MORPHOLOGICAL DEGRADATION MODELS AND THEIR USE IN DOCUMENT IMAGE RESTORATION

Qigong Zheng and Tapas Kanungo

Center for Automation Research
University of Maryland College Park
College Park, MD 20742

ABSTRACT

Document images undergo various degradation processes. Numerous models of these degradation processes have been proposed in the literature. In this paper we propose a model-based restoration algorithm. The restoration algorithm first estimates the parameters of a degradation model and then uses the estimated parameters to construct a lookup table for restoring the degraded image. The estimated degradation model is used to estimate the probability of an ideal binary pattern, given the noisy observed pattern. This probability is estimated by degrading noise-free document images and then computing the frequency of corresponding noise-free and noisy pattern pairs. This conditional probability is then used to construct a lookup table to restore noisy images. The impact of the restoration process is then quantified by computing the decrease in OCR word and character error rate.

We find that given the estimated degradation model parameter values, the restoration algorithm decreases the character error rate by 16.1% and the word error rate by 7.35%. In some categories of degradation (e.g. model parameters that give rise to broken characters) there is a 41.5% reduction in character error rate and 20.4% reduction in word error rate.

1. INTRODUCTION

Document images are usually corrupted by various types of noises during the document generation and copying processes. We wish to design a filter to restore a class of document images with similar structural features and degradation conditions. A traditional approach to this problem is by means of linear filters [1]. Although linear filters are mathematically simple, their use usually results in distortion of many important image characteristics. In this paper we propose an algorithm to create a look-up-table that can be used for restoring degraded images.

The issue of morphological filter design has been studied by numerous researchers [2, 3, 4]. These algorithms do not incorporate prior noise model characteristics into the filter design. This suggests that the restoration algorithms may be further improved by using the image noise model.

A survey of document image degradation models proposed in the literature can be found in [5]. We use the model proposed by Kanungo et al. [6, 7] for our restoration algorithm.

2. DOCUMENT DEGRADATION MODEL

Our degradation model [6] has six parameters: $\Theta = (\eta, \alpha_0, \alpha, \beta_0, \beta, k)$. We model the probability of a pixel flipping from foreground to background or vice-versa as an exponential function of its distance from the nearest boundary pixel. The foreground and the background 4-neighbor distance d are computed using a standard distance transform algorithm. The flipping probabilities of the foreground and background pixels are controlled by α_0, α and β_0, β respectively. The parameters α_0, β_0 are the initial values for the exponentials and decay speed of the exponentials are controlled by the parameters α, β. Parameter η is the constant probability of flipping for all pixels. Parameter k is the size of the disk used in the morphological closing operation. This operation normally simulates the correlation introduced by the point-spread functional of the optical system. The procedure to degrade an ideal binary image is as follows:

1. Compute the distance d of each pixel from the character boundary.
2. Flip each foreground pixel with probability
$$p(0|1, d, \alpha_0, \alpha) = \alpha_0 e^{-\alpha d^2} + \eta.$$
3. Flip each background pixel with probability
$$p(1|0, d, \beta_0, \beta) = \beta_0 e^{-\beta d^2} + \eta.$$
4. Perform a morphological closing operation with a disk structuring element of diameter k.

Figure 1 illustrates the ideal and degraded images with different model parameters. Note that the two degraded images differs in the speed of decay of exponential functions. If $\alpha < \beta$, more foreground pixels change to the background so the images look like being corrupted by subtractive noise. If $\alpha > \beta$, more background pixels change to the foreground so the images are more like having additive noise.

3. THE ESTIMATION ALGORITHM

In this section, we describe a parameter estimation algorithm for the abovementioned degradation model [8]. Inputs to the estimation algorithm are the ideal and degraded images. The basic idea behind our estimation algorithm is that we assume two images that looks alike should possess the same model parameters. Let I be the ideal image and R be the given degraded image. The problem is to estimate the model parameter Θ such that if we degrade I with the parameter fixed at Θ, we will get an image S_θ that looks similar to R. It's quite important for us first to characterize the measure of similarity between two images. The simplest way to measure the similarity is to compute the absolute difference between two images but this also requires perfect alignment between two images, which is nearly impossible if the character level geometric groundtruth is not available. We drive a more robust similarity measure by looking at the distribution of the noise patterns. We say that two image R and S are similar if the corresponding pattern distributions are similar.

Let P be a set of neighborhood bit patterns and p be an arbitrary element in the set P. If we choose (3×3) neighborhood,we will have totally 512 different pattern distributions. Let $H_R(p)$ $(p \in P)$ denotes the number of times the pattern p occurs in the binary image R. In terms of morphological morphology, we could define $H_R(p)$ more precisely as:

$$H_R(p) = \#\{R \ominus p\}. \tag{1}$$

We say that two image R and S are similar if the corresponding pattern distributions H_R and H_{S_θ} are similar. To test similarity of two pattern distributions, we use the Kolmogorov-Smirnov test of the two 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 will use this p-value as the objective function that the estimation process tries to maximumize. That is,

$$\hat{\theta} = arg_\theta max KS(H_R, H_{S_\theta}) \tag{2}$$

Conventional optimization algorithm normally need a closed form to minimize or maximize the objective function. But in our case, since S_θ is computed by simulation, it's impossible to use the standard derivative approach to solve the problem. Thus we choose the simplex optimization algorithm to minimize KS. To prevent the problems of local minimum, we select multiple random starts and then pick the one with the lowest P-value.

4. THE RESTORATION ALGORITHM

Compared to other morphological restoration algorithms [3, 4], our method is model-based. We always assume that the degraded image can be characterized by a set of parameters such that it can be estimated by using the algorithm we described in the previous chapter. Our algorithm has two stages, a training stage and a restoration stage.

Suppose we have an ideal image I and a corresponding degraded image $S_{\hat{\theta}}$ where $\hat{\theta}$ is the estimated parameter set used to generate $S_{\hat{\theta}}$ from I. The training stage is responsible for computing the conditional distribution between the noise pattern pairs in the image pair: $(I, S_{\hat{\theta}})$. During the training stage, we first scan $S_{\hat{\theta}}$. Next we obtain its noise pattern $P_S(x, y)$ at the location (x, y). We also obtain the point pattern at location (x, y) in the ideal image I: $P_I(x, y)$. From the pattern pairs $(P_I(x, y), P_S(x, y))$, we form the pattern distribution of an ideal image I conditioned on the degraded image $S_{\hat{\theta}}$: $H_{\hat{\theta}}(P_I|P_S)$. The restoration stage is conducted after estimating the model parameters of the degraded image. Let Q represents the restored image version of $S_{\hat{\theta}}$. Given the pattern $P_S(x, y)$ at location (x, y) of the degraded image $S_{\hat{\theta}}$, the restored pattern $P_Q(x, y)$ in Q is computed as:

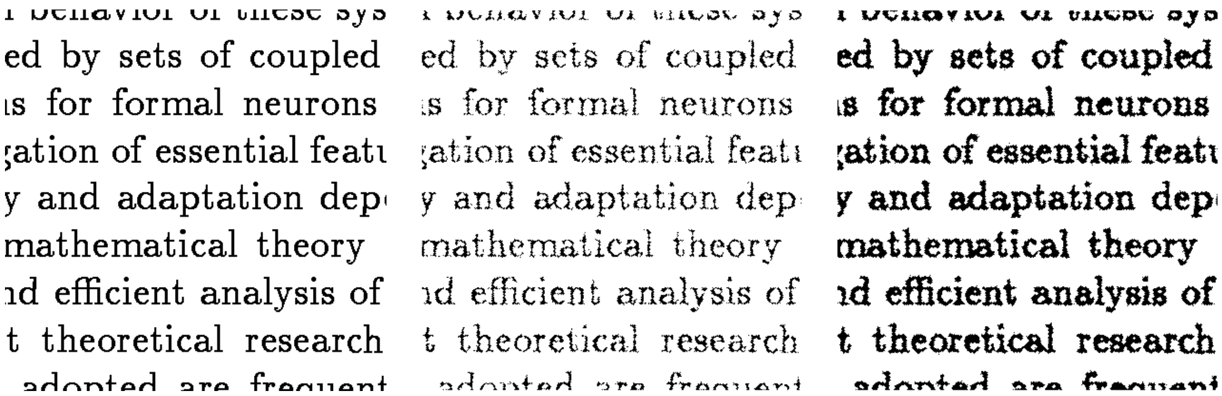
$$P_Q(x, y) = \arg\max_{p \in P_I} H_{\hat{\theta}}(p|P_S(x, y)) \tag{3}$$

Equation (3) is essentially the Maximum Likehood (ML) estimate of the pattern based on the estimated parameter $\hat{\theta}$. Figure 2 shows four typical noise patterns in degraded image in Figure 1(b) and its conditional pattern distribution based on the corresponding ideal image in Figure 1(a). Figure 3(a) and Figure 3(b) are the restored images corresponding to the degraded images in Figure 1(b) and Figure 1(c).

5. EXPERIMENT PROTOCOL AND RESULTS

The experiment outline is illustrated in Figure 4. The basic idea is to compare the OCR result of the degraded image with that of the restored one. The evaluation software is provided by the University of Maryland. It compares the OCR outputs and the corresponding groundtruth information and generate statistical information such as character-level or word-level accuracy in a batch mode. We believe that the OCR accuracy rate is a good and objective indicator for showing how well our algorithm improves the overall image quality.

The test images were 100 pages (one-column) of English Bible that were typeset using LaTeX. The image size



ed by sets of coupled
ıs for formal neurons
;ation of essential featı
y and adaptation dep
mathematical theory
ıd efficient analysis of
t theoretical research

(a) (b) (c)

Fig. 1. (a) A typical ideal image; (b) Degraded version of (a) with parameters $(1.0, 0.7, 1.0, 3.0)$; (c) Degraded version of (a) with parameters $(1.0, 3.0, 1.0, 0.7)$.

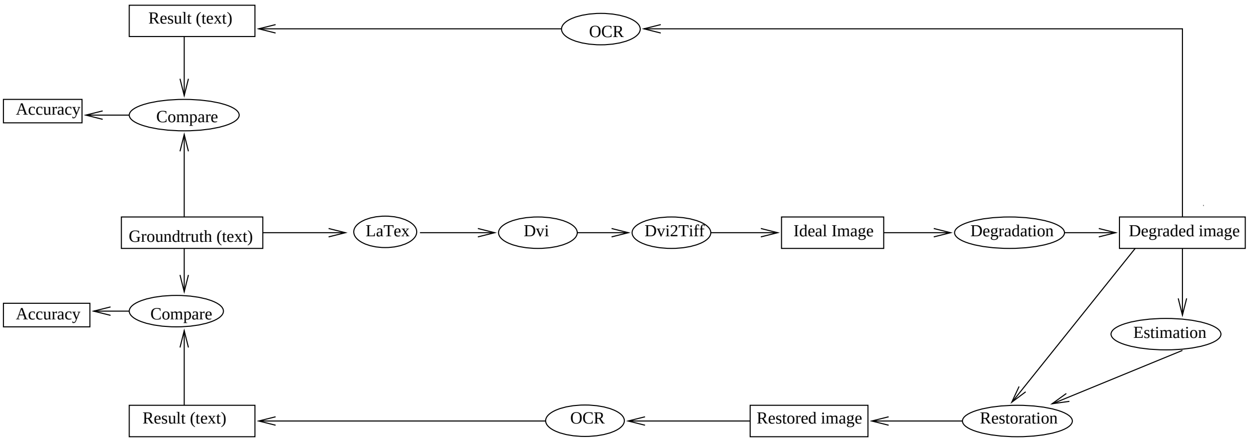


Fig. 4. Illustration of the experimental setup to compare OCR accuracy on restored versus unrestored images.

is A4 with 12-point font size. The 100 test images were degraded and then categorized into ten groups with each group possessing an unique parameter set. The OCR product was FineReader4.0, manufactured by ABBYY. Table 1 gives the OCR accuracy improvement before and after our restoration algorithm with the specific parameter set $(\alpha_0, \alpha, \beta_0, \beta) = (0.8, 0.8, 1.0, 3.0)$.

From the Table 1, we see that our restoration algorithm decreases both the OCR error rate and image noise level. In this special case, the OCR error rate at the character and word levels get improved by 19% and 9.0%. For all of 100 images we tested, the decreases in OCR accuracy error rate at the character and word levels range from 3.4% to 41.5% and from 1.0% to 20.4% respectively, depending on what model parameters are associated with the degraded images. The average improvement are 16.1% at character level and 7.35% at word level. In particular, we find that our algorithm performs better in restoring the images suffering from broken characters (Figure 1(b)) than those that have blurred chracters (Figure 1(c)). This gives us the impression that the OCR product seems to be more vulnerable to broken characters which have more subtractive noise. In addition to the OCR error rate, our algorithm significantly decreases the image noise level by amount, ranging from 13.1% to 52.7%.

6. REFERENCES

[1] A. K. Jain, *Fundamentals of Digital Image Processing*, Prentice Hall, Englewood, Cliffs, NJ, 1989.

[2] J. Serra, *Image Analysis and Mathematical Morphology*, Academic Press, London, 1982.

[3] R. P. Loce and E. R. Dougherty, “Facilitation of optimal binary morphological filter design via structuring

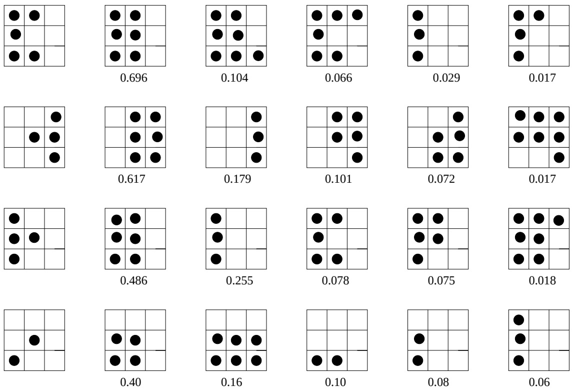
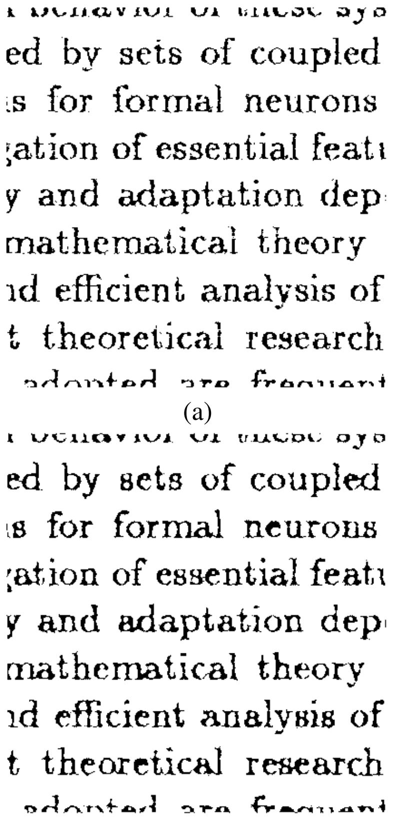


Fig. 2. Four typical noise patterns are shown in the left-most column. The pattern entries in the other columns show the possible ideal pattern and the corresponding probabilities. The ideal image was degraded with parameter set $(1.0, 0.7, 1.0, 3.0)$.



ed by sets of coupled
s for formal neurons
ation of essential feat
y and adaptation dep
mathematical theory
d efficient analysis of
t theoretical research

(a)

ed by sets of coupled
s for formal neurons
ation of essential feat
y and adaptation dep
mathematical theory
d efficient analysis of
t theoretical research

(b)

Fig. 3. (a) The restored version of the image shown in Figure 1(b). (b) The restored version of the image shown in Figure 1(c).

Table 1. OCR error improvement with parameters $(\alpha_0, \alpha, \beta_0, \beta) = (0.8, 0.8, 1.0, 3.0)$.

OCR Result	Degraded	Restored	Improvement
Num. of Chars	24391	24806	
Num. of Correct Chars	23935	23999	
Num. of Char Errors	996	807	19.0%
Num. of Words	4953	4953	
Num. of Correct Words	3737	3846	
Num. of Word Errors	1216	1107	9.0%
Foreground Noise Level	22.2%	14.7%	
Background Noise Level	0.18%	0.24%	
Num. of Error Flipping Pixels	625228	516481	17.4%

element libraries and design constraints," *Optical Engineering*, vol. 31, pp. 1008–1025, 1992.

[4] J. Liang and R. M. Haralick, "Document image restoration using binary morphological filters," in *Proceedings of SPIE*, 1996, vol. 2660, pp. 274–285.

[5] H. S. Baird, "Document image quality: Makeing fine discriminations," in *Proceedings of the International Conference on Document Analysis and Recognition*, 1999, pp. 459–462.

[6] T. Kanungo, R. M. Haralick, and I. Phillips, "Nonlinear global and local document degradation models," *International Journal of Imaging Systems and Technology*, vol. 5, pp. 220–230, 1994.

[7] T. Kanungo, R. Haralick, H. Baird, W. Stuetzle, and D. Madigan, "A statistical, nonparameteric methodology for document degradation model validation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1209–1223, 2000.

[8] T. Kanungo and Q. Zheng, "Estimation of morphological degradation model parameters," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001, (To appear.).