Recognition and Extraction of Tabular Data from Scanned Document Images

Information Technologies Institute
Centre of Research & Technology - Hellas
6th km Xarilaou - Thermi, 57001, Thessaloniki, Greece

***Abstract*— In this paper we propose a heuristics-based method for automatic table detection in document images. Our method works on the output of Tesseract, Google's open-source OCR engine, which provides the bounding box and the text of each word in the document. These elements are grouped on a bottom-up way, in order to identify tabular arrangements of data. The approach aims at improving the results of OCR, when used on documents with tables.**

***Index Terms*—table recognition, OCR, scanned document images.**

I. Introduction

Tables are visually oriented arrangements of text, widely used to present complex information and data to human readers. While for regular text it is sufficient to acquire just the textual content of each word, in the case of tables it is critical to identify their basic building blocks (table cells) and group them together in rows and columns. Thus, in order to extract and understand the information stored in a table, we have to process the raw data generated by an OCR engine and recreate the table's layout and structure.

Kieninger & Dengel[1] propose a bottom-up approach for table recognition, by grouping words in segments and looking for horizontally overlapping structures in the text. Oro & Ruffolo[2] further build on this approach, by adding line-by-line analysis and selecting initial table areas, before moving to the bottom-up grouping of the data. Our method is partially based on the above approaches, however we propose a novel method for the grouping of the word segments in table columns and multiple-line table rows.

Section 2 describes in detail our approach, section 3 presents some experimental results in comparison with the state-of-the-art commercial OCR engine ABBYY FineReader 11 and section 4 proposes some future improvements.

II. Proposed Method

The goal of our approach is to export in HTML format the image of a document, while reconstructing any tables present in the document.

The steps of the algorithm are the following:

A. Image Processing

Initially, some image processing is required, for the OCR engine to generate better results. The input image is binarized and resized (300 dpi, average letter height 40-60 px). Since Tesseract does not support multiple-column text, we check for continuous vertical white areas in the image and segment it in blocks, each one corresponding to a text column. Finally, we remove any horizontal and vertical grid lines, because the OCR engine has trouble distinguishing text when it is surrounded by thick dense grid lines.

B. Optical Character Recognition

The processed image is given as input to the OCR engine, which in return exports for each recognized word:

- A string containing the word characters.
- The Cartesian coordinates of the bounding box of the word.
- The recognition confidence of the word
- Font information for the word (font size, bold, italic, underlined)
- Dictionary information (whether the word is found in Tesseract's dictionary or not).

C. Text Lines, Word Segments and Line Types

Using the coordinates for each word, we create text lines from words that overlap vertically. For each line, we create word segments, by merging together words whose horizontal distance is below a threshold T_{seg} (dense text). Finally, we assign each line to a specific type:

- Type 1 / Text – Lines with a single long segment (longer than half of the text width).
- Type 2 / Table – Lines with more than one segments.
- Type 3 / Unknown – Lines with a single short segment.

In case of multiple continuous images, the top and bottom areas of each image are checked for similar repetitive word segments. When found, these segments are considered footers/headers and are removed, except from the headers on the first image. Next, the images are processed as a single continuous document.

D. Initial Table Areas

We create initial table areas, by grouping together adjacent type 2 and type 3 lines. However, for a type 3 line to be assigned to a table area, it must have at least one type 2 line above it (a table cannot begin with a type 3 line at the top).

E. Table Column Generation

Next, we select word segments as column generators, for each table area. The selection algorithm is the following:

1. Select the left-most word segment that is not assigned to a column
2. Find all the word segments that horizontally almost align with it (the horizontal distance between the left limit of the word segments is below a threshold T_{col}
3. Find the average length of these word segments.
4. Select as a column generator, the word segment that is closest to the average length found on step 4.
5. Assign to a new column all the word segments that horizontally overlap with the column generator (some longer word segments can be assigned to more than one column).
6. Go back to step 1, until all the segments in a table area are assigned to a column.

Next, the newly generated column are customized, based on the following criteria:

- If a column has only one segment, which is on the 1st table row (possibly misaligned table header), while the column on its left does not have a segment in the same row, these two columns are merged.
- If a column has more multiple-column segments than single-column segments, it is merged to the column on its left
- Tables that end up with a single column are discarded and treated as simple text.
- Type 3 lines at the bottom of a table, with a multiple-column segment, are removed from the table
- Scrambled tables, i.e. tables that have a very inconsistent format, usually generated by random word segments with big white spaces between them (justified text alignment), are discarded and treated as simple text
- If all the columns of a table have more empty cells than cells with data, the table is discarded (simple formatted text).

F. Multiple-Line Table Rows

Some of the rows in the table areas are merged to multiple-line rows, based on the following criteria:

- If a table row does not have a word segment in the first column, and there is one-to-one correspondence with the word segments of the table row above it, these two rows are merged together.
- If a type 3 table row has its single segment assigned to the first column, and the row below it has a segment in the first column which is indented to the right, these two rows are merged together.
- Tables that end up with a single row or with two single-line rows are discarded and treated as simple text.

G. Data Extraction

Finally the text and tabular data are extracted in HTML format, using a custom made HTML writer.

III. Experimental Results

Below are displayed some experimental results, covering multiple table layouts. The figures displayed are the input image, the HTML generated by our and the HTML generated by state-of-the-art commercial OCR engine ABBYY FineReader11 is displayed on the right.

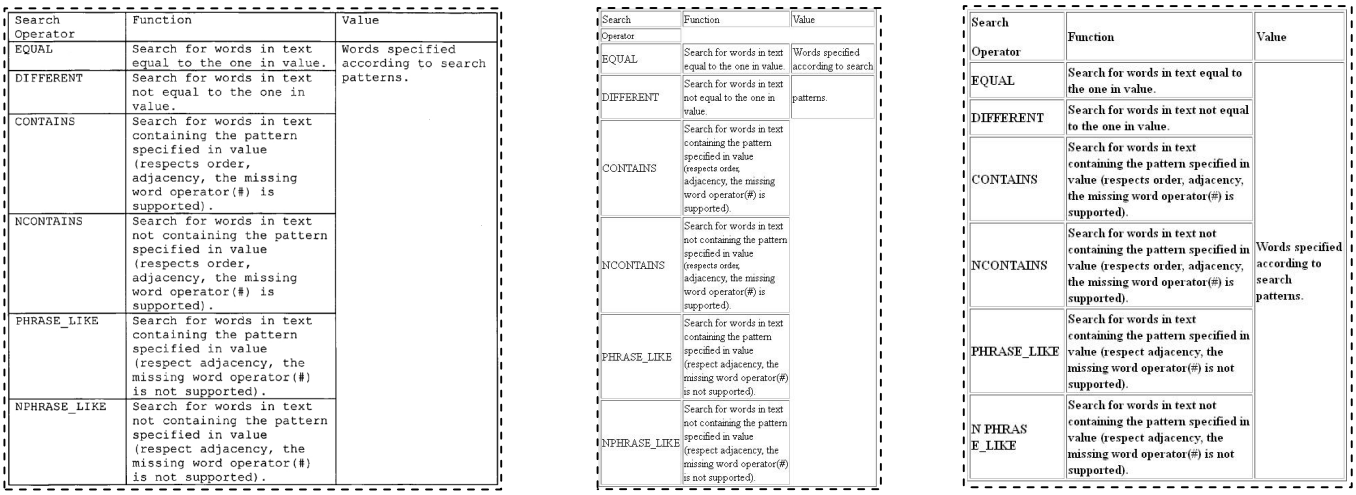
The algorithm successfully recognizes and reconstructs the table in most cases, and in some cases it even outperforms the commercial OCR engine.

Spot		mean.co	mean.pd	SD.co	SD.pd	log2.fold	p.val.Wilcox	LDA.Coef	Delta.norm
1	86	4,48E-03	1,56E-03	2,88E-03	9,72E-04	-1,521	3,32E-04	-425,75	1,244
2	87	9,34E-03	4,16E-03	7,19E-03	3,91E-03	-1,166	1,26E-02	-236,86	1,226
3	335	9,41E-03	2,73E-03	9,70E-03	1,83E-03	-1,786	1,96E-03	178,98	1,196
4	362	9,56E-03	4,38E-03	6,24E-03	3,48E-03	-1,126	2,72E-03	382,08	1,980
5	365	3,95E-02	1,40E-02	2,77E-02	1,01E-02	-1,493	8,10E-04	39,32	1,001
6	368	1,44E-02	6,63E-03	9,32E-03	4,55E-03	-1,122	7,82E-03	177,71	1,387
7	369	4,31E-02	1,47E-02	2,29E-02	1,02E-02	-1,557	6,06E-05	-165,94	4,724
8	382	1,86E-02	7,94E-03	1,64E-02	3,27E-03	-1,224	3,47E-04	-100,27	1,064
9	657	5,92E-03	2,10E-03	7,31E-03	1,37E-03	-1,497	9,97E-03	-299,77	1,146

Spot		mean.co	mean.pd	SD.co	SD.pd	log2.fold	p.val.Wilcox	LDA.Coef	Delta.norm
1	86	4,48E—03	1,56E—03	2,88E—03	9,72E—04	—1,521	3,32E—04	—425,75	1,244
2	87	9,34E—03	4,16E—03	7,19E—03	3,91E—03	—1,166	1,26E—02	—236,86	1,226
3	335	9,41E—03	2,73E—03	9,70E—03	1,83E—03	—1,786	1,96E—03	178,98	1,196
4	362	9,56E—O3	4,38E—03	6,24E—O3	3,48E—03	—1,126	2,72E—03	382,08	1,980
5	365	3,95E—O2	1,40E—02	2,77E—02	1,01E—02	—1,493	8,10E—04	39,32	1,001
6	368	1,44E—02	6,63E—03	9,32E—03	4,55E—03	—1,122	7,82E—03	177,71	1,387
7	369	4,31E—02	1,47E—02	2,29E—02	1,02E—02	—1,557	6,06E—05	—165,94	4,724
8	382	1,86E—02	7,94E—03	1,64E—02	3,27E—03	-1 ,224	3,47E—04	—100,27	1,064
9	657	5,92E—03	2,10E—03	7,31E—03	1,37E—03	—1,497	9,97E—03	—299,77	1,146

Spot	mean.co	mean.pd	SD.co	SD.pd	log2.fold	p.val.Wilcox	LDA.Coef	Delta.norm
1	86	4,48E-03	1,56E-03	2,88E-03	9,72E-04	-1,521	3,32E-04	-425,75	1,244
2	87	9,34E-03	4,16E-03	7,19E-03	3,91 E-03	-1,166	1,26E-02	-236,86	1,226
3	335	9,41 E-03	2,73E-03	9,70E-03	1,83E-03	-1,786	1,96E-03	178,98	1,196
4	362	9,56E-03	4,38E-03	6,24E-03	3,48E-03	-1,126	2,72E-03	382,08	1,980
5	365	3,95E-02	1,40E-02	2,77E-02	1,01 E-02	-1,493	8,10E-04	39,32	1,001
6	368	1,44E-02	6,63E-03	9,32E-03	4,55E-03	-1,122	7,82E-03	177,71	1,387
7	369	4,31 E-02	1,47E-02	2,29E-02	1,02E-02	-1,557	6,06E-05	-165,94	4,724
8	382	1,86E-02	7,94E-03	1,64E-02	3,27E-03	-1,224	3,47E-04	-100,27	1,064
9	657	5,92E-03	2,10E-03	7,31 E-03	1,37E-03	-1,497	9,97E-03	-299,77	1,146

Fig. 1. Table with misaligned headers, no gridlines, no text. Top – input image, centre - HTML generated by our algorithm, bottom – HTML generated by ABBYY FineReader11

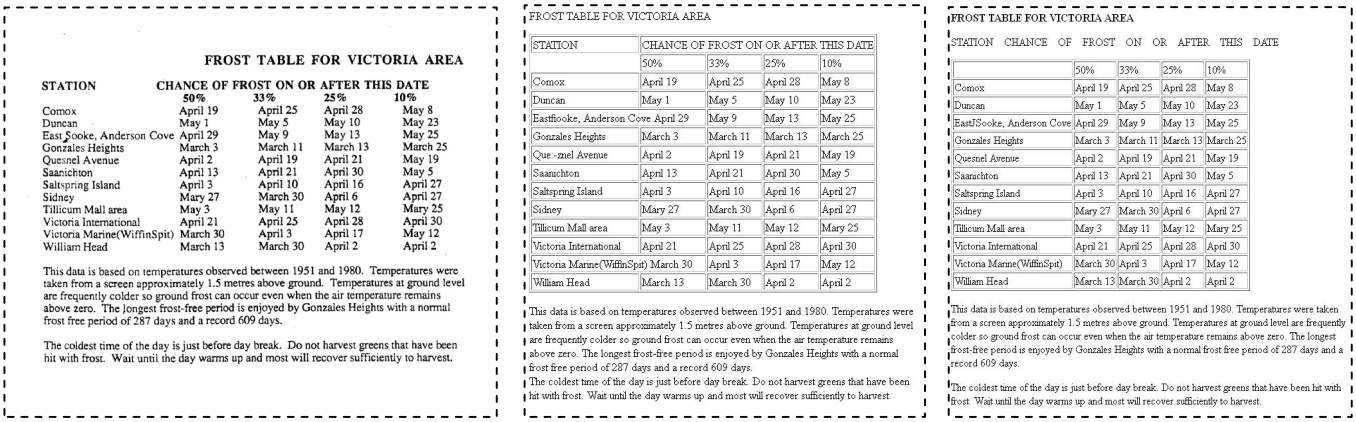


Search Operator	Function	Value
EQUAL	Search for words in text equal to the one in value.	Words specified according to search patterns.
DIFFERENT	Search for words in text not equal to the one in value.	
CONTAINS	Search for words in text containing the pattern specified in value (respects order, adjacency, the missing word operator(#) is supported).	
NCONTAINS	Search for words in text not containing the pattern specified in value (respects order, adjacency, the missing word operator(#) is supported).	
PHRASE_LIKE	Search for words in text containing the pattern specified in value (respect adjacency, the missing word operator(#) is not supported).	
NPHRASE_LIKE	Search for words in text not containing the pattern specified in value (respect adjacency, the missing word operator(#) is not supported).	

Search Operator	Function	Value
EQUAL	Search for words in text equal to the one in value.	Words specified according to search patterns.
DIFFERENT	Search for words in text not equal to the one in value.	
CONTAINS	Search for words in text containing the pattern specified in value (respects order, adjacency, the missing word operator(#) is supported).	
NCONTAINS	Search for words in text not containing the pattern specified in value (respects order, adjacency, the missing word operator(#) is supported).	
PHRASE_LIKE	Search for words in text containing the pattern specified in value (respect adjacency, the missing word operator(#) is not supported).	
NPHRASE_LIKE	Search for words in text not containing the pattern specified in value (respect adjacency, the missing word operator(#) is not supported).	

Search Operator	Function	Value
EQUAL	Search for words in text equal to the one in value.	Words specified according to search patterns.
DIFFERENT	Search for words in text not equal to the one in value.	
CONTAINS	Search for words in text containing the pattern specified in value (respects order, adjacency, the missing word operator(#) is supported).	
NCONTAINS	Search for words in text not containing the pattern specified in value (respects order, adjacency, the missing word operator(#) is supported).	
PHRASE_LIKE	Search for words in text containing the pattern specified in value (respect adjacency, the missing word operator(#) is not supported).	
N PHRAS E_LIKE	Search for words in text not containing the pattern specified in value (respect adjacency, the missing word operator(#) is not supported).	

Fig. 2. Table with multiple-line rows, with gridlines, no text. Left – input image, centre - HTML generated by our algorithm, right – HTML generated by ABBYY FineReader11



FROST TABLE FOR VICTORIA AREA

STATION	CHANCE OF FROST ON OR AFTER THIS DATE			
	50%	33%	25%	10%
Comox	April 19	April 25	April 28	May 8
Duncan	May 1	May 5	May 10	May 23
East Sooke, Anderson Cove	April 29	May 9	May 13	May 25
Gonzales Heights	March 3	March 11	March 13	March 25
Quesnel Avenue	April 2	April 19	April 21	May 19
Saanichton	April 13	April 21	April 30	May 5
Saltspring Island	April 3	April 10	April 16	April 27
Sidney	Mary 27	March 30	April 6	April 27
Tillicum Mall area	May 3	May 11	May 12	Mary 25
Victoria International	April 21	April 25	April 28	April 30
Victoria Marine(WiffinSpit)	March 30	April 3	April 17	May 12
William Head	March 13	March 30	April 2	April 2

This data is based on temperatures observed between 1951 and 1980. Temperatures were taken from a screen approximately 1.5 metres above ground. Temperatures at ground level are frequently colder so ground frost can occur even when the air temperature remains above zero. The longest frost-free period is enjoyed by Gonzales Heights with a normal frost free period of 287 days and a record 609 days.

The coldest time of the day is just before day break. Do not harvest greens that have been hit with frost. Wait until the day warms up and most will recover sufficiently to harvest.

FROST TABLE FOR VICTORIA AREA

STATION	CHANCE OF FROST ON OR AFTER THIS DATE			
	50%	33%	25%	10%
Comox	April 19	April 25	April 28	May 8
Duncan	May 1	May 5	May 10	May 23
Eastfiooke, Anderson Cove	April 29	May 9	May 13	May 25
Gonzales Heights	March 3	March 11	March 13	March 25
Que:-znel Avenue	April 2	April 19	April 21	May 19
Saanichton	April 13	April 21	April 30	May 5
Saltspring Island	April 3	April 10	April 16	April 27
Sidney	Mary 27	March 30	April 6	April 27
Tillicum Mall area	May 3	May 11	May 12	Mary 25
Victoria International	April 21	April 25	April 28	April 30
Victoria Marine(WiffinSpit)	March 30	April 3	April 17	May 12
William Head	March 13	March 30	April 2	April 2

This data is based on temperatures observed between 1951 and 1980. Temperatures were taken from a screen approximately 1.5 metres above ground. Temperatures at ground level are frequently colder so ground frost can occur even when the air temperature remains above zero. The longest frost-free period is enjoyed by Gonzales Heights with a normal frost free period of 287 days and a record 609 days.
The coldest time of the day is just before day break. Do not harvest greens that have been hit with frost. Wait until the day warms up and most will recover sufficiently to harvest.

FROST TABLE FOR VICTORIA AREA

STATION CHANCE OF FROST ON OR AFTER THIS DATE

	50%	33%	25%	10%
Comox	April 19	April 25	April 28	May 8
Duncan	May 1	May 5	May 10	May 23
EastJSooke, Anderson Cove	April 29	May 9	May 13	May 25
Gonzales Heights	March 3	March 11	March 13	March 25
Quesnel Avenue	April 2	April 19	April 21	May 19
Saanichton	April 13	April 21	April 30	May 5
Saltspring Island	April 3	April 10	April 16	April 27
Sidney	Mary 27	March 30	April 6	April 27
Tillicum Mall area	May 3	May 11	May 12	Mary 25
Victoria International	April 21	April 25	April 28	April 30
Victoria Marine(WiffinSpit)	March 30	April 3	April 17	May 12
William Head	March 13	March 30	April 2	April 2

This data is based on temperatures observed between 1951 and 1980. Temperatures were taken from a screen approximately 1.5 metres above ground. Temperatures at ground level are frequently colder so ground frost can occur even when the air temperature remains above zero. The longest frost-free period is enjoyed by Gonzales Heights with a normal frost free period of 287 days and a record 609 days.

The coldest time of the day is just before day break. Do not harvest greens that have been hit with frost. Wait until the day warms up and most will recover sufficiently to harvest.

Fig. 3. Table with a multiple-column header, no gridlines, with text. Left – input image, centre - HTML generated by our algorithm, right – HTML generated by ABBYY FineReader11

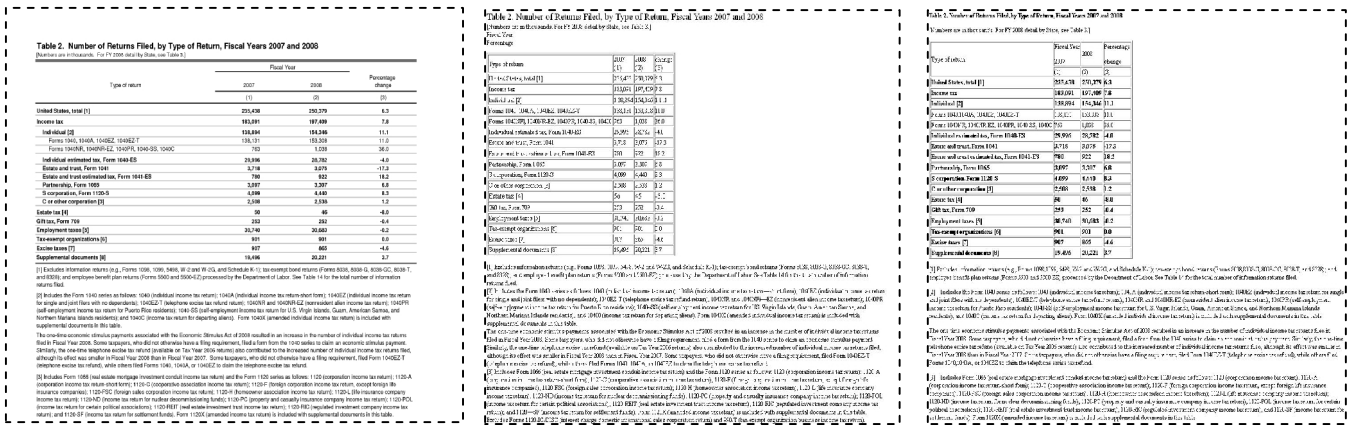


Table 2. Number of Returns Filed, by Type of Return, Fiscal Years 2007 and 2008

[Numbers are in thousands. For FY 2008 detail by State, see Table 5.]

Type of return	Fiscal Year 2007	Fiscal Year 2008	Percentage change
	(1)	(2)	(3)
United States, total [1]	**255,438**	**260,379**	**1.9**
Income tax	**183,091**	**197,499**	**7.9**
Individual [2]	**138,894**	**154,346**	**11.1**
Forms 1040, 1040A, 1040EZ, 1040EZ-T	136,131	153,308	11.0
Forms 1040PR, 1040PR-EZ, 1040PR, 1040-SS, 1040C	763	1,039	36.0
Individual estimated tax, Form 1040-ES	**23,996**	**20,782**	**-4.0**
Estate and trust, Form 1041	**3,718**	**3,079**	**-17.0**
Estate and trust estimated tax, Form 1041-ES	**780**	**922**	**18.2**
Partnership, Form 1065	**3,207**	**3,297**	**0.8**
S corporation, Form 1120-S	**4,098**	**4,440**	**8.3**
C or other corporation [3]	**2,598**	**2,536**	**1.2**
Estate tax [4]	**50**	**46**	**-8.0**
Gift tax, Form 709	**252**	**252**	**-0.4**
Employment taxes [5]	**30,748**	**30,663**	**-0.2**
Tax-exempt organizations [6]	**901**	**901**	**0.0**
Excise taxes [7]	**937**	**865**	**-4.0**
Supplemental documents [8]	**16,496**	**20,231**	**2.7**

[1] Excludes information returns (e.g., Forms 1098, 1099, 5498, W-2 and W-2G, and Schedule K-1); tax-exempt bond returns (Forms 8038, 8038-G, 8038-GC, 8038-T, and 8328); and employee benefit plan returns (Forms 5500 and 5500-EZ) processed by the Department of Labor. See Table 14 for the total number of information returns filed.

[2] Includes the Form 1040 series as follows: 1040 (individual income tax return); 1040A (individual income tax return-short form); 1040EZ (individual income tax return for single and joint filers with no dependents); 1040EZ-T (telephone excise tax refund return); 1040PR and 1040PR-EZ (nonresident alien income tax return); 1040PR (self-employment income tax return for Puerto Rico residents); 1040-SS (self-employment income tax return for U.S. Virgin Islands, Guam, American Samoa, and Northern Mariana Islands residents); and 1040C (income tax return for departing aliens). Form 1040S (amended individual income tax return) is included with supplemental documents in this table.

The one-time economic stimulus payments associated with the Economic Stimulus Act of 2008 resulted in an increase in the number of individual income tax returns filed in Fiscal Year 2008. Some taxpayers, who did not otherwise have a filing requirement, filed a form from the 1040 series to claim an economic stimulus payment. Similarly, the one-time telephone excise tax refund (available on Tax Year 2006 returns) also contributed to the increased number of individual income tax returns filed, although its effect was smaller in Fiscal Year 2008 than in Fiscal Year 2007. Some taxpayers, who did not otherwise have a filing requirement, filed Form 1040EZ-T (telephone excise tax refund), while others filed Forms 1040, 1040A, or 1040EZ to claim the telephone excise tax refund.

[3] Includes Form 1066 (real estate mortgage investment conduit income tax return) and the Form 1120 series as follows: 1120 (corporation income tax return); 1120-A (corporation income tax return-short form); 1120-C (cooperative association income tax return); 1120-F (foreign corporation income tax return, except foreign life insurance companies); 1120-FSC (foreign sales corporation income tax return); 1120-H (homeowner association income tax return); 1120-L (life insurance company income tax return); 1120-ND (income tax return for nuclear decommissioning funds); 1120-PC (property and casualty insurance company income tax return); 1120-POL (income tax return for certain political associations); 1120-REIT (real estate investment trust income tax return); 1120-RIC (regulated investment company income tax return); and 1120-SF (income tax return for settlement funds). Form 1120X (amended income tax return) is included with supplemental documents in this table.

Table 2. Number of Returns Filed, by Type of Return, Fiscal Years 2007 and 2008

[Numbers are in thousands. For FY 2008 detail by State, see Table 5.]

Fiscal Year

Type of return	2007	2008	Percentage change
	(1)	(2)	(3)
United States, total [1]	**255,438**	**260,379**	**1.9**
Income tax	**183,091**	**197,499**	**7.9**
Individual [2]	**138,894**	**154,346**	**11.1**
Forms 1040, 1040A, 1040EZ, 1040EZ-T	136,131	153,308	11.0
Forms 1040PR, 1040PR-EZ, 1040PR, 1040-SS, 1040C	763	1,039	36.0
Individual estimated tax, Form 1040-ES	**23,996**	**20,782**	**-4.0**
Estate and trust, Form 1041	**3,718**	**3,079**	**-17.0**
Estate and trust estimated tax, Form 1041-ES	**780**	**922**	**18.2**
Partnership, Form 1065	**3,207**	**3,297**	**0.8**
S corporation, Form 1120-S	**4,098**	**4,440**	**8.3**
C or other corporation [3]	**2,598**	**2,536**	**1.2**
Estate tax [4]	**50**	**46**	**-8.0**
Gift tax, Form 709	**252**	**252**	**-0.4**
Employment taxes [5]	**30,748**	**30,663**	**-0.2**
Tax-exempt organizations [6]	**901**	**901**	**0.0**
Excise taxes [7]	**937**	**865**	**-4.0**
Supplemental documents [8]	**16,496**	**20,231**	**2.7**

[1] Excludes information returns (e.g., Forms 1098, 1099, 5498, W-2 and W-2G, and Schedule K-1); tax-exempt bond returns (Forms 8038, 8038-G, 8038-GC, 8038-T, and 8328); and employee benefit plan returns (Forms 5500 and 5500-EZ) processed by the Department of Labor. See Table 14 for the total number of information returns filed.

[2] Includes the Form 1040 series as follows: 1040 (individual income tax return); 1040A (individual income tax return-short form); 1040EZ (individual income tax return for single and joint filers with no dependents); 1040EZ-T (telephone excise tax refund return); 1040PR and 1040PR-EZ (nonresident alien income tax return); 1040PR (self-employment income tax return for Puerto Rico residents); 1040-SS (self-employment income tax return for U.S. Virgin Islands, Guam, American Samoa, and Northern Mariana Islands residents); and 1040C (income tax return for departing aliens). Form 1040S (amended individual income tax return) is included with supplemental documents in this table.

The one-time economic stimulus payments associated with the Economic Stimulus Act of 2008 resulted in an increase in the number of individual income tax returns filed in Fiscal Year 2008. Some taxpayers, who did not otherwise have a filing requirement, filed a form from the 1040 series to claim an economic stimulus payment. Similarly, the one-time telephone excise tax refund (available on Tax Year 2006 returns) also contributed to the increased number of individual income tax returns filed, although its effect was smaller in Fiscal Year 2008 than in Fiscal Year 2007. Some taxpayers, who did not otherwise have a filing requirement, filed Form 1040EZ-T (telephone excise tax refund), while others filed Forms 1040, 1040A, or 1040EZ to claim the telephone excise tax refund.

[3] Includes Form 1066 (real estate mortgage investment conduit income tax return) and the Form 1120 series as follows: 1120 (corporation income tax return); 1120-A (corporation income tax return-short form); 1120-C (cooperative association income tax return); 1120-F (foreign corporation income tax return, except foreign life insurance companies); 1120-FSC (foreign sales corporation income tax return); 1120-H (homeowner association income tax return); 1120-L (life insurance company income tax return); 1120-ND (income tax return for nuclear decommissioning funds); 1120-PC (property and casualty insurance company income tax return); 1120-POL (income tax return for certain political associations); 1120-REIT (real estate investment trust income tax return); 1120-RIC (regulated investment company income tax return); and 1120-SF (income tax return for settlement funds). Form 1120X (amended income tax return) is included with supplemental documents in this table.

Table 2. Number of Returns Filed, by Type of Return, Fiscal Years 2007 and 2008

[Numbers are in thousands. For FY 2008 detail by State, see Table 5.]

Type of return	Fiscal Year 2007	Fiscal Year 2008	Percentage change
	(1)	(2)	(3)
United States, total [1]	**255,438**	**260,379**	**1.9**
Income tax	**183,091**	**197,499**	**7.9**
Individual [2]	**138,894**	**154,346**	**11.1**
Forms 1040, 1040A, 1040EZ, 1040EZ-T	136,131	153,308	11.0
Forms 1040PR, 1040PR-EZ, 1040PR, 1040-SS, 1040C	763	1,039	36.0
Individual estimated tax, Form 1040-ES	**23,996**	**20,782**	**-4.0**
Estate and trust, Form 1041	**3,718**	**3,079**	**-17.0**
Estate and trust estimated tax, Form 1041-ES	**780**	**922**	**18.2**
Partnership, Form 1065	**3,207**	**3,297**	**0.8**
S corporation, Form 1120-S	**4,098**	**4,440**	**8.3**
C or other corporation [3]	**2,598**	**2,536**	**1.2**
Estate tax [4]	**50**	**46**	**-8.0**
Gift tax, Form 709	**252**	**252**	**-0.4**
Employment taxes [5]	**30,748**	**30,663**	**-0.2**
Tax-exempt organizations [6]	**901**	**901**	**0.0**
Excise taxes [7]	**937**	**865**	**-4.0**
Supplemental documents [8]	**16,496**	**20,231**	**2.7**

[1] Excludes information returns (e.g., Forms 1098, 1099, 5498, W-2 and W-2G, and Schedule K-1); tax-exempt bond returns (Forms 8038, 8038-G, 8038-GC, 8038-T, and 8328); and employee benefit plan returns (Forms 5500 and 5500-EZ) processed by the Department of Labor. See Table 14 for the total number of information returns filed.

[2] Includes the Form 1040 series as follows: 1040 (individual income tax return); 1040A (individual income tax return-short form); 1040EZ (individual income tax return for single and joint filers with no dependents); 1040EZ-T (telephone excise tax refund return); 1040PR and 1040PR-EZ (nonresident alien income tax return); 1040PR (self-employment income tax return for Puerto Rico residents); 1040-SS (self-employment income tax return for U.S. Virgin Islands, Guam, American Samoa, and Northern Mariana Islands residents); and 1040C (income tax return for departing aliens). Form 1040S (amended individual income tax return) is included with supplemental documents in this table.

The one-time economic stimulus payments associated with the Economic Stimulus Act of 2008 resulted in an increase in the number of individual income tax returns filed in Fiscal Year 2008. Some taxpayers, who did not otherwise have a filing requirement, filed a form from the 1040 series to claim an economic stimulus payment. Similarly, the one-time telephone excise tax refund (available on Tax Year 2006 returns) also contributed to the increased number of individual income tax returns filed, although its effect was smaller in Fiscal Year 2008 than in Fiscal Year 2007. Some taxpayers, who did not otherwise have a filing requirement, filed Form 1040EZ-T (telephone excise tax refund), while others filed Forms 1040, 1040A, or 1040EZ to claim the telephone excise tax refund.

[3] Includes Form 1066 (real estate mortgage investment conduit income tax return) and the Form 1120 series as follows: 1120 (corporation income tax return); 1120-A (corporation income tax return-short form); 1120-C (cooperative association income tax return); 1120-F (foreign corporation income tax return, except foreign life insurance companies); 1120-FSC (foreign sales corporation income tax return); 1120-H (homeowner association income tax return); 1120-L (life insurance company income tax return); 1120-ND (income tax return for nuclear decommissioning funds); 1120-PC (property and casualty insurance company income tax return); 1120-POL (income tax return for certain political associations); 1120-REIT (real estate investment trust income tax return); 1120-RIC (regulated investment company income tax return); and 1120-SF (income tax return for settlement funds). Form 1120X (amended income tax return) is included with supplemental documents in this table.

Fig. 4. Table with formatted text, partial gridlines, with text. Left – input image, centre - HTML generated by our algorithm, right – HTML generated by ABBYY FineReader11

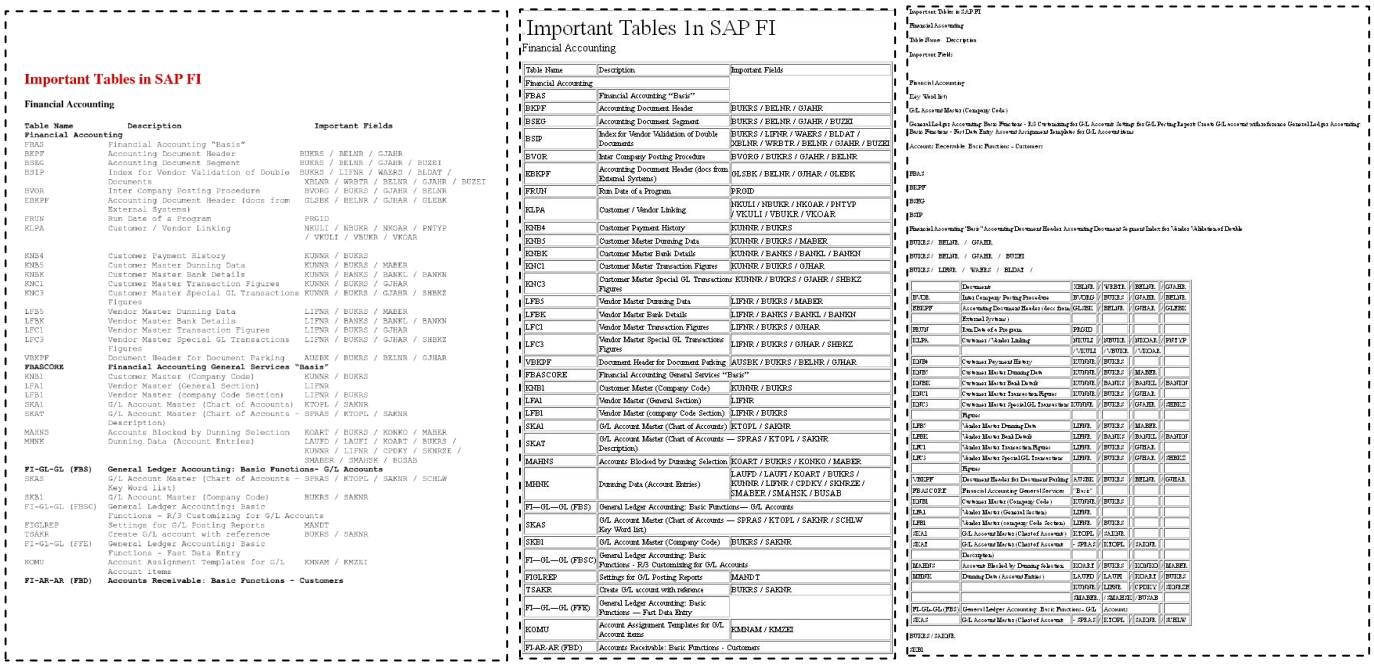


Fig. 5. Table with inconsistent layout, no gridlines, no text. Left – input image, centre - HTML generated by our algorithm, right – HTML generated by ABBYY FineReader11

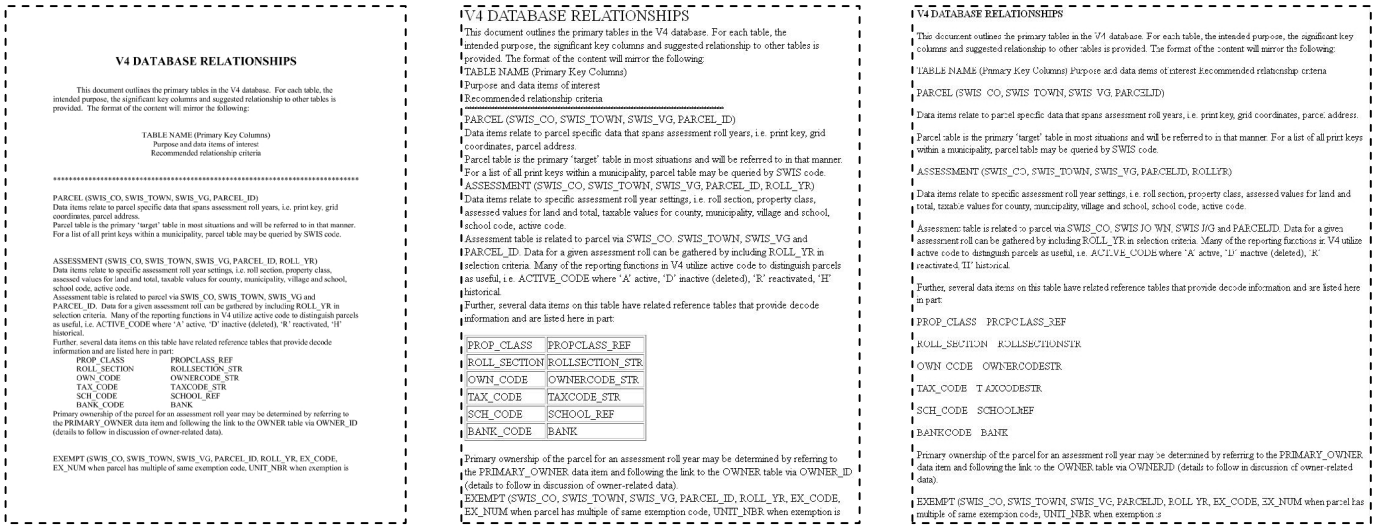


Fig. 6. Table with formatted text, no gridlines, with text. Left – input image, centre - HTML generated by our algorithm, right – HTML generated by ABBYY FineReader11

IV. Future Improvements

The success of the algorithm relies heavily on the correct recognition of the words by the Tesseract OCR engine. Specifically, some of the problems that lead to failure of the algorithm because of limitations of the OCR engine are:

- Documents with both multiple-column and single-column text. As Tesseract cannot distinguish multiple-column text, these parts are reconstructed as big 2-column tables.
- Documents with non-Manhattan Layout give wrong results as Tesseract is unable to segment them
- Documents with vertical text. Tesseract cannot correctly recognize text that has both vertical and horizontal alignment, leading to wrong results.
- Documents with images. Tesseract tries to recognize images as text, leading to "imaginary" words.

Some other minor problems:

- The algorithm sometimes fails to merge consecutive table lines to a single row, as they do not always meet all the line-merging criteria stated in section 2.
- Some words are merged to a single word segment, despite the fact that they are part of different columns, because the horizontal distance between them is very small (below the T_{seg} threshold).
- Formatted text sometimes tends to be recognized as a table (e.g. source code aligned to the left and comments aligned to the right)

References

[1] T. Kieninger and A. Dengel, "The T-Recs table recognition and analysis system," DAS '98, LNCS 1655, pp. 255-270, 1999

[2] E. Oro and M. Ruffolo, "PDF-TREX: An approach for recognizing and extracting tables from pdf documents," ICDAR'09, pp. 906-910, 2009