

# Multiple Outputs Techniques Evaluation for Arabic Character Recognition

Zeyad Q. Al-Zaydi<sup>1</sup>, Dr. Hisham Salam<sup>2</sup>

<sup>1,2</sup>(Department of computer science, *University of Technology*, Iraq)

\*\*\*\*\*

## Abstract:

Different OCR systems generate multiple outputs for same input image. The difference between several outputs of OCR is used to select the best features from them. Multiple outputs techniques show a significant improvement in OCR accuracy for images that have low scanning resolution and noises. Arabic language often produces high OCR error rate compared to the Latin character-based language. The reason is that, the unique characteristics of Arabic language. For this reason, this paper evaluates the performance of main multiple outputs techniques. The goal of the evaluation process is to test and analysis these techniques on Arabic dataset, and chooses the best for this language. The testing dataset used in the evaluation process is large in order to make the reliability and the validity of the testing are higher. The result of experiments shows that the best performance obtained among comparative techniques is the one that used multiple scanning for same input image to produce several outputs of OCR.

**Keywords** —Multiple OCR Texts, Accuracy evaluation, Character recognition, Arabic language

\*\*\*\*\*

## I. INTRODUCTION

Extracting words, sentences, and text from image is a process called optical character recognition (OCR). Post-processing stage is used to detect and correct the errors of OCR output text [1, 2]. Multiple outputs techniques are widely used in the OCR post-processing stage. These techniques present a significant improvement in accuracy of OCR system for low scanning images [3] and noisy image [4-6]. This paper will evaluate the performance of main multiple outputs OCR techniques for Arabic language. Evaluation process is important to select the best between them for this language [3-6].

The idea of multiple outputs techniques is that if there are several OCR outputs of same input image, then it can select the best features from them [3, 5, 7]. Fig. 1 displays example on the process of multiple outputs generation of OCR using different OCRs. Assuming, the input image in Fig. 1 includes only four words, which are “Arabic language is complex”.

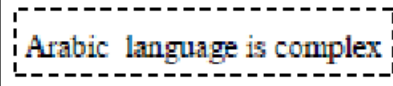
|                 |                                                                                      |
|-----------------|--------------------------------------------------------------------------------------|
| Reference Text  | Arabic language is complex                                                           |
| Image           |  |
| Output of OCR 1 | Arabic language is comaiex                                                           |
| Output of OCR 2 | Anbic language ii complex                                                            |

Fig. 1 Multiple outputs Alignment for a single input image

Fig. 1 shows several variances between two outputs of OCR texts for the a single image [5]. Therefore, a process can be implemented to choose the best features among them.

This study is arranged in five sections. Section one explained the introduction. Section two presents the comparative techniques that will be used in this study. Section three discussed the settings of evaluation process. In section four, a prototype description, results, analysis, and

discussion are explained. The last section includes the summary of this research and future work for further improvement in this research.

**II. COMPARISON TECHNIQUES**

This section of study will clarify three different techniques used in the evaluation process. It explains how existing techniques produce multiple outputs of OCR. Furthermore, it will present the strengths and weakness of each multiple outputs technique. In addition to that, the existing techniques are used by several researchers. However, each researcher used additional technique to choose the best features from resulted OCR texts.

For instance, the methods suggested by [7] and [8] performed same multiple outputs generation technique. Nevertheless, both researchers used different additional technique to select the best features between the OCR outputs text [7, 8]. Therefore, this paper will test only these three main techniques. However, it will perform a unified technique (UT) for all experiments to select the best features between multiple outputs of OCR. This is important due to the using various techniques in selecting the best will not determine which multiple outputs technique is the best. A unified technique (UT) will be explained in section 3.

**A. MOUMS**

This technique is based on idea that scanning a document several times will certainly generate various features for same input document [3]. This research will denote to this technique in this paper as MOUMS. It means multiple outputs by multiple scanning. After scanning process, a document file and their versions will become inputs to same OCR system. This will lead to generate one output for each version of document.

Characters in OCR system may be misrecognized, inserted, or deleted. For instance, the character “H” may be identified by OCR system as two characters “ll”. Another case, the two characters “uu” may be identified by OCR system as “w”. The deleted, inserted, and misrecognized characters will produce different number of words in each OCR output text [3]. Therefore, alignment process is required to handle unequal words.

Alignment process requires complex calculations. Furthermore, it causes long processing time when number of words in text is large. In addition to that, existing techniques handle alignment process approximately. In other words, the alignment process between multiple outputs of OCR can produce wrong alignment of words, and this will increase OCR error rate [8, 9]. Fig. 2 displays alignment process between multiple outputs of OCR.

|                                        |                  |
|----------------------------------------|------------------|
| Reference text                         | Arbic language   |
| OCR output1                            | Arabic lang age  |
| OCR output2                            | Aralcic language |
| Alignment between two OCR outputs text |                  |
| A r a b - i c                          | l c n g - a g e  |
| A r a l c i c                          | l a n g u a g e  |

Fig. 2: Alignment process for two OCR outputs text

After alignment process is completed, a additional technique is implemented to vote the best features among resulting OCR texts [10]. Finally, the testing dataset that used in this technique is English, and the scanning resolution of images 300 dpi.

**B. MOUMO**

This research will denote to this technique in this paper as MOUMO. It means multiple outputs by multiple systems of OCR. This technique is proposed to avoid scanning a document several times. It will pass a document to dissimilar OCR systems in order to generate a single output for each OCR system. The differences will be occurred because each OCR system has different designs, algorithms, and training datasets [6]. English dataset also used in testing this technique, Furthermore, the dataset consists from noisy images and the scanning resolution of images is 400 dpi.

The only strength of this technique is that it does not require scanning images several times. However, the weakness includes, the alignment problem, and

the process of combining different OCR systems is a complex process and time-consuming [3, 5].

**C. MOUMT**

This research will denote to this technique in this paper as MOUMT. It means multiple outputs by multiple values of threshold. This technique is proposed to avoid scanning a document several times as MOUMS, and it proposed to avoid combining different OCR systems as MOUMO. It will produce several versions of same documents by using seven values of threshold [7]. For example, it will convert a document to binary-image using “x” value of threshold, and it will convert same document to another binary-image using “y” value of threshold, and so on. The main disadvantage of this technique is that it also suffers from alignment problem [3, 7].

**III. PLAN OF EVALUATION**

This paper is based on experimental approach in conducting evaluation process. It includes three stages: experimental design, measurement, and reporting. Experimental design will explain how to conduct the experiments. A measurement will determine the metrics for measuring the performance. Reporting is related to the approach used in representing the results [11].

**A. Experimental Design**

The goal of all experiments is to find the best existing multiple outputs technique for Arabic language. This paper will compare output of four experiments in order to achieve the goal of this study. Each experiment is implemented to convert scanning documents to a text. Figs. 3 and 4 display how to perform each experiment.

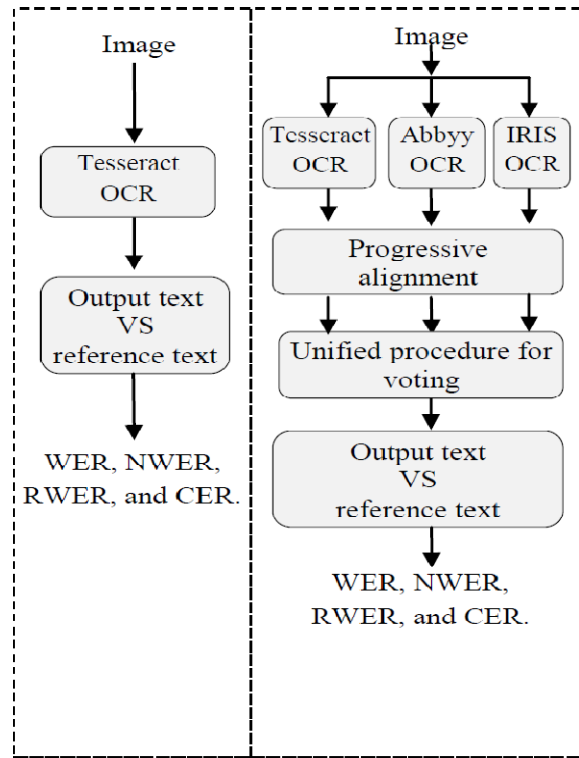


Fig. 3: Experimental one and two

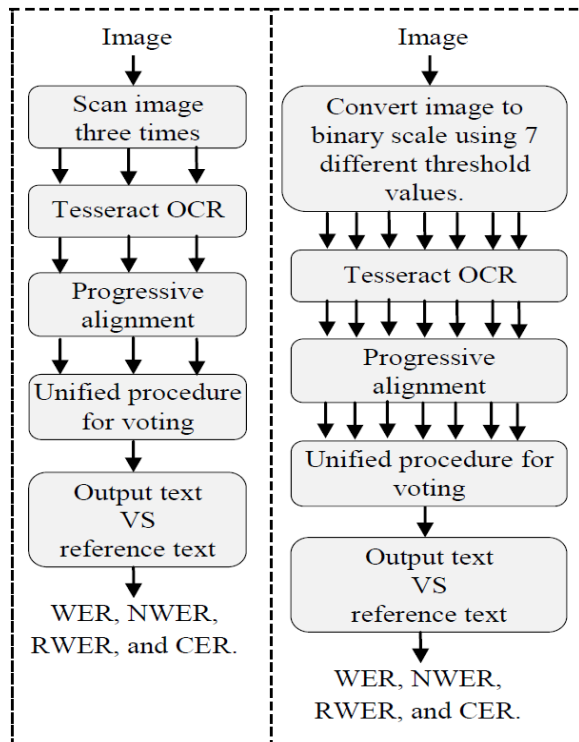


Fig. 4: Experimental three and four

Fig. 3 shows that experiment one will not use alignment operation and will not use any voting between resulted texts, while other experiments two, three, and four will perform both alignment operation and a unified technique (UT). Experiment two, three, and four will perform MOUMO, MOUMS, and MOUMT respectively.

ABBYY OCR [12], TESSERACT OCR [13], and IRIS OCR [14] will be used in experiment two. They are chosen because many researchers used them in their methods [8, 15]. Figs. 3 and 4 also show that progressive alignment algorithm will be used to align the output texts of each experiment. This algorithm is used by most researchers to align output texts of OCR [6, 7, 15]. After alignment process, UT will be performed to select the best features between resulted texts. Fig. 5 displays UP details.

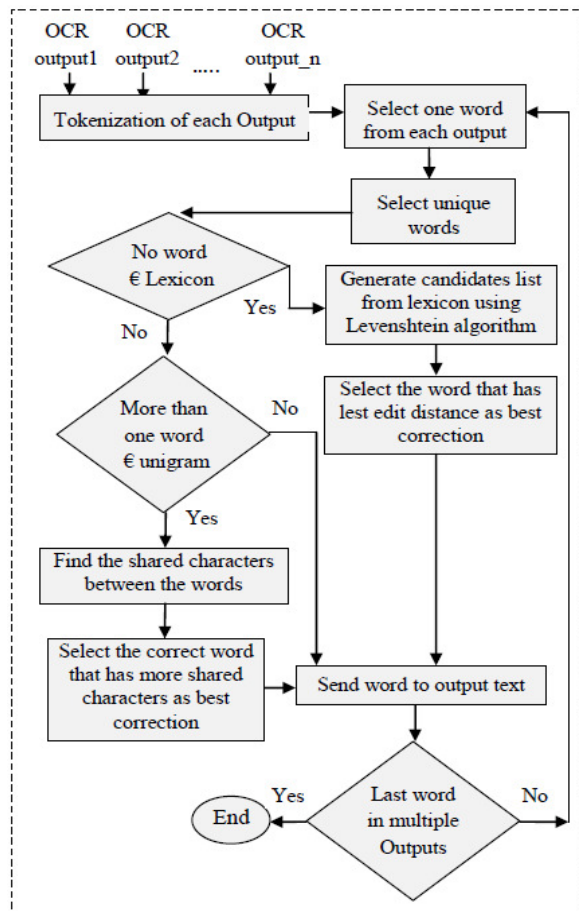


Fig. 5: UT flowchart

UT will use Levenshtein algorithm to produce candidates list for any incorrect word. This algorithm is chosen because most researchers used it in generating candidates list [16, 17]. However, it requires much processing time when number of the errors of text is high [16, 18, 19].

**B. Metrics**

Character error rate (CER), word error rate (WER), real word error rate (RWER), and non-word error rate (NWER) are the main metrics in measuring error rate of OCR [2, 5, 15]. CER counts the wrong characters in a text of OCR. Non-word error refers to any word does not find in a lexicon, while the real word error refers to any word exist in lexicon, but it is inappropriate for the phrase. Finally, wrong word error counts both NWER and RWER together [3, 5, 15, 16]. Equations one, two, three, and four can be used to measure CER, RWER, NWER, and WER [2, 3, 16].

$$WER = \frac{\text{No. of wrong words in output OCR text}}{\text{No. of all words in reference text}} * 100 \quad (1)$$

$$NWER = \frac{\text{No. of non - words in output OCR text}}{\text{No. of all words in reference text}} * 100 \quad (2)$$

$$RWER = \frac{\text{No. of real words in output OCR text}}{\text{No. of all words in reference text}} * 100 \quad (3)$$

$$CER = \frac{\text{No. of wrong characters in output OCR text}}{\text{No. of all characters in reference text}} * 100 \quad (4)$$

To calculate CER, RWER, NWER, and WER, the reference text will be aligned with OCR text using algorithm of Smith-Waterman. This algorithm can align two texts accurately [6, 20]. Next, CER can be measured by counting number of errors occurred when any letter in OCR text is not equal to the letter in reference text. WER can be measured by counting number of errors occurred when any word in OCR text is not equal to the word in reference text. Lastly, if any wrong word does not exist in a dictionary then it considers NWER, otherwise, it considers RWER.

C. Testing Images

It is difficult to find standard datasets for Arabic language that can be used in measuring accuracy of OCR [1-3, 16]. The main reason is that, most existing datasets are small in size. Therefore, the reliability and the validity of the testing will become inaccurate. Second reason is that, images of most existing datasets include only one word or one sentence in each image. This will lead to not reflect the real documents that contain large number of words. Last reason is that, most OCR methods used different Arabic datasets in measuring OCR accuracy. This is because Arabic datasets are different in types and sizes [1-3, 16].

For the previous reasons mentioned above, this paper will perform the same steps implemented by [1] to produce large testing images. These testing images have five properties. Firstly, it includes 192856 characters, commas, brackets, etc. Secondly, it is selected by chance from Arabic websites. Thirdly, it consists of 8 various fonts: Times New Roman, Simplified Arabic, Microsoft sans Serif, Tahoma, Courier new, Arial, Adobe Arabic, and Traditional Arabic. Fourthly, Fonts size of dataset is ranging from 10 to 20. To create testing images from the dataset, a text is printed. After that, the papers are scanned at 300 dpi using new scanner.

IV. RESULTS AND ANALYSIS

This section will show the results of comparative techniques. To measure the results of experiments, a prototype is designed. The test results of experiments one, two, three, and four are shown in Figs. 6, 7, 8, and 9 respectively.

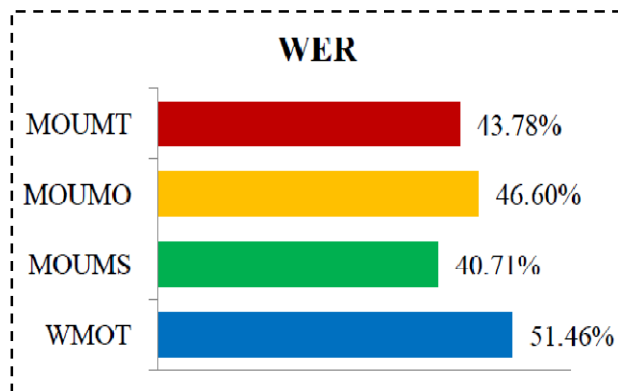


Fig. 6: WER for all experiments

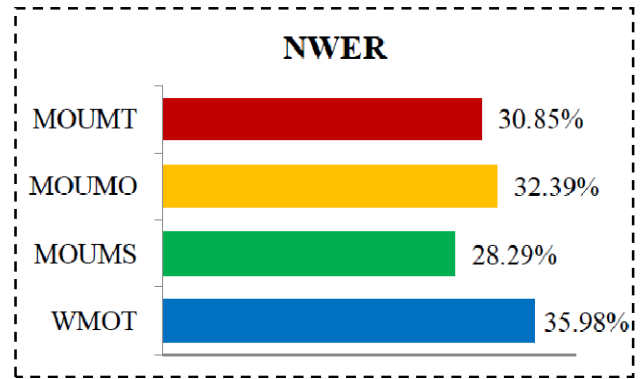


Fig. 7: RWER for all experiments

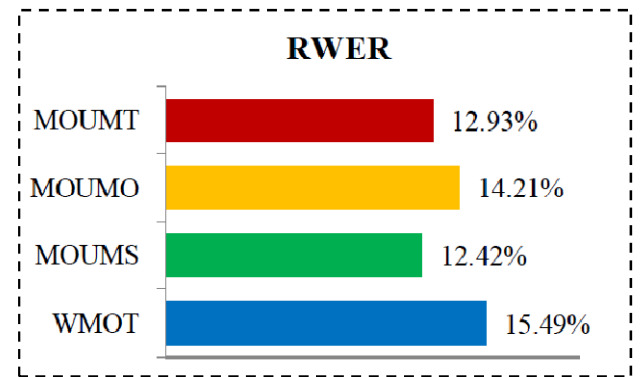


Fig. 8: RWER for all experiments

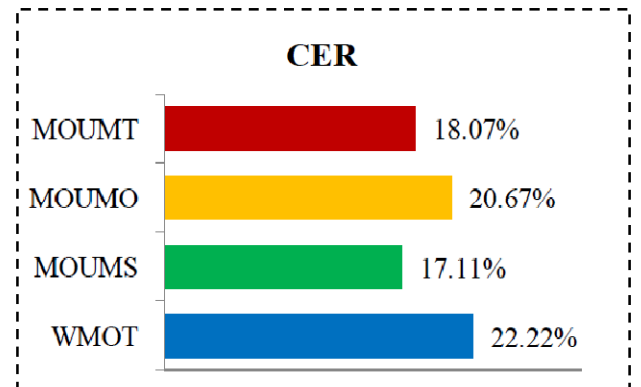


Fig. 9: CER for all experiments

Note this study will use the word “WMOT” to refer to the results of OCR process when no multiple outputs technique is used. Figs. 6, 7, 8, and 9, show that the values of the metrics WER, NWER, RWER, and CER for WMOT, MOUMS, MOUMO, and MOUMT are different from each other.

Furthermore, they show that results of WMOT had the highest values of WER, CER, NWER, and RWER, than the other experiments, with rates of 51.46%, 22.22%, 35.98%, and 15.49% respectively. This indicates that error rate for Arabic language is high.

In addition to that, it can be seen clearly that MOUMS had the least values of WER, CER, NWER, and RWER, than the other experiments, with rates of 40.71%, 17.11%, 28.29%, and 12.42% respectively. This indicates that MOUMS is the most OCR accuracy among comparative techniques. Lastly, high error rate resulted from implementing the experiments shows that techniques and methods used in OCR system need more improvement to suit the characteristics of this language. This can be achieved by combining several existing methods or techniques to benefit from their strengths together.

## V. CONCLUSION

This paper presented the performance evaluation of main techniques used in producing multiple outputs of OCR. Evaluation process implements and tests three main techniques using a large number of testing images. Number of testing images is large in order to make the reliability and the validity of the testing are higher. The testing results presented that MOUMS was the best technique among comparative techniques. In addition to that, the testing results also show that Arabic language has a high OCR error rate.

Further research can be performed to increase the OCR accuracy of the existing techniques used in generating multiple outputs of OCR. This can be done by handling the weakness of current techniques. For examples, all existing techniques that used in generating multiple outputs of OCR have alignment problem. This problem as described in section 2 increases OCR error rate. Therefore, it can suggest effective solutions to handle it.

Another potential improvement can be done by proposing and developing new techniques for generating multiple outputs of OCR that can give better features between OCR resulting texts. For examples, MOUMS suffers from time consuming due to the scanning input image several times. MOUMO suffers from complexity and time

consuming, because it needs processing each OCR system separately by a user. This will make combining them in a one auto system is difficult. Lastly, MOUMT converts image into binary-scale by assigning value of zero to all image pixels under a threshold, and assigning value of 250 to all image pixels above a threshold. This is not a strong approach to generate multiple outputs of OCR. It needs more improvement to give better features.

## REFERENCES

- [1] M. S. M. El-Mahallawy, "A large scale HMM-based omni front-written OCR system for cursive scripts," (PhD thesis, Cairo University, Cairo, Egypt), 2008.
- [2] Y. Bassil and M. Alwani, "Ocr post-processing error correction algorithm using google online spelling suggestion," *arXiv preprint arXiv:1204.0191*, 2012.
- [3] I. Q. Habeeb, S. A. Yusof, and F. B. Ahmad, "Improving Optical Character Recognition Process for Low Resolution Images," *IJACT: International Journal of Advancements in Computing Technology*, vol. 6, pp. 13 - 21, May 30 2014.
- [4] W. B. Lund and E. K. Ringger, "Error Correction with In-Domain Training Across Multiple OCR System Outputs," in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, 2011, pp. 658-662.
- [5] W. B. Lund and E. K. Ringger, "Improving optical character recognition through efficient multiple system alignment," in *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, 2009, pp. 231-240.
- [6] W. B. Lund, D. D. Walker, and E. K. Ringger, "Progressive alignment and discriminative error correction for multiple OCR engines," in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, 2011, pp. 764-768.
- [7] W. B. Lund, D. J. Kennard, and E. K. Ringger, "Why multiple document image binarizations improve OCR," presented at the Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing, Washington, District of Columbia, 2013.

- [8] W. B. Lund, D. J. Kennard, and E. K. Ringger, "Combining multiple thresholding binarization values to improve OCR output," in *IS&T/SPIE Electronic Imaging*, 2013, pp. 86580R-86580R-11.
- [9] C. Notredame, "Recent evolutions of multiple sequence alignment algorithms," *PLoS computational biology*, vol. 3, p. e123, 2007.
- [10] D. Lopresti and J. Zhou, "Using consensus sequence voting to correct OCR errors," *Computer Vision and Image Understanding*, vol. 67, pp. 39-47, 1997.
- [11] M. Ramanan, A. Ramanan, and E. Charles, "A performance comparison and post-processing error correction technique to OCRs for printed Tamil texts," in *Industrial and Information Systems (ICIIS), 2014 9th International Conference on*, 2014, pp. 1-6.
- [12] ABBYY Production LLC. (2015, January 02). *ABBYY FineReader 12 Professional*. Available: <http://www.abbyy.com/finereader/>
- [13] Google Inc. (2015, January 02). *Tesseract-ocr v3.02*. Available: <https://code.google.com/p/tesseract-ocr/>
- [14] Readiris technologies. (2015, January 10). *Readiris Pro 15*.
- [15] W. B. Lund, E. K. Ringger, and D. D. Walker, "How well does multiple OCR error correction generalize?," in *IS&T/SPIE Electronic Imaging*, 2013, pp. 90210A-90210A-13.
- [16] I. Q. Habeeb, S. A. Yusof, and F. B. Ahmad, "Two Bigrams Based Language Model for Auto Correction of Arabic OCR Errors," *JDCTA: International Journal of Digital Content Technology and its Applications*, vol. 8, pp. 72 - 80, February 28 2014.
- [17] J. F. Daðason, "Post-Correction of Icelandic OCR Text," (Master's thesis, University of Iceland, Reykjavik, Iceland), 2012.
- [18] K. U. Schulz and S. Mihov, "Fast string correction with Levenshtein automata," *International Journal on Document Analysis and Recognition*, vol. 5, pp. 67-85, 2002.
- [19] P. Mitankin, "Universal levenshtein automata. building and properties," Master's thesis, Sofia University, Bulgaria, 2005.
- [20] E. Ayguade, J. J. Navarro, and D. Jimenez-Gonzalez, "Smith-Waterman Algorithm," ed, 2007.