A Two-state Markov Chain Model of Degraded Document Images

Shamik Sural, NIIT Limited, 6B Pretoria Street, Calcutta 700 071, India.
P.K.Das, Dept. of Computer Science & Engineering, Jadavpur University, Calcutta 700 032, India.

Abstract

We propose a two-state Markov chain model of degraded document images. The model generates random and burst noise to simulate isolated pixel reversal as well as blurring of a larger document region. In the Random state, the probability of pixel inversion is low compared to that in the Burst state. However, the model remains in the Random state for a much longer period of time. Validation of the model has been done using the statistical methodology of Kanungo et al [9]. To estimate the parameters efficiently, we use Genetic Algorithm (GA) to search through the parameter space in which the model parameter values are encoded into a concatenated bit string to form the chromosomes. We also show how the accuracy of an optical character recognition system with dictionary search varies with two derived parameters of the proposed noise model.

Keywords : Image degradation model, Markov chain, Model validation, Parameter estimation, Genetic algorithm.

1. Introduction

Over the last few years, a lot of importance is being given to the problem of modeling document image defects so that a formal evaluation of the different optical character recognition (OCR) systems can be done. Most of the character recognition systems of today are found to be suitable for a specific type and quality of image. The methods and algorithms used in the development of these OCRs are often biased by the researcher's choice of the training and the test data sets. As a result, such systems perform excellently for the data sets chosen by the researcher. In many cases, however, the recognition accuracy falls sharply when even a slightly degraded image is chosen. The fall in recognition accuracy is often high compared to the visual nature of the degradation, i.e., the degradation as perceived by the human eye. It has, therefore, been felt necessary to model the defects quantitatively and experiment with extensive simulation to determine the nature of image defects that result in higher failure rate.

Baird [1,2] first described a defect model which includes a number of document image parameters, namely, size, resolution, horizontal and vertical scaling factors, translational offsets, jitter, defocussing, sensitivity and binarization threshold. Methods of calibrating image defect models were later discussed by him [3]. Review of the then state of the art of calibration methods and further discussions on document image defect models, specially the defects due to the physics of apparatus for printing and imaging, was also made by Baird [4]. This line of research has resulted in the "Bell Labs Image Defect Model Database". Ho and Baird [8] have presented a more recent report on similar defect models.

An alternative approach to modeling is the morphological document degradation model proposed by Kanungo et al [10]. Their model simulates both the statistically independent pixel inversion that occurs in images and the blurring caused by point spread function of the scanner optical system. Kanungo et al [9] have proposed a statistical methodology for the validation of document image models and estimation of model parameters.

We model the random pixel inversion that occurs in the form of so called "salt and pepper" noise as well as blurring of a large region of a document image using a two-state Markov chain. In document images, black dots appear on the white background while white specks are formed on the foreground connected regions. Usually, one or two consecutive pixels are reversed due to the presence of random noise and a few document regions may be totally distorted in the form of a patch of bit reversal. Since both the types of noise affect the foreground as well as the background pixels, we treat them uniformly.

In the next section, we describe the proposed two-state model of degraded document images. The model validation procedure and a genetic algorithm for the estimation of model parameters are discussed in section 3. We present the experimental results in section 4 and summarize our work in the last section.

2. Proposed document image degradation model

Traditionally, the effect of different types of noise has been studied in communication sciences, especially in

Fifth International Conference on Document Analysis and Recognition (ICDAR99), Bangalore, India, 1999, pp. 463-466

coding theory. Noise in communication channels occurs both randomly as well as in a burst. One of the models used for simulation of such communication channels is the Gilbert-Elliott burst noise channel model [5,6]. A burst error correcting Viterbi algorithm was later developed based on this channel model [11]. We have used this idea of random and burst noise to model a degraded document image.

To model a degraded document image, we use a two-state Markov chain in which one state produces errors in the image with a probability 'r' while the other state corrupts the image pixels with a probability 'b' as shown in figure 1.

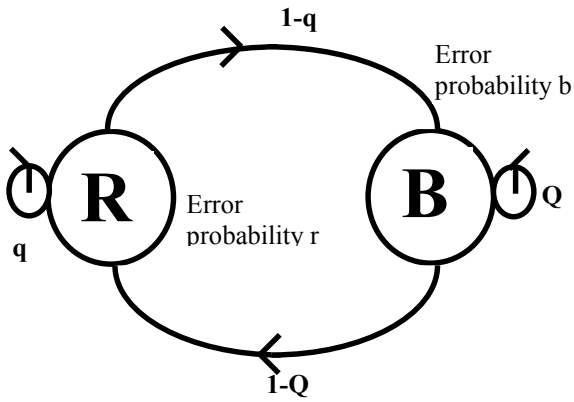

Figure 1. Two-state Markov chain model of a degraded document image

Here r << b and the state with lower error probability is called the Random state (R) while the other state is referred to as the Burst state (B). The transition probabilities are 'q' and 'Q' where q is the conditional probability that the image remains in the Random state for the next pixel position, given that it is in the Random state for the current pixel. With probability (1-q), it makes a transition to the Burst state. Q is also defined similarly for the Burst state. The transition probability matrix for the Markov chain is $P = \begin{bmatrix} q & 1-q \\ 1-Q & Q \end{bmatrix}$. Here p_{ij}, i,j = 1,2 is the transition probability from state i to state j in one step. The steady state probabilities of the document image being in the Random state and in the Burst state are P_R= (1-Q)/(2-Q-q) and P_B= (1-q)/(2-Q-q). The average pixel error probability P_e on the document is $P_e = bP_B + rP_R$

In a document image, which is inherently two-dimensional in nature, we define a Burst start as an event when pixels lying on the neighborhood of the current pixel are affected with higher error probability. The burst first propagates to the 8 neighboring pixels of the current pixel, then to the next 16 neighbors, followed by the next 24 neighbors and so on. Thus, propagation of Burst is spatial in a noisy document unlike communication channels where it is temporal in nature. The parameters of the noise model proposed by us, are thus $\theta = (q, Q, r, b)^T$. Here the parameters q and r control the random noise in the document. Q, on the other hand, controls the duration and propagation of a noise Burst while b determines the probability of a pixel getting affected during the Burst. It should be noted that a Burst does not necessarily mean that all the pixels are reversed in this state. Rather, it signifies that pixels falling in a Burst sequence have a higher probability of getting reversed.

From the primary parameters of the noise model, two other parameters can be derived which are suitable for the actual process of simulation. We have also made an attempt to link the character recognition efficiency using dictionary search with these two derived parameters. Since most of the currently available commercial OCR packages employ the dictionary search step, we feel it is important to undertake this type of simulation study. The error density ratio, which is the ratio of the Burst state error probability and the Random state error probability, is defined as Δ=b/r. For different document images with the same pixel error probability P_e, Δ is an indicator of the severity of the bursts in the image. The average burst length, λ is defined as the average number of pixels for which the image remains in the burst state. Here, λ=Q/(1-Q). The simulation test results using the derived parameters Δ and λ will be presented in section 4.

3. Model validation and parameter estimation

We use the statistical validation method proposed by Kanungo et al [9] for the validation of our noise model. The model validation procedure uses two degraded character sequences of which one is the *real* image sequence $X = \{\bar{x}_1, \bar{x}_2, ..., \bar{x}_N\}$ and the other is a *synthetic* sequence $Y = \{\bar{y}_1, \bar{y}_2, ..., \bar{y}_N\}$ where $\bar{x}_i, \bar{y}_i \in \{0,1\}^D$ D, being the dimension of a vector generated from a two dimensional character pattern. The model validation procedure is to test the Null hypothesis: X and Y have the same statistical distribution. The test is carried out in the following steps.

Step 1 – *Initialize*
Two data sets : Real data $X = \{\bar{x}_1, \bar{x}_2, ..., \bar{x}_N\}$ and Synthetic data $Y = \{\bar{y}_1, \bar{y}_2, ..., \bar{y}_M\}$ are given. Size of test $\varepsilon = 0.05$.
A set distance metric $\rho(X,Y)$ is defined as :
$\rho(X,Y) = (\rho(Y|X) + \rho(X|Y))/(N+M)$, where
$\rho(Y|X) = \sum_{\bar{x} \in X} (\min_{\bar{y} \in Y} \delta(\bar{x}, \bar{y}))$ and

$\rho(X|Y) = \sum_{\bar{y}\in Y} (\min_{\bar{x}\in X} \delta(\bar{x},\bar{y}))$, $\delta(\bar{x},\bar{y})$ = Hamming distance $(\bar{x},\bar{y})$. Compute initial distance $d_0 = \rho(X,Y)$.

Step 2 – *Partition*

Mix the samples X and Y and randomly select N vectors to form a new sequence X'. The remaining M vectors form another sequence Y'. Compute $d_i = \rho(X',Y')$.

Step 3 *Repeat Partition*

Repeat step 2, K number of times.

Step 4 *Compute Probability*

Compute the probability $p_0 = P(d \geq d_0)$ as

$$p_0 = \frac{\#\{k \mid d_k \geq d_0\}}{K}$$

Step 5 *Test Null Hypothesis*

Reject Null hypothesis that the two samples X and Y come from the same population if $p_0 < \varepsilon$.

A power function of the validation procedure is generated by taking both the samples X and Y from synthetic distribution using a reference parameter set θ_r and a probe parameter set θ_p, respectively. The above mentioned 5 steps are repeated T number of times, each time generating reference samples and probe samples by keeping θ_r fixed and varying θ_p. The reject rate $\gamma(\theta_p)$ is computed as the ratio of the number of times the null hypothesis is rejected. A plot of $\gamma(\theta_p)$ vs. θ_p is the power function of the validation procedure for the set of fixed values for M, N, K, T and the distance metric $\rho(X,Y)$.

The parameter estimation step is carried out to estimate the values of the parameters $\theta' = (q', Q', r', b')^T$ for which the degraded document image generated by this model is statistically similar to a real image in the sense of the validity of the Null hypothesis. A procedure similar to power function generation is repeated except that the reference sample is not generated for each new value of θ_p i.e., the reference sample is fixed and the parameter space is searched to determine the values of the parameters which best fit the reference sample. In the experiments, however, so far we have used synthetic samples generated by known values of parameters for the reference sample and checked that the parameter estimation method is working correctly.

It has been observed that an exhaustive search through the parameter space is highly computation intensive. We, instead, use a genetic algorithm [7] to speed up this search process. In our model, all the four parameters are real numbers. We encode the parameters as binary bit strings and concatenate them to form an encoded parameter string. As the parameters q, Q, r, b represent finite probabilities, their domain is [0,1]. The process of encoding them is, therefore, simply a process of decimal to binary conversion. We have used five bits to represent each of these parameters so that the precision is 2^{-5} = 0.03125. The complete binary string representing each parameter set thus becomes :

$q_1q_2q_3q_4q_5Q_1Q_2Q_3Q_4Q_5r_1r_2r_3r_4r_5b_1b_2b_3b_4b_5$ where

$q_i, Q_i, r_i, b_i \in \{0,1\}\ \forall\ i \in \{1,2,3,4,5\}$.

Starting with an initial random population of 20 chromosomes, the genetic algorithm moves through a number of generation creation steps using the operators *reproduction*, *crossover* and *mutation* where more viable chromosomes are retained in the new generation. By design, a more viable chromosome is a probe parameter set with lower rejection rate. We keep track of the best fitting parameters detected till any point of time and use a standard stopping criteria for the genetic algorithm.

4. Experimental results

We have used the two-state Markov chain model mentioned above to generate noise in a good quality document image and studied the variation of OCR performance with the noise parameters. Our character recognition system uses a number of fuzzy features for training a Multilayer Perceptron (MLP) and the outputs are obtained as fuzzy sets denoting the belongingness of an input pattern to a number of fuzzy character pattern classes [12]. A dictionary search is used to ascertain the characters if more than one MLP output has a fuzzy set membership value higher than a threshold value *th_opt*.

The ability to correctly identify a character through dictionary search depends on the noise distribution in the document image. If, within a word, one or only a few characters are affected and the rest are recognized correctly, a dictionary search can uniquely identify the undetected characters. Recognized characters in their positions within the word and unrecognized possibilities in their positions are considered for this search. However, if the noise distribution is such, that a number of short bursts occur in more than one character position within a word, then the number of possible choices may become quite large, increasing the search time. It may not also be possible to uniquely determine the character from word level knowledge only. The characters remain unresolved and are identified by user intervention.

Figure 2 shows the variation in unresolved error after dictionary search for different values of λ, the average burst length. We keep the total probability of error P_e = 0.005, while b = 0.5 and r = 0.0005. λ is varied from 10 to 200. From the figure, it is seen that for small values of λ, the recognition error percentage is higher and it goes down with increase in the value of λ. When λ is small, there are a large number of noise bursts, each with short length. As a result, the number of recognized characters is less resulting in ambiguous choice from the dictionary search. As λ increases, the number of bursts goes down and the affected characters are resolved by the dictionary search, reducing the recognition error.

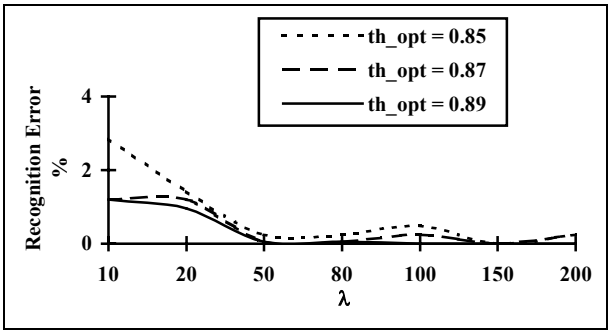

Figure 2. Variation of recognition error with average burst length

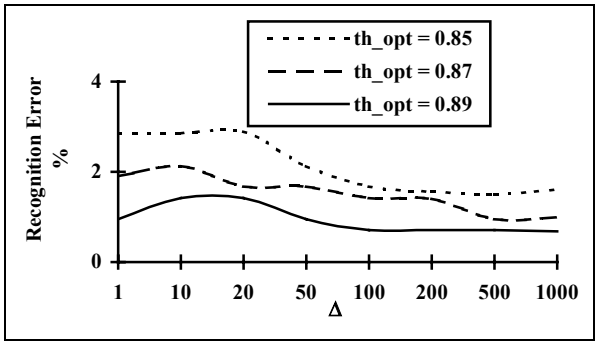

Figure 3. Variation of recognition error with error density ratio

We have also varied the noise distribution in the document image for different values of Δ, the ratio of burst state error probability to the random state error probability. The recognition results are shown in figure 3. Here also, the total probability of error P_e = 0.005. We maintain q = 0.999 and Q = 0.99. Δ is varied from 1 to 1000. From the figure it is seen that for small values of Δ, unresolved error is higher. The error percentage then decreases and remains almost constant for higher values of Δ. For small values of Δ, error occurs with almost equal probability in both the random state and the burst state, increasing the number of unresolved errors. For higher values of Δ, error due to random noise is corrected by the fuzzy MLP itself while the errors caused by dense bursts are resolved in the dictionary search step.

From the figures 2 and 3, it is also seen that the unresolved error percentage is lower for higher values of MLP output threshold *th_opt*. For low threshold values, the number of possible choices is high for each character position and, hence, the dictionary search often cannot uniquely determine the characters. However, for very high threshold values, all the MLP outputs may fall below the threshold so that the dictionary search is to be made with wildcards, increasing the search time.

5. Summary

We have proposed a two-state Markov chain model of degraded document images. The model generates noise both randomly as well as in a burst. In the random state, isolated pixels of the image get reversed while in the Burst state, there is a larger patch of affected pixels. A statistical validation of the model has been done and a genetic algorithm has been suggested for the estimation of model parameters. The genetic algorithm can be applied to the estimation of parameters for other document defect models also. We have presented simulation results of our experiments using an OCR with dictionary search step for resolving ambiguous characters.

References

1. Baird, H.S. (1990). Document image defect models. Proc., *IAPR Workshop on Syntactic and Structural Pattern Recognition*. Murray Hill, NJ., 38-46.
2. Baird, H.S. (1992). Document image defect models. In H.S.Baird, H.Bunke and K.Yamamoto (Eds.) *Structured Document Image Analysis*. Springer Verlag, New York.
3. Baird, H.S. (1993). Calibration of document image defect models. *Proc. Fourth Annual Symposium on Document Analysis and Information Retrieval.* Las Vegas, Nevada, 1-16.
4. Baird, H.S. (1993). Document image defect models and their uses. *Proc. Second International Conference on Document Analysis and Recognition*, Japan, IEEE Computer Society Press, CA, USA, 62-67.
5. Elliott, E.O. (1963). Estimation of error rates for codes on burst channels. *Bell Systems Technical Journal,* 1977-1997.
6. Gilbert, E.N. (1960). Capacity of a burst-noise channel. *Bell Systems Technical Journal,* 1253-1265.
7. Goldberg, D.E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley. Reading, Mass.
8. Ho, T.K. and H.S. Baird (1997). Large scale simulation studies in image pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1067-1079.
9. Kanungo T, H.S. Baird and R.M. Haralick (1995). Validation and estimation of document degradation models. *Proc. Fourth Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, Nevada, 217-225.
10. Kanungo T, R.M.Haralick and I.Phillips (1993). Global and local document degradation models. *Proc. Second International Conference on Document Analysis and recognition*, Tsukuba, Japan, 730-734.
11. Schlegel, C.B. and M.A.Herro (1990). A burst error correcting Viterbi algorithm. *IEEE Transactions on Communication* 38, 285-291.
12. Sural S. and P.K.Das (1999). Fuzzy Hough transform and an MLP with fuzzy input/output for character recognition. *Fuzzy Sets and Systems* (to appear).