Journal of Universal Computer Science, vol. 14, no. 2 (2008), 252-265
submitted: 15/6/07, accepted: 17/11/07, appeared: 28/1/08 © J.UCS

Table-form Extraction with Artefact Removal

Luiz Antônio Pereira Neves
(PUCPR, Brazil,
neves@ppgia.pucpr.br)

João Marques de Carvalho
(UFCG, Brazil
carvalho@dee.ufcg.edu.br)

Jacques Facon
(PUCPR, Brazil,
facon@ppgia.pucpr.br)

Flávio Bortolozzi
(PUCPR, Brazil
fborto@ppgia.pucpr.br)

Abstract: In this paper we present a novel methodology to recognize the layout structure of handwritten filled table-forms. Recognition methodology includes locating line intersections, correcting wrong intersections produced by what we call artefacts (overlapping data, broken segments and smudges), extracting correct table-form cells and using as little previous table-form knowledge as possible. To improve layout structure recognition, a novel artefact identification and deletion method is also proposed. To evaluate the effectiveness of the methodology, a database composed of 350 handwritten filled table-form images damaged by different types of artefacts was used. Experiments show that the artefact identification method improves performance of the table-forms structure extractor that reached a success rate of 85%.

Keywords: Table-form recognition, Table-form extraction, Handwritten data, Document segmentation
Categories: I.4, I.4.6, I.7, I.7.m

1 Introduction

A table-form can be generally defined as a structured document composed by cells delimited by horizontal and vertical line segments. Cells can be blank or filled with data, either printed or handwritten. Figure 1 shows two examples of table-forms.

A table-form recognition system aims to automatically identify the document structure and extract meaningful data from it. Several approaches have been proposed for table-form recognition [Arias et al., 1996] [Couasnon 2001] [Hori at al. 1995] [Hu at al., 2002] [Kieninger at al., 1998] [Thom 1997] [Watanabe at al., 1993a]. Some of the authors use table-form models or not damaged tables (composed by perfect horizontal and vertical line segments) to reduce the complexity of the problem. Although such approaches solve the challenge posed by many table-forms, they are not able to handle almost any variation on the document physical structure. This

drawback increases in the case of damaged tables. For further complexity reduction some researchers use *a priori* knowledge, which is not always effective to overcome damages. For instance, Figure 2 shows some artefacts which may be present in a table-form image, like handwritten draft (a), overlapping of handwritten data with table cells line segments (b), and flaws of the line segments (c).

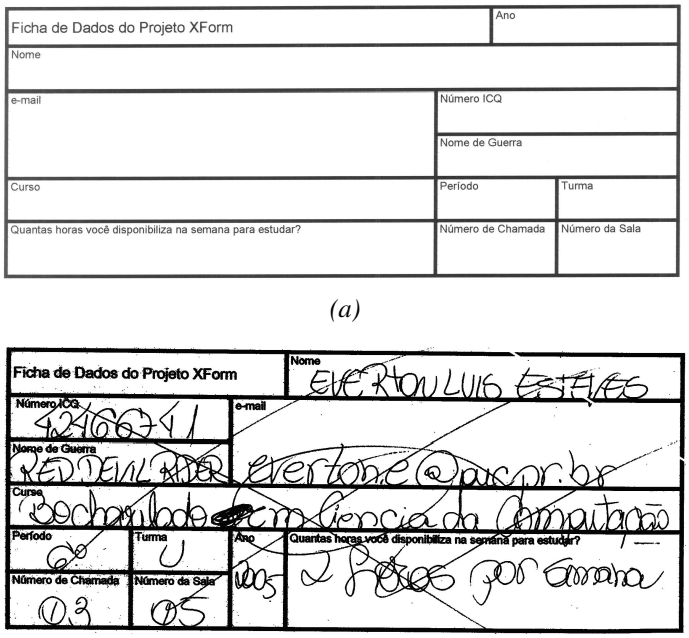


Figure 1: Examples of table-forms with pre-printed fields: (a)blank, and (b) filled with handwritten data.

Figure 2 helps to see the lack of predictive pattern in artefacts and to understand why this can interfere in the performance of a table-form recognition system. We list three main problems related to table interpretation:

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

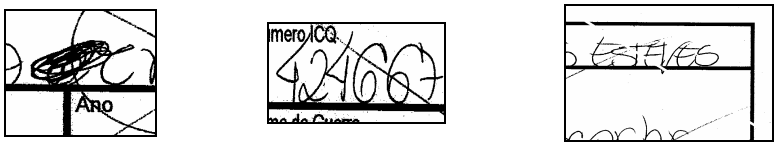


Figure 2: Examples of artefacts in table-forms

Several researchers have partially solved these problems:

- P1: Shinjo [Shinjo at al., 2001] use previous knowledge to detect and correct damaged table corners. Shimotsuji and Asano [Shimotsuji at al., 1996] use table models with imperfections as previous knowledge in order to ease the interpretation process. Hu [Hu at al., 2002] uses a table analyzer with information of the distances among the lines of the table. Fan et al. [Fan at al., 1995] analyze the known distances among points in clusters, through a grouping technique.
- P1 and P3: Arias and Kasturi [Arias et al., 1996] [Arias et al., 1995] use the morphological closing operator to eliminate imperfections and to recover extinguished segment lines for the analysis of table-form intersections. Liang [Liang at al., 1996] considers noise and imperfections as previous knowledge. Doermann [Hori at al., 1995] and Hirano et al. [Hirano at al., 2001] use table models with noise and imperfections. Pizano [Pizano 1992] reduces the image to eliminate noisy segments and uses the minimum parameters of width and distance to eliminate the remaining noise;
- P1, P2 and P3: Watanabe [Watanabe at al., 1993b] [Watanabe at al., 1995] presents two procedures, one to be used when no information is previously known and another which stores artefacts (noise) characteristics in a knowledge base. Couasnon [Couasnon 2001] uses previous noise and imperfections knowledge as grammar rules. Tran van Thom [Thom 1997] reduces the image and uses threshold for detecting and correcting segments with imperfections.

In this paper, we present a table-form recognition methodology able to treat damaged tables. Damages are produced by what we generically call artefacts (overlapping data, broken segments, smudges, etc.), as depicted in Figure 2. Our aim is to solve the above listed problems (P1, P2, P3) and our challenge is to reduce the use of previous table-form knowledge. The only knowledge we assume beforehand is that we deal with closed table-forms with corners formed by line segments that intersect orthogonally. We call that "little knowledge". Our method does not need beforehand information about number of cells, document skew, handwritten and preprinted data, interrupted segments or data overlaps.

The rest of the paper is organized as follows: our approach, consisting of the intersections identification, corner detection and correction and cell extraction stages, is described in Section 2. Experimental results and discussions are presented in Section 3. Section 4 presents conclusions and final remarks.

2 Methodology

The methodology is developed in three steps, namely: Identification of Table-form Intersections, Corner Detection and Correction, and Table Cell Extraction, as illustrated in the Figure 3:

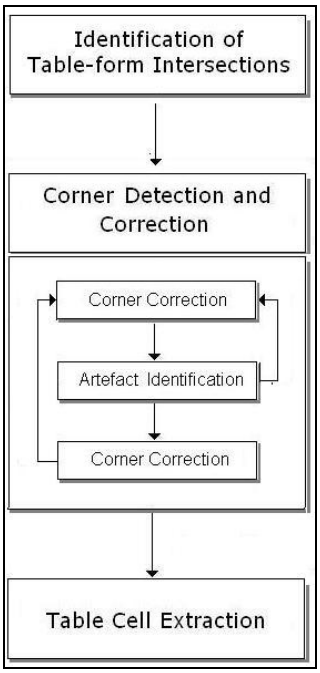


Figure 3: Proposed methodology.

2.1 Step 1. Identification of Table-Form Intersections.

Morphological structuring elements are used to detect seeds corresponding to the cell corner types shown in Figure 4, from the horizontal and vertical line segments intersections in the table. These structuring elements were built with 36 pixels because in our experiments this was found to be the best size to process most types of table-forms. The cell extraction method consists in initially locating and extracting the line intersections, in order to determine cell position and shape. In this step, 9 intersection models are considered, represented hierarchically by numbers (Figure 4) [Arias et al., 1995]. The intersection location method is based on the use of binary mathematical dilation [Neves at al., 2003a] [Neves at al., 2003b] and 9 structuring elements having the same shapes of the intersections. In order to cut down on memory and calculating time, only the structuring elements corresponding to the first 4 intersections (1,2,3,4) shown in Figure 4 are used to find all corner types [Neves at al., 2003a] [Neves at al., 2003b] . This simplification is possible, because the remaining intersection types (5, 6, 7, 8, 9) can be obtained from combinations of the first four types.

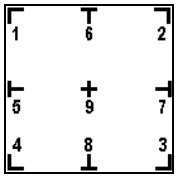
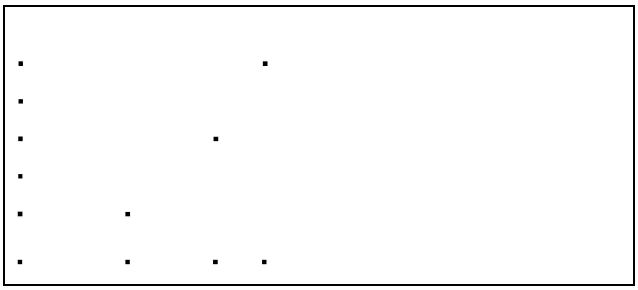
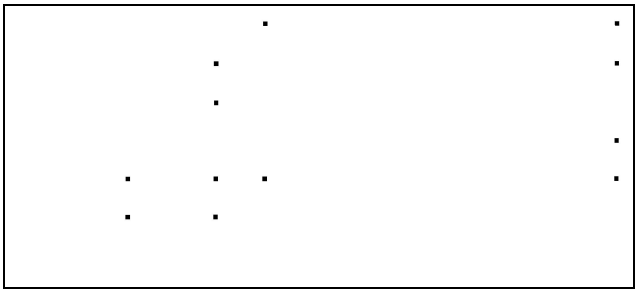


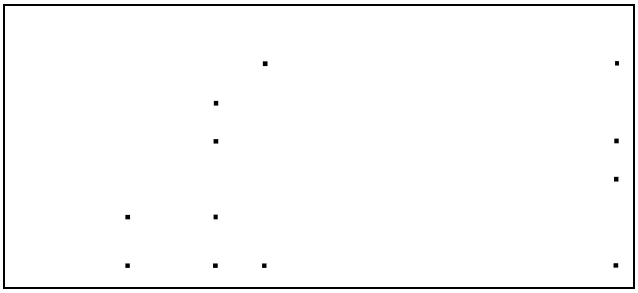
Figure 4: Representation of the nine intersection types



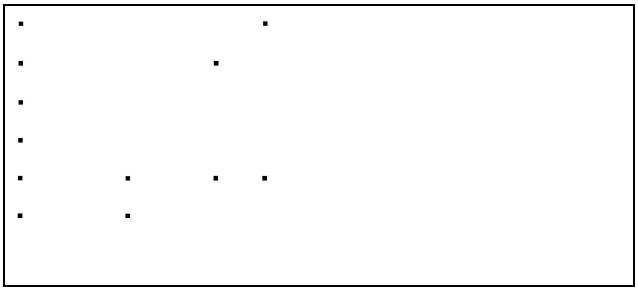
(a)



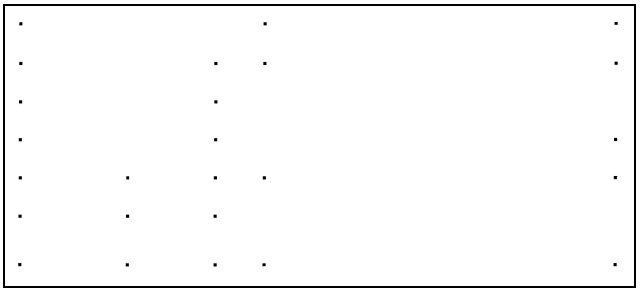
(b)



(c)



(d)



(e)

Figure 5: Identified intersections for table-form in Figure 1.b: (a) image with intersections type 1, 6, 5 and 9; (b) image with intersections type 2, 6, 7 and 9; (c) image with intersections 3, 8, 7 and; (d) image with intersections type 4, 8, 5 and 9; (e) union of all intersections for binary reconstruction.

After the initial locating stage, an image is created with the union of all line intersections previously found. Figures 5 (a, b, c, d and e) show the images obtained from the table-form in Figure 1.b. This image will be used for generation of the physical and real arrays.

The first array, called physical array, contains information about the physical table structure, such as, number of rows, number of columns and lines position, as shown in Figure 6. Data for this array is obtained from horizontal and vertical projection profiles of image union, after submitting it to a morphological binary reconstruction. The second array, called real array, contains the identification and location of each intersection found in the table-form. Data for the real array is obtained from evaluation of the physical array and from the line intersection images generated in the previous steps. Each number in the real array corresponds to an intersection in the table-form, as shown in Figure 7.

2.2 Step 2. Corner detection and correction.

Some of the identified intersections in the table-form may not be genuine ones. Wrong intersections produced by overlapping data, broken segments and smudges can exist, this being the reason why detecting and solving errors is a fundamental step.

Detecting errors in the physical structure is performed by analyzing the real array, trying to verify and identify the possible errors originated in the previous identification steps (Figures 6 and 7). To allow automation of searching and detecting errors in the physical structure, Rejection Tables for the North-South, West-East, North-East, North-West, South-East and South-West neighborhood directions of each table-form intersection were prepared for all intersection types (1 to 9). The Rejection Tables store the incompatible neighboring corners of the analyzed intersection. The underlying principle of this process is to analyze intersection neighborhoods in the six defined directions, by comparing real array neighborhoods with the rejection table neighborhoods (error detection). From this comparison, acceptance tables are generated for the error correction process.

Error counters were created to register the errors found in this step. Each time an identification error is detected the respective counter is incremented, as illustrated in Figure 8.

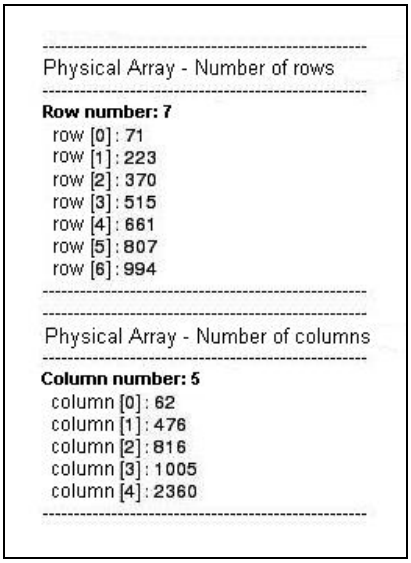


Figure 6: Physical array of interpreted image 1.

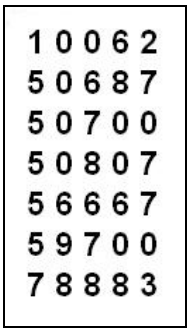
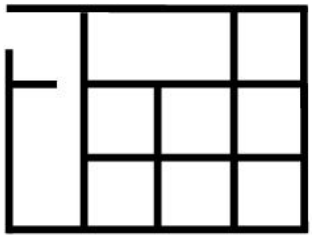


Figure 7: Real array of interpreted image 1.



5	0	0	0	0
2	3	0	0	0
0	0	0	0	0
0	0	0	0	0

(a) (b)

Figure 8: Examples of error counters: (a) table with error, and (b) error identification with error counters, using Rejection Tables

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

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. 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 equation 1.

$$Fc = \frac{P^2}{4\pi S} \qquad (1)$$

The circle 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 [Tukey 1977] [Neves at al. 2006a] [Neves at al. 2006b] , 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 considered as artefacts. Therefore, all handwritten data that presents compactness factor below 1.4 is considered an artefact. Figure 9 shows several types of artefacts with the respective compactness factors.

Figure 10 shows some table segments with compactness factor values above the established threshold. For Figure 10.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 9 and 10, one can conclude that the artefact identification method can make the correct distinction between a handwritten artefact and a table segment.

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

Figure 9: Examples of Artefacts with their compactness factors.

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

Figure 10: Examples of segments that are not artefacts with their compactness factors.

This artefact identification method is inserted into Step 2 (corner detection and correction), to allow for correct corner detection. The method used in the corner correction module is based on the idea that a wrong intersection has correct neighbouring intersections that will allow re-establishing the correct situation. For that purpose, acceptance tables were developed for each one of the intersections, as shown in Figure 11. Finally, the strategy used for error detection is performed again, as illustrated in Figure 3.

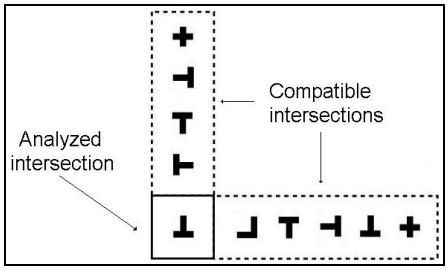


Figure 11: Example of Acceptance table of an intersection.

2.3 Step 3. Table Cell Extraction.

Extracting the table cells consists of interpreting the table logical structure. Cells interpretation is performed through the analysis of the identified corners, verifying which corner makes up the cell, as illustrated in Figure 12. Therefore, we use an algorithm for the validation of the analyzed corners, using the respective corner information.

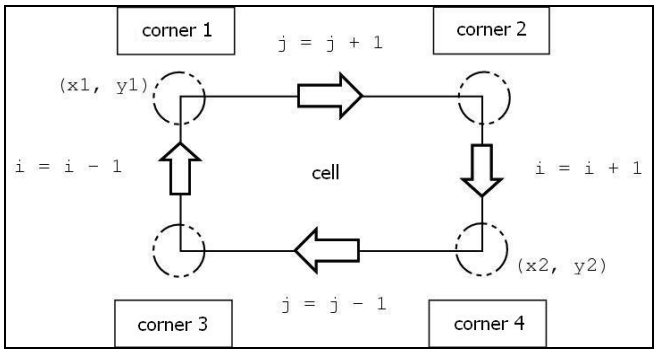


Figure 12: Algorithm of cell extraction.

For verifying the result of the interpretation, an interpretative image is created that shows the extracted cells and the logical structure of the table-form, as illustrated in Figures 13 and 14 (for the analyzed table of Figure 1.b), without and with the use of artefact analysis, respectively. It can be seen that the use of artefact identification, Figure 14, produces the correct table interpretation.

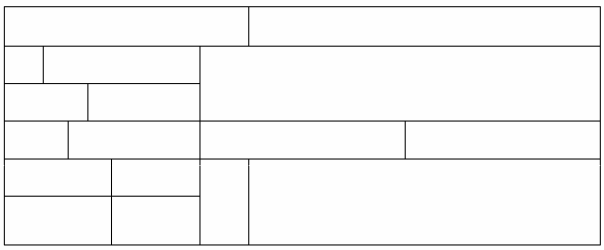


Figure 13: Interpretative image for table-form in Figure 1.b, without artefact analysis

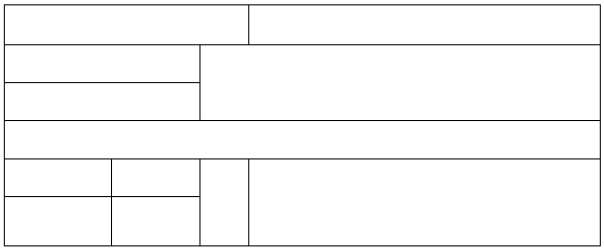


Figure 14: Interpretative image for table-form in Figure 1.b, with artefact analysis

2.4 Artefact cases.

Figures 15 and 16 illustrate cases where the artefact analysis method does not perform correct artefact identification. This happens because the handwritten letter ***f*** (Figure 15), as well as the handwritten digits ***¨1¨*** **(**Figure 16) is 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.

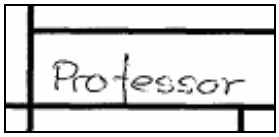


Figure 15: Case 1 with artefacts.

Figure 16: Case 2 with artefacts.

2.5 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. These table-form images, scanned at 300 dpi, are filled with handwritten data with and without overlap and contain different types of artefacts.

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% by applying artefact analysis, with an index of 85% for correctly interpreted images.

These results are summarized in Table 1.

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%)

Table 1: Summarized results of tests with 350 images

3 Conclusions

A novel methodology for extracting the structure of handwritten filled table-forms has been presented. The approach is able to identify and to remove wrong intersections produced by overlapping data, faulty segments and smudges. One strong contribution of this paper is the analysis and interpretation of artefacts, which have not been investigated in depth until now.

The experiments carried out on a database of 350 table-form images show that the methodology is efficient for filled-in forms, with an 85% rate of correct identification. 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 intersection defects, all that with very little use of *a priori* knowledge and being appropriate to most existing types of table-forms.

Acknowledgements

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).

References

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

[Arias et al., 1996] Arias J. F., Chhabra A., and Misra V. 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.

[Couasnon 2001] Couasnon B. Dmos: A generic document recognition method, application to an automatic generator of musical scores, mathematical formulae and table structures

recognition systems. ICDAR 2001 - Sixth International Conference on Document Analysis and Recognition, pages 215–220, 2001.

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

[Hirano at al. 2001] Hirano T., Okada Y., and Yoda F. 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.

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

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

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

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

[Neves at al. 2003a] Neves, L.A.P.; Carvalho, J. M. ; Facon, J.. Bit Block Transfer and Structuring Element Decomposition for Table-form Physical Structure. SIBGRAPI 2003 - XVI Brazilian Symposium on Computer Graphics and Image Processing, 2003, São Carlos, SP.

[Neves at al. 2003b] Neves, L.A.P.; Carvalho, J. M. ; Facon, J.. Recognition of Deteriorated Table-form Documents: A New Approach. SIBGRAPI 2003 - XVI Brazilian Symposium on Computer Graphics and Image Processing, 2003, São Carlos, SP.

[Neves at al. 2006a] Neves, L.A.P.; Carvalho, J. M.; Facon, J.; Bortolozzi, F.; Ignacio, S. A. Handwritten Artefact Identification Method In Table Interpretation With Little Use of Knowledge. LNCS - DAS 2006 - Seventh International Association For Pattern Recognition on Document Analysis Systems, Nelson, Nova Zelândia, 2006.

[Neves at al. 2006b] Neves, L.A.P.; Carvalho, J. M.; Facon, J.; Bortolozzi, F.. A New Table Interpretation Methodology with Little Knowledge Base. ACM SAC 2006 - The 21th Annual ACM Symposium on Applied Computing - Document Engineering Track, Dijon, França, 2006.

[Pizano 1992] Pizano A. 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.

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

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

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

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

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

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

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