

1-1-2005

Text Degradations and OCR Training

Elisa H. Barney Smith
Boise State University

Tim Andersen
Boise State University

Text Degradations and OCR Training

Elisa H. Barney Smith and Tim Andersen

Boise State University

Boise, ID 83725-2075 USA

EBarneySmith@boisestate.edu, tim@cs.boisestate.edu

Abstract

Printing and scanning of text documents introduces degradations to the characters which can be modeled. Interestingly, certain combinations of the parameters that govern the degradations introduced by the printing and scanning process affect characters in such a way that the degraded characters have a similar appearance, while other degradations leave the characters with an appearance that is very different. It is well known that (generally speaking) a test set that more closely matches a training set will be recognized with higher accuracy than one that matches the training set less well. Likewise, classifiers tend to perform better on data sets that have lower variance. This paper explores an analytical method that uses a formal printer/scanner degradation model to identify the similarity between groups of degraded characters. This similarity is shown to improve the recognition accuracy of a classifier through model directed choice of training set data.

1. Introduction

It is relatively common to design classifiers with the assumption that there is large within class similarity and low between class similarity. The within class similarity is often increased by restricting the problem to a “common” domain, or dividing a larger non-homogeneous problem into multiple problems each of which exhibit larger homogeneity. This has often been done in OCR problems by assuming common font in machine printed text or a single writer in handwritten text. The problem of high between-class similarity is often solved or lessened by the same efforts just mentioned to increase within class similarity.

To improve classifier performance when the input is not guaranteed to be homogeneous, it is also desirable to have a large and varied training set to improve the ability of the classifier to generalize. This however leads to lower within class similarity.

For optical character recognition problems, the typical

approach that is used to maximize within class similarity and minimize between class similarity for classifier training is to restrict the training data to a particular font and style. This approach requires that some form of font/style detection be employed in order to select the appropriate classifier (the one that was trained on the font style being recognized) during execution. For example, [8] uses a nearest neighbor classifier in conjunction with the tangent distance to perform font recognition in order to improve classification accuracy for a multi-font OCR engine. In [11] an error rate reduction of 20% is achieved on a handwriting recognition problem when writer style is taken into consideration.

This paper will further study the problem of training data selection from within the context of degradations introduced by a common text degradation model. The degradation model used for this research is based on the model developed by Baird [1]. The PSF width and the binarization threshold are the two most significant parameters in Baird’s model affecting degradations of bilevel images [6]. The PSF accounts for the blurring caused by the optics of the scanner. Its functional form is not constrained, but needs to be specified. In this work, the PSF is assumed to be a bivariate Gaussian with the width, w , equal to the standard deviation. The size is in units of pixels, which allows the model to be used for scanning at any optical resolution. The resulting gray level image is converted to a bilevel image with a global threshold, Θ . The units for the threshold are absorbance. Additive noise is also incorporated in this degradation model. The noise is Gaussian distributed with a standard deviation, s . This is added independently to each pixel in the image prior to thresholding. In addition to variations from the PSF width, threshold level and noise, variations in the resulting bilevel bitmaps also come from phase effects [10]. The character samples will be divided within the degradation space. We will not directly estimate the parameters to the degradation model. Several papers address this topic [4, 5].

Each pair of degradation parameter values will typically affect an image differently. However, there are several combinations of these parameters that produce degraded

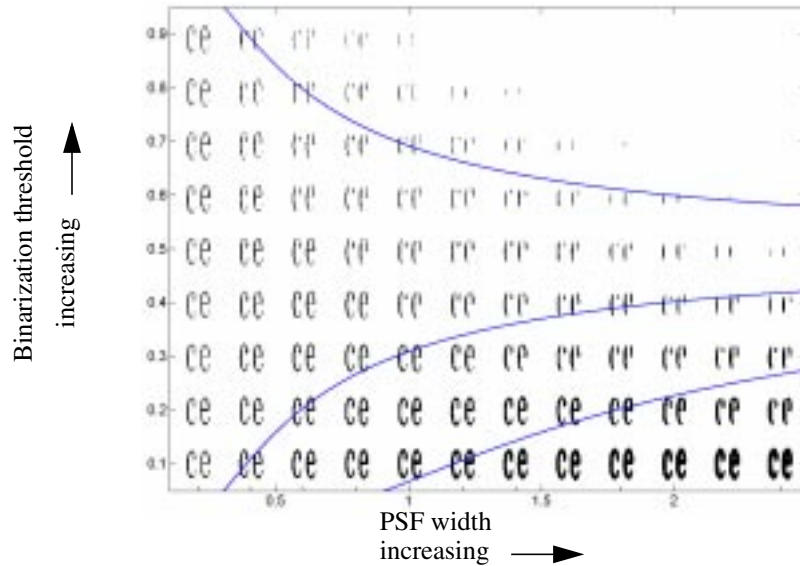


Figure 1: Characters after blurring and thresholding over a range of PSF widths, w , and binarization thresholds, Θ . A broad range of character appearances can be seen, but certain characters have some general similarities.

character images that are highly similar in appearance. Two primary image degradations associated with bilevel processes were defined in [3, 4]. These are the edge displacement, δ_c , and the erosion of a black or white corner. Both these degradations are functions of the degradation model parameters, w and Θ . The edge displacement,

$$\delta_c = -w \text{ESF}^{-1}(\Theta), \quad (1)$$

is the amount that an isolated edge would be displaced if it is subjected to the blurring and thresholding degradation model [3]. $\text{ESF}()$ is the Edge Spread Function which is the integral of the PSF.

A statistical test was conducted in [2] to compare the similarity between groups of characters synthetically generated with parameters (w, Θ) varying over the parameter space. This test showed that the amount of variation in the characters correlated highly with the change in the edge spread degradation. Changes of the degradation parameters that remain along constant δ_c isolines will not produce as large a difference in the characters as other changes of the degradation model parameters. Figure 1 shows the characters c and e each degraded over a variety of PSF widths and binarization thresholds without noise. Superimposed on this are lines of equal edge displacement at $\delta_c = \{-1/2, 1/2, 1\}$. Based on the similarity around the common edge spread, it is hypothesized that if the characters used for training and testing are chosen from the regions of the parameter space that exhibit similar edge displacements, then the training set will have less variance. Furthermore, due to the decreased variance of the training set data, recognition accuracy of a classifier

trained on this data will be greater than the accuracy of a classifier trained on data taken from regions based on more arbitrary divisions, such as spatial locality. The experiments that follow test this hypothesis.

2. Experiments and Data

The experiment used in this paper compares the performance of an artificial neural network based classifier as the training set is divided in different ways based on parameters of the degradation model. The two-class problem of distinguishing characters c and e was chosen since these two characters are commonly confused by OCR engines and experiments using these characters will thus serve to illustrate the effect of using different training sets (pairs of characters that are easier to distinguish may not show enough difference in performance across the various regions to be conclusive). The characters are 190,000 c 's and 190,000 e 's at 300dpi 12 point, Times Roman font. Noise was added at three different noise levels, $s = \{0, 0.15, 0.21\}$. Some examples are shown in Figure 2a. Synthetic characters were used since the exact effect of the degradation was of most interest. The experiment will test the performance of our training data selection method under the ideal condition where character degradation is due solely to the parameters of the degradation model (without considering, for example, degradations due to poor quality paper, poor lighting, smudges, dust, etc.).

Classification was done using an artificial neural network trained with 1 hidden layer containing 4 hidden nodes. The features used are 16x16 pixel normalized char-

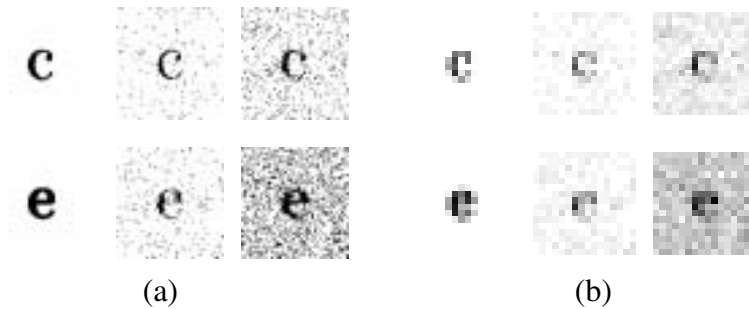


Figure 2: Examples of (a) original characters and (b) character resized for 16x16 grid of input features.

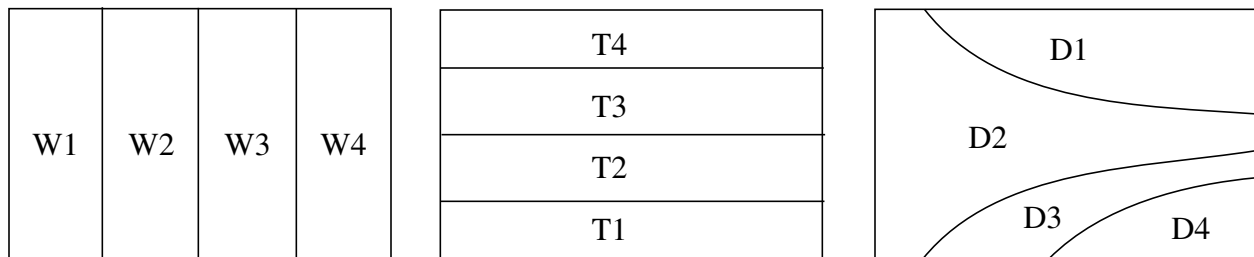


Figure 3: Divisions of degradation space used in this paper.

acters, thus giving 256 inputs to the neural network. The value of each resulting 'pixel' in the normalized character is the average amount of black pixels in that portion of the original character corresponding to each pixel in the normalized window, Figure 2b. The neural network used a single output node, with a high output value indicating the prediction of a c, and a low output indicating an e.

The reported results for each region are averages obtained using 10 fold cross-validation. In order to make the 10 fold cross-validation results comparable, and in order to avoid the problem of overlapping training/test data between each data region's 10 fold splits, the initial division of the data into the 10 fold cross-validation partitions was done with the base data set (the data set that contained all of the samples). From here, each 10 fold split was filtered according to the desired degradation parameter settings to form the region specific training/test set pairs. This guarantees that for any two regions A and B in the degradation space, the intersection of region A's i^{th} training set with B's i^{th} test set is empty, and so the test results of classifiers trained on A and tested on B are valid and comparable across regions.

Experiments were run using this data. Each involved dividing the dataset into 4 zones using three different division methods shown in Figure 3. First the degradation space was divided based on blur width alone, Figure 3a. Second it was divided into 4 regions based on threshold value, Figure 3b. Then finally it was divided into regions

having nearly common edge spread: $\delta_c = \{(-\infty, -0.5], (-0.5, 0.5], (0.5, 1.5], (1.5, \infty)\}$, Figure 3c. The entire degradation space, A13, was also used for comparison with not doing the subdivision. Thirteen different ANN's were trained on representing characters in each of the 12 degradation regions plus one using all the characters with mixed degradation parameters from the entire degradation space.

The results are summarized in Table 1. The middle column shows the results from training on the whole dataset and testing on each partitioned region. The right column shows the results from training on a partitioned region and testing on that same region. As expected, in all but one case (T1) the recognition results were better when the classifier was trained on data from a limited region in the degradation space and tested on data from the same region, versus when trained on data from the whole degradation space and tested in any particular subset of the degradation.

The next question is how the classifier performance changed when the degradation space is divided based on the degradation parameters w and Θ versus when the degradation space is divided according to the edge spread δ_c . Here the raw percentages can not be compared since they are not comparing the performance of like data sets. Instead the total performance of all four classifiers over the combined dataspace is compared. To do this, the fraction of characters in each sub-region was used to weight the performance. Doing this, when the data sets were divided

Table 1: Results comparing classifier trained with broad versus narrow training.

Region of Degradation Space	Training on data from whole degradation space	Training on data from limited part of degradation space
W1	99.3	99.6
W2	95.7	96.5
W3	94.6	94.9
W4	94.1	94.5
T1	99.5	99.5
T2	98.5	98.7
T3	94.1	94.6
T4	89.6	90.8
D1	86.3	87.8
D2	98.0	98.3
D3	99.1	99.4
D4	99.4	99.6
A13	96.2	96.2

by width, the net performance was 96.6% correct, the performance on the data when divided by threshold was 96.6% and when divided on edge spread, the performance was 97.6%. Also when no divisions were used, training on the whole dataset and testing on all the data resulted in 96.2% recognition accuracy. Therefore dividing the data set by edge spread has the greatest benefit to recognition performance.

3. Conclusions and Future Work

The results of our experiments show that, under certain conditions, the use of the δ_c isolines to partition the degradation data space into separate regions for training can improve the overall accuracy of the set of classifiers produced by training on these regions. This improvement in accuracy is likely due to the increased homogeneity of the partitioning that the δ_c isolines produces over other, more arbitrary divisions of the degradation data space.

In future work the authors intend to expand on the results of these experiments to more deeply investigate the benefits of dividing the degradation space. First experiments will be tried with different features to see if the effect is feature independent. While this particular set of features has been used in other experiments and has shown to have comparable recognition performance to other features, it is worthy to question which features are most 'immune' and which react most strongly to the division of the degradation space.

The choice of δ_c boundaries used in these experiments was somewhat arbitrary. Further experiments to see which division is optimal will follow. Also the choice for the number of regions into which the space should be divided needs to be further explored.

The edge spread was defined as the response for isolated edges. The effect of dividing based on edge spread is greatest if the characters are large enough and the strokes are wide enough to allow the original definition of the edge spread to hold. 300dpi 12 point characters are at the limit of this pure assumption. The edge spread function may need to be modified to take into account the nearness of the neighboring edges as in [9].

The characters c and e are one of the more difficult recognition problems and have been used in several experiments to compare classifiers [6, 7]. Still it would be worthy to try the full Roman character set, A-Za-z0-1 etc., to see if these results hold.

Synthetic characters were used in this paper because the exact effects of the degradation model were to be explored. While methods to accurately estimate the degradation parameter from characters are being developed, there are likely other ways to get a broad sense of which characters have similar edge spread degradations in a training set, particularly if the labels are already known. Methods to estimate these features and divide the dataset along these characteristics would likely be more practical and could produce equivalent results.

4. References

- [1] Henry S. Baird, "Document Image Defect Models," *Proc. IAPR Workshop on Syntactic and Structural Pattern Recognition*, Murry Hill, NJ, June 1990, pp. 13-15. Reprinted in H. S. Baird, H. Bunke, and K. Yamamoto (Eds.), *Structured Document Image Analysis*, Springer Verlag: New York, 1992, pp. 546-556.
- [2] Elisa H. Barney Smith and Xiaohui Qiu, "Statistical image differences, degradation features and character distance metrics," *International Journal of Document Analysis and Recognition*, Vol.6, No. 3, 2004, pp. 146-153.
- [3] Elisa H. Barney Smith, "Characterization of Image Degradation Caused by Scanning," *Pattern Recognition Letters*, Vol. 19, No. 13, 1998, pp. 1191-1197.
- [4] Elisa H. Barney Smith, "Estimating Scanning Characteristics from Corners in Bilevel Images," *Proc. SPIE Document Recognition and Retrieval VIII*, Vol. 4307, S.an Jose, CA, 21-26 January 2001, pp.176-183.
- [5] Elisa H. Barney Smith, "Scanner Parameter Estimation Using Bilevel Scans of Star Charts," *Proc. International Conference on Document Analysis and Recognition 2001*, Seattle, WA, 10-13 September 2001, pp. 1164-1168.

- [6] Tin Kam Ho and Henry S. Baird, "Large-Scale Simulation Studies in Image Pattern Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, No. 10, October 1997, pp. 1067-1079.
- [7] Dz-Mou Jung, Mukkai S. Krishnamoorthy, George Nagy and Andrew Shapira, "N-tuple features for OCR revisited," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 18, No. 7, July 1996, pp. 734-745.
- [8] S. La Manna, A.M. Colla, A. Sperduti, "Optical Font Recognition for Mult-Font OCR and Document Processing," 10th International Workshop on Database and Expert Systems Applications, 1-3 September, 1999, Firenze, Italy, pp. 549-553.
- [9] Theo Pavlidis, Minghua Chen, and Eugene Joseph, "Sampling and Quantization of Bilevel Signals," Pattern Recognition Letters, Vol. 14, July 1993, pp. 559-562.
- [10] Prateek Sarkar, George Nagy, Jiangying Zhou, Dan Lopresti, "Spatial sampling of printed patterns," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, No. 3, March 1998, pp. 344-351.
- [11] Sriharsha Veeramachaneni, George Nagy, "Style context with second-order statistics," IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 27, No. 1, January 2005, pp. 14 - 22.