Zoning Tabular Documents

Heath E. Nielson **William A. Barrett**
Department of Computer Science, Brigham Young University

Extended Abstract

With improvements in scanning technology, and the increasing connectivity and availability of the Internet, millions of genealogical documents have been made available. However, in order to exploit the content of these documents, the granularity of the indexing must move from the image level to individual fields within the document. Field-level addressing gives us a means of partitioning the document into meaningful and relevant components with the attendant benefits of speed and the economy of data transfer and storage. Rather than transferring and searching through the entire document, selected fields could be transmitted instead. This allows users to focus only on the information important to them.

Segmentation of a document into its respective fields also allows each field's contents to be contextually analyzed. For example, a field that contains printed text would be sent to an OCR engine. Fields containing handwriting would be stored for subsequent semi-automated, user-assisted interpretation or pattern matched indexing. To perform automated field-level indexing and addressability, automated zoning techniques are needed to partition the document and identify the location and content of regions and fields. We have developed a zoning algorithm which allows regions of interest in a document to be identified and partitioned, and we also propose a method to determine the content of these regions.

Where we can anticipate multiple instances of documents which have the same geometric layout (such as successive images of census-like records on a roll of microfilm), we would like to exploit the intra and inter-document consistency in automatically discovering the layout and generating a flexible template (Fig. 3) of that layout. By establishing consensus across multiple documents and fields within a document, we intend to greatly increase the signal-to-noise ratio (SNR) and produce a more robust statistic with respect to the derived template.

Partitioning a tabular document is based on the assumption that different regions within the document are delimited by lines. By searching for and identifying these lines, the document can then be partitioned and analyzed. We propose a relatively simple approach to finding lines by identifying the signature created by the line within the image's horizontal and vertical profile. Profiles have the property of representing lines in a document as peaks even where the line may be broken or intersects other lines or writing. For an image with width M and height N, the horizontal and vertical profiles are defined to be:

$$p_h(y) = \Sigma_{i=0}^{M} image(i, y)$$
$$p_v(x) = \Sigma_{i=0}^{N} image(x, i)$$

These profiles are convolved with a matched filter to increase the SNR of the signatures created by the lines.

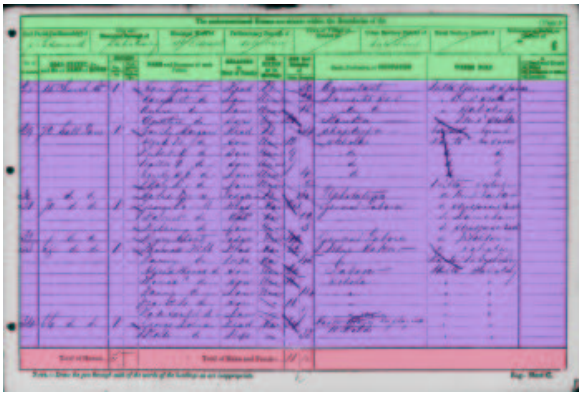

Figure 1: The three components of a tabular document: Header, Body, and Footer

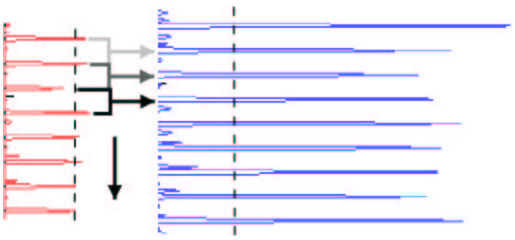

Figure 2: A 2-prong probe sums input from the filtered profile (red) to increase the signal-to-noise (SNR) of lines in the body (blue). This allows lines to be identified that would have been missed in the original profile.

Profiles of an entire image can only be used to find global lines within the document. Local lines would not produce a large enough peak in the profile to correctly segment it against the surrounding noise. In order to find these local lines within a profile, the profile must be constrained to specific areas of the image. In order to identify local lines, we propose dividing the document into several logical sections corresponding to similar geometric layouts found within the document. For now we will assume 3 sections: the header, body and footer (Fig. 1).

To split the document into its component parts we first identify the location of the body within the image. The body of a tabular document represents the largest of the 3 sections and presents the most intra-document consistency. Rows are equally spaced and the columns remain the same throughout the body.

Body identification is performed using a frequency analysis of the horizontal profile. By exploiting the consistent line spacing in the body of the document, the equally spaced rows in the body produce an identifiable frequency which can be used to identify line spacing between rows. To identify the locations of the rows in the body, we increase the SNR of the horizontal profile by using a 2-prong test (Fig. 2) defined by the following algorithm:

$$C(i) = \Sigma_{j=i-\delta}^{i+\delta} p_h(j) + \Sigma_{j=i-\delta}^{i+\delta} p_h(j-w)$$

Where δ is some small integer value and w is the spacing between rows.

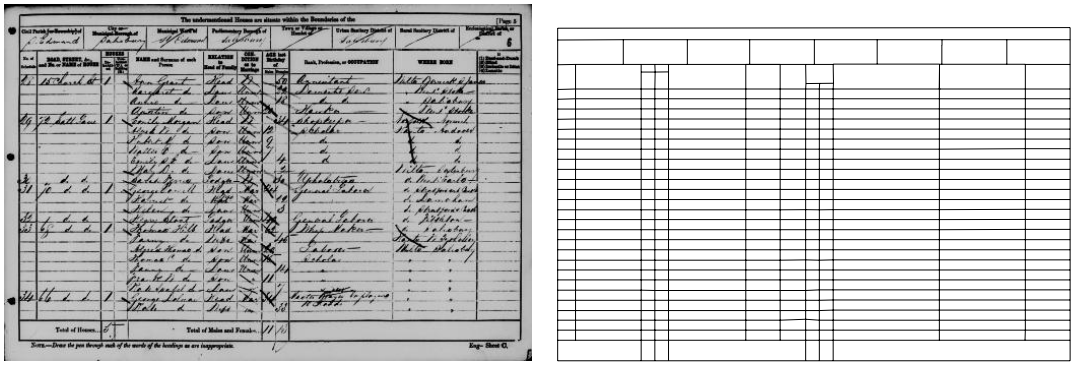

Figure 3: Source image and its corresponding template

Having identified the body of the document, the header and footer of the document are subsequently labeled as anything above or below the body respectively.

With the document split into its 3 component parts, each section is, in turn, analyzed for vertical lines. A mesh of the entire document is created from the horizontal and vertical lines derived from each of the 3 sections of the document independently. To account for geometric distortion, the mesh is snapped to the lines within the document.

To exploit the inter-document consensus we propose combining the meshes generated from each document. This is accomplished by voting on line positions as found in each document. Line segments with a low vote count are discarded. As additional documents of the same geometric layout are zoned and their meshes are combined, we are able to generate a flexible template describing the geometric layout of the document (Fig. 3). Spurious zoning errors resulting from images of poor quality can be eliminated by combining the layout information from multiple documents. Once a robust template has been created, subsequent documents can be zoned simply by snapping the document template to the image.

With the creation of the document template, the content in each region is classified into one of three classes:

1. Empty
2. Printed Text, or
3. Handwriting

Classification occurs by analyzing the content of each region from multiple documents. Regions containing printed text will contain the same printed text from document to document (Fig. 4), while the content of regions containing handwriting will vary.

Being able to determine the regions of a document and automatically identifying the content of those regions frees us from the repetitive task of finding all the regions interactively. For large batches of digital documents, this can result in increased speed and better consistency.

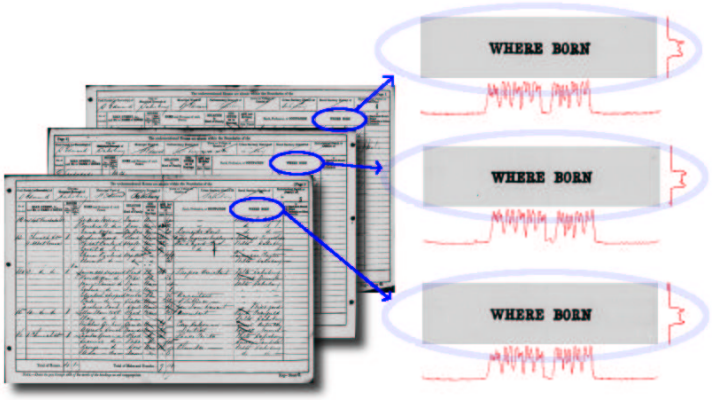

Figure 4: The horizontal and vertical profiles of the same cell in the table from three separate documents