Estimation of Morphological Degradation Parameters

Tapas Kanungo and Robert M. Haralick

Intelligent Systems Laboratory
Department of Electrical Engineering, FT-10
University of Washington
Seattle, WA 98195, USA
{tapas,haralick}@ee.washington.edu

Abstract

We have discussed a morphologically based nonlinear document degradation model to characterize the perturbation process associated with the printing and scanning process [KHP93, KHP94]. In this paper we use the nonparametric estimation algorithm discussed in [KHP93, KHP94] for estimating the the sizes of the structuring elements of the degradation model. Other parameters of the model can be estimated in a similar fashion.

Thus, given a small sample of (real) scanned documents, we can estimate the parameters of the model using the nonparametric estimation algorithm, and use the estimated parameters to create a large sample of simulated documents with degradation characteristics similar to that of the real scanned documents. The large simulated sample can then be used for various purposes, for example, training classifiers, estimating performance of OCR algorithms, choosing parameter values in noise removal algorithms, etc.

Keywords: document degradation models, noise models, model validation, parameter estimation.

1 Introduction

We have earlier proposed a nonlinear document degradation model to characterize the perturbation process associated with the printing and scanning process [KHP93, KHP94]. There are other degradation models such as [Bai90, Bai93, LL90, MS88], but we will not be discussing them in this paper. Models are useful only if there are methods of estimating the parameters of the model given a sample of degraded documents. Furthermore, the model should be validated. That is, the degradations produced by the simulation model should be similar to the degradations appearing in reality and this similarity should be tested statistically.

Since most document degradation models are nonlinear in nature – our model has a nonlinear morphological closing step – it is not possible to have a closed-form maximum likelihood estimate of the model parameter. However, a nonparametric estimate of the parameters is possible, and in this paper we use a nonparametric estimation methodology, which we proposed earlier [KHB+94]. In particular, we use the nonparametric estimation technique to estimate the the sizes of the structuring elements used in the degradation modeling of the algorithm.

In section 2 we describe our document degradation algorithm. We state the statistical problem of model validation and estimation in section 3. In section 4 and 5 we review our nonparametric validation and estimation techniques. Finally we give results of our estimation procedure in section 6.
2 The Morphological Document Degradation Model

In this section we briefly describe a document degradation model for the local degradation that occurs when documents are printed, scanned and digitized. For more details, please see [KHP93, KHP94].

The model accounts for (i) 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 (ii) the blurring that occurs due to the point-spread function of the optical system of the scanner.

The probability of a pixel flipping from being a foreground (binary one) pixel to a background (binary zero) pixel depends on the distance of that pixel from the boundary of a character. Let \( d \) be the distance of a foreground or background pixel from the boundary of the character and let \( \theta \) 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.

\[
P(1|d, \theta, b) = P(1|\alpha_0, \alpha, \eta_b) \\
= \alpha_0 e^{-\alpha d^2} + \eta_b \\
P(0|d, \theta, f) = P(0|\beta_0, \beta, \eta_f) \\
= \beta_0 e^{-\beta d^2} + \eta_f
\]

Here \( \alpha_0 \) and \( \beta_0 \) are the scalar multipliers for the exponentials that decide the probabilities at the boundary; \( \alpha \) and \( \beta \) control the decay speed of the exponentials; \( \eta_f \) and \( \eta_b \) are the base probabilities of 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 model the probability of pixels flipping after the closing operation.

The degradation model parameter vector \( \theta \) 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.

An example of simulated degraded document is shown in figure 3.

3 Statistical Problem Definition

Once a document has been scanned in, it is a matrix of ones and zeros. If we scan a document of size 8.5 x 11 inch, at a resolution of 300 dpi, there results a 2500 x 3300 binary matrix. A scanned character, typically, is represented by 30 x 30 binary matrix. This matrix can be represented as 1000 x 1 vector (30 x 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 \( \omega \). That is, each of these \( x_i \)'s are produced from the same ideal pattern \( \omega \) (say the ideal character 'e') and the same degradation parameters \( \theta \).

Two problems that need to be addressed are:
Model Validation: Suppose we are given a set of real degraded instances \( x_1, \ldots, x_N \in B \) of the pattern \( \omega \) and the another set of simulated degraded instances \( y_1, \ldots, y_M \in B \) of the pattern \( \omega \). The model is validated at a significance level \( \epsilon \) if the distribution of \( y_1, \ldots, y_M \) is statistically insignificantly different from \( x_1, \ldots, x_N \), at a significance level \( \epsilon \).

Parameter Estimation: Suppose we are given a set of degraded instances \( x_1, \ldots, x_N \in B \) of the pattern \( \omega \). 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 \( \omega \), such that the distribution of \( y_1, \ldots, y_M \) is close to that of \( x_1, \ldots, x_N \).

First we will briefly describe how to validate two samples \( X \) and \( Y \). Then we will describe in some detail a method to estimate one of the parameter components of our model – the disk structuring element size, \( k \), used in the morphological closing operation. For more discussion on the validation problem, please see our papers [KHB+94, KBH95].

4 Model Validation

In this section we briefly describe the validation method we proposed earlier, [KHB+94, KBH95]. This validation procedure will be used in the estimation algorithm. 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 \( Y = \{y_1, y_2, \ldots, y_M\} \). The question that needs to be addressed is whether the distribution of \( x_i \)'s come is the same as the distribution from which the that of \( y_i \)'s come. Note that at this point it does not matter where the \( x_i \)'s and the \( y_i \)'s come 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.

A statistical hypothesis test can be performed to test the null hypothesis that the underlying distributions for the \( x_i \)'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 \( B \), \( \rho(X, Y) \), where \( X, Y \subseteq B \). (iv) size of test \( \epsilon \), (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 \( p_0 < \epsilon \).
Figure 1: A visual representation of the validation procedure. The computation of the null distribution $F$ of the distance function $\rho(X, Y)$ is shown in figure 2. Here the oval shapes represent processes and the rectangular boxes represent data structures or objects.
Figure 2: A visual representation of the permutation method for computing the null distribution of the set distance function $\rho(X, Y)$. 
In figure 1 we give a visual representation of the procedure.

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 $\rho$.

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

(5)

where,

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

(6)

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

(7)

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

(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 distances, $\delta(x, y)$, such as Hausdorff 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 Hausdorff distance, etc.

Each choice of $\rho$ and $\delta$ results in a particular test. Furthermore, there are other hypothesis testing methods, such as Monte Carlo techniques, that could have been used for computing the null distribution of $\rho$. Thus each combination of $\rho$, $\delta$, and null distribution computation method will give rise to a particular validation procedure. Although each of these procedures will have the same significance level (probability of misdetection) $\epsilon$, their false alarm rate when null hypothesis is not true, that is, the power, will differ. Thus, the best validation procedure is the one that is most powerful. (See [Arn90, KHB+94] definition of power of tests.)

5 Estimation of the Degradation Model Parameters

Given a degraded document we would like to estimate the parameters of the document degradation model, $\hat{\theta} = (\hat{\alpha}_0, \hat{\alpha}, \hat{\eta}_f, \hat{\beta}_0, \hat{\beta}, \hat{\eta_b}, \hat{k})'$, that could be used to create similar looking degraded documents. The notion of "similar" degradation was addressed in the previous section where we defined a set distance $\rho(X, Y)$.

We will use the following procedure to estimate the structuring element (s.e.) size $k$.

1. Given (i) a fixed reference sample of degraded characters $X = \{x_1, x_2, \ldots, x_N\}$, (iii) assumption that the degradation model is correct, (ii) unknown structuring element (s.e.) size $k = k_r$, (one of the components of the model parameter $\theta$, (iv) a distance metric on sets, $\rho(X, Y)$, where $X$ and $Y$ are sets of characters. (v) size of test $\epsilon$, (usually 0.05).

2. Generate a probe sample $Y = \{y_1, y_2, \ldots, y_M\}$ with s.e. size fixed at $k = k_p$.

3. Run the validation procedure at $\epsilon$ significance level and record whether the null hypothesis that $k_r = k_p$ is accepted or rejected.

4. Compute the expected value, $\gamma(k_p)$, of the reject rate by repeating steps 2 and 3, $N$ Repeat number of times.

5. Vary $k_p$ over some reasonable set and repeat steps 2 through 4.

6. Compute the estimate: $\hat{k} = \arg_{k_p} \min \gamma(k_p)$
The variance of the estimate, $\hat{k}$, can be computed by resampling methods, such as the bootstrap. The estimation procedure is a general technique and can be used for estimating other parameters.

6 Experimental Protocol and Results

In this section we outline the protocol we use to conduct the experiments. Here we give all the sample sizes used, the number of trials that were run at different stages, the exact model parameter values used for the synthetically degraded characters, etc. In the following discussion we assume that all the components of the degradation model, except the structuring element size $k$, are known.

The following protocol was used for creating the reference sample $X$ and probe samples $Y$. The reference distribution, $X$, was generated using parameter $\theta = \theta_r$ where the structuring element size $k$, was fixed at $k = k_r = 5$ and rest of the components of parameter $\theta_r$ were fixed at following values: $\eta_I = \eta_o = 0$, $\alpha_0 = \beta_0 = 1$, and $\alpha = \beta = 1.5$. All the components of the probe distribution parameter $\theta_p$ were same as that of the reference distribution parameter $\theta_r$ except the structuring element size, $k_p$, which was varied over the set $k_p \in \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. 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 parameters used were as follows:

1. Sizes of samples selected from bootstrap pool, $N = M = 40$.
2. Number of permutations, $K = 1000$.
3. Significance level of the test, $\epsilon = 0.05$.
4. Number of repetition trials, $N_{\text{Repeat}} = 100$.
5. The character-to-character distance, $\delta(x, y)$, used was the Hamming distance.
6. The set-to-set distance, $\rho(X, Y)$, used was the mean nearest-neighbor distance.

The noise free document is shown in figure 3 (a). The reference degraded document generated with model parameter $\theta_r = 5$ is shown in figure 3 (b). The expected value of the reject rates as a function of probe distribution structuring element size, $k_p$, for sample size 40 is shown in figure 4. Note the function has a misdetection (reject) rate better than $\epsilon = 0.05$ when the probe distribution has a parameter value close to that of the reference distribution ($k_r = k_p = 5$). Furthermore, when $k_p$ is far from 5, $\gamma(k_p)$ is close to 1, which implies that the validation procedure can distinguish the probe samples $Y$ from the reference sample $X$ with high probability. Thus, according to our estimation procedure, $\hat{k} \in \{5, 6\}$.

7 Summary

We used a nonparametric parameter estimation procedure to estimate the structuring element size, $k$, used in the document degradation model we had proposed earlier [KHP93, KHP94]. For a given sample of degraded characters, $X$, generated with unknown structuring element size $k$, we (i) generate a sample $Y$, with structuring element size fixed at $k$; (ii) run our validation procedure that accepts or rejects the null hypothesis that $X$ and $Y$ are samples from the same underlying population with a significance level $\epsilon$; (iii) find the expected value of the reject rate by repeating steps (i) and (ii) many times; (iv)
mical behavior of these systems described by sets of coupled equations for formal neurons with investigation of essential features of plasticity and adaptation depends on the mathematical theory to be applied. The efficient analysis of dynamic systems is often performed in theoretical research where theoretical results are frequently obtained from formal models. The neurodynamic systems due to various biological facts amount to a degree as large as .

(a)

(b)

Figure 3: Sample images for the Morphological Document Degradation Model. (a) Subimage of the noise-free document. (b) Reference degraded document simulated with \( k = k_r = 5 \). All the true letter ‘e’ of the correct font type and size were extracted out and made into the sample \( X \). The samples size was 60.

repeat steps (i) through (iii) for various values of \( k \); (v) use the value for which the expected reject rate is minimum as the best estimate of \( k \). The variance of the estimate can be computed by resampling methods, such as the bootstrap. The estimation procedure is a general technique and can be used for estimating other parameters.

Acknowledgement: We would like to thank Werner Stuetzle and David Madigan of Statistics Department, University of Washington for many discussions on nonparametric tests. We would also like to thank Henry Baird of AT&T Bell Labs, Murray Hill, for discussions on the validation problem.

References

<table>
<thead>
<tr>
<th>Author(s)</th>
<th>Title</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>S. F. Arnold</td>
<td><em>Mathematical Statistics</em></td>
<td>1990</td>
</tr>
</tbody>
</table>


Figure 4: Plots of expected values of reject, $\gamma(k_p)$, as a function of the probe distribution parameter $k_p$. The reference distribution had $k_r = 5$. Notice that $\gamma(k_p)$ has a minimum near $k = k_p = k_r = 5$. The sample sizes were: $N = M = 40$. The minimum is at sizes 5 and 6. The correct size is 5. We get a minimum at both sizes because in our implementation disk structuring elements of diameter 6 and diameter 5 do not differ much after discretization.