MSC 90C35, 90C27 DOI: 10.14529/mmp240107

TABLE RECOGNITION TECHNOLOGY IN TAX DOCUMENTS OF THE RUSSIAN FEDERATION

O.A. Slavin[1,2]
[1]Federal Research Center "Computer Science and Control" RAS, Moscow, Russian Federation
[2]LLC "Smart Engines Service", Moscow, Russian Federation
E-mail: oslavin@isa.ru

This paper investigates the problem of cell recognition in the image of a table using the example of the Russian tax document (2-NDFL). Despite the simple structure of the tables, the printing method is based on a flexible template. The flexibility of the form is observed in the modifications of textual information and in the table area. The flexibility of tables lies in the modification of the number and size of columns. A structural method was proposed for table detection. The input data are the detected horizontal and vertical segments. Segments were searched by the Smart Document Reader system. Implementing and testing the method were also carried out in the Smart Document Reader system. In addition to detecting the area where tables can be placed, the following objectives were achieved: searching for table cells, naming table cells, and validating the table area. Validation of the table area was performed for separate tables and for table sets. The application of table aggregate descriptions showed the high reliability of linking table sets.

Keywords: table recognition; line detection; table layout.

Introduction

A text table is defined as a set of rows and columns. Borders of rows and columns can be defined by a limited number of ways of data representation: separating segments (lines), separating areas between text cells, and highlighting by color.

Affordable scanning devices appeared in the late 80's and early 90's of the 20th century. Simultaneously commercial programs for text recognition (OCR) were developed. In OCR it was possible to recognize tables. Tables were extracted from pages of both arbitrary and administrative documents, such as tax, banking, or insurance forms. Such documents contained tables with a known or typed structure. Often administrative document designers are limited to simple tables in the form of matrices.

Currently, there is ongoing research not only in document recognition [1, 2], but also in table recognition. The paper [3] states that optical recognition for data recovery from financial documents using text regression analysis is an expensive and impractical solution. A well-known methodology is pattern matching. However, for classes of documents such as invoices, there is no predefined set of samples, which has been known to limit the accuracy. The authors [3] claim that the use of recurrent neural networks and graph neural networks solves most of these problems. Attention is drawn to the problem of document image noisiness, which leads to incorrect extraction or recognition of characters in images and PDF files. It is stated in [3] It is stated in [3] that most systems recognize tables with errors due to lack of antialiasing, skew correction, rotation correction, etc.

In paper [4] table parsing is reduced to two tasks: table detection and table structure recognition [5]. The task of table detection can be solved by detecting a set of pixels representing the table area in a document. Effective methods are known to solve this problem [6–9], providing high detection results in publicly available datasets. Other tasks in table recognition are table structure identification, table structure comprehension, and cell area detection [10].

An obvious way to identify table structure is to detect grid boundaries [11–13]. Detecting cell areas can be based on identifying the rows and columns that form the set of table cells [4]. The works [11, 14–17] describe mechanisms for predicting the area of rows and columns of a table. When processing simple tables, non-visible grid lines are predicted [18]. The row/column split operation can also extract cells containing multiple lines of text. The authors of [16] describe a group of methods that reconstruct relationships between retrieved table cells using GNN (Graph Neural Networks). It is noted that GNN-based methods depend on substantial training costs based on large volume samples.

Although the topic is well-developed, many logical problems remain relevant. For example, researchers are looking for more efficient solutions to the problem of finding table cell boundaries and the problem of identifying the table structure.

Some of the issues are solved by training on representative datasets. For example, a system CloudScan is described in [19] that extracts data from a dataset using recurrent neural network (LSTM) and provides high-precision extraction. CloudScan does not rely on invoice layout templates. CloudScan provides extraction of 8 fields.

A dataset of 326 471 invoices was used to train CloudScan [19]. Another high-volume dataset is an image-based dataset of documents with tables, TableBank [20]. TableBank consists of 417 000 labelled tables and original documents. A smaller ICDAR dataset is available – 124 documents from the ICDAR 2013 table detection competition [21].

1. Background

Let us consider the task of recognizing tables contained in 2-NDFL tax documents used in the Russian Federation. The 2-NDFL document is a single-column document and can be either single-page or multi-page. 2-NDFL tables are simple, their structure is known in advance. Templates in XLS format are used for printing them out. Features of 2-NDFL tables are the presence of several types of tables (four types of tables are used t_1, t_2, t_3, t_4), several types of tables (four types of tables are used t_1, t_2, t_3, t_4 are used), the possibility of repeating one type of table, transfer of tables to the next page. The source data are images of 2-NDFL documents scanned or digitized by mobile devices. The result of recognition is data from the cells of each table. Recognition of tables should work in the Smart ID Engine document recognition system [22]. For training, validation, and testing there were about 2575 images of 2-NDFL pages available with different digitization quality.

The following functions of the Smart ID Engine were used to solve the problem:
- page boundary search;
- page shape normalization;
- image improvement;
- detection of lines (segments).

Smart ID Engine features were also used to recognize text objects. A two-stage process of flexible document recognition was implemented in this system. In the first stage, the detection of graphic primitives in the normalized page image and recognition without using the description of the document and its parts were performed. After that, the boundaries of the fields that contain variable information of the document were predicted. In the second stage, the fields were recognized again using the parameters of the detected fields. It was decided to detect table and cell boundaries before the text recognition stage. This significantly accelerated the processing of one page by eliminating the recognition of words in the table areas at the first stage.

The task of table recognition is to extract the maximum number of cell boundaries of all tables on a page. The following attributes must be known for each cell:
- table type t;
- number of the given table type t;
- column type f.

2. Algorithms for Solving Table Cell Detection Problems

To detect table areas and determine their structure, we used sets of lines

$$S_h = \{S_{h_1}, \ldots, S_{h_n}, S_v = S_{v_1}, \ldots, S_{v_m}\}.$$

Image vectorization was based on methods relying on morphological operations on image pixels.

A large volume of examples, exceeding 100 000 samples, was used to train vectorization mechanisms. This allowed to creation of an algorithm that detects segments of different lengths and with different distortions. The shapes of real segments can be different from ideal ones. The segments may be noisy and partially lost during digitization.

Only a set of vertical segments S_v was used to search for table areas. Each segment of S_v was described by a quadrilateral

$$Q(S_v) = (P_1(S_v), P_2(S_v), P_3(S_v), P_4(S_v)),$$

where each point $P(S_v)$ consisted of two coordinates $P_x(S_v)$ and $P_y(S_v)$ in the normalized page image. For vertical segments the following relation is true

$$|P^{x_1}(S_v) - P^{x_2}(S_v)| \ll |P^{y_1}(S_v) - P^{y_4}(S_v)|. \tag{1}$$

For horizontal segments the following relation is true

$$|P^{x_1}(S_v) - P^{x_2}(S_v)| \gg |P^{y_1}(S_v) - P^{y_4}(S_v)|.$$

The pages were pre-normalized to a size of N_h pixels in height and N_w pixels in width. The projection $\{v_1, \ldots v_{N_h}\}$ of all S_{vj} segments on the vertical axis was calculated:

$$v_q = \sum_{j=1}^{m} \vartheta(S_j^v, q), \tag{2}$$

where $\vartheta(S_j^v, q)$ equals 1, if the quadrilateral S_{v_j} intercepts with the segment

$$(1, q, N_w, q).$$

Otherwise $\vartheta\left(S_j^v, q\right)$ equals 0. In the projection $\{v_1, \ldots, v_{N_h}\}$, non-intersecting areas were convsidered

$$J = (j_1 + j_1 + 1, \dots, j_2)$$

with close values:

$$|V_z(J) - \eta(t)| < \varepsilon(t),$$

where $\eta(t)$ is mode of the value series $\{v_1, \dots, v_{N_h}\}$, and $\varepsilon(t)$ is specified proximity parameter. The parameter $\varepsilon(t)$ is necessary to account for image distortions, leading to segment detection errors.

Month	Revenue Code	Revenue Amount	Deduction Code	Deduction Amount	Month	Revenue Code	Revenue Amount	Deduction Code	Deduction Amount
1	2000	17133.79			1	2010	5.20		
2	2000	18270.65			2	2010	5.20		

Deduction Code	Deduction Amount	Deduction Code	Deduction Amount	Deduction Code	Deduction Amount	Deduction Code	Deduction Amount

Total Profit	370316.44	**Tax Deduction Amount**	48141
Tax Base	370316.44	**Tax Transfer Amount**	48141
Tax Amount	48141	**Tax Overpayment**	0
Fixed Amount Advance Payment	0	**Tax Amount (not withheld)**	0

Income Amount (not withheld)	
Amount of withheld Tax	

Fig. 1. Examples of 2-NDFL document tables t_1, t_2, t_3, t_4

For ideal tables, the values of $\eta(t)$ are equal to the number of table columns, increased by 1 (see. Fig. 1). For ideal tables, the value $\eta(t)$ is invariant: $\eta(t_1) = 12, \eta(t_2) = 9, \eta(t_3) = 5, \eta(t_4) = 3$. However, due to changes in the design of the 2-NDFL document, the number of vertical segments of $\eta(t)$, may also change, see the example in Fig. 2.

Month	Revenue Code	Revenue Amount	Deduction Code	Deduction Amount	Month	Revenue Code	Revenue Amount	Deduction Code	Deduction Amount
01	2000	32163.63			11	2012	3716.26		
02	2000	33414.44			12	2000	150532.38		

Deduction Code	Deduction Amount	Deduction Code	Deduction Amount	Deduction Code	Deduction Amount	Deduction Code	Deduction Amount

Fig. 2. Example of redesigning 2-NDFL document tables (merging t_1 segments and breaking t_2 into separate tables)

Another likely cause of the change in $\eta(t)$ is an error in setting the scan area, resulting in the loss of the leftmost or rightmost segment; see the example in Fig. 3.

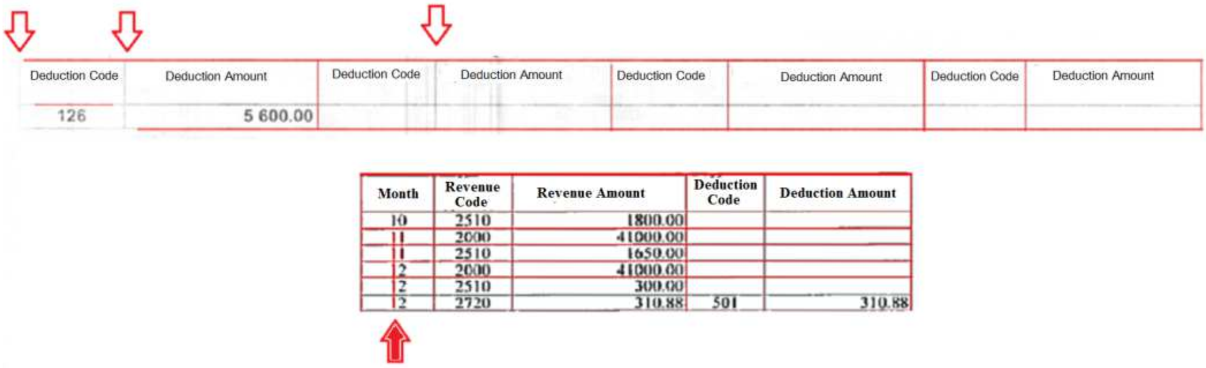
Thus, the acceptable value $\eta(t)$ for each table lies within a certain range $\eta_1(t) \div \eta_2(t)$. In this case, the ranges for some tables intersect $\eta(t_3) \in [3, 12]$ and $\eta(t_4) \in [1, 3]$. Detection errors of noisy segments and false segments are possible, see the example in Fig. 4.

The result of the analysis is represented by a set of candidate areas of the table. Each candidate area J using of characteristic $\eta(t)$ is assigned to one or more types of tables $(\tau_1(J), \tau_2(J), \dots, \tau_{k(j)}(J))$. Qualification precision of several tables on one page

Month	Revenue Code	Revenue Amount	Deduction Code	Deduction Amount
01	2000	22.000,00	-	-
02	2000	17.000,00	-	-
02	2300	5.000,00	-	-
03	2000	22.000,00	-	-
04	2000	22.000,00	-	-
05	2000	22.000,00	-	-
06	2000	22.000,00	-	-

Month	Revenue Code	Revenue Amount	Deduction Code	Deduction Am
-	-	-	-	-
-	-	-	-	-
-	-	-	-	-
-	-	-	-	-
-	-	-	-	-
-	-	-	-	-
-	-	-	-	-

Fig. 3. Example of the redesign of 2-NDFL document tables (loss of vertical segment during scanning)



Deduction Code	Deduction Amount	Deduction Code	Deduction Amount	Deduction Code	Deduction Amount	Deduction Code	Deduction Amount
126	5 600.00						

Month	Revenue Code	Revenue Amount	Deduction Code	Deduction Amount
10	2510	1800.00		
11	2000	41000.00		
11	2510	1650.00		
12	2000	41000.00		
12	2510	300.00		
12	2720	310.88	501	310.88

Fig. 4. Examples of table segment detection errors

was enhanced by the use of a layout – an ordered set of candidate areas $J_1, J_2, \ldots, J_n$. During the training process, acceptable layouts are specified. Some possible permissible layouts are given in Table 1. The following layouts are allowed on the used datasets: (t_1, t_2, t_3, t_4), (t_1, t_2, t_3), (t_1, t_2), (t_4, t_3, t_2, t_1), and others. Layout (t_4, t_3, t_2, t_1), (t_3, t_2, t_1) and their analogues correspond to the case when the document image is rotated by 180 degrees.

Table 1

Permissible table layouts of the 2-NDFL document

t_1	t_1	t_1	t_1	t_1	t_1	t_3	t_4	t_1	t_1	t_1	t_1	t_1	t_1	t_1	t_3
t_2	t_2	t_2	t_2	t_2	t_2	t_2	t_3	t_1	t_2	t_2	?	t_3	t_3	?	t_4
t_3	t_3	t_3	t_3	t_3	t_3	t_1	t_2	t_3	t_2		t_3	t_4		t_6	
t_1	t_1	t_1	t_4	t_4			t_1								
t_2	t_3	t_2	t_1												
t_3															
						rotate	rotate								

Consider the layout $J_1, J_2, \ldots, J_n$ a word with symbols from the alphabet $\{t_1, t_2, t_3, t_4\}$, and possible layouts – a dictionary over the same alphabet. After checking the eligibility of the layout, the table areas are considered to be bound. When specifying the ranges

$\eta_1(t) \div \eta_2(t)$ it is necessary to take into account errors of detection of vertical segments. The main cause of such errors is the lightening of a part of the page area.

The next task is to find the cell boundaries of each table. Horizontal segments are used for this purpose. Similar to the projection of vertical segments, projections of horizontal segments are made in the linking areas of tables with known type. Table sections separated by sufficient gaps are specified (see Fig. 2). In the area of table t_1 there can be one or two sections, in the area of table t_2 – one, two, or four sections. Other types of tables have one section. Further, it is possible to define a simple matrix structure in each section.

False columns or rows may be detected in the lightened areas due to possible loss of part of the segments. Such false cells are the merge of several real cells. However, fused columns or rows can be successfully processed at a later stage. This is the case when it is known that a column (row) is a merge of several columns (rows) with known widths (heights). To a group of such cells, the algorithms forming a column mask are applied [4], which separates a column in two.

After cell boundary detection, cell naming is performed, i.e. attributes are assigned to each cell such as table type t, number of the given table type t, column type f. The column type can be composite for cases of merged columns. When naming, cells from the table caps of classes t_1 and t_2 are excluded. All table cells, including empty ones, are involved in the naming procedure. Cells with constant information are excluded from table cells. Recognition by artificial neural networks and post-processing of the recognition results by LM models are performed within the boundaries of the remaining cells.

3. Experiments

The method parameters were selected on their private dataset D_1. On another proprietary dataset D_2, the layout accuracy, table linking accuracy, and average table cell recognition accuracy were evaluated. The volumes of datasets D_1 and D_2 were 1575 and 1000 documents. The results are summarized in Table 2.

Table 2

Table linking accuracy and average accuracy of table cells recognition

Type of Table		t_1	t_2	t_3
Reject of tables (%)	dataset D_1	0,7	1,1	1,3
	dataset D_2	1,7	1,8	2
Average cell recognition accuracy (%)	dataset D_1	1,35	1,86	1,73
	dataset D_2	2,85	2,81	3,5

The data in Table 2 show that table t_1 is most accurately bound. Tables t_2 and t_3 are bound worse due to the deterioration of vertical line detection in the table zone. This deterioration is due to the lower height of the zones of tables t_2 and t_3 compared to the height of the zone of table t_1.

From the data in Table 2 it can be seen that the tables are detected worse in the test sample images of dataset D_1 than in the training sample images of dataset D_1. Nevertheless, it is impossible to speak about the retraining of the model. First, the number of layout linking failures includes several images of old versions of 2-NDFL documents that do not correspond to the described model. Second, images with table linking errors from the D_2

dataset are actually more distorted. The linking errors for each of the tables are due to the inability to identify the table elements for the overexposed areas. In the overexposed areas, some of the segments are detected inaccurately. When detecting the table area, this leads to unacceptable values of $\eta(t)$.

The achieved accuracy of table linking and recognition was high. Datasets D_1 and D_2 were not synthesized datasets but were taken from real document archives. Datasets D_1 and D_2 can be characterized as noisy compared to the mentioned datasets [20, 21]. These datasets [20, 21] contain document images not only without noise but also without rotation (see examples in Fig. 5). Another test dataset D_3 was synthesized. Dataset D_3 consisted of clean documents printed on 2 printers and scanned on 2 scanners. Dataset D_3 was free of digitization defects and the rotation angle ranged from $1-7°$. The results of table linking and recognition on dataset D_3 are shown in Table 3. It can be concluded that the model is workable when scanning at a resolution of 75 dpi or higher.

DAY AND DATE	SUBJECT CODE	SUBJECT
TUESDAY,6TH MARCH,2018	002	HINDI COURSE-A
MONDAY,12TH MARCH,2018	101	ENGLISH COMMUNICATIVE
FRIDAY,16TH MARCH,2018	086	SCIENCE-THEORY
THURSDAY,22ND MARCH,2018	087	SOCIAL SCIENCE
WEDNESDAY,28TH MARCH,2018	041	MATHEMATICS
MONDAY,2ND APRIL,2018	122	COMM.SANSKRIT

a)

Table A.3: Overall interview and examination response rates, NHANES III, 1988-94

Status	Sample size	Percent	Weighted Percent
Total	39695	100.0	100.0
Not interviewed	5701	14.4	18.2
Interviewed, not examined	2683	6.8	7.5
MEC examined	30818	77.6	73.4
Home examined	493	1.2	0.8

Source: The NHANES III data file, 1988-94

b)

	Utfall 2015	Årsplan 2016	Årsplan 2017
Intäkter	2 139	1 813	
Kostnader	-8 512	-8 822	
Avskrivningar	-244	-239	
S:a verksamhetens nettokostnader	**-6 617**	**-7 248**	**-7 629**
Skatteintäkter	6 066	6 415	6 730
Statligt utjämningssystem, generella statsbidrag	875	881	1 049
Finansnetto	121	52	42
Årets resultat	**445**	**100**	**192**

c)

Fig. 5. Examples of noise-free and non-rotated images from datasets [20, 21]

4. Discussion

The proposed technology for table search and table cell recognition consists of the following steps:

- page localization;
- page normalization;
- image processing;
- vectorization (line detector);

Table 3

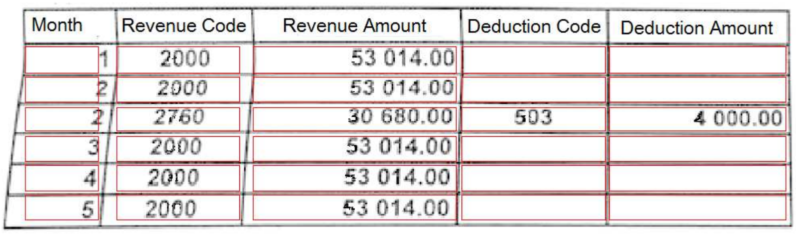
Table linking accuracy and average table cell recognition accuracy for dataset D_2

dpi	percentage rejection tables	accuracy of recognition
300	0%	100%
200		
150		
100		
75	0%	99,80%
50	30 – 40%	50 – 60%

- table layout detector;
- if the layout is rotated then the image is rotated by 180° and the first step is performed;
- table cell detector;
- recognition cells via ANN (LSTM).

The method described in this paper (table detector) links the line detector mechanism and OCR. To a large extent, the success of table and table cell extraction is based on vectorization capabilities. The used table detector can extract not only clear segments, but also segments distorted by foreign objects, and segments broken into parts. The used table detector is adjusted in such a way that the errors of selecting segments broken into parts are about 2 times greater than the errors of selecting segments distorted by foreign objects. This results in the fact that more failures of the method are observed on overexposed document images than on noisy images.

The proposed method is focused on images in which the dividing lines of the sheet line system are converted into segments after normalization. However, the method also works on images in which the segments are represented by curves. Fig. 6 shows a table digitized using the camera of a mobile device. In Fig. 6 it can be seen that all the segments were found using Table detector. However, due to digitization defects, condition (1) is not always satisfied. In other words, in the left part of the table the vertical segments $Q(S_v)$ have width $|P^{x_1}(S_v) - P^{x_2}(S_v)|$ comparable to the height $|P^{y_1}(S_v) - P^{y_4}(S_v)|$. Construction of the projection according to formula (2) allows to find approximate cell boundaries for such segments as well.



Month	Revenue Code	Revenue Amount	Deduction Code	Deduction Amount
1	2000	53 014.00		
2	2000	53 014.00		
2	2760	30 680.00	503	4 000.00
3	2000	53 014.00		
4	2000	53 014.00		
5	2000	53 014.00		

Fig. 6. Example of a digitized 2-NDFL table using a mobile device camera

Conclusion

The paper proposes an algorithm for detecting tables and table cell boundaries for a 2-NDFL tax document. This document is of great interest because it contains several simple tables. Each table has a constant number of columns. The set of tables is not constant. The cell boundaries of the tables are not constant as well. It is possible to move a part of the table to the next page of the document.

A simple method of checking whether a sequence of image rows belongs to a table of a certain type is proposed. For this purpose, the structural method of analyzing the projection of segment areas is applied. A layout model is proposed for table area detection. A layout consists of a sequence of several tables with a previously known description of each table. The use of a dictionary of possible layouts ensures the reliability of the linking of all tables in the document image.

The proposed algorithm uses a set of segments found by the vectorizer in the normalized image as input data. The result of the algorithm is either a set of table cells, or an indication of the need to rotate the image by 180°, or an indication that the set of tables is incorrect.

The peculiarity of the algorithm is that word recognition is not applied to find tables. Recognition is applied only to valid images after the algorithm is completed. This explains the high speed of the algorithm (0,02 – 0,1 milliseconds on Intel® Core™ i9-9900 3.60 GHz, DDR 2666 MHz).

The limits of the algorithm's applicability are determined by the ability of the vectorizer to detect segment boundaries in noisy and distorted images. The conducted experiments have proved that

- for medium and high-quality scans, table cells of the 2-NDFL are detected with no errors;
- for noisy and distorted images, the table area detection error of the 2-NDFL does not exceed 2%, and the table cell detection error does not exceed 2,5%;
- for digital photos of 2-NDFL documents the table search error depends on the success of solving the tasks of searching and restoring the document sheet boundaries.

The proposed algorithm can be applied to find table cells in documents containing tables with a known set of columns.

References

1. Vasiliev S.S., Korobkin D.M., Kravets A.G., Fomenkov S.A., Kolesnikov S.G. Extraction of Cyber-Physical Systems Inventions Structural Elements Of Russian-Language Patents. *Cyber-Physical Systems: Advances in Design and Modelling. Studies in Systems, Decision and Control*, 2020, vol. 259, pp. 55–68. DOI: 10.1007/978-3-030-32579-4_5

2. Slavin O., Arlazarov V., Tarkhanov I. Models and Methods Flexible Documents Matching Based on the Recognized Words. *Cyber-Physical Systems: Advances in Design and Modelling. Studies in Systems, Decision and Control. Springer Nature Switzerland AG*, 2021, vol. 350, pp. 173–184. DOI: 10.1007/978-3-030-67892-0_15

3. Varma O., Srivastava S., Gayathri M. Technical Invoice Data Extraction System: State of the Art, Research Challenges and Countermeasures. *Ambient Communications and Computer Systems. Lecture Notes in Networks and Systems*, 2022, vol. 356, pp. 201–210. DOI: 10.1007/978-981-16-7952-0_19

4. Pegu B., Singh M., Agarwal A., Mitra A., Singh K. Table Structure Recognition Using CoDec Encoder-Decoder. *Document Analysis and Recognition – ICDAR 2021 Workshops. Lecture Notes in Computer Science*, Lausanne, 2021, vol. 12917, pp. 66–80. DOI: 10.1007/978-3-030-86159-9_5
5. Siddiqui S.A., Khan P.I., Dengel A., Ahmed S. Rethinking Semantic Segmentation for Table Structure Recognition in Documents. *Proceedings of the International Conference on Document Analysis and Recognition*, Sydney, 2019, pp. 1397–1402. DOI: 10.1109/ICDAR.2019.00225
6. Gilani A., Qasim S.R., Malik I., Shafait F. Table Detection using Deep Learning. *Proceedings of the International Conference on Document Analysis and Recognition*, Kyoto, 2017, pp. 771–776. DOI: 10.1109/ICDAR.2017.131
7. Gatos B., Danatsas D., Pratikakis I., Perantonis S.J. Automatic table detection in document images. *Pattern Recognition and Data Mining, Third International Conference on Advances in Pattern Recognition*, Bath, 2005, vol. 3686, pp. 609–618. DOI: 10.1007/11551188_67
8. Siddiqui S.A., Malik M.I., Agne S., Dengel A., Ahmed S. DeCNT: Deep Deformable CNN for Table Detection. *A Multidisciplinary Open Access Journal*, 2018, vol. 6, pp. 74151–74161, DOI: 10.1109/ACCESS.2018.2880211
9. Prasad D., Gadpal A., Kapadni K., Visave M., Sultanpure K. CascadeTabNet: an Approach for End to End Table Detection and Structure Recognition from Image-Based Documents. *Conference on Computer Vision and Pattern Recognition Workshops*, Seattle, 2020, pp. 572–573. DOI: 10.1109/CVPRW50498.2020.00294
10. Gobel M., Hassan T., Oro E., Orsi G. ICDAR 2013 Table Competition. *Proceedings of the International Conference on Document Analysis and Recognition*, Washington, 2013, pp. 1449–1453. DOI: 10.1109/ICDAR.2013.292
11. Qiao Liang, Li Zaisheng, Cheng Zhanzhan, et al. LGPMA: Complicated Table Structure Recognition with Local and Global Pyramid Mask Alignment. *Document Analysis and Recognition*, 2021, article ID: 12821. DOI: 10.1007/978-3-030-86549-8_7
12. Liu Ying, Bai K., Mitra P., Giles C.L. Improving the Table Boundary Detection in PDFs by Fixing the Sequence Error of the Sparse Lines. *International Conference on Document Analysis and Recognition*, Barcelona, 2009, pp. 1006–1010. DOI: 10.1109/ICDAR.2009.138
13. Liu Ying., Mitra P., Giles C.L. Identifying Table Boundaries in Digital Documents via Sparse Line Detection. Proceedings of the Conference on Information and Knowledge Management, Napa Valley, 2008, pp. 1311–1320. DOI: 10.1145/1458082.1458255
14. Paliwal S.S., Vishwanath D., Rahul R., Sharma M., Vig L. Tablenet: Deep Learning Model for End-To-End Table Detection and Tabular Data Extraction from Scanned Document Images. *International Conference on Document Analysis and Recognition*, 2020, pp. 128–133. DOI: 10.1109/ICDAR.2019.00029
15. Schreiber S., Agne S., Wolf I., Dengel A., Ahmed S. Deepdesrt: Deep Learning for De-Tection and Structure Recognition of Tables in Document Images. *International Conference on Document Analysis and Recognition*, Kyoto, 2017, pp. 1162–1167. DOI: 10.1109/ICDAR.2017.192
16. Siddiqui S.A., Fateh I.A., Rizvi S.T.R., Dengel A., Ahmed S. Deeptabstr: Deep Learning Based Table Structure Recognition. *International Conference on Document Analysis and Recognition*, Sydney, 2019, pp. 1403–1409. DOI: 10.1109/ICDAR.2019.00226
17. Siddiqui S.A., Khan P.I., Dengel A., Ahmed S. Rethinking Semantic Segmentation for Table Structure Recognition in Documents. *International Conference on Document Analysis and Recognition*, Sydney, 2019, pp. 1397–1402. DOI: 10.1109/ICDAR.2019.00225

18. Tensmeyer C., Morariu V.I., Price B.L., Cohen S., Martinez T.R. Deep Splitting and Merging for Table Structure Decomposition. *International Conference on Document Analysis and Recognition*, Sydney, 2019, pp. 114–121. DOI: 10.1109/ICDAR.2019.00027
19. Palm R.B., Winther O., Laws F. CloudScan – A Configuration-Free Invoice Analysis System Using Recurrent Neural Networks. *International Conference on Document Analysis and Recognition*, Kyoto, 2017, pp. 406–413. DOI: 10.1109/ICDAR.2017.74
20. Li Minghao, Cui Lei, Huang Shaohan, Wei Furu, Zhou Ming, Li Zhoujun. TableBank: A Benchmark Dataset for Table Detection and Recognition. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, 2020, pp. 1918–1925.
21. *ICDAR 2013.* Available at: https://paperswithcode.com/dataset/icdar-2013 (accessed on 08.10.2023)
22. *Smart Document Engine – Automatic Analysis and Data Extraction from Business Documents for Desktop, Server and Mobile Platforms.* Available at: https://smartengines.com/ocr-engines/document-scanner (accessed on 09.10.2023)

Received November 14, 2023

УДК 004.932.72'1 **DOI: 10.14529/mmp240107**

ТЕХНОЛОГИЯ РАСПОЗНАВАНИЯ ТАБЛИЦ В НАЛОГОВЫХ ДОКУМЕНТАХ РФ

О.А. Славин[1,2]
[1]Федеральный исследовательский центр «Информатика и управление» РАН, г. Москва, Российская Федерация
[2]ООО «Смарт Энджинс Сервис», г. Москва, Российская Федерация

Рассматривается известная задача распознавания ячеек таблиц на изображении. Исследуется обработка налогового российского документа 2-НДФЛ. Несмотря на простую структуру таблиц, способ печати основан на гибком шаблоне. Гибкость формы наблюдается как в части модификаций текстовой информации, так и в области таблиц. Гибкость таблиц состоит в изменении числа и размеров столбцов. Для детектирования таблиц был предложен структурный метод. Входными данными метода являются детектированные горизонтальные и вертикальные отрезки. Поиск отрезков проводился механизмами, реализованными в системе Smart Document Reader. Апробация и внедрение предложенного метода также осуществлялось в системе Smart Document Reader. Кроме детектирования области предполагаемого размещения таблиц решены следующие задачи: поиск ячеек таблиц, именование ячеек таблиц, валидация области таблицы. Валидация области таблицы проводилась для отдельных таблиц, а также для совокупностей таблиц. Применение описаний совокупностей таблиц обеспечило высокую надежность привязки набора таблиц.

Ключевые слова: распознавание таблиц; детектирование отрезка; раскладка таблиц.

Олег Анатольевич Славин, доктор технических наук, доцент, главный научный сотрудник, федеральный исследовательский центр «Информатика и управление» РАН (г. Москва, Российская Федерация); старший научный сотрудник-программист, ООО «Смарт Энджинс Сервис» (г. Москва, Российская Федерация), oslavin@isa.ru.

Поступила в редакцию 14 ноября 2023 г.