Experimental evaluation of Arabic OCR systems

Mansoor Alghamdi
University of Tabouk, Tabouk, Saudi Arabia, and
William Teahan
School of Computer Science, Bangor University, Bangor, UK

Received 4 May 2017
Revised 2 August 2017
Accepted 6 September 2017

Abstract

Purpose – The aim of this paper is to experimentally evaluate the effectiveness of the state-of-the-art printed Arabic text recognition systems to determine open areas for future improvements. In addition, this paper proposes a standard protocol with a set of metrics for measuring the effectiveness of Arabic optical character recognition (OCR) systems to assist researchers in comparing different Arabic OCR approaches.

Design/methodology/approach – This paper describes an experiment to automatically evaluate four well-known Arabic OCR systems using a set of performance metrics. The evaluation experiment is conducted on a publicly available printed Arabic dataset comprising 240 text images with a variety of resolution levels, font types, font styles and font sizes.

Findings – The experimental results show that the field of character recognition for printed Arabic still requires further research to reach an efficient text recognition method for Arabic script.

Originality/value – To the best of the authors' knowledge, this is the first work that provides a comprehensive automated evaluation of Arabic OCR systems with respect to the characteristics of Arabic script and, in addition, proposes an evaluation methodology that can be used as a benchmark by researchers and therefore will contribute significantly to the enhancement of the field of Arabic script recognition.

Keywords Performance evaluation, Performance metrics, Arabic OCR, OCR

Paper type Research paper

Introduction

Optical character recognition (OCR) is a technique that aims to automatically convert a machine-printed or handwritten text image into an editable text format (Alghamdi *et al.*, 2016). This technique is highly desirable in various real-world applications, such as digitising learning resources to assist visually impaired people, bank cheque processing and mail sorting (Alginahi, 2013; Al-Badr and Mahmoud, 1995). Generally, the process for developing OCR systems involves five stages: pre-processing, segmentation, feature extraction, classification and post-processing. In each stage, specific techniques are applied; for more details, see Khorsheed (2002).

Previous research on text recognition has focused primarily on Latin scripts, such as English and Chinese, and it has not been until the last two decades that recognition of non-Latin scripts, such as Arabic, have been researched (Alginahi, 2013). Although

PSU Research Review
Vol. 1 No. 3, 2017
pp. 229-241
Emerald Publishing Limited
2399-1747
DOI 10.1108/PRR-05-2017-0026

handwritten script is significantly more challenging than printed Arabic text for OCR, Arabic printed text OCR still poses significant challenges (Alghamdi *et al.*, 2016). Therefore, this study will deal only with Arabic printed text. Figure 1 illustrates the characteristics of Arabic printed text that contribute to the inadequate development of Arabic script recognition. Arabic script is written cursively through a baseline and contains loop-shaped characters, zigzag-shaped characters, dot characters and diacritics. Moreover, a character might have up to four different shapes in relation to its position in a word. Therefore, research is being undertaken seeking more solutions for Arabic OCR systems (Parvez and Mahmoud, 2013; Slimane *et al.*, 2013).

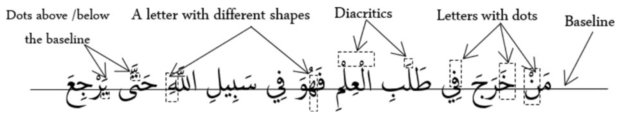
To produce an efficient Arabic OCR system, effective performance evaluation of current OCR systems is essential. Furthermore, evaluating OCR performance contributes to monitoring progress in OCR system development, analysing the effectiveness of OCR systems, identifying open areas and providing a scientific explanation for the performance of OCR systems (Kanungo *et al.*, 1999a; Mihov *et al.*, 2005).

Despite the significance of Arabic OCR system performance evaluation, relatively little work has been published on empirical analysis of the effectiveness of Arabic OCR systems. For instance, two studies provide an evaluation of two Arabic OCR systems (Kanungo *et al.*, 1999a, 1999b). However, as these studies were conducted in 1999, over 17 years ago, they do not reflect current progress in the field of Arabic OCR development (Alginahi, 2013). A more recent empirical study provides a comparative evaluation of the most common Arabic OCR systems (Saber *et al.*, 2016). However, the study by these authors only investigated the effectiveness of input quality images on the performance of Arabic OCR systems. In addition, exploring Arabic OCR systems in relation to their sensitivity to different levels of page quality may not be adequate in fairly assessing their success, as some OCR systems include a combination of image enhancement techniques. To the best of the authors' knowledge, no established work has gauged the current progress in the enhancement of Arabic OCR in terms of the challenges of Arabic script.

Moreover, the performance assessments of Arabic OCR systems are only reported by their developers: their results are derived from different datasets that might be small or might be used in developing the systems (Al-Badr and Mahmoud, 1995; Alginahi, 2013). Consequently, as these performance tests are statistically invalid, they cannot be used to compare the performance between Arabic OCR systems (Margner and El Abed, 2009). Evaluating the performance of Arabic OCR systems is also challenging as no standard dataset is available nor is a set of performance metrics freely available to the community of Arabic OCR developers (Al-Muhtaseb and Qahwaji, 2011; Abdelraouf *et al.*, 2008; Al-Badr and Mahmoud, 1995; Ahmad *et al.*, 2016). In addition, most reports on the performance of Arabic OCR systems are in terms of the general, standard performance measurement of character accuracy, such as in Dahi *et al.* (2015) and Ahmad *et al.* (2016). However, this performance metric is insufficient to assess how Arabic OCR systems are coping with the challenges of Arabic script.

Thus, the current work first attempts to provide a better insight into the effectiveness of the state-of-the-art printed Arabic OCR systems with possible interpretations for future performance enhancement. It then aims to propose a standard protocol with a set of metrics

Figure 1.
Printed Arabic script characteristics


Dots above /below the baseline
A letter with different shapes
Diacritics
Letters with dots
Baseline
مَنْ خَرَجَ فِي طَلَبِ الْعِلْمِ فَهُوَ فِي سَبِيلِ اللَّهِ حَتَّى يَرْجِعَ

for measuring the effectiveness of Arabic OCR systems which we hope will be used as a benchmark by researchers in comparing between OCR algorithms.

This paper is organised as follows: in the first section below, the most common Arabic OCR systems are introduced. The Arabic OCR system evaluation background and performance metrics for Arabic OCR system evaluation are discussed in the second and third sections, respectively. An experimental protocol is presented in the fourth section. The experimental results are then presented and discussed. In the final section, future work is suggested and the conclusion is presented.

Arabic OCR systems

Only a handful of OCR systems claim that they are capable of recognising Arabic script. Our evaluation study is limited to the four most well-known Arabic OCR systems, namely, Automatic Reader 11.2 produced by the Sakhr Software Company; FineReader 12 produced by the ABBYY Company; Clever Page produced by RDI (Research & Development International) and Tesseract produced originally by Hewlett-Packard (HP). In the following subsections, the four Arabic OCR engines are briefly discussed.

Automatic reader 11.2 software

Automatic Reader is a commercial product first developed by the Sakhr Software Company in 1982 for text recognition of Arabic script. It supports Arabic language and several Arabic character-based languages, such as Arabic, Farsi and Urdu. Sakhr claims that Automatic Reader has been ranked as the best existing Arabic OCR software for high-quality text images by US government evaluators (Sakhr Software OCR, 2017). It supports multi-font type and multi-resolution images. However, font size 8 is not supported by the Automatic Reader OCR software (henceforth, referred to as Sahkr OCR).

FineReader 12 software

FineReader is produced commercially by a global company, called ABBYY, as advanced OCR software. The performance of FineReader has been enhanced by ABBYY for many years. FineReader 12 supports 190 languages including Arabic script using dictionary support (Abbyy OCR, 2017). It supports multi-font types, multi-size and multi-resolution images. Henceforth, FineReader will be referred to as ABBYY OCR.

Clever page software

Clever Page originally began as a PhD research study by El-Mahallawy (2008). Since 2008, Clever Page has been designed and developed as an Omni font-written Arabic OCR engine by Research & Development International (RDI). It supports multi-font types and multi-size text images. However, it is worth mentioning that the Clever Page OCR software only works on pages with 300 dots per inch (dpi). Henceforth, Clever Page will be referred to as RDI OCR.

Tesseract software

Tesseract is an OCR engine designed at Hewlett-Packard (HP) between 1984 and 1994. Since late 2005, it has been maintained by Google and released as open source OCR software. However, Arabic support has only been added recently (Sabbour and Shafait, 2013). Tesseract is the only Arabic OCR software that is freely available. It supports multi-font types, and multi-size and multi-resolution images.

Arabic OCR system performance evaluation background

Evaluation of OCR systems can be classified into two types: black-box evaluation and white-box evaluation (Kanungo *et al.*, 1999b). In black-box evaluation, an entire OCR system is treated as an indivisible unit; thus, the evaluators do not have access to the submodules of the OCR system. Furthermore, this evaluation type is only concerned with the output of the OCR system rather than how it is produced. On the other hand, in white-box evaluation, the evaluator must have access to the submodules of the OCR system to evaluate each submodule (Kanungo *et al.*, 1999a). As accessing the submodules of commercial Arabic OCR systems was not possible for the authors of this paper and as our interest is only in error analysis of Arabic OCR system output, the white-box evaluation type is outside the scope of this evaluation study.

To evaluate OCR systems, comparison between an observed variable, which is the output text of the OCR system, and a reference variable, which is the original text called “ground-truth”, is required (Kanai *et al.*, 1993; Teahan *et al.*, 1998).

Metrics for Arabic OCR system performance evaluation

Generally, Arabic OCR systems are evaluated in terms of character accuracy with this obtained by identifying the differences between the ground-truth text and the OCR output text. These differences can be determined by the edit distance which is the minimum number of edit operations required to correct the OCR output text to be matched with the ground-truth text. These edit operations are: character insertion, character deletion and character substitution; for more details refer to Levenshtein (1966).

This general performance metric is not sufficient to measure the performance of an Arabic OCR system, as the accuracy rates of Arabic OCR systems are comparable (Saber *et al.*, 2016). Moreover, the character accuracy metric only provides us with a measure of how well an Arabic OCR system performs in text recognition in general terms. Thus, it cannot provide us with insight into which systems have overcome the various challenges of Arabic script.

On the other hand, a recent study (Alghamdi *et al.*, 2016) tackles this problem by suggesting various objective performance metrics for evaluating Arabic OCR systems which can provide us with more insight into the effectiveness of Arabic OCR systems. Therefore, the current study adopts these metrics to evaluate the performance of Arabic OCR systems. The adopted performance metrics are defined in the following subsections.

Overall character accuracy

Overall character accuracy determines the accuracy of Arabic OCR over all of the tested text images. Character accuracy is the percentage of ground-truth characters that are recognised correctly on an Arabic text image, by comparing the ground-truth text file with the OCR output text file. In accordance with Alghamdi *et al.* (2016), character accuracy is determined by equation (1):

$$\frac{m-e}{m} \times 100 \tag{1}$$

where m is the number of characters in the ground-truth text file and e is the edit distance. The cost for each edit operation is defined as 1.

Character accuracy based on character position

As previously mentioned, an Arabic character may have one to four shapes depending on its position in a word; for example, see Table I. The four possible shapes are isolated, initial, middle and end. Thus, it is valuable to analyse the effectiveness of Arabic OCR systems in recognising isolated, initial, middle and end characters. To analyse accuracy, Arabic characters have been categorised into isolated, initial, middle and end classes by Alghamdi *et al.* (2016), as shown in Table I. The accuracy of each class is determined by equation (1).

Dot character accuracy and no-dot character accuracy

One of the challenges of Arabic OCR is the presence of dots in Arabic script (Alghamdi and Teahan, 2017). Assessing the impact of dot and no-dot characters on the performance of Arabic OCR systems is therefore of interest. To do so, Arabic characters have been categorised into four classes: one dot class, two dot character class, three dot character class and no-dot character class by Alghamdi *et al.* (2016), as illustrated in Table II. The accuracy of each class is determined by equation (1).

Dot character accuracy based on baseline

The presence of a baseline is specific to Arabic script characteristics. The baseline is significant in developing Arabic OCR systems. Alghamdi *et al.* (2016) classify Arabic characters into two classes: dot character above the baseline and dot character below the baseline to compare the performance of Arabic OCR in each class, as illustrated in Table III. The accuracy of each class is also determined by equation (1).

Zigzag-shaped character accuracy

One of the distinguishing characteristics of Arabic script is the presence of a zigzag-shaped mark (ء), called *Hamza*, with some Arabic characters. The aim of using this metric is to expose the sensitivity of Arabic OCR systems to zigzag-shaped characters. Alghamdi *et al.* (2016) compute the zigzag-shaped character accuracy by using equation (1).

Isolated	Initial	Middle	End
ه	هـ	ـهـ	ـه

Table I. An Arabic character with dissimilar shapes

One dot	Two dots	Three dots	No-dot
ب ج خ ذ ز ض ظ غ ف ن	ت ق ي ة	ث ش	ح د ر س ص ط ع ل م ه و ك

Table II. Examples of dot characters and no-dot characters

Dot character above baseline	Dot character below baseline
ق ن ف غ ظ ض ش ز ذ خ ث ت	ي ج ب

Table III. Examples of dot characters based on the baseline

Loop-shaped character accuracy

Several Arabic characters have a loop shape, such as *Saad* (ص), *Dhad* (ض), Fa (ف, *Meem* (م) and *Qaf* (ق). According to Alghamdi *et al.* (2016), an obstacle for Arabic OCR is recognising Arabic characters that contain a loop shape. To assess the effectiveness of Arabic OCR systems for recognising these characters, the accuracy of loop-shaped characters is also computed by equation (1).

Diacritics accuracy

Some Arabic text may be written with diacritical marks, as previously illustrated in Figure 1. Thus, it is essential to evaluate the performance of Arabic OCR systems in recognising Arabic diacritical marks. As before, diacritical mark accuracy is determined by equation (1).

Digit accuracy

The digit accuracy metric is used to determine the performance accuracy of Arabic OCR systems in recognising Hindi–Arabic digits (Alghamdi *et al.*, 2016). Digit accuracy is determined by equation (1).

Punctuation accuracy

The punctuation accuracy metric is used to assess the performance accuracy of Arabic OCR systems in recognising punctuation symbols (Alghamdi *et al.*, 2016). Punctuation accuracy is determined by equation (1).

Experimental protocol

Very few Arabic datasets are freely available to researchers. The most widely used dataset for evaluating Arabic OCR approaches is the Arabic Printed Text Image (APTI). This public dataset was developed by Slimane *et al.* (2009). However, APTI is a word-level dataset where each text image contains only one Arabic word. Another Arabic dataset is ALPH-REGIM which is provided by Ben Moussa *et al.* (2010). This dataset contains about 5,000 printed and handwritten Arabic text images. Compared to the APTI dataset, ALPH-REGIM is a paragraph-based text image dataset. However, the text images are only available in one font size and at one resolution level. Thus, the KAFD dataset, developed by Luqman *et al.* (2014), is used for evaluating the performance of the four Arabic OCR systems.

The KAFD dataset is freely available and is a page-level dataset where each text image consists of a text that resulted in 2,576,024 line images. It has Arabic printed text images and corresponding ground-truth text files. These text images are available at 100 dpi, 200 dpi and 300 dpi. Furthermore, it comprises text images of 40 Arabic font types, 10 pitch sizes and four styles, with resolutions of 100 dpi, 200 dpi, 300 dpi and 600 dpi.

For the current work, 10 different font types of the text image, in which the forms of the characters are of various types, are selected, namely, Andalus, Arabic Transparent, AdvertisingBold, Diwani Letter, DecoType Thuluth, Simplified Arabic, Tahoma, Traditional Arabic, DecoType Naskh and M Unicode Sara. These font types are selected based on the level of complexity of the writing style in printed Arabic script, ranging from simple fonts with no overlaps and ligatures, such as AdvertisingBold, to more complex fonts with overlaps, such as Diwani Letter. For each font type, 8, 12, 18 and 24 pitch sizes have been used to highlight the effectiveness of Arabic OCR systems on specific pitch sizes. Moreover, to enable the evaluation of the performance of Arabic OCR systems in terms of font style, normal and italic font styles are used. To study the influence of the page resolution level on Arabic OCR system performance, all text images in the above set have been randomly selected at 100 dpi, 200 dpi and 300 dpi. This resulted in 80 text images for each resolution. Thus, this experiment is performed on 240 text images in TIFF format.

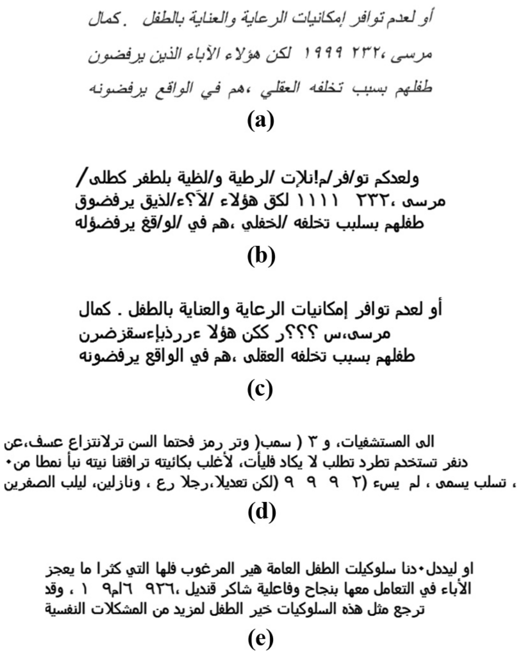
The experiment was conducted on the four Arabic OCR systems, as previously discussed, namely, Sakhr, ABBYY, RDI and Tesseract OCR systems. Sakhr and ABBYY OCR systems were kindly provided to the authors by their developers, whereas the RDI OCR system output was obtained by sending the dataset to the system's developer to apply the system to the test images.

To statistically assess the performance of each Arabic OCR system, the performance metrics described previously were used. In particular, the output text files of each Arabic OCR system were used to compute the quantities of each performance metric, corresponding to the ground-truth text files for the dataset. To eliminate human error, improve speed and precision and reduce repetition, an automated open access tool for evaluating Arabic OCR systems, provided by Alghamdi *et al.* (2016), was used to obtain the statistical data. A sample of a text image from the dataset and the corresponding Arabic OCR output are visually illustrated in Figure 2.

Experimental results and discussion

The experimental results, obtained from the evaluation experiment discussed in the previous section, are presented in this section to analyse the effectiveness of the evaluated Arabic OCR systems in printed Arabic text recognition.

The overall character accuracy scores for the four evaluated Arabic OCR systems over different resolutions are presented in Figure 3. It is apparent that the Arabic OCR systems are affected by the resolution of the text images. In particular, the results shows a gradual increase in the overall character accuracy of all Arabic OCR systems when increasing the



أو لعدم توافر إمكانيات الرعاية والعناية بالطفل .كمال
مرسى، ٢٣٢ ١٩٩٩ لكن هؤلاء الآباء الذين يرفضون
طفلهم بسبب تخلفه العقلي ،هم في الواقع يرفضونه

(a)

ولعدكم تو/فر/م!نلإت /الرطية و/لظية بلطفر كطلى/
مرسى ،٢٣٢   ١١١١ لكق هؤلاء /لآ؟ء/لذيق يرفضوق
طفلهم بسلبب تخلفه /لخفلي ،هم في /لو/قغ يرفضؤله

(b)

أو لعدم توافر إمكانيات الرعاية والعناية بالطفل . كمال
مرسى،س ؟؟؟ر ككن هؤلا ءررذبإءسقزضرن
طفلهم بسبب تخلفه العقلى ،هم في الواقع يرفضونه

(c)

الى المستشفيات، و ٣ (سمب(وتر رمز فحتما السن ترلانتزاع عسف،عن
دنفر تستخدم تطرد تطلب لا يكاد فليأت، لأغلب بكائيته ترافقنا نيته نبأ نمطا من٠
، تسلب يسمى ، لم يسء (٢ ٩ ٩ ٩ ٩ (لكن تعديلا،رجلا رع ، ونازلين، ليلب الصفرين

(d)

او ليددل٠دنا سلوكيلت الطفل العامة هير المرغوب فلها التي كثرا ما يعجز
الأباء في التعامل معها بنجاح وفاعلية شاكر قنديل ،٩٣٦ ٦ام٩ ١ ، وقد
ترجع مثل هذه السلوكيات خير الطفل لمزيد من المشكلات النفسية

(e)

Figure 2.
(a) A text image from the KAFD dataset; (b) output of the Sakhr OCR system; (c) output of the ABBYY OCR system; (d) output of the RDI OCR system and (e) output of the tesseract OCR system

resolution of text images from 100 dpi to 300 dpi. In addition, a clear upward trend is apparent in the overall character accuracy of Sakhr OCR when the resolution of text images increased from 100 dpi to 200 dpi. Interestingly, high scores of character accuracy at 100 dpi, 200 dpi and 300 dpi were obtained by the ABBYY OCR system. Crucially, this system among the four Arabic OCR systems has the highest character accuracy at 100 dpi, 200 dpi and 300 dpi of 46, 60 and 62 per cent, respectively.

The accuracy of Arabic OCR systems in recognising Arabic characters based on their position in a word are provided in Figure 4. As has been mentioned, the connectivity feature of Arabic script is an obstacle to Arabic text recognition. This is supported by the experiment data which indicate that all of the evaluated Arabic OCR systems have higher accuracy in recognising isolated characters when they are not connected with other characters in a word. On the other hand, the performance of the Arabic OCR systems decreased in recognising initial, middle and end characters, as shown in Figure 4. Compared to the Arabic OCR systems' accuracy in recognising isolated characters, the low recognition accuracy of initial, middle and end characters by the evaluated systems is possibly because of the segmentation algorithms implemented by these systems, as the segmentation process is not required for recognising isolated characters. However, a new methodology is needed to evaluate the segmentation stage in Arabic OCR systems to gain a better understanding of these results.

The results of the recognition accuracy performance of the Arabic OCR systems in terms of one dot, two dot, three dot and no-dot characters are illustrated in Figure 5. It is obvious that the recognition accuracy rates of no-dot characters are significantly better for all evaluated Arabic OCR systems, compared to one, two and three dot characters. This confirms that one of the challenges in Arabic OCR development is the presence of dots in Arabic script, as discussed previously. It has been hypothesised that a thinning algorithm, as a pre-processing technique, has an influence on the recognition accuracy of dot characters (Hosseini, 1997; Alghamdi and Teahan, 2017). In particular, some thinning algorithms may remove the dots of Arabic characters which results in misrecognising these characters, as

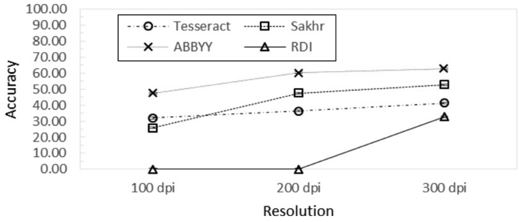


Figure 3.
Overall accuracy versus resolution results from the OCR system evaluation

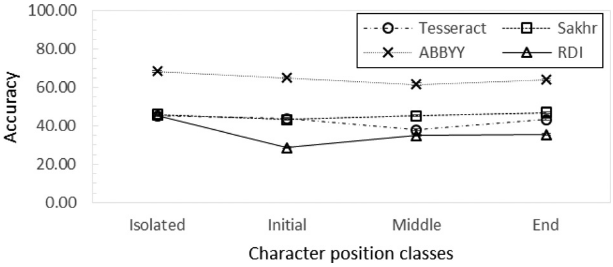


Figure 4.
Character accuracy in terms of character position results from the OCR system evaluation

stated in Alghamdi and Teahan (2017). Therefore, these results are likely to be related to thinning algorithms used in the pre-processing stage of the Arabic OCR system.

The results of analysing the performance of the evaluated Arabic OCR systems on characters that have a dot above or below the baseline are presented in Figure 6. These results highlight that Arabic OCR systems perform much better in identifying characters with a dot below the baseline than characters with a dot above the baseline. The reasons for these results are not entirely clear. However, one technique used in the pre-processing stage for developing Arabic OCR systems is page decomposition which separates the lines of a text block in a text image. To be specific, some researchers emphasise that considerable attention must be paid in the page decomposition of Arabic text images to ensure that dots placed above or below a line of text are not separated (Al-Badr and Mahmoud, 1995; Sami El-Dabi *et al.*, 1990). Thus, a possible reason for this finding may be because of the page decomposition process.

Figure 7 compares the recognition accuracy for zigzag-shaped characters and loop-shaped characters by the evaluated Arabic OCR systems. It is apparent from the results that the performance rates of Arabic OCR systems in accurately recognising loop-shaped characters are higher than in recognising zigzag-shaped characters. It can thus be assumed that recognising characters with a zigzag shape is more challenging than recognising characters that have a loop shape.

Figure 8 presents the recognition accuracy of the evaluated Arabic OCR systems in terms of the digits, punctuation and diacritics groups. Surprisingly, Tesseract and RDI OCR

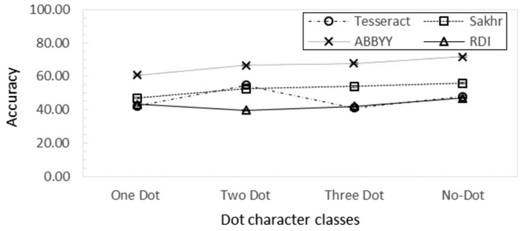


Figure 5. Dot character accuracy versus no-dot character accuracy results from the OCR system evaluation

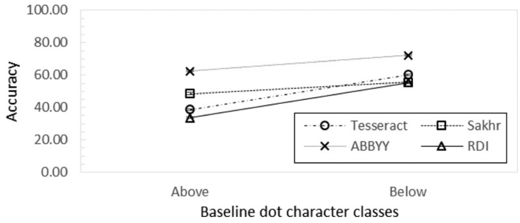


Figure 6. Dot character accuracy based on baseline results from the OCR system evaluation

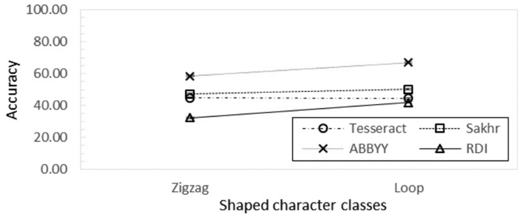


Figure 7. Character accuracy based on shaped character results from the OCR system evaluation

systems were not able to recognise any diacritical marks. However, one possible reason is that these OCR systems were not trained to recognise diacritics. In addition, Figure 8 indicates that the Sakhr OCR system was unable to recognise Arabic digits (Hindu digits). As this result was not expected, we undertook further investigation to find an explanation. After manual inspection of the output of the Sakhr OCR system, we discovered that the system recognises Arabic digits as English digits. In other words, the system replaced the Arabic digits with English digits when producing the output. It is worth mentioning that Sakhr OCR system obtained an 84 per cent digit accuracy rate when allowing English digits as not error. Thus, improved accuracy rates for the Sakhr OCR system could be achieved if the system were to overcome this problem.

The respective accuracy of the Arabic OCR systems in relation to different font pitch sizes is plotted in Figure 9. It can be seen that a drop in character recognition accuracy occurs when reducing the font pitch size. It can thus be concluded that the evaluated systems struggle when the font pitch size is too small, such as a font pitch of 8. Another important finding from this result is that the RDI OCR system performs better for a font pitch of 12 than for a font pitch of 24. The reason for this finding is unclear:

The recognition performance analysis of the evaluated Arabic OCR systems on fonts with an italic style compared to those with a non-italic font style is shown in Figure 10. Our

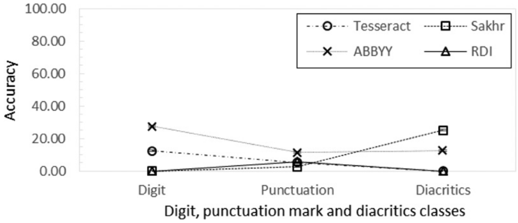


Figure 8.
Digit, punctuation and diacritics accuracy results from the OCR system evaluation

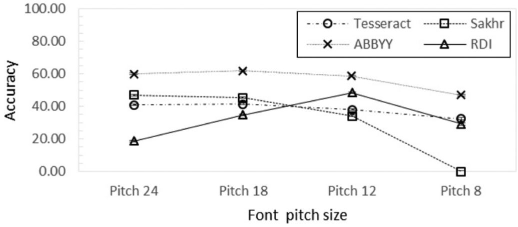


Figure 9.
Arabic OCR accuracy versus font pitch size results from the OCR system evaluation

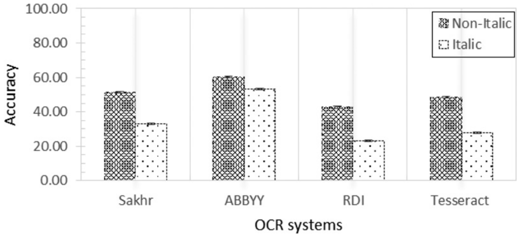


Figure 10.
Italic and non-italic font accuracy comparison results from the OCR system evaluation

experimental results show that, for all evaluated systems, the recognition accuracy rates for a non-italic font style are higher than for an italic font style. The low recognition accuracy percentages for an italic font style are likely to be related to the fact that the italic font style is a cursive font.

Table IV reports on the accuracy of the evaluated Arabic OCR systems in terms of font types. As expected, the low accuracy rates for all evaluated systems in recognising the Diwani Letter font are because of the complexity of this font which contains overlaps.

In summary, most of the performance accuracy rates of the evaluated Arabic OCR systems are below 75 per cent, indicating the continuing need for improvements in the recognition of printed Arabic script. Moreover, the correlation between the performance accuracy of Arabic OCR systems and the features of Arabic script have been highlighted. In particular, the experimental analysis indicates that the characteristics of printed Arabic script, such as the connectivity, and the presence of dots and zigzag shapes, all contribute to the challenge for Arabic OCR systems. Overall, the results indicate that recognition of Arabic script remains an open research problem.

Conclusion and further work

This paper presents the results of a performance evaluation of four well-known Arabic OCR systems: Sakhr, ABBYY, RDI and Tesseract. In addition, this study provides an experimental protocol with a set of objective performance metrics that enables the effectiveness of different Arabic OCR systems to be compared. The results of this study show that all the evaluated Arabic OCR systems have low performance accuracy rates, below 75 per cent, which means that the time which would take to manually correct the OCR output would be a prohibitive. On the other hand, the recognition accuracy rates for isolated characters are higher than for initial, middle and end characters. Moreover, the recognition accuracy rates of no-dot characters are significantly better compared to one, two and three dot characters. The current evaluation study highlights several open areas and major factors that need to be considered when either developing Arabic OCR systems or enhancing the current systems. For instance, the experimental results indicate that recognition of Arabic text with complex font, such as the Diwani Letter font, and text with diacritics are still open research problems. For future work, it will be interesting to assess the effectiveness of applying post-processing methods, such as spelling correction, on the performance accuracy rates of Arabic OCR systems. Furthermore, a methodology should be developed to evaluate the performance of each stage of Arabic OCR which would enable

	OCR System			
Font Type	Sakhr (%)	ABBYY (%)	RDI (%)	Tesseract (%)
Traditional Arabic	48.54	67.66	51.88	47.04
Tahoma	40.52	69.91	26.38	38.37
Simplified Arabic	52.97	67.69	44.94	46.75
M Unicode Sara	36.03	59.40	25.92	33.72
Diwani letter	18.13	18.47	18.13	23.32
DecoType Thuluth	36.12	37.71	24.26	32.48
DecoType Naskh	48.88	50.22	41.63	40.92
Arabic transparent	51.56	75.19	46.00	48.61
Andalus	28.07	37.53	21.68	25.34
AdvertisingBold	57.35	70.26	27.20	39.39

Table IV. Arabic OCR performance accuracy according to 10 different font types

the effects of each stage on the overall performance of Arabic OCR systems to be determined. Further work is also needed to evaluate other Arabic OCR systems. Furthermore, it will be interesting to assess the effectiveness of evaluation experiment of Arabic OCR systems on handwritten Arabic script.

References

Abbyy OCR (Optical Character Recognition) (2017), available at: www.abbyy.com/en-gb/ (accessed 15 April 2017).

Abdelraouf, A., Higgins, C.A. and Khalil, M. (2008), "A database for Arabic printed character recognition", *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Berlin Heidelberg, pp. 567-578.

Ahmad, I., Mahmoud, S.A. and Fink, G.A. (2016), "Open-vocabulary recognition of machine-printed Arabic text using hidden Markov models", *Pattern Recognition*, Vol. 51, pp. 97-111, available at: www.sciencedirect.com/science/article/pii/S0031320315003428

Al-Badr, B. and Mahmoud, S.A. (1995), "Survey and bibliography of Arabic optical text recognition", *Signal Processing*, Vol. 41 No. 1, pp. 49-77.

Alghamdi, M.A. and Teahan, W.J. (2017), "A new thinning algorithm for Arabic script", *International Journal of Computer Science and Information Security*, Vol. 15 No. 1, pp. 204-211, available at: http://search.proquest.com/openview/5c29856709508552a5566d2504966d54/1?pq-origsite=gscholar&cbl=616671 (accessed 22 April 2017).

Alghamdi, M.A., Alkhazi, I.S. and Teahan, W.J. (2016), "Arabic OCR evaluation tool", *Proceedings of 7th International Conference on Computer Science and Information Technology (CSIT), IEEE, Amman*, pp. 1-6.

Alginahi, Y.M. (2013), "A survey on Arabic character segmentation", *International Journal on Document Analysis and Recognition (IJDAR)*, Vol. 16 No. 2, pp. 105-126.

Al-Muhtaseb, H. and Qahwaji, R. (2011), "Arabic optical character recognition: recent trends and future directions", *Applied Signal and Image Processing: Multidisciplinary Advancements*, IGI Global, pp. 324-346.

Ben Moussa, S., Zahour, A., Benabdelhafid, A. and Alimi, A.M. (2010), "New features using fractal multi-dimensions for generalized Arabic font recognition", *Pattern Recognition Letters*, Vol. 31 No. 5, pp. 361-371.

Dahi, M., Semary, N.A. and Hadhoud, M.M. (2015), "Primitive printed Arabic optical character recognition using statistical features", *the IEEE Seventh International Conference on Intelligent Computing and Information Systems (ICICIS), IEEE*, pp. 567-571.

El-Mahallawy, M.S.M. (2008), "A large scale HMM-based Omni font-written OCR system for cursive scripts", PhD thesis, Faculty of Engineering, Cairo University Giza.

Hosseini, H.M.M. (1997), *Analysis and Recognition of Persian and Arabic Handwritten Characters*, Department of Electrical and Electronic Engineering, University of Adelaide, available at: https://books.google.co.uk/books?id=iJrcNwAACAAJ

Kanai, J., Rice, S., Nagy, G. and Nartker, T. (1993), "Performance metrics for printed document understanding systems", Proceedings of the International Conference on Document Analysis and Recognition (ICDAR-93), *IEEE, Tsukuba*, pp. 424-427, available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=395703

Kanungo, T., Marton, G.A. and Bulbul, O. (1999a), "OmniPage vs. Sakhr: paired model evaluation of two Arabic OCR products", *Electronic Imaging'99*, San Jose, CA, pp. 109-120.

Kanungo, T., Marton, G.A. and Bulbul, O. (1999b), "Performance evaluation of two Arabic OCR products", *Proceedings of SPIE*, pp. 76-83, available at: http://link.aip.org/link/?PSI/3584/76/1&Agg=doi

Khorsheed, M.S. (2002), "Off-line Arabic character recognition – a review", *Pattern Analysis & Applications*, Vol. 5 No. 1, pp. 31-45.

Levenshtein, V. (1966), "Binary codes capable of correcting deletions, insertions, and reversals", *Soviet Physics Doklady*, Vol. 10 No. 8, pp. 707-710.

Luqman, H., Mahmoud, S.A. and Awaida, S. (2014), "KAFD Arabic font database", *Pattern Recognition*, Vol. 47 No. 6, pp. 2231-2240.

Margner, V. and El Abed, H. (2009), "Arabic word and text recognition – current developments", *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, MEDAR Consortium.

Mihov, S., Schulz, K.U., Ringlstetter, C., Dojchinova, V. and Nakova, V. (2005), "A corpus for comparative evaluation of OCR software and postcorrection techniques", *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, *IEEE*, pp. 162-166.

Parvez, M.T. and Mahmoud, S.A. (2013), "Offline Arabic handwritten text recognition: a survey", *ACM Computing Surveys*, Vol. 45 No. 2, pp. 1-23, available at: http://doi.acm.org/10.1145/2431211.2431222%5Cnhttp://dl.acm.org/ft_gateway.cfm?id=2431222&type=pdf

Sabbour, N. and Shafait, F. (2013), "A segmentation-free approach to Arabic and Urdu OCR", Proceeding of Document Recognition and Retrieval XX Conference, *SPIE*, Vol. 8658, pp. 1-12, available at: http://reviews.spiedigitallibrary.org/data/Conferences/SPIEP/72454/86580N.pdf (accessed 15 April 2017).

Saber, S., Ahmed, A., Elsisi, A. and Hadhoud, M. (2016), "Performance evaluation of Arabic optical character recognition engines for noisy inputs", in Gaber T., Hassanien, A., El-Bendary, N. and Dey, N. (Eds), *1st International Conference on Advanced Intelligent Systems and Informatics (AISI2015)*, November 28-30, 2015, Springer, *Beni Suef, Cham*, Vol. 407, pp. 449-459.

Sakhr Software OCR (Optical Character Recognition) (2017), available at: www.sakhr.com/index.php/en/solutions/ocr (accessed 15 April 2017).

Sami El-Dabi, S., Ramsis, R. and Kamel, A. (1990), "Arabic character recognition system: a statistical approach for recognizing cursive typewritten text", *Pattern Recognition*, Vol. 23 No. 5, pp. 485-495, available at: http://linkinghub.elsevier.com/retrieve/pii/003132039090069W (accessed 22 April 2017).

Slimane, F., Ingold, R., Kanoun, S., Alimi, A.M. and Hennebert, J. (2009), "A new Arabic printed text image database and evaluation protocols", *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, *IEEE*, pp. 946-950.

Slimane, F., Kanoun, S., Hennebert, J., Alimi, A.M. and Ingold, R. (2013), "A study on font-family and font-size recognition applied to Arabic word images at ultra-low resolution", *Pattern Recognition Letters*, Vol. 34 No. 2, pp. 209-218, available at: http://dx.doi.org/10.1016/j.patrec.2012.09.012

Teahan, W.J., Inglis, S., Cleary, J.G. and Holmes, G. (1998), "Correcting English text using PPM models", *Proceedings of the Data Compression Conference (DCC'98)*, *IEEE*, pp. 289-298.

Further reading

Saber, S., Ahmed, A. and Hadhoud, M. (2014), "Robust metrics for evaluating Arabic OCR systems", *First International Image Processing, Applications and Systems Conference (IPAS)*, *IEEE*, pp. 1-6.

Corresponding author

Mansoor Alghamdi can be contacted at: malghamdi@ut.edu.sa