1-16-2006

Partitioning of the Degradation Space for OCR Training

Elisa H. Barney Smith
Boise State University

Tim Andersen
Boise State University

 DOI: 10.1117/12.641229

Partitioning of the degradation space for OCR training

Elisa H. Barney Smith, Tim Andersen
EBarneySmith@boisestate.edu, tim@cs.boisestate.edu
Boise State University
Boise, ID, USA

ABSTRACT

Generally speaking optical character recognition algorithms tend to perform better when presented with homogeneous data. This paper studies a method that is designed to increase the homogeneity of training data, based on an understanding of the types of degradations that occur during the printing and scanning process, and how these degradations affect the homogeneity of the data. While it has been shown that dividing the degradation space by edge spread improves recognition accuracy over dividing the degradation space by threshold or point spread function width alone, the challenge is in deciding how many partitions and at what value of edge spread the divisions should be made. Clustering of different types of character features, fonts, sizes, resolutions and noise levels shows that edge spread is indeed shown to be a strong indicator of the homogeneity of character data clusters.

Keywords: character degradations, OCR training, character homogeneity, clustering.

1. INTRODUCTION

It is often possible to improve the recognition accuracy for general, non-homogeneous recognition problems by subdividing the problem into several smaller, more homogeneous problems. This approach has been used extensively to achieve improved performance in the field of optical character recognition (OCR). For example, [2, 13] used font as a basis for subdividing the data for OCR, and the problem was divided by writing style in [16, 17]. This paper looks at using the level and type of degradation present in the character images as the basis for dividing the recognition problem into smaller homogeneous sub-problems.

In this paper we use an automated clustering algorithm to verify that the similarity between characters seen with degradation model parameters matches the clusters that result when clustering the data with standard OCR features. Clustering algorithms have been used in a number of ways as part of the document understanding process. For example, clustering has been used in conjunction with cryptoanalysis to achieve high-accuracy OCR in the presence of previously unseen character fonts [8, 9, 11], it has been used as a means to perform general document classification [10] and automatic text classification [12], and also used directly as a mechanism to perform OCR [7].

The degradations considered are those that result from scanning. This is viewed in context of the degradation model based on the model by Baird [1]. Here character templates are blurred (2-D convolution) by a point spread function (PSF) that is an isotropic bivariate Gaussian with width, w, defined by the standard deviation. This width is measured in number of pixels, not in inches or mm. This blurred character is then globally thresholded at a threshold Θ to form the bilevel character. An edge, when degraded, re-forms in a location parallel but displaced from the original. The edge displacement was defined in [3] to be

$$\delta_c = -w\, \mathrm{ESF}^{-1}(\Theta)\,, \tag{1}$$

where the ESF is the edge spread function. Examples of different values of the edge spread function corresponding to degradation with a Gaussian PSF are shown in Figure 1.

In [6] the recognition problem was divided into sub-problems by parameters of this degradation space. Division by PSF width, by threshold level, and by edge spread were all explored. Division by PSF width and threshold gave recognition rate of 96.6%. This was an improvement from the 96.2% recognition rate that resulted when no divisions were made. When the division was made based on edge spread, the recognition rates of the problem improved even more to 97.6%. These improved recognition rates are consistent with observations that characters that had the same edge spread both appeared similar [4] and were statistically similar [5]. What remained un-clear was how to choose the values of the edge degradation with which to divide the degradation space. This paper explores choosing the values of the edge spread at

Document Recognition and Retrieval XIII, edited by Kazem Taghva, Xiaofan Lin,
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 6067, 606705, © 2006 SPIE-IS&T · 0277-786X/06/$15

which to divide the degradation space.

To use this knowledge of degradations to improve recognition, the best partition of the degradation space is needed. This requires evaluation of the characters and making decisions of how many regions into which the character set should be divided. The changes inherent to the degradation of blur (point spread function) and threshold are of most interest but the effect of additive noise also matters. To help guide future OCR experiments and to better understand how character similarity corresponds to divisions based on the ESF, clustering is employed to view how the characters cluster based on features over a range of degradations. The effect on the divisions of the degradation space when changing aspects of the clustering, including the number of clusters, is also explored.

2. CLUSTERING EXPERIMENTS

It was assumed that characters that are similar should group together when grouped by clustering. Here the c-means clustering algorithm is used to divide the degradation space. The base experiment uses 300 dpi, sans serif, 12 pt letter 'c' degraded over a threshold range of [0.05, 0.95] and a PSF width range of [0.5, 3.0]. Examples of the characters that result from the degradation model at various points in the degradation space are shown in Figure 2a. Each experiment uses 1000 characters randomly and uniformly distributed over the degradation space. As the base experiment, eight RMS normalized central moment features (m00, m02, m11, m20, m20, m30, m12, m21, m30) are calculated for each character and clustered with the c-means clustering algorithm. Characters with no black pixels are excluded, but highly degraded characters with even a single pixel remaining were considered. These results will be used to guide OCR experiments. We explore how the degradation space divides when

(1) the number of clusters is changed,

(2) the resolution or character size is changed (300dpi and 600dpi),

(3) noise is included in the model,

(4) the font is changed between sans-serif and serif,

(5) the character is varied (c, e, o, x, w), or

(6) non-moment features or different number of moment features are used.

A discussion of each of these cases and their results follows.

2.1 Number of Clusters

The characters were clustered with the c-means clustering algorithm. Figure 4 shows the divisions that result for 300 dpi sans serif 12 pt 'c' without noise at each of C={2, 3, 4, 5} clusters. In each case, the clusters do have boundaries approximately along the edge spread boundaries. Because the character strokes are significantly smaller than the width of the PSF for most of the degradation space, the assumption that the edges are affected by the PSF independently is violated and the boundaries have a slope that is more negative than the theoretical ESF lines.

Three cluster validity metrics were used to determine the best number of clusters to see if a particular number of clusters was valid [15]. Each of the three cluster validation metrics preferred a different number of clusters. To visualize how the features grouped, the features were projected into two dimensions using principle components, Figure 3. As expected, there is more of a continuum of features within the feature space than a distinct natural clumping, so determining the best number of clusters will need to rely on results from future OCR experiments.

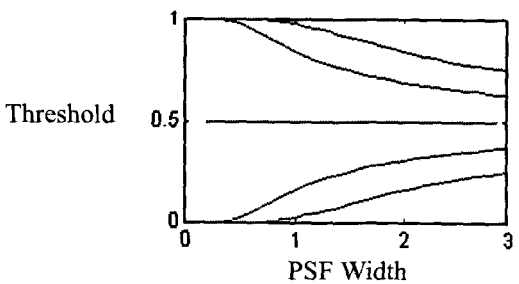


Figure 1: Loci of constant edge spread. δ_c =[-2 -1 0 1 2] (from top to bottom) for Gaussian PSF function.

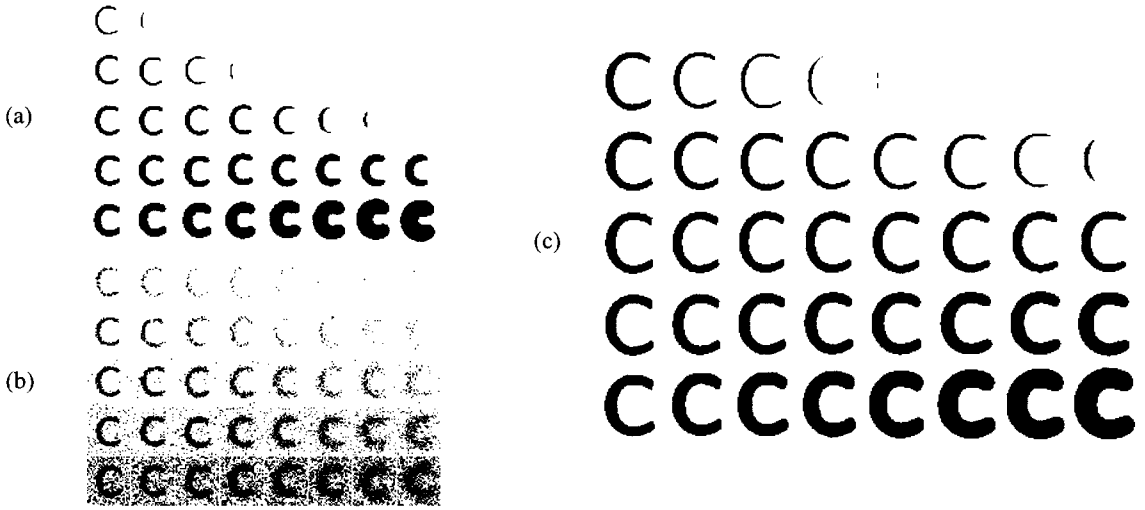


Figure 2: Characters generated in the degradation space for 12 point sans serif 'c' at (a) 300 dpi no noise, (b) 300dpi serif, and (c) 600 dpi no noise. For each, the PSF width ranges [0.5, 3.0] and the threshold [0.05, 0.95].

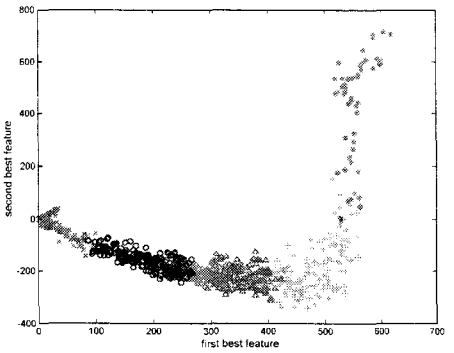


Figure 3: Projection of features into 2 best dimensions. The divisions resulting from the clustering into C=5 clusters are shown by different marker symbols.

2.2 Resolution or Character Size

Since changes in scanning resolution and character size have the same net effect on characters (with the exception of small hinting effects), only resolution was experimented with here. Two different resolutions were considered: 300dpi and 600dpi. Examples of 300 and 600 dpi characters are shown in Figures 2a and c. Figure 4 shows the 300 dpi cluster results and Figure 5 shows the 600 dpi cluster results. The higher resolution characters are essentially twice as large in each direction, and thus have thicker strokes (greater number of pixels). The thicker strokes require much larger point spread functions and higher thresholds to fully erode, therefore less of the degradation space is empty. The region in the degradation space that formerly was empty then occupies part of a cluster. This results in the centroids of the cluster moving upward and thus the degradation boundaries "rising" to regions of higher thresholds, or to be more precise, more negative edge spread. Also the thicker strokes mean that the assumption that the strokes are affected by the edge spread independently is more valid so the cluster boundaries now have a greater resemblance to the theoretical edge spread lines predicted in [5].

2.3 Font

The sans-serif font was used as a baseline in this series of experiments because most characters have fairly constant stroke widths. However, the fonts with serifs are more commonly used in document analysis. Here the strokes vary in width, and are often quite narrow in places, and the serifs themselves are also quite narrow. It is expected that scanning degradations will affect these characters more than the sans-serif fonts. Figure 6 shows the clusters that result for C={3, 5}. For C=2 clusters (results not shown) the results are similar to the results for the sans-serif case in Figure 4. When C=3, the effect of large PSF width and low thresholds causes a greater negative slope in the cluster boundaries than was seen in Figure 4.

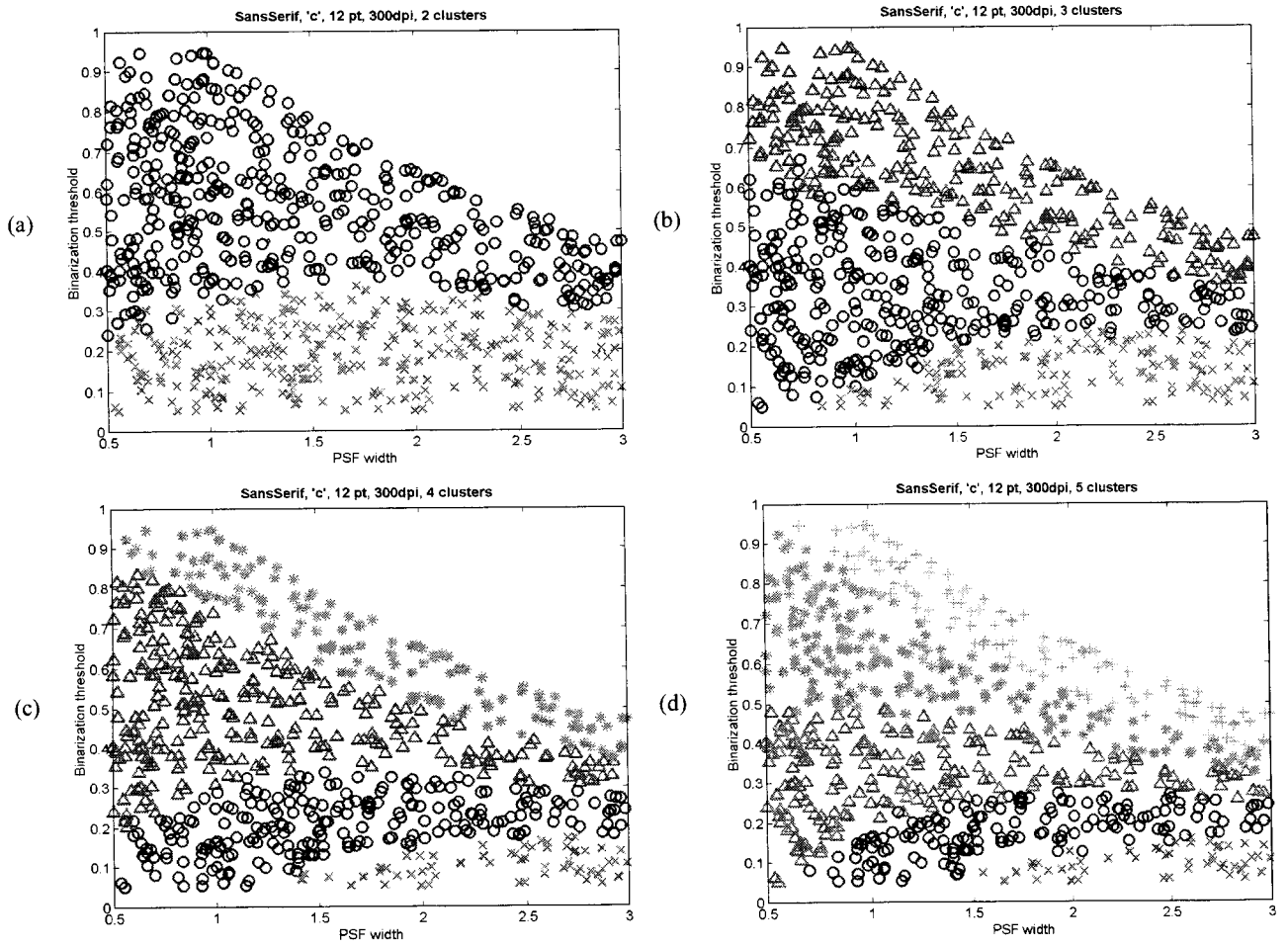


Figure 4: Clustering results for SansSerif 12pt 'c' at 300dpi with (a) 2 clusters, (b) 3 clusters (c) 4 clusters and (d) 5 clusters. Results are for 8 moment features, no noise.

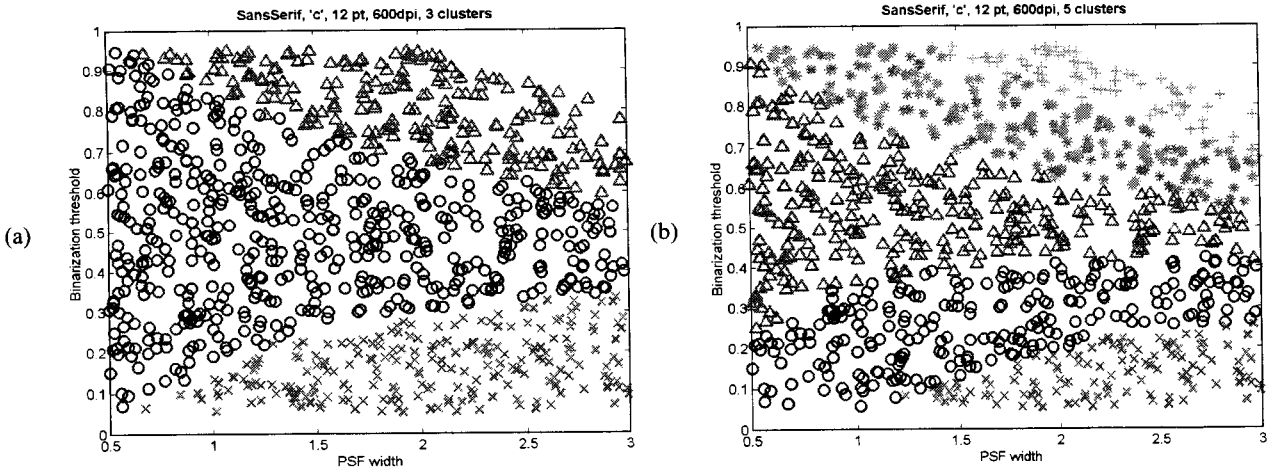
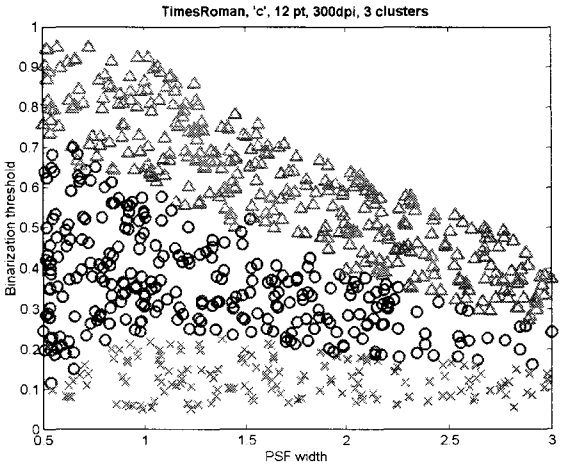


Figure 5: Clustering results for SansSerif 12pt 'c' at 600dpi with (a) 3 clusters (b) 5 clusters. Results are for 8 moment features, no noise.

(a)



(b)

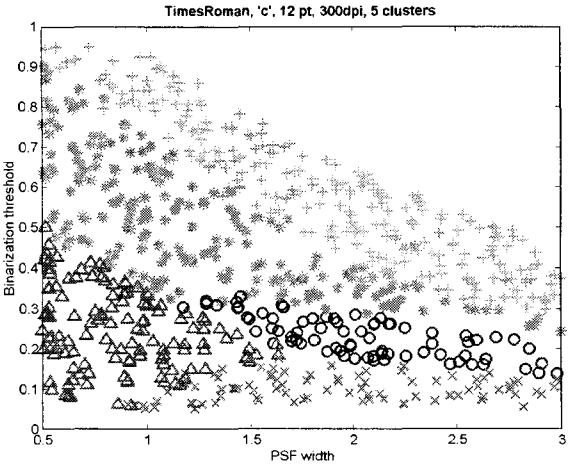


Figure 6: Clustering results for Times Roman 12pt 'c' at 300dpi with (a) 3 clusters (b) 5 clusters. Results are for 8 moment features, no noise.

For C=4 and C=5, the difference in characters for small PSF versus large PSF is greater due to the closeness of the edges so the clusters prefer to divide at an intermediate PSF width value instead of dividing solely on adjusted ESF boundaries.

2.4 Character

The previous experiments have looked solely at the character c. The experiments were rerun with the characters e, o, x, and w. The e and o have similar structure to the c, and are likely to be confused with the c during OCR. The x and w were chosen because, unlike the c and e, they contain straight lines and have sharp corners. The results for the w with C={2, 5}are shown in Figure 7. The results for the e, o and x are not shown. The clusters for the e and o are very similar to the clusters for the c. The top most boundary is a little higher because the e and o have more black area in their original character template than the c. This results in most of the other cluster boundaries shifting slightly higher, otherwise the structure of the cluster boundaries is the same. The cluster boundaries for the x and the w are quite similar to the cluster boundaries for the c and the e even given the significant structural differences in the underlying characters. The largest difference is in the C=2 case where the similarity for small PSF widths is stronger for the x and w, causing the cluster boundary to encompass a larger portion of the PSF width = 0.5 region.

2.5 Features

The experiments to this point have used 8 moment features to determine the cluster boundaries. Since OCR packages rely on a variety of features, the features used were changed to see the effect that alternate features would have on the clusters. The NxN normalized pixel features were used in [6] with N=16 to show how dividing by degradation features improves recognition and in [14] 16x16 normalized pixel map features were used with synthetic data from the same model to illustrate the effects of style improving OCR performance. The results, Figure 8a, were virtually identical to the results for the 8 moment features case shown in Figure 4. Another experiment looked at only using the first 4 of the 8 moment features, Figure 8b, and here again, great similarity to the 8 moment feature case in Figure 4b can be seen.

3. CONCLUSIONS AND FUTURE WORK

Several experiments that examine how characters in the degradation space should be clustered have been completed. The effects of changing a number of variables have been explored. In general the partitioning of the boundaries based on edge spread holds, however for small and serifed fonts some additional compensation for the narrowness of the strokes should be incorporated. For these cases, the clustering algorithm occasionally subdivides an existing cluster, and sometimes it appears to completely move the boundaries of several clusters (this is not a hierarchical clustering algorithm). However, each case shows rather distinct boundaries that all appear to follow the edge spread iso-lines.

Since the degradations occur continuously, and where one degradation region ends there are no major changes where the next one begins, there is no natural number of clusters into which the degradation space should be divided. For future work, some sort of gradual change from membership in one cluster to another will be considered, for example utilizing a fuzzy clustering framework.

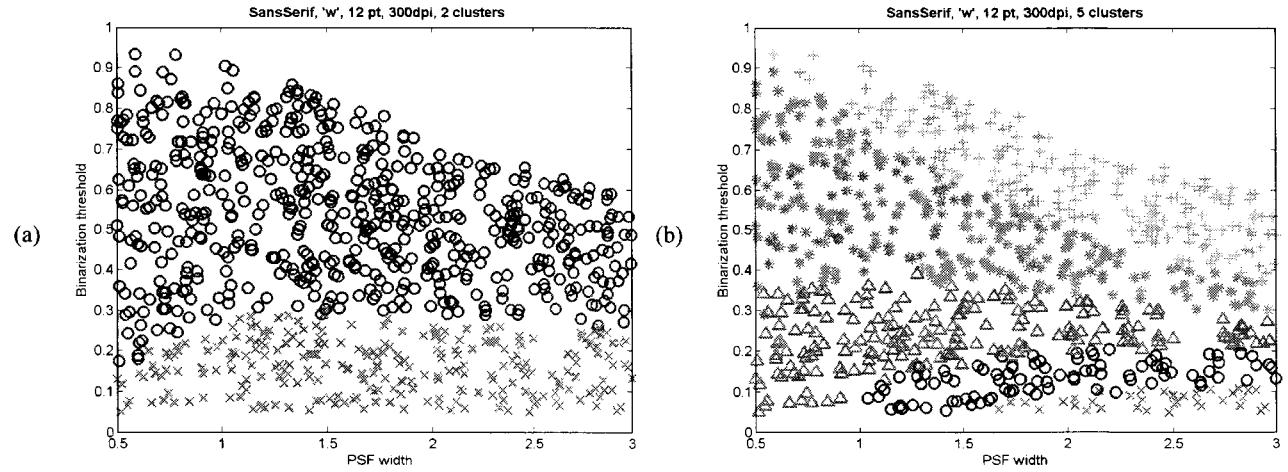


Figure 7: Clustering results for SansSerif 12pt 'w' at 300dpi with (a) 2 clusters (b) 5 clusters. Results are for 8 moment features, no noise.

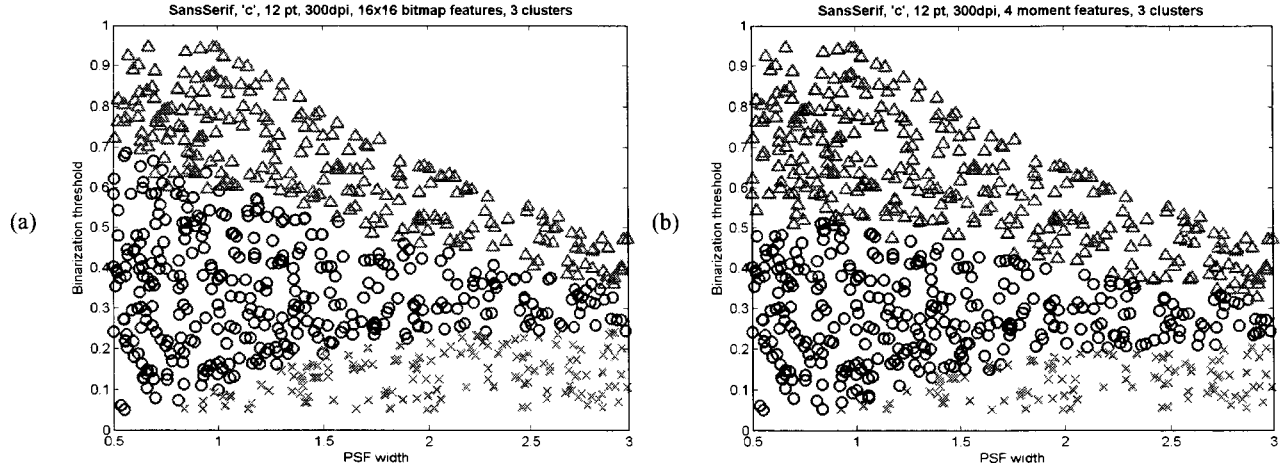


Figure 8: Clustering results for SansSerif 12pt 'c' at 300dpi with 3 clusters (a) for 16x16 normalized pixel map and (b) the first 4 moment features, no noise.

In addition to the results reported in this paper, we also ran experiments that included noise. The inclusion of noise does not have a large effect on the location of the boundaries, but it does reduce the number of characters that are totally blank (no black pixels). This then moves the boundary higher. Increasing the size of the characters to 600dpi makes the characters more closely follow the model assumptions that the edges when being blurred are independent. Therefore the cluster boundaries more closely follow theoretical constant edge spread lines.

Fonts with serifs cause the PSF width to have a larger effect in the degradation. The 300 point serif 'c' was the only case where some cluster boundaries were based more on PSF width than on edge spread alone, however this was only with low thresholds, and the edge spread was still present in the other boundaries.

The shape of the clusters was predominantly independent of the characters being examined. Most significant was the amount of area contained in the character template, since larger areas lead to less empty space which has the greatest effect on cluster boundaries. Clusters for rounded characters like c, e and o were more similar to each other than clusters for sharp clusters such as x and w, however all shared significant similarities.

The choice between the NxN normalized window features and the moment features appeared to have very little effect on the shape of the clusters. While we plan on conducting more experiments to confirm that this result will hold in general, it is expected that this will hold for a variety of OCR features (or at least those OCR features that capture relevant information about the shape and look of the characters). This means that one can use features to cluster the data into homogeneous

training sets which are independent of the features used by the OCR engine used to classify the data, making it possible to produce homogeneous training sets that are tuned to off-the-shelf OCR engines where the features used are unknown.

In conclusion, these experiments show that the edge spread function is a strong indicator of character similarity regardless of the types of OCR features being considered. Thus, the use of the edge spread function as a mechanism for subdividing OCR character data, both during training and during execution, is a natural way to homogenize the data in order to improve OCR accuracy. In addition to this, the edge spread function could be used in the following areas:

- Feedback to auto-adjust scanner parameters for OCR or other purposes.
- Generation of realistic artificial data sets for OCR engine training.
- Identification of the printer/scanner that produced a document - forensic science, document origin verification.
- Improved scanner hardware/printer hardware.

In the future the c's and e's should be mixed together to see if they have some of the same features that are common to a degradation. This will lead to the ability to find features that can help divide a standard character set to help OCR on real character problems. When c's and e's are projected into the best feature space they were separate in pairs, except at the one and two pixel remaining cases where the clusters merged. Using the clusters to divide the training sets should allow greater discrimination where the characters are less eroded, and thus increase OCR results.

4. ACKNOWLEDGEMENT

This material is based upon work supported by the National Science Foundation under Grant No. CCR-0238285.

5. 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. Henry S. Baird and George Nagy, "A Self-Correcting 100-font Classifier," Proc. SPIE Document Recognition, Vol. 2181, San Jose, CA, March 1994, pp. 106-115.
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, "Uniqueness of Bilevel Image Degradations," Proc. SPIE Document Recognition and Retrieval IX, Vol. 4670, San Jose, CA, January 2002, pp. 174-180.
5. 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.
6. Elisa H. Barney Smith and Tim Andersen, "Text Degradations and OCR Training," International Conference on Document Analysis and Recognition 2005, Seoul, Korea, August 2005.
7. Thomas M. Breuel, "Modeling the sample distribution for clustering OCR," In Proceedings of IS&T/SPIE Electronic Imaging 2001: Document Recognition and Retrieval VIII, January 2001, pp 193-200.
8. Thomas M. Breuel, "Classification by Probabilistic Clustering," 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (ICASSP 2001), 2001 pp. 1333-1336.
9. Thomas M. Breuel, "Character Recognition by Adaptive Statistical Similarity," Seventh International Conference on Document Analysis and Recognition, ICDAR 2003, vol 1. 2003, pp. 158-162.
10. Eui-Hong (Sam) Han and George Karypis, "Centroid-Based Document Classification: Analysis & Experimental Results" Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery 2000, pp. 424 - 431.
11. Tin Kam Ho, George Nagy, "OCR with No Shape Training," Proceedings of the 15th International Conference on Pattern Recognition, March 2000, vol. 4, pp 27-30.
12. Iwayama Makato & Tokunaga Takenobu, "Cluster-Based Text Categorization" A Comparison of Category Search Strategies," Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle, Washington, USA, July 1995, pp. 273-280.
13. 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.

14. Charles Mathis & Thomas Breuel, "Classification using a Hierarchical Bayesian Approach," In Proceedings of the International Conference on Pattern Recognition (ICPR'02), Quebec City, Quebec, Canada, 2002.
15. Ujjwal Maulik and Sanghamitra Bandyopadhyay, "Performance Evaluation of Some Clustering Algorithms and Validity Indices," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 12, December 2002, pp. 1650-1654.
16. Prateek Sarkar and George Nagy, "Style consistent classification of isogenous patterns", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 1, January 2005, pp. 88 - 98.
17. Sriharsha Veeramachaneni, George Nagy, "Style context with second-order statistics," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 1, January 2005, pp. 14 - 22.