Evaluating Binarization Techniques for Optical Character Recognition

Donald B. Curtis; The Generations Network, Inc.; Provo, Utah

Abstract

Keying data from digital images is time-consuming and costly and is subject to human error. Archivists are often limited in their content production by their keying budget and by the cost of auditing keyed index data. To increase both quality and production, one alternative is to OCR-process machine-printed documents. Today's OCR technologies are only as good as the bitonal (black and white) documents they process, so a high-quality, high-performance binarizer (a tool to convert color or grayscale images to bitonal ones) is critical to the success of OCR-processing historical records.

Discussed are the challenges binarizers face, the methodology used to test a new binarizer, and the results of the new binarizer, compared with a small sampling of other binarization technologies. Not discussed are the proprietary details of the new binarization algorithm.

Introduction

Binarizers are typically bundled with image scanners, OCR software, and image processing tools. Binarizers may face many challenges in determining how to convert images to bitonal ones. Different binarizers handle these challenges in different ways. Typically, binarizers perform well with one type of content but may perform poorly with another.

Organizations often have a goal to achieve a consistent quality level for all their indexing projects. Because using different binarizers results in differing OCR quality, it may be desirable to use the same binarizer for all documents associated with a project, regardless of which scanners are used to capture those documents. Thus, a methodology for determining which binarizer to use for a particular project is important.

Binarization Challenges

Here are a few examples of the challenges that a binarizer may encounter.

Lights and Shadows

Images may contain bright and/or dark areas that have been caused by inconsistent lighting, as in Figure 1, or because of damage to the original document, as in Figure 2.

Noisy Background

Images may have a noisy background that confuses some binarizers, as in Figure 3.

Bleed-through

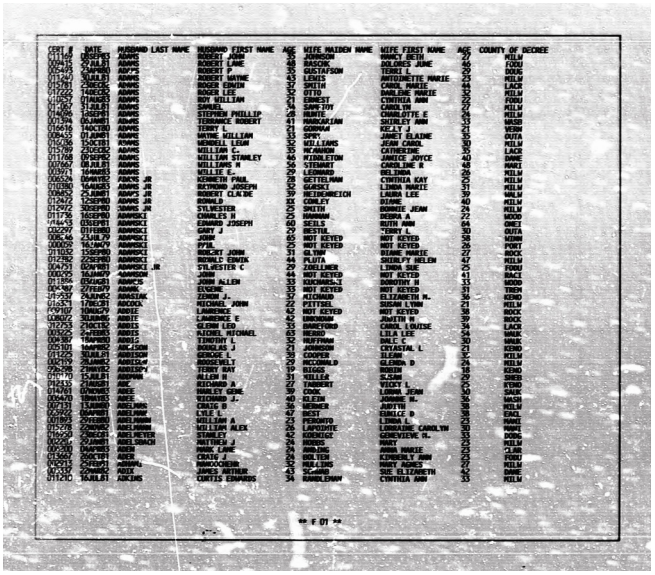
Images may show data from the opposite side of the page, as in Figure 4.

Ink Dispersion

Papers may be dark and may allow the ink to disperse, resulting in images with blurry, low-contrast text, as in Figure

Figure 1. *Shadowed Corners*

Figure 2. *Mildew Damage*



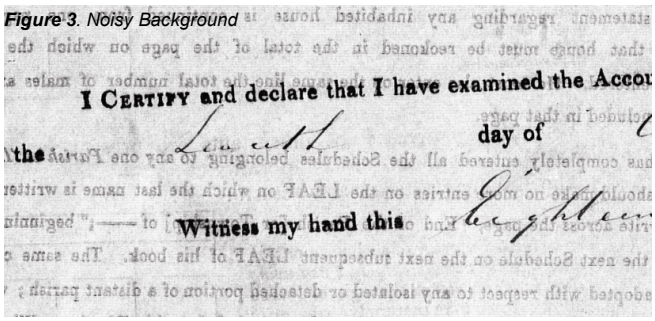


***Figure 3**. Noisy Background*

***Figure 4**. Bleed-through*

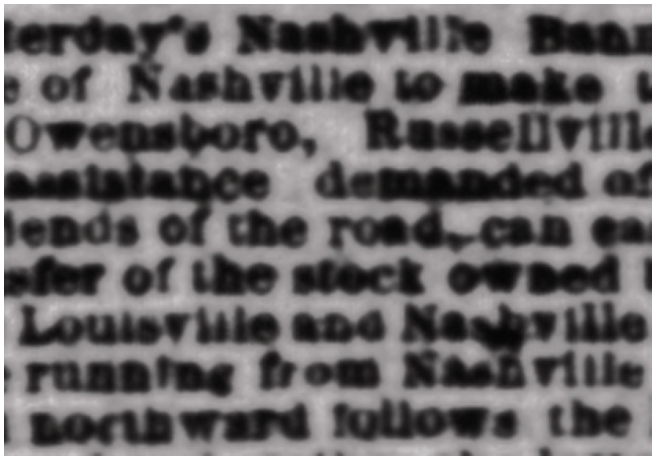


***Figure 5**. Ink Dispersion*

Choosing a Binarizer

When defining the requirements for an OCR project, one of the considerations should be which binarizer to use. To make that decision, a number of factors could be considered, such as:

- What type of content is in the project; what challenges will the binarizers face? If the images are challenging to binarize, then a binarizer that can meet those challenges should be selected.
- What digital images will be presented to the customer? If color or grayscale images are to be presented, then the documents should be scanned that way and binarized later to avoid the cost of scanning the documents twice.
- Is the speed of completing the project more important than the quality of the images and text? If so, then the images should be binarized by the scanner during the scanning process.
- Is quality of utmost importance? If so, then several binarizers should be tested with a representative sample of documents to determine which binarizer produces results with the greatest quality. Using that binarizer will provide a consistent, high quality level for the project.

Evaluating the Quality of a Binarizer

With the completion of a new software binarizer, there was a desire to measure its quality. For the purpose of this assessment, binarization error is measured in terms of the OCR results. A set of source images with a defined set of textual data represented by those images is used for the evaluation. Each deviation in the text produced by a specific OCR engine from the actual text, using the bitonal images produced by the binarizer, is considered an error. The binarizer that results in the fewest OCR errors is considered to have the greatest quality.

The following steps were used to compare the quality of the new binarizer with that of a few others:

1. A book was obtained for which the actual text of the book was available. This was used as control text to measure the OCR results against. The tested 85 pages of the book contain 220,421 characters.
2. The pages of the book were scanned as grayscale images.
3. The grayscale images were converted to bitonal images using the new binarizer.
4. The bitonal images were OCR-processed.
5. The OCR-produced text was compared with the actual text on a per-character basis. Each difference was recorded as an error.
6. The process was repeated, using the previous software binarizer.
7. Binarized images were created from the same book using two scanners from different vendors and steps 4 and 5 were repeated using those images.

The OCR errors were categorized into three types: *added* characters, not in the actual text; *changed* characters, different from the actual text; and *deleted* characters, present in the actual text but not in the OCR-produced text. Table 1 documents the OCR errors resulting from the bitonal images produced by each of the four binarization sources.

Table 1. OCR Errors from Four Binarization Sources

Binarization Source	Chars Added	Chars Changed	Chars Deleted	Total Errors
New Software	43	20	29	92
Old Software	1128	262	80	1470
Scanner 1	2529	73	102	2704
Scanner 2	44	34	36	114

This test verified that there is a measurable difference in quality between the four binarizers, in terms of the number of OCR errors, even for a contemporary book with few binarization challenges.

The difference in quality between binarizers is more pronounced when the binarizers are given historical content that is challenging to binarize, as Figure 6 demonstrates with two binarizations of a mildew-damaged document. In this example, the new binarizer was able to distinguish the text from the damaged, darkened portions of the image better than the other binarizer. The binarizer chosen for a particular OCR project can have a significant impact on the OCR accuracy for that project.

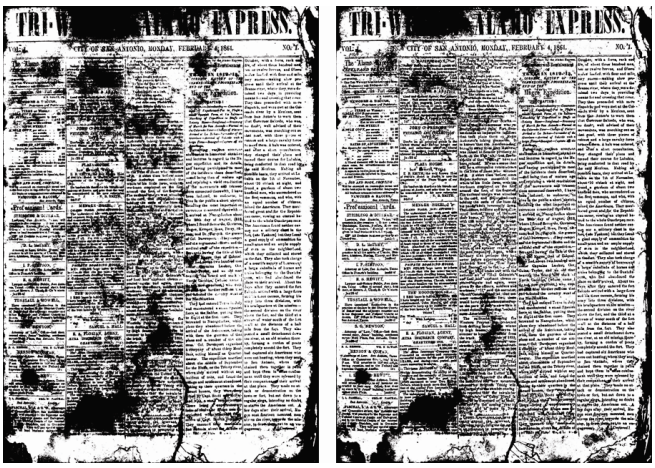


***Figure 6**. Comparison of Old Binarizer Results (left) with New Binarizer (right)*

If canonical text is not available or is too costly to create for evaluating OCR-processed text, how does one test binarization results? A solution to this problem is derived from the recognition that for evaluation purposes, it is less important how the resulting text compares to a correct set of text than how the resulting text for each binarizer compares with the results from the other binarizers. Thus, each set of resulting text from OCR-processing the bitonal images from each binarizer can be compared with each other, and only the differences need be examined. The binarizer associated with the differences that are most correct is the binarizer with the greatest quality.

Conclusion

Many binarization methods are available. Different methods yield different OCR accuracy results, depending on the characteristics of the source images and the strengths and weaknesses of the binarization methods employed.

To increase OCR accuracy, careful consideration should be made as to which binarizers are used. Selecting the "best" binarizer for a particular project can make a significant difference in the quality of the resulting data.

Author Biography

Donald B. Curtis received both his BS in computer science (1984) and his MS in computer science (1985) from Brigham Young University. Since then he has worked for AT&T Bell Laboratories, WordPerfect Corporation, and Microsoft Corporation. Don is currently employed by The Generations Network as a development architect, where he focuses on image processing and digital preservation technologies.