Binarization for OCR
Abstract for 6th Annual Workshop on Technology for Family History and Genealogical Research

Submitted by Donald B. Curtis, MyFamily.com

Keying data from digital images is time-consuming and costly and is subject to human error. Genealogical content providers are often limited in their production by their keying budget and by the cost of auditing the keyed data. To increase both quality and production, one alternative is to OCR machine-printed documents. Today's OCR technologies are only as good as the bitonal documents they processes (garbage-in, garbage-out), so a high-quality, high-performance binarizer is critical to the success of OCR'ing genealogical data.

Genealogical data sources are, more often than not, scanned from low-contrast microfilm or from old, worn/faded/yellowed books and documents. The paper is often thin enough that bleed-through from the opposite side of the page is common. These digital images present a huge challenge for a binarizer. To distinguish faded strokes from a murky background and yet ignore bleed-through text is more than most binarizers can handle.

Some adaptive binarizing algorithms do well at getting the text to come through, but in the process also blacken many pixels that, though darker than their neighbors, are just noise in the background. This noise is often interpreted by OCR as punctuation or other characters and thus dilutes the quality of the results and can make automatic processing of punctuation-delimited text more error prone. Despeckling or other noise-removal algorithms can help but sometimes do more damage than good, removing legitimate punctuation or dots over characters.

Convolution-based binarizers can be very slow, sometimes taking minutes to binarize a single image. Just the time to set up such a binarizer, testing the parameters for the optimal results for a project, can be taxing.

Relying on the binarizer of a scanner has its own set of problems. Tuning such binarizers for each project may not be possible, and the variety of binarizers present with different scanners makes the quality of the results inconsistent and dependant on the scanner chosen. If images are to be presented to customers in grayscale or color, the images must be scanned twice with potentially differing coordinates for the content within the image.

Developing a binarizer presents additional challenges. How does each algorithmic change or modification to a controlling variable affect the quality of the results? How does one decide when the optimal algorithm has been achieved?

At MyFamily.com, a new binarizer has been developed. This paper and its corresponding conference session presentation explore the problems associated with binarizers, document how the new binarizer was tested, and show the results of the new binarizer, compared with previous technology. Not discussed are the details of the binarizing algorithm itself.