Performance Evaluation of Two Arabic OCR Products

Tapas Kanungo, Gregory A. Marton, Osama Bulbul

Center for Automation Research
University of Maryland
College Park, MD 20742
Email: kanungo@cfar.umd.edu
Web: http://www.cfar.umd.edu/~kanungo

ABSTRACT

Numerous Optical Character Recognition (OCR) companies claim that their products have near-perfect recognition accuracy (close to 99.9%). In practice, however, these accuracy rates are rarely achieved. Most systems break down when the input document images are highly degraded, such as scanned images of carbon-copy documents, documents printed on low-quality paper, and documents that are n-th generation photocopies. Besides, the end user cannot compare the relative performances of the products because the various accuracy results are not reported on the same dataset.

In this article we report our evaluation results for two popular Arabic OCR products: i) Sakhr OCR and ii) OmniPage for Arabic. In our evaluation we establish that the Sakhr OCR product has 15.47% lower page error rate relative to the OmniPage page error rate. The absolute page accuracy rates for Sakhr and Omnipage are 90.33% and 86.89% respectively. Our evaluation was performed using the SAIC Arabic image dataset, and we used only those pages for which both OCR systems produced output. A scatter-plot of the page accuracy-rate pairs reveals that Sakhr in general performs better on low-accuracy (degraded) pages. The scatter-plot visualization technique allows an algorithm developer to easily detect and analyze outliers in the results.

Keywords: OCR, Arabic, performance evaluation, Sakhr, OmniPage, SAIC dataset.

1. INTRODUCTION

Characterizing the performance of OCR systems is important for many reasons:

- Predict performance: Typically OCR is part of a bigger system, e.g., an information retrieval (IR) system or a machine translation (MT) system. Since the overall performance depends on the performances of the individual subsystems, the overall performance of the MT/IR system is a function of the OCR recognition rate. Knowledge of end-to-end performance as a function of OCR accuracy rate will allow us to predict the minimum recognition rate required for achieving a specified overall MT/IR system performance rate.
- Monitor progress: In order to monitor progress in research/development of OCR systems, we need quantitative measures. Periodic quantitative performance evaluation of OCR systems will allow us to assess progress in the field.
- Provide scientific explanations: Understand the contributions to the accuracy improvement by specific sub-modules. That is, explain *why* an OCR system achieves a particular accuracy.
- Identify open problems: Determine areas that need improvement/research and the impact of these improvements on the entire system.

Numerous OCR companies claim that their products have near-perfect recognition accuracy (close to 99.9%). In practice, however, these accuracy rates are rarely achieved. Most systems break down when the input document images are highly degraded, such as scanned images of carbon-copy documents, documents printed on low-quality paper, and documents that are n-th generation photocopies. Besides, the end user cannot compare the relative performances of the products because the various accuracy results are not reported on the same dataset.

In this article we report our evaluation results for two most commonly used Arabic OCR products: i) Sakhr OCR and ii) OmniPage for Arabic. In Section 2 we describe various methods for evaluating performance of OCR systems, and OCR evaluation results that have been reported earlier. In Section 3 we describe the experimental protocol we used to conduct our evaluation and in Section 4 we discuss the results.

2. PERFORMANCE EVALUATION BACKGROUND

OCR evaluation can be broadly categorized into two types: i) blackbox evaluation and ii) whitebox evaluation. In blackbox evaluation an entire OCR system is treated as an indivisible unit and its end-to-end performance of the system is characterized. The performance of the system is evaluated as follows. First a corpus of scanned document images is selected. Next, the text zones are delineated. Then, for each text zone, the correct text string is keyed in by humans. The process of delineating the zones and keying in the text is very laborious, expensive, and prone to errors. Finally the OCR algorithm is run on each text zone and the results are compared with the keyed in groundtruth text using a string matching routine. In theory the corpus should be a representative sample of the population of images for which the algorithm was designed. In practice, however, factors like time and cost forces us to limit the size of the dataset to something feasible. This process was adopted by the UNLV OCR evaluation program[1] and the UW evaluation process.[2] The UNLV evaluation corpus consisted of English annual reports, documents from department of Energy, magazines, business letters, legal documents, Spanish newspapers, and German business letters. The UW dataset[4] consisted of English technical journals.

Whitebox evaluation, on the other hand, characterizes the performance of individual submodules. Most OCR systems have submodules for skew detection and correction, page segmentation, zone classification, and text extraction. Zone segmentation evaluation has been attempted earlier by Vincent *et al.*[6,7] Whitebox evaluation is possible only if the evaluator has access to the input and output of the submodules of the OCR system. Thus for segmentation evaluation, access to coordinates of zones produced by OCR is crucial. While blackbox evaluation does not require access to intermediate results, it does not provide performance analysis at the submodule level. Furthermore, the blackbox evaluations described above do not take into account the errors due to segmentation.

More recently, researchers have advocated the use of synthetically generated data for OCR evaluation. In this methodology (see Kanungo *et al.*[8,9]) documents are first typeset using a standard typesetting system such as LaTeX or Word. Then a noise-free bitmap image of the document and the corresponding groundtruth is automatically generated. The noise-free bitmap is then degraded using a parametrized degradation model.[10,8,9] The degradation level is controlled by varying the parameters of the model. This methodology has the advantage that the laborious process of manually typing in the data is completely avoided. Furthermore, no manual scanning is required, and the process is entirely independent of language (up to the limits of the typesetting software). Since the typesetting software is available to us, the effects of page layout, font size and type on OCR accuracy can be studied by conducting controlled experiments. A variant of the above methodology proposed by Kanungo and Haralick[11,9] by printing the ideal document, scanning it, and then transforming the ideal groundtruth to match the real image. This process allows a researcher to generate groundtruth at a geometric level (character bounding boxes, identity, font, etc.) in any language, which is essential for building classifiers.

In this article, we conduct a blackbox evaluation of two Arabic OCR products. In the next section we describe the details of our experimental setup, and in Section 4 we summarize our results.

3. EXPERIMENTAL PROTOCOL

In this section we describe the experimental setup. We selected the SAIC dataset as our corpus for performance evaluation. The corpus has binary images of Arabic text and the corresponding "groundtruth." By groundtruth we mean manually typed correct Arabic ASCII strings that OCR systems should ideally produce. We then run both OCR products on the dataset and compute the accuracy rate of the OCR engines, which is defined as the percentage of groundtruth characters correctly recognized, by comparing the outputs with the groundtruth.

The two Arabic OCR products that were i) Sakhr's Automatic Reader 3.01 and ii) Caere's OmniPage Pro v2.0. Both products were run on a Pentium 166MHz PC with 32Mb RAM, 256Kb cache, and running Microsoft Windows 95 (Arabic version). The DOD error counter was used for counting errors in the OCR-generated text; the software was run a Sun Ultra 2 running Solaris 5.5. On UNIX, AraMosaic – a public-domain Arabic browser – was used for viewing the OCR-generated text. In order to reduce manual errors, scripts were written to automate the process as much as possible.

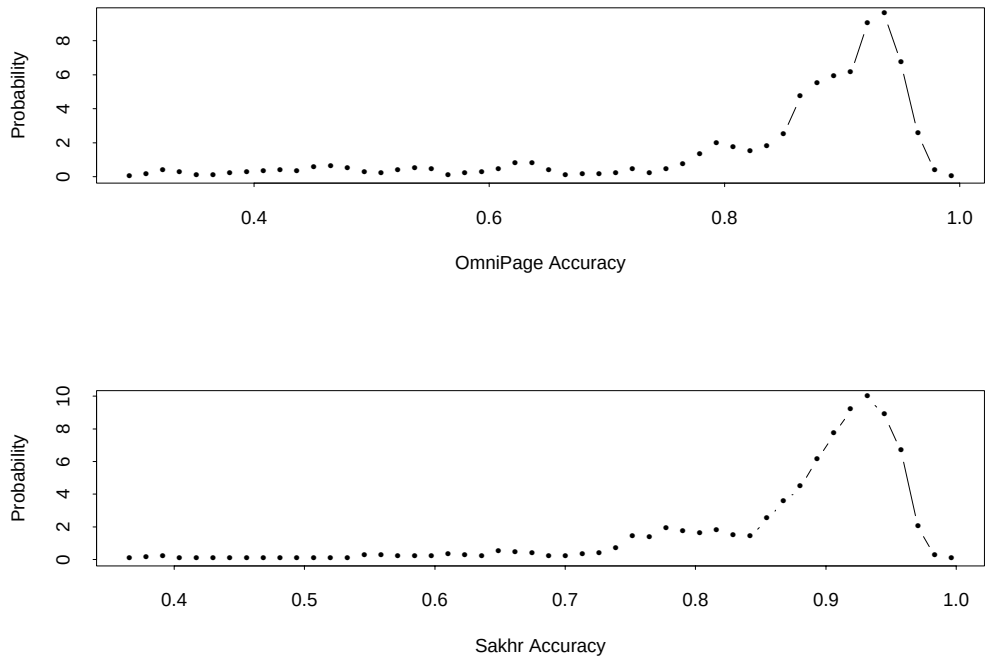


Figure 1. The first plot is the distribution of page accuracies of OmniPage for images at 300 dpi. The second plot is the corresponding distribution of Sakhr page accuracies. Notice that the accuracies are not distributed as Gaussians.

The Department of Defense provided us with the DARPA/SAIC dataset.[5] It originally contained 345 images with groundtruth. Three of these were unusable and were removed (ATI0746 did not have image, ATI0116 did not have groundtruth, and ATI0286 image and groundtruth did not match), leaving 342 images with groundtruth. Groundtruth text was encoded in CP1256 format. TIFF images, originally at 600 dpi, were then sampled at 300 dpi using the public-domain utility `convert`. Images in the DARPA/SAIC dataset are zones with single column of text. The images are relatively clean and are scanned from books, magazines and computer generated documents.

4. RESULTS

In our evaluation we computed the page accuracy rate, which is defined as the average page accuracy rate. Sakhr achieved 90.33% accuracy whereas OmniPage achieved 86.89% accuracy. The 95% confidence interval for mean accuracy of Sakhr is 90.33 ± 0.9 and that of OmniPage is 86.89 ± 1.54. The absolute page accuracy of Sakhr is on the average 3.44% higher than that of OmniPage. The 95% confidence interval on the difference between the two means is 3.44 ± 1.13%. In relative terms, the page error rate (1 minus page accuracy rate) of Sakhr is 26% lower relative to the page error rate of OmniPage.

A histogram of the accuracies is shown in Figure 1. It can be seen that the empirical distribution of the accuracies is not Gaussian. In fact, the accuracy distribution of OmniPage has a fatter tail than that of Sakhr. A scatter plot of accuracy pairs for Sakhr and OmniPage is shown in Figure 2. Each point on the plot corresponds to a document image in the dataset. The x-coordinate corresponds to the OmniPage accuracy for that image and the y-coordinate corresponds to the Sakhr accuracy. Points on the diagonal represent document images for which both products achieved similar accuracies. Points very far from the diagonal represent images for which the accuracies differed a lot. It can be seen that there are many images for which Sakhr performed better. Numerous subimages from the dataset images, and the corresponding OCR output for both products, are shown in Figures 3-9.

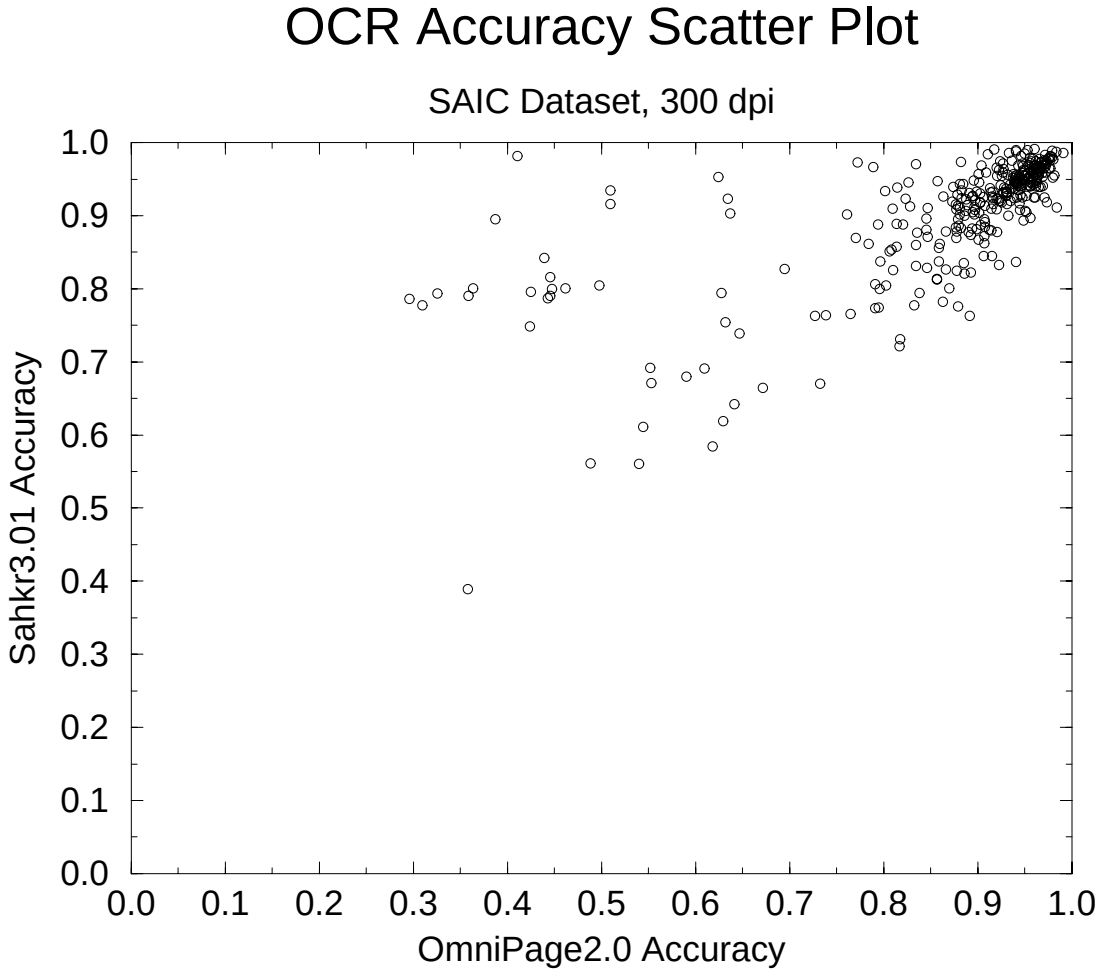


Figure 2. Scatter plot of OCR accuracies of OmniPage and Sakhr at 300 dpi resolution. Each data point represents a specific image. The x-coordinate of the data point represents OmniPage accuracy, whereas the y-coordinate of the data point represents Sakhr accuracy. Points along the diagonal represent document images for which both products achieved similar accuracy. Off-diagonal points indicate that one product performed better than the other.

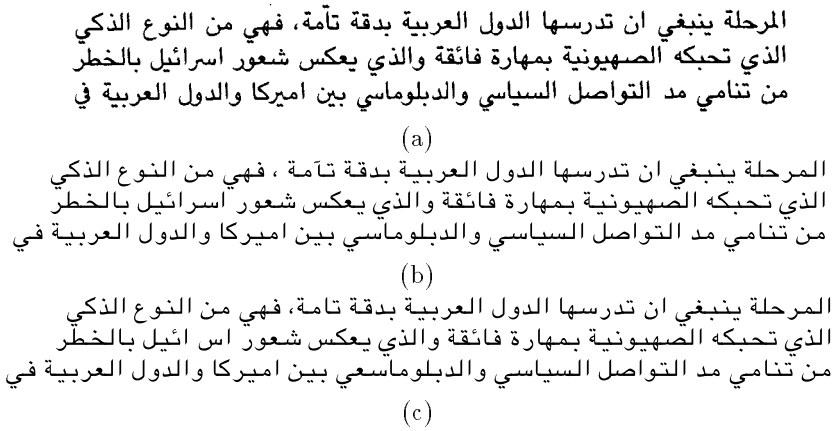
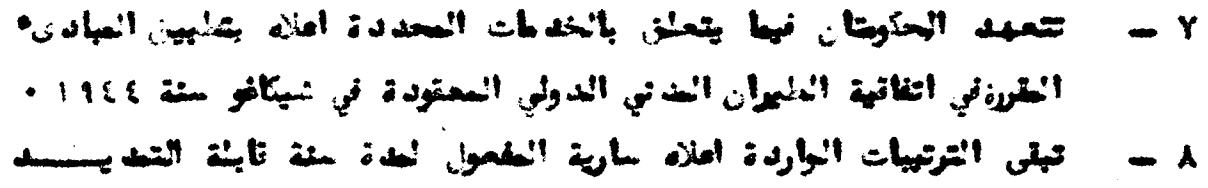


Figure 3. Image ATI0290. Both Sakhr and OmniPage performed well on this image. Sakhr achieved 98.08% accuracy and OmniPage achieved 97.7% accuracy.



(a)

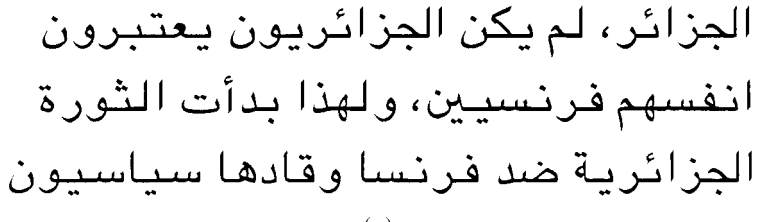
م +؛ م ,0م ؛ م + د م , 0م 0 0م م ؛ 0م م ؛ م م
سزتسر ؛ سم م ث كن م ؛ م د؛ ء ؛مم ز كل
)1نوو من في ا محه ؛ انترت كبلى تث لاً 4نمعر مز لأرمن

(b)

شضصمه "1حلأمت نحا ثه!ا كلتعئض ، بانحي هاشه 5ئ !!ددج اتي !مم!ئي تم ث ان!!هاس ى ه
انطي ؟ير5ظ قهة ؟؟ظئهض ض ،نط ت اى!دوليئ ائمعش دج مض مثمشاش صشة 944؟ .
س اترشط شي ؟نصالم دة5علأ، طالة *نممعو(صه ه ش!دة سشى كمايئة 9ش!!ه كله

(c)

Figure 4. (a) Image ATI0082. Both Sakhr and OmniPage performed poorly on this image. Sakhr achieved 38.88% accuracy and OmniPage achieved 35.79% accuracy.(b) Output of OmniPage. (c) Ouput of Sakhr.



(a)

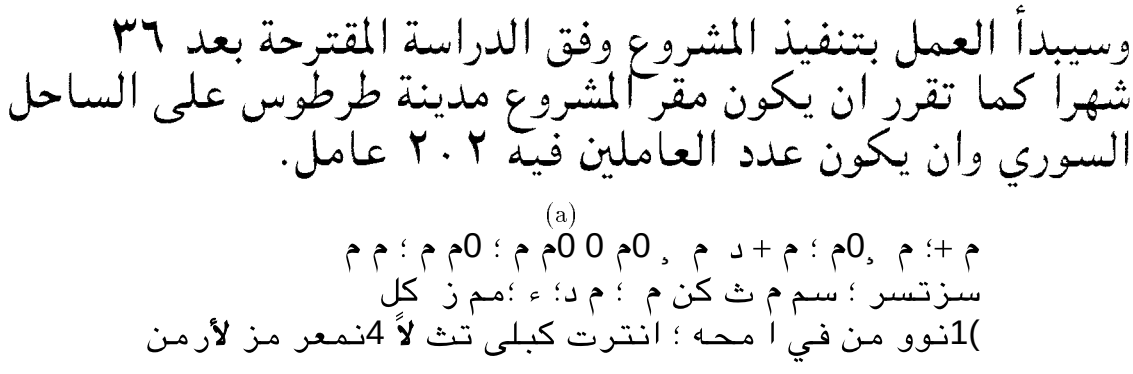
! ل نل
نفسم في نسيين ، و بد رة
ية ضد في و ن

(b)

الجزائر، لم يكن الجزائريون يعتبرون
انفسهم فرنسييت ، ولهذا بدأت ا لثورة
الجزائرية ضد فرثسا وقادها سياسيون

(c)

Figure 5. (a) Image ATI0012. Sakhr performed better than OmniPage on this image. Sakhr achieved 98.13% accuracy whereas OmniPage achieved 41.09% accuracy. (b) Output of OmniPage. (c) Ouput of Sakhr.

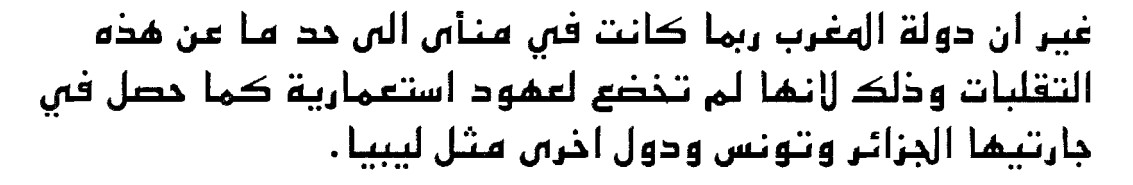


(a)

(b)

(c)

Figure 6. (a) Image ATI0239. Sakhr performed better than OmniPage on this image. Sakhr achieved 89.52% accuracy whereas OmniPage achieved 38.75% accuracy. (b) Output of OmniPage. (c) Ouput of Sakhr.

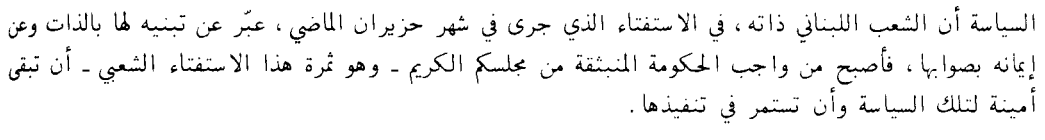


(a)

(b)

(c)

Figure 7. a) Image ATI0285. Sakhr performed better than OmniPage on this image. Sakhr achieved 78.61% accuracy whereas OmniPage achieved 29.62% accuracy. (b) Output of OmniPage. (c) Ouput of Sakhr.


السياسة أن الشعب اللبناني ذاته، في الاستفتاء الذي جرى في شهر حزيران الماضي، عبّر عن تبنيه لها بالذات وعن إيمانه بصوابها، فأصبح من واجب الحكومة المنبثقة من مجلسكم الكريم - وهو ثمرة هذا الاستفتاء الشعبي - أن تبقى أمينة لتلك السياسة وأن تستمر في تنفيذها.

(a)

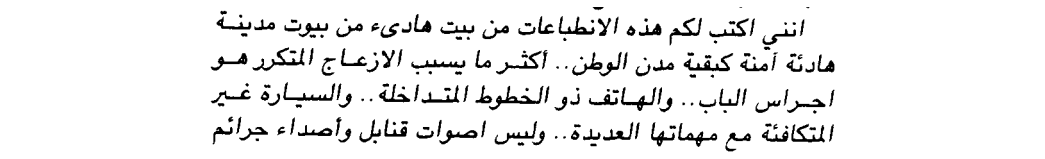
اليامة أن الشعب اللبناني ذاته ، في الامتفتا ,لذي جرى في شهر حزيران الماني ، عبر عن تبنيه لها بالذات وي
إ يأنه بعوام 1، فأسبح ء .واجب المكومة المنبثقة من مجللا الكرم _وهو ثرة هذا الاستفتاء الشعبي _أن تبق
أمينة لتلك اليات وأن تتمو فى تنفيذها .

(b)

السماسة أر السصط اللسايى ذاته ، ي الاسسصاء الدي جرى ى شهر حريراد المامي ، ككمر س تبسبه لها بالذات وكل إبمأنه لصواكا ، لاصبح ص ط واجط ألحكومة المشمة س محلسم الكريم -وهو ثمرة هدا الاسسمعاء ال!عى -أن تبقى أمسة لمللث السماسة وأر نسممر ي تمدها .

(c)

Figure 8. (a) Image ATI0078. OmniPage performed better than Sahkr on this image. OmniPage achieved 89.14% accuracy whereas Sakhr achieved 76.3% accuracy. (b) Output of OmniPage. (c) Ouput of Sakhr.


انني اكتب لكم هذه الانطباعات من بيت هادىء من بيوت مدينة هادئة أمنة كبقية مدن الوطن.. أكثر ما يسبب الازعاج المتكرر هو اجراس الباب.. والهاتف ذو الخطوط المتداخلة.. والسيارة غير المتكافئة مع مهماتها العديدة.. وليس اصوات قنابل وأصداء جرائم

(a)

انني اكتب لكم هذ؛ الانطباعات من بيت هادىء من بيوت مدينة هادئة أمنة كبقية مدن الوطن . .أكثرها يسبب الازعاج المتكررهو اجراس الباب . .والهاتف ذوالخطوط المتداخلة . .والسيارة غير المتكافئة مع مهماتها العديدة . .وليس اصوات قنابل وأصداء جرائم

(b)

لم ننني /كتب لكم هده لم لانطباعات من بلت مادىء من بليرت مدلية هادئة لم منة كبقية مدن لم لوطن . .لم كثرما لشبب /الازعاج لم لتكررهو /جرلم س لم لباب . .رلم لهاتف ذو لم لخطوط /التداخلة ..و/السيارة غير لم لمتكافئة مع مهاتها ،لعدب ة ..بميم لم صؤت قنابل ولم صدلمء جزئم

(c)

Figure 9. (a) Image ATI0446. OmniPage performed better than Sahkr on this image. OmniPage achieved 94.06% accuracy whereas Sakhr achieved 83.67% accuracy. (b) Output of OmniPage. (c) Ouput of Sakhr.

5. SUMMARY

We reported on our evaluation results on two Arabic OCR products: Sakhr and OmniPage. We showed that on the SAIC dataset Sakhr performed better than OmniPage. The average page accuracy rate of Sakhr is 90.333% while that of OmniPage is 86.89%. The average page accuracy of Sakhr is $3.44 \pm 1.13\%$ higher than that of OmniPage. In relative terms, Sakhr has 26% lower page error rate (defined as 1 minus page accuracy rate) than OmniPage. A scatter plot is used to visualize the page accuracies. This visualization technique allows an algorithm developer to easily detect and analyze outliers.

6. ACKNOWLEDGEMENT

We would like to DOD for providing us with the OCR error counting software. This research is supported in part by Army Research Lab (ARL 01-5-29294) and the Department of Defense (DOD 01-5-29177).

REFERENCES

1. S. V. Rice, F. R. Jenkins, and T. A. Nartker, "The fifth annual test of OCR accuracy," Tech. Rep. TR-96-01, Information Science Research Institute, University of Nevada, Las Vegas, NV, 1996.
2. S. Chen, S. Subramaniam, and R. M. H. I. T. Phillips, "Performance evaluation of two ocr systems," in *Proc. of Annual Symp. on Document Analysis and Information Retrieval*, pp. 299–317, April 1994.
3. T. Kanungo, G. Marton, and O. Bulbul, "OmniPage vs. Sakhr: Paired model evaluation of two Arabic OCR products," in *Proc. of SPIE Conf. on Document Recognition and Retrieval VI*, D. Lopresti and Y. Zhou, eds., (San Jose, CA), 1999.
4. R. M. Haralick, I. Phillips, *et al.*, "UW-CDROM-I."
5. R. Davidson and R. Hopely, "Arabic and persian OCR training and test data sets," in *Proc. of Symp. on Document Image Understanding Technology*, April 30 – May 2 1997.
6. B. A. Yanikoglu and L. Vincent, "Pink panther: a complete environment for ground-truthing and benchmarking document page segmentation," *Pattern Recognition* **31**, pp. 1191–1204, September 1998.
7. S. Randriamasy and L. Vincent, "Benchmarking page segmentation algorithms," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, June 1994.
8. T. Kanungo, R. M. Haralick, and I. Phillips, "Non-linear local and global document degradation models," *Int. Journal of Imaging Systems and Technology* **5**(4), 1994.
9. T. Kanungo, *Document Degradation Models and a Methodology for Degradation Model Validation*. PhD thesis, University of Washington, Seattle, WA., 1996. http://www.cfar.umd.edu/ kanungo/pubs/phdthesis.ps.Z.
10. H. S. Baird, "Document image defect models," in *Structured Document Image Analysis*, Springer-Verlag, New York, 1992.
11. T. Kanungo and R. M. Haralick, "An automatic closed-loop methodology for generating character groundtruth for scanned images," *IEEE Transactions on Pattern Analysis and Machine Intelligence* , December 1998.
12. R. M. Haralick, I. Phillips, *et al.*, "U.W. English Database I," 1994.
13. T. Kanungo and P. Resnik, "The bible, truth, and multilingual ocr evaluation," in *Proc. of SPIE Conf. on Document Recognition and Retrieval VI*, D. Lopresti and Y. Zhou, eds., (San Jose, CA), 1999.