1-1-2008

Correlating Degradation Models and Image Quality Metrics

Darrin K. Reed
Boise State University

Elisa H. Barney Smith
Boise State University

 DOI: 10.1117/12.766784

Correlating Degradation Models and Image Quality Metrics

Darrin K. Reed and Elisa H. Barney Smith

Electrical and Computer Engineering

Boise State University

Boise, ID 83725-2075 USA

E-mail: EBarneySmith@boisestate.edu

URL: http://coen.boisestate.edu/EBarneySmith

ABSTRACT

OCR often performs poorly on degraded documents. One approach to improving performance is to determine a good filter to improve the appearance of the document image before sending it to the OCR engine. Quality metrics have been measured in document images to determine what type of filtering would most likely improve the OCR response for that document image. In this paper those same quality metrics are measured for several word images degraded by known parameters in a document degradation model. The correlation between the degradation model parameters and the quality metrics is measured. High correlations do appear in many places that were expected. They are also absent in some expected places and offer a comparison of quality metric definitions proposed by different authors.

Keywords: Degradation model, Quality metrics, Edge spread, Bilevel image, Synthetic word images

1. INTRODUCTION

High quality Optical Character Recognition (OCR) depends on high quality images. Unfortunately, these high quality images are not always available. Since low quality document images have high OCR error rates, corrections are required making the OCR process tedious and expensive. Throssel in 1972 developed a system to measure the paper and ink with goals of assessing the print quality of documents in a way that would correlate with OCR accuracy [15]. This required special hardware and an additional scanning step. In the mid 1990's an approach was taken to identify for which pages it would be more efficient to process through OCR and make corrections, versus through hand entering the text [7, 10]. These methods utilized quality metrics calibrated from the image scanned for OCR. More recently an approach has surfaced where authors are designing systems to automatically choose a filter that when applied to the degraded document image improves the image quality and therefore the OCR performance [9, 13, 14]. Here the choice of filter is related to the level and type of degradation present in the document image.

The methods used to predict the OCR accuracy and to choose the best image filter both rely on measuring features, called quality metrics, from the document image that indicate the types of degradations present in the image. Broken and touching characters have been identified as leading causes of OCR errors [8, 12]. Based on this, many quality metrics attempt to measure the presence of broken or touching characters along with background speckle in the image.

Other research to improve OCR focuses on modeling the degradations that occur in document images. It is believed that the OCR engines can be designed using the information provided by these models to be more immune to the degradations that could be present in document images. The degradation model in [1] has parameters to cover all aspects of a document image production: resolution, blur (Point Spread Function or PSF width), binarization threshold, sensitivity (additive noise), jitter, skew, character width and height, baseline location, and kerning. Of these model elements, the PSF width, w, binarization threshold, Θ, and additive noise, s, have been identified as the three most significant parameters in this degradation model affecting degradations of bilevel images [11].

Some secondary parameters based on the model from [1] have been defined that give quantitative parameters to measure observed differences in the images. The amount that an edge is displaced, δ_c [3], and the amount a black or white corner is

Document Recognition and Retrieval XV, edited by Berrin A. Yanikoglu, Kathrin Berkner,
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 6815, 681508, © 2008 SPIE-IS&T · 0277-786X/08/$18

eroded, d_b and d_w [4, 16]. Later work showed that the images are more affected by the edge spread [2] than they are by the corner erosion. It was also shown that OCR accuracy could be improved by grouping characters with common edge spreads [5].

This paper explores whether the quality metrics are correlated to either the primary or secondary degradation model parameters. The existence of a relationship could influence decisions on how to take measurements from degraded documents. Strong enough correlations could allow for the use of quality metrics as a more sophisticated tool for OCR training. Since the degradation model parameters would be directly available, the model could be directly used to choose the best filtering system. If the model more closely represents the degradations to which a document has been subjected, pre-processing can be chosen to better fit the image. Parameter knowledge could also be used to determine the global degradation space. Division of the global degradation space allows the OCR system to be trained on data that more closely represents the testing data. Experiments from [5] showed that fewer errors will occur if training data more specific to the document being sent through OCR is used. Further advancements could lead to ideas of other metrics to measure and to the efficient estimation of degradation model parameters. This would open up a new direction of research drawing from both of these fields.

The specifics of the degradation model used in this paper are introduced in Section 2, and the quality metrics being used are described in Section 3. A description of the experiments are provided in Section 4 with the resulting correlations presented in Section 5. The results are summarized in Section 6.

2. DEGRADATION MODEL

The degradation model of [1] was implemented with just the PSF width, w, the binarization threshold, Θ, and the additive noise, s, components as shown in Figure 1. The bilevel image is blurred with the PSF through two dimensional convolution. Gaussian noise of standard deviation, s, is added to the resulting gray scale image. Then a global threshold of Θ is applied to return the image to a bilevel image. The PSF accounts for the blurring caused by the optics of the scanner. The blurring model is implemented by convolving the PSF with a high resolution image template and then sub-sampling the blurred image. In this work, the PSF is assumed to be a bivariate Gaussian with the width, w, equal to the standard deviation, measured in units of pixels. Additive noise is also incorporated in this degradation model. The noise is Gaussian distributed with a standard deviation, s, and is added independently to each pixel in the image prior to thresholding. The resulting gray level image is converted to a bilevel image with a global threshold, Θ. The units for the threshold are absorptance so paper has values of 0 and ink has values of 1.

The stroke width is determined by the location of the edges of the stroke. The stroke width will change as the edge locations move due to effects of the degradation model. The distance an edge is displaced depends on the threshold, the PSF width, and the functional form of the PSF. During scanning, an edge separated from other edges by a distance greater than the support of the PSF changes from a step to an edge spread function, ESF, through convolution with the PSF. This is then thresholded to reform a step edge, as shown in Figure 2. The amount an edge was displaced after scanning, δ_c, was shown in [3] to be

$$\delta_c = -w\, \mathrm{ESF}^{-1}(\Theta) \tag{1}$$

and is called edge spread. This is a secondary degradation parameter and is used in the experiments and analysis.

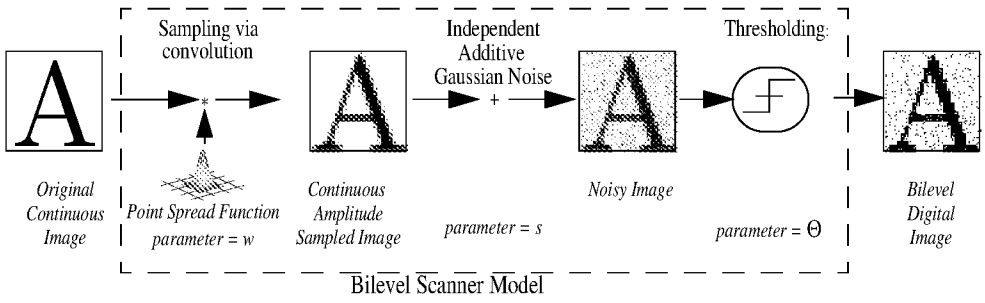

Figure 1: Degradation Model

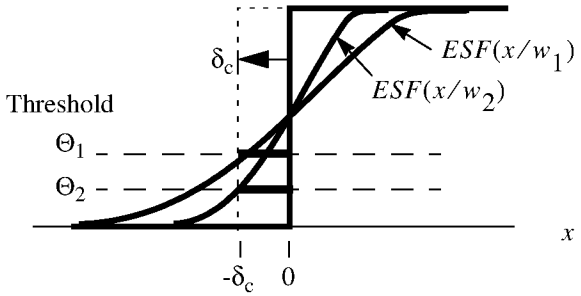

Figure 2: Edge after blurring with a generic PSF at two widths, *w*. The two thresholds shown produce the same edge shift δ_c.

3. QUALITY METRICS

Quality metrics are not based on a theoretical degradation model, but instead are empirically measured from an image. Most are defined to quantitatively capture the different types of degradations seen in large document collections. Eight different quality measures defined in [7, 9, 10, 13] were implemented for these experiments. Some quality metrics aim to measure the same effect, but the method used to calculate the degradation effect has been defined differently by different authors. These metrics were chosen by the original authors because they experimentally had a high correlation with the OCR error rate. In this paper we will see if some of them also have a high correlation with either primary or secondary degradation model parameters. All quality metrics used in this paper are described next. The reader is encouraged to use the original source for a detailed description of the algorithms used to calculate these metrics.

Font Size (FS): Font Size is not a quality metric per se; its use is strictly for calculating the quality metrics described next. Many of the quality metric definitions are made relative to the size of the font used in the document being evaluated. This can affect whether dots on i's or apostrophes, commas and periods are considered noise or text. For these metrics, the Font Size (FS) refers to the average measured height in pixels of x-height characters in the text.

Small Speckle Factor (SSF) [9]: This measures the amount of black background speckle in the image. It identifies all the black connected components in an image that contain between 6 and FS pixels. This is normalized by the area under a histogram of connected component sizes between 6 and the FS squared.

Small Speckle Factor (SSF) [13]: This is similar to SSF defined in [9], but counts the number of connected components with fewer than 0.5*FS pixels.

Touching Character Factor (TCF) [9, 13]: The TCF measures the degree to which neighboring characters touch. This is looking for long and low connected components. This was defined by the connected component height to width ratio being less than 3/4. Additionally, it is required that the number of pixels in the connected component must be greater than 3*FS, the height of the connected component must be between 0.75*FS and 2*FS.

Stroke Thickness Factor (STF) [13]: This is the most frequently measured thickness in the horizontal direction. At each row of the text lines, the number of consecutive black pixels is counted. The peak is extracted from a histogram of all stroke thicknesses measured.

White Speckle Factor (WSF) [9]: The WSF measures the degree to which fattened character strokes have shrunken existing holes in or gaps between characters causing several small white islands to form. This is the area under the histogram of white connected components between 1 pixel and $0.01*(FS)^2$ pixels, normalized by the area under the same histogram between 1 pixel and $(FS)^2$ pixels. This is a modification of the BCF defined in [7].

White Speckle Factor (WSF) [7, 13]: The quality metric defined by [13] was based on and defined identical to that WSF defined by [7]. The number of white connected components less than 3x3 pixels in size is divided by the number of white connected components in the total image.

Broken Character Factor (BCF) [9]: The BCF measures the number of thin connected components based on the assump-

(a) (b) (c) (d) (e)

Figure 3: Sample words from the 18,144 words depicting high measured values of the five different quality metrics, (a) SSF (b) WSF (c) TCF (d) BCF (e) STF

tion that if characters are broken, there will be many connected components that are thinner than if the characters were not broken. Here the connected component height and width must both be less than 0.75*FS and the number of pixels in the connected component must be greater than FS.

Broken Character Factor (BCF) [13]: This version of BCF uses a 2-D histogram of the heights and widths of black connected components. Only connected components with height and width smaller than 0.75*FS are counted. The number of cells in that histogram that are occupied (area of footprint) is divided by the maximum possible footprint, $(FS)^2$. This is a modification of the BCF defined in [7, 10].

Words degraded through the model by [1] illustrating five different quality metrics can be seen in Figure 3. Each image was measured to have above average values for it's particular quality metric. The SSF can most easily be described as a measurement of the amount of background noise in an image as clearly shown in Figure 3(a). WSF is generally used to measure the closing of closed loops due to thicker stroke-widths, but as shown in Figure 3(b) some contribution due to noise near the edges of the characters can be seen. The connection of neighboring characters will change a connected components block for a single character from approximately a tall, slender rectangle to a shorter, wider connected components rectangle of multiple characters, thus, increasing the TCF metric. This tendency can be seen in the touching of the 'r' and 'e' in Figure 3(c). Regions of thinner character strokes are quicker to break and contribute to BCF. Seriffed words have thinner stroke widths in curved regions; Figure 3(d) shows a Times Roman version of the word 'are' where the letters 'a' and 'e' break in the thinned, curved sections. Because multiple degradations are affected by the thickness of the stroke, STF quantifies this characteristic. The version of 'but' in Figure 3(e) shows a much thicker stroke thickness which was measured to have a higher STF value.

4. IMPLEMENTATION

Experiments to measure each quality metric were conducted on a set of isolated words degraded by the degradation model [1]. Thirty-two words were selected for study, Figure 4. These are words of at least three characters in length selected from the most common 50 English words [6]. The 25 most common words make up about one third of all printed material in English and the first 100 make up about one-half of all written material in English, so even with the one and two letter words being omitted, this is representative of a good portion of the text that would be in a common scanned document.

Images of the words in Figure 4 were generated in both Times New Roman and Arial fonts to see if there were any differences between seriff and sans-seriff characters in the relationship between the quality metrics and the degradation model

1. the	18. his	31. but	40. said
3. and	19. they	32. not	41. there
8. you	23. this	33. what	42. use
9. that	24. have	34. all	44. each
12. was	25. from	35. were	45. which
13. for	27. one	37. when	46. she
15. are	28. had	38. your	48. how
17. with	30. word	39. can	49. their

Figure 4: Words of three or more letters from the 50 most common English words with their occurrence ranking indicated [6].

parameters. The words were created in Microsoft Word at 260 point and output directly to an image file at 400 dpi using the Doc to Image Converter software [17]. This high resolution image was then used as the model word input to the degradation model. This allowed the natural spacing that would be found in these words to be present and also include effects of kerning where appropriate. No ligatures were present in this set of words, but had there been, this method would have included those where appropriate. The small differences in the character shapes due to hinting are not expected to be significant in this work.

The words were degraded by the degradation model of [1] with a wide range of model parameters for PSF width, threshold and noise to produce word samples at 12 point, 300 dpi. The characters were degraded by 9 thresholds, Θ, ranging from 0.1 to 0.9, three widths, w, 1, 1.5 and 2, and 21 levels of noise, s, ranging from 0 to 0.1 producing 567 instances of each word and 18,144 total degraded word images.

Several of the quality metrics of [9] required normalization by a quantity in a histogram of all the values measured for that metric on a given page. As the experiments in this paper were run on individual words instead of full pages the normalization was not practical to implement, so the raw quality metric values were used instead.

5. RESULTS AND ANALYSIS

A number of synthetic word images were created based on the degradation model with known model parameters. Each of the 8 quality metrics was measured for each of the degraded word images. Both N_4 and N_8 were implemented because the quality metric authors did not specify which type of connected component they intended for their algorithms. Correlation analysis was conducted between both the primary and secondary degradation model parameters and each of the 8 quality metrics. Calculations of the two word sets, Arial and Times New Roman, were completed separately. The resulting correlations are shown in Tables 1 and 2.

The small speckle factor is expected to be higher when the noise is larger. A correlation was found between the noise added and the SSF. This was only found consistently in the SSF as defined by [16]. This was likely because [9] ignored any connected components with six connected pixels or less.

Because the edge spread parameter, δ_c, measures the amount a character stroke is increased or decreased, some correlation between this secondary parameter and the quality metrics described above is expected. Likewise high thresholds are expected to produce thin and broken characters and low thresholds are expected to produce fat and touching characters. Since an increased character stroke is the result of a reduced threshold which holds consistent with the definition of edge spread in Equation 1, the edge spread and threshold are inversely related. In all cases the threshold showed a correlation negative of the correlation measured with edge spread.

Table 1: Correlations between quality metrics and degradation model parameters for Arial words.

	SSF-N_4 [9]	SSF-N_8 [9]	SSF-N_4 [13]	SSF-N_8 [13]	TCF-N_4 [9,13]	TCF-N_8 [9,13]	STF [13]
Noise	0.179	0.189	**0.530**	**0.453**	-0.229	-0.199	-0.220
Width	-0.025	-0.030	0.249	0.199	0.178	0.213	-0.053
Threshold	0.414	0.337	0.081	0.062	-0.314	-0.351	**-0.468**
Edge Spread	-0.330	-0.255	-0.035	-0.034	0.364	0.403	**0.467**
	WSF-N_4 [9]	WSF-N_8 [9]	WSF-N_4 [7,13]	WSF-N_8 [7,13]	BCF-N_4 [9]	BCF-N_8 [9]	BCF [13]
Noise	0.376	0.217	**0.538**	0.319	-0.085	-0.071	0.411
Width	0.238	0.180	0.311	0.326	0.093	0.109	0.164
Threshold	**-0.466**	**-0.560**	-0.345	-0.415	0.174	0.186	-0.246
Edge Spread	**0.527**	**0.623**	**0.409**	**0.491**	-0.162	-0.176	0.280

Table 2: Correlations between quality metrics and degradation model parameters for Times New Roman words.

	SSF-N_4 [9]	SSF-N_8 [9]	SSF-N_4 [13]	SSF-N_8 [13]	TCF-N_4 [9,13]	TCF-N_8 [9,13]	STF [13]
Noise	0.147	0.178	**0.585**	**0.505**	-0.250	-0.227	-0.263
Width	-0.048	-0.029	0.109	0.103	0.063	0.081	-0.129
Threshold	0.259	0.189	-0.051	-0.039	-0.342	-0.374	**-0.588**
Edge Spread	-0.183	-0.112	0.131	0.104	0.360	0.395	**0.578**
	WSF-N_4 [9]	WSF-N_8 [9]	WSF-N_4 [7,13]	WSF-N_8 [7,13]	BCF-N_4 [9]	BCF-N_8 [9]	BCF [13]
Noise	0.328	0.193	**0.463**	0.259	-0.132	-0.100	0.364
Width	0.218	0.177	0.294	0.295	-0.063	-0.042	0.027
Threshold	**-0.483**	**-0.530**	**-0.504**	**-0.497**	-0.014	-0.001	-0.430
Edge Spread	**0.543**	**0.588**	**0.548**	**0.558**	0.052	0.041	**0.482**

The WSF, TCF and STF are expected to increase as δ_c increases because a larger positive δ_c indicates the stroke is thicker than in the original template. Thicker character strokes are expected to create more enclosed loops that should lead to white speckle, thus yielding a positive relationship between edge shift and WSF that was observed for the WSF by [9]. The WSF in [7, 13] also showed correlation with δ_c and Θ, but was not as strong as the correlation exhibited by WSF by [9]. The STF also proved to be correlated to edge shift and threshold, which is because a larger positive edge shift would result in a thicker character stroke.

It is surprising that TCF showed little correlation to the global threshold or edge shift. However, TCF did manage to show some relationship to threshold and edge shift on a single-word instance such as the Arial version of the word 'which'. The correlation between WSF and noise comes unexpectedly, but as shown in the noisy instance of 'which' (s=0.070, w=2.0, Θ=0.1) in Figure 5(a), one will observe that the excess noise created small white N_4 connected components around the edge of the character. Similarly, the Times Roman version of 'was' (s=0.080, w=2.0, Θ=0.1) in Figure 5(b) has a larger contribution to WSF due to noise around the character edges versus due to fattened character strokes in the letters 'a' & 's'. The correlation patterns that were observed were fairly consistent across all words in the data set. Many of the correlations were slightly stronger for the Times New Roman word set than they were for the Arial word set.

The PSF width parameter did not show much relationship to any of the quality metrics. This is likely because at small widths, the characters are not degraded much regardless of threshold, however characters generated with large widths at high threshold will be very thin and broken and at low thresholds will be very fat and touching.

For bilevel images N_4 components contribute to the connected component if a pixel is directly adjacent to another pixel of the same value, whereas, N_8 components contribute if a pixel is directly adjacent and/or diagonally adjacent. The results, in general, showed that w, δ_c and Θ were more strongly correlated to N_8 implementations of the metrics, excluding SSF.

(a) (b)

Figure 5: Sample word instance supporting the unexpected correlation found between WSF and noise

This makes sense because the N_8 variation would contribute more pixels to most metric measurements, but would contribute fewer pixels to SSF. Likewise, s showed stronger correlations to N_4 implementations of the metrics. This is with the exception of [9]'s implementation of SSF since their definition requires at least six pixels to contribute to SSF.

6. CONCLUSIONS AND FUTURE WORK

Many of the correlations that were expected were found, but some of the expected ones were absent and in the case of the WSF [7, 13], it has a reaction to noise that was unexpected. The threshold does represent the thinning and thickening of character strokes and is correlated with the quality metrics that relate to thinning and thickening of characters as expected. Since thresholding has a stronger effect at larger PSF widths than it does at smaller PSF widths, it has a smaller (absolute) correlation to these quality metrics than the edge spread does.

By using individual words instead of individual characters the interaction between neighboring characters was considered. Images are highly unlikely to have degradations that are influenced by characters in neighboring words in machine printed text. Small noise between words is included in these experiments, but large regions of speckle are not considered. Unless the large regions are interpreted as 'garbage' text causing OCR errors, this is not a problem as those speckles will not interfere with the recognition of the true characters. This experiment has also allowed a side-by-side comparison of the quality metrics by different authors.

Extensive experiments to estimate the OCR error rate as a function of w, Θ and s were run in [11]. As the original goal of the quality metrics was to estimate either the OCR error rate or the filtering necessary to most improve the error rate, the results from those experiments or repeating one similar to it would confirm the results presented here.

In future work the authors intend to expand on the results of these experiments to consider different point sizes and non-Latin scripts. The analysis structure in place lends itself to a quick transformation to evaluate degradations in non-Latin scripts. It will be determined if the correlations related to degradation model parameters hold only for Latin or also for others. Because the structure of various scripts is different, it may be possible to identify scripts in multi-script documents based on the quality metric results. For example, due to the abundance of enclosed loops in Chinese scripts, it is perceivable that the WSF quality metric would be greater for Chinese scripts than WSF for Latin scripts. Also, short and long connected scripts such as Arabic could yield TCF quality metric results different than any other script types. The relationship between the degradation model and the OCR algorithms for non-Latin fonts should be similar, but since different features are used for different scripts, script identification through quality metrics would be useful prior to OCR.

Some of the quality metric algorithms called for normalization based on the statistics observed across a whole page image. That was not feasible to do under the structure set out for this experiment so the raw, un-normalized values were used instead. Future experiments using full pages of text in the correlation analysis with the quality metric author's intended normalizations implemented are planned.

Most documents that give poor OCR performance are those that have been photocopied. Future experiments running the degradation model twice or thrice in succession would enable analysis in that direction, although inclusion of a printer model would be beneficial.

7. ACKNOWLEDGEMENT

This material is based on work supported by the National Science Foundation under grant No. CCR-0238285. The authors would like to thank Jim Steele for his help in coding the Quality Metric extraction functions.

8. REFERENCES

1. Henry S. Baird, "Document Image Defect Models," *Proc. IAPR Workshop on Syntactic and Structural Pattern Recognition*, Murry Hill, NJ, June 1990, pp. 13-15. Reprinted in H. S. Baird, H. Bunke, and K. Yamamoto (Eds.), *Structured Document Image Analysis*, Springer Verlag: New York, 1992, pp. 546-556.
2. Elisa H. Barney Smith and Xiaohui Qiu, "Statistical image differences, degradation featuresand character distance metrics," International Journal of Document Analysis and Recognition, Vol. 6, No. 3, 204, pp. 146-153.
3. Elisa H. Barney Smith, "Characterization of Image Degradation Caused by Scanning," Pattern Recognition Letters, Vol. 19, No. 13, 1998, pp. 1191-1197.

4. Elisa H. Barney Smith, "Estimating Scanning Characteristics from Corners in Bilevel Images," Proceedings SPIE Document Recognition and Retrieval VIII, Vol. 4307, San Jose, CA, 21-26 January 2001, pp. 176-183.
5. Elisa H. Barney Smith and Tim Andersen, "Text Degradations and OCR Training," Proceedings International Conference on Document Analysis and Recognition 2005, Seoul, Korea, 29 August - 1 September 2005, pp. 834-838.
6. Edward Bernard Fry, Jacqueline E. Kress and Dona Lee Fountoukidis, "The Reading Teachers Book of Lists," Third Edition, http://www.duboislc.org/EducationWatch/First100Words.html .
7. Luis R. Blando, Junichi Kanai, and Thomas A. Nartker, "Prediction of OCR Accuracy Using Simple Features," Proceedings of the Third International Conference on Document Analysis and Recognition, Montreal, Canada, 14-16 August 1995, pp. 319-322.
8. Mindy Bosker, "Omnidocument Technologies," Proceedings of the IEEE, vol. 80, no. 7, July 1992, pp. 1066-1078.
9. Michael Cannon, Judith Hochberg, Patrick Kelly, "Quality assessment and restoration of typewritten document images," International Journal on Document Analysis and Recognition, vol 2., 1999, pp. 80-89.
10. Juan Gonzalez, Junichi Kanai, Thomas A. Nartker, "Prediction of OCR Accuracy Using a Neural Network," Proceedings International Workshop on Document Analysis Systems, 1996, pp. 323-337.
11. Tin Kam Ho and Henry S. Baird, "Large-Scale Simulation Studies in Image Pattern Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, No. 10, October 1997, pp. 1067-1079.
12. Stephem V. Rice, Frank R. Jenkins, Thomas Nartker, "The Fourth Annual Test of OCR Accuracy," Information Science Research Institute 1995 Annual Report, 1995, pp. 11-49.
13. Andrea Souza, Mohamed Cheriet, Satoshi Naoi, Ching Y. Suen, "Automatic Filter Selection Using Image Quality Assessment," Proceedings International Conference on Document Analysis and Recognition, 2003, pp. 508-511.
14. Kristen Summers, "Document image improvement for OCR as a classification problem," Proceedings SPIE Document Recognition and Retrieval X, Vol. 5010, 2003, pp. 73-83.
15. W. R. Throssell and P. R. Fryer, "The Measurement of Print Quality for Optical Character Recognition Systems," Pattern Recognition, Vol. 6, 1974, pp. 141-147.
16. Hok Sum Yam and Elisa H. Barney Smith, "Estimating Degradation Model Parameters from Character Images," Proceedings International Conference on Document Analysis and Recognition 2003, Edinburgh, Scotland, 3-6 August 2003, pp. 710-714.
17. Doc to Image Converter - www.pdf-convert.com/doc2img/index.htm .