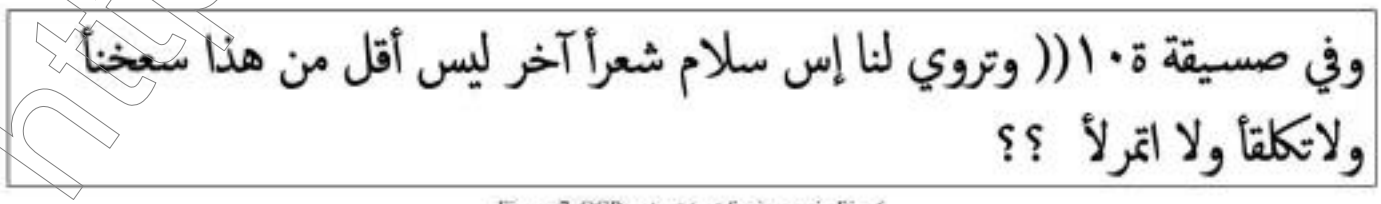

Figure 7. OCR output text for image in Fig. 6

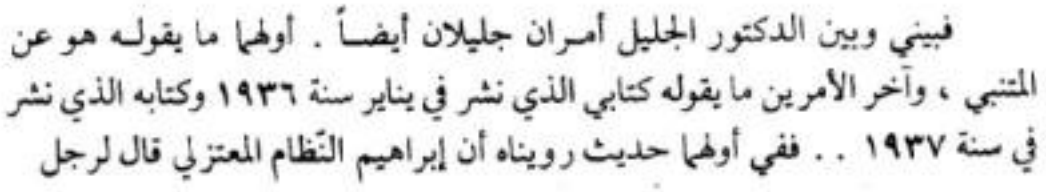

Figure 8. Second example of an Arabic text image

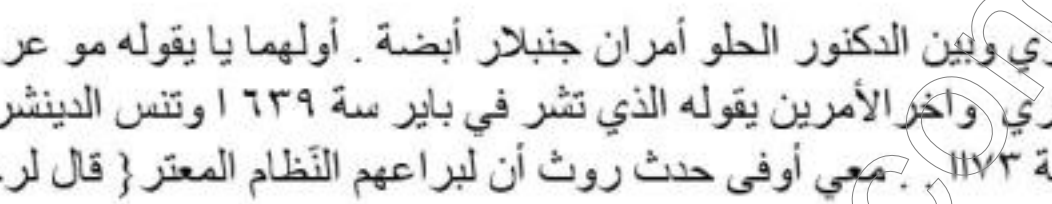

Figure 9. OCR output text of second image text in Fig. 8

This example illustrates the idea that evaluation of Arabic OCR systems according to different accuracy measures is required. In other words, it can be seen from the results in Fig. 5 that the total character accuracy is about 78%. However, the accuracy results for individual character classes are significantly different. In particular, the accuracy of the one dot characters group and the two dots characters group are around 67%, whereas, the accuracy of the three dot character class is 100%. Moreover, the observed difference between the accuracy of the zigzag characters group and the loop characters group is significant, 50% and 80% respectively. Another important finding is that the accuracy of the above baseline characters class (80%) is very different to the accuracy of below baseline characters (67%). Also, it can clearly be seen that the punctuation class has the lowest accuracy rate.

Another example is provided in order to show the importance of the accuracy metrics, which have been provided in this paper, in evaluating the performance of Arabic OCR systems. The text image and the corresponding OCR-generated text are shown in Fig. 8 and Fig. 9 respectively. Fig. 10 presents the results obtained from the evaluation of the Tesseract OCR system on the text image which are clearly different to the previous example.

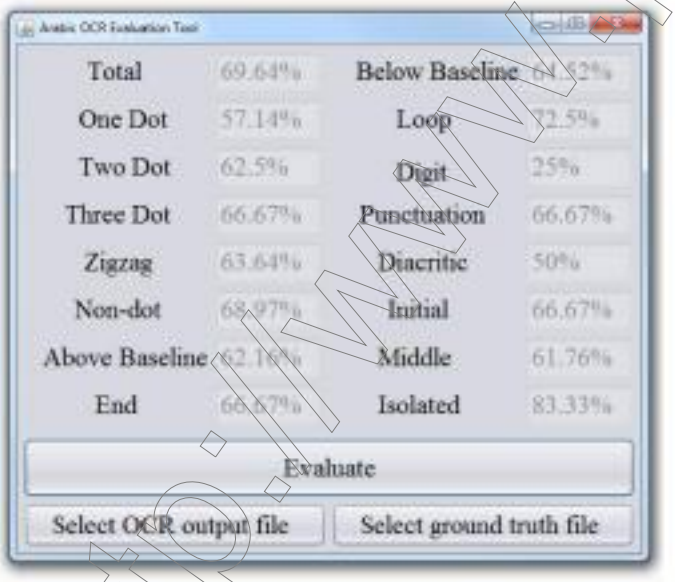

Figure 10. Evaluation results of the second test of the Arabic Tesseract OCR software

Based on the findings above, it is possible to state that measuring Arabic OCR performance in terms of the accuracy metrics, which are implemented in the Arabic OCR evaluation tool, can help researchers to analyse the strengths and weaknesses of Arabic OCR systems, helps to determine open problems in Arabic OCR systems, and compare alternative OCR systems in order to enhance the accuracy of OCR systems.

VI. CONCLUSION

A software tool for performance evaluation, based on different objective accuracy metrics, has been described. This tool has been specifically developed to assist Arabic OCR researchers to calculate the accuracy of different Arabic OCR systems. Furthermore, by using this automated tool, an OCR evaluator can quickly process an enormous number of testing experiments. Thus, the authors believe that the Arabic OCR evaluation tool that has been presented in this paper will be very useful by the OCR community.

For future work, the evaluation tool will be enhanced. For example, more string similarity metric algorithms other than Levenshtein edit distance algorithms would allow comparison of the results of different algorithms.

REFERENCES

[1] B. Al-Badr and R. M. Haralick, "Segmentation-free word recognition with application to Arabic," *Proc. 3rd Int. Conf. Doc. Anal. Recognit.*, vol. 1, pp. 355–359, 1995.

[2] M. T. Parvez and S. a. Mahmoud, "Offline arabic handwritten text recognition: A Survey," *ACM Comput. Surv.*, vol. 45, no. 2, pp. 23:1–23:35, 2013.

[3] F. Slimane, S. Kanoun, J. Hennebert, A. M. Alimi, and R. Ingold, "A study on font-family and font-size recognition applied to Arabic word images at ultra-low resolution," *Pattern Recognit. Lett.*, vol. 34, no. 2, pp. 209–218, 2013.

[4] S. V Rice, "Measuring the accuracy of page-reading systems," University of Nevada, Las Vegas, 1996.

[5] T. Kanungo, G. A. Marton, and O. Bulbul, "OmniPage vs. Sakhr: Paired model evaluation of two Arabic OCR products," in *Electronic Imaging '99*, 1999, pp. 109–120.

[6] S. Mihov, K. U. Schulz, C. Ringlstetter, V. Dojchinova, V. Nakova, K. Kalpakchieva, O. Gerasimov, A. Gotscharek, and C. Gercke, "A corpus for comparative evaluation of OCR software and postcorrection techniques," *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, vol. 2005, pp. 162–166, 2005.

[7] S. Taha, Y. Babiker, and M. Abbas, "Optical character recognition of arabic printed text," *SCOReD 2012 - 2012 IEEE Student Conf. Res. Dev.*, pp. 235–240, 2012.

[8] I. Supriana and A. Nasution, "Arabic Character Recognition System Development," *Procedia Technol.*, vol. 11, no. Iceei, pp. 334–341, 2013.

[9] H. Pirsiavash, R. Mehran, and F. Razzazi, "A robust free size OCR for omni-font persian/arabic printed document using combined MLP/SVM," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 3773 LNCS, pp. 601–610, 2005.

[10] M. Dahi, N. A. Semary, and M. M. Hadhoud, "Primitive printed Arabic Optical Character Recognition using statistical features," in *2015 IEEE Seventh International Conference on Intelligent Computing and Information Systems (ICICIS)*, 2015, pp. 567–571.

[11] I. Ahmad, S. A. Mahmoud, and G. A. Fink, "Open-vocabulary recognition of machine-printed Arabic text using hidden Markov models," *Pattern Recognit.*, vol. 51, pp. 97–111, 2016.

[12] S. Saber, A. Ahmed, A. Elsisi, and M. Hadhoud, "Performance Evaluation of Arabic Optical Character Recognition Engines for Noisy Inputs," in *The 1st International Conference on Advanced Intelligent System and Informatics (AISI2015), November 28-30, 2015, Beni Suef, Egypt*, 2016, pp. 449–459.

[13] R. C. Carrasco, "An Open-source OCR Evaluation Tool," in *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 2014, pp. 179–184.

[14] T. A. Nartker, S. V Rice, and S. E. Lumos, "Software tools and test data for research and testing of page-reading OCR systems," in *In International Symposium on Electronic Imaging Science and Technology*, 2005.

[15] T. Kanungo, G. a Marton, and O. Bulbul, "Performance evaluation of two Arabic OCR products," *Proc. AIPR*, pp. 76–83, 1999.

[16] S. Saber, A. Ahmed, and M. Hadhoud, "Robust metrics for evaluating arabic OCR systems," in *Image Processing, Applications and Systems Conference (IPAS), 2014 First International*, 2014, pp. 1–6.

[17] J. Kanai, T. Nartker, and S. Rice, "Performance Metrics for Document Understanding Systems," *Doc. Anal.*, pp. 424–427, 1993.

[18] W. J. Teahan, S. Inglis, J. G. Cleary, and G. Holmes, "Correcting English text using PPM models," in *Data Compression Conference Proceedings*, 1998.

[19] E. Boruvikov and W. Lane, "A survey of modern optical character recognition techniques (DRAFT)," vol. 1, no. 301, pp. 1–37, 2004.

[20] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Sov. Phys. Dokl.*, vol. 10, no. 8, pp. 707–710, 1966.

[21] S. Tanner, T. Muñoz, and P. H. Ros, "Measuring Mass Text Digitization Quality and Usefulness," *D-Lib Mag.*, vol. 15, no. 7/8, 2009.

[22] S. V Rice, J. Kanai, and T. A. Nartker, "An evaluation of OCR accuracy," *Inf. Sci. Res. Institute, 1993 Annu. Res. Rep.*, vol. 9, p. 20, 1993.

[23] Y. Batawi and O. Abulnaja, "Accuracy Evaluation of Arabic Optical Character Recognition Voting Technique: Experimental Study," *Int. J. Electr. Comput. Sci.*, vol. 2, no. 1, pp. 29–33, 2012.