Interactive degraded document enhancement and ground truth generation

G. Bal[a], G. Agam[a], O. Frieder[a], G. Frieder[b]

[a]Illinois Institute of Technology, Chicago, IL 60616
[b]The George Washington University, Washington, DC 20052

ABSTRACT

Degraded documents are frequently obtained in various situations. Examples of degraded document collections include historical document depositories, document obtained in legal and security investigations, and legal and medical archives. Degraded document images are hard to to read and are hard to analyze using computerized techniques. There is hence a need for systems that are capable of enhancing such images. We describe a language-independent semi-automated system for enhancing degraded document images that is capable of exploiting inter- and intra-document coherence. The system is capable of processing document images with high levels of degradations and can be used for ground truthing of degraded document images. Ground truthing of degraded document images is extremely important in several aspects: it enables quantitative performance measurements of enhancement systems and facilitates model estimation that can be used to improve performance. Performance evaluation is provided using the historical Frieder diaries collection.[1]

Keywords: degraded documents, image enhancement, historical documents, document image analysis, document degradation models, image analysis

1. INTRODUCTION

Degraded documents are archived and preserved in large quantities worldwide. Electronic scanning is a common approach in handling such documents in a manner which facilitates public access to them. Such document images are often hard to read, have low contrast, and are corrupted by various artifacts. Thus, given an image of a faded, washed out, damaged, crumpled or otherwise difficult to read document, one with mixed handwriting, typed or printed material, with possible pictures, tables or diagrams, it is necessary to enhance its readability and comprehensibility. Documents might have multiple languages in a single page and contain both handwritten and machine printed text. Machine printed text might have been produced using various technologies with variable quality.

The approach described herein is concerned with semi-automatic enhancement of such documents and is based on several steps: the image foreground is separated from the background, the foreground image is enhanced, the original image is enhanced, and the two enhanced images are blended using a linear blending scheme. The use of the original image in addition to the foreground channel allows for foreground enhancement while preserving qualities of the original image. In addition, it allows for compensation for errors that can occur in the foreground separation. The foreground-background separation component of our system is based on a probabilistic model estimated through expectation maximization (EM). Multiple probabilistic models are estimated at different neighborhoods and at multiple resolutions. Parameter knowledge at specific locations and resolutions is used to initiate parameter values at neighboring locations. Complete details of our foreground-background separation component may be found in Reference.[2] The overall system architecture is depicted in Figure 1.

The focus of this paper is foreground enhancement based on inter- and intra-document coherence, where in this process characters are segmented, clustered, and matched. The proposed approach is based on several key observations as follows: Global foreground document-level enhancement is inferior to local character-level enhancement which is capable of adapting to local degradation models; Local character-level coherence can be exploited to improve severely degraded characters with missing parts which can not be inferred directly from a local neighborhood; The automated enhancement of severely degraded documents may assisted by human intervention and so there is a need to facilitate such intervention beyond simple pixel-level editing tools.

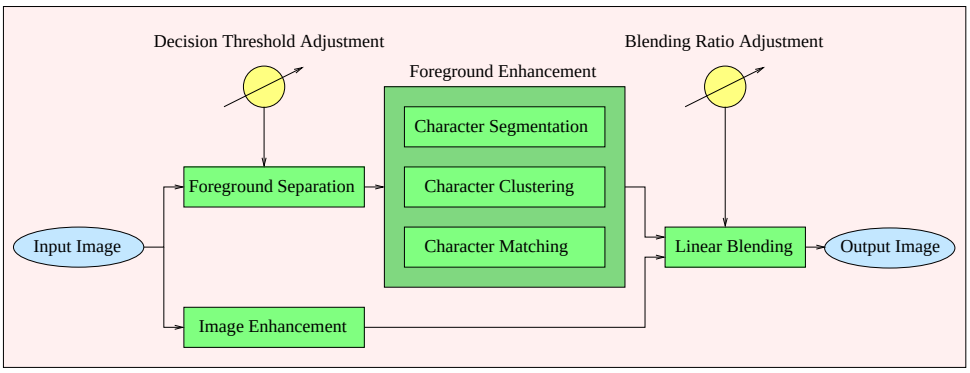

Figure 1: Description of the document enhancement system architecture (see text). After initial processing, the user can use two adjustable thresholds to control both the foreground separation decision threshold and the blending level.

The system we have developed supports automated enhancement and with semi-automated editing can produce high quality results. Such high quality results can be used as ground truth data for the corresponding degraded document image. While ground truthed data may be generated synthetically from good quality document images, ground truthing of actual degraded document images can be used to learn actual degradation models which can be used to develop degradation-specific enhancement techniques. In addition, actual-data ground truthing is crucial to performance evaluation of enhancement techniques.

2. RELATED WORK

Document image enhancement has been studied mostly from the point of view of image segmentation. That is, given a degraded document image, segment text and graphics in it so that the binarized image can be processed more successfully by various document image understanding algorithms. A thorough review of segmentation techniques for grayscale document images is provided in.[3,4] The comparison is based on OCR accuracy rate after segmentation. Using this metric, it is concluded that local segmentation techniques have higher performance. In particular, a method which is based on a local minimum and maximum values in each sub-window[5] is shown to be one of the most efficient and effective techniques. Approaches for text segmentation in color images are described in by Perroud et al.[6] and Loo and Tan.[7] A component based approach for foreground-background separation in low-quality color document images is described by Garain et al..[8,9] In this approach, connected components are labeled and organized in tree structures. Nodes in the tree are then segmented using K-means. The performance of this approach is evaluated by measuring the improvement in word and line segmentation algorithms before and after enhancement.

Binarization of historical documents based on adaptive threshold segmentation and various pre- and post-processing steps is described by Gatos et al..[10] In this approach, a background surface is estimated and used to segment the image. An iterative approach for segmenting degraded document images is described by Kavallieratou et al..[11] There, a global thresholding technique is used to obtain an initial segmentation. Areas with likely incorrect segmentation are detected, and a local thresholding is applied in them. This approach is efficient in that local thresholds are computed only at selected locations. It is also noted that general-purpose segmentation techniques provided better performance on historical documents as compared with document-specific segmentation techniques.

An improved binarization for historical documents is described by Bar-Yosef et al..[12] The process starts by thresholding the image and detecting connected components. The detected components are used as a seed image. The distance between the seed image and a thresholded image is used to distinguish between good quality and poor quality characters. A region growing process is then used to correct poor quality characters. The region growing process is an iterative process in which the distance of candidate pixels from the foreground and background is used. A method for binarization of historical documents that is based on local operations in

detected character boxes is described in.[13–15] An approach for the enhancement of low-resolution, binarized, faxed documents is described by Hobby and Ho.[16] In this approach, multiple low-resolution characters are used to estimate high resolution ones. When a global degradation model such as a geometric deformation model or an illumination model is known, degraded documents may be enhanced by canceling out the degradation. A thorough discussion and evaluation of document degradation models is provided by Kanungo et al..[17] Photometric correction in camera captured documents which is based on geometric structure and an illumination model is discussed in.[18]

To evaluate the performance of document enhancement systems, ground truth data is necessary. Existing work on ground truth generation focus on synthetic data generation.[19] Synthetic ground truth generation has the advantage of being able to produce large quantities of ground truthed degraded document images. However, synthetically generated degraded document images requires the knowledge of the degradation model and so may have difficulties in producing degraded data corresponding to actual document degradations.

3. THE PROPOSED APPROACH

The proposed approach for document image enhancement is composed of several steps including foreground segmentation, foreground enhancement, image enhancement, and linear blending (see Figure 1). A description of the foreground-background separation component of our system can be found in Reference.[2] Once the foreground and background are separated, the foreground image is binarized and enhanced. The foreground enhancement step is based on inter- and intra-document coherence, where in this process characters are segmented, clustered, and matched. The basic approach is to segment characters and cluster them in groups based on some similarity metric. Clusters of similar characters can then be used to produce better quality characters which can replace degraded instances. The advantage of character clusters is the ability to produce maximum likelihood estimates of character data from degraded character images. This in turn can be used to correct severe degradations which cannot be corrected by simply looking at local neighborhoods. The enhanced foreground image is blended with the original image (after enhancement) to produce the final enhancement result.

Character segmentation

Given a segmented foreground image, the purpose of the character step is to identify bounding boxes of characters that can then be processed individually. The first step in this process is concerned with character segmentation. Character segmentation is performed by identifying bounding boxes where characters may reside. It should be noted that the process of extracting character boxes in degraded document images requires special attention to be able to cope with noise inherent to such images.

The character segmentation process begins by detecting gaps between text lines in the image using a vertical projection histogram of the foreground image (cf.[20]) and using the gaps to determine the baseline of each text line. The average distance between text lines is used to form an estimate of average character height and width. Given the extracted character baselines, connected components in the segmented foreground image are identified along each text line and a bounding box is computed for each character candidate. A histogram of character height is used to estimate the parameters of a two component Gaussian mixture distribution of character heights. Similarly, a histogram of character width is used to estimate the parameters of a two component Gaussian mixture distribution of character widths. These mixtures are used in conjunction with the baseline-based width/height estimate to identify character box outliers. Finally, merge process is employed in an attempt to combine small outlier boxes, and a split process is employed in an attempt to split large outlier boxes. An example of the character segmentation results obtained by following the above process is presented in Figure 2 where green boxes indicate identified character boxes, red boxes indicate outliers, and blue boxes indicate identified word gaps.

Character Clustering

Given a segmented foreground image with detected character bounding boxes, the purpose of the character clustering step is to identify groups of characters that are likely to originate from the same character model. To perform the clustering there is a need to define a metric that can quantify shape similarity. Generally speaking, shape matching techniques can be classified[21] into feature-based and intensity-based techniques. Intensity-based

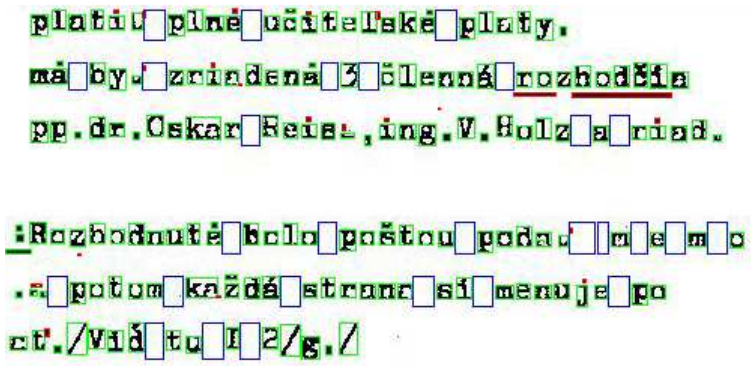

Figure 2: Example of character box extraction using the proposed approach.

techniques often employ some global transformation to obtain a translation/rotation/scale invariant representation. The metric necessary for character clustering needs to be efficient and can be simplified under assumptions of a linear transformation between characters. Specifically, we assume that the transformation between characters is limited to translation without rotation or scale.

A particularly successful shape matching technique is the shape context approach of Belongie et al..[22] This approach is based on the characterization of shape points using a 2D histogram of pixel distributions as a function of distance and angle. Given two shapes, the matching of points between them is treated as a linear assignment problem in which the points in one shape are matched to the points in the second shape based on the shape context distance between points. As the linear assignment will most likely produce errors, an iterative solution is employed in which the shape context and the assignment between points is refined. This refinement is done by using the initial correspondence solution to warp one of the shapes using thin-plate splines, recomputing the shape context descriptors of points in the warped shape, and repeating the assignment and warping steps. Due to the performance requirements and the constraint on the set of admitted transformation group, our implementation of the shape context algorithm is restricted to a single iteration. Furthermore, we have evaluated the shape context distance when applied to multiple locations in each character and when applied to the center of mass alone.

Perhaps the simplest possible metric for shape similarity is direct shape correlation where the Euclidean distance between shapes is used. The shapes are aligned based on their center of mass. Direct shape correlation is sensitive to incorrect alignment. Consequently, an extension of this metric involves a search in a local neighborhood for a translation that would minimize the Euclidean distance between shapes. The size of the local neighborhood that needs to be searched is a parameter that needs to be set where larger search regions result in slower performance. Performance evaluation of the shape context metric with single/ multiple points and the correlation metric with different parameters is presented in Section 4.

Using the different metrics, the characters in each document are clustered. The characters in each cluster are then combined to produce a maximum likelihood estimate of the original character model. The maximum likelihood estimate is produced using the distribution of black and white pixels at each pixel location. In severely degraded documents user intervention can be employed to correct the automated clustering results by merging and splitting clusters as necessary. The quality of the produced character model depends on the number of instances that are used to generate it. Hence, to increase the number of instances in each cluster, clusters of multiple documents are merged and used to reproduce the character model. This is performed by first producing character clusters for each document in the collection and then clustering the character cluster sets of the various documents.

Character matching

Given a document with character boxes and character models produced as described above, character models can be matched to individual characters in the document using the same metric. Matched character models

can then replace the degraded instances to which they correspond. The alignment of a character model with a degraded character instance is performed by aligning their center of mass and using the correlation metric with shift to find the minimum distance correspondence.

The process of matching character models to degraded character instances is somewhat simpler than OCR since character models are document specific. It is well known that OCR performance is adversely affected by document degradation. Similarly, the more severe the degradation is the less likely it is that a character model will match a degraded instance. It is, therefore, not possible to assume that an automated process will successfully enhance the complete document and so user intervention may be required. User intervention is of particular importance for the purpose of ground truth generation. User correction of the automated results is performed by letting the user add and remove character model assignments to degraded character instances. When assigning a character model to a degraded character instance by user interaction, the character model is aligned with the location specified by the user and a search in a local neighborhood is performed. The local search is used to determine a position refinement that would minimize the Euclidean distance between the model and the instance.

4. EXPERIMENTAL RESULTS

The performance of different metrics for clustering was evaluated quantitatively using various degradation models on a synthetic image with known character models. The degradation was produced using a combination of several degradation processes modeled after degradations observed in actual documents.[2] These processes include local-brightness degradation, blurring degradation, noise degradation, and texture-blending degradation. The local-brightness degradation simulates effects such as uneven key pressure in typewriter produced documents, or faded ink in handwritten documents. This degradation was produced by randomly selecting rectangular windows in the image and increasing their brightness by adding a constant brightness and clamping the obtained intensity. The blurring degradation simulates effects such as fading or writing with imprecise writing instruments, and was produced by convolving the image with a Gaussian. The noise degradation simulates effects such as imperfect typing and dirt, and was produced by randomly flipping the values of pixels in the image. Finally, the texture-blending degradation simulates effects such as textured paper or stained paper, and was produced by linearly blending the document with a texture image.

A sequence of increasingly degraded images was generated by varying the degradation model parameters. In changing the degradation model parameters we made sure that the distance between the known truth and the generated image was monotonically increasing and used this distance as a quantitative measure of the degradation level. The error between the known ground truth clusters and the automatically obtained clusters was measured. The automatically generated clusters were generated using different metrics. The metrics evaluated include: shape context distance at a single point (the center of mass); shape context distance at multiple points; correlation distance without any shift; correlation distance with a shift in a 3×3 neighborhood; and correlation distance shift in a 5×5 neighborhood. The results of this evaluation are presented in Figure 3. As can be observed, the correlation metric resulted in a smaller error rate compared with the shape context metric. The 5×5 shift neighborhood performed better than smaller neighborhoods. The improved performance of the correlation metric compared with the shape context metric can be explained by the fact that the correlation is measured at all locations whereas the shape context is computed only at selected locations. Qualitative results for the enhancement of a degraded document image using the proposed approach is shown in Figure 4. Performance as a function of degradation level is shown in Figure 5. As can be observed, the proposed approach is capable of handling increasing levels of degradations. Error histograms before and after enhancement are shown in Figure 6. As can be observed, the error histogram after enhancement is narrower and even for high levels of degradation.

5. CONCLUSION

We proposed a novel approach to enhance the quality of degraded document with human interaction. The system perform the foreground separation, character clustering and suggest initial labeling automatically and the let user perform labeling on bad characters manually. In this process we get an output which is best that we can get from current image or we can get ever better output using the ASCII map from other images of same resolution and

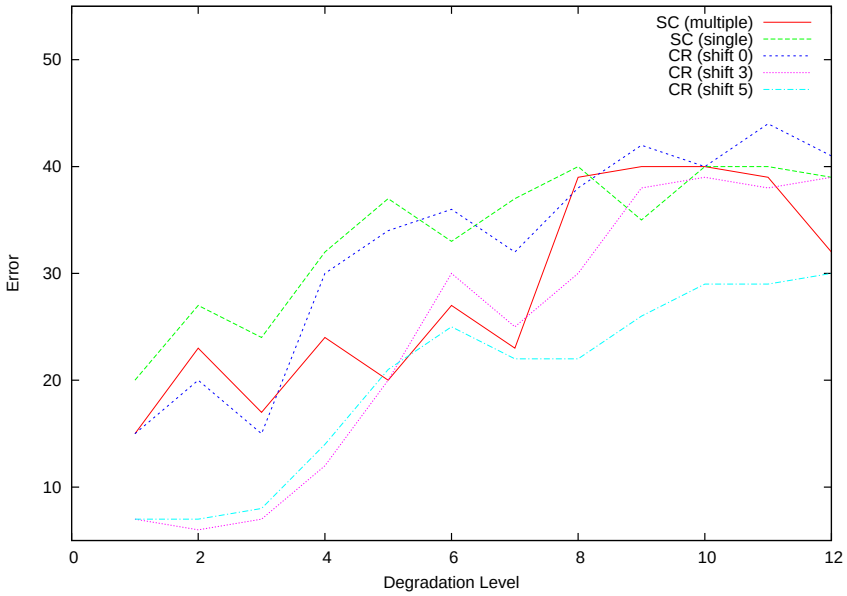

Figure 3: Evaluation of different metrics for clustering: shape context distance (SC) at a single and multiple points, and correlation distance (CR) using a zero shift, a 3×3 shift, and a 5×5 shift.

Figure 4: Results using the ground truth labeling system. (a) Original Image, (b) Foreground separation produce by using MinMax segmentation, (c) Binary Ground truth labeled Image, (d) Final Labeled Image produced by blending original image and binary ground truth labeled image with blending coefficient = 0.5.

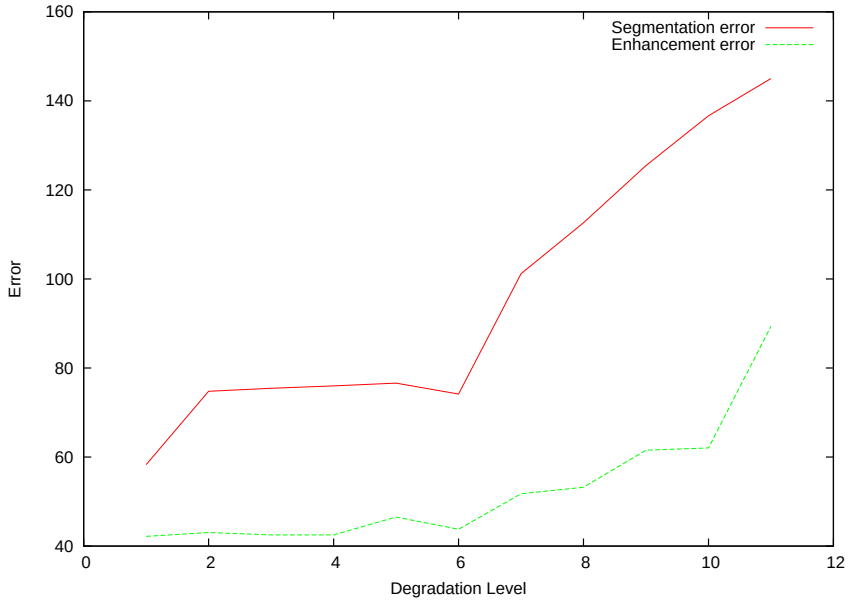

Figure 5: Performance as a function of degradation level.

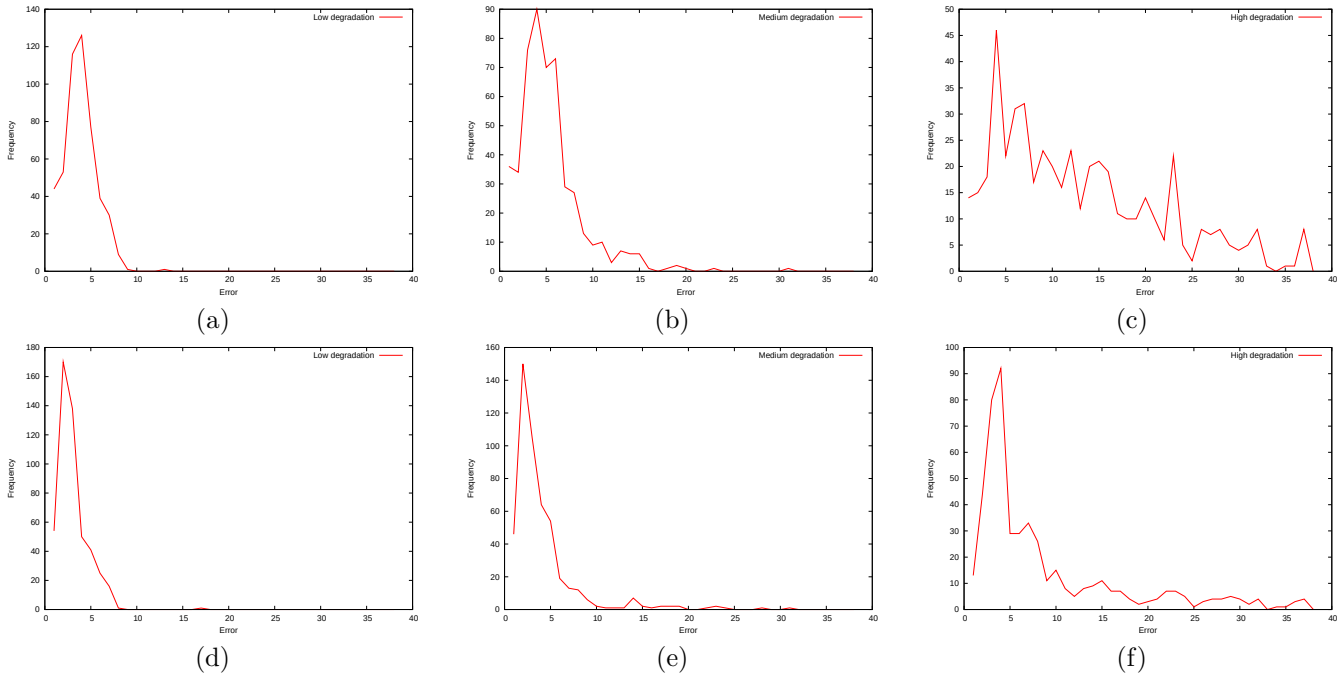

Figure 6: Comparison of Error Histogram. (a)-(c) Error Histogram before using interactive enhancement system (low/medium/high degradation), (d)-(f) Error Histogram after using interactive enhancement system (low/medium/high degradation)

font. Future work will address the incorporation of additional features, development of quantitative perceptual quality measures, and development of methods for selecting particular enhancement models for individual cases.

REFERENCES

1. "The diaries of Rabbi Dr. Avraham Abba Frieder." http://ir.iit.edu/collections/.
2. G. Agam, G. Bal, G. Frieder, and O. Frieder, "Degraded document image enhancement," in *Document Recognition and Retrieval XIV*, X. Lin and B. A. Yanikoglu, eds., *Proc. SPIE* **6500**, pp. 65000C–1 – 65000C–11, 2007.
3. O. Trier and T. Taxt, "Evaluation of binarization methods for document images," *IEEE Trans. Pattern Analysis and Machine Intelligence* **17**(3), pp. 312–315, 1995.
4. O. Trier and A. Jain, "Goal-directed evaluation of binarization methods," *IEEE Trans. Pattern Analysis and Machine Intelligence* **17**(12), pp. 1191–1201, 1995.
5. J. Bernsen, "Dynamic thresholding of gray-level images," in *Proc. Int'l Conf. Pattern Recognition (ICPR)*, **2**, pp. 1251–1255, 1986.
6. T. Perroud, K. Sobottka, and H. Bunke, "Text extraction from color documents-clustering approaches in three and four dimensions," in *Proc. Int'l Conf. Document Analysis and Recognition (ICDAR)*, pp. 937–941, 2001.
7. P. K. Loo and C. L. Tan, "Adaptive region growing color segmentation for text using irregular pyramid," in *Proc. Int'l Workshop Document Analysis Systems (DAS)*, pp. 264–275, (Florence, Italy), 2004.
8. U. Garain, T. Paquet, and L. Heutte, "On foreground-background separation in low quality color document images," in *Proc. Int'l Conf. Document Analysis and Recognition (ICDAR)*, **2**, pp. 585–589, (Seoul, Korea), 2005.
9. U. Garain, T. Paquet, and L. Heutte, "On foreground - background separation in low quality document images," *Int'l J. Document Analysis and Recognition* **8**(1), pp. 47–63, 2006.
10. B. Gatos, I. Pratikakis, and S. J. Perantonis, "An adaptive binarization technique for low quality historical documents," in *Proc. Int'l Workshop Document Analysis Systems (DAS)*, pp. 102–113, (Florence, Italy), 2004.
11. E. Kavallieratou and E. Stamatatos, "Improving the quality of degraded document images," in *Proc. Int'l Conf. Document Image Analysis for Libraries (DIAL)*, (Lyon, France), 2006.
12. I. Bar-Yosef, I. Beckman, K. Kedem, and I. Dinstein, "Binarization, character extraction, and writer identification of historical hebrew calligraphy documents," *Int'l J. Documents Analysis and Recognition* **9**, p. 8999, 2007.
13. A. Antonacopoulos and D. Karatzas, "Document image analysis for world war ii personal records," in *Proc. Int'l Workshop on Document Image Analysis for Libraries*, pp. 336–341, 2004.
14. A. Antonacopoulos and D. Karatzas, "A complete approach to the conversion of typewritten historical documents for digital archives," in *Document Analysis Systems VI*, A. Dengel and S. Marinai, eds., *Lecture Notes in Computer Science* **3163**, pp. 90–101, Springer, 2004.
15. A. Antonacopoulos and C. C. Castilla, "Flexible text recovery from degraded typewritten historical documents," in *Proc. ICPR*, pp. 1062–1065, 2006.
16. J. Hobby and T. K. Ho, "Enhancing degraded document images via bitmap clustering and averaging," in *Proc. Int'l Conf. Document Analysis and Recognition (ICDAR)*, **1**, pp. 394–400, (Ulm, Germany), 1997.
17. T. Kanungo, R. Haralick, H. Baird, W. Stuezle, and D. Madigan, "Statistical, nonparametric methodology for document degradation model validation," *IEEE Trans. Pattern Analysis and Machine Intelligence* **22**(11), pp. 1209–1223, 2000.
18. G. V. Landon, Y. Lin, and W. B. Seales, "Towards automatic photometric correction of casually illuminated documents," in *Proc. CVPR*, 2007.
19. G. Zi, "Groundtruth generation and document image degradation," Master's thesis, University of Maryland, College Park, 2005.
20. L. Likforman-Sulem, A. Zahour, and B. Taconet, "Text line segmentation of historical documents: a survey," *Int'l J. Documents Analysis and Recognition* **9**, pp. 123–138, 2007.

21. M. Hagedoorn, *Pattern matching using similarity measures*. PhD thesis, Utrecht University, the Netherlands, 2000.
22. S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pattern Analysis and Machine Intelligence* **24**(4), pp. 509–522, 2002.