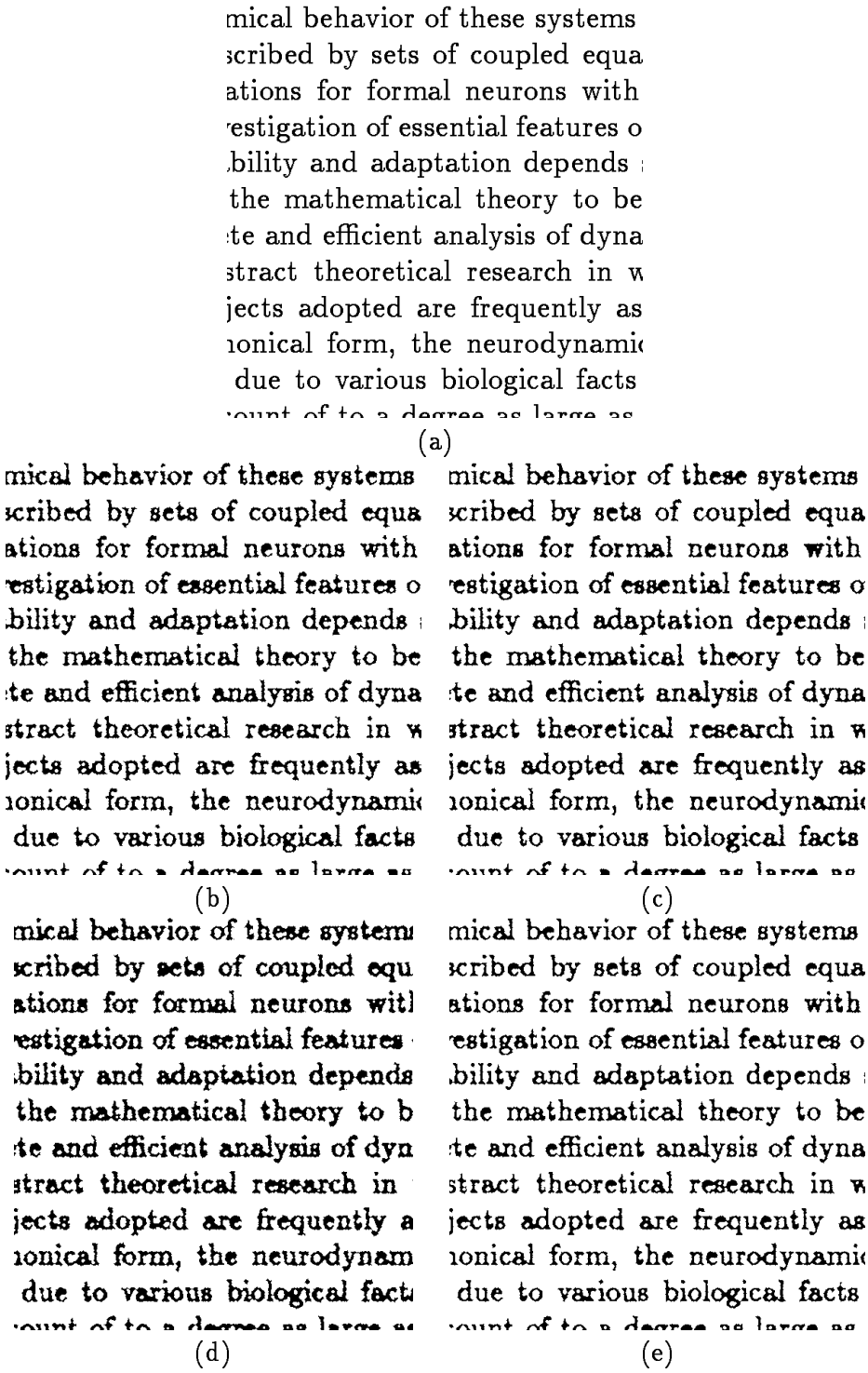

Figure 2: (a)Subimage of the noise free document. (b) Reference degraded document simulated with $\alpha = \beta = 2.0$. (c) Probe sample accepted, $\alpha = \beta = 1.7$. (d) Probe sample rejected, $\alpha = \beta = 0.9$. (e) Probe sample rejected, $\alpha = \beta = 2.0$. Sample size used was 60.

reject the null hypothesis easily, are shown in figure 2 (d) and figure 2 (e), respectively.

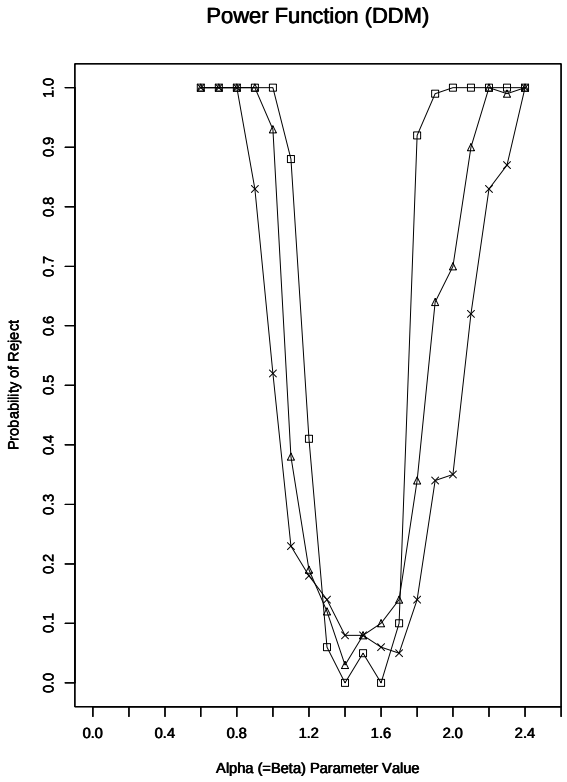

Figure 1: Power plots. The reference distribution had $\alpha = \beta = 1.5$. Notice that the power function has a minimum near $\alpha = \beta = 1.5$. The power function corresponding to sample size of 60, is sharper (marked with square boxes), and that corresponding to a sample size of 10 is broader (marked with crosses).

6 Summary

We proposed a statistical methodology for validating a document degradation model proposed elsewhere [KHP93, KHP94]. We constructed the power function of the validation procedure and used it to estimate the model parameters. This validation methodology is very general and can be used for validating other document degradation models, e.g., [Bai90]. Different validation procedures can be compared by comparing their power functions. Currently we are working on applying this methodology to real data.

References

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

[Bai90] H. Baird. Document image defect models. In *Proc. of IAPR Workshop on Syntactic and Structural Pattern Recognition*, pages 38–46, Murray Hill, NJ, June 1990.

[HS92] R.M. Haralick and L.G. Shapiro. *Robot and Computer Vision (vols. 1 and 2).* Addison-Wesley Publishing Co., Inc., Reading, Massachusetts, 1992.

[KHP93] T. Kanungo, R. M. Haralick, and I. Phillips. Global and local document degradation models. In *Proc. of Second International Conference on Document Analysis and Recognition*, pages 730–734, Tsukuba, Japan, October 1993.

[KHP94] T. Kanungo, R. M. Haralick, and I. Phillips. Non-linear local and global document degradation models. *Int. Journal of Imaging Systems and Technology (to appear)*, 1994.

tances, $\delta(x,y)$, such as Hausdorf distance could have been used. Similarly, other set distance functions, $\rho(X,Y)$, could have been used, e.g., squared difference of the set means, the Hausdorf distance, etc. The combination of character distance $\delta(x,y)$, and set distance, $\rho(X,Y)$, that give rise to a power function that is higher (more powerful) than all other power functions, is the best pair of distances to use for the validation procedure.

4 Estimation of the Degradation Model Parameters

Given a degraded document we would like to estimate the parameters of the document degradation model, $\hat{\Theta} = (\hat{\alpha}_0, \hat{\alpha}, \hat{\eta}_f, \hat{\beta}_0, \hat{\beta}, \hat{\eta}_b, \hat{k})^t$, that could be used to create similarly degraded documents. The notion of "similar" degradations was addressed in the pervious section on model validation.

We will use the following procedure to estimate the parameter vector $\hat{\Theta}$.

1. Construct the power function $\Phi(\Theta)$ for small sample sizes N and M. This gives a shallow power function and allows one to sample the parameter space coarsely.
2. Find the parameter value where the power function is minimum.
3. Construct the power function in the neighborhood of the minimum, but with larger sample sizes, and finer sampling of the parameter space.
4. Find the parameter value where the power function is minimum.
5. Repeat steps 3 and 4 until local minimum is found.

5 Experimental Results

In this section we give experimental results on simulated data: (i) we use the validation procedure to distinguish two degraded documents that were simulated with different parameter values, and (ii) given a sample of degraded documents, we estimate the model parameter value. Both the problems are addressed by constructing the power function $\Phi(\Theta)$. The simulation was done using the model described in section 1.

The following protocol used for creating the reference and probe samples X and Y. The reference distribution parameter Θ_f was fixed with the following parameter component values: $\eta_f = \eta_b = 0$, $\alpha_0 = \beta_0 = 1$, $\alpha = \beta = 1.5$, and the structuring element size $k = 5$. The probe distribution parameter Θ_p was varied by changing $\alpha = \beta$. Other probe distribution parameter components, $\eta_f, \eta_b, \alpha_0, \beta_0, k$ were same as that in the reference model parameter. In all cases the noise free document was the same (a Latex document page formatted in IEEE Transaction style) and the same set of 340 character 'e' (Computer Modern Roman 10 point font) were extracted from the page, for creating the reference population X and the probe population Y.

The validation procedure protocol was as follows: the significance level ϵ was fixed at 0.05; the sample sizes $N = M$ used were 10, 20, and 40; the number of permutation K for creating the empirical null distribution was 1000; the number of trials T for estimating the misdetection rate was 100.

The noise free document is shown in figure 2 (a). The reference degraded document generated with model parameter Θ_r is shown in figure 2 (b). The power function for the sample sizes 10, 20, 60 are shown in figure 1. The power function corresponding to sample size 10 is the widest, and the power function corresponding to sample size 60 is the narrowest. Note all the three power functions give a misdetection (reject) rate close to $\epsilon = 0.05$ when the probe distribution has a parameter value close to that of the reference distribution ($\alpha = \beta = 1.5$). Furthermore, when the $\alpha = \beta$ are far from 1.5, the misdetection rate is close to 1, which implies that the validation procedure can distinguish the two samples with high probability. An image generated with $\alpha = \beta = 1.7$ that the validation procedure accepted with a probability close to 0.9, is shown in figure 2 (c). Two images of documents generated with parameter values $\alpha = \beta = 2.0$ and $\alpha = \beta = 0.9$, which the validation procedure could

$\{y_1, y_2, \ldots, y_M\}$. The question that needs to be addressed is whether the distribution of x_i's is same as that of y_i's or not. Note that at this point it does not matter where the x_i's and the y_i's came from. Thus, x_i's and y_i's could both be artificially generated, or both be real instances, or any one of them could be artificial and the other could be real. Furthermore, N and M are typically on the order of 10,000, which makes the sample size very small compared to 10^{300}, the size of the space B. A statistical hypothesis test can be performed to test the null hypothesis that the distributions x_i's and y_i's are the same. We now describe a procedure that will perform this test.

1. Given (i) real data $X = \{x_1, x_2, \ldots, x_N\}$, (ii) simulated data $Y = \{y_1, y_2, \ldots, y_M\}$, (iii) a distance metric on sets, $\rho(X, Y)$, where $X, Y \subseteq B$. (iv) size of test ϵ, (usually 0.05).

2. Create a new sample $Z = \{x_1, \ldots, x_N, y_1, \ldots, y_M\}$. Thus Z has $N + M$ elements labeled z_i, $i = 1, \ldots, N + M$.

3. Randomly partition the set Z into two sets as follows. Randomly select N elements $z_{i_1}, \ldots, z_{i_N}$ as the first set X', and the rest as the second set, Y'.

4. Compute $d_i = \rho(X', Y')$.

5. Now repeatedly permute the elements of Z, create new partitions X' and Y' and compute d_i. Let us say we make K repetitions.

6. Empirically compute a distribution of d_i's as follows $P(d \geq v) = \#\{k | d_k \geq v\}/K$

7. Compute $d_0 = \rho(X, Y)$.

8. Compute the P-value: $p_0 = P(d \geq d_0)$.

9. Reject the hypothesis that the two samples come from the same population if $p_0 < \epsilon$.

Power Functions: If the above procedure is repeated T times, each time with true null hypothesis, the procedure will reject the true null hypothesis, on the average, $\epsilon \cdot T$ number of times. That is, the misdetection rate will be ϵ. In fact, one can generate the *power function,* which is the plot of the misdetection rate as a function of the parameter Θ, (see [Arn90] for details on power functions), of the testing procedure as follows. First, generate the reference sample with the model parameter Θ_r. Now, generate the probe sample with model parameter Θ_p and compute the reject rate $\Phi(\Theta_p)$, by repeating the validation procedure T times, and computing the percentage of times the hypothesis was rejected. Next, keep varying Θ_p and for each value of Θ_p record the reject rate. The plot of the reject rate $\Phi(\Theta_p)$ versus the parameter value Θ_p is the power function. This function should have a minimum at $\Theta_p = \Theta_r$, and should increase on either side and go upto 1 when Θ_p is very far from Θ_r. The sensitivity, i.e, the width of the notch, is a function of the sample sizes N and M and the various metrics used. When the sample size is small, the notch is broader and when the sample size is large, the notch is sharper. If we compute the power functions Φ_1 and Φ_2 for two different validation procedure, for same sample sizes, and $\Phi_1(\Theta) > \Phi_2(\Theta)$, then Φ_1 is a better validation procedure. Note, by design $\Phi_1(\Theta = \Theta_r) = \Phi_2(\Theta = \Theta_r) = \epsilon$. This fact can be used to compare two validation procedures: the validation procedure with a higher power function is better. See [Arn90] for details and justifications.

Distance functions: Various distance functions $\rho(X, Y)$ can be used for computing the distance between the sets of characters X and Y. We used the following distance function for ρ.

$$\rho(X, Y) = (\rho(Y) + \rho(X))/(N + M) \quad (5)$$

where,

$$\rho(Y) = \sum_{x \in X} \left(\min_{y \in Y} \delta(x, y) \right) \quad (6)$$

$$\rho(X) = \sum_{y \in Y} \left(\min_{x \in X} \delta(x, y) \right) \quad (7)$$

$$\delta(x, y) = \text{HammingDistance}(x, y). \quad (8)$$

Hamming distance mentioned above is computed by counting the number of pixels where the characters x and y differ after their centroids have been registered. A variety of other character dis-

$$P(0|d,\Theta,f) = P(0|\beta_0,\beta,\eta_f) \quad (3)$$
$$= \beta_0 e^{-\beta d^2} + \eta_f \quad (4)$$

Here α_0 and β_0 are the initial values for the exponentials; α and β control the decay speed of the exponentials; η_f and η_b are the uniform probability of a foreground and background pixels flipping, respectively. The independent pixel degradation is followed by a morphological closing operation with a disk of diameter k to account for the correlation introduced by the optical point spread function preceding the thresholding operation which produces the noisy image. Since the closing operation is a nonlinear, it is difficult to difficult to model the probability of pixels flipping after the closing operation.

The degradation model parameter vector Θ is a vector of seven parameters, $\Theta = (\alpha_0, \alpha, \eta_b, \beta_0, \beta, \eta_f, k)^t$, where the last entry k is the size of the disk used in the morphological closing operation.

Software for simulating degraded documents using the above degradation model has been written and is now available from University of Washington. Few examples of simulated degraded documents are shown in figure 2.

2 Statistical Problem Definition

In this section we formulate the degradation model parameter estimation problem and the model validation problem as statistical problems. Although degradation of the document is over the entire page, the degradation process itself is local. That is, degradation in a one area does not influence the degradation process in another area, if it is sufficiently far. In particular, the degradation at a pixel is influenced only by pixels within a disk of diameter k, which is the size of the disk structuring element used in the morphological closing process. Thus, one way to characterize the degradation process is to study the degradation of local patterns. Since the most common patterns that occur on a documents page are characters, we will statistical characterize the degradation of individual characters on the page and use this characterization to estimate the parameters of the degradation model that could produce similar degradations.

Assume that a scanned character is represented by 30×30 matrix with 0 or 1 entries. This matrix can be represented as 1000×1 vector ($30 \times 30 \approx 1000$). Let, B be the space of $D = 1000$ dimensional binary vectors, that is, $B = \{0,1\}^D$. Now, let $x_1, x_2, \ldots, x_N \in B$ be independent and identically distributed D-dimensional vectors representing instances of degraded characters produced from the same class ω. That is, each of these x_i's were produced from the same ideal pattern ω (say the ideal character 'e') and the same degradation parameters Θ. Now, in our case D is large, typically on the order of 1000. Thus, the number of possible values x_i can take up is 2^{1000}, which is approximately equal to 10^{300}, a very large number. Furthermore, the sample size, N, is typically on the order of 1000, which is much less than 10^{300}. Thus, the samples x_i fill the space B very sparsely.

Two problems we need to address are:

Model Validation: Suppose we are given a set of *real* degraded instances $x_1, \ldots, x_N \in B$ of the pattern ω and the another set of *simulated* degraded instances $y_1, \ldots, y_M \in B$ of the pattern ω. Test the null hypothesis that the distribution of $y_1, \ldots, y_M$ is same as that of $x_1, \ldots, x_N$, to a specified significance level ϵ.

Parameter Estimation: Suppose we are given a set of degraded instances $x_1, \ldots, x_N \in B$ of the pattern ω. Estimate the degradation model parameter $\hat{\Theta}$, which can be used to simulate degraded instances $y_1, \ldots, y_M \in B$ from the ideal pattern ω, such that the distribution of $y_1, \ldots, y_M$ is close to that of $x_1, \ldots, x_N$.

3 Model Validation

In this section we describe a method can be used to statistically validate the degradation model. Suppose we are given a sequence of real degraded characters $X = \{x_1, x_2, \ldots, x_N\}$, and another sequence of artificially degraded characters $Y =$

Document Degradation Models: Parameter Estimation and Model Validation

Tapas Kanungo†, Robert M. Haralick†, Henry S. Baird*,
Werner Stuetzle‡ and David Madigan‡

†Department of Electrical Engineering, FT-10
‡Department of Statistics, GN-22
University of Washington
Seattle, WA 98195, USA
{tapas,haralick}@ee.washington.edu
{wxs,madigan}@stat.washington.edu

*AT&T Bell Laboratories
600 Mountain Avenue, Room 2C-322
Murray Hill, NJ 07974, USA
hsb@research.att.com

Abstract

Scanned documents are noisy. Recently, [KHP93, KHP94, Bai90], document degradation models were proposed that model the local distortion introduced during the printing and scanning process. In this paper propose a statistical methodology that can be used to validate the degradation models. That is, we show how to test whether two samples of degraded documents are from the population or not. Although we demonstrate the methodology on simulated documents, degraded document populations could in general be: (i) both artificially generated, (ii) one artificial and one real, or (iii) both real. This hypothesis testing methodology is independent of the degradation model and can be used to validate other document degradation models (e.g. [Bai90]). Furthermore, we construct the power function of the validation procedure and use it to the estimate the parameters for the document degradation model from a degraded document page image.

1 A Document Degradation Model

In this section we summarize a document degradation model that was proposed earlier [KHP93, KHP94]. The model accounts for (i) the pixel inversion (from foreground to background and viceversa) that occurs independently at each pixel due to light intensity fluctuations and thresholding level, and (ii) the blurring that occurs due to the point-spread function of the optical system the scanner.

We model the probability of a pixel changing from its ideal value as a function of the distance of that pixel from the boundary of a character. Let d be the distance (four connected or eight connected) of a foreground or background pixel from the boundary of the character and let Θ be the parameters of the model. Let $P(1|d,\Theta,f)$ and $P(0|d,\Theta,f)$ be the probability of a foreground pixel at a distance d from the background to remain as 1 and to change to a 0, respectively. Similarly, let $P(1|d,\Theta,b)$ and $P(0|d,\Theta,b)$ be the probability of a background pixel at a distance d changing to a 1 and remaining a 0, respectively. The foreground and background 4-neighbor distance can be computed using any distance transform algorithm (see [HS92]). The random perturbation process then proceeds to change pixel values in a pixel by pixel independent manner. The following forms for the background and foreground conditional probabilities were used in the model.

$$P(1|d,\Theta,b) = P(1|\alpha_0,\alpha,\eta_b) \quad (1)$$

$$= \alpha_0 e^{-\alpha d^2} + \eta_b \quad (2)$$