16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

MODELING IMAGE DEGRADATIONS FOR IMPROVING OCR

Elisa H. Barney Smith

Electrical and Computer Engineering, Boise State University
Boise, Idaho, 83725, USA
email: EBarneySmith@BoiseState.edu web: coen.boisestate.edu/EBarneySmith

ABSTRACT

Clean documents are relatively easy to recognize. However, when digitizing collections of documents, the clean ones are rarely the documents that are encountered. The processes of printing and scanning documents introduce image degradations that interfere with the segmentation and recognition processes. Mathematical models of the degradation processes are presented. From these the types of degradations that are seen can be quantitatively and qualitatively described. Included in the discussion are sampling, edge spread, corner erosion, and edge noise. The relationship between these degradations and common OCR errors is described. By considering the degradation model, a theoretical foundation is available to improve the document recognition process.

1. INTRODUCTION

Document image analysis is the process of taking paper documents, digitizing them and then interpreting the contents of those images. There are many types of documents that can be analyzed, ranging from simple office documents, to historical documents hundreds of years old. The reasons these documents are digitized varies. Modern documents are digitized often so their contents can be reused without the tedious task of retyping them. Sometimes they are digitized to save storage space, or so the contents of large collections can be searched or summarized. Historical documents are often digitized so a broader population can access and utilize the contents of the document. For both new and old documents, the World Wide Web has made document collections accessible to a broad audience. The larger population eager to view these documents has increased the demand for further document collections to be digitized.

New and old documents will vary in content and composition, but the general process for analyzing them is the same. An image of the paper document is created by either a desktop scanner, or a digital camera. The contents of the image is broken into smaller pieces. This usually starts by separating images from text regions, then breaking text regions into columns, then paragraphs, lines, words, and finally characters. The characters are then processed by an Optical Character Recognition (OCR) engine that will determine what letter or ASCII symbol is represented by that image segment. The recognition process depends on the segmentation being properly conducted and symbol shape being adequately preserved.

The segmentation process generally relies on the content consisting of dark inks on solid light backgrounds. Extremely old documents are likely to have a lower contrast between the foreground and background as the paper is colored and the ink is lighter. These are also more likely to be scanned into a gray level or color digital image as the composition of these documents is of as much interest as their content. Documents from the past century will likely have a black ink on a white background. These are more likely to be scanned into a black and white image. The signal processing for the gray level of color images has been well studied. The theoretical foundation for studying the formation of the bilevel images is less thoroughly studied.

This paper will describe the image acquisition process mathematically through introducing a prominent document degradation model. Based on that degradation model, an analysis of the how acquisition of documents through this process affects the bilevel contents of documents will be described. There are several degradations that have been observed both by researchers who focus on the pure degradation and by those who develop methods to recognize document content dispite these degradations. These degradations and how they are related to the degradation process will be described.

2. IMAGE DEGRADATION MODEL

A degradation model provides a theoretical platform through which the document acquisition process can be analyzed. While ultimately all system components must work well on actual samples and with all components contributing to the success or failure as they do, it is beneficial to have a theoretical framework to guide the design and evaluation. For instance, filters are often applied to the acquired image before segmentation to remove noise or smooth contours aiming to increase both the segmentation and the recognition accuracies. Their design and evaluation is based on samples of data, and the performance of OCR engines at hand. This leads to the effects of the filter and the OCR classifier being coupled. While some coupling is necessary for the complete system to perform optimally, the design of the individual parts can be better studied when the sources of the degradations are analyzed with a solid signal processing component.

Degradation models also provide a platform in which to run tests. Large quantities of labeled data can be generated inexpensively. Understanding the types of degradations that occur can lead to better design of algorithms, as well as generation of targeted datasets.

The degradation model that is described next is based on the physics of the image acquisition process. It is a portion of the degradation model presented by Baird in 1990 [1]. It is described schematically in Figure 1. It will be used as the basis for the studies reported in the remainder of the paper.

Images are converted to digital form using either a photographic camera or a document scanner, such as a desktop scanner. At each digital sensor light reflecting off the page is accumulated. The assumption is nominally that the doc-

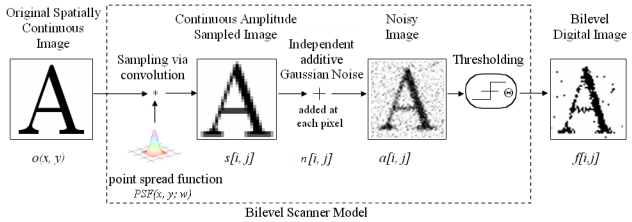

Figure 1: Scanner model used to determine the value of the pixel $[i, j]$ centered on each sensor element.

ument is uniformly lighted, in which case the light that is acquired is proportional to the reflectance of the paper image itself. At each discrete sensor light is collected from a region of the paper. The contribution of each region is the Point Spread Function, PSF, $PSF(x,y)$, or impulse response of the optical system. The acquired image is the convolution of the source image, $o(x,y)$, with the PSF,

$$s[i,j] = \int\int PSF(x_i - u, y_j - v; w)o(u,v)dudv. \quad (1)$$

The PSF is usually chosen to be circularly symmetric and describable by a single width parameter, w. The most common functional forms are bivariate Gaussian and bivariate Cauchy. Gaussian is often used because of its many nice mathematical properties. Cauchy more accurately represents the physics of the scanner, but unlike Gaussian, its heavy tails do not lend itself to compact numerical simulation.

During the acquisition process, noise is often included. Some of this comes from sensor noise, some of this comes from the unevenness in the response of the sensors, and some is from variations in the source image. This noise is modeled as being additive independent Gaussian noise, $n[i,j]$, with a standard deviation, σ_{noise},

$$a[i,j] = s[i,j] + n[i,j]. \quad (2)$$

For gray level images, usually 256 gray levels are recorded, and they are not linearly proportional to the reflectance or received signal. In document imaging, it is more common for bilevel, black and white, images to be acquired. This is partially due to the majority of images consisting of black print on white paper, and partially due to the historical limitations in disk space and memory, making 1-bit black and white images preferable to 8-bit gray scale images. When bilevel digital images are created, an extreme form of the intensity quantization process is implemented through global thresholding,

$$f[i,j] = \begin{cases} 1 & a[i,j] \geq \Theta \\ 0 & a[i,j] < \Theta. \end{cases} \quad (3)$$

Prior to the thresholding, the processes are linear and can be analyzed using standard linear signal processing methods. When the thresholding is included, the non-linearity from the thresholding requires different analysis methods. These have been used to describe the relationship between the source image content and the output image content. Degradations observable in the resulting image and their relation to the degradation model are described in the next section.

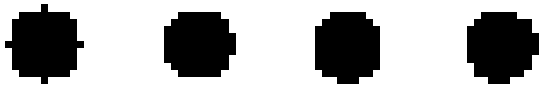

Figure 2: Examples of images of circles of radius 5. Shown are 4 different sampling phases, each of which produces a different bitmap.

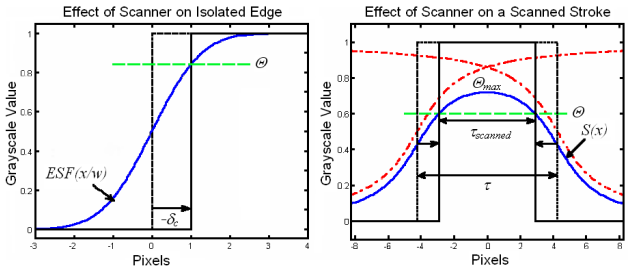

Figure 3: Effect of blurring on edges (a) an isolated edge (b) a pair of edges separated by less than the support of the PSF.

3. OBSERVED IMAGE DEGRADATIONS

For document images, the high contrast input image is predominantly composed of strokes of various width, lengths, orientations and concavities. These strokes are defined by their edges. Intersecting strokes are defined by the angle at which they meet. Between the strokes are areas of nominally solid background (white) or foreground (black) areas. These are characterized by their color. Degradation is now defined to be changes in the image from the source (paper) image to the digital image. The degradations affecting the image content have been grouped into four categories: spatial sampling, edge effects, corner effects and noise effects.

3.1 Digitization

When digitizing a document image, even if ideal sampling is used, there will be some differences seen in the digitized image relative to the original due to spatial quantization. The roundness of a curved edge, and even the smoothness of a straight edge will not be preserved in the resulting discrete image. Even if using ideal sampling, there are multiple possible resulting bitmaps for digitizing any given document, as shown in Figure 2. Each of these different combinations corresponds to a different sampling phase and will produce a different bitmap. The differences are at the edges of an image, and the uncertainty and differences are often treated as noise. Sarker et. al [13] described the number of possible patterns and their respective frequency of occurence using modulo-grid diagrams. Dorst and Duin [8] described the same using spirographs. For a circle of radius 5, there are approximately 236 possible bitmaps, of which four are shown in Figure 2.

3.2 Edge Spread

When non-ideal sampling is used, meaning that the Point Spread Function is not an impulse, additional degradations are introduced into the image. One of the more prevalent and highly observable degradations is the change in the edge

location. Generally the exact edge position of the digitized edge relative to the edge in the original continuous image is not measurable, but when two edges are considered together, the change in position of each edge results in a change of stroke width. This will affect the image appearance.

The amount that a single edge in isolation will change can be calculated by treating the single edge source image $o(x)$ as a unit step function. The response to blurring will be

$$s(x) = ESF\,(x/w)\,, \tag{4}$$

where ESF is the edge spread function which is the cumulative marginal of the functional form of the PSF used in the blurring, Figure 3(a). The edge occurs where the amplitude $s(x)$ equals the threshold, Θ, so

$$\delta_c = -wESF^{-1}(\Theta). \tag{5}$$

As the edge spread depends on two system parameters, there are an infinite number of (w,Θ) pairs that can form any specific edge spread, δ_c, amount [2].

If two edges forming a stroke are located further than the distance of the support of the PSF apart, then the stroke width will change by $2\delta_c$. If the two edges are closer than the support of the PSF, then the effects of both edges will affect the output image in a way that is not simply additive [10].

The amount of change can be calculated by treating the one dimensional cross section of a scanned stroke as a square pulse with a width τ, Figure 3(b). Assuming that the square pulse is centered on the origin, the value of pixels as a function of their position, $s(x)$, can be found using

$$s(x) = ESF\left(\frac{(\tau/2)-x}{w}\right) - ESF\left(\frac{(-\tau/2)-x}{w}\right). \tag{6}$$

As with isolated edges the new edges are located where $s(x)$ is equal to Θ. If the thickness of the stroke after thresholding is given by $\tau_{scanned}$, then the threshold is

$$\Theta = ESF\left(\frac{\tau-\tau_{scanned}}{2w}\right) - ESF\left(\frac{-\tau-\tau_{scanned}}{2w}\right). \tag{7}$$

The change of stroke width Δ can be defined as

$$\Delta = \tau_{scanned} - \tau, \tag{8}$$

which leads to

$$\Theta = ESF\left(-\frac{\Delta}{2w}\right) - ESF\left(\frac{-2\tau-\Delta}{2w}\right). \tag{9}$$

Inverting this to solve for Δ must be done through numerical methods. When using ideal sampling, or very small PSF widths, the edge spread is zero. This also occurs for isolated edges if the threshold $\Theta = 0.5$. If w is small or τ is large $\Delta = 2\delta_c$.

For large PSF widths, w, Δ is smaller than $2\delta_c$. For positive Δ the strokes are not widened as much as they would be if they were further apart, and for negative Δ they are thinned even more. As shown in Figure 3(b), for many high thresholds the strokes can totally vanish. A similar and opposite effect would occur for white strokes on a black background, but this situation is less often encountered in document images.

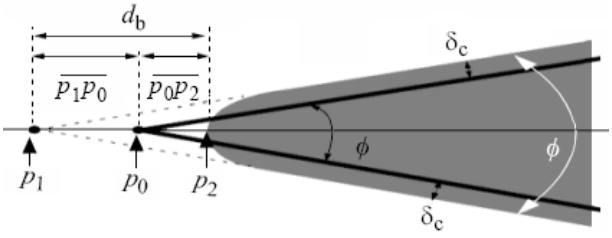

Figure 4: Effect of blurring on corners

3.3 Corner Rounding

Another bilevel image degradation is the amount of erosion seen in a black or white corner after scanning [3]. This degradation is caused by the interaction of the two edges, but also includes the displacement of the individual edges. In Figure 4, point p_0 is the apex of the original corner. Point p_2 is the point along the angle bisector of the new rounded corner where the blurred corner equals the threshold value. Point p_1 is the point where the new corner edges would intersect if extrapolated. The distance

$$d_b = \overline{p_1p_2} = \overline{p_1p_0} + \overline{p_0p_2} \tag{10}$$

is not the erosion from the original corner location, but it does represent the degradation actually seen on the corner. This quantity can be measured from bilevel document images and can be predicted from the model. This corner erosion distance, d_b, depends on the threshold, the PSF width, and the functional form similar to the edge displacement above.

The corner erosion distance is a combination of the distance along the angle bisector from the original corner to where the amplitude of the blurred corner equals the threshold, $\overline{p_1p_2}$, and the distance from the original corner to the extrapolated corner, $\overline{p_1p_0}$, which is based on the edge spread δ_c. Thus

$$d_b = \frac{-wESF^{-1}(\Theta)}{sin(\phi/2)} + f_b^{-1}(\Theta; w, \phi) \tag{11}$$

where the intensity along the angle bisector is

$$f_b(\tilde{x}; w, \phi) = \int_{x=0}^{x=\infty}\int_{y=-xtan\frac{\phi}{2}}^{y=xtan\frac{\phi}{2}} PSF(x-\tilde{x}, y; w)dydx. \tag{12}$$

As with edge displacement, a given amount of corner erosion can also occur for an infinite number of (w, Θ) values. The erosion of a white corner is defined similarly and results in

$$d_w(w,\Theta) = d_b(w, 1-\Theta). \tag{13}$$

One attractive aspect of corner erosion is that it can be relatively easily measured from the degraded image, and requires no a priori information from the original image other than that the adjacent edges are straight. If the measurements are accurate, or enough can be collected, the combination of measurements from black and white corners can be used to estimate the degradation model parameters, w and Θ [15].

It was stated that the edge spread will be zero if the threshold is 0.5 for isolated edges. For corners, the edges are not isolated and the corners will erode, and they will erode more for larger w. Thinning followed by thickening on isolated edges and broad strokes is reversible, but the loss due to corner rounding or strokes breaking is not reversible.

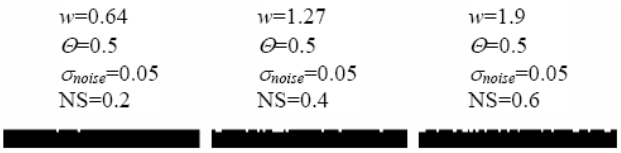

Figure 5: Effect of noise on edges. The additive noise used to produce all images is equal, but the observed effect changes.

3.4 Noise Spread

For grey level images noise is usually described by the standard deviation of the additive noise. However the amount of noise present in a bilevel scanned image is not dependent purely on the level of noise added prior to thresholding. This can be seen clearly by looking at Figure 5. The three images all result from additive noise with the same standard deviation, σ_{noise}, however the appearance of the noise and the effect of the noise on the images is noticeably different among the three images.

A better descriptor of the amount of noise in a bilevel image is the noise spread (NS) [10]. The basic idea behind noise spread is that when an image is thresholded the noise is concentrated on the edges of the objects in the image. The noise spread for a given edge is the size of the domain in which pixels are affected by additive noise. Typically this domain, called the noise spread region, is less than a pixel thick.

Isolated edges can be represented in one dimension as step functions. Section 3.2 discussed how straight edges are affected by scanning, but that section focused on the deterministic effects of scanning. Figure 3(a) shows how an edge is affected by scanning when noise is disregarded. As was discussed in Section 3.2 the edge shifts by δ_c. However, as shown in Figure 6, when noise is added there is a region in which the value of pixels after thresholding is uncertain. This region is called the noise spread region [10]. The size of this region is called the noise spread (NS), and as the examples in Figure 5 show, it is a good quantitative measure of how noisy a bilevel image is.

The precise definition of NS starts with the probability that a pixel at a certain distance from the edge will be above the threshold. This threshold probability (THP) depends on the cumulative distribution function (CDF) of the noise,

$$THP(x) = CDF\left(\frac{ESF(x/w) - \Theta}{\sigma_{noise}}\right). \tag{14}$$

The noisy edge will be above the threshold with probability near 0 on one side of the NS region and with a probability of near 1 on the other side of the NS region. By defining the noise spread as the breadth of the domain over which the threshold probability is in the range $(\alpha, 1-\alpha)$, a parameter Z_α can be defined such that $1-\alpha = CDF(Z_\alpha)$. Since most CDFs are odd symmetric, the noise spread region will be centered on δ_c so

$$1-\alpha = THP\left(\frac{NS}{2}\cdot\delta_c\right). \tag{15}$$

Two terms of a linear approximation of the ESF are substituted into Equation 14, which is combined with Equation 15.

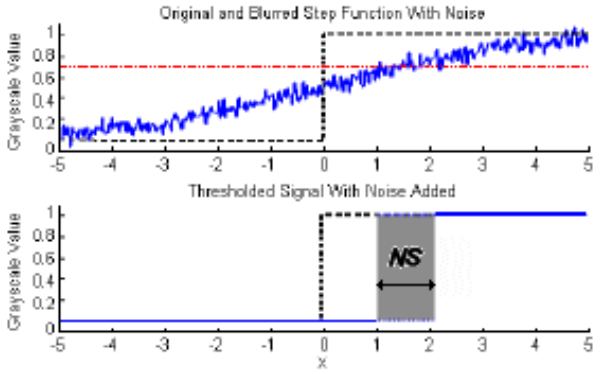

Figure 6: (top) Blured edge with noise. (bottom) After thresholding the uncertain boundary shown in gray, is the noise spread region. The effects of sampling are not shown.

Z_α is then equated with the arguments of the CDF which when solved for NS results in

$$NS = \frac{2Z_\alpha \cdot \sigma_{noise} \cdot w}{LSF(ESF^{-1}(\Theta))}. \tag{16}$$

Details of this derivation are shown in [10]. A choice of $Z_\alpha = \sqrt{2\pi}/2$ has been found to be a good value producing a threshold probability in the range (0.105, 0.895).

Noise Spread through the THP also represents the probability that an edge pixel has been affected by noise. It has also been shown that the NS is also directly proportional to the Hamming distance between an edge without noise and an edge with noise.

4. EFFECTS OF IMAGE DEGRADATION

Several different ways of quantifying degradation effects on images relative to a model that describes how bilevel images are acquired and created have been described. Understanding the degradations can assist researchers in development of OCR classifiers and document image processing algorithms.

Even when no blurring or additive noise is present, just the sampling alone leads to a lot of variation in the possible bitmaps. From a practical standpoint, this limits the use of certain classification methods such as templates.

Touching and broken characters are one of the leading causes of OCR errors [6][12], largely because they can cause the segmentation algorithm in an OCR system to improperly segment the characters. This can cause the letters 'rn' to be interpreted as an 'm' or other pairs of characters when joined to be interpreted as nonsense. The breaking of characters can lead to the division of the 'm' reversing the above example, or the division of an 'o' into what may be interpreted as two parentheses, '()'. There have been developments in the field of word recognition [14][7] as an attempt to overcome these common errors. Absent of such word recognition techniques, the OCR system is unlikely to properly identify the label of the pair of merged characters, and is not designed to return two labels. Many OCR systems will include a post-processing step after the classifier to look for words in a lexicon that can correct for errors introduced by degraded characters [9]. A priori knowledge that the characters are likely to be broken or alternatively merged, can be used to bias these techniques to favor the appropriate corrections.

In [4] the homogeneity of characters with respect to the edge spread, δ_c was shown. This homogeneity was used in [5] to train a classifier in sections defined by the edge spread. Four different classifiers were constructed, one for each of the following edge spread partitions $\delta_c \in (-\infty,-0.5],(-0.5,0.5],(0.5,1.5],(1.5,\infty)$. The test characters were channeled into the appropriate classifier based on their edge spread. Classification on the mixed data set by a single classifier resulted in 96.2% accuracy. When the dataset was partitioned into four edge spread regions, the recognition rate increased to 97.6%, which was shown to be a statistically significant improvement.

The sharpness of the corners in character images are not generally used as significant features in character recognition, but are something that can be observed by the human readers of the digitized document images. High amounts of eroded corners will change the characteristics of the image appearance and how human readers perceive the document image.

Additive noise, optical blurring or low thresholds of the sensed light signal can cause adjacent characters to touch. The Noise Spread can quantify and predict the amount of edge noise that will be present. From this the likelihood of touching characters could be anticipated. For other symbols being recognized, a technique of skeletonization is often performed to get the basic symbol shape in the form of a "stick figure". Noise along the edge of shapes will often interfere with the skeletonization process causing extraneous spurs to be created. A priori knowledge of this tendency can influence follow-on processing approaches.

5. CONCLUSION

The degradation model allows considerable analysis of the content of document images and ties the observed image features to a theoretical description of their sources.

Originally degradation models were proposed solely as a source to generate large quantities of labeled data to use in training OCR systems. Their potential uses, however, extend well beyond this purpose. They can also be used to design and test filtering algorithms through a theoretical understanding of the degradation sources. Classifier design can be affected with a priori information about the degradations likely to occur.

The degradation categories of sampling phase, edge spread, corner erosion and edge noise describe the major processes. Sampling effects can't be avoided, but provide a great variety in the resulting bitmaps. Edge Spread describes when and why the strokes of characters are likely to become wider or thinner. Corner erosion describes when the details of the font flourishes are likely to be absent or preserved. The Noise Spread discusses when the noise around the edges is likely to be notable.

Methods to measure some of these parameters exist, such as corner erosion. Techniques such as quality metrics [11] have some potential for use in measuring some of these degradations. A method for measuring Noise Spread is also in development. The measurement tools will bridge the last gap between the theory and regular utilization.

6. ACKNOWLEDGEMENTS

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

REFERENCES

[1] H. S. Baird, "Document Image Defect Models," in *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] E. H. Barney Smith, "Characterization of Image Degradation Caused by Scanning," *Pattern Recognition Letters*, Vol. 19, No. 13, 1998, pp. 1191–1197.

[3] E. H. Barney Smith, "Estimating Scanning Characteristics from Corners in Bilevel Images," in *Proc. SPIE Document Recognition and Retrieval VIII*, Vol. 4307, San Jose, CA, 21-26 Jan. 2001, pp. 176–183.

[4] E. H. Barney Smith, and X. Qiu, "Statistical image differences, degradation features and character distance metrics," *Int. J. of Document Analysis and Recognition*, Vol.6, No. 3, 2004, pp. 146–153.

[5] E. H. Barney Smith and T. Andersen, "Text Degradations and OCR Training," in *Proc. Int. Conf. on Document Analysis and Recognition 2005*, Seoul, Korea, 29 Aug. - 1 Sept. 2005, pp. 834–838.

[6] M. Bosker, "Omnidocument Technologies," *Proc. of the IEEE*, Vol. 80, No. 7, 1992, pp. 1066–1078.

[7] C. H. Chen, J. L. DeCurtins, "Word recognition in a segmentation-free approach to OCR," in *Proc. Int. Conf. on Document Analysis and Recognition*, Tsukuba Science City, Japan, 20-22 Oct. 1993, pp. 573–576.

[8] L. Dorst and A. W. M. Smeulders, "Discrete Representation of Straight Lines," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 6, No. 4, July 1984, pp. 450–463.

[9] G. Ford, S. E. Hauser, D. X. Le, G. R. Thoma, "Pattern matching techniques for correcting low-confidence OCR words in a known context," in *SPIE Document Recognition and Retrieval VIII*, Vol. 4307, 24-25 Jan. 2001, pp. 241–249.

[10] C. McGillivary, *Quantifying Noise Effects In Bilevel Document Images*, Masters Thesis, Boise State University, Boise, Idaho, USA, Dec., 2007.

[11] Darrin K. Reed and Elisa H. Barney Smith, "Correlating degradation models and image quality metrics," Proc. SPIE Document Recognition and Retrieval XV, San Jose, CA, Jan. 2008.

[12] S. V. Rice, F. R. Jenkins, T. A. Nartker, "The Fourth Annual Test of OCR Accuracy," in *Information Science Research Institute 1995 Annual Report*, 1995, pp. 11–49.

[13] P. Sarkar, G. Nagy, J. Zhou, D. Lopresti, "Spatial sampling of printed patterns," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 3, Mar. 1998, pp. 344–351.

[14] Y. Xu, G. Nagy, "Prototype extraction and adaptive OCR," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 21, No. 12, Dec. 1999, pp. 1280–1296.

[15] H. S. Yam and E. H. Barney Smith, "Estimating Degradation Model Parameters from Character Images," in *Proc. Int. Conf. on Document Analysis and Recognition 2003*, Edinburgh, Scotland, 3-6 Aug. 2003, pp. 710–714.