Presented at IS&T/SPIE EI'99, Conference 3651, Document Recognition and Retrieval VI
Jan 26-28, 1999, San Jose, CA.

Determining the Resolution of Scanned Document Images

Dan S. Bloomberg

Xerox Palo Alto Research Center
Palo Alto, CA 94304

Abstract

Given the existence of digital scanners, printers and fax machines, documents can undergo a history of sequential reproductions. One of the most important determiners of the quality of the resulting image is the set of underlying resolutions at which the images were scanned and binarized. In particular, a low resolution scan produces a noticeable degradation of image quality, and produces a set of printed fonts that cause omnifont OCR systems to operate with a relatively high error rate. This error rate can be reduced if the OCR system is trained on text having fax scanner degradations, but this also requires that the OCR system can determine in advance if such degraded fonts are present.

Methods are found for determining the lowest resolution that a binary document image has been scanned (or printed) in the past. Observing that the signature for a low resolution scan is contained in the boundary contours of the black components, the connected components that constitute the contours are measured, and both the size histogram and its power spectrum are used to determine the underlying low resolution, if any. The primary application is to fax, both standard and fine. Degradation of both font appearance and OCR performance is far more severe for standard than fine fax, so the most important practical problem is to identify documents of binary scanned text that have a standard fax signature. Only a small part of the page image is required to determine the existence of such a signature; consequently, the image is pre-processed to locate a subregion that is well suited for analysis. Results for discriminating between originals, standard fax and fine fax on a small but diverse test suite are given.

Keywords: image resolution, facsimile, fax, pattern spectrum, power spectrum, image analysis

1 Introduction

The purpose of this study is to find accurate and efficient methods for determining if a digital image has been previously scanned at low resolution, measured in pixels/inch (ppi). The most common fax resolutions are *standard*, approximately 200 x 100 ppi, and *fine*, approximately 200 x 200 ppi. Images scanned at standard fax resolution suffer obvious visible degradation. This is also reflected by the generally poor performance of omni-font OCR systems on such images [3]. For example, in the OCR "bakeoff" from the 1996 SDAIR Conference, the character error rate of the two *best* recognizers on standard fax was nearly 6 percent; the others were more than 10 percent. This should be compared with best character error rates of less than 2 percent on 300 ppi binary, and less than 3 percent on 200 ppi binary and fine fax. OCR performance can be improved using special recognizers trained on low-resolution fonts, but their use requires prior determination of the effective resolution of the image. If the fax is stored as an electronic image, it is easy to infer the scan resolution from the image size. But if the fax is printed and then scanned at high resolution, with or without prior copying, the inherent low resolution of the image contents must be inferred from the image itself.

Our approach is morphological and relies on the statistics that can be accumulated from a region of the image. The methods have general validity, but we apply them for the common situations where the low resolution scan is fax (either about 100 ppi or 200 ppi) and the high resolution scan is either 300 ppi or 400 ppi. For the standard fax identification problem, it is necessary to extract a set of measurements from the image for which a discriminator can unambiguously place the image into one of two categories: high resolution or low resolution. The approach is empirical. Histograms constructed from data extracated by various measurements are analyzed for structural or periodic characteristics. Where low and high resolution images show significant differences, the feasability of building a discriminator is investigated. It is reasonable to vary the discrimination function depending on the final scan resolution, which is typically known or can be inferred from the image size.

Fax documents are typically scanned in standard mode and in fine mode. In the sequel, although all images we study have been most recently scanned at high resolution (300 or 400 ppi), for simplicity we refer to an image that was never scanned at low resolution as an *original*, an image previously scanned at standard fax resolution as a *standard-fax*, and an image previously scanned at fine fax resolution, which is about 200 x 200 ppi, as a *fine-fax*. Naturally, it should be easier to discriminate a standard-fax image from an original than a fine-fax image. Fortunately, as a practical matter, it is far more important to identify standard-fax images, because omni-font OCR systems do fairly well on fine-fax.

2 Approach

Visually, the difference in appearance between standard fax and original images is striking. Three significant components of the difference are:

- Thickening introduced by the low resolution fax scanner.
- A tendency for larger parts of the fax boundaries to be horizontal and vertical.

- The original boundaries are smoother.

The thickness of strokes varies greatly between scanned documents, and is not a robust parameter for discriminating between original and fax images. The other two observations relate to the boundary characteristics, and much of the signature of a low resolution scan is expected to reside in the edge pixels. Thus, the focus of investigation will be on the statistics of boundary pixels, and in particular, functions $f(l)$ of the length l of runs or connected components of the boundary. The most important operations for generating $f(l)$ are expected to be image morphology, granulometries, and connected component labelling. There are a number of questions to consider, and the final arbiter is empirical:

1. Should the images be filtered morphologically for noise before extracting the edge pixels?
2. Should all edge pixels be considered together, or only horizontal or vertical edge pixels separately?
3. For standard fax, because the lowest resolution is in the vertical direction, should we concentrate exclusively on vertical edge pixels?
4. What functions $f(l)$ should be derived from the edge pixels?
5. Of what use is the power spectrum of these functions?
6. How are the measurements combined to form robust and size-invariant discriminators?

Regarding the methodology, rather than using a model-based theoretical approach, where you try to guess the salient features in detail *a priori* and build a specific set of operations to extract exactly those features, we use an iterative method where you start with a guess about which features might be salient and take measurements where differences are expected to arise. Results are noted, and the extraction method is varied in an attempt to improve the results. Through such iterations, a subspace of methods and parameters is investigated. We show a few of the parameters that were varied, along with the best methods found.

Aiming for saliency on the low-resolution vertical scan for standard fax, we begin by extracting the *vertical* edge pixels[1]. The image is eroded with a horizontal structuring element of length 3, with the center at the center of the structuring element, and the result is XORed back with the uneroded image. Fig. 1 shows the vertical edge pixels extracted from the original, fine fax, and standard fax images.

Possibilities for $f(l)$ include the number of runlengths of each size, $r(l)$, perhaps weighted by a power of the length of the runs, or the number of connected components at each size, $c(l)$, again optionally weighted. In Fig. 1, the original appears to have many more isolated single pixels, and far fewer short vertical 8-connected components (8-cc), than the standard fax; the fine fax

[1] If we use *all* boundary pixels, the 8-connected components of the boundary are just the size of the connected components in the image.

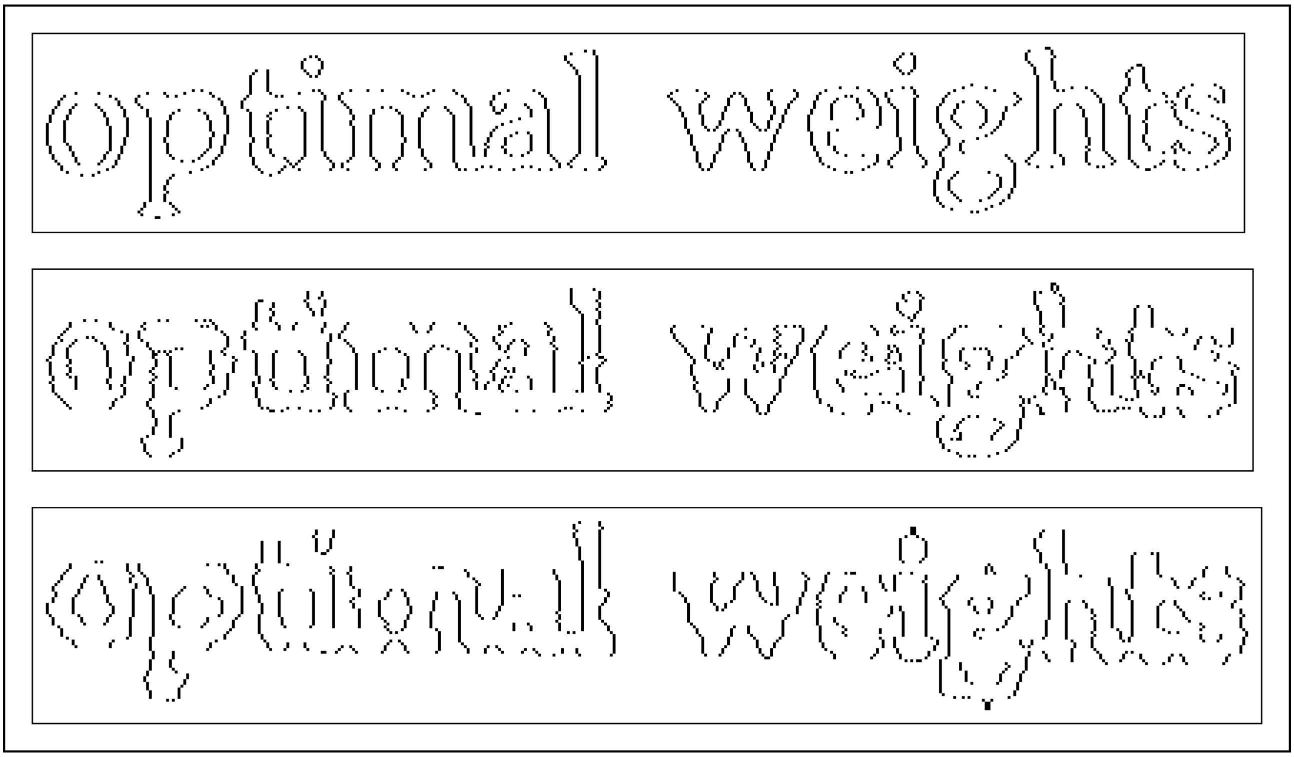

Figure 1: Vertical edge pixels: top (original); middle (fine fax); lower (standard fax)

is somewhere in between. For these boundary images, the number of vertical runlengths $r_v(l)$ is equal to the number of 4-connected components (4-cc) of that size, $c_4(l)$. Further, the pattern spectrum[4], which is the number of pixels within runs of a given length, is just the first moment $lr_v(l)$ of the runlength histogram, $r_v(l)$. The histogram of 8-cc cannot be derived simply from vertical runlengths. In general, we measure the distribution of d-connected component heights, both 4-cc and 8-cc, and with different moments m for the weighting factors; namely,

$$f(l) \equiv l^m c_d(l) \tag{1}$$

2.1 Typical component height data

Fig. 2 shows the distribution of 4-cc heights, $f(l) = c_4(l)$, from a typical original and standard fax, both taken from images sampled at 400 ppi. There is relatively little structural difference between the two. Both histograms are dominated by components with very low counts, as is evident from Fig. 1.

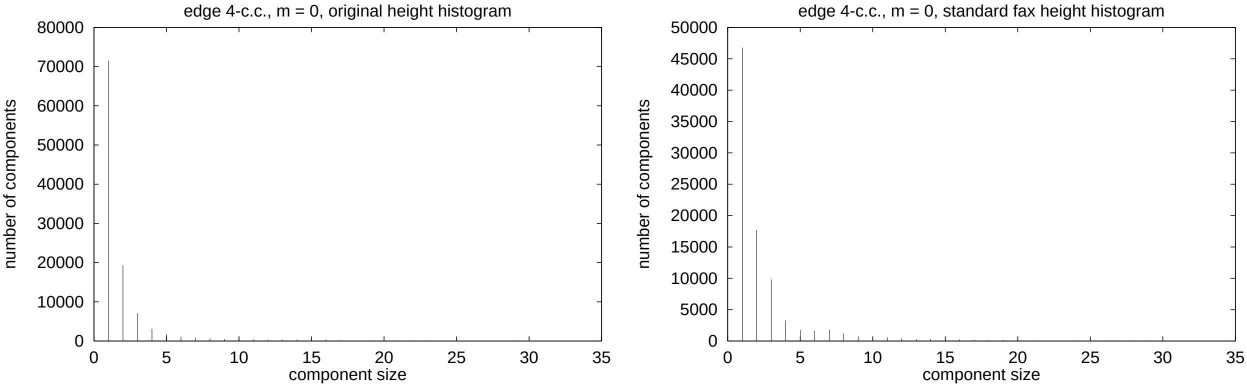

Figure 2: Counts of 4-cc heights, as a function of height, from original and standard fax images, at 400 ppi.

Fig. 3 shows the first moment of the distribution of 4-cc heights, $f(l) = lc_4(l)$, from a typical original and standard fax. Except for $l = 1$ and $l = 2$, the counts are significantly larger for standard fax, and this data may be useful for discrimination.

Fig. 4 shows the distribution of 8-cc heights, $f(l) = c_8(l)$, from a typical original and standard fax. The differences are striking. Most of the original counts are in $l = 1$; the signal for $l > 1$ is very small. For the fax, a significant part of the signal is in $l > 1$, and the underlying periodicity of the low-resolution scan at multiples of 4 pixels is evident.

Fig. 5 shows the first moment of the distribution of 8-cc heights, $f(l) = lc_8(l)$, from a typical original and standard fax. Use of the first moment spreads the data in the original image beyone $l = 1$, and the broad peak between 20 and 30 is due to components extending to the x-height of the text. (See Fig. 1). The periodicity of the fax at higher multiples of 4 is clearly evident, accentuated by the weighting factor.

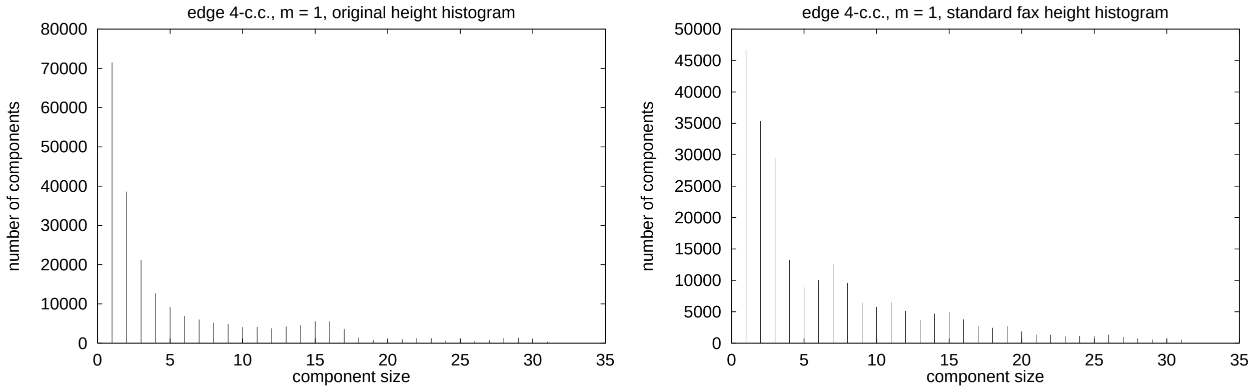

Figure 3: *Counts of the first moment of 4-cc heights, as a function of height, from original and standard fax images, at 400 ppi.*

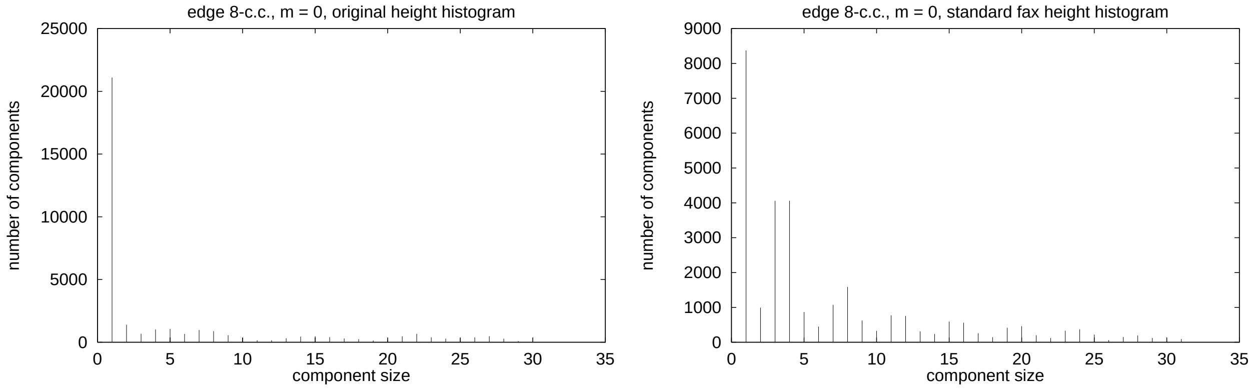

Figure 4: *Counts of 8-cc heights, as a function of height, from original and standard fax images, at 400 ppi.*

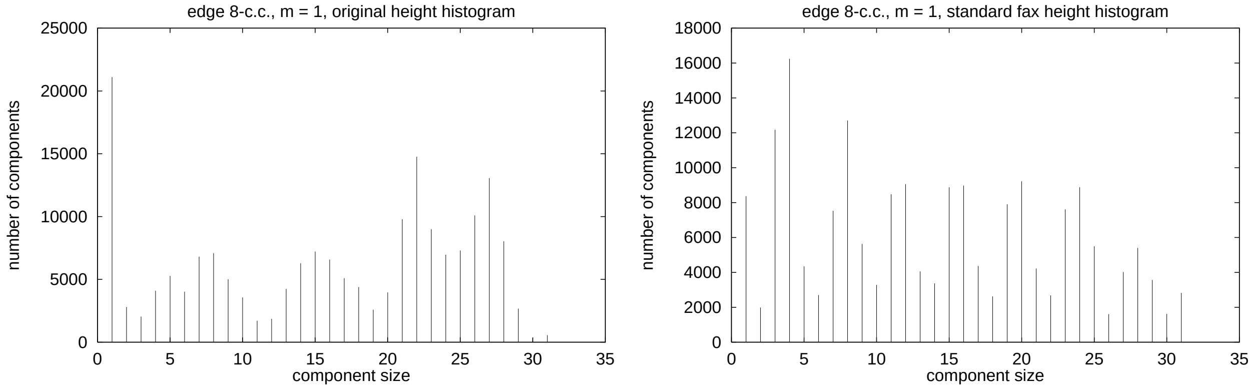

Figure 5: *Counts of the first moment of 8-cc heights, as a function of height, from original and standard fax images, at 400 ppi.*

Finally, in Fig. 6 we show comparable 8-cc histograms from original and standard fax images that have been rescanned at 300 ppi. The periodicity of the low resolution scan, this time in units of 3 pixels, is evident.

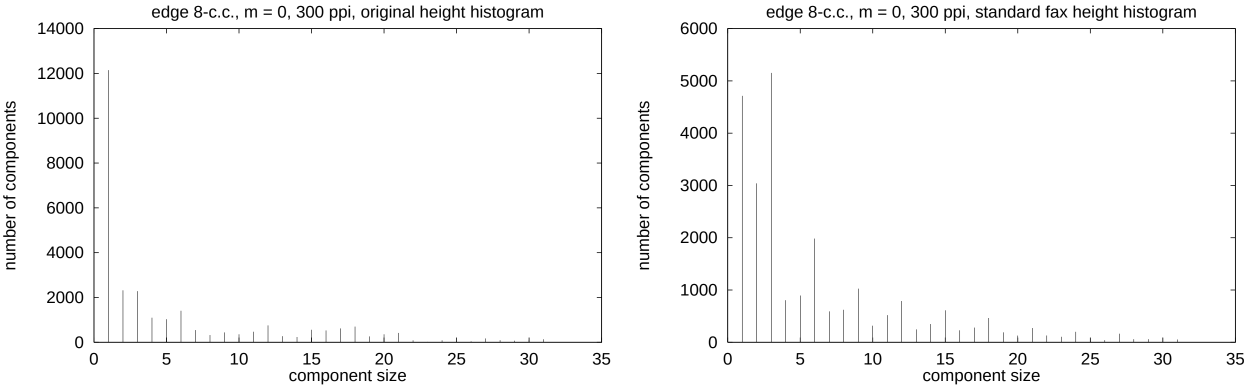

Figure 6: Counts of 8-cc heights, as a function of height, from original and standard fax images, at 300 ppi.

Because the differences between original and fax are much larger for 8-cc than 4-cc, we use them exclusively. Pre-filtering with small vertical openings and closings can be done to remove small pixel noise on the vertical edges, before the edges are extracted from the image. However, it was observed that such filtering reduced the differences between original and fax histograms, for both 4-cc and 8-cc; consequently, it is not used.

2.2 Discrimination functions and power spectra

The discrimination functions formed from these "features" must be invariant with respect to the size of the image and the number of runs or connected components in the image. The score s, which is to be compared with a threshold, can be constructed by taking the ratio of two linear combinations of the $f(l)$, or more simply by dividing a linear combination by any one of the numbers,

$$s = \frac{\sum f(l)}{f(l_k)} \tag{2}$$

By inspection, from Fig. 4 for 400 ppi, we choose

$$s_{400} = \frac{-f(2) + \sum_{l=3}^{8} f(l)}{f(1)} \tag{3}$$

and from Fig. 6 for 300 ppi,

$$s_{300} = \frac{\sum_{l=2}^{8} f(l)}{f(1)} \tag{4}$$

Note that these discrimination functions do not explicitly use the low-resolution periodic structure. Where such structure exists, discriminators that are sensitive to the specific wavelengths can be constructed. To emphasize the periodicity, it is useful to take the power spectrum of the histogram data, and use appropriate terms to form discriminating functions. Because the data is real, the power spectrum is symmetric about the center.

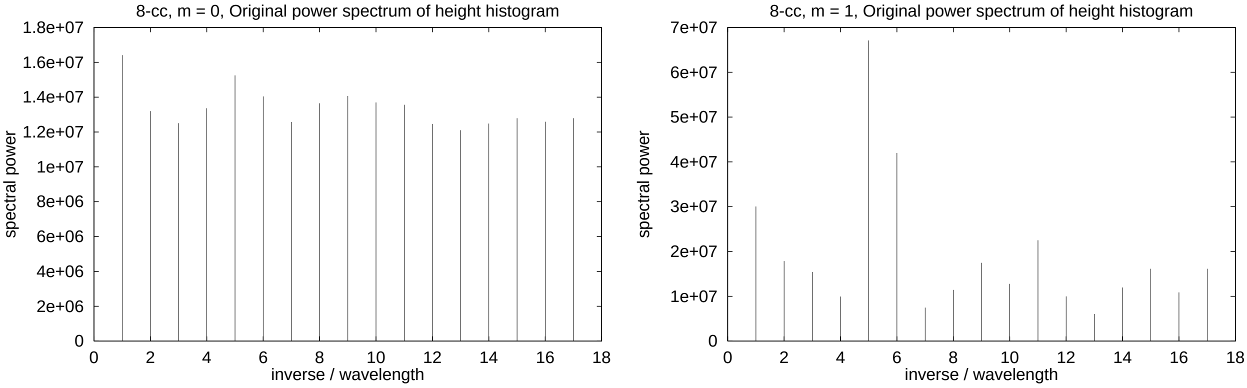

Figure 7: *Power spectra of the zeroth and first moments of 8-cc heights for an original at 400 ppi, as a function of inverse distance.*

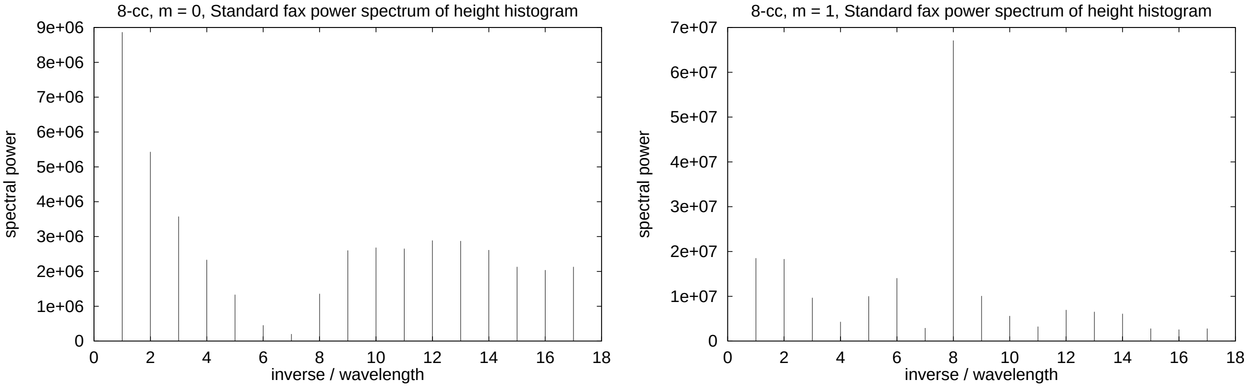

Figure 8: *Power spectra of the zeroth and first moments of 8-cc heights for a standard fax at 400 ppi, as a function of inverse distance.*

Fig. 7 shows the power spectrum of 8-cc heights with zeroth and first moments, for a typical *original* scanned at 400 ppi. These spectral components

$$p(k) = \frac{1}{N} \mid F(k) \mid^2 \tag{5}$$

are the absolute value squared of the fourier components

$$F(k) = \sum_{l=0}^{N-1} f(l) e^{2\pi ikl/N} \tag{6}$$

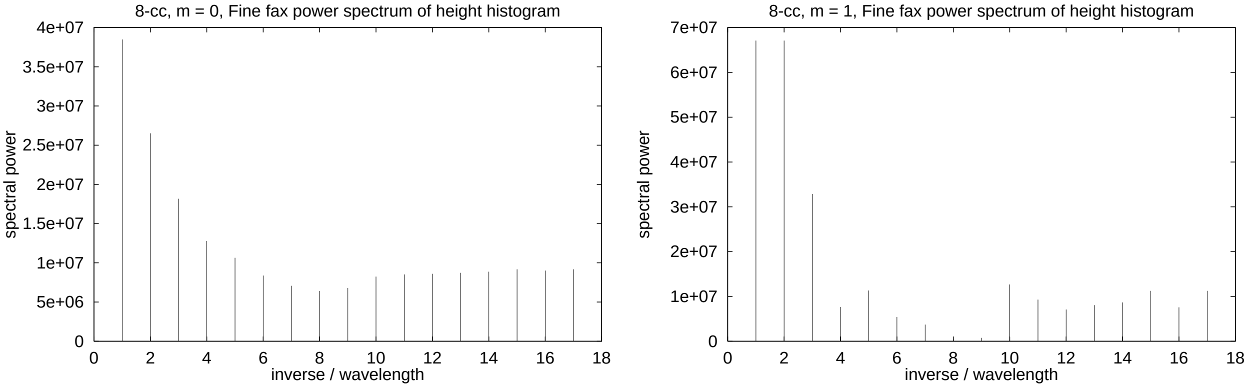

Figure 9: *Power spectra of the zeroth and first moments of 8-cc heights for a fine fax at 400 ppi, as a function of inverse distance.*

The power spectra were constructed from $N = 32$ samples, and are shown up to the Nyquist limit at $k = 16$, where the periodicity, λ, is given by $\lambda \equiv N/k = 2$ pixels. The *original* has a very uniform spectrum for m = 0, and structure with a concentration of weight at a frequency corresponding to about 7 or 8 pixels for m = 1. The power spectra of the zeroth and first moments of 8-cc heights for a typical *standard fax* are shown in Fig. 8. In contrast to the m = 0 spectrum from the original, the fax spectrum displays considerable structure. The m = 1 spectrum has a notable signature, where the large weight at exactly 4 pixels ($k = 8$, half the Nyquist cutoff) is highly prominent. The power spectra of zeroth and first moments of 8-cc heights for a typical *fine fax* are given in Fig. 9. In contrast to the standard fax spectrum for m = 1, this shows nearly zero weight at the 4 pixel periodicity. It also has a concentration of weight at a long periodicity of 10 to 12 pixels. For m = 1, the standard fax is more similar to the original than to the fine fax, and these differences will be used to make a discriminator between standard and fine fax.

Because the spectral shapes are quite uniform from image to image, discrimination functions can be generated to separate these images reliably. For the zeroth moment spectra, a size-invariant discrimination function can be made using just the ratio of the maximum to minimum values in the spectrum.

3 Subregion selection

Empirically, it is observed that the statistics gathered from a small part of a scanned page suffice for making an accurate determination of the underlying resolution. In fact, there are two reasons for *not* using the entire image. First, and of most importance, stipples and halftones can have an uncharacteristically large population of very small components, which can bias the edge pixel statistics to simulate a higher resolution. Second, the analysis of a subimage is faster than the full page.

The image is arbitrarily divided into N equal tiles, where N is typically taken to be 16, and the tile with statistics most similar to text is selected for resolution analysis. The operation to identify "textness" is as follows. For each tile:

- The image is threshold reduced[1] by a factor of 8, using a threshold of 1 on each 2x2 reduction. This is equivalent to performing an 8x8 dilation before subsampling, and consolidates the image pixels.
- The textlines are enhanced by closing with a horizontal structuring element (of size 15).
- The halftone and stippled regions are removed by first opening with a vertical structuring element (of size 10) to remove the text only, and then XORing to restore only the text.
- The tile with the largest number of ON pixels is selected.

These operations, taking place on a 8x reduced image, are much faster than doing a the resolution analysis on the full image. On a 167 MHz Sun Ultra-Sparc, for a 400 ppi image, the selection of a subregion takes 0.06 seconds. The resolution analysis requires computation of connected component bounding boxes and, on a subregion of size 1/16 (about 1 million pixels) of a typical image, takes 0.015 seconds. The total time is less than 100 ms.

4 Discrimination results

A set of 20 diverse pages was scanned and printed on a plain paper fax, at both standard and fine fax resolution. The 20 originals and the two sets of fax pages were then scanned at 300 ppi and 400 ppi. The image set is shown in Fig. 10. The discrimination functions given in (4) and (3) were used for the 300 ppi and 400 ppi sets, respectively, and applied to the zeroth and first moments of 8-cc heights.

The results using the 8-cc height histogram for 400 ppi originals, fine fax and standard fax are given in Fig. 11, with the zeroth moment on the left and the first moment on the right. In these and following plots, the standard fax is the solid line with diamonds, the originals are the dashed line with "+" marks, and the fine fax is the dotted line with open squares. Here, the standard fax is far above the originals, with the fine fax in between. A threshold is easily chosen to separate standard fax from originals; thresholds can also be chosen to separate fine fax from each of the others. Similar scores, generated at 300 ppi from the 8-cc height histograms, are shown in Fig. 12. Again the standard fax is easily separated by a single threshold from the original.

Using the power spectra for 8-cc, with the zeroth and first moments as shown in Fig. 7, Fig. 8 and Fig. 9, discrimination functions can be made from which the underlying resolution can be identified by application of a threshold. The result on the set of 20 pages is seen in Fig. 13. On the left, for m = 0, we use as a discrimination function the ratio of the max/min of the spectral values, capping the standard fax ratio at 100 to make visible the lower scores of the other plots. A threshold at about 15 separates the standard fax from fine fax and originals. On the right, for m = 1, we use the spectral discrimination function

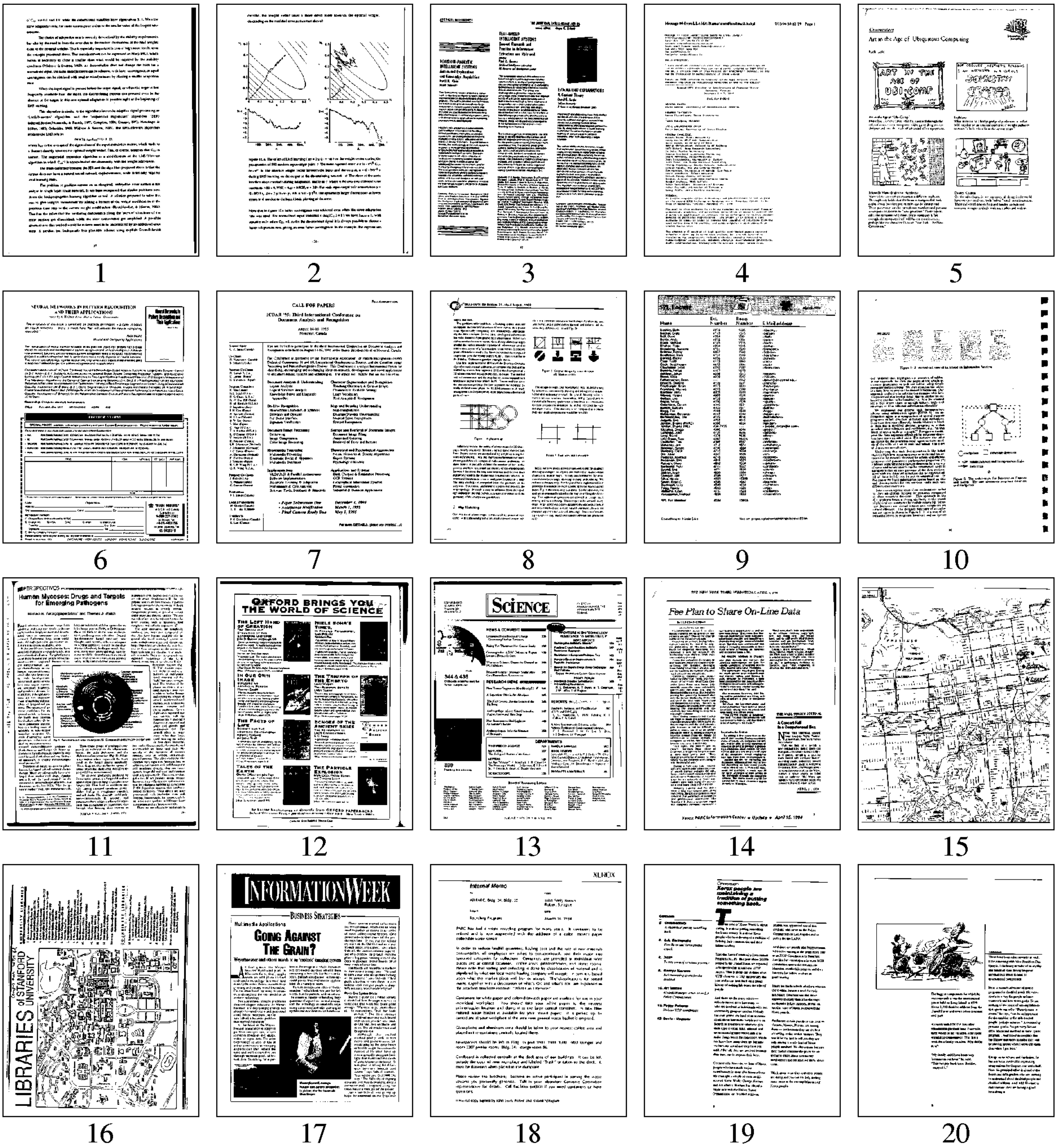

Figure 10: *20 images analyzed in this section*

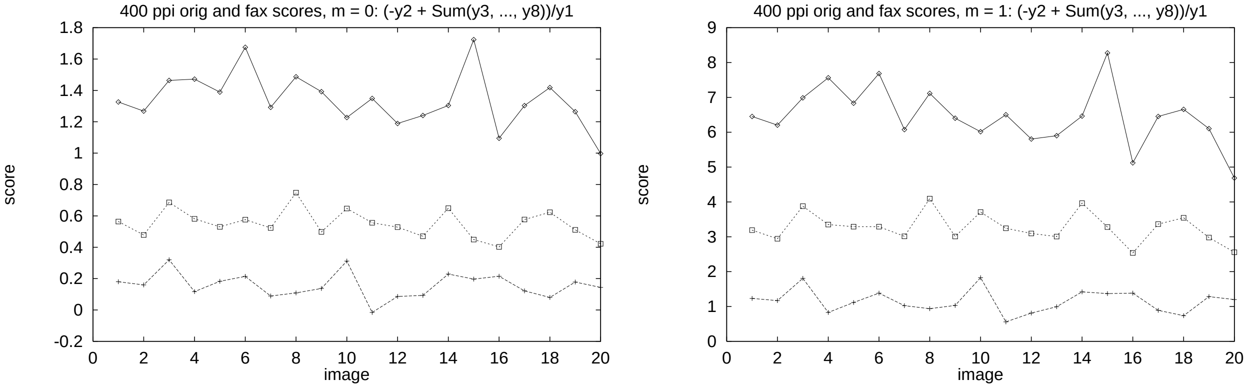

Figure 11: *Scores of 400 ppi original, fine and standard fax images, using (3). Zeroth and first moments of the 8-cc height histograms are used on the left and right.*

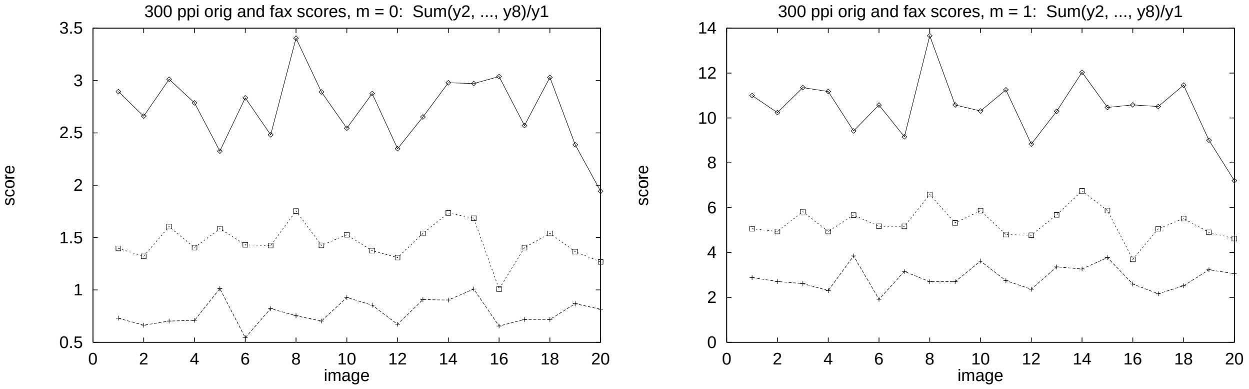

Figure 12: *Scores of 300 ppi original, fine and standard fax images, using (4). Zeroth and first moments of the 8-cc height histograms are used on the left and right.*

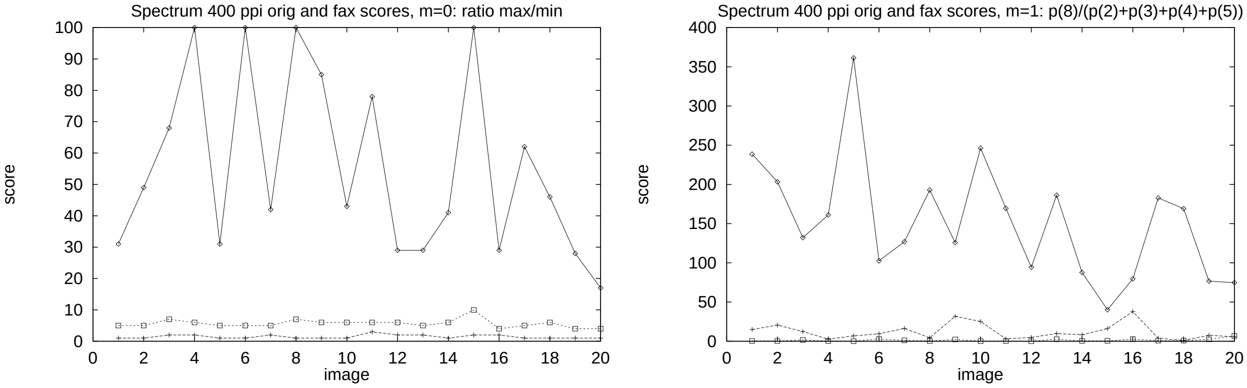

Figure 13: *Spectral power scores of 400 ppi original, fine and standard fax images, using the max/min ratio for m = 0 on the left, and using (7) for m = 1 on the right.*

$$s_{400}^{spect} = \frac{100p(8)}{p(2) + p(3) + p(4) + p(5)} \tag{7}$$

The result is particularly interesting because the fine fax spectrum scores are *lower* than those for the original image, for all images in the set. This inversion allows a threshold around 20 to make a clear separation between fine fax and standard fax documents.

5 Conclusions

Measurements of the vertical edge pixels have been used in various ways to identify an underlying low resolution structure in binary document images. Vertical, rather than horizontal, edge pixels have been chosen for illustration because standard fax has the lowest resolution in the vertical direction. The histogram of 8-cc heights is most sensitive to such structure, clearly displaying the periodicity. Note also the presence of single pixel 8-cc primarily in the original (top) of Fig. 1. By contrast, in all three images, 4-cc heights have many single pixel runs near corners where the boundary direction is changing between vertical and horizontal.

Images with many small components, such as fine stipples or halftones, have most vertical edge pixels near corners. There are very few longer runs, and as a result, the 8-cc height histogram is significantly different from that for text and graphics. Consequently, it is important to avoid such areas, and to select the most text-like regions for spectral analysis. This is done using morphological filters on a highly reduced version of the image, and counting the resulting pixels. Because the selection operations assume that most of the text on the page is horizontal, it is important to verify the orientation first, and efficient methods exist to do this[2].

Once a subregion is selected and the 8-cc statistics are accumulated, a discrimination function with a threshold is used to separate images with differing underlying resolutions. The discriminator must give a result that is (1) independent of the size of the image and (2) independent of the size and number of components in the image. Normalization is easily achieved by taking, in general, a ratio of linear combinations of histogram values or power spectral values. The most important distinction is between standard fax and higher resolutions, and this can be done either directly on the histogram or on its power spectrum. The power spectrum can be used as a narrow spectral filter to determine the underlying resolution. As an example, if the high/low resolution ratio is 4, most of the spectral power can be put into the 4 pixel mode, whereas for an image with high/low resolution ratio of 2, the 4 pixel mode will be depleted. With an appropriate discrimination function, this permits accurate separation between standard fax and fine fax images.

There are interesting and practical questions that have not been addressed here. For example, a relatively high quality fax scanner was used, on the assumption that the low frequency signature of cheaper scanners will be at least as easy to see. What image quality variations are possible with various low-resolution scanners and fax printers? How is the low-frequency resolution information changed with successive copies? And one possible application is to provide information for image restoration that will smooth boundaries and improve the appearance of degraded text. This may also be helpful as a pre-processing step for OCR.

6 Acknowledgments

I am grateful to Ron Danisewicz at Xerox ScanSoft for encouraging work on this interesting problem, and to Judy Bares for helpful discussions.

References

[1] D. S. Bloomberg, “Image analysis using threshold reduction,” *SPIE Conf. 1568, Image Algebra and Morphological Image Processing II,* pp. 38-52, San Diego, CA, July 23-24, 1991.

[2] D. S. Bloomberg, G. E. Kopec and L. Dasari, “Measuring document image skew and orientation,” *SPIE Conf. 2422, Document Recognition II,* pp. 302-316, San Jose, CA, Feb 6-7, 1995.

[3] S. V. Rice, F. R. Jenkins and T. A. Nartker, “The Fourth Annual Test of OCR Accuracy,” an addendum to *Fifth Annual Symp. on Document Analysis and Information Retrieval,* Las Vegas, NV, Apr 15-17, 1996.

[4] L. Vincent, “Fast opening functions and morphological granulometries,” *SPIE Conf. 2300, Image Algebra and Morphological Image Processing V,* San Jose, CA, July 1994.