LAMP-TR-030
CAR-TR-903
CS-TR-3972

MDA 9049-6C-1250
December 1998

Paired Model Evaluation of OCR Algorithms

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 system behavior and identifying open problems. While research has been done in the past to compare the performances of OCR systems, all methods assume that the accuracies achieved on individual documents in a dataset are independent. In this paper we argue that accuracies reported on any dataset are *not independent* and invoke the appropriate statistical technique — the paired model — to compare the accuracies of two recognition systems. We show theoretically that this method provides tighter confidence intervals than the methods used in the 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 functions of image 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 Sakhr OCR algorithm resamples the high-resolution images 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 paper are very general and can be used for comparing the accuracies of any two recognition systems, not just OCR systems.

This research was funded in part by the Department of Defense and the Army Research Laboratory under Contract MDA 9049-6C-1250.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
DEC 1998		**00-12-1998 to 00-12-1998**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Paired Model Evaluation of OCR Algorithms	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Language and Media Processing Laboratory,Institute for Advanced Computer Studies,University of Maryland,College Park,MD,20742-3275**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**19**	

LAMP-TR-030
CAR-TR-903
CS-TR-3972

MDA 9049-6C-1250
December 1998

Paired Model Evaluation of OCR Algorithms

Tapas Kanungo, Gregory A. Marton, Osama Bulbul

Paired Model Evaluation of OCR Algorithms

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 system behavior and identifying open problems. While research has been done in the past to compare the performances of OCR systems, all methods assume that the accuracies achieved on individual documents in a dataset are independent. In this paper we argue that accuracies reported on any dataset are *not independent* and invoke the appropriate statistical technique — the paired model — to compare the accuracies of two recognition systems. We show theoretically that this method provides tighter confidence intervals than the methods used in the 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 functions of image 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 Sakhr OCR algorithm resamples the high-resolution images 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 paper are very general and can be used for comparing the accuracies of any two recognition systems, not just OCR systems.

This research was funded in part by the Department of Defense and the Army Research Laboratory under Contract MDA 9049-6C-1250.

1 Introduction

Performance evaluation and characterization of OCR systems is crucial for many reasons: i) To predict the overall performance of any system that employs an OCR submodule — for example, an information retrieval (IR) system or a machine translation (MT) system. ii) To monitor progress in research/development of OCR systems, quantitative performance measures are essential. iii) To scientifically understand the contributions of specific submodules to the accuracy improvement, and thus to 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 system.

In this article we use a statistical technique — the paired model — to compare the accuracies of two recognition systems. We show theoretically that this method provides tighter confidence intervals than the methods used in the OCR 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 a researcher to visually compare the OCR performances on individual document images.

We start the discussion by providing background information on the OCR performance evaluation literature. Metrics for quantifying errors are discussed in Section 3. In Section 4 we discuss the statistical theory of paired models, which we use to compare the performances of two Arabic OCR systems. In Section 5 we describe the datsets and experimental protocol we use to conduct our evaluation. Scatter plots and evaluation results are discussed in Section 6.

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 force us to limit the size of the dataset to something feasible. This process was adopted in the UNLV OCR evaluation program [RJN96] and the UW evaluation process[CSHP94]. The UNLV evaluation corpus consisted of English annual reports, documents from the Department of Energy, magazines, business letters, legal documents, Spanish newspapers, and German business letters. The UW dataset [HP$^+$] consisted of English technical journals.

Whitebox evaluation, on the other hand, characterizes the performances 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 by Vincent *et al.* [YV98, RV94].) Whitebox evaluation is possible only if the evaluator has access to the inputs and outputs of the submodules of the OCR system. Thus for segmentation evaluation, access to the coordinates of the zones produced by the system 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.* [KHP94, Kan96]) 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 [Bai92, KHP94, Kan96]. 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 [KH99, Kan96] is based on 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 the measurements.

3 Metrics for Performance Evaluation

What metrics are good for evaluating OCR systems? In this section we describe a few metrics that we consider important and give their advantages and disadvantages. 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 for which another symbol is substituted, and T the number of groundtruth symbols. We now define five metrics based on these quantities.

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: 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 are computed using the DOD error counter, which is based on a string matching routine. In this article we have not reported the above metrics at the word level. We are currently in the process of computing the word-level metrics. 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 systems 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 comparison of measurements. One of the problems encountered while comparing the OCR results of two algorithms is that of comparing the means of two accuracy samples that are obtained by running the two algorithms on a specific dataset. In general, the underlying true accuracy populations are not distributed as Gaussians and thus making such assumptions is not justified. However, since the datasets are usually large (greater than 30), certain statistical techniques can be used for comparing average accuracies. We now describe a few of these techniques; refer to [Arn90] for details.

4.1 Large sample inference about means

Let $x_1, x_2, \ldots, x_n$ be a set of OCR accuracy measurements obtained by processing n document images. Let the underlying distribution of the accuracies have 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 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 α is the significance level and $z^{\alpha/2}$ is a number such that $P(z > z^{\alpha/2} | z \sim N(0,1)) = \alpha$,

4.2 Inference about the means of two independent samples

Let $x_1, x_2, \ldots, x_m$ be a sample of OCR accuracy measurements obtained by processing m document images. Let $y_1, y_2, \ldots, y_n$ be another *independent* sample of accuracy measurements. Let $\bar{x}$ and S^2 be the sample mean and variance of x_i, and $\bar{y}$ and T^2 the sample mean and variance of y. Let the underlying x population have mean μ and variance σ^2, and the y population have 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:

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

It can be shown that

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

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

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 the 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 OCR accuracy measurements on the *same* document image. The results in the previous subsection required the two samples to be independent and so these results 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 the paired and unpaired estimators are identical, and equal to the difference between the population means. The variances, however, differ. The variance of the paired estimator is $(\sigma^2 + \tau^2 - 2\rho\tau\sigma)/n$ while the variance of the unpaired estimator is $(\sigma^2 + \tau^2)/n$. Thus the paired estimator is better since its variance is smaller and the uncertainty is lower for a fixed sample size. The paired estimator uses the correlation information to reduce the uncertainty in its estimate. When the correlation coefficient ρ is equal to zero, the estimation methods have identical variances.

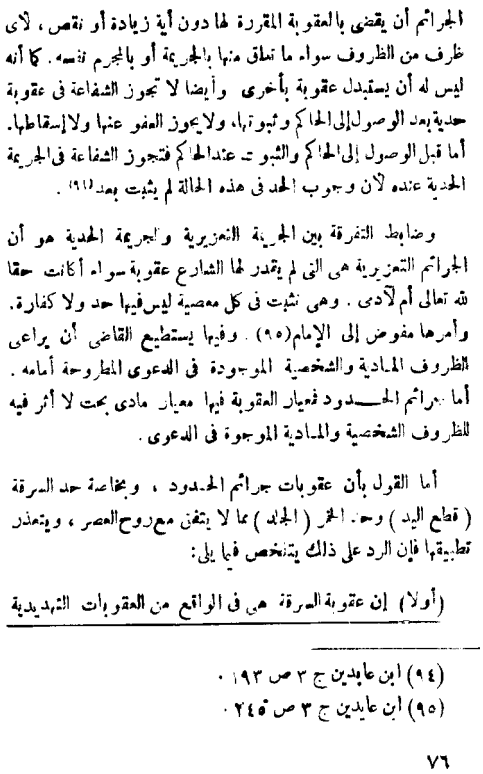
5 Experimental Protocol

The SAIC dataset[DH97], which was provided to us by the Department of Defense, was used for evaluating the performance of two Arabic OCR systems. The SAIC corpus has binary images of printed Arabic text and corresponding "groundtruth." By groundtruth we mean manually typed correct Arabic ASCII strings that the OCR systems should ideally produce. The images in the dataset are zones with a single column of text. In Figure 1 we show a sample image from the dataset. The images are relatively clean and are scanned from books, magazines and computer-generated documents. The dataset contains 345 image/groundtruth pairs. Three of these pairs (ATI0746, ATI0116, and ATI0286) are unusable: ATI0746 does not have an image, ATI0116 does not have the groundtruth, and the image and the groundtruth for ATI0286 do not match. Groundtruth text is encoded in CP1256 format. TIFF images, originally at 600 dpi, were resampled at 100, 200, 300 and 400 dpi using the public-domain utility `convert`.

The Arabic OCR products that were evaluated are Sakhr's Automatic Reader 3.01 and Shonut's OmniPage Pro 2.0. 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.

6 Results and Discussion

For both products we computed histograms of the accuracy measurements on the SAIC dataset at 300 dpi. These histograms are shown in Figure 2. It can be seen that the empirical distributions of the accuracies are not Gaussian and that the accuracy distribution of OmniPage has a fatter tail than that of Sakhr. Scatter plots of accuracy pairs for Sakhr and OmniPage at 100, 200, and 300 dpi are shown in Figure 3(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 along the diagonal represent document images for which both products achieved similar accuracies. Points very far from the diagonal represent images for which one product performed much better than the other. In Figure 3(c) it



الجرائم أن يقضى بالعقوبة المقررة لها دون أية زيادة أو نقص ، لأى ظرف من الظروف سواء ما تعلق منها بالجريمة أو بالمجرم نفسه . كما أنه ليس له أن يستبدل عقوبة بأخرى وأيضا لا تجوز الشفاعة فى عقوبة حدية بعد الوصول إلى الحاكم وثبوتها، ولا يجوز العفو عنها ولا إسقاطها. أما قبل الوصول إلى الحاكم والثبوت عند الحاكم فتجوز الشفاعة فى الجريمة الحدية عنده لأن وجوب الحد فى هذه الحالة لم يثبت بعد(٩٤) .

وضابط التفرقة بين الجريمة التعزيرية والجريمة الحدية هو أن الجرائم التعزيرية هى التى لم يقدر لها الشارع عقوبة سواء أكانت حقا لله تعالى أم لآدمى . وهى تثبت فى كل معصية ليس فيها حد ولا كفارة. وأمرها مفوض إلى الإمام(٩٥) . وفيها يستطيع القاضى أن يراعى الظروف المادية والشخصية الموجودة فى الدعوى المطروحة أمامه . أما جرائم الحـــدود فمعيار العقوبة فيها معيار مادى بحت لا أثر فيه للظروف الشخصية والمادية الموجودة فى الدعوى .

أما القول بأن عقوبات جرائم الحدود ، وبخاصة حد السرقة (قطع اليد) وحد الخمر (الجلد) مما لا يتفق مع روح العصر ، ويتعذر تطبيقها فإن الرد على ذلك يتلخص فيما يلى:

(أولا) إن عقوبة السرقة هى فى الواقع من العقوبات التهديدية

(٩٤) ابن عابدين ج ٣ ص ١٩٣ .
(٩٥) ابن عابدين ج ٣ ص ٢٤٥ .

٧٦

Figure 1: A sample image from the SAIC dataset.

can be seen that there are more points above the diagonal than below it. This implies that at 300 dpi there are many images for which Sakhr performed better than OmniPage. A scatter plot of Sakhr at 300dpi and 600dpi is shown in Figure 3(d). It can be seen that contrary to what one might expect, the Sakhr algorithm performs worse at 600dpi than at 300dpi. A paired model analysis reveals that the 95% confidence interval for the difference in the means is 5.4453 ± 0.6557. Thus the difference is statistically significant.

Accuracy, precision, and error are plotted as functions of document image resolution in Figures 4 and 5. Accuracy (also known as *recall* in the information retrieval community) is the number of correctly recognized symbols normalized by the number of groundtruth symbols. Precision is the number of correctly recognized symbols normalized by the number of symbols in the OCR output. Error is the sum of the numbers of insertion, deletion and substitution errors normalized by the number of groundtruth symbols. The paired differences in accuracy and precision between OmniPage and Sakhr at various resolution are summarized in Table 1. Notice that at 300 dpi Sakhr has a higher accuracy but OmniPage has a higher precision. Although the 95% confidence intervals for precision overlap, it is shown in Table 1 that the difference between the precision means is statistically significant. Higher precision means that OmniPage has fewer insertion errors than Sakhr. This can be seen in Table 2, which summarizes the differences between insertion, deletion, and substitution errors for OmniPage and Sakhr.

In Figure 6 the average time taken to OCR an image is plotted. The average time taken to process an image is lower for Sakhr than for OmniPage. Furthermore, Sakhr's

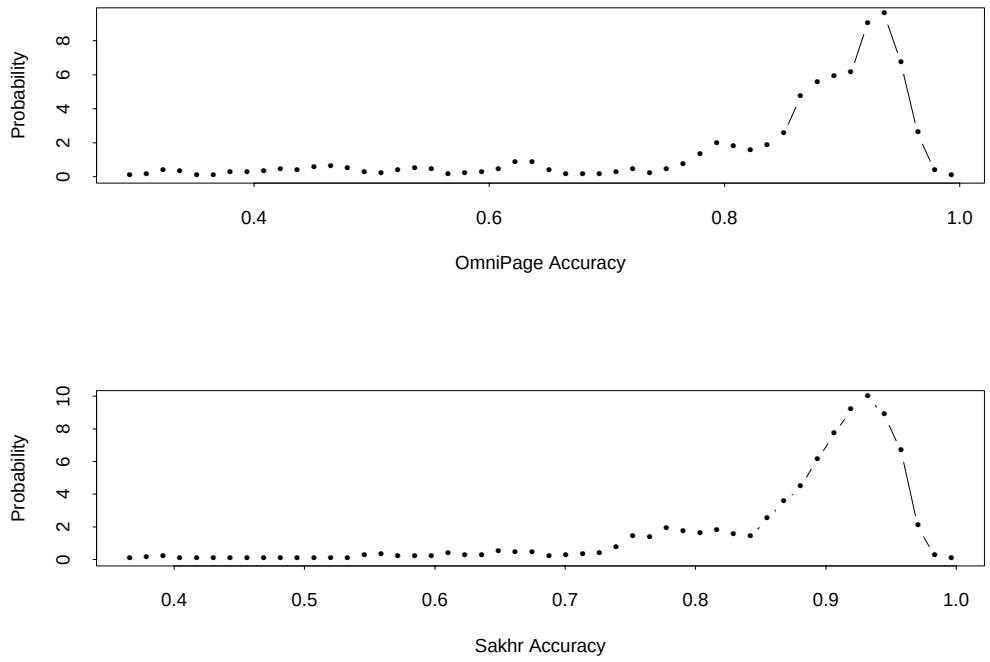


Figure 2: The first plot is the distribution of page accuracies of OmniPage for images at 300 dpi. The second plot is the corresponding distribution of Sakhr page accuracies. Notice that the accuracies are not distributed as Gaussians.

processing 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 then does the OCR processing. In Table 3 we provide the average paired difference in processing time per 100 symbols. Again it can be seen that processing time for Sakhr is lower than that for OmniPage.

Numerous subimages from the dataset images, and the corresponding OCR outputs at 300 dpi for both products, are shown in Figures 7-12. In Figure 4 we provide the number of images on which the products crashed or required manual intervention. Sakhr performed poorly at 100dpi, but was quite stable at all other resolutions. OmniPage was quite stable at 300 dpi and lower, but crashed on almost all the pages at higher resolutions.

7 Summary

We have shown that the paired model approach to performance comparison gives rise to tighter confidence intervals than unpaired methods when computing the 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 better precision. The average page accuracy rate of Sakhr is 90.333% while that of OmniPage

Table 1: Accuracy and precision differences as functions 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

is 86.89%. That is, the average page accuracy of Sakhr is $3.44 \pm 1.13\%$ higher than that of OmniPage. However, at 300 dpi, OmniPage has $0.9917 \pm 0.4672\%$ higher precision than Sakhr. We have also characterized the accuracy, precision, and error as functions of resolution and shown that the accuracy of Sakhr drops when the image resolution is increased beyond 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 Sakhr algorithm samples the high-resolution images to a lower resolution prior to the OCR process. A scatter plot is used to visualize, compare, and summarize 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 Defense for providing us with the OCR error counting software. This research was supported in part by the Army Research Lab and the Department of Defense under Contract MDA 9049-6C-1250.

References

[Arn90] S. F. Arnold. *Mathematical Statistics*. Prentice-Hall, New Jersey, 1990.

[Bai92] H. S. Baird. Document image defect models. In *Structured Document Image Analysis*. Springer-Verlag, New York, 1992.

[CSHP94] S. Chen, S. Subramaniam, R. M. Haralick, and I. T. Phillips. Performance evaluation of two OCR systems. In *Proc. of Annual Symp. on Document Analysis and Information Retrieval*, pages 299–317, Las Vegas, NV, April 1994.

[DH97] R. Davidson and R. Hopely. Arabic and Persian OCR training and test data sets. In *Proc. of Symp. on Document Image Understanding Technology*, Annapolis, MD, April 30 – May 2 1997.

[HP+] R. M. Haralick, I. Phillips, et al. UW-CDROM-I.

[Kan96] 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.

[KH99] 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*, 1999.

[KHP94] T. Kanungo, R. M. Haralick, and I. Phillips. Non-linear local and global document degradation models. *Int. Journal of Imaging Systems and Technology*, 5(3):220–230, 1994.

[KR99] T. Kanungo and P. Resnik. The Bible, truth, and multilingual OCR evaluation. In D. Lopresti and Y. Zhou, editors, *Proc. of SPIE Conf. on Document Recognition and Retrieval VI*, San Jose, CA, 1999.

[RJN96] S. V. Rice, F. R. Jenkins, and T. A. Nartker. The fifth annual test of OCR accuracy. Technical Report TR-96-01, Information Science Research Institute, University of Nevada, Las Vegas, NV, 1996.

[RV94] S. Randriamasy and L. Vincent. Benchmarking page segmentation algorithms. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, June 1994.

[YV98] B. A. Yanikoglu and L. Vincent. PinkPanther: A complete environment for ground-truthing and benchmarking document page segmentation. *Pattern Recognition*, 31:1191–1204, 1998.

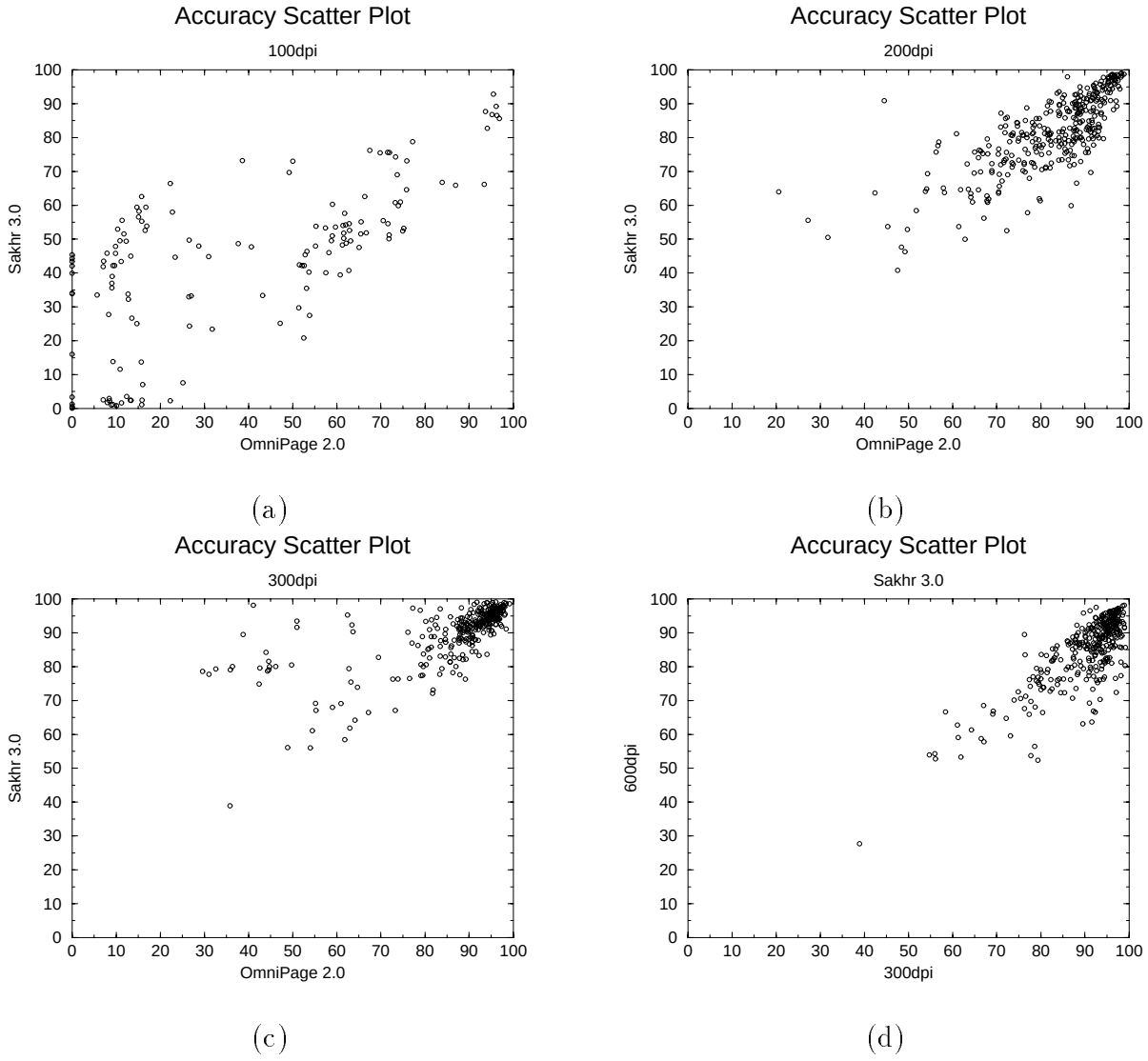
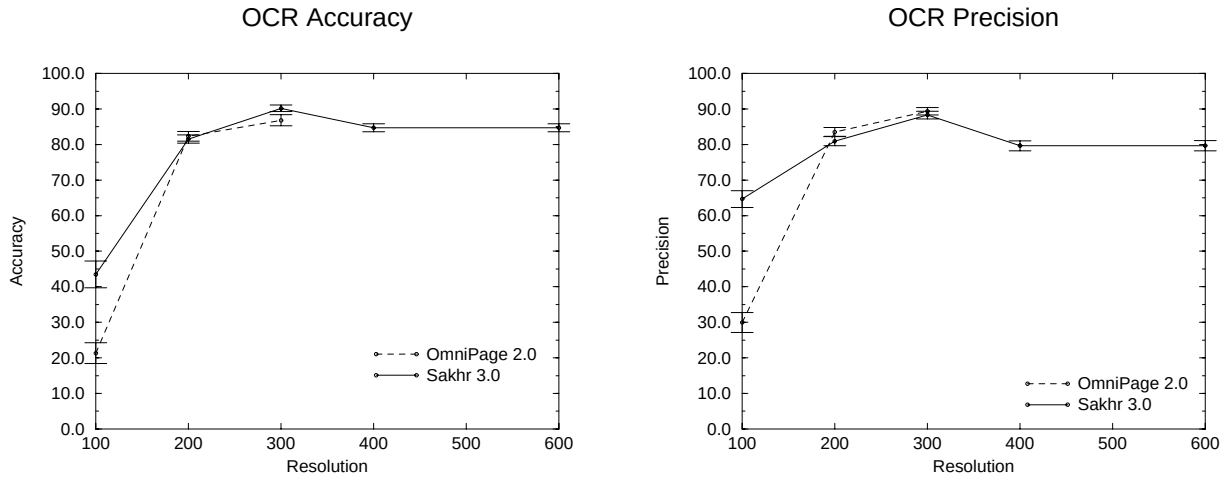


Figure 3: Scatter plot of the OCR accuracies of OmniPage and Sakhr at 300 dpi resolution. Each data point represents a specific image. The x-coordinate represents OmniPage accuracy and the y-coordinate represents Sakhr accuracy. Points along the diagonal represent document images for which both products 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 a larger number of images.



(a) (b)

Figure 4: Accuracy and precision as functions of document image resolution. Accuracy (also known as *recall* in the information retrieval community) is the number of correctly recognized symbols normalized by the number of groundtruth symbols. Precision is the 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 1 that the difference between the means is statistically significant.

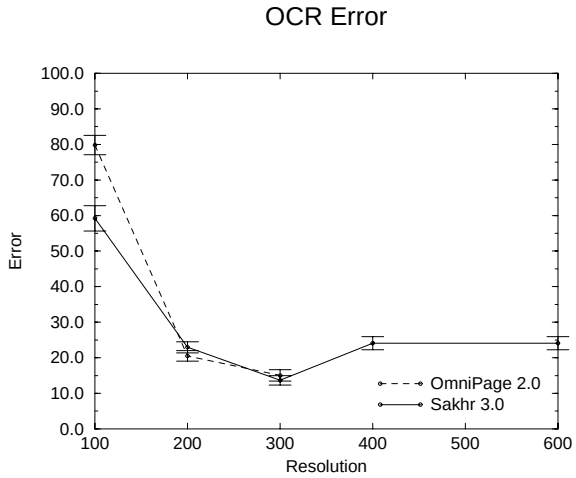


Figure 5: Error (sum of the numbers of insertion, deletion and substitution errors normalized by the number of groundtruth symbols) as a function of document image resolution.

Table 2: 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 OmniPage has 1.4596% $\pm$ 1.036 higher total error than Sakhr, whereas Sakhr has 1.9803% $\pm$ 0.3849 higher insertion error. 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 4 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	$\pm$ 2.7445	1.9355	$\pm$ 3.0867
200	-1.8611	$\pm$ 0.6112	-1.8611	$\pm$ 1.4196
300	-0.4556	$\pm$ 0.3334	-0.4556	$\pm$ 1.0876

Deletion Differences

Res	Paired		Unpaired	
100	3.0277	$\pm$ 4.8687	3.0277	$\pm$ 7.2253
200	1.0998	$\pm$ 0.4240	1.0998	$\pm$ 0.5720
300	3.8956	$\pm$ 1.0232	3.8956	$\pm$ 1.0605

Insertion Differences

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

Error Differences

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

Table 3: Timing differences between OmniPage and Sakhr per 100 characters.

Res	Paired		Unpaired	
100	0.0217	$\pm$ 0.0145	0.0217	$\pm$ 0.0201
200	0.0329	$\pm$ 0.0082	0.0329	$\pm$ 0.0111
300	0.0775	$\pm$ 0.0131	0.0775	$\pm$ 0.0173

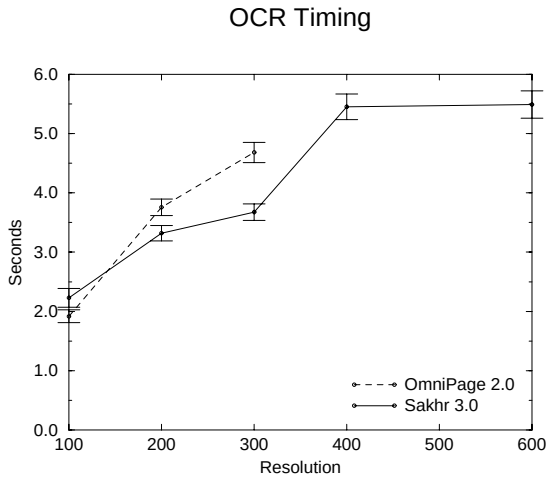


Figure 6: Average time taken to OCR an image. The times are on a 400 MHz Intel Pentium processor with 128MB RAM and 256K cache. Notice that Sakhr's Automatic Reader takes the same amount of time to process 600 and 400 dpi images. This is most likely because the Sakhr Automatic Reader samples the 400 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)

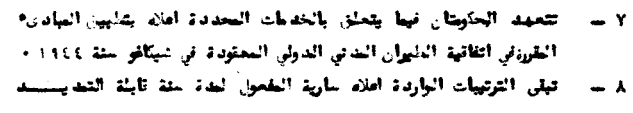
المرحلة ينبغي ان تدرسها الدول العربية بدقة تآمة ، فهي من النوع الذكي
الذي تحبكه الصهيونية بمهارة فائقة والذي يعكس شعور اسرائيل بالخطر
من تنامي مد التواصل السياسي والدبلوماسي بين اميركا والدول العربية في

(b)

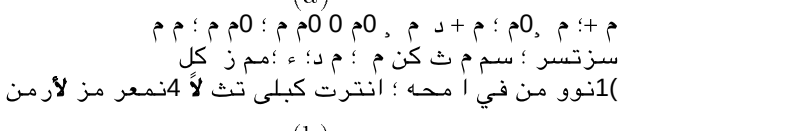
المرحلة ينبغي ان تدرسها الدول العربية بدقة تامة، فهي من النوع الذكي
الذي تحبكه الصهيونية بمهارة فائقة والذي يعكس شعور اس ائيل بالخطر
من تنامي مد التواصل السياسي والدبلوماسعي بين اميركا والدول العربية في

(c)

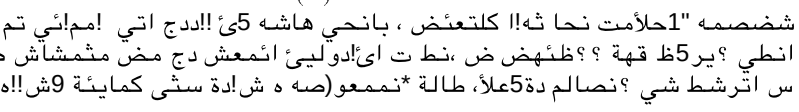
Figure 7: Subimage from ATI0290. Both Sakhr and OmniPage performed well on this image. Sakhr achieved 98.08% accuracy and OmniPage achieved 97.7% accuracy. A subimage of the original image is shown in (a), the OmniPage output is shown in (b) and the Sakhr output is shown in (c).



(a)

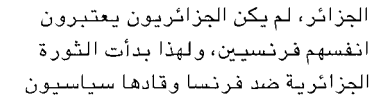


(b)

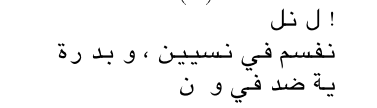


(c)

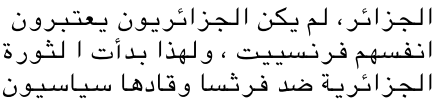
Figure 8: (a) Subimage from ATI0082. Both Sakhr and OmniPage performed poorly on this image. Sakhr achieved 38.88% accuracy and OmniPage achieved 35.79% accuracy.(b) Output of OmniPage. (c) Ouput of Sakhr.



(a)



(b)



(c)

Figure 9: (a) Subimage from ATI0012. Sakhr performed better than OmniPage on this image. Sakhr achieved 98.13% accuracy whereas OmniPage achieved only 41.09% accuracy. (b) Output of OmniPage. (c) Ouput of Sakhr.

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

وسيبدأ العمل بتنفيذ المشروع وفق الدراسة المقترحة بعد ٣٦
شهرا كما تقرر ان يكون مقر المشروع مدينة طرطوس على الساحل
السوري وان يكون عدد العاملين فيه ٢٠٢ عامل.

(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 **لأ** 4نمعر مز **لأ**رمن

(b)

وسيبدأ العمل بتمفيذ المنتمووكا وفق ألدواسة المقترحة بعد 36
شهرا كما تقرو ان يكون مقرم لمشو؟ع مدينة طرطوس على الساحل
السوري وان يكون عدد العاملش فيه 252عامل .

(c)

Figure 10: (a) Subimage from ATI0239. Sakhr performed better than OmniPage on this image. Sakhr achieved 89.52% accuracy whereas OmniPage achieved only 38.75% 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)

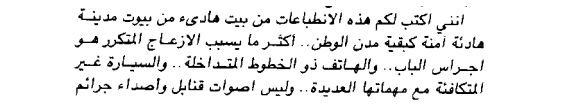
اليامة أن الشعب اللبناني ذاته ، في **الامتفتا** ,لذي جرى في شهر حزيران الماني ، عبر عن تبنيه لها بالذات وي
إ يأنه بعوام 1، فأسبح ء .واجب المكومة المنبثقة من مجللا الكرم _وهو ثرة هذا **الاستفتاء** الشعبي _أن تبق
أمينة لتلك اليات وأن تتمو فى تنفيذها .

(b)

السماسة أر السصط اللسايى ذاته ، ي **الاسسصاء** الدي جرى ى شهر حريراد المامي ، ككمر س تبسبه لها بالذات وكل إبمأنه لصواكا ، **لا**صبح 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 11: (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.


انني اكتب لكم هذه الآنطباعات من بيت هادىء من بيوت مدينة
هادئة آمنة كبقية مدن الوطن.. اكثر 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 /التدلخلة ..و/السيارة غير
لم لمتكافئة مع مهاتها ،لعدب ة ..بميم لم صؤت قنابل ولم صدلمء جزئم

(c)

Figure 12: (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.