OmniPage vs. Sakhr: Paired Model 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

Characterizing the performance of Optical Character Recognition (OCR) systems is crucial for monitoring technical progress, predicting OCR performance, providing scientific explanations for the system behavior and identifying open problems. While research has been done in the past to compare performances of two or more OCR systems, all assume that the accuracies achieved on individual documents in a dataset are independent when, in fact, they are not. In this paper we show that accuracies reported on any dataset are correlated and invoke the appropriate statistical technique — the paired model — to compare the accuracies of two recognition systems. Theoretically we show that this method provides tighter confidence intervals than methods used in OCR and computer vision literature. We also propose a new visualization method, which we call the accuracy scatter plot, for providing a visual summary of performance results. This method summarizes the accuracy comparisons on the entire corpus while simultaneously allowing the researcher to visually compare the performances on individual document images. Finally, we report on the accuracy and speed performances as a function of scanning resolution. Contrary to what one might expect, the performance of one of the systems degrades when the image resolution is increased beyond 300 dpi. Furthermore, the average time taken to OCR a document image, after increasing almost linearly as a function of resolution, suddenly becomes a constant beyond 400 dpi. This behavior is most likely because the OCR algorithm samples the images at resolutions 400 dpi and higher to a standard resolution. The two products that we compare are the Arabic OmniPage 2.0 and the Automatic Page Reader 3.01 from Sakhr. The SAIC Arabic dataset was used for the evaluations. The statistical and visualization methods presented in this article are very general and can be used for comparing accuracies of any two recognition systems, not just OCR systems.

Keywords: OCR, paired models, confidence intervals, Arabic, performance evaluation, Sakhr, OmniPage, SAIC dataset.

1. INTRODUCTION

Performance evaluation and characterization of OCR systems is crucial for many reasons: i) 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 allows us to predict the minimum recognition rate required for achieving a specified overall MT/IR system performance rate. ii) In order to monitor progress in research/development of OCR systems, we need quantitative measures. Periodic quantitative performance evaluation of OCR systems allows us to assess progress in the field. iii) To scientifically understand the contributions to the accuracy improvement by specific submodules. That is, explain *why* an OCR system achieves a particular accuracy. iv) To determine areas that need improvement/research and the impact of these improvements on the entire system.

In this article we report our evaluation results for two most commonly used Arabic OCR products: i) Sakhr Automatic Reader version 3.01 and ii) OmniPage for Arabic version 2.0 from Shonut. We start the discussion by providing a background of OCR evaluation literature in Section 2. Metrics for quantifying errors are discussed in Section 3. In Section 4 we provide the statistical theory of paired models, which we use to compare the performance of the two Arabic OCR systems. In Section 5 we describe the experimental protocol we use to conduct our evaluation and in Section 6 we discuss our 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 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[3] 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.*[4,5] 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.*[6,7]) 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.[8,6,7] 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[9,7] 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 metrics we use for evaluating the OCR systems, and in Section 4 we describe the statistical techniques we use for comparing measurements.

3. METRICS FOR PERFORMANCE EVALUATION

What metrics are good for evaluating OCR results? In this section we describe a few metrics that we consider important and give advantages and disadvantages of using them. Let O represent the number of symbols in the OCR-generated text, M the number of correctly recognized symbols, D the number of symbols deleted, I the number of symbols inserted, S the number of symbols in the groundtruth replaced by another symbol, and T the number of groundtruth symbols.

Accuracy: The number of symbols correctly recognized on a page normalized by the total number of symbols in the groundtruth. Thus accuracy is M/T. This is also called *recall* in the information retrieval (IR) literature. Notice that this number does not reflect the number of extraneous symbols that get introduced.

Precision: This is the number of symbols correctly recognized on a page normalized by the number of symbols in the OCR-generated text. Thus precision is M/O. If two systems have the same accuracy but one has higher precision than the other, the system with higher precision generates fewer extraneous symbols.

Insertion: The number of symbols inserted normalized by the number of groundtruth symbols on the page: I/T.

Deletion: This is the number of symbols deleted normalized by the number of groundtruth symbols on the page: D/T.

Substitution: The number of symbols substituted normalized by the number of groundtruth symbols on the page: S/T.

The above character-level metrics were computed using the DOD error counter, which is based on a dstring matching routine. In this article we have not reported the above metrics on using words. We are currently in the process of computing the metrics using words. While character level metrics are useful for predicting improvements in information retrieval systems based on OCR-generated text, word metrics are better for judging improvements in i) ease of human readability and manual correction, and ii) machine translation that accept OCR-generated text as input.

4. STATISTICAL COMPARISON OF SAMPLE MEANS

If a computed metric for one OCR algorithm is better than that for another, is the result statistically significant? In this section we describe the theory behind statistical comparisons of measurements. One of the problems we encounter while comparing OCR results of two algorithms or products is that of comparing means of two samples, which are obtained by running the two algorithms on a dataset. The underlying accuracy populations are typically not distributed as Gaussians and making such an assumption is not justified. However, one can assume that a dataset is large. Large for the current discussion is means greater than 30. There are certain statistical techniques that can be used with these basic assumptions for comparing accuracies. We now describe the techniques. Please refer to Arnold[11] for details.

4.1. Large sample inference about means

Let $x_1, x_2, \ldots, x_n$ be the set of OCR accuracy numbers that are obtained by processing n document images. Let the underlying distribution of the accuracies have a mean μ and variance σ^2. Let $\bar{x}$ and S^2 be the sample mean and variance. An unbiased estimator for the population mean μ is the sample mean $\bar{x}$, and

$$E[\hat{\mu}] = E[\bar{x}] = \mu, \tag{1}$$

$$Var[\hat{\mu}] = \frac{\sigma^2}{n}. \tag{2}$$

These results hold because for large samples ($n > 30$), the distribution of the mean asymptotically gets close to the Gaussian distribution:

$$\frac{n^{1/2}(\bar{x} - \mu)}{S} \sim N(0, 1).$$

This fact is due to the Central Limit Theorem and can be used to construct a confidence interval for the estimated mean:

$$\mu \in \bar{x} \pm \frac{z^{\alpha/2} S}{\sqrt{n}},$$

where $z^{\alpha/2}$ is the error function — $P(z > z^{\alpha/2} | z \sim N(0,1)) = \alpha$, and α is the significance level.

4.2. Inference about means of two independent samples

Let $x_1, x_2, \ldots, x_m$ be a sample of OCR accuracy numbers that are obtained by processing m document images. Let $y_1, y_2, \ldots, y_n$ be another *independent* sample of accuracies. Let $\bar{x}$ and S^2 be the sample mean and variance of x_i and $\bar{y}$ and T^2 be the sample mean and variance of y. Let the underlying x population have a mean μ and variance σ^2, and the y population have a mean ν and variance τ^2. We are interested in drawing conclusions about the difference between the means $\delta = \mu - \tau$. An estimator of δ is the difference between the sample means:

$$\hat{\delta} = \bar{x} - \bar{y}.$$

Then,

$$E[\hat{\delta}] = \mu - \nu, \tag{3}$$

$$Var[\hat{\delta}] = \frac{\sigma^2}{m} + \frac{\tau^2}{n}. \tag{4}$$

As in the previous subsection, the confidence interval for the estimated difference in means δ is

$$\delta \in \hat{\delta} \pm z^{\alpha/2} \left(\frac{S^2}{m} + \frac{T^2}{n} \right)^{1/2}.$$

4.3. Paired model inference about difference in means

Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, be n correlated pairs of OCR accuracy values such that $E[x_i] = \mu$, $E[y_i] = \nu$, $Var(x_i) = \sigma^2$, $Var(y_i) = \tau^2$, and $cov(x_i, y_i) = \rho\sigma\tau$. This correlation occurs because x_i and y_i are the accuracies on the same document image. Results in the previous subsection required the two samples to be independent and so results in that section cannot be used. We proceed by constructing a new variable $u_i = x_i - y_i$, with sample mean $\bar{u}$ and sample variance V^2. We are again interested in drawing inferences about $\delta = \mu - \nu$. An estimator for δ is $\bar{u}$. Thus,

$$E[\hat{\delta}] = E[\bar{u}] = E[\bar{x} - \bar{y}] = \mu - \nu, \tag{5}$$

$$Var[\hat{\delta}] = \frac{\sigma^2 + \tau^2 - 2\rho\tau\sigma}{n}. \tag{6}$$

The confidence interval for the estimated difference of means $\hat{\delta}$ is given by

$$\delta \in \hat{\delta} \pm \frac{z^{\alpha/2} V}{\sqrt{n}}.$$

4.4. Discussion

In the previous sections we saw that the expected values of both the paired as well as unpaired estimators are equal to the difference between the population means. The variances, however, differ. The paired estimator variance is $(\sigma^2 + \tau^2 - 2\rho\tau\sigma)/n$ while the unpaired estimator has a variance equal to $(\sigma^2 + \tau^2)/n$. Thus the paired estimator is better since its variance is smaller and uncertainty is lower for the same sample size. The paired estimator uses the correlation information to reduce the uncertainity in its estimate.

5. 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) Shonut's OmniPage Pro v2.0. Both products were run on a Pentium 400 PC with 128 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.

The Department of Defense provided us with the DARPA/SAIC dataset.[12] 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 100, 200, 300 and 400 dpi using the public-domain utility `convert`. Images in the DARPA/SAIC dataset are zones with a single column of text. Subimages from these images are shown in Figures 6-11. The images are relatively clean and are scanned from books, magazines and computer generated documents.

6. RESULTS

In our evaluation we computed the page accuracy rate, which is defined as the average page accuracy rate. At 300 dpi 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%. However, although Sakhr has higher accuracy, OmniPage has higher precision. In Table 2 we see that at 300dpi OmniPage has 0.9917% ± 0.4672 higher precision than Sakhr.

Histograms of the accuracies of the two products at 300 dpi are shown in Figure 1. It can be seen that the empirical distribution of the accuracies is not Gaussian and the accuracy distribution of OmniPage has a fatter tail than that of Sakhr. A scatter plot of accuracy pairs for Sakhr and OmniPage at 100, 200, 300, is shown in Figure 2(a)-(c). 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 at 300 dpi there are many images for which Sakhr performed better. A scatter plot of Sakhr at 300dpi and 600dpi is shown in Figure 2(c). It can be seen that the Sakhr algorithm performs worse at 600dpi than at 300dpi.

In Figure 5 the average time taken to OCR an image is plotted. It can be seen that Sakhr's time does not increase when the resolution is increased from 400 dpi to 600 dpi. This is probably because the algorithm first samples the image to a standard resolution and them does the OCR processing. Numerous subimages from the dataset images, and the corresponding OCR output at 300 dpi for both products, are shown in Figures 6-11.

7. SUMMARY

We have shown that paired model approach to performance comparison gives rise to tighter confidence intervals than unpaired methods when computing difference in OCR accuracies. We have used this methodology to evaluate two Arabic OCR products: Sakhr Automatic Reader 3.0 and OmniPage 2.0. We have shown that on the 300 dpi SAIC dataset Sakhr has higher accuracy than OmniPage but OmniPage has a better precision. 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. But at 300 dpi OmniPage has 0.9917 ± 0.4672 higher precision than Sakhr. We also characterized the accuracy, precision, and error as a function of resolution and noted that accuracy of Sakhr drops when the image resolution is increased beyound 300 dpi. Furthermore, the average time taken for Sakhr to OCR a page does not increase when the image resolution is increased from 400 dpi to 600 dpi. This could be because the algorithm might be sampling the images to a standard resolution prior to the OCR process. A scatter plot is used to visualize the page accuracies. This visual summarization technique allows an algorithm developer to easily detect and analyze outliers.

8. ACKNOWLEDGEMENT

We would like to thank the Department of Defence for providing us with the OCR error counting software. This research is supported in part by the 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. R. M. Haralick, I. Phillips, *et al.*, "UW-CDROM-I."
4. 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.
5. S. Randriamasy and L. Vincent, "Benchmarking page segmentation algorithms," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, June 1994.

6. 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.
7. 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.
8. H. S. Baird, "Document image defect models," in *Structured Document Image Analysis*, Springer-Verlag, New York, 1992.
9. 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.
10. 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.
11. S. F. Arnold, *Mathematical Statistics*, Prentice-Hall, New Jersey, 1990.
12. 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.

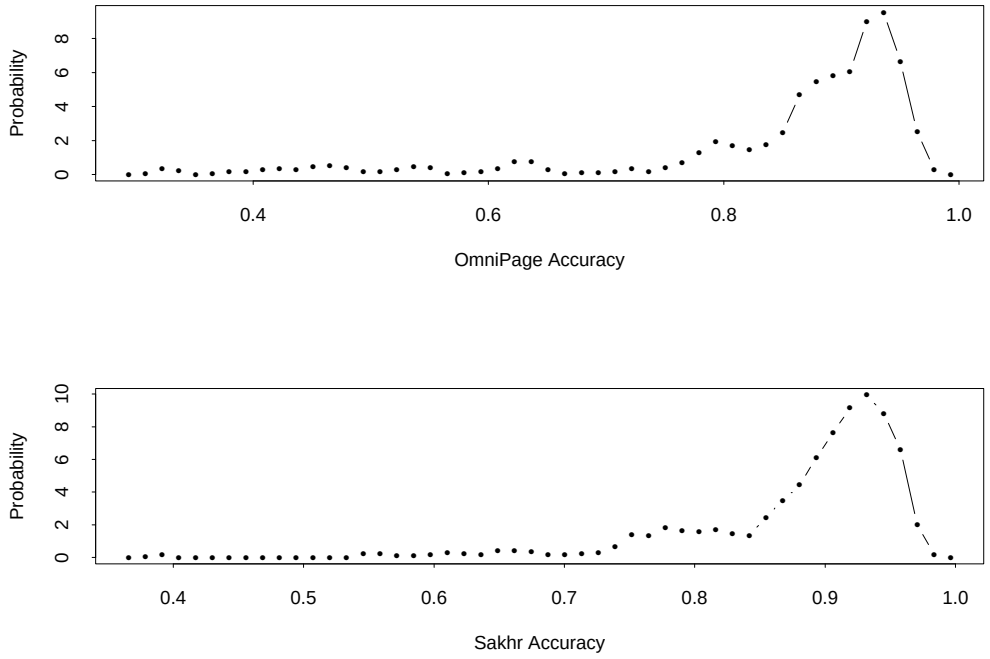


Figure 1. The first plot is the distribution of page accuracies of OmniPage for images at 300 dpi. The second plot is the corresponding distribution of Sakhr page accuracies. Notice that the accuracies are not distributed as Gaussians.

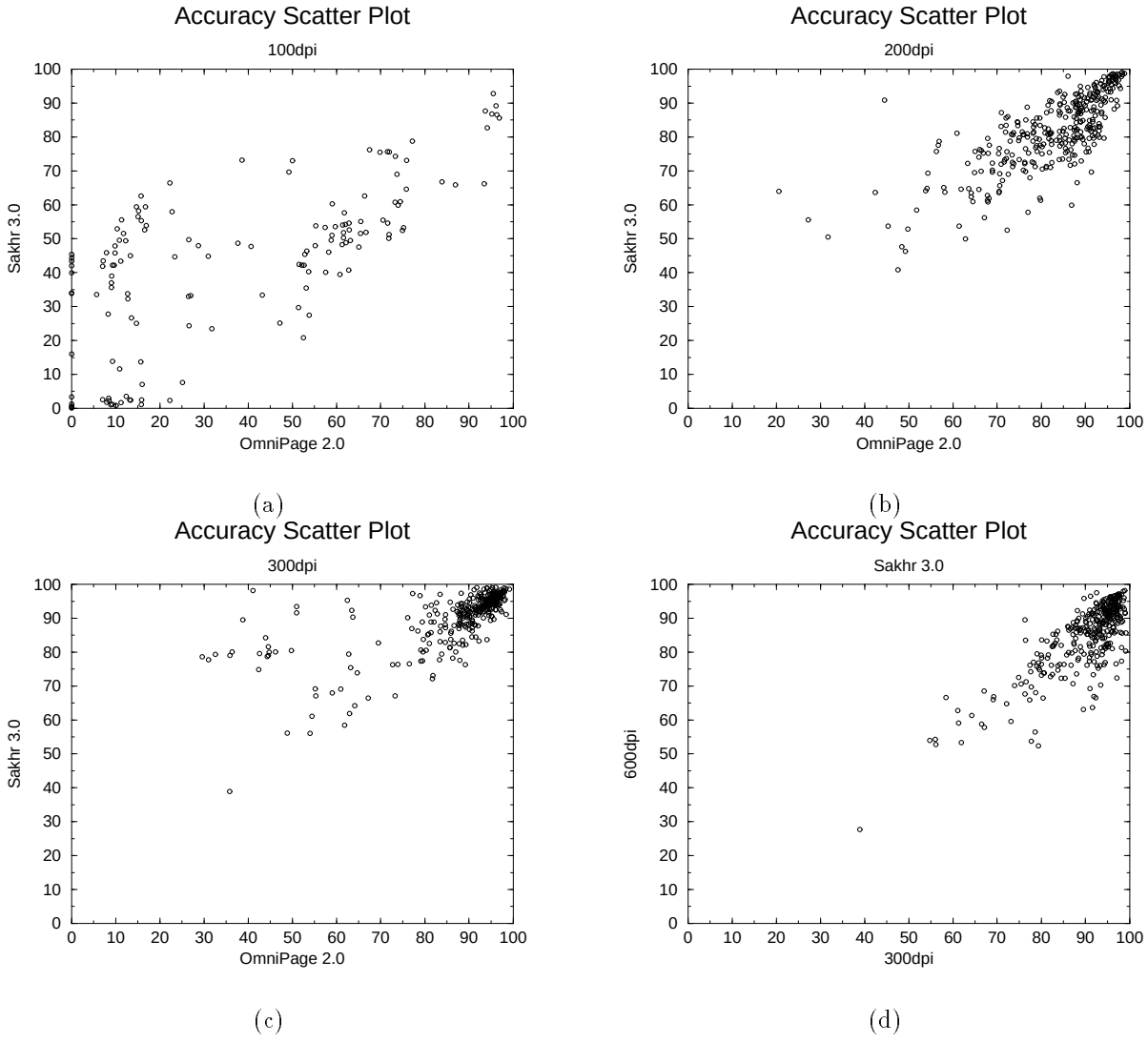
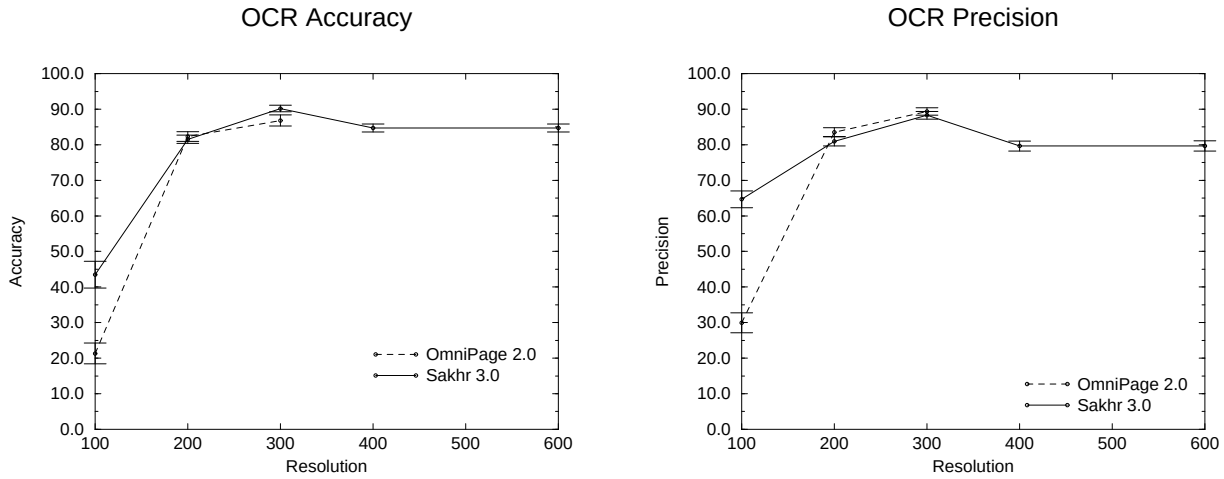


Figure 2. Scatter plot of OCR accuracies of OmniPage and Sakhr at 300 dpi resolution. Each data point represents a specific image. The x-coordinate of the data point represents OmniPage accuracy, whereas the y-coordinate of the data point represents Sakhr accuracy. Points along the diagonal represent document images for which both products achieved similar accuracy. Off-diagonal points indicate that one product performed better than the other. If most points are to one side of the diagonal, then one product is better than other. For example in (c) it can be seen that Sakhr is better than OmniPage on more number of images.



(a) (b)

Figure 3. Accuracy and precision as a function of document image resolution. Accuracy (also known as *recall* in IR community) is number of correctly recognized symbols normalized by the number of groundtruth symbols. Precision is the number number of correctly recognized symbols normalized by the number of symbols in the OCR output. Notice that at 300 dpi, although Sakhr has a higher accuracy, OmniPage has a higher precision. Although the 95% confidence intervals overlap, it is shown in Table 2 that the difference between the means is statistically significant.

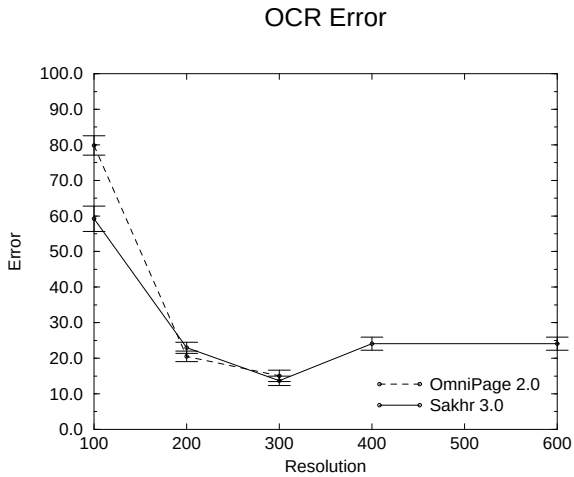


Figure 4. Error (sum of insertion, deletion and substitution errors normalized by the number of groundtruth symbols) as a function of document image resolution.

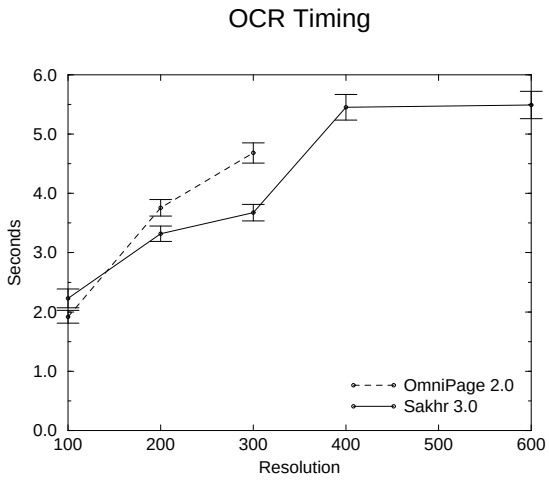


Figure 5. Average time taken to OCR an image. The times are on 400 MHz Intel Pentium processor with 128MB RAM and 256K cache. Notice that Sakhr's Automatic Reader takes the same amount of time to process 600 and 400 dpi images. This is most likely because Sakhr Automatic Reader samples the 400 and 600 dpi images to a standard resolution and then does the OCR processing.

المرحلة ينبغي ان تدرسها الدول العربية بدقة تآمة، فهي من النوع الذكي
الذي تحبكه الصهيونية بمهارة فائقة والذي يعكس شعور اسرائيل بالخطر
من تنامي مد التواصل السياسي والدبلوماسي بين اميركا والدول العربية في

(a)

المرحلة ينبغي ان تدرسها الدول العربية بدقة تآمة ، فهي من النوع الذكي
الذي تحبكه الصهيونية بمهارة فائقة والذي يعكس شعور اسرائيل بالخطر
من تنامي مد التواصل السياسي والدبلوماسي بين اميركا والدول العربية في

(b)

المرحلة ينبغي ان تدرسها الدول العربية بدقة تامة، فهي من النوع الذكي
الذي تحبكه الصهيونية بمهارة فائقة والذي يعكس شعور اس ائيل بالخطر
من تنامي مد التواصل السياسي والدبلوماسعي بين اميركا والدول العربية في

(c)

Figure 6. Subimage from ATI0290. Both Sakhr and OmniPage performed well on this image. Sakhr achieved 98.08% accuracy and OmniPage achieved 97.7% accuracy. A subimage of the original image is shown in (a), the OmniPage output is shown in (b) and the Sakhr output is shown in (c).

٧ ـ تتعهد الحكومتان فيما يتعلق بالخدمات المحددة اعلاه بتطبيق المبادىء
المقررة في اتفاقية الطيران المدني الدولي المعقودة في شيكاغو سنة ١٩٤٤ .
٨ ـ تبقى الترتيبات الواردة اعلاه سارية المفعول لمدة سنة تالية التجديـــد

(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 دة5علاً، طالة *نممعو(صه ه ش!دة سشى كمايئة 9ش!!ه كله

(c)

Figure 7. (a) Subimage from 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.

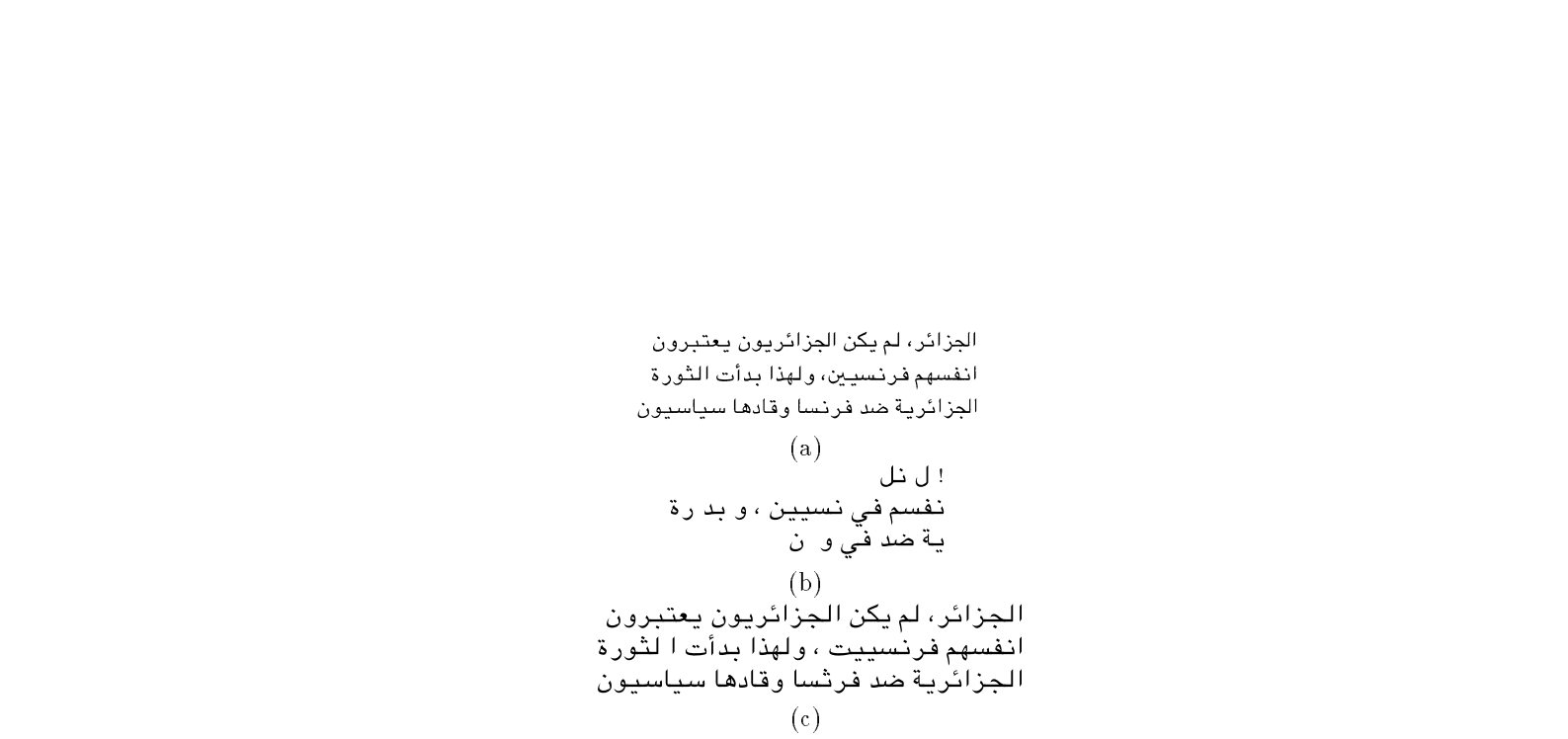


Figure 8. (a) Subimage from ATI0012. Sakhr performed better than OmniPage on this image. Sakhr achieved 98.13% accuracy whereas OmniPage achieved 41.09% accuracy. (b) Output of OmniPage. (c) Ouput of Sakhr.

Figure 9. (a) Subimage from 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.

Figure 10. (a) Subimage from 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.

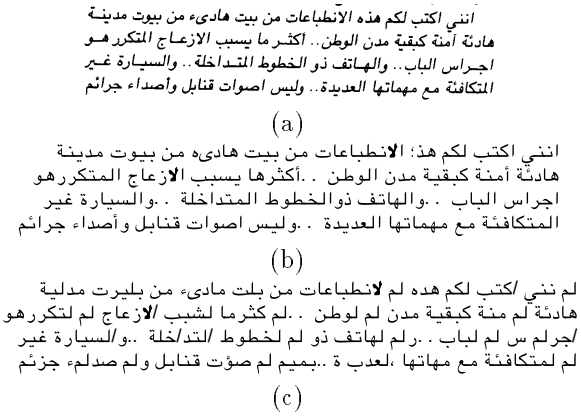


Figure 11. (a) Subimage from ATI0446. OmniPage performed better than Sahkr on this image. OmniPage achieved 94.06% accuracy whereas Sakhr achieved 83.67% accuracy. (b) Output of OmniPage. (c) Ouput of Sakhr.

Table 1. Substitution, deletion, insertion, and total error paired differences. The numbers reported below are the mean paired differences between OmniPage and Sakhr and the corresponding 95% confidence intervals. For example, at 300 dpi the OmniPage has 1.4596% ± 1.036 higher total error than Sakhr, whereas Sakhr has 1.9803% ± 0.3849 higher insertion errors. The intervals are estimated using two techniques. We can see that the paired intervals are smaller than the unpaired ones. A point to note is that at 100 dpi, Sakhr did not generate text on 198 images (required manual intervention). Since the paired differences are reported on images for which both products produced results, and the accuracy plots in Figure 3 report on all the files for which a product generated output, the results can look different if the number of files on which a product crashed is large.

Substitution Differences

Res	Paired		Unpaired	
100	1.9355	± 2.7445	1.9355	± 3.0867
200	-1.8611	± 0.6112	-1.8611	± 1.4196
300	-0.4556	± 0.3334	-0.4556	± 1.0876

Deletion Differences

Res	Paired		Unpaired	
100	3.0277	± 4.8687	3.0277	± 7.2253
200	1.0998	± 0.4240	1.0998	± 0.5720
300	3.8956	± 1.0232	3.8956	± 1.0605

Insertion Differences

Res	Paired		Unpaired	
100	-0.3148	± 0.3533	-0.3148	± 0.5861
200	-1.6079	± 0.3635	-1.6079	± 0.5766
300	-1.9803	± 0.3849	-1.9803	± 0.5273

Error Differences

Res	Paired		Unpaired	
100	4.6487	± 3.3356	4.6487	± 5.7481
200	-2.3692	± 0.8931	-2.3692	± 2.1350
300	1.4596	± 1.0356	1.4596	± 2.0619

Table 2. Accuracy and precision differences as a function of resolution. At 300 dpi, Sakhr has 3.4401% ± 1.1257 higher accuracy than OmniPage whereas OmniPage has 0.9917% ± 0.4672 higher precision than Sakhr.

Accuracy

Res	Paired		Unpaired	
100	-4.9631	± 3.5644	-4.9631	± 6.1339
200	0.7612	± 0.8929	0.7612	± 1.8019
300	-3.4401	± 1.1257	-3.4401	± 1.7859

Precision

Res	Paired		Unpaired	
100	-18.8328	± 4.7582	-18.8328	± 4.8472
200	2.5524	± 0.7212	2.5524	± 1.7795
300	0.9917	± 0.4672	0.9917	± 1.4738

Table 3. Timing differences between OmniPage and Sakhr per 100 character.

Res	Paired		Unpaired	
100	0.0217	± 0.0145	0.0217	± 0.0201
200	0.0329	± 0.0082	0.0329	± 0.0111
300	0.0775	± 0.0131	0.0775	± 0.0173

Table 4. Number of crashed files. At 100dpi Sakhr required human intervention on most of the 198 files that are listed as crashed below. A message popped up asking the user to manually zone the image. Thus we listed them as crashed since we assume that the OCR system is to work in a completely automatic mode.

Res	Opa	Sakhr
100	28	198
200	3	1
300	9	1
400		6
600		6