HEWLETT PACKARD

Adaptive Thresholding for OCR: A Significant Test

Ray Smith, Chris Newton, Phil Cheatle
Personal Systems Laboratory
HP Laboratories Bristol

HPL-93-22
March, 1993

image thresholding, image binarization, character recognition

Although many adaptive thresholding algorithms have been published, few have ever been rigorously tested on a significant number of images. Also, few have been applied to OCR. The results of testing a new algorithm against three previously published algorithms are given. Performance for OCR applications is tested objectively by running an OCR system on the thresholded images of 350 A4 pages. The results show our new algorithm to be superior to the others tested. The differences measured would happen with probability less than 1% if they were occurring randomly.

Of the three existing algorithms tested, only Ots's algorithm achieves a performance better than choosing a fixed threshold at 50% of the greyscale range. This has an important implication for other algorithms derived from the commonly used model of images - simple objects corrupted by Gaussian noise. Algorithms tested only with synthetic images derived from the model can easily fail when applied to real world images.

Adaptive Thresholding for OCR: A Significant Test

1 Introduction

This paper describes the results of a comprehensive test of a new algorithm for adaptive thresholding for OCR. The new algorithm is compared to three previously published algorithms, by testing on 350 real document images from varied sources. The tests are made strictly objective by measuring the error rate of a good OCR system on the thresholded images.

Adaptive thresholding involves selecting a good greyscale at which to re-quantize an image to a small number of greyscales. We will only discuss binary thresholding algorithms - i.e. the objective is to create an image of just black and white. Adaptive thresholding algorithms operate either *globally* or *locally*. A global thresholding algorithm sets a single threshold for the entire page. A local algorithm varies the threshold across the page; often by partitioning the page into a set of non-overlapping tiles. The threshold is then held constant within each tile.

Numerous adaptive thresholding algorithms have been published over the years, and a large proportion of papers discuss thresholding for computer vision, rather than for character recognition. The vast majority give results for a small number of images, usually less than 6, and the images are often synthetic. The synthetic images commonly consist of a spot of one colour on a background of another colour with additive Gaussian noise. Unfortunately, this is a very poor model of the degradation modes encountered on real images of text. Very few papers report much in the way of a systematic analysis of the results against other algorithms.

Meanwhile, cheap "omnifont" OCR packages have been available for more than 5 years, and are not yet pervasive. The fact is that customers find that not all the documents they have can be read by these systems. For this reason, adaptive thresholding algorithms are also starting to appear in the marketplace. When a good evaluation procedure is followed, these algorithms can be of excellent practical value. For example, HP has introduced an adaptive thresholding algorithm as part of Accupage^{TM} which increases the fraction of documents which can be read accurately. These types of product have become very important in the OCR industry.

One possible reason why OCR has been so neglected by authors of adaptive thresholding algorithms is the technical difficulty of doing significant objective testing. There are several problems, some of which have only recently become easy to solve:

- Obtaining a large yet representative database of images and obtaining accurate correct text of the documents. The sources must span several organizations as there is a remarkable difference in the kind and quality of documents in different organizations.
- Storing the images on a sufficiently high capacity, high bandwidth device to enable high throughput for testing. (Our test database currently occupies about 7GB.)

- Obtaining an OCR system close enough to the leading edge which can be controlled fully automatically and finding enough compute power to get through the tests in a reasonable time. The exhaustive search discussed in section 2 involved processing around 80000 page images.

Reports of significant tests of OCR on real pages are just starting to appear.[4] It is our belief that such testing of the "industrial strength" of algorithms in this field is where industry can make a significant contribution by collaboration with academic researchers.

2 Why is adaptive thresholding needed for OCR?

Not every document in the office is an immaculate original printed on snow-white paper. The kinds of documents where adaptive thresholding can help are:

- Originals printed with very fine strokes. The thin strokes combine with the limited resolution of the scanner to produce pixels which can easily disappear when thresholded.
- Badly reproduced copies. Lightening of text can have the same effect as thin strokes. Darkening can have a corresponding effect on the gaps between strokes.
- Text printed on a coloured or halftoned background, or under a coffee stain. Such text can easily be lost by thresholding. Even if the text is retained, allowing halftone dots through can cause havoc with the OCR system.

Results from some documents which exhibit these effects can be seen in Figure 1.

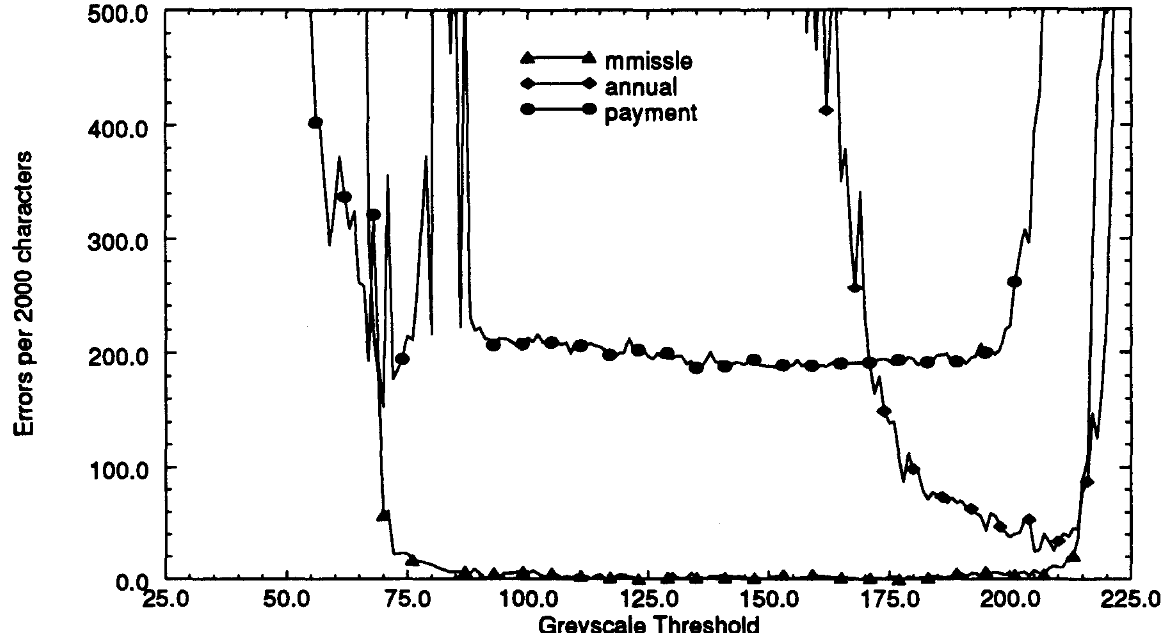

Figure 1 Exhaustive search showing error rate for each possible threshold.

Figure 1 shows a graph of error rate for each possible threshold between 32 and 224 of 3 documents in 8 bit greyscale. Document "mmissle" has a wide, flat-bottomed valley and is typical of the easiest possible page - almost any threshold will work.

Document "annual" is a faded photocopy and only a very narrow range of thresholds close to one end produces anything like a reasonable result. The fading is fairly uniform however, and a global threshold will suffice. Document "payment" is an example of a page where no global threshold will ever achieve a good result, and only a threshold which is variable across the page will be adequate. This is because it contains a halftoned box which cannot be picked out simultaneously with the body text. These documents are deliberately chosen examples of course. The results across a set of 350 documents are unmistakable however.

Figure 2 shows how much effect a good choice of threshold can have on the error rate over a large set of documents. The X-axis shows the error rate, and the Y-axis the percentage of documents in the test set with less than or equal to X errors per 2000 characters in the correct text. This cumulative graph is useful as it makes it easy to draw multiple lines on one graph and see a difference in the average performance. It is typically clearer to view than a histogram and provides more information than just an average error rate figure, which is so heavily weighted by the poor quality documents.

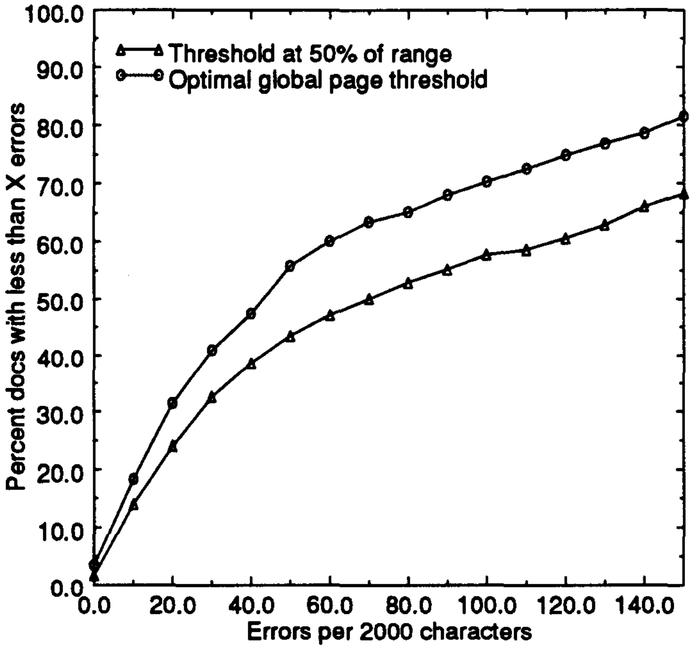

Figure 2 Cumulative graph showing number of documents with less than a given error rate

In Figure 2, the lower curve is the result of thresholding all the images at half the grey scale range. The upper curve results from an exhaustive search for the optimal global threshold for each page. This was found by running the OCR software on all the images thresholded at all possible thresholds. The conclusion must be that some kind of adaptive thresholding is important.

The curves show that our test set contains a significant proportion of documents which are difficult to recognize. They are not just difficult for our internal OCR research tool. One of the leading commercial packages achieves only 45% of the 350 documents with less than 50 errors per 2000 characters.

3 Previous Work

To make space for the all-important results, the reader is directed to the recent surveys[5][6] for detail on previous work. Most adaptive thresholding algorithms fit the general description of applying some kind of computation to the histogram of greyscales in an image or a tile of an image. The analysis may be applied to a simple histogram, or some techniques may be applied to improve the histogram first.[8] These include raising the peaks or deepening the valley by weighting the pixels or collecting a histogram of a subset of the pixels.

An alternative way of separating the classes is employed by Taxt, Flynn & Jain,[6] who use clustering to classify pixels into black and white and then treat thresholding as a classification problem. Our algorithm is related to both the traditional techniques and the latter, but unfortunately it is not possible to disclose details at this stage for commercial reasons.

4 Test Results

Our new adaptive thresholding algorithm, called mmB, has been tested on the same 350 documents which were shown in Figure 2. To give comparison with some existing methods we chose 3 popular algorithms due to Otsu[3], the moment-preserving method of Tsai[7] and the minimum error method of Kittler & Illingworth[1].

4.1 Global Thresholding

The results of all 4 algorithms to produce a global threshold are shown in Figure 3, together with the original curves which were shown in Figure 2.

It is clear from these curves that mmB outperforms the other three. In fact the mean difference between mmB and Otsu is 14.9 errors per 2000 characters and is statistically significant at better than the 1% level. Otsu's method itself achieves a mean difference of 54.1 errors per 2000 characters better than the 50% threshold, also significant at better than the 1% level. Of the existing algorithms only Otsu's method gets a better result than the fixed threshold at 50% of the greyscale.

The poor showing of the minimum error algorithm is due to a flaw in the process used to derive the criterion function. Where one of the classes has a very high, narrow peak in the histogram, and the valley floor contains a fairly flat distribution of pixels, the histogram is so far removed from the model of two Normal distributions, that the algorithm chooses a threshold "in the foothills" of the sharp peak. This choice of threshold has the effect of thickening and joining the text, and filling in the holes. Such distributions are typical of clean images of text.

4.2 Local Thresholding

None of the algorithms shown so far have approached the curve corresponding to the optimal global threshold. Although we would ideally expect to get close, the optimal global threshold can not be beaten without allowing the threshold to vary across the page. With local thresholding would expect to be able to get better results on a certain class of documents.

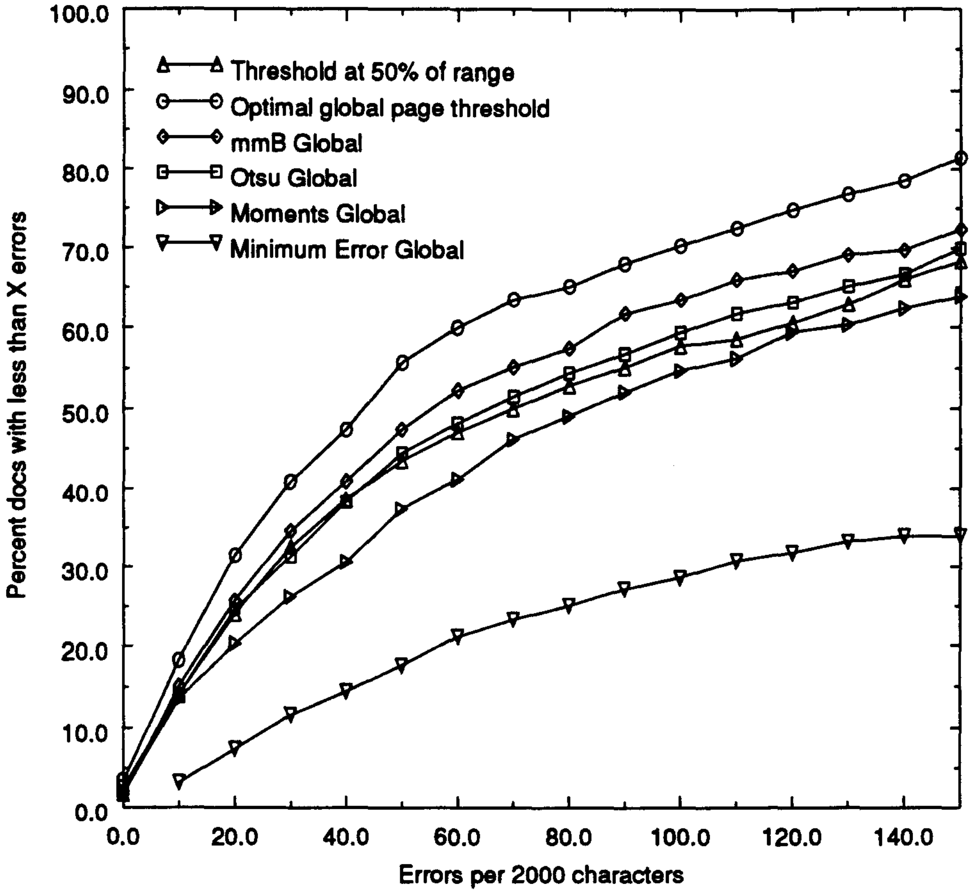

Figure 3 Results of global thresholding.

The final experiment therefore involved running mmB with tiling. These results are shown in Figure 4.

While the curves cross, making the advantage of tiling not quite clear cut, there is a distinct advantage for documents in the more commercially interesting area of <=40 errors per 2000 characters. The mean error difference over all 350 test pages between tiling and global thresholding is 10.4 errors per 2000 characters, with a significance of 2.5%

5 Conclusion

We have designed a new algorithm for adaptive thresholding, which can be applied to generate either a global or a locally variable threshold. The new algorithm has been tested against 3 previously published algorithms on 350 A4 page images, by running an OCR system on the output, which is sufficient to be able to show that mmB is superior to the best of the other algorithms with an average difference in error rate which would happen with probability less than 1% if the differences were occurring randomly.

Of the other algorithms tested, Otsu's algorithm is the best, being the only one able to produce better results than a threshold fixed at half the greyscale range. The minimum error algorithm has startlingly poor performance. This has an important implication for other algorithms derived from the commonly used model of images

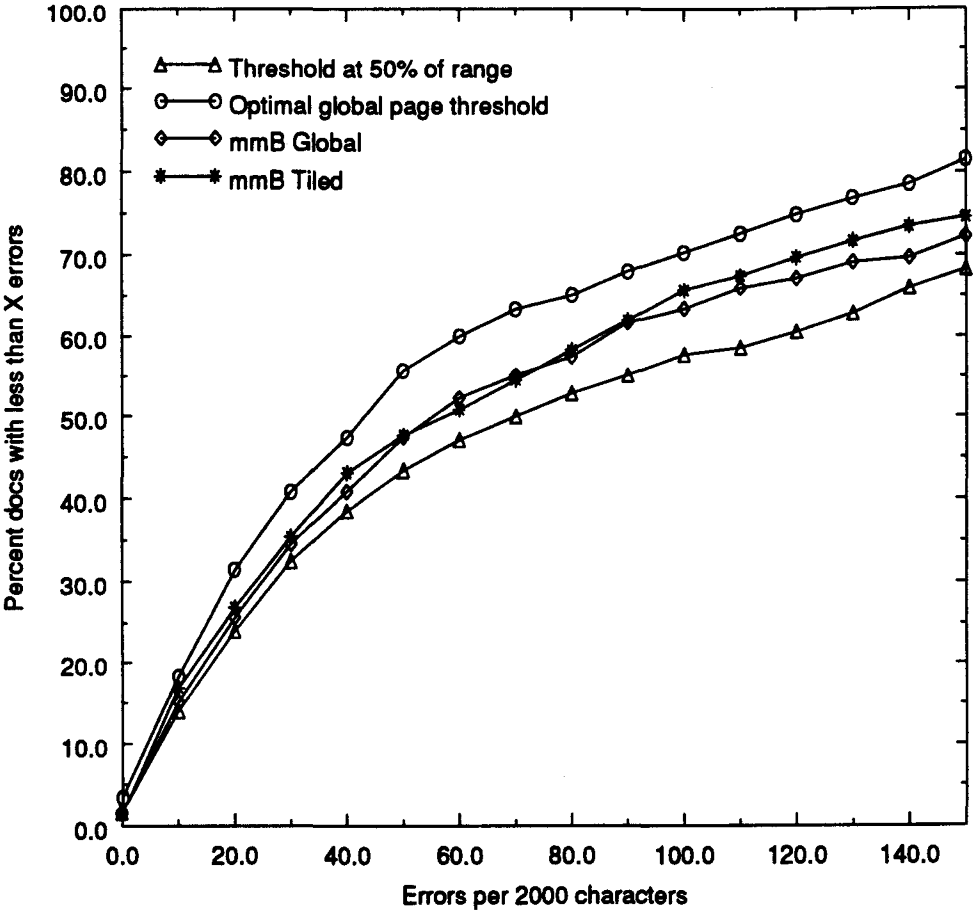

Figure 4 Comparison of global threshold with local tiled threshold.

– simple objects corrupted by Gaussian noise. Images of text exhibit distributions which are far removed from this model. We can therefore expect that algorithms produced from such a model, and tested only on synthetic images corresponding to the model, could fail dramatically when tested on a significant number of real images.

6 References

[1] Kittler J, Illingworth J. Minimum Error Thresholding. *Pattern Recognition* Vol. **19** No. 1 pp. 41-47. 1986

[2] Kittler J, Illingworth J. On Threshold Selection Using Clustering Criteria. *IEEE Trans. SMC.* Vol. **SMC-15** No. 5 pp. 652-655 Sept./Oct. 1985.

[3] Otsu N. A Threshold Selection Method from Gray-Level Histograms. *IEEE Trans. SMC.* Vol. **SMC-9** No.1 pp.62-66 Jan. 1979.

[4] Rice S.V, Kanai J, Nartker T. A Report on the Accuracy of OCR Devices. *University of Nevada Internal Report.* Information Science Research Institute, Las Vegas, USA.

[5] Sahoo P.K, Soltani S, Wong A.K.C. A Survey of Thresholding Techniques. *Computer Vision, Graphics & Image Processing* **41**, pp. 233-260 1988.

[6] Taxt T, Flynn P.J, Jain A.K. Segmentation of Document Images. *IEEE Trans. PAMI,* Vol. **PAMI-11**, No. 12 pp. 1322-1329 Dec. 1989.

[7] Tsai W. Moment-preserving Thresholding: A new approach. *Computer Vision, Graphics & Image Processing* **29**, pp. 377-393 1984.

[8] Weszka J.S, Rosenfeld A. Histogram Modification for Threshold Selection. *IEEE Trans. SMC.* Vol. **SMC-9** No. 1 pp. 38-52 Jan. 1979.