2003-08-01

Consensus-Based Table Form Recognition

William A. Barrett
william_barrett@byu.edu

Heath E. Nielson

Original Publication Citation

H. Nielson and W. A. Barrett, "Consensus-Based Table Form Recognition," 7th International Conference on Document Analysis and Recognition (ICDAR), pp. 96-91, Edinburgh, Scotland, August, 23.

BYU ScholarsArchive Citation

Barrett, William A. and Nielson, Heath E., "Consensus-Based Table Form Recognition" (2003). *Faculty Publications*. 484.
https://scholarsarchive.byu.edu/facpub/484

Consensus-Based Table Form Recognition

Heath E. Nielson
Brigham Young University
heath@cs.byu.edu

William A. Barrett
Brigham Young University
barrett@cs.byu.edu

Abstract

Zoning documents increases the resolution of indexing from the image level to the field level. A line-delimited tabular document forms a well defined series of regions. However, as image quality decreases, accurate zoning becomes increasingly difficult. Given a sequence of documents with the same layout, we present a robust zoning method which exploits both intra- and inter-document consensus to form a more accurate combined result (template) that can be applied to any other document with the same layout.

1. Introduction

With improvements in scanning technology, storage capacity, and Internet connectivity, millions of digital documents are becoming accessible on line. 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 allows the document to be partitioned into meaningful and relevant components. Rather than transferring and searching through the entire document, selected fields can be transmitted instead, targeting only relevant information.

Segmentation of a document into its respective fields allows field contents to be contextually labeled as printed text or handwriting. Text could be sent to an OCR engine and handwriting stored for subsequent semi-automated, user-assisted interpretation or pattern matched indexing. To perform automated field-level indexing, automated zoning techniques are needed to partition the document and identify the location and content of regions and fields.

Many current zoning techniques attempt to completely and accurately segment a single image at a time. With images of poor quality, zoning accuracy suffers. Where we can anticipate multiple instances of the same document, many methods fail to take full advantage of the combined features of each document. We present a novel method for combining geometric information extracted from multiple documents. By making use of both intra- and inter-document consensus we can construct a robust geometric layout (template) of a document that is more accurate which can be extracted from a single document and that can be applied to successive documents of the same layout.

2. Background

There has been much work in document understanding in general and zoning in particular [8, 14]. There are typically two basic approaches: Top-down, or model driven, approaches and bottom-up, or data driven, approaches. Some hybrid approaches contain elements of both.

Top-down approaches divide a document into its component parts through a divide-and-conquer strategy, starting at the global level and recursively subdividing large areas into smaller ones. The recursive X-Y cut originally proposed by Nagy [9] is representative of this. More complex algorithms have been developed [3, 2, 1, 6, 10] which employ rules to determine how the document is to be divided. The most common feature used to subdivide a document is either a line or the empty space between rows commonly referred to as a "white stream". Profiles are used extensively to find these delimiters in the document.

Bottom-up techniques generally rely on a connected component strategy building a document hierarchy starting at the character level and working up to word, line and paragraph [7, 5, 11, 12, 13]. Wavelets are also used for table segmentation and identification [16, 15].

3. Consensus-Based Zoning

In our consensus-based zoning algorithm, partitioning a tabular document is based on the assumption that different regions within the document are delimited by lines (Fig. 1). By identifying these lines, an editable mesh representing the geometric layout of the document is created. Individual meshes are combined to form a single mesh (template) through consensus. Each region of interest (ROI) in the template is classified according to its content: printed text, handwriting, or empty. The template is then used to zone new documents of that layout.

Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR'03)
0-7695-1960-1/03 $17.00 © 2003 IEEE

3.1. Candidate Line Identification

Peaks in horizontal and vertical profiles are used to identify lines in a document, even where the line may be broken or intersects with other lines or writing. For an image with width M and height N, the horizontal and vertical profiles are defined to be

$$p_h(y) = \Sigma_{i=0}^{M} image(i, y) \qquad (1)$$
$$p_v(x) = \Sigma_{i=0}^{N} image(x, i) \qquad (2)$$

A matched filter is applied to each $p(x)$ to localize peaks by increasing the signal-to-noise ratio between line peaks and the remainder of the profile. The filter $f(x)$ is created by sampling N peaks from the profile.

Each profile is convolved and normalized as follows:

$$\begin{aligned} c(x) &= p(x) * f(x) & (3) \\ p_f(x) &= \frac{S}{c_{MAX}}(MAX(c(x), 0) + 1) & (4) \end{aligned}$$

where c_{MAX} is the largest value in $c(x)$ and S defines the scaled range of $p_f(x)$.

Any peak exceeding a preset threshold is identified as a candidate line.

3.2. Region Splitting

The image is split horizontally into three separate logical sections corresponding to regions of similar geometric layout: the header, body, and footer.

The body is assumed to always be present in a document while the header and footer regions are optional. The body, which presents the most intra-document consensus, is identified first. The Fourier transform of the horizontal profile produces a conspicuous peak frequency that identifies the spacing between rows in the body. Pairwise matching of lines satisfying this row spacing identifies the body as the largest group. Any candidate lines found within the body not satisfying the body row spacing are labeled as false positives and removed. Lines above the body are labeled as header and lines below as footer.

With the document split into its three component parts, each section is analyzed for vertical lines. Using the same process discussed in Sec. 3.1, vertical and horizontal lines are combined to form an editable mesh (Fig. 1).

3.3. Local Snapping

Although the mesh consists of strictly horizontal or vertical line segments, the image itself often manifests geometric distortion due to imaging optics or the acquisition process. To make sure that the lines in the mesh correspond to the lines in the image, each line segment's position is adjusted, or "snapped", to the location in the image presenting the strongest line support (Fig. 2).

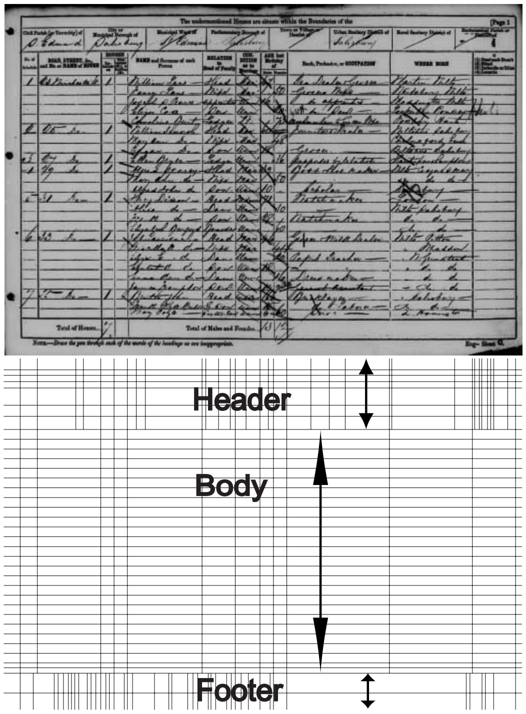

Figure 1. Initial mesh created from above image.

Each column and row is snapped by identifying the line segment maximizing $p(x)$ over the interval defined by the segment. It is labeled the "seed edge", with its two vertices labeled pivot vertices (v_p). Beginning with one pivot vertex and moving away from the seed edge, the next adjacent vertex becomes the snap vertex (v_s) which is snapped to the location maximizing

$$s(x) = e^{\frac{-(p_g - x)^2}{2\sigma^2}} fp_l(x) \qquad (5)$$

The first term is a Gaussian weighting where p_g represents the global line position. Sigma defines the width of the Gaussian and is $\frac{1}{2s_r}l$ where l is the snap neighborhood height for rows or width for columns and s_r represents the snap resistance. $fp_l(x)$ is the filtered local profile over the line segment $(v_p, \frac{1}{2}(v_s + v_{s+1}))$ where v_{s+1} is the next adjacent vertex.

Using a Gaussian weighting gives the vertex flexibility to adjust its position to locations close to the line's global position, but becomes increasingly restrictive the farther away it gets. By adjusting the value of s_r we can restrict how far we will allow snapping to occur from the global position.

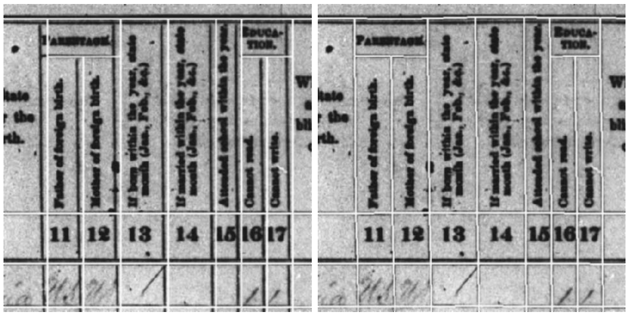

Figure 2. Unsnapped mesh (left), local snap (right).

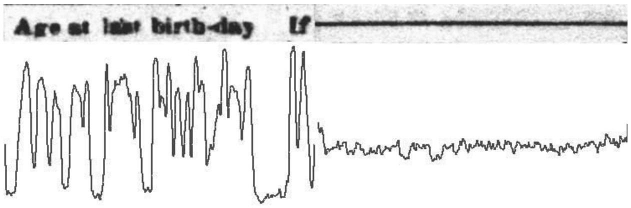

Figure 3. Vertical profile of a false positive (left) and an actual line (right).

The location maximizing $s(x)$ is clipped to the snap neighborhood to prevent overlapping with adjacent rows or columns and v_s is moved to that position. v_p now becomes v_s and v_s advances to the next adjacent vertex in the line. This process continues until there are no more vertices to snap in that direction. The algorithm repeats with the remaining pivot vertex, moving in the opposite direction.

3.4. False Positive Detection

Peaks found in a profile often correspond to items other than actual lines in the image, primarily rows or columns of printed text. In addition, candidate lines initially extend through the length of the document which may not be the case. These false positives need to be identified and removed from consideration as lines.

To identify the false positives we examine the perpendicular profile of each line segment in the header and footer (Fig. 3). False body lines were already identified in Sec. 3.2. For a line of text, alternating characters and white-space create a high amount of variability in the profile compared with an actual line.

The line profile (lp), a one pixel wide profile of the pixels under the line segment scaled by a neighborhood min and max, is generated. If $mean(lp) < T_m$ or $variance(lp) > T_v$ then the line is labeled a false positive and removed.

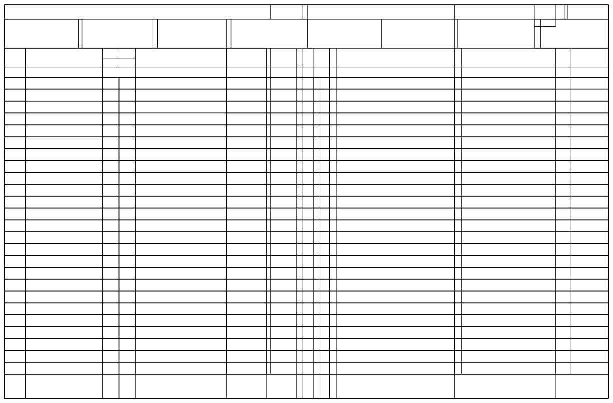

Figure 4. The combined template.

Having removed the false positives, any remaining line segment which doesn't contribute to a closed ROI is removed.

3.5. Template Creation

To exploit the inter-document consensus, we combine the meshes generated from each document. To combine two meshes, m_1 and m_2, we arbitrarily choose to merge m_2 into m_1 using profiles (Sec. 3.1) of "images" of m_1 and m_2. First, by correlating the profiles of m_1 and m_2, m_2 is moved to the location of highest correlation so that it overlaps m_1. Then, an equivalency table is created matching the rows and columns in m_2 to the rows and columns of m_1. The lines from m_2 are merged into the corresponding lines in m_1 with each line segment maintaining a count of the number of times it was merged or votes received. If there was no match for a column or row in m_2, the line is added to m_1.

When all meshes have been combined, a simple Otsu threshold is applied to all line segment votes to remove light line segments (i.e. those with a low vote count) (Fig. 4).

3.6. Global Snapping

With a robust template of the document's geometric layout, subsequent images can be zoned by snapping the template to the documents in subsequent images. Identifying the document's position is accomplished by correlating the horizontal and vertical profiles of the image with the profile of the template. The point of highest correlation identifies the location of the document.

The horizontal and vertical profiles of the image are generated using the approach discussed in Sec. 3.1. The template's profile is created by establishing a peak at every line

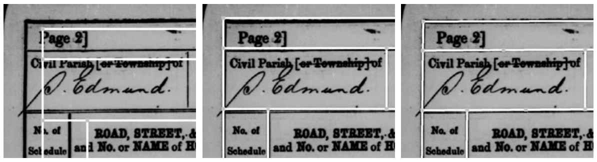
Page 2]
Civil Parish [or Township] of
No. of Schedule
ROAD, STREET, &c. and No. or NAME of H

Figure 5. White template on new document (left). Global snap (middle). Local snap (right).

location. The peak's intensity for line x is determined by

$$I(x) = \frac{v}{NV} l(x) \tag{6}$$

where v is the number of votes received for each segment in the line, N is the number of segments in the line, V is the total number of votes, and $l(x)$ is the length of the line.

The peak falls linearly from the peak position $I(x)$ to zero corresponding to the estimated width of the line.

Scale is also an issue, especially on large, high resolution images. To find the optimal scale, the template's signatures are generated at several different scales. Beginning with the scale range $[0.96, 1.04]$ at increments of 0.01, the scale with the maximum peak (s) is identified. The scale is further refined over the range $[s - 0.005, s + 0.005]$ at increments of 0.001.

With the optimum scale identified, the offset (dx, dy) is

$$dx = P_v - s_v \tag{7}$$

$$dy = P_h - s_h \tag{8}$$

where P is the peak location in the correlated profile and s is the size of the profile for vertical and horizontal profiles.

With the optimum scale and offset, the template is snapped into position and the mesh undergoes a local snap as discussed in Sec. 3.3. To be less susceptible to noise, local snapping at this stage is more restrictive with an increased s_r value. This is to prevent snapping to neighboring signals, such as a text line, which might prove stronger than the actual line.

3.7. ROI Content Classification

With the creation of the document template, the content in each field or ROI is classified into one of three classes: empty, printed text, or handwriting.

Classification occurs by sampling the profiles of the ROI's dominant axis for each ROI from multiple documents. Empty ROIs are identified by their relatively linear profile measured by calculating the standard error of estimate from the least squares regression line of the profile. If the ROI is empty, it is removed from consideration as a candidate printed text ROI.

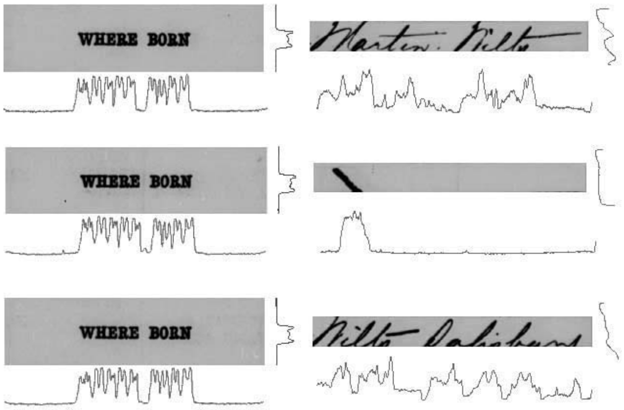

Figure 6. Horizontal and vertical profiles of corresponding ROIs from three separate documents are correlated to discriminate between machine print and handwriting.

The remaining candidate ROI profiles are compared with each other resulting in N choose 2 comparisons. Each comparison is made by calculating the difference

$$d(x) = \Sigma_{i=0}^{N} |p_1(i-x) - p_2(i)| \tag{9}$$

For those ROIs which contain printed text, p_1 and p_2 will be very similar (Fig. 6) and $d(x)$ will have a minimum around $\frac{N}{2}$. Those ROIs identified with high similarity are classified as printed text ROIs while the remainder are classified as handwriting.

At this point we have a template describing the geometric layout of the table with each region's content classified and awaiting further processing (Fig. 7).

4. Results

Three different data sets were used to evaluate the presented approach: British 1841, 1881, and U.S. 1870 census. Each document group represents a line-delimited table, each with their own deficiencies in image quality.

To measure the accuracy of the templates generated we use the metrics of efficiency and coverage as proposed by Garris [4]. Given a reference mesh representing the ground truth of the document's geometric structure, it is compared with the resultant document template called the hypothesis mesh. The hypothesis mesh is measured by two criteria: efficiency and coverage. Efficiency measures the number of ROIs found compared with the number of ROIs in the reference mesh. Coverage measures the similarity between the hypothesis and reference ROIs. ROIs in the reference mesh are matched to similar ROIs in the hypothesis mesh. Any reference ROIs which do not have a corresponding match

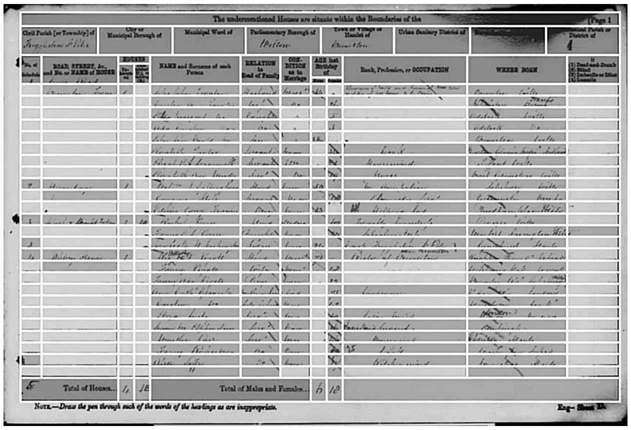
The undermentioned Houses are situate within the Boundaries of the
[Page 1
Civil Parish [or Township] of
City or Municipal Borough of
Municipal Ward of
Parliamentary Borough of
Town or Village or Hamlet of
Urban Sanitary District of
Ecclesiastical Parish or District of
No. of Schedule
ROAD, STREET, &c., and No. or NAME of HOUSE
HOUSES
NAME and Surname of each Person
RELATION to Head of Family
CONDITION as to Marriage
AGE last Birthday of
Rank, Profession, or OCCUPATION
WHERE BORN
If (1) Deaf-and-Dumb (2) Blind (3) Imbecile or Idiot (4) Lunatic
Total of Houses
Total of Males and Females
Note.—Draw the pen through such of the words of the headings as are inappropriate.
Eng.—Sheet B.

Figure 7. Zoned document: Printed text (dark gray), Handwriting (light gray).

Table 1. Efficiency Error

Census	Avg.	Tmpl.
1841	0.100	0.008
1870	0.079	0.003
1881	0.033	0.012

Table 2. Coverage Error

Census	Avg.	Tmpl.
1841	0.102	0.027
1870	0.106	0.005
1881	0.081	0.012

are called "deleted" ROIs and ROIs in the hypothesis mesh which do not have a match are called "inserted" ROIs.

Efficiency error (e) is defined as

$$e = \frac{d+i}{N+d+i} \tag{10}$$

where d is the number of "deleted" ROIs, i is the number of "inserted" ROIs, and N is the total number of ROIs in the reference template.

Coverage error (c) is defined as

$$c = \frac{u+o}{A+u+o} \tag{11}$$

where u is the amount of underage, or the area in the reference ROI which does not overlap with the hypothesis ROI and includes the area of "deleted" ROIs, o is the overage, or the area in the hypothesis ROI which does not overlap with the reference ROI and includes the area of "inserted" ROIs, and A is the sum of the reference ROI's area.

Tables 1 and 2 show the average calculated efficiency and coverage error across all images compared to that of the template. In every case the template's error rate is significantly lower, demonstrating the power of consensus with sequences of similar documents.

5. Conclusion

We have presented a method which relies on intra- and inter-document consensus to build a robust template of the geometric layout of a tabular document and have briefly shown that combining information from multiple images provides superior results to zoning the images separately.

References

[1] S. Chandran, S. Balasubramanian, T. Gandhi, A. Prasad, and R. Kasturi. Structure recognition and information extraction from tabular documents. *International Journal of Imaging Systems and Technology*, 7:289–303, 1996.

[2] S. Chandran and R. Kasturi. Structural recognition of tabulated data. In *ICDAR*, pp. 516–519, 1993.

[3] J. Chen and H. Lee. An efficient algorithm for form structure extraction using strip projection. *Pattern Recognition*, 31(9):1353–1368, 1998.

[4] M. Garris. Evaluating spatial correspondence of zones in document recognition systems. In *ICIP*, pp. 304–307, Oct. 1995.

[5] E. Green and M. Krishnamoorthy. Model-based analysis of printed tables. In *ICDAR*, pp. 214–217, Aug. 1995.

[6] D. J. Ittner and H. S. Baird. Language-free layout analysis. In *ICDAR*, pp. 336–340, Oct. 1993.

[7] T. Kieninger and A. Dengel. The T-recs table recognition and analysis system. In *Document Analysis Systems: Theory and Practice, Third IAPR Workshop*, pp. 255–269, Nov. 1998.

[8] G. Nagy. Twenty years of document image analysis in PAMI. *PAMI*, 22(1):38–62, 2000.

[9] G. Nagy and S. Seth. Hierarchical representation of optically scanned documents. In *Seventh ICPR*, pp. 347–349, Jul. 1984.

[10] G. Nagy, S. C. Seth, and S. D. Stoddard. Document analysis with an expert system. In *Pattern Recognition Practice II*, pp. 149–159, Jun. 1986.

[11] L. O'Gorman. The document spectrum for page layout analysis. *PAMI*, 15(11):1162–1173, 1993.

[12] T. Saitoh, T. Yamaai, and M. Tachikawa. Document image segmentation and layout analysis. *IEICE Transactions on Information and Systems*, E77-D(7):778–888, 1994.

[13] A. Simon, J. Pret, and A. Johnson. A fast algorithm for bottom-up document layout analysis. *PAMI*, 19(3):273–277, 1997.

[14] Y. Tang, S. Lee, and C. Suen. Automatic document processing: A survey. *Pattern Recognition*, 29(12):1931–1952, 1996.

[15] Y. Tang, J. Liu, B. F. Li, and D. Xi. Multiresolution analysis in extraction of reference lines from documents with gray level background. *PAMI*, 19(8):921–926, 1997.

[16] D. Xi and S. Lee. Table structure extraction from form documents based on gradient-wavelet scheme. In *Document Analysis Systems: Theory and Practice, Third IAPR Workshop*, pp. 240–254, Nov. 1998.