

Relating Statistical Image Differences and Degradation Features

Elisa Barney Smith¹ and Xiaohui Qiu²

¹ Boise State University, Boise, Idaho 83725, USA,
EBarneySmith@boisestate.edu,

<http://coen.boisestate.edu/EBarneySmith>

² Nanjing University of Post and Telecommunication, China

Abstract. Document images are degraded through bilevel processes such as scanning, printing, and photocopying. The resulting image degradations can be categorized based either on observable degradation features or on degradation model parameters. The degradation features can be related mathematically to model parameters. In this paper we statistically compare pairs of populations of degraded character images created with different model parameters. The changes in the probability that the characters are from different populations when the model parameters vary correlate with the relationship between observable degradation features and the model parameters. The paper also shows which features have the largest impact on the image.

1 Introduction

Document images can be degraded through processes such as scanning, printing and photocopying. This paper discusses bilevel degradations in the context of the scanning process. For the bilevel processes, two observable image degradations were described in [3]–[7]. These degradations are the amount an edge is displaced from its original location and the amount of erosion in a black or white corner. The variables that cause these image degradations can be related to the functional form of the degradation model: the PSF, the associated PSF width, and the binarization threshold.

From a calibrated model, one can predict how a document image will look after being subjected to the appropriate printing and scanning processes and, therefore, predict system performance. Large training sets of synthetic characters can be created using the model when the model parameters are matched to the source document. This can increase recognition accuracy. Models of the degradation process, along with estimates of the parameters for these models, can be combined to make a decision on whether a given document should be entered by hand or sent to an OCR routine [8,13,14]. A model will allow researchers to conduct controlled experiments to improve OCR and DIA performance. Knowledge of the system model parameters can also be used to determine which documents originated from the same source and, when the model includes multiple

printing/scanning steps, which document was the original and which was a later generation copy.

The degradation model used for this research is convolution followed by thresholding [1]. The two most significant parameters affecting degradations of bilevel images are the point spread function (PSF) width and the binarization threshold [9]. Each pair of these values will affect an image differently. However, several combinations of these parameters will affect images in a similar fashion. The PSF accounts for the blurring caused by the optics of the scanner. Its functional form is not constrained, but needs to be specified. A form that is circularly symmetric is usually chosen so its width is determined by one parameter. The size is in units of pixels, which allows the model to be used for scanning at any optical resolution. The threshold converts the image to a bilevel image. This is often done in software, and a global threshold is assumed. The units for the threshold are absorbance. The variations in the resulting bilevel bitmaps come largely from phase effects [12].

Several methods have been proposed to calibrate this model from bilevel images [2,4,6,7]. The resulting parameter estimates will never be error-free. However, not all errors are equally bad. Some will produce characters that have similar appearances and that are more likely to have the same response from an OCR system. This type of estimation error can be treated differently than estimation errors that result in a larger change in the character appearance. This paper explores the amount that characters made with different model parameters will differ as the true model parameters change.

Kanungo et al. [10] proposed a method of validating degradation models. This was achieved through a nonparametric two-sample permutation test. It decided whether two images are close enough to each other to have originated from the same source, having passed through the same sequence of systems. The application he proposed was to decide whether a model of a character degradation produced characters that were “close” to a set of “real” characters generated by physical printing and scanning. This testing could validate the degradation model and the choice of model parameters. The underlying statistical method is not restricted to comparing real and synthetic characters. Consequently, the two images could also be two real images, or two synthetic images. This statistical testing procedure can also be used to determine which parameters of the degradation model created the sample of characters. Another statistical device, the power function, was used to choose between algorithm variables.

Kanungo et al. demonstrated their method using a bit flipping and morphological degradation model. Their approach of statistically comparing character populations is applied in this paper to the convolution and thresholding degradation model shown in Figure 1. The parameters in this model are the point spread function (PSF) width, w , and the binarization threshold, θ . Both populations of character images were synthetically generated to see by how much the characters created with different model parameters will vary over the regions of the parameter space.

This paper starts by describing two image degradations and how they relate quantitatively to the degradation model parameters. It then describes the experiment conducted using Kanungo’s non-parametric permutation test to mathematically illustrate the size of the difference between two sets of degraded characters created using our model with different parameters. We then describe how the difference between characters relates to the degradations.

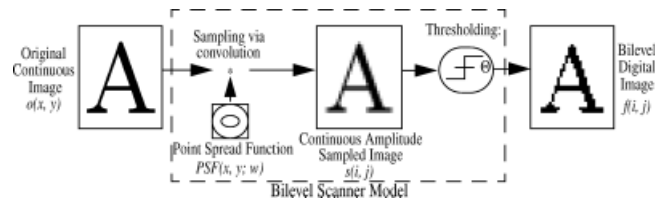


Fig. 1. Scanner model used to determine the value of the pixel (i, j) centered on each sensor element.

2 Image Degradations

Each model parameter set will produce a different character image. Examples of the characters that are produced for 600 dpi 12-point sans-serif font ‘W’ over a range of PSF widths and binarization thresholds are shown in Figure 2. Some of the degradations that are introduced are common to multiple characters, such as the final thickness of the character strokes, but each character is slightly different. Two primary image degradations associated with bilevel processes were defined in [4,5,6]. These are the edge displacement and the erosion of a black or white corner. All these degradations are functions of the degradation model parameters, w and Θ .

During scanning, the profile of an edge changes from a step to an edge spread function, ESF, through convolution with the PSF. This is then thresholded to reform a step edge, Figure 3. The amount an edge was displaced after scanning, δ_c , was shown in [3,4] to be related to w and Θ by

$$\delta_c = -wESF^{-1}(\Theta). \quad (1)$$

The edge spread determines the change in a stroke width after scanning. An infinite number of (w, Θ) values could produce any one δ_c value. Equation (1) holds when edges are considered in isolation, for example when the edges are separated by a distance greater than the support of the PSF. Figure 4 shows how the values of (w, Θ) vary for 5 different constant δ_c values for each of four PSF shapes. A positive threshold value will produce a negative edge displacement. The curves for δ_c and $-\delta_c$ are symmetric around the $\Theta=1/2$ line. If $\Theta=1/2$, then $\delta_c=0$ for all values of w .

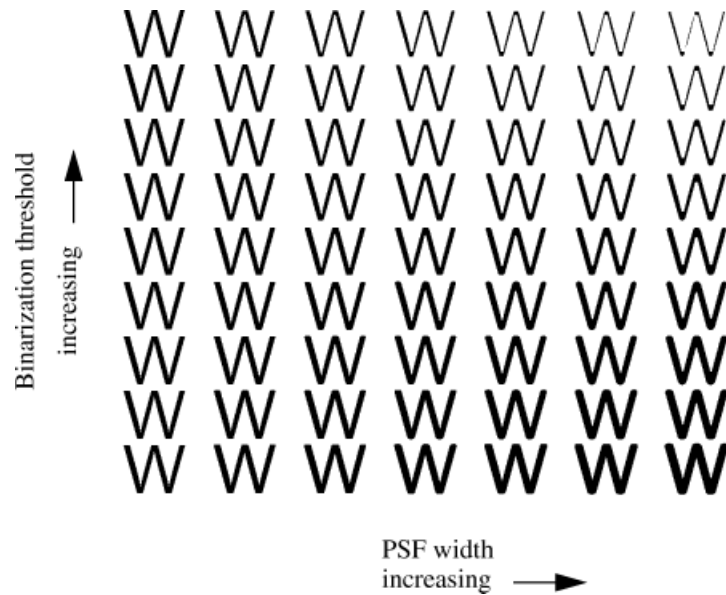


Fig. 2. Characters after blurring and thresholding over a range of PSF widths, w , and binarization thresholds, Θ . A broad range of character appearances can be seen, but some characters have general similarities.

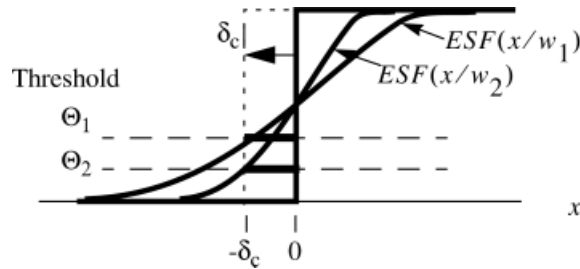


Fig. 3. Edge after blurring with a generic PSF of two widths, w . Two thresholds that produce the same edge shift δ_c are shown.

The other pair of bilevel image degradations are the amount of erosion seen in a black or a white corner after scanning [4,6]. This degradation is caused by the interaction of the two edges, but also includes the displacement of the individual edges. The erosion of a corner can occur in any of the three forms shown in Figure 5. Point p_0 is the apex of the original corner. Point p_2 is the point along the angle bisector of the new rounded corner where the blurred corner equals the threshold value. Point p_1 is the point where the new corner edges would intersect if extrapolated. The distance

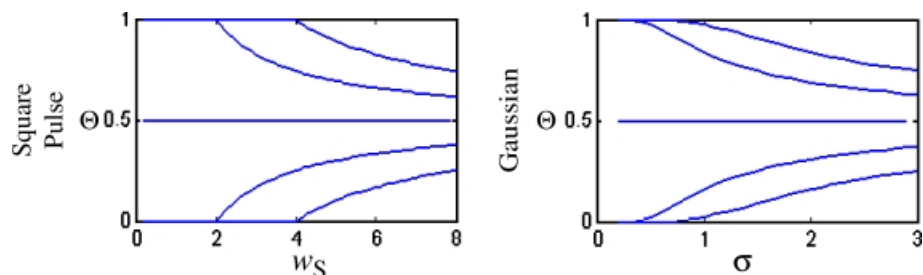


Fig. 4. Contours showing constant edge spread of $\delta_c = [-2 \ -1 \ 0 \ 1 \ 2]$ (from top to bottom) for two PSF functions.

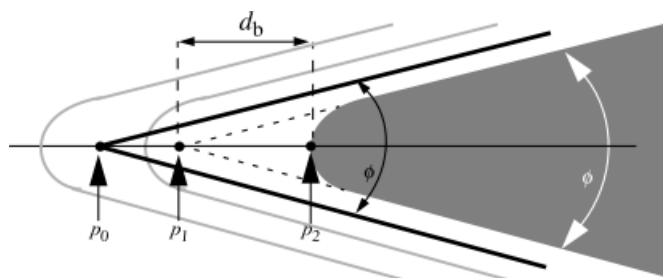


Fig. 5. The blurred corner (grey area and lines) may be displaced from the original corner position (black line) in three different ways. The visible erosion, d_b , is calculated the same for all three.

$$d_b = \overline{p_1 p_2} \quad (2)$$

is not the erosion from the original corner location, but it does represent the degradation actually seen on the corner, and this quantity can be measured from bilevel document images. The corner erosion distance, d_b , depends on the threshold, the PSF width, and the functional form similar to the edge displacement above.

The corner erosion distance is a combination of the distance from the original corner to the extrapolated corner, $\overline{p_1 p_0}$, which is based on the edge spread δ_c , and the distance along the angle bisector from the original corner to where the amplitude of the blurred corner equals the threshold, $\overline{p_1 p_2}$. Thus

$$\begin{aligned} d_b &= \overline{p_1 p_2} = \overline{p_1 p_0} + \overline{p_0 p_2} \\ &= \frac{-wESF^{-1}(\Theta)}{\sin(\phi/2)} + f_b^{-1}(\Theta; w, \phi) \end{aligned} \quad (3)$$

where

$$f_b(d_{0b}; w, \phi) = \int_{x=0}^{x=\infty} \int_{y=-x\tan\frac{\phi}{2}}^{y=x\tan\frac{\phi}{2}} PSF(x - d_{0b}, y; w) dy dx. \quad (4)$$

As with edge displacement, a given amount of corner erosion can also occur for an infinite number of (w, Θ) values. The erosion of a white corner is defined similarly and results in

$$d_w(w, \Theta) = d_b(w, 1 - \Theta). \quad (5)$$

Samples of constant d_b and d_w are shown in Figure 6.

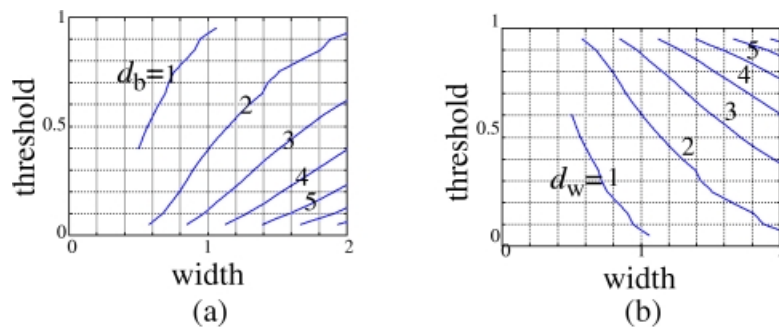


Fig. 6. Observable erosion contours for constant erosion on (a) a black corner, d_b , and (b) on a white corner, d_w . Loci are for a Gaussian PSF and $\phi = \pi/4$.

3 Experiment

Experiments were run to statistically compare characters in pairs of populations each made with different parameters based on the method proposed by Kanungo et al. [10]. In the experiments presented in this paper, the two populations, X and Y, are both composed of synthetically generated characters created by the blurring and thresholding model with varying phase offsets [12]. The characters in population X were created with PSF width and binarization threshold parameters (w_0, Θ_0) , and those in population Y with (w_1, Θ_1) . The null hypothesis that these sets of characters have been drawn from populations with a common set of parameters was compared to the alternate hypothesis that they have been drawn from populations with different parameters:

$$H_N : (w_0, \Theta_0) = (w_1, \Theta_1) \quad (6)$$

$$H_A : (w_0, \Theta_0) \neq (w_1, \Theta_1). \quad (7)$$

Each experiment consisted of the following steps:

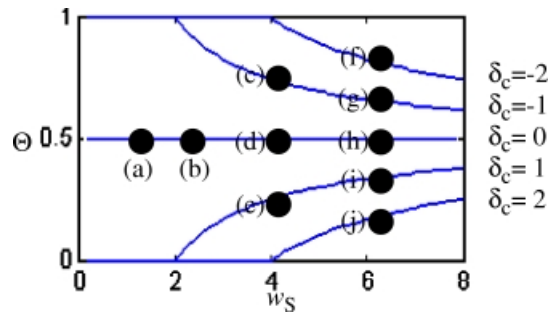


Fig. 7. Set of (w_0, Θ_0) values used as null hypotheses in the sequence of experiments.

1. Create a set of synthesized characters $X = \{x_1, x_2, \dots, x_{2M}\}$ with the model parameters of $\{w_0, \Theta_0, PSF\}$.
2. Using the permutation test method, calculate the null distribution of the population and choose a threshold, d_0 , to make the misdetection rate or significance level, ε , about 5%.
3. Create a set of synthesized degraded characters $Y = \{y_1, y_2, \dots, y_{2M}\}$ of the same character class, using parameters $\{w_1, \Theta_1, PSF\}$.
4. Randomly permute the sets X and Y and select M characters from each.
5. Compute the distance D_k between the sets of $\{x_{k1}, x_{k2}, \dots, x_{kM}\}$ and $\{y_{kM+1}, y_{kM+2}, \dots, y_{k2M}\}$.
6. Repeat steps (4) and (5) K times and get K distances D_1, D_2, \dots, D_K .
7. Compute the probability of $P\{D_k > d_0\} = \#\{k | D_k \geq d_0\} / K$.

The Hamming distance was used to calculate the distance between individual characters, and the distance between sets of characters was calculated using the truncated mean nearest-neighbor distance. K was set to 1000. Steps (3)-(7) were repeated for several parameter sets (w_1, Θ_1) in the vicinity of (w_0, Θ_0) to generate a two-dimensional power function. This will show how likely it is that a change in system parameters will cause the characters to differ.

Experiments were conducted around several initial parameter combinations (w_0, Θ_0) to see how the location in the (w, Θ) space affects the results. The combinations of initial points (w_0, Θ_0) that were used are shown in Figure 7. These points were chosen to give a range of edge displacements $\delta_c = \{-2, -1, 0, 1, 2\}$ for the initial characters and to fill the (w, Θ) space. The initial character image used was a 600-dpi 12-point sans-serif ‘W’. The PSF form was a square pillbox with a base width w_s .

The two-dimensional power functions for several (w_0, Θ_0) are shown as contour images in Figure 8. These contours show the place where the probability of rejecting the null hypothesis is constant over a range of alternate parameters (w_1, Θ_1) . The probability of rejecting the null hypothesis is less than 0.1 in the shaded region. It is 1 in the area outside of the contour lines. The blockiness in

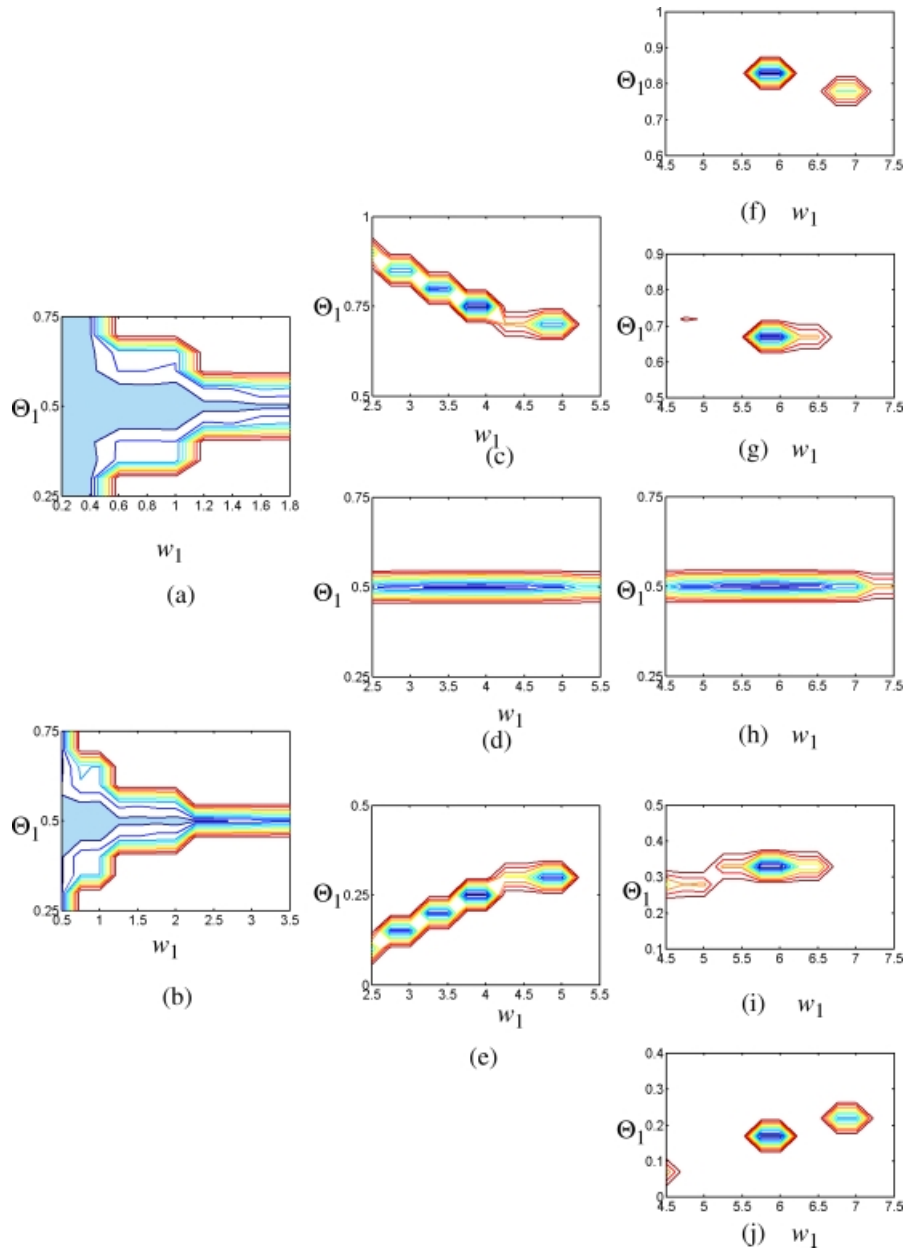


Fig. 8. Probability of rejecting null hypotheses with the letter ‘W’ (a) $(w_0, \Theta_0) = (1.0, 0.5)$, (b) $(2.0, 0.5)$ (c) $(4.0, 0.75)$, (d) $(4.0, 0.5)$, (e) $(4.0, 0.25)$, (f) $(6.0, 0.83)$, (g) $(6.0, 0.67)$, (h) $(6.0, 0.5)$, (i) $(6.0, 0.33)$, (j) $(6.0, 0.17)$. The shaded region has a probability of less than 0.1.

the contour shapes is caused by the quantization in the range of (w_1, Θ_1) values used in the experiments and the Matlab interpretation of the contour.

The constant reject probabilities have a shape similar to the constant edge spread contours shown in Figure 4. This is more easily seen in Figure 9, where the hypothesis reject probability contours have been superimposed on the δ_c contours. The edge spread degradation has the predominant effect on the appearance of a character visually [5] and, from these results, also statistically.

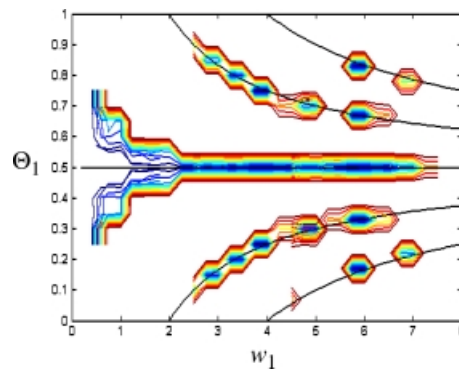


Fig. 9. Composite showing results from Figure 8 superimposed over constant δ_c lines.

To show the sensitivity of this procedure, consider the corresponding sets of characters in the right and left columns of Figure 10 which are created with parameters that are very close. These appear similar, however, the null hypothesis that the populations from which these characters came were generated with the same parameters was rejected with probability equal to 1. In [5] it was proposed that characters with a common δ_c value would appear most similar to humans, while other degradation features, such as d_b and d_w , change the character's appearance less. This similarity is now quantified through statistical testing.

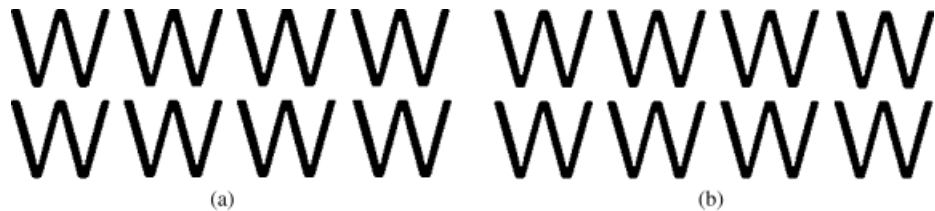


Fig. 10. Synthetic characters created at two (w, Θ) combinations with varying phase offsets. (a) $(w, \Theta) = (0.4, 0.50)$, (b) $(w, \Theta) = (0.4, 0.55)$. The characters in the two sets look the same but are decided to be from different parameter sets with probability of 1.

Maintaining a constant δ_c increases the probability of the characters appearing similar, but they are only similar within a small range of (w, Θ) values. Figure 11 shows sample characters with pairs having a common δ_c . The first column shows $\delta_c < 0$, the middle $\delta_c = 0$, the right $\delta_c > 0$. For characters with a positive δ_c (low threshold), the characters have thicker strokes, whereas with negative δ_c , the characters have thinner strokes. The pairs of characters in Figure 11 look similar, but the differences can be easily seen because the model parameters used to create them are very different. The places where the characters with common δ_c differ is at the corners.

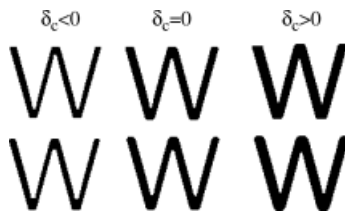


Fig. 11. Characters degraded with (w, Θ) values to produce negative, zero and positive δ_c values. Each character has a different (w, Θ) .

While the δ_c value has remained the same, the corner erosion and thus the character appearance is different. For $\delta_c > 0$, ($\Theta < 1/2$) the d_b isolines are almost perpendicular to the δ_c isolines, and for $\delta_c < 0$ ($\Theta > 1/2$) the d_w isolines are almost perpendicular to the δ_c isolines. When w and $|\Theta - 1/2|$ are large, a small change in (w, Θ) will produce a larger change in the d_b and d_w values (see Figure 6). This causes the size of the region of low probability of rejecting the null hypothesis in Figure 8f,g,i and j to be smaller than the corresponding regions in Figure 8c and e.

A similar set of experiments was run using a 12-point sans-serif ‘O’ over a subset of the cases used for the letter ‘W’. This character has approximately the same stroke width for the whole character but contains no corners. The resulting power function contours are shown in Figure 12. When the plots are compared to the plots for the corresponding null hypothesis for the ‘W’ shown in Figure 8a,c,e,g,h and i, the appearance of the same general shape can be seen. What is different, particularly for $(w_0, \Theta_0) = (6, 0.67)$ and $(6, 0.33)$, is the probability of rejecting the null hypothesis being less than 1 extends for a larger range of values for the letter ‘O’. This is due to the absence of corners. The degradation seen in the characters is only due to the edge spread for a large range of (w, Θ) values. With an absence of corners, no corner erosion is present. However, the edge spread was defined for edges that are isolated from each other, and when the PSF width is large enough, this premise is no longer valid [11]. The edges will interfere, and an effect similar to the corner erosion will occur degrading the character images. The corner erosion is a special case of two edges

spreading with interference, where the overlap occurs at any PSF support width because the distance between the edges at the corners is zero.

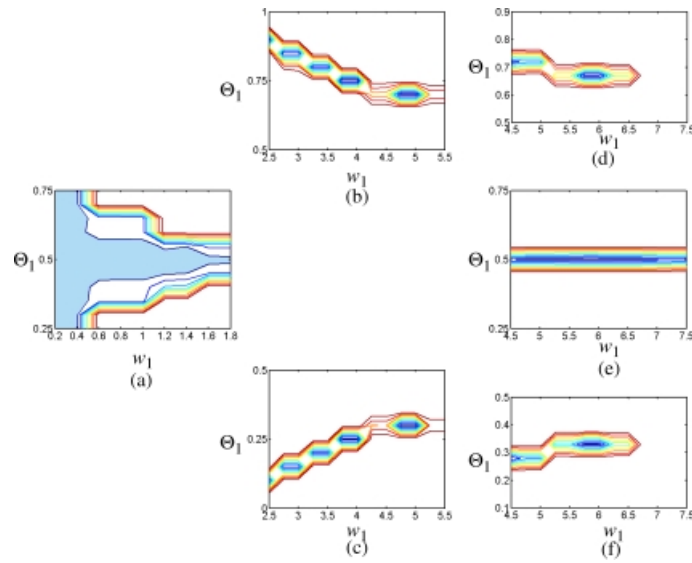


Fig. 12. Probability of rejecting null hypotheses with the letter ‘O’ (a) $(w_0, \theta_0) = (1.0, 0.5)$, (b) $(4.0, 0.75)$, (c) $(4.0, 0.25)$, (d) $(6.0, 0.67)$, (e) $(6.0, 0.5)$, (f) $(6.0, 0.33)$. The shaded region has a probability less than 0.1.

4 Conclusion

A statistical test was conducted to compare the similarity between groups of characters synthetically generated with parameters (w, Θ) varying over the parameter space. The amount of variation in the characters correlated highly with the change in the edge spread degradation. This change can be quantified in terms of the degradation system model parameters. When estimating the degradation model parameters, errors along the δ_c isolines will not produce as large a difference in the characters generated with the model as would an error perpendicular to these isolines.

The effects of an estimation error will be less for characters with fewer corners. Characters with many corners, thin strokes, or variable width strokes will remain similar over a smaller range of (w, Θ) values. This can be used to decide how much of an effect an error in estimating the system parameters will have when using those parameters to generate synthetic characters for choosing an OCR structure, training OCR systems, or predicting OCR performance. This can also be used to decide how to distribute model parameters if we want to

experiment with characters with small differences, or larger differences that are evenly distributed. These experiments have also given more insight on how the shape of a character will influence the variation in the resulting bitmap.

The statistical difference between characters could also be used as a metric of model parameter estimation error. Because w and Θ are not in the same units, conventional metrics like euclidean or city block aren't reasonable for combining errors in these two estimates. Also, just adding a scaling factor won't necessarily help because we don't know how to equate width and threshold units. But if we measure error in units of character difference, that would be meaningful.

References

1. H. S. Baird. Document image defect models. In H.S.Baird, H. Bunke, and K. Yamamoto(eds), editors, *Structured Document Image Analysis*. Springer-Verlag, June 1992.
2. H. S. Baird. Calibration of document image defect models. In *Proc. of 2nd annual symposium on document analysis and information retrieval, Las Vegas, Nevada*, pages 1–16, April 1993.
3. E. H. Barney Smith. Characterization of image degradation caused by scanning. *Pattern Recognition Letters*, 19(13):1191–1197, 1998.
4. E. H. Barney Smith. *Optical Scanner Characterization Methods Using Bilevel Scans*. PhD thesis, Rensselaer Polytechnic Institute, December 1998.
5. E. H. Barney Smith. Bilevel image degradations: Effects and estimation. In *Proc. 2001 Symposium on Document Image Understanding Technology*, pages 49–55, Columbia, MD, 2001.
6. E. H. Barney Smith. Estimating scanning characteristics from corners in bilevel images. In *Proc. SPIE Document Recognition and Retrieval VIII*, volume 4307, pages 176–183, San Jose, CA, 2001.
7. E. H. Barney Smith. Scanner parameter estimation using bilevel scans of star charts. In *Proc. International Conference on Document Analysis and Recognition 2001*, pages 1164–1168, Seattle, WA, 2001.
8. L. R. Blando, J. Kanai, and T. A. Nartker. Prediction of OCR accuracy using simple features. In *Proc. of the Third International Conference on Document Analysis and Recognition*, pages 319–322, Montreal, Quebec, Canada, 1995.
9. T. K. Ho and H. S. Baird. Large-scale simulation studies in image pattern recognition. *IEEE PAMI*, 19(10):1067–1079, 1997.
10. T. Kanungo, R. M. Haralick, H. S. Baird, and D. M. Werner Stuezle. A statistical, nonparametric methodology for document degradation model validation. *IEEE PAMI*, 22(11):1209–1223, 2000.
11. T. Pavlidis, M. Chen, and E. Joseph. Sampling and quantization of bilevel signals. *Pattern Recognition Letters*, 14:559–562, 1993.
12. P. Sarkar, G. Nagy, J. Zhou, and D. Lopresti. Spatial sampling of printed patterns. *IEEE PAMI*, 20(3):344–351, 1998.
13. T. Sziriányi and Á. Böröczki. Overall picture degradation error for scanned images and the efficiency of character recognition. *Optical Engineering*, 30(12):1878–1884, 1991.
14. W. R. Throssell and P. R. Fryer. The measurement of print quality for optical character recognition systems. *Pattern Recognition*, 6:141–147, 1974.