



HAL
open science

A New Table Extraction and Recovery Methodology with Little Use of Previous Knowledge

Luiz Antônio Pereira Neves, João Marques De Carvalho, Jacques Facon,
Flávio Bortolozzi

► **To cite this version:**

Luiz Antônio Pereira Neves, João Marques De Carvalho, Jacques Facon, Flávio Bortolozzi. A New Table Extraction and Recovery Methodology with Little Use of Previous Knowledge. Tenth International Workshop on Frontiers in Handwriting Recognition, Université de Rennes 1, Oct 2006, La Baule (France). inria-00104719

HAL Id: inria-00104719

<https://inria.hal.science/inria-00104719>

Submitted on 9 Oct 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A New Table Extraction and Recovery Methodology with Little Use of Previous Knowledge

Luiz Antônio Pereira Neves*,** João Marques de Carvalho*

Jacques Facon**
Flávio Bortolozzi**

*UFCG - Universidade Federal de Campina Grande, Brazil

**PUCPR - Pontifícia Universidade Católica do Paraná, Brazil

neves@ppgia.pucpr.br, carvalho@dee.ufcg.edu.br, facon@ppgia.pucpr.br, fborto@ppgia.pucpr.br

Abstract

A new methodology for table-form extraction and recovery with little previous knowledge is presented. The first module performs the identification of line intersections in a table-form, the second module detects and corrects wrong intersections produced by fault intersection segments or by table artefacts (smudges, overlapping of handwritten data and fault segments). In this module, an artefact identification method for handwritten filled table-forms is proposed. The proposed method aims to detect, identify and remove table-form artefacts with little use of previous knowledge. The third module performs the table-form cell extraction.

The evaluation of the efficiency is carried out from a total of 305 table-form images. Experiments showed significant and promising results. The artefact identification method improves table-form interpretation rates. The proposed approach reached a successful rate up to 85%. The main advantage of the presented methodology is requiring little knowledge from documents, being able to apply for most of the table-forms.

Keywords: Table-form recognition, Table-form extraction, handwritten data, Document segmentation.

1. Introduction

Tables, or table-forms, are documents composed by cells, which are determined by intersections of straight line segments, as illustrated in figure 1.

TEMPO		FLUXO		
		AUTOMÓVEL	ÔNIBUS	CAMINHÃO
HORA	MINUTO	BARRAS	BARRAS	BARRAS
	:00			
	:15			
HORA	MINUTO	BARRAS	BARRAS	BARRAS
	:15			
	:30			

Figure 1. Example of a table-form document.

Several studies have been presented on table recognition [1],[3],[6],[7],[8],[15],[16]. Some of these researches use tables without imperfections in the horizontal and vertical segments to reduce the complexity of the problem. In the damaged table case, many researchers use previous knowledge for their interpretations, aiming to minimize its complexity. Figure 2 shows some artefacts, which may be present in a table-form image, which can be a handwritten draft (a), overlapping of handwritten data with the segments of lines that make up the cell (b), and flaws of these line segments (c).

Figure 2 shows a form titled 'Ficha de Dados do Projeto XForm' with the year '2020'. The form contains fields for Name, e-mail, Número ICG, Nome da Guerra, Curso, Período, Turma, Quantas horas você disponibiliza na semana para estudar?, Número de Chamada, and Número de Sala. Handwritten entries are present in several fields. Three callouts labeled (a), (b), and (c) point to specific artefacts: (a) points to a smudge in the 'Número ICG' field, (b) points to overlapping handwritten data in the 'Período' field, and (c) points to a flaw in the 'Quantas horas' field.

Figure 2. Example of artefacts.

Therefore, the important table interpretation problems, shown in figure 2, are:

- Problem 1 – P1 - Presence of overlapping data.
- Problem 2 – P2 - Presence of handwritten drafts.
- Problem 3 – P3 - Imperfections of the table-form segments.

These problems have been partially solved by other researchers, such as:

1. P1, P2 and P3: Watanabe [17] [18] presents two procedures, one when the information is not previously known and another when it stores artefacts (noise) characteristics in the knowledge base. Couasnon [3] uses previous noise and imperfections knowledge as grammar rules. Tran van Thom [15] makes the reduction of the image and uses some thresholdings for detecting and correcting the segments with imperfections.
2. P2 and P3: Arias and Kasturi [1][2] use the morphological closing operator to eliminate

imperfections and to recover the extinguished segment lines for the analysis of the table-form intersections. Liang et al. [9] consider noise and imperfections as previous knowledge. Doermann [6] and Hirano et al. [5] use table models with noise and imperfections. Pizano [12] makes the reduction of image to eliminate the noise and segments. Later, he uses the minimum parameters of width and distance to eliminate the remaining noise;

3. P3: Shinjo et al. [14] use previous knowledge to detect and correct damage of table corners. Shimotsuji and Asano [13] use table models with imperfections as previous knowledge in order to identify the interpretation process. Lopresti [7] uses a table analyzer which has information of the distances among the lines of the table. Fan et al. [4] analyze the known distances among the points of clusters through the grouping technique.

The challenge is to reduce the use of knowledge for table-form understanding. The use of term “little knowledge” means processing closed table forms without rounded corners. The present work does not need previous knowledge about number of cells, document skew, handwritten and preprinted data, interrupted segments or data overlaps.

To solve problems P1, P2, and P3, we present, in this paper, a methodology that allows the table extraction and recovery using little priori knowledge. This article is organized as follows:

In Section 2, we describe our approach for table-form extraction and recovery with little knowledge. Section 3 presents some experimental results and discussions. Finally, the conclusions are given in Section 4.

2. Methodology

The methodology is developed in three steps, as shown in the figure 3:

2.1. Step 1. Identification of table-form intersections

The morphological structuring elements of the table corners as in figure 4 are used aiming to detect the intersection seeds from the segments of horizontal and vertical lines of the table. These structuring elements were built with 36 pixels because, in this search, this size was analyzed as the best size to process most types of table-forms. The cell extraction method consists in initially locating and extracting the intersections of those lines, so as to deduct cell position and shape. In this step, 10 intersection models could be used, 9 of them are representing hierarchically according to [2] by means of numbers (figure 4). The intersection location method is based on the use of the binary mathematical dilation [10] [11] and on the use of 9 structuring elements having the same aspect of the intersection.

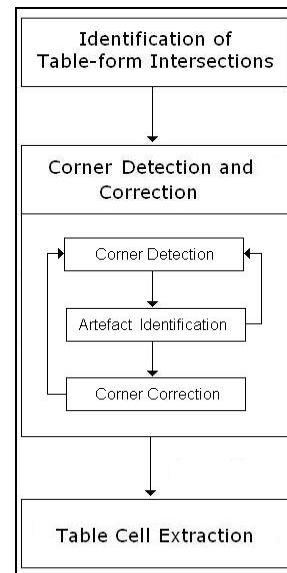


Figure 3. Proposed Methodology.

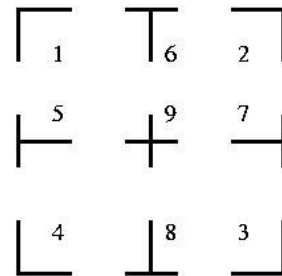


Figure 4. Representation of the nine intersections.

After the localization step, the creation of image union is made, with the union of all intersection images, as shown in figure 5 from processed image of figure 1, for generation of physical and real arrays.

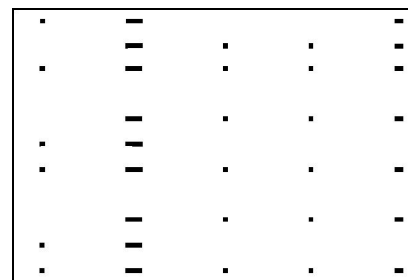


Figure 5. Union of all intersections.

The first array, called **physical array**, contains information about the physical table structure, as well as, number of rows, number of columns and positions of lines as shown in figure 6. This array has been built after using the morphological binary reconstruction on the image union and the data were obtained from its contents by horizontal and vertical projection profiles [10][11]. The second array, called **real array**, contains the

identification and location of each intersection found into the used table-form. Based on the evaluation of physical array and on the created images mentioned above, the data were extracted to build the real array. Each number in real array corresponds to an intersection of the table-form as shown in figure 7.

Physical Array	
Number of rows: 9	
row [0]:	19
row [1]:	61
row [2]:	99
row [3]:	183
row [4]:	226
row [5]:	269
row [6]:	353
row [7]:	396
row [8]:	439
Number of columns: 5	
column [0]:	46
column [1]:	200
column [2]:	353
column [3]:	497
column [4]:	645

Figure 6. Physical array of interpreted image 1.

1	6	0	0	2
0	5	6	6	7
5	9	9	9	7
0	5	9	9	7
5	7	0	0	0
5	9	9	9	7
0	5	9	9	7
5	7	0	0	0
4	8	8	8	3

Figure 7. Real array of interpreted image 1.

2.2. Step 2. Corner detection and correction.

The process of detecting errors in the physical structure aims to analyze, verify and identify the possible errors originated in the previous identification step from real array. To allow the automation of the search and detection of errors in the physical structure, Rejection Tables following the North-South, West-East, North-East, North-West, South-East and South-West directions of the neighborhood of each intersection were prepared for each type of intersections 1 to 9 (figures 8 and 9). The Rejection Tables store the incompatible neighboring corner of analyzed intersection as shown in figure 10. The underlying principle of this process is the neighborhood analysis of each intersection in these six directions and compare real array neighborhood with the rejection tables neighborhood (error detection); and acceptance tables in the error correction process. Analyzing the figure 2-c, the imperfections of the table-form segments are corrected by means of the Rejection Tables.

To register the errors, which were found in this step, the error counters were created. Every time an

identification error is detected, the respective counter increases.

Since the table-forms processed here can be filled in by machines or by hand, overlapping printed or handwritten information (seen figure 2(a) and figure 2(b)) might create false intersections. These occurrences are called artefacts. In next section, the artefact identification method is described.

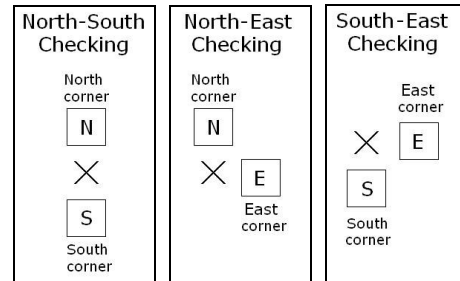


Figure 8. Types of rejection tables in North-South, North-East and South-East directions.

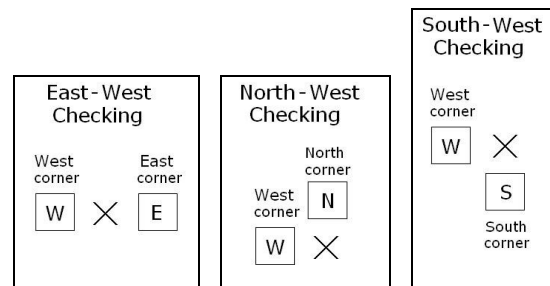


Figure 9. Types of rejection tables in East-West, North-West and South-West directions.

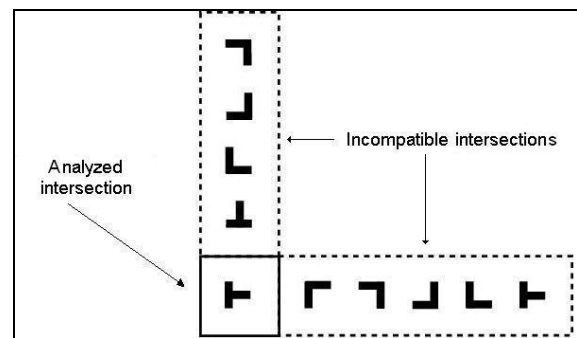


Figure 10. Example of rejection table of an intersection.

2.2.1. Artefact Identification

The proposed artefact identification method is based on compactness analysis. Compactness is a property that expresses how large the area concentrated inside a given perimeter is, as shown in figure 11. Compactness is measured by the compactness factor, computed from the perimeter and the area of the analyzed shape. Given a shape of perimeter P and area S , its compactness factor is given by FC , as shown in the equation 1.

$$F_c = \frac{P^2}{4\pi S} \quad (1)$$

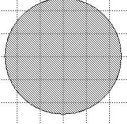
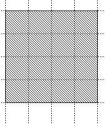
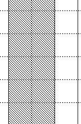







Shapes/Data				
	(a)	(b)	(c)	(d)
Perimeter	16	16	16	20
Area	20.3721	16	12	10
Compactness Factor	1	1.2732	1.6976	3.1830







Figure 11. Compactness analysis for artefact identification.

Verifying the shapes in the figure 11, the circle (figure 11-a) presents the best compactness and we can say that, in general, table-form artefacts present high compactness, with values equal or around 1. Thereby, a threshold has been created for distinguishing if the value calculated for the compactness factor corresponds to that of an artefact or to a straight line segment of a table cell. For determining threshold value, compactness factors from more than 30 different artefacts were submitted to exploratory data analysis [19], characterizing a homogenous distribution with a confidence level of 99%. The range of variation $\mu \pm 2.576\sigma$, where μ and σ are the mean and standard deviation respectively, produces inferior and superior limits of 1.21688 and 1.37419, respectively. The 0.5% of values above the superior limit are not contemplated as artefacts. Therefore, the handwritten data that presents compactness factor below 1.4 is considered an artefact.

Figures 12-a and 12-b show several types of artefacts with the respective compactness factors.

index	artefact	compactness factor
(a)		1.28998
(b)		1.27324
(c)		1.34486
(d)		1.28857
(e)		1.27947
(f)		1.28547

(a)

index	artefact	compactness factor
(a)		1.29841
(b)		1.28477
(c)		1.27404
(d)		1.28803
(e)		1.29628
(f)		1.28137

(b)

Figure 12. Examples of Artefacts with their compactness factors.

Figure 13 shows some table segments with compactness factor values above the established threshold. For figure 13-d, for instance, the compactness factor is 5.27407. This value indicates that the analyzed

object is not an artefact, but rather a segment. Therefore, by observing figures 12-a, 12-b and 13, one can conclude that the artefact identification method can make the correct distinction between a handwritten artefact and a table segment.

Then, the proposed artefact identification approach is inserted into **Step 2**.







index	segment	compactness factor
(a)		1.94365
(b)		4.98904
(c)		1.59781
(d)		5.27407
(e)		3.17814
(f)		1.96736

Figure 13. Examples of segments that are not artefacts with their compactness factors.

The used method in the correction module is based on the idea that a wrong intersection has correct neighboring intersections that will allow the reestablishment of the correct situation. For that purpose, acceptance tables were developed for each one of the intersections. The strategy used during error detection is used again, as illustrated in figure 3.

2.3. Step 3. Table cell extraction.

Extraction of the table cells consists on the interpretation of the table logical structure [10]. The interpretation of the cells is made through the analysis of the identified corners, verifying which corner makes up the cell. Therefore, we propose the use of an algorithm for the validation of the analyzed corners, using the information of each corner.

The extraction of all the cells is made through scanning in order to cover all intersections of the table. For verifying the result of the interpretation, the interpretative image is created that shows the extracted cells and the logical structure of table-form, as illustrated in figures 15 and 16 from analyzed table of figure 14, without and with the use of artefact analysis respectively. However, in figure 16, the use of the artefact identification produces the correct table interpretation.

2.4. Artefact cases

Figures 17 and 18 illustrate cases where the artefact analysis method does not perform correct artefact identification. This happens because the handwritten letter *f* (figure 17), as well as the handwritten digits *1* (figure 18) are similar to the table lines. The resulting shapes produce high compactness factors and the method does not consider them as artefacts. These cases represent challenges that will be the subject of further studies.

TEMPO		FLUXO		
1*		AUTOMÓVEL	ÔNIBUS	CAMINHÃO
HORA	MINUTO	BARRAS	BARRAS	BARRAS
7:00	:15	15	3	10
8:15	:30	30	10	4

Figure 14. Example of table-form with artefacts.

Figure 15. Interpretative image of figure 14, without artefact analysis.

Figure 16. Interpretative image of figure 14, with artefact analysis.

3952870	
Nome de Guerra	
Professor	
Período	Tur
5 ^o	

Figure 17. Case 1 of table with artefacts.

01	08	144	01	18
BARRAS	BARRAS	BARRAS	BARRAS	BARRAS

Figure 18. Case 2 of table with artefacts.

3. Experimental results and analysis

To evaluate the performance of the proposed artefact identification approach, 305 table-form images were used to compose the test database, as exemplified in figures 19, 20 and 21. These table-form images, scanned at 300 dpi, are filled with handwritten data, handwritten overlap, and contain artefacts.

TEMPO	FLUXO 1 (1)			FLUXO 2 (2)			FLUXO 3 (3)		
	AUTOMÓVEL	ÔNIBUS	CAMINHÃO	AUTOMÓVEL	ÔNIBUS	CAMINHÃO	AUTOMÓVEL	ÔNIBUS	CAMINHÃO
HORA	MINUTO	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS
08:00	08:15	59	2	24	1	1	13	1	01
08:15	08:30	100	1	9	3	8	1	5	1
08:30	08:45	104	1	5	14	-	3	18	-
08:45	09:00	110	-	14	20	-	3	15	-

Figure 19. Example of table-form in the base of tests.

Ficha de Dados do Projeto XForm		Ano
Nome		2005
e-mail		EVERTON LUIS ESTEVES
Número ICG		52466741
Nome de Guerra		Red Devil Rider
Curso	Período	Turma
Eng. Ciência da Computação	6 ^o	U
Quantas horas você disponibiliza na semana para estudar?	Número de Chamada	Número de Sala
2 Horas	03	05

Figure 20. Example 2 of table-form in the base of tests.

Ficha de Dados do Projeto XForm		Nome
Número ICG		EVERTON LUIS ESTEVES
e-mail		52466741
Nome de Guerra		erertone@pac.pn.br
Cursos		Red Devil Rider
Eng. Ciência da Computação	Período	Turma
6 ^o	U	
Quantas horas você disponibiliza na semana para estudar?	Número de Chamada	Número de Sala
2 Horas por semana	03	05

Figure 21. Example 3 of table-form in the base of tests.

Tests were carried out with and without artefact analysis in order to quantify the improvement produced by the proposed approach. The rate of processed images, shown in table 1, indicates the percentage of images that went through all steps of the methodology. Rejected images are those that did not reach the final processing stage of the methodology. Correctly interpreted images are images that presented no interpretation errors, i.e., their contents were 100% correctly interpreted. Initially, with no artefact analysis, 211 images (69%), were correctly processed and 94 images (31%) were rejected.

From the 211 correctly processed images, 196 (64%) were correctly interpreted. The process was then repeated applying artefact analysis. 299 images (98%) were correctly processed and 6 images (2%) were rejected. For the 299 processed images, 260 (85%) were correctly interpreted. A significant result that can be observed is that without artefact analysis, 31% of the table-form images in the base were rejected, whereas this index decreased to 2%, keeping an index of 85% for correctly interpreted images, by applying artefact analysis. These results are summarized in table 1.

Table 1. Summarized results of tests with 350 images.

Method	Rate of processed images	Rate of rejected images	Rate of correctly interpreted images
Without using artefact analysis	211 (69%)	94 (31%)	196 (64%)
With using artefact analysis	299 (98%)	6 (2%)	260 (85%)

4. Conclusions

The results show that using “little knowledge” does not damage the efficiency of the methodology (85% rate of correct identification). The unsolved errors derive mainly from the handwritten data. Based on the variation interval $\mu \pm 2.576\sigma$, on the coefficient of Pearson and on the compactness property, the proposed artefact identification method has shown to be effective in identifying different kinds of artefacts.

Summarizing the advantages of the approach, we mention the possibility of applying it to different types of handwritten filled table-forms for identification of handwritten smudges, as well as the intersection defects, all that with very little use of *a priori* knowledge.

Acknowledgments

We would like to acknowledge support for this research from UFCG, PUCPR and the PROCAD Program from CAPES/MEC (Brazilian government, project number 153/01-1).

5. References

[1] J. F. Arias, A. Chhabra, and V. Misra. Interpreting and Representing Tabular Documents. CVPR 1996 - IEEE - In: Proceedings of the Conference on Computer Society Conference on Computer Vision and Pattern Recognition, pages 600–605, 1996.

[2] J. F. Arias, R. Kasturi, and A. Chhabra. Efficient Techniques for Telephone Company Line Drawing Interpretation. ICDAR 1995 - IEEE - In: Proceedings of the Third International Conference on Document Analysis and Recognition, pages 795–798, 1995.

[3] B. Couasnon. Dmos: A generic document recognition method, application to an automatic generator of musical scores, mathematical formulae and table structures recognition systems. ICDAR 2001 - In: Proceedings of the Sixth International Conference on Document Analysis and Recognition, pages 215–220, 2001.

[4] K.-C. Fan, J.-M. Lu, L.-S. Wang, and H.-Y. Liao. Extration of characters from form documents by feature point clustering. Pattern Recognition Letters, 1995.

[5] T. Hirano, Y. Okada, and F. Yoda. Field extraction method from existing forms transmitted by facsimile. ICDAR 2001 - In: Proceedings of the Sixth International Conference on Document Analysis and Recognition, pages 738–742, 2001.

[6] O. Hori and D. S. Doermann. Robust table-form structure analysis based on box-driven reasoning. ICDAR 1995 - In: Proceedings of the Third International Conference on Document Analysis and Recognition, pages 218–221, 1995.

[7] J. Hu, R. S. Kashi, D. Lopresti, and G. T. Wilfong. Evaluating the performance of table processing algorithms. International Journal on Document Analysis and Recognition, 4:140–153, 2002.

[8] T. Kieninger and A. Dengel. The t-recs table recognition and analysis system. In: DAS’98 - Proceedings of the Sixth International Conference on Document Analysis Systems, pages 255–269, 1998.

[9] J. Liang, J. Ha, R. M. Haralick, and I. T. Phillips. Document layout structure extraction using bounding boxes of different entities. WACV 1996 In: Proceedings of the Third IEEE Workshop on Applications of Computer Vision, pages 278–283, 1996.

[10] L. A. P. Neves. Extração de células de dados manuscritos em tabelas. Master’s thesis, Pontifícia Universidade Católica do Paraná - PUCPR, Brazil, 1999.

[11] L. A. P. Neves. Metodologia de Extração e Recuperação de Tabelas. PHD’s thesis, Universidade Federal de Campina Grande – UFCG, Paraíba, Brazil, 2006.

[12] A. Pizano. Extracting line features from images of business forms and tables. IAPR - In: Proceedings of the 11th International Conference on Pattern Recognition, 3:399–403, 1992.

[13] S. Shimotsuji and M. Asano. Form Identification based on Cell Structure. ICPR 1996 - IEEE - In: Proceedings of the 12th IAPR International Conference on Pattern Recognition, pages 793–797, 1996.

[14] H. Shinjo, E. Hadano, K. Marukawa, Y. Shima, and H. Sako. A recursive analysis for form cell recognition. ICDAR2001-In:Proceedings of the Sixth International Conference on Document Analysis & Recognition, 2001.

[15] R. T. V. Thom. Modelisation de Tableaux pour le traitement Automatique des Formulaire. Laboratoire PSI, Universit de Rouen, 1997.

[16] T. Watanabe, Q. Luo, and N. Sugie. Structure recognition methods for various types of documents. Machine Vision and Applications, 1993.

[17] T. Watanabe, Q. Luo, and N. Sugie. Toward a practical document understanding of table-form documents: Its framework and knowledge representation. In: Proceedings of the Second Conference on Document Analysis and Recognition, pages 510–515, 1993.

[18] T. Watanabe, Q. Luo, and N. Sugie. Layout recognition of multi-kinds of table-form documents. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1995.

[19] Tukey, J.W.: Exploratory Data Analysis. Addison-Wesley, 1977.