Estimating Degradation Model Parameters Using Neighborhood Pattern Distributions: An Optimization Approach

Tapas Kanungo, *Senior Member*, *IEEE*, and Qigong Zheng

Abstract—Noise models are crucial for designing image restoration algorithms, generating synthetic training data, and predicting algorithm performance. There are two related but distinct estimation scenarios. The first is model calibration, where it is assumed that the input ideal bitmap and the output of the degradation process are *both* known. The second is the general estimation problem, where only the image from the output of the degradation process is given. While researchers have addressed the problem of calibration of models, issues with the general estimation problems have not been addressed in the literature. In this paper, we describe a parameter estimation algorithm for a morphological, binary, page-level image degradation model. The inputs to the estimation algorithm are 1) the degraded image and 2) information regarding the font type (italic, bold, serif, sans serif). We simulate degraded images using our model and search for the optimal parameter by looking for a parameter value for which the local neighborhood pattern distributions in the simulated image and the given degraded image are most similar. The parameter space is searched using a direct search optimization algorithm. We use the p-value of the Kolmogorov-Smirnov test as the measure of similarity between the two neighborhood pattern distributions. We show results of our algorithm on degraded document images.

Index Terms—Degradation models, parameter estimation, direct search algorithms, neighborhood pattern distributions.

1 Introduction

Numerous document image degradation models have been proposed in the literature [1], [10], [11]. However, prior to using these models, it is important to 1) validate the models—that is, verify that the simulations generated by these models are similar to real-world examples, and 2) provide algorithms for estimating the model parameters from real samples. The issue of validation was addressed by Kanungo et al. [8], [9] by converting the validation problem into a statistical hypothesis testing problem and then using a statistical permutation test to test the null hypothesis that a synthetic sample of degraded characters and another sample of real degraded characters come from the same underlying distribution. Lopresti et al. [14] instead proposed to study the differences in the error characteristics of the OCR output for the real and synthetic samples. This method, however, considers the degradation coupled with the OCR system and not just the degradation process.

The issue of model parameter estimation has been studied to a lesser extent in the literature. There are two related but distinct estimation problems: 1) model calibration and 2) general parameter estimation. In a model calibration scenario, you have a particular device (for example, photocopier/scanner system) with you that you have control over. You can provide it with any input and observe the corresponding output. The problem is to estimate the model parameters given the input and the output. If these parameters are known, one can use them to make, for example, the output of the scanner less noisy.

- *T. Kanungo is with the IBM Almanden Research Center, 650 Harry Rd., San Jose, CA 95120. E-mail: kanungo@almaden.ibm.com.*
- *Q. Zheng is with the Department of Electrical Engineering, University of Maryland, College Park, MD 20742. E-mail: qzheng@cfar.umd.edu.*

Manuscript received 27 Aug. 2001; revised 5 Nov. 2002; accepted 8 Sept. 2003.
Recommended for acceptance by L. Vincent.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 114853.

In the general parameter estimation problem, one picks up a document lying on the desk that was perhaps typeset on some unknown device, printed on some unknown printer, photocopied on unknown photocopiers (unknown number of times), and asks if there is a set of parameters of a model that can create simulated noisy documents that have degradation characteristics of the entire process. The ideal bitmap is *not* provided and the process sequence is not known.

Of the two estimation problems, the calibration problem has been studied more since it is more tractable. Kanungo and Haralick [7] reported results of some preliminary experiments that they conducted to calibrate the degradation model parameters using an objective function based on the power function. Baird [2] used the same power function approach proposed in [7] to calibrate the parameters of another physics-based degradation model, and Sural and Das [18] calibrated the parameters of a two-state Markov chain document degradation model using the same power function approach. Finally, Kanungo and Zheng [12] used the overall methodology described in [7], but replaced the brute-force optimization by the direct-search optimization to calibrate a degradation model.

All the above papers assumed that an ideal document image and the corresponding degraded image were given, as is the case in all calibration scenarios. Furthermore, since most of the methods use aligned ideal and degraded bitmaps, they cannot be used for the general model parameter estimation problem.

In this paper, we propose an estimation algorithm for the general model parameter estimation problem that does not require the ideal images and does not require character-level geometric groundtruth either. The algorithm is based on computing similarity between the distributions of neighborhood patterns in the observed degraded images and synthetically degraded images of document with similar text content. The calibration version of this paper appeared in [12], and an application of our estimation algorithm for restoration of document images appeared in [20].

In Section 2, we describe our document degradation model. The notion of neighborhood pattern distribution is introduced in Section 3 and, in Section 4, we study the impact of changing document font and text properties on these distributions.We outline the estimation algorithm in Section 5 and provide simulation results in Section 6.

2 The Morphological Document Degradation Model

In this section, we briefly describe a document degradation model for the local degradation that is introduced when documents are printed, scanned, and digitized [8], [10], [11]. The model accounts for 1) the pixel inversion (from foreground to background and vice versa) that occurs independently at each pixel due to light intensity fluctuations, pixel sensitivity, and thresholding level, and 2) the blurring that occurs due to the point-spread function of the optical system of the scanner. We model the probability of a background pixel flipping as an exponential function of its distance from the nearest boundary pixel. The parameter α_0 is the initial value for the exponential, and the decay speed of the exponential is controlled by the parameter α. The foreground and background 4-neighbor distance are computed using a standard distance transform algorithm [6] . The flipping probabilities of the foreground pixels are similarly controlled by β_0 and β. The parameter η is the constant probability of flipping for all pixels. Finally, the last parameter k, which is the size of the disk used in the morphological closing operation [6], accounts for the correlation introduced by the point-spread function of the optical system.

The degradation model thus has six parameters:

$$\Theta = (\eta, \alpha_0, \alpha, \beta_0, \beta, k).$$

These parameters are used to degrade an ideal binary image as follows:

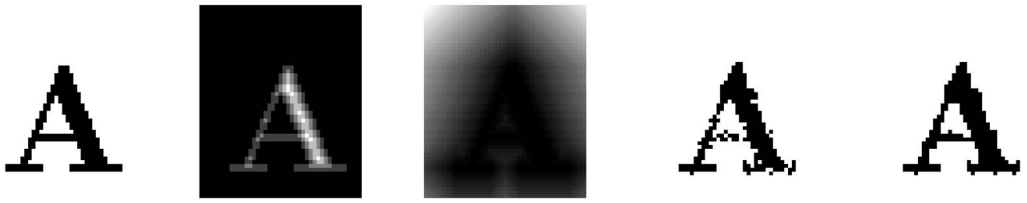

Fig. 1. Local document degradation model: (a) Ideal noise-free character. (b) Distance transform of the foreground. (c) Distance transform of the background. (d) Result of the random pixel-flipping process (the probability of a pixel flipping is $p(0|d,\beta,f) = p(1|d,\alpha,b) = \alpha_0 e^{-\alpha d^2}$; here, $\alpha = \beta = 2$, $\alpha_0 = \beta_0 = 1$). (e) Morphological closing of the result in (d) by a 2×2 binary structuring element. This model is typically applied on an entire typeset page and, thus, accounts for the intercharacter interactions.

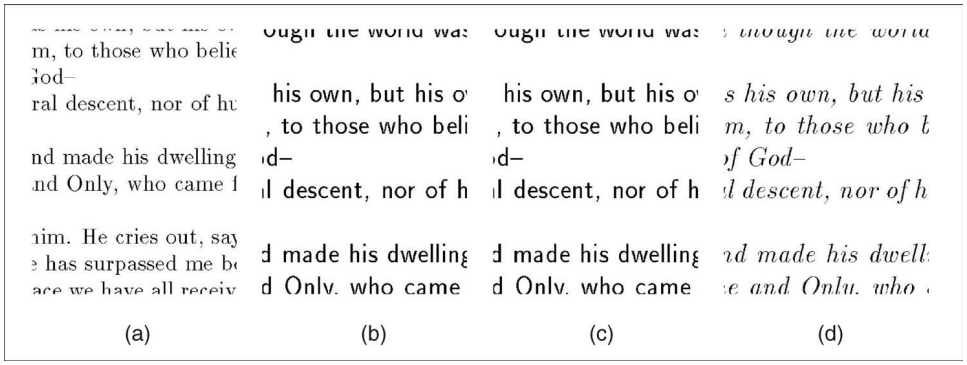

Fig. 2. Text typeset in computer modern Roman font. (a) Serif text. (b) Sans Serif text. (c) Serif bold text. (d) Serif italic text.

1. Flip each foreground pixel with probability

$$p(0|1, d, \alpha_0, \alpha) = \alpha_0 e^{-\alpha d^2} + \eta.$$

2. Compute the distance d of each pixel from the character boundary.
3. Flip each background pixel with probability

$$p(1|0, d, \beta_0, \beta) = \beta_0 e^{-\beta d^2} + \eta.$$

4. Finally, perform morphological closing with a disk element of diameter k.

The application of the various steps of the model is illustrated in Fig. 1. The procedure described above works on bit-mapped images. Since there is no restriction on the size of the image that can be degraded, or the language of the written text, an entire document page image can be degraded using this model. Application of the model on the entire typeset document automatically accounts for the intercharacter interactions in the degradation process.

3 Neighborhood Pattern Distributions

Our estimation algorithm is based on the assumption that, if the degradation parameters are estimated correctly, the local degradations in a simulated image generated using the estimated parameters will look similar to those of a real image. The way we capture this fact is by looking at the distribution of neighborhood patterns.

Let P be a set of neighborhood bit patterns and p be an arbitrary element in the set P. For example, p could be a 3×3 neighborhood with all 1s, or it could be a 5×5 neighborhood with a 1 in the middle and 0s everywhere else. Now, we define the neighborhood pattern distribution of an image R. Let H_R denote a neighborhood pattern distribution, so that $H_R(p)$, where $p \in P$, is the number of times the pattern p occurs in the binary image R. Using mathematical morphology [6], we can define $H_R(p)$ more precisely: $H_R(p) = \#\{R \ominus p\}$. A sample of document subimages is shown in Fig. 5 and their corresponding neighborhood pattern distributions are shown in Fig. 6.

4 Sensitivity to Font Classes and Text Properties

We conducted three experiments to study whether the pattern distributions could discriminate various font and language characteristics. In particular, we studied whether the change in 1) fonts, 2) font size, or 3) text or statistical language properties could be detected using the neighborhood pattern distributions in ideal images. In Fig. 2 and Table 1a, we show subimages of same text typeset in serif, sans serif, bold, and italic fonts. We find that the Kolmogorov-Smirnov test can easily detect the differences in the neighborhood pattern distributions. In Fig. 3 and Table 1b, we show that, even if we change the font size of serif text, the neighborhood pattern distributions are quite indistinguishable. Finally, in Fig. 4 and Table 1c, we show that if we replace the original text with another from the same source, and keep the font characteristics identical, the Kolmogorov-Smirnov test cannot detect the difference. However, if we change the underlying language properties (e.g., bigram probabilities) drastically, the method can discriminate easily. This is quite an interesting result because it says that, in order to compare the noise pattern distributions of two images, the two images need not have the same underlying ideal image—bitmaps can be generated from texts that have similar statistical language and font properties. We will use this fact in the estimation algorithm described in the next section.

5 The Estimation Algorithm

Let R be the given degraded image. Let I be the corresponding (unknown) ideal image. The problem is to estimate the degradation model parameter θ such that, if we were to degrade the ideal image I with the degradation model with parameter fixed at θ, we will get a degraded image S_θ that looks similar to R. For our purposes, we say that two images R and S are similar if the corresponding neighborhood pattern distributions H_R and H_{S_θ} are similar. In the previous section, we saw that the neighborhood patterns of

TABLE 1
Kolmogorov-Smirnov Test Statistics and Significance Level (T, P-value) for Text Fragments in (a) Fig. 2, (b) Fig. 3, and (c) Fig. 4

T / P-value	Serif	Sans Serif	Serif Bold	Serif Italic
Serif	0.0	0.193	0.120	0.078
	1.0	0.00	0.00	0.09
Sans Serif	0.193	0.0	0.096	0.25
	0.00	1.0	0.02	0.00
Serif Bold	0.12	0.096	0.0	0.19
	0.00	0.02	1.0	0.00
Serif Italic	0.078	0.25	0.19	0.0
	0.09	0.00	0.00	1.0

(a)

T / P-value	6pt	8pt	12pt	17pt
6pt	0.0	0.053	0.057	0.086
	1.0	0.472	0.382	0.045
8pt	0.053	0.0	0.062	0.076
	0.472	1.0	0.268	0.101
12pt	0.057	0.062	0.0	0.049
	0.382	0.268	1.0	0.572
17pt	0.086	0.076	0.049	0.0
	0.045	0.101	0.572	1.0

(b)

T / P-value	Fig 4(a)	Fig 4(b)	Fig 4(c)	Fig 4(d)
Fig 4(a)	0.0	0.025	0.392	0.384
	1.0	0.996	0.00	0.00
Fig 4(b)	0.025	0.0	0.412	0.404
	0.996	1.0	0.0	0.00
Fig 4(c)	0.392	0.412	0.0	0.075
	0.00	0.0	1.0	0.118
Fig 4(d)	0.384	0.404	0.075	0.0
	0.00	0.0	0.118	1.0

(c)

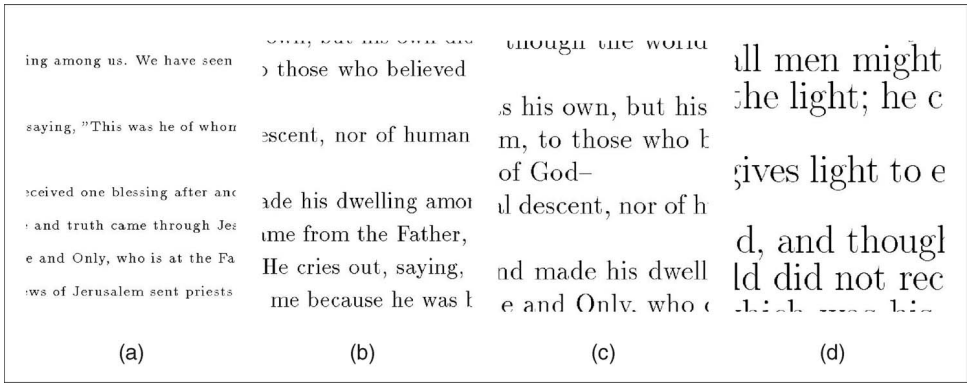

Fig. 3. Text in various font sizes. (a) 6pt font. (b) 8pt font. (c) 12pt font. (d) 17pt font.

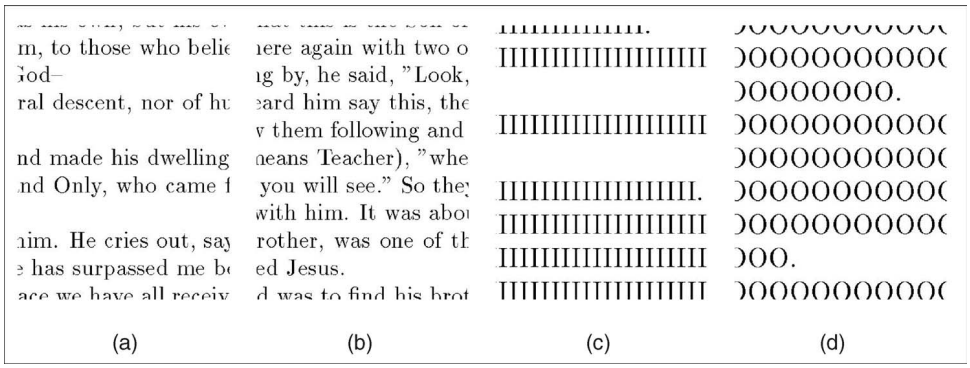

Fig. 4. Texts in 12pt serif Computer Modern Roman font. (a) A fragment of text from one document. (b) Another fragment from a similar document. (c) A fragment of text containing only "I"s. (d) A fragment of text containing only "O"s.

document images with *different* text, but similar statistical language and font class properties have similar neighborhood pattern distributions. Thus, in our estimation problem, we will generate an ideal image $\tilde{I}$ by typesetting some text that has statistical language properties and font that are similar to the original text in R and use $\tilde{I}$ as a surrogate for I. Such text can be searched from various collection of electronic texts in various genres using information retrieval systems and keywords selected from R.

Next, we convert the estimation problem to an optimization problem. The search space is the degradation model parameter space. The objective function is computed as follows: We use the Kolmogorov-Smirnov test [15] (see the appendix for an overview of the KS test) to compute the similarity of the two neighborhood pattern distributions. Let $KS(H_R, H_{S_\theta})$ denote the KS test p-value for the null hypothesis that the two distributions are same. We use this p-value as the objective function that the optimization process maximizes: $\hat{\theta} = \max_\theta KS(H_R, H_{S_\theta})$.

Notice that the degraded image S_θ is computed by the algorithm described in Section 2. Thus, the derivatives of the objective function cannot be computed in closed form. Hence, standard derivative approaches to maximizing KS are not applicable. The optimization problem falls under the category of maximization of *multivariate nonsmooth function* [5]. Direct search methods are typically used to solve such optimization problems [13], [19], [17]. These optimization algorithms have been used in practice for very high-dimensional space, and also in the case where the space has both discrete and continuous dependent variables [4], [3]. We used the Nelder-Mead derivative-free optimization algorithm [16] to maximize KS. There

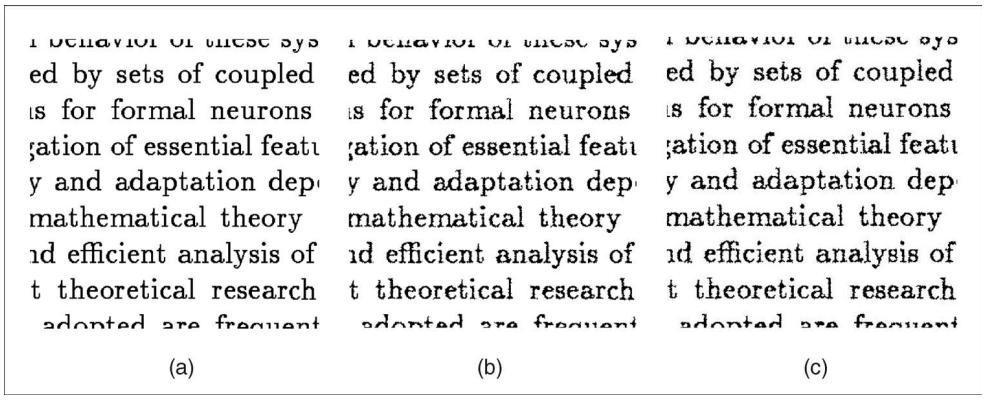

Fig. 5. (a) A typical ideal image. (b) A degraded image with parameters $(0.0, 0.6, 1.5, 0.8, 2.0, 3)$. (c) Image generated using the estimated parameters.

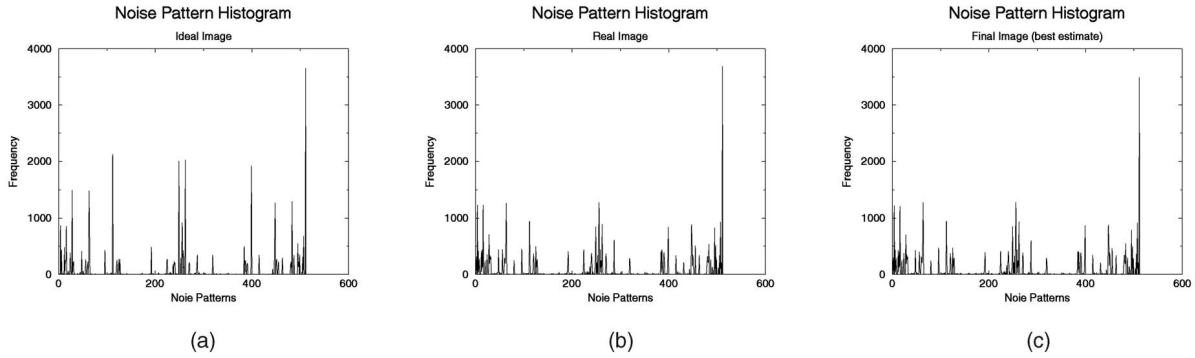

Fig. 6. Neighborhood pattern distributions corresponding to Figs. 5a, 5b, and 5c. Each bin along the x-axis corresponds to a different 3×3 neighborhood pattern.

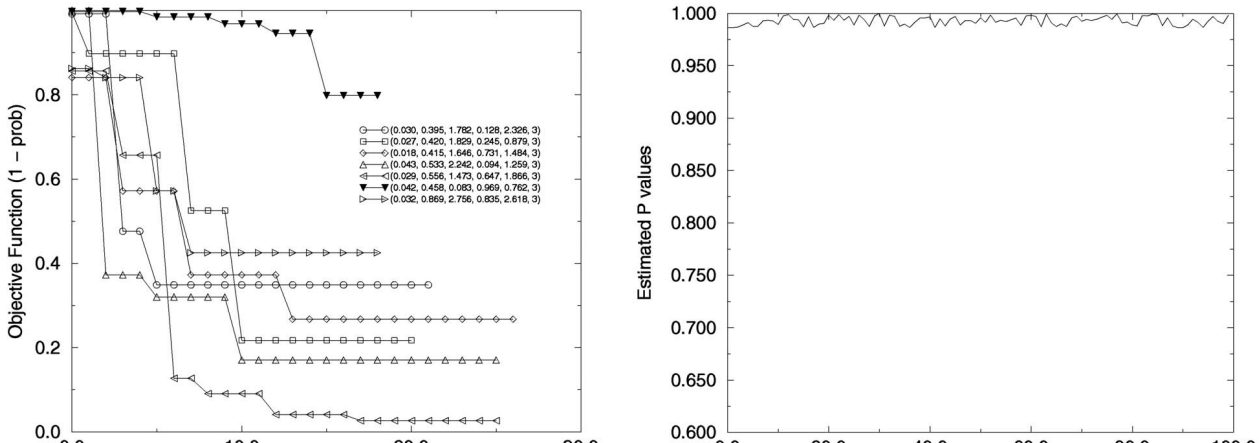

Fig. 7. (a) Downhill simplex convergence for 10 different (random) starting locations. Three of the 10 curves that did not converge and remained at 1.0 have been eliminated for clarity. (b) P-values associated with 100 different model parameter estimations. Since the p-values are high, we can infer that the neighborhood pattern distributions of the degraded image and the simulated image corresponding to the estimated parameters are similar.

is no reason to believe that KS is unimodal over the model parameter space. To circumvent this problem, we perform multiple random starts and then pick the solution corresponding to the highest maximum value.

It is important to note that the three main components of the estimation algorithm: feature set (neighborhood patterns), the objective function (KS test), the optimization algorithm (Nelder-Mead), can be substituted by other algorithms in the literature. How to choose amongst the algorithms is an interesting issue. Furthermore, given the multimodal nature of the objective function, computation of the variance of the estimate is another issue not addressed here.

6 Experimental Protocol and Results

We start with a 400×400 ideal binary image I such as that shown in Fig. 5a. The given degraded image R shown in Fig. 5b was created using the model parameters

$$\theta = (\eta, \alpha_0, \alpha, \beta_0, \beta, k) = (0.0, 0.6, 1.5, 0.8, 2.0, 3).$$

The neighborhood pattern set P was chosen to be all the possible binary patterns in a 3×3 window. Thus, P has 512 patterns. The neighborhood pattern distribution corresponding to Figs. 5a, 5b, and 5c are shown in Figs. 6a, 6b, and 6c. Notice that some patterns occur more frequently than others, and that the distributions of the ideal and degraded images are different. The search was done for $\eta, \alpha_0, \alpha, \beta_0, \beta$, and k. The Nelder-Mead algorithm was started 10 times with random start locations. The objective function value $(1 - p\text{value})$ is plotted as a function of iterations in Fig. 7a. The best optimal solution is found to be

$$\hat{\theta} = (0.029, 0.556, 1.473, 0.647, 1.866, 3).$$

In Fig. 5c, we show the image $R_{\hat{\theta}}$ generated using the optimal solution $\hat{\theta}$. Notice that the neighborhood pattern distribution corresponding to the estimated image, which is shown in Fig. 6c, is quite similar to the histogram of the original degraded image shown in Fig. 6b. Note that the ideal image need not correspond to the degraded image. In fact, one can use any other ideal image that has 1) the same font type as that of the degraded image (the font size can be different, however) and 2) statistical language (e.g., bigram probabilities) properties similar to those of the degraded text. A more careful experimental study that characterizes the behavior of the objective function as a function of the statistical language property difference between the two text samples might be interesting.

Finally, we used the estimation algorithm to estimate the degradation model parameters for 100 *different* images that were generated with parameters chosen randomly over the parameter space. The p-value associated with each run is shown in Fig. 7b. Note that the estimation algorithm resulted in very high p-values, suggesting very similar neighborhood pattern distributions.

7 Conclusion

We have described an algorithm for estimating the parameters of a degradation model. The algorithm assumes that we know the font type (serif, sans serif, bold, italic) of the degraded image and then typeset an arbitrary ideal text image in the same font and similar statistical language properties. The ideal image is then degraded with various parameters of the degradation model. For each parameter value, the neighborhood pattern distributions of the ideal and the degraded images are compared using the Kolmogorov-Smirnov test. The parameter value that maximizes the p-value is used as an estimate of the model parameters. The search for the optimal parameters is done using the Nelder-Mead algorithm.

Appendix A

Kolmogorov-Smirnov Test for Similarity

The Kolmogorov-Smirnov test [15] is widely used as the measure of similarity of two distributions. It is a statistical procedure that uses the maximum distance between two distribution functions as a measure of how well the functions resemble each other. Let $x_i, i = 1, ..., N$, be N values and let $S_N(x)$ represent the fraction of data points to the left of x. Kolmogorov-Smirnov statistic T is defined as the maximum value of the absolute difference between two cumulative distribution functions. For comparing two different cumulative distribution functions $S_{N_1}(x)$ and $S_{N_2}(x)$, the KS statistic is given by $T = \max_{-\infty<x<+\infty} |S_{N_1}(x) - S_{N_2}(x)|$.

Under the null hypothesis, the distribution of the KS statistic can be theoretically derived and, hence, one can compute the significance level (as a disproof of a true null hypothesis) of any observed value of T. While the accuracy of the significance level increases as the sample size N increases, in practice, $N = 20$ is large enough. See [15] for more details.

Acknowledgments

The authors would like to thank Azriel Rosenfeld, Margaret Wright, Virginia Torczon, Charles Audet, and the anonymous reviewers for their comments. This research was funded in part by Science Applications International Corporation under Contract 4400019848, the Defense Advanced Research Projects Agency under Contract N660010028910, and the US National Science Foundation under Grant IIS9987944.

References

[1] H. Baird, "Document Image Defect Models," *Proc. IAPR Workshop Syntactic and Structural Pattern Recognition*, pp. 38-46, June 1990.
[2] H.S. Baird, "Document Image Quality: Making Fine Discriminations," *Proc. Int'l Conf. Document Analysis and Recognition*, pp. 459-462, Sept. 1999.
[3] A.J. Booker, J.E. Dennis, P.D. Frank, D.B. Serafini, V. Torczon, and M.W. Trosset, Optimization Using Surrogate Objectives on a Helicopter Test Example. pp. 49-58, 1998.
[4] A.J. Booker, J.E. Dennis, P.D. Frank, D.B. Serafini, V. Torczon, and M.W. Trosset, A Rigorous Framework for Optimization of Expensive Functions by Surrogates, *Structural Optimization*, vol. 17, pp. 1-13, 1999.
[5] P.E. Gill, W. Murray, and M.H. Wright, *Practical Optimization*. London and New York: Academic Press, 1993.
[6] R.M. Haralick and L.G. Shapiro, *Robot and Computer Vision*, vols. 1 and 2, Reading, Mass.: Addison-Wesley, 1992.
[7] T. Kanungo and R.M. Haralick, "Morphological Degradation Parameter Estimation," *Proc. SPIE Conf. Nonlinear Image Processing*, vol. 2424, pp. 86-95, Feb. 1995.
[8] T. Kanungo, R.M. Haralick, H. Baird, W. Stuezle, and D. Madigan, "A Statistical, Nonparametric Methodology for Document Degradation Model Validation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1209-1223, 2000.
[9] T. Kanungo, R.M. Haralick, H.S. Baird, W. Stuetzle, and D. Madigan, "Document Degradation Models: Parameter Estimation and Model Validation," *Proc. Int'l Workshop Machine Vision Applications*, Dec. 1994.
[10] T. Kanungo, R.M. Haralick, and I. Phillips, "Global and Local Document Degradation Models," *Proc. Int'l Conf. Document Analysis and Recognition*, pp. 730-734, Oct. 1993.
[11] T. Kanungo, R.M. Haralick, and I. Phillips, "Non-Linear Local and Global Document Degradation Models," *Int'l J. Imaging Systems and Technology*, vol. 5, pp. 220-230, 1994.
[12] T. Kanungo and Q. Zheng, "Estimation of Morphological Degradation Model Parameters," *Proc. IEEE Int'l Conf. Speech and Signal Processing*, May 2001.
[13] R.M. Lewis, V. Torczon, and M.W. Trosset, "Why Pattern Search Works," *OPTIMA*, vol. 59, pp. 1-7, 1998.
[14] Y. Li, D. Lopresti, G. Nagy, and A. Tomkins, "Validation of Document Defect Models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, pp. 99-107, 1996.
[15] F.J. Massey, "The Kolmogorov-Smirnov Test for Goodness," *J. Am. Statistical Assoc.*, vol. 46, pp. 68-78, 1951.
[16] J. Nelder and R. Mead, "A Simplex Method for Function Minimization," *Computer J.*, vol. 7, pp. 308-313, 1965.
[17] M.J.D. Powell, Direct Search Algorithms for Optimization Calculations. vol. 7, pp. 287-336, 1998.
[18] S. Sural and P.K. Das, "A Two-State Markov Chain Model of Degraded Document Images," *Proc. Int'l Conf. Document Analysis and Recognition*, pp. 463-466, Sept. 1999
[19] M.H. Wright, "Direct Search Methods: Once Scorned, Now Respectable," *Numerical Analysis*, pp. 191-208. D.F. Griffiths and G.A. Watson, eds., Addison Wesley, Longman (Harlow), 1996.
[20] Q. Zheng and T. Kanungo, "Morphological Degradation Models and Their Use in Document Image Restoration," *Proc. IEEE Int'l Conf. Image Processing*, Oct. 2001.