Optimizing OCR accuracy for bi-tonal, noisy scans of degraded Arabic documents

Paul Herceg, Ben Huyck, Chris Johnson, Linda Van Guilder, Amlan Kundu[*]
The MITRE Corporation, 7515 Colshire Drive, McLean, VA 22102-7508

ABSTRACT

Acquiring foreign language from degraded hardcopy documents is of interest to military and border control applications. Bi-tonal image scans are desirable because file size is small. However, the nature of hardcopy degradations and the scanner or image enhancement software capabilities used directly affect the quality of the captured image and the extent of language acquisition. We applied a collection of manual treatments to hardcopy Arabic documents to develop a corpus of bi-tonal images. We then used this corpus in an exploratory study to derive conclusions about how bi-tonal images could be enhanced. This paper discusses the manually degraded Arabic document corpus, the image enhancement study, and the significant optical character recognition (OCR) improvements obtained with simple scanner driver adjustments.

Keywords: OCR, Arabic, image enhancement, foreign language, degraded documents.

1. INTRODUCTION

Foreign language data acquired via Arabic OCR is of vital interest to military and border control applications. Various hardcopy paper types and machine- and environment-based treatments introduce artifacts in scanned images. Artifacts such as speckles, lines, faded glyphs, dark areas, shading, etc. complicate OCR and can significantly reduce the accuracy of language acquisition. For example, Sakhr Automatic Reader, a leader in Arabic OCR, performed poorly in initial tests with noisy document images. We hypothesized that performing image enhancement of bi-tonal images prior to Arabic OCR would increase the accuracy of OCR output. We also believed that increased accuracy in the OCR would directly correlate to the success of downstream machine translation.

We applied a wide variety of paper types and manual treatments to hardcopy Arabic documents. The intent was to artificially model how documents degrade in the real world. Four hardcopies of each document were created by systematically applying four levels of treatments. Subsequent scanning resulted in images that reflect the progressive damage in the life-cycle of each document – the Manually Degraded Arabic Document (MDAD) corpus.

We then used this corpus in an exploratory study to derive conclusions about how bi-tonal images could be enhanced. This paper discusses the manually degraded Arabic document corpus, the image enhancement study, and the significant optical character recognition (OCR) improvements obtained with simple scanner driver adjustments. We also discuss lessons learned in developing a manual degradation approach for those who wish to create their own manually degraded corpus.

2. CREATING THE MANUALLY DEGRADED ARABIC DOCUMENT CORPUS

Arabic was the target language, therefore we selected the Arabic newswire and broadcast news documents of the Linguistic Data Consortium (LDC) Automatic Content Extraction (ACE) corpus[1]. These documents include mostly Arabic with some Romanized text.

[*] {pherceg,bhuyck,cj,lcvg,akundu}@mitre.org; phone 1 703 883-4539; fax 1 703 883-1379; mitre.org

Visual Information Processing XIV, edited by Zia-ur Rahman, Robert A. Schowengerdt,
Stephen E. Reichenbach, Proc. of SPIE Vol. 5817 (SPIE, Bellingham, WA, 2005)
0277-786X/05/$15 · doi: 10.1117/12.606447

The ACE corpus is in UTF-8 SGML format, therefore we used a simple Java script to strip the tags, preserve the paragraph formatting, and create the UTF-8 ground truth for the MDAD corpus. A UTF-8 to CP1256 converter was used to transform the UTF-8 ground truth to CP1256.

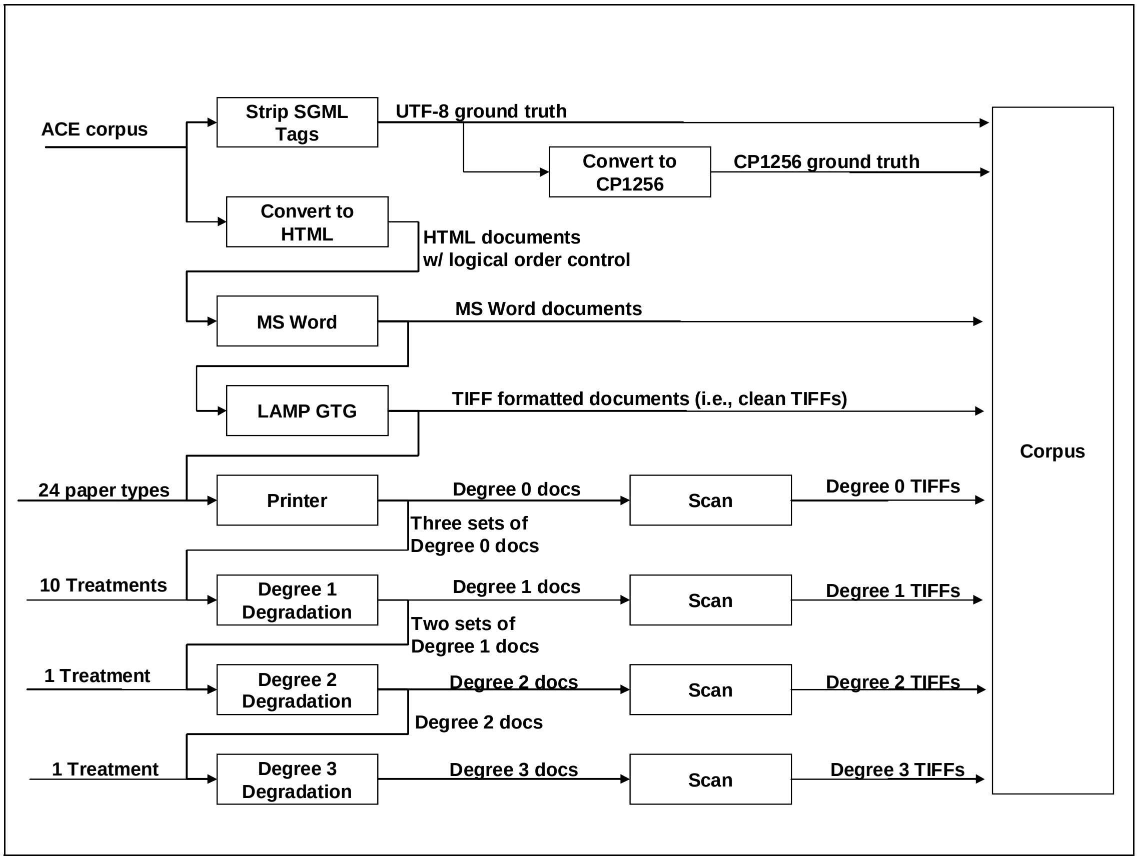

Figure 1. Creating the MDAD corpus.

We converted the UTF-8 truth files to HTML, then to Microsoft (MS) Word format, then to TIFF format. We wanted to add font formatting to the UTF-8 truth files, so HTML tags were inserted to control font face, size, and weight. Rendering direction was also a formatting objective due to the presence of English words in corpus documents. The formatting was specified with cascading stylesheet (CSS) tags. The default direction of each document was specified as right-to-left and left-to-right tags were placed around English text. One of the capabilities of the University of Maryland's (UMD) Language and Media Processing Laboratory (LAMP) Ground Truth Generator (GTG) is conversion of MS Word documents to Tagged Image File Format (TIFF). So we used MS Word to convert each HTML document to an MS Word document, then used GTG to convert each MS Word document to TIFF.

We devised a degradation scheme of combining paper types and treatments. The 24 paper types used included graph paper, parchments, tracing paper, textured paper, lined paper, and construction paper. The 12 treatments applied to hardcopies included rubber stamp, food stain, annotation, three hole punch, edge tears, dog-eared, fire-damaged, highlighter pen, fax, photocopy, soil, and wrinkles/folds. The first ten treatments were considered to be associated with both useful and discarded documents, and were used as *initial treatments*. Soil and wrinkles/folds were considered to be associated with discarded documents.

Degradation degree was our method of denoting a snapshot in the life-cycle of damaging a document. Degradation degree 0 denoted the ground truth image printed on paper. Degradation degrees 1, 2, and 3 denoted the cumulative application of an *initial treatment*, soil treatment, and wrinkle/fold treatment, respectively.

Hardcopy documents were needed at each degree (i.e., stage) in a document's life-cycle. Also, rescanning of the hardcopy was required to investigate the effects of scanner driver adjustments at each degree in a document's life-cycle. This applied to tools that were tightly-coupled with the scanning process. Therefore, we began by printing the ground truth TIFF files on the 24 paper types to create four identical sets of Degree 0 documents. Some paper had to be trimmed to standard letter size. A corporate laser printer was used for printing on standard 20 lb. white paper. An HP 970Cxi inkjet printer was used to print on other paper types.

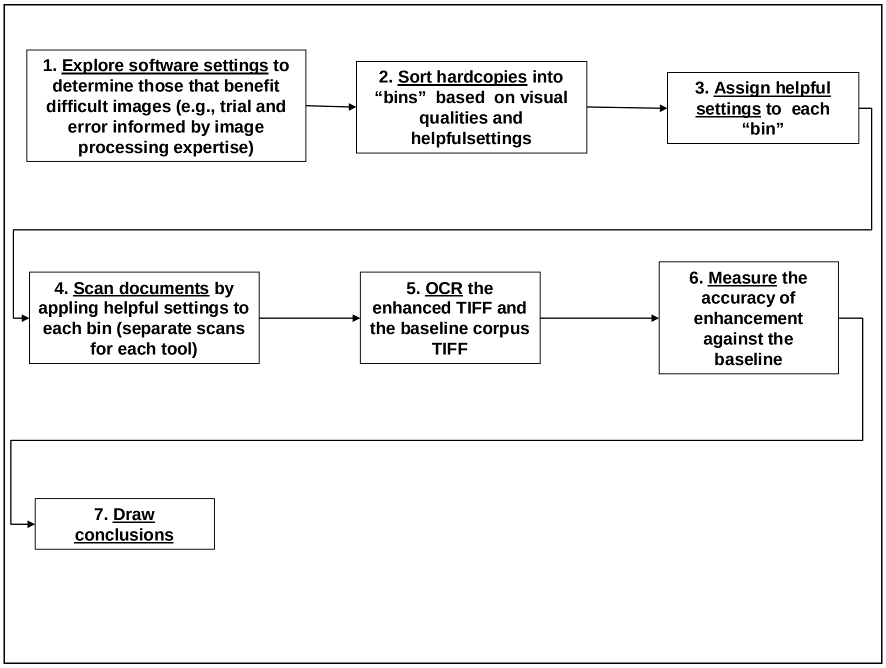
1. Explore software settings to determine those that benefit difficult images (e.g., trial and error informed by image processing expertise)
2. Sort hardcopies into "bins" based on visual qualities and helpfulsettings
3. Assign helpful settings to each "bin"
4. Scan documents by appling helpful settings to each bin (separate scans for each tool)
5. OCR the enhanced TIFF and the baseline corpus TIFF
6. Measure the accuracy of enhancement against the baseline
7. Draw conclusions

Figure 2. Approach to the exploratory study.

This enabled us to easily correct paper jams and reflected the ink type of interest. Initial tests of water-based damage proved that inkjet-printed text would not bleed during the anticipated degradations.

Of these four sets, the first set was preserved – *Degree 0*. *Initial treatments* were applied to the second set – *Degree 1*. Initial treatments and soil were applied to the third set – *Degree 2*. Initial treatments, soil, and wrinkles/folds were applied to the fourth set – *Degree 3*.

Dog-ear treatment was applied by placing documents inside text-books and tossed inside a backpack. Photocopy treatment was performed with a dark copy setting, introducing skew at times. Fax treatment was performed with a Sharp UX-P200 and HP OfficeJet. The OfficeJet performed better than the Sharp at transmitting dark color documents at default settings (e.g., red, green, blue construction paper).

For degree 2, each document was placed face down into wet garden soil until a thin residue covered the text. Once fully dried the residue was brushed off, leaving a soil stain (i.e., soil treatment).

For degree 3, wrinkle/fold treatment was combined with previous treatments. Wrinkling was performed by crushing the document in the hand one, two, or three times. Care was applied to ensure crushing did not introduce severe edge tears or holes. Four combined folds produced interesting results.

The resulting hardcopy documents were scanned with a Fujitsu M3091DC scanner using Fujitsu's TWAIN driver v9.11.47. Default settings for OCR were used to maintain scan consistency for the subsequent image enhancement study (including the default 300dpi, Simplified Dynamic Thresholding [SDTC], etc.). It appeared the scanner auto-adjusted the scan threshold then used that setting for all subsequent pages. Therefore, each document was scanned individually rather than in batch to ensure the proper adjustment.

The 206 ACE corpus documents vary in length from one to nine pages in length. Therefore, each page of a document was captured in a individual TIFF file. Version 2 of the MDAD corpus includes 1352 TIFF files.

For some of the darker documents, default scans performed poorly so that none of the text can be deciphered. These documents were rescanned and added to version 2 of the MDAD corpus as a separate document set. An additional set of documents reflects combining the rescans with the default scans. The entire process is depicted in Figure 1.

3. IMAGE ENHANCEMENT EXPLORATORY STUDY

Image enhancement capabilities can be described as image processing capabilities targeted to enhancing a photograph or document image. They are found in multiple software types: image enhancement software, scanner driver software, OCR software, and image editing software. Features applicable to bi-tonal text documents include thresholding (e.g., brightness), contrast, gamma correction, deskew, rotate, margin processing, noise removal, line removal, text processing, and smoothing such as dilation and erosion.

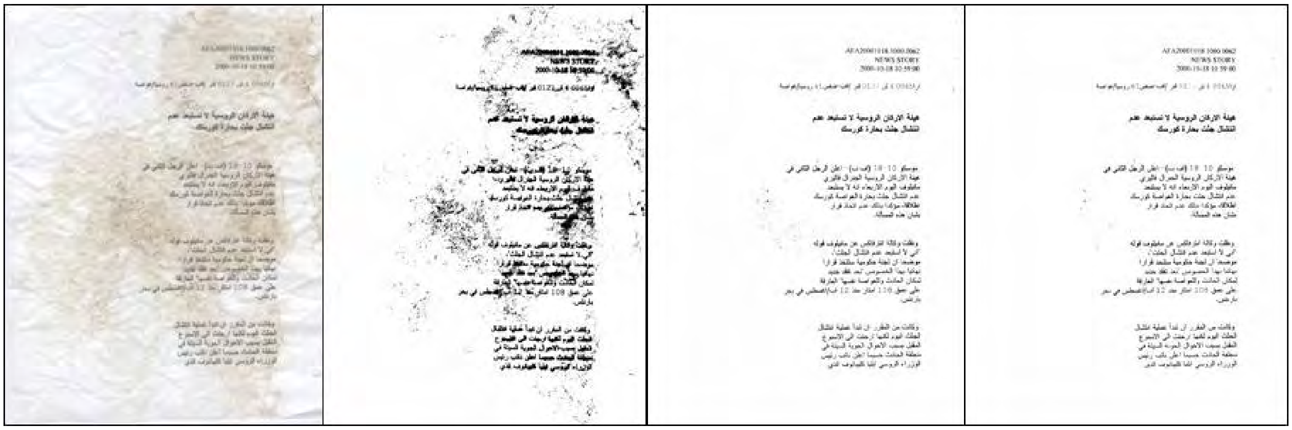

Figure 3. Example image enhancement sequence (a.original document, b. default scan, c. after Fujitsu enhancement, d. after both Fujitsu and ScanFix enhancement).

We conducted an exploratory study to investigate the benefit of image enhancement capabilities for OCR of the MDAD corpus (i.e., bi-tonal Arabic document images). We used the Fujitsu TWAIN32 driver software for the M3091DC scanner, TMSSequoia Scanfix, Kofax VRS, and Sakhr Automatic Reader (i.e., Arabic OCR). To quantify the OCR accuracy improvement we used the UMD Accuracy tool, which implemented an approximate string matching algorithm[4]. This tool enabled us to compare each OCR output text file against the corresponding ground truth text file.

The approach to the exploratory study is depicted in Figure 2.

3.1. Exploring software settings

We selected a small document set from the MDAD corpus that included some of the more degraded documents. Various settings in Fujitsu TWAIN32, Scanfix, and Kofax VRS were explored in iterative trial and error tests to determine which image enhancement capabilities provided visual improvement. The best results were processed by OCR and the accuracy tool to quantify the improvement.

Out of all of the capabilities tested, quantifiable improvements were observed in only a few settings. Lightening the scanner brightness and isolated despeckle (up to a 7x7 window) generally improved noisy documents. Dropping out the correct RGB color value for color paper documents made previously unreadable documents into high quality documents. Scanfix's erosion operation worked well for cleaning up glyphs after despeckling noisy documents. Scanfix's sand and fill operation smoothed jagged-edged characters inherent in faxed documents or resulting from erosion or reduction in scan brightness.

The Kofax Virtual Re-scan (VRS) tool enabled us to explore the effects of gamma correction on scans. On documents with color damage areas (e.g., food stains, burns), a high gamma correction setting virtually eliminated the degradation. Also, it was fairly effective to eliminate the effects of soil and wrinkles. However, it had the side-effect of washing out the inside of glyphs (leaving only an outline), which resulted in poor OCR results.

Scanfix provided a line detection and removal function. However when lines were on the top of characters (e.g., with graph paper), the characters were destroyed during line removal. Blob removal was applied to hole-punches and fire-damaged areas, however it did not perform well; it consistently removed characters along with blobs.

The OCR tool was particularly sensitive to speckles and jagged-edged characters. Characters that were slightly jagged resulted in recognition failures. Speckles were erroneously recognized as spurious characters or resulted in failures. The tool skew-corrected some documents based on the slant of the rubber stamp text, rather than the body text. Soiled and wrinkled fax documents caused a paradox for the OCR tool; if the scan brightness was reduced to remove the soil/wrinkle degradation the inside of glyphs were washed out. As mentioned above, this posed problems for OCR.

Paper/ Damage Color	Glyphs	Other Attribute	Document	Default Scan	After Fujitsu Enhancement	After both Fujitsu and ScanFix Enhancement	Increase due to Fujitsu	Increase after Fujitsu and ScanFix	Increase due to ScanFix
Blue	Smooth	Soiled/Wrinkled/Folded	ANN20001201.0900.0069	0.00%	0.00%	8.93%	0.00%	8.93%	8.93%
Blue	Smooth	Isolated Damage	AFA20001019.1800.0178	69.72%	75.04%	78.45%	5.32%	8.73%	3.41%
Blue	Smooth	Soiled/Wrinkled/Folded	NTV20001222.1530.1553	0.00%	75.84%	76.06%	75.84%	76.06%	0.22%
Blue	Smooth	Isolated Damage	ANN20001011.1500.0073	7.41%	76.16%	79.85%	68.75%	72.44%	3.69%
Blue	Smooth	Gray Shaded	VAR20001118.1100.2219	77.10%	90.72%	91.60%	13.62%	14.50%	0.88%
Green	Smooth	Gray Shaded	VAR20001212.1100.1070	67.28%	0	64.49%	0.00%	0.00%	0.00%
Green	Smooth	Soiled/Wrinkled/Folded	ANN20001202.1500.0082	0.00%	42.61%	57.67%	42.61%	57.67%	15.06%
Green	Smooth	Isolated Damage	AFA20001208.1400.0080	54.61%	72.04%	74.05%	17.43%	19.44%	2.01%
Green	Smooth	Soiled/Wrinkled/Folded	ALH20001121.1300.0096	0.00%	76.32%	77.07%	76.32%	77.07%	0.75%
Green	Smooth	Isolated Damage	NTV20001215.1530.0490	82.71%	91.10%	92.02%	8.39%	9.31%	0.92%
None	Jagged	Gray Shaded	VAR20001114.1100.0587	0.00%	0	0	0.00%	0.00%	0.00%
None	Jagged	Gray Shaded + Soiled	AFA20001207.1000.0039	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
None	Jagged	Gray Shaded + Soiled	VAR20001210.1100.1230	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
None	Smooth	Gray Shaded + Soiled	VAR20001109.1200.1544	0.00%	0.00%	58.94%	0.00%	58.94%	58.94%
None	Jagged	Soiled/Wrinkled/Folded	AFA20001223.1800.0112	0.00%	0.00%	18.01%	0.00%	18.01%	18.01%
None	Jagged	Soiled/Wrinkled/Folded	ALH20001208.1900.0105	0.00%	19.61%	43.04%	19.61%	43.04%	23.43%
None	Smooth	Gray Shaded	VAR20001205.1100.0773	0.00%	36.66%	77.60%	36.66%	77.60%	40.94%
None	Jagged	None	ALH20001104.0700.0055	48.60%	45.81%	46.77%	0.00%	0.00%	0.00%
None	Jagged	None	AFA20001124.1800.0159	55.84%	56.46%	58.44%	0.62%	2.60%	1.98%
None	Smooth	Soiled/Wrinkled/Folded	AFA20001018.1000.0062	0.00%	63.47%	71.43%	63.47%	71.43%	7.96%
None	Smooth	Gray Shaded	VAR20001224.1100.1102	0.00%	71.57%	77.61%	71.57%	77.61%	6.04%
None	Smooth	Soiled/Wrinkled/Folded	NTV20001013.1530.0001	0.00%	72.62%	57.65%	72.62%	57.65%	0.00%
None	Smooth	Isolated Damage	AFA20001019.1400.0115	70.43%	76.57%	78.17%	6.14%	7.74%	1.60%
None	Smooth	Isolated Damage	ALH20001106.1300.0100	80.61%	83.84%	84.20%	3.23%	3.59%	0.36%
None	Smooth	Gray Shaded + Soiled	VAR20001111.1200.1204	0.00%	85.99%	85.99%	85.99%	85.99%	0.00%
Red	Smooth	Soiled/Wrinkled/Folded	ANN20001109.0900.0083	0.20%	0.00%	1.38%	0.00%	1.18%	1.18%
Red	Smooth	Soiled/Wrinkled/Folded	AFA20001205.1400.0100	1.04%	67.70%	81.12%	66.66%	80.08%	13.42%
Red	Smooth	Isolated Damage	AFA20001112.1400.0162	67.98%	77.69%	79.51%	9.71%	11.53%	1.82%
Red	Smooth	Isolated Damage	AFA20001205.0000.0010	0	85.94%	86.50%	85.94%	86.50%	0.56%

Table 1. Accuracy of degraded document OCR, before and after image enhancement.

3.2. Sorting hardcopies and assigning settings

Next, the documents were sorted into bins based on visible document qualities and prior observations. The bins were designed to exploit the strengths of the enhancement tools and compensate for Sakhr's character recognition weaknesses, while keeping the number of bins reasonable. The sort order was determined to be: 1. paper color or damage color[†], 2. presence of jagged-edged glyphs, and 3. other attribute[‡]. Noting the dominant quality of each document simplified the sorting process.

The prior trial and error testing gave us insight to tailor helpful settings for each bin. Settings were assigned to each bin and included combinations of brightness level adjustments, dropout color, despeckle, erode, and sand and fill – settings found in Fujitsu TWAIN32 and Scanfix.

3.3. Scanning, OCR, and measurement

Applying the assigned image enhancement settings, three types of images were captured for each document:

- Without image enhancement,
- With Fujitsu TWAIN32 image enhancement, and
- With both Fujitsu TWAIN32 and ScanFix image enhancement.

The MDAD corpus default scans already established the images without image enhancement. The dynamic threshold capability (i.e., SDTC) was disabled in order to gain full control of the scan brightness. Discovering the ideal brightness setting involved re-scanning and reducing the brightness setting repeatedly until white pixels appeared inside glyphs. The last scan with solid black glyphs was selected as the optimal scan.

The three types of images for each document were then processed through the OCR tool. CP1256 files were output and compared against the ground truth using the UMD accuracy tool. The results are shown in Table 1.

We discovered that the evaluation metrics may not be reflecting the OCR output well. We have already mentioned that the OCR tool expects clean documents and on noisy documents it attempts to recognize speckles as characters. For noisy documents, the OCR tool produced several failure characters in the output file or caused Automatic Reader to abnormally end. Since accuracy was calculated as the number of correct characters minus error characters, divided by the number of correct characters, the tool produced negative and zero values. No output was obtained when Sakhr abnormally ended. Therefore, we interpreted both the lack of output and negative values to be zero accuracy.

3.4. Drawing conclusions

Analyzing the data resulted in several conclusions about scanning and image enhancement. Selecting the appropriate dropout color had a tremendous benefit to the OCR accuracy of color paper documents. This benefit will depend on the presence of color paper in the document set. The general OCR accuracy and related improvement of faxes was low. Although Scanfix sand and fill provided some improvement on clean faxes, the overall improvement is questionable. All of the non-color, smooth glyph documents could be improved with a brightness adjustment. This includes documents with gray shading, gray shading and soil, soil/wrinkles/folds, and isolated damage. We concluded that significant OCR accuracy improvement on degraded documents can be primarily gained with a brightness level adjustment. Since brightness adjustment is a function of the scanner driver, scanner users already have the best tool for image enhancement at their disposal.[§] We also concluded that image enhancement of degraded documents to produce high quality images is vital for Sakhr OCR performance. Further research is necessary to isolate which settings are helpful and prove their benefit.

4. CONCLUSIONS

In this paper we presented development of a manually degraded Arabic document corpus and discussed its use in an exploratory study of image enhancement. The corpus includes 206 Arabic documents in 1352 TIFF files. The exploratory study concluded that significant OCR accuracy improvement on degraded documents can be primarily gained with brightness level adjustment. Since huge gains can be obtained simply with brightness adjustment, scanner users already have the best tool for image enhancement at their disposal[**].

[†] Red, green, blue, no color (including bright color paper such as yellow lined paper, manilla, yellow construction paper)

[‡] For example, soil stain, gray shading, and isolated damage, including the degradation categories mentioned previously

[§] Third party image enhancement resulted in small OCR improvements (despeckle, erode, sand and fill).

[**] The methods of manual or automatic brightness control operations vary among scanner models.

Applying more time on developing the degradation approach might have avoided some of the challenges we experienced. Although the design of a degradation approach is tailored to the application of interest, it seems all approaches should include determining the following:

- the set of treatments to be applied
- the desired combination and sequencing of treatments
- the quantity of treatment combinations per document
- whether the quantity of combinations per document should be consistent or vary across the corpus
- whether a hardcopy is required at each point in a document's life cycle
- the granularity to which applied treatments will be tracked (e.g., if applying annotations, is it necessary to track the quantity of annotations per document?)
- treatment variations (e.g., if applying fax treatment, is it necessary to track whether the faxes are thermal transfer or inkjet?)
- whether treatments leaving residue should be cleaned prior to scanning[††].

The Fujitsu TWAIN32 scanner driver used only allows control of a single threshold for creating bi-tonal images[‡‡]. We briefly investigated the use of two thresholds[§§] for image enhancement on a color image, followed by conversion to bi-tonal. With the MDAD corpus, we found that text appears in a relatively constrained tonal range. Keeping only pixels within this range and pixels for which the RGB color values were close (e.g., 1/100 of the constrained tonal range) was useful for segmenting text from document noise. We plan to continue investigation into the use of dual thresholds on color images to create high quality bi-tonal images of degraded documents.

REFERENCES

1. TIDES Extraction (ACE) 2003 Multilingual Training Data. LDC 2004T09 / ISBN 1-58563-292-9) http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2004T09

2. Y. Zheng, H. Li, D. Doermann, "Machine Printed Text and Handwriting Identification in Noisy Document Images." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(3), pages 337-353, 2004.

3. D. Doermann, G. Zi, "Groundtruth Image Generation from Electronic Text (Demonstration)". *Symposium on Document Image Understanding Technology*, pages 309-312, APR 2003.

4. E. Ukkonen, "Algorithms for approximate string matching". *Information and Control* 64 (1985), 100-118.

5. D. A. Lehotsky (DALSA Inc.), "Intelligent High-sensitivity CCD Line Scan Camera with Embedded Image Processing Algorithms". *Proceedings of the International Society for Optical Engineering (SPIE)* Vol. 3966 Machine Vision Applications in Industrial Inspection VIII. January 2000.

6. Y. Yang, K. Summers, M. Turner, "A Text Image Enhancement System Based on Segmentation and Classification Methods". In *Proceedings of the First ACM Hardcopy Document Processing (HDP) Workshop 2004.*

[††] With some degraded documents, residue from the treatment can obstruct the text. The judgment to clean the residue depends on the resiliency of the paper and ink. When building a corpus, the effect of cleaning treatments can be tested in advance. Cleaning was a reasonable step for the MDAD corpus and did not harm the paper or ink.

[‡‡] Below the threshold pixels are scanned as black, above the threshold pixels are scanned as white.

[§§] Similar to the one described in Lehotsky[5].

APPENDIX

This appendix shows examples from the MDAD corpus. In each figure the image on the left reflects the original document, and the image on the right is the corresponding bi-tonal image.

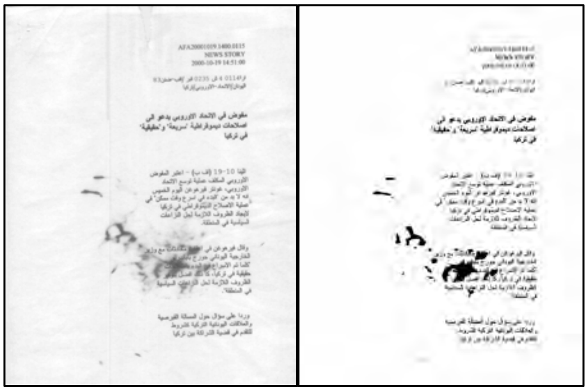

Figure A1. Degree 1, fire-damaged, thin paper.

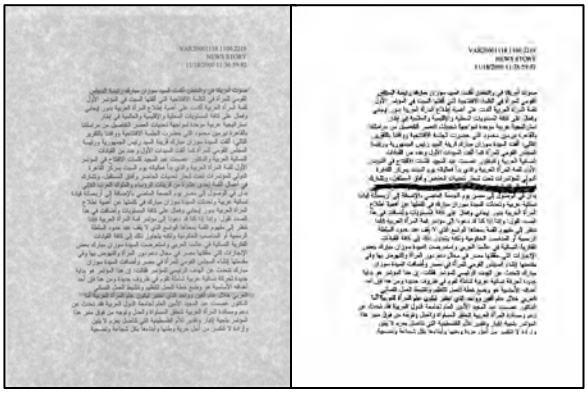

Figure A2. Degree 1, highlighter pen, textured paper.

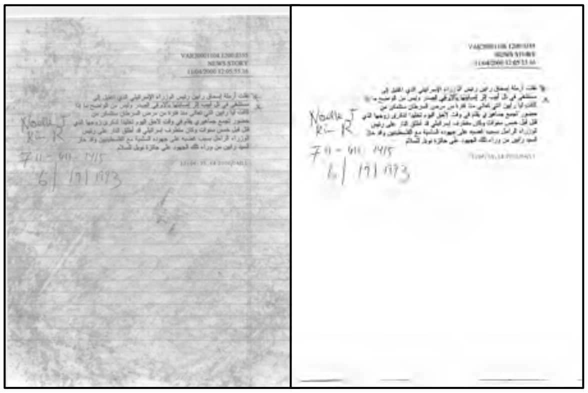

Figure A3. Degree 2, annotation, lined paper.

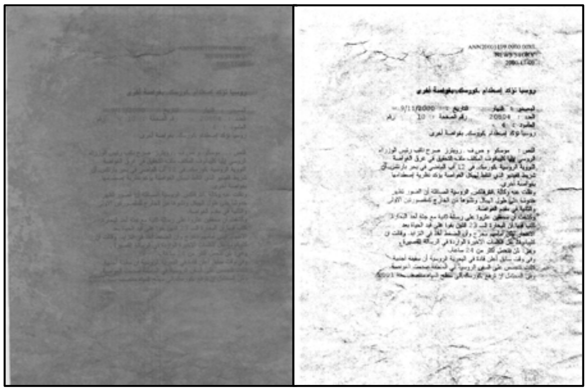

Figure A4. Degree 2, highlighter pen, colored paper.

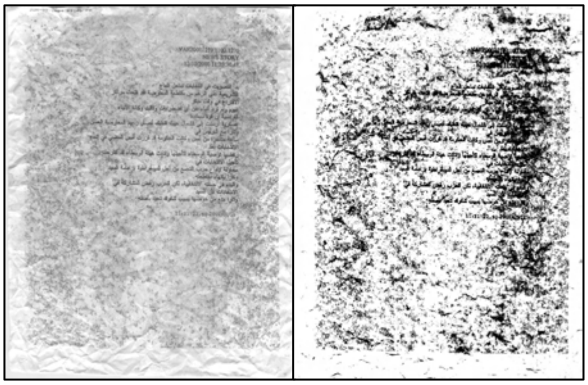

Figure A5. Degree 3, fax, textured paper.

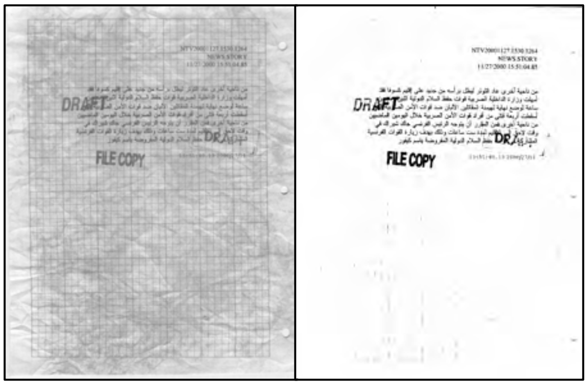

Figure A7. Degree 3, rubber stamp, graph paper.

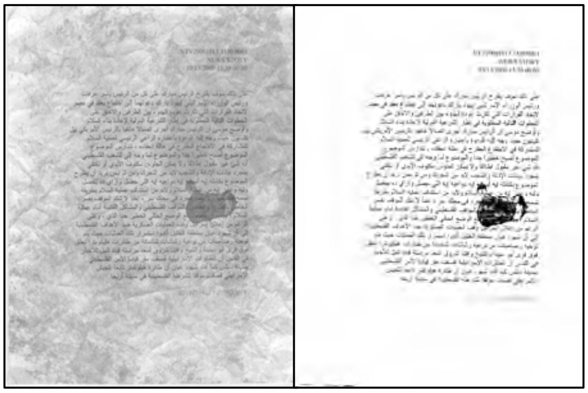

Figure A6. Degree 3, food stain.

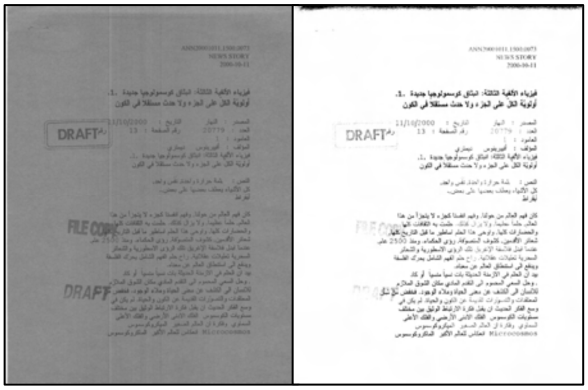

Figure A8. Degree 1, rubber stamp, colored paper.