Automatic Zoning of Digitized Documents

Heath Nielson and William A. Barrett

Recent improvements in scanning technology have made available (over the web) millions of scanned genealogical documents. 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. Being able to search or browse through individual fields of a document rather than the whole image is the first step in indexing and understanding the content of those fields (e.g. name, age, sex, location, etc.). In addition, field-level addressing gives us a means of partitioning the document into meaningful, and relevant components with all of the other 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. Sending portions of an image also provides for a quicker response time allowing the user to efficiently locate the information they are searching for. 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 or user-assisted interpretation. To perform automated field-level indexing and addressability, automated zoning techniques are needed to automatically partition the document and identify the location and content of regions and fields. We have developed a zoning algorithm which allows rectangular regions of interest in a document to be identified and partitioned. We also propose methods to determine whether these regions contain handwriting or printed text.

Identifying regions within a document is based on the assumption that such regions 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 identifying lines by identifying the signature created by the line within the image's profile. By taking the horizontal and vertical profiles of an image we attempt to identify those signatures created by the lines. In order to facilitate better recognition of these lines' signatures, a matched filter is created. The filter is dynamically created and can be adapted to the specifics of each document. By convolving the matched filter with the profiles, signatures similar to the matched filter are augmented while everything else is dampened. Once the initial pass has been made, the document is cut up into rectangular regions. Each line segment is analyzed at the local level where one of three decisions can be made:

(1) either a line does exist and the correct location has been identified,
(2) a line exists but in a slightly different location, or
(3) no line exists there at all and neighboring regions are merged.

Additionally, the region itself can be scanned for any additional lines which may not have shown up on the first pass. The algorithm continues to be applied recursively to smaller, more localized regions of the document to dynamically determine the document's layout and output the best partition.

Having determined the document's layout, that layout could be applied as a template to additional documents possessing the same layout This has two main advantages:

(a) It speeds up batch document processing. Since the document layout is the same, repetitive work can be eliminated.
(b) Accuracy can be increased. Documents of poor quality may be difficult to zone. By looking at several documents, the resulting templates can be averaged together to create a more accurate result.

Once the document has been successfully partitioned into regions, the content of each region is labeled as belonging to one of three classes:

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

This is done using features similar to those used in identify lines: horizontal and vertical profiles. We first attempt to determine if the region is empty. If the region is not empty, we then try to determine whether the region contains printed text. Since fixed-width fonts produce a regular, box-like structure, the use of matched filters can also be applied here to identify the printed text. Failing the first two tests, it is assumed that the region either contains handwriting or something requiring user-assisted identification.

Being able to determine the regions of a document and automatically identify the content of those regions frees us from the repetitive task of finding all regions interactively. For large batches of digital documents, this can result in increased speed and, potentially, better consistency. This allows for a quicker turnaround from the time a document is digitally acquired to being placed into a searchable database.