Segmentation Methodology of Table-Form Documents

Luiz Antônio Pereira Neves
IPPUC–Curitiba, PR, Brazil
neves@ippuc.curitiba.pr.gov.br
Jacques Facon
PUCPR–Curitiba, PR, Brazil
facon@ppgia.pucpr.br

Abstract

This article presents a method for the automatic extraction of the contents of passive and/or active cells in forms. The approach is based on the analysis and recognition of the types of intersection of the lines that make up such cells. Very little a priori knowledge of the form is required. The performance of this approach depends on the correction module mechanisms for detection and correction of errors generated during the intersection identification phase. The potentialities and advantages of this approach are described and illustrated with tests carried out on different form bases.

1 Introduction

Though document processing systems already provide several solutions to automate the extraction and recognition of printed and handwritten information, several problems remain partially solved. One of them is the segmentation of information contained in forms. The segmentation consists in defining the physical structure of the document, and then recognize its hierarchical structure by analyzing the relationships among the blocks to create its logical structure.

This article will focus on the physical structure of binary images of closed forms (delimited by horizontal and vertical lines). A method for the automatic extraction of passive and/or active cells will be presented based on the extraction of the lines that make up those forms, associated to a module of analysis and correction of the poorly identified intersections. The structure of this article is the following: section 2 will present an analysis of the state of the art in the identification of the physical structure of form images; section 3 will provide an overview of the methodology employed, section 4 will detail the location and identification of intersections. Section 5 will detail the detection and correction of identification errors in the physical structure. Section 6 will present the results obtained from three different form databases.

2 State of the art in form segmentation

A form is made up of horizontal and vertical lines crossing each other and delimiting small areas called cells. There are two types of cells, a passive cell where the printed information is part of the blank form, and the active cell containing the information added when the form was filled in. The identification of the physical structure of a form consists in determining characteristics such as the position, size and contents of its cells. Such information is important to define the logical structure. [7] presents an approach to recognize any given document, starting with the classification by means of a heuristic knowledge base. In the case of a form, the identification of the physical structure will forcibly go through the location of the upper left intersections of the form before locating the other intersections, using a binary tree graph structure to store the information on that physical structure. The cells are detected by locating the upper left intersections by means of masks. The time to process this identification is approximately 30 seconds. The process of determining the relationships between the cells is called local structure description tree construction phase. A description tree of the global structure is used to formalize the relationships between blocks that will serve to set up the logical structure of the document. The disadvantage of that approach lies in the fact that only the upper left intersections are used. The authors do not offer any solution for missing intersections. [7] only process unskewed forms and does not deal with skewness problems.

[1] presents a practical approach that enhances the recognition of the physical structure of the form by using nine intersections represented hierarchically by numbers (Figure 1). The approach consists in assessing the document skewness and then proceeding to the identification of the type 1, 2, 3 and 4 intersections of the form, followed by type 5, 6, 7 and 8 intersections and finally type 9 intersection (cross). According to [1] , this type of recognition improves the previous approach. Furthermore, the authors identify horizontal and vertical lines, even when interrupted. As in the case of [7], no solution is offered for when one of the 1, 2, 3 or 4 intersections is deleted, and therefore decreases the form recognition performance. The trial results obtained are 95% good recognition of synthetic images with a skewness under 4 degrees.

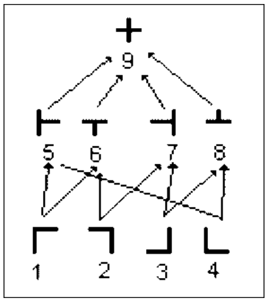


Figure 1: Hierarchical representation of the nine intersections [1]

[5] uses a different approach, also employing the 9 intersection concept. The strategy consists in first identifying the cross-type intersection (number 9). The image size is initially reduced to accelerate processing. The cell extraction phase employs the [7]´s

technique, i.e. the same masks, but instead of storing the information in the form of a binary tree, [5] uses a two-dimensional matrix. This allows a reduction in the time of form recognition.

3 Overview of the methodology for the location and correction of errors in the physical structure

The general process to segment the physical structure of a form is divided into three steps:

a) Location and identification of line intersections ;

b) Systematic detection of the identification errors;

c) Error analysis and correction.

In view of the complexity of a form, a corrected error might generate another error somewhere else in the form during the correction phase. Therefore it is necessary that steps *b*) and *c*) follow each other and be repeated so as to correct everything systematically.

4 Intersection location and identification phase

As mentioned before, a form is made up of horizontal and vertical lines crossing each other and delimiting small areas called cells. The cell extraction method employed consists in initially locating and then extracting the intersections of those lines, so as to deduct cell position and shape.

In this phase, 10 intersection models will be used, 9 of them represented hierarchically according to [1] by means of numbers (Figure 1). The tenth, represented by a ”0” corresponds to no intersection, but simply to a part of a horizontal or vertical line.

4.1 Location

The intersection location method is based on the use of the binary Mathematical Morphology ([2]) and on the use of structuring elements having the same aspect of the intersections in Figure 1. The basic idea consists in eroding the binary image of the form on the basis of each one of those structuring elements so as to preserve only the central nucleus of the corresponding intersection. Even though theoretically, 9 structuring elements are needed (one per intersection); in practice, to save memory and calculation time, according to [4], only 4 structuring elements are used corresponding to the first 4 intersections from the hierarchical standpoint in Fig 1. Called IMGi (i = 1 to 9) as the result of the erosion by the structuring element i (i = 1 to 9), the combination of the erosions corresponding to intersections 1, 2, 3 and 4 allows the deduction of the results of the erosions by structuring elements 5, 6, 7, 8 and 9:

$$\begin{aligned} IMG5 &= IMG1 \cap IMG4 & IMG6 &= IMG1 \cap IMG2 \\ IMG7 &= IMG3 \cap IMG2 & IMG8 &= IMG3 \cap IMG4 \\ IMG9 &= IMG1 \cap IMG3 \quad or & IMG9 &= IMG2 \cap IMG4 \end{aligned}$$

In the application under study here, the size of those structuring elements is 35 pixels on each side of the line crossing.

The location phase is organized as follows:

- Obtainment of images IMG1, IMG2, IMG3 and IMG4 by erosion of the binary image of the form on the basis of the structuring elements corresponding to intersections 1, 2, 3 and 4;
- Deduction of images IMG5, IMG6, IMG7, IMG8 and IMG9 corresponding to intersections 5, 6, 7, 8 and 9;
- Creation of image IMGUNION, the union of all IMGi (i = 1 to 9) images;
- Vertical and horizontal normalization of the nuclei of all the intersections in image IMGUNION.

4.2 Identification

The identification of the intersection types is founded on image IMGUNION and is organized as follows:

1: Identification of the highest level intersection still present in image IMGUNION, starting with the cross intersection (9);

2: Elimination by deduction of the highest level intersection found in all images IMGi (i = 9 to 1) and in image IMGUNION;

3: Return to phase 1 until all types of intersections are completely eliminated (9 to 1).

In order to"leave no residue" between the different images that could be misinterpreted as an intersection, the elimination phase must be the most efficient possible and employ exhaustively the binary morphological reconstruction technique [6] [2] The final result of intersection location and identification is the generation of the physical structure of the processed form.

5 Detection and correction of identification errors in the physical structure

The quality of the binary images of the form is generally very variable; two cases might happen:

- Certain line intersection might not appear, due to the poor quality of the acquisition, or to binarization problems, or simply due to the poor quality of the digitized document (torn, dirty, etc...);

- Since the forms processed here can be filled in by machines or by hand, overlapping printed andor handwritten information might create false intersections.

Those factors can eliminate real intersections and/or add false intersections that must be dealt with and corrected, so as not to ”fool” the identification module. A method for the detection and correction of error in the physical structure was developed. The principle underlying this process is the analysis of the neighborhood of each intersection in the North-South, West-East, Northeast, Northwest, Southeast and Southwest directions and comparison to reference neighborhoods congregated in the form of tables called rejection tables in error detection, and acceptance tables in error correction.

5.1 Detection of Errors in the Physical Structure

5.1.1 Basic Principle

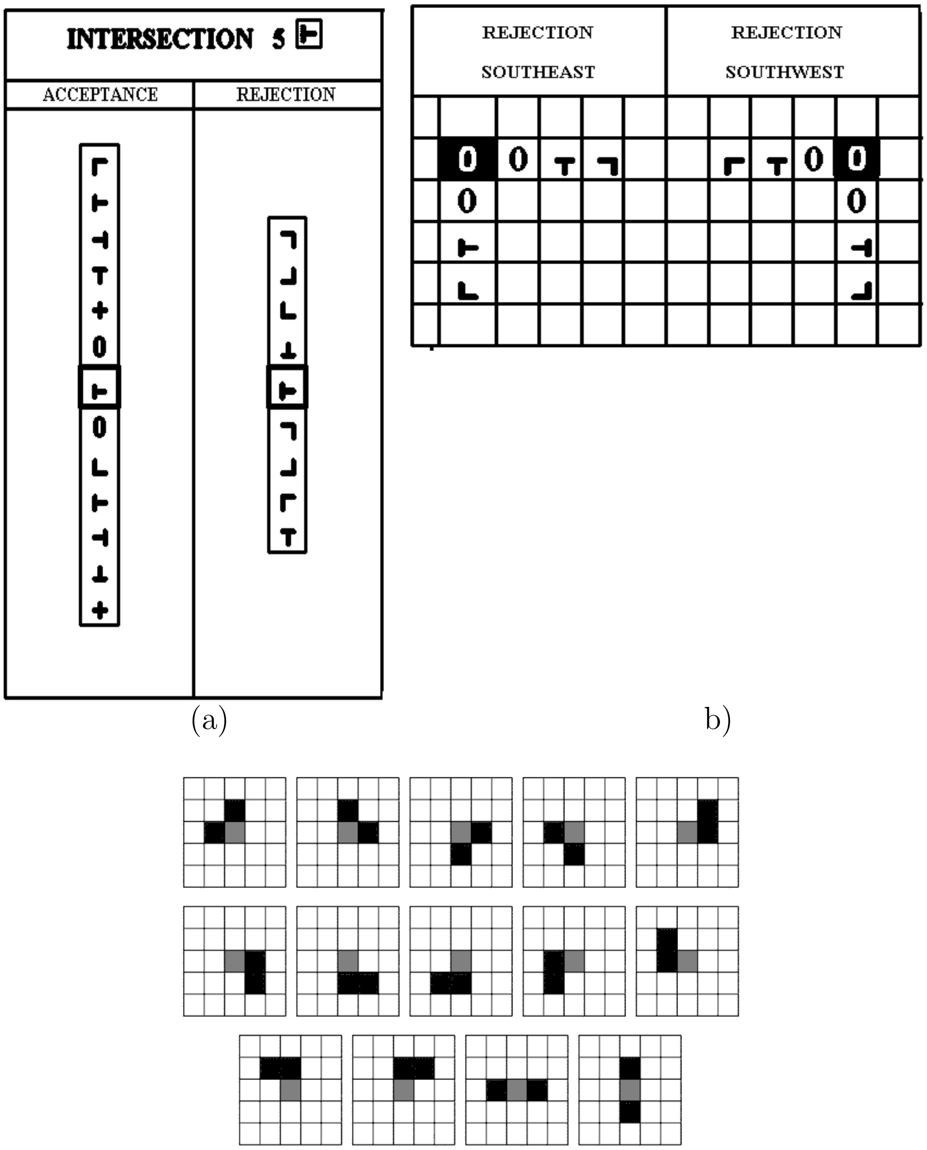
The process of detecting errors in the physical structure aims to analyze, verify and identify the possible errors originated in the previous identification phase. To allow the automation of the search and detection of errors in the physical structure, rejection tables following the North-South, West-East, Northeast, Northwest, Southeast and Southwest directions of the neighborhood of each intersection were prepared for each type of intersection 1 to 9. Every time an identification error is detected, the respective counter goes up. In the case of a wrong intersection located in the first/last line or column (meaning there is no intersection before or after) the final value of its counter is increased by 1. Figure 2-a) depicts an example of acceptance and rejection tables in the case of intersection 5 in the North-South direction.

5.1.2 The particular case of type 0 intersections

A type 0 intersection, as previously defined (a portion of a horizontal or vertical line without any indication as to its direction), is accepted by all its neighbors; a rejection table was especially defined to allow the detection of other possible errors. As in the case of intersections type 1 to 9, the neighborhood of each type 0 intersection is analyzed in the Northeast, Northwest, Southeast, Southwest, Northeast-South, Northwest-South, Southwest-East, Northwest-East and North-South-West-East directions. Figure 2-(b) depicts an example of rejection tables in the Southeast and Southwest directions.

5.2 Module of correction of errors in the physical structure of the form

Once the detection of intersection identification errors is completed, the correction is the next phase. For that purpose, two questions must be answered: What are the



(a)

b)

(c)

Figure 2: a) Acceptance and rejection tables in the case of intersection 5 in the North-South direction (b) Rejection tables in the case of type 0 intersection, (c) The 14 verification masks.

intersections to be corrected and how? In terms of the first, it is evident that all errors must be corrected. But certain identification errors do not exist in fact, they just appear because other, more important errors "radiate" to their neighbors. This means that the correction of certain identification errors must have priority over others. Those priorities are established as follows:

- In the event of detection of several errors, the correction starts with the intersection with the highest error (corresponding in fact to the highest-value counter),
- In the event of detection of errors with the same value (identical counters), the correction starts with the intersection with the highest hierarchical value (Figure 1).

5.2.1 First correction level

As to the second question, the method employed 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 again employed. The neighborhood of the error is analyzed, taking into account all pairs of two immediately adjacent neighbors in the sense of the 8- neighborhood, i.e. a total of 14 analyses is possible (Figure 2-(c)). In the case of intersections located in the first/last line or column, the number of possible analyses drops to 7.

The basic idea of this first level is to perform corrections in an iterative way: only the poorly identified top priority intersection (in the sense of the priorities of section 5.2) is corrected. After that all counters are initialized by 0 to return to the module of identification error detection in the physical structure of the form. In the case of error detection, the first correction level is again carried out, until all errors are corrected.

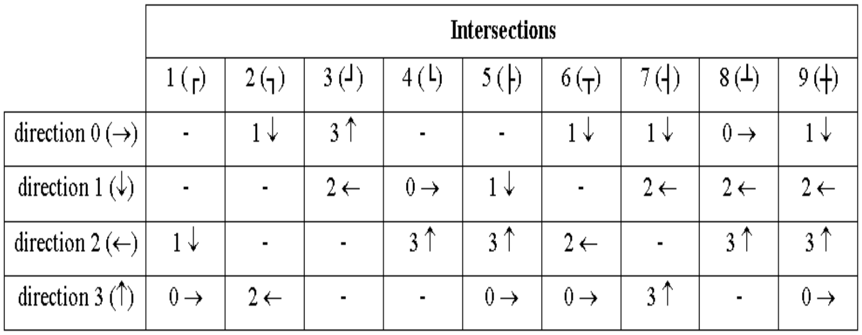
When searching for a solution, it might happen that among all possibilities several solutions are possible. The counter for each intersection goes up. At the end of the analysis, the results of the counters of all possible solutions are again studied and the intersection considered incorrect is replaced for a correct one, following three priority rules:

- If several solutions are possible, the choice of the solution is based on the counter with the highest value;
- If there are several solutions with equal values (identical counters), the solution chosen corresponds to the intersection with the highest hierarchical level (Figure 1),
- The search for that solution in a list of solutions already used is carried out as a way to avoid the following problem: it may happen that, in certain cases, a corrected error gives rise to another error and that the first level of correction will provide the same solution always. This could result in an endless loop. In

this case, the solution chosen corresponds to the intersection with the counter showing the highest value at an hierarchically inferior level.

5.2.2 Second correction level

Once the first correction level is attained, it might happen that all errors have disappeared without the physical structure of the form having a logical yet. From the local or microscopic standpoint, everything looks correct, but macroscopically the structure generated makes no sense. This means a more global type of correction of the physical structure must be performed. The method chosen consists in extracting the cells from the image under analysis, that is in fact the objective of this work. For that, [5] 's approach was used. It consists in the recursive generation of cells on the basis of rules described in Figure 3. If the impossibility to create a cell appears at an intersection, its error counter goes up. There being errors, the system will return to the first correction level upon exiting this second level.



	Intersections								
	1 (┌)	2 (┐)	3 (┘)	4 (└)	5 (├)	6 (┬)	7 (┤)	8 (┴)	9 (┼)
direction 0 (→)	-	1↓	3↑	-	-	1↓	1↓	0→	1↓
direction 1 (↓)	-	-	2←	0→	1↓	-	2←	2←	2←
direction 2 (←)	1↓	-	-	3↑	3↑	2←	-	3↑	3↑
direction 3 (↑)	0→	2←	-	-	0→	0→	3↑	-	0→

Figure 3: Second correction level rules.

5.2.3 Module of extraction of form cells

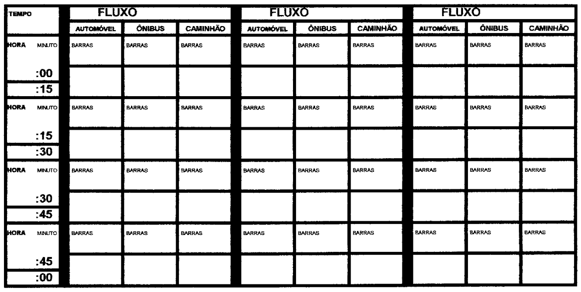
The approach structure presented allows the automatic generation of form cells in the second correction level, when all errors present in the physical structure of the form have been corrected.

6 Trial results

To demonstrate the performance of the proposed approach, three trials were carried out, two of them on a database of forms of a single type (Figure 4) and a third on a database of miscellaneous forms.

6.1 Standard blank-type forms

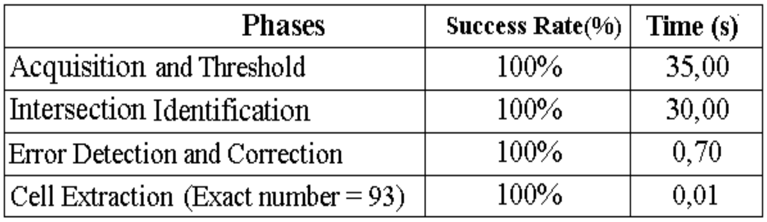
A first database made up of 50 grayscale images of blank forms of a single type (Figure 4) with 3800×3000 pixels digitized in 300 dpi was generated. This type of form contains a total of 93 cells.



TEMPO		FLUXO			FLUXO			FLUXO		
		AUTOMÓVEL	ÔNIBUS	CAMINHÃO	AUTOMÓVEL	ÔNIBUS	CAMINHÃO	AUTOMÓVEL	ÔNIBUS	CAMINHÃO
HORA	MINUTO	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS
	:00 :15									
HORA	MINUTO	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS
	:15 :30									
HORA	MINUTO	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS
	:30 :45									
HORA	MINUTO	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS
	:45 :00									

Figure 4: Example of a standard form used in the trials.

To check the soundness of the methodology, those images were artificially skewed from 4 to 8 degrees to be later corrected by [3]'s algorithm. The results obtained (Figure 5) by the proposed approach are 100% good cell segmentation (the total number of cells extracted is 93) thus showing that the approach is insensitive to slight skewness.



Phases	Success Rate(%)	Time (s)
Acquisition and Threshold	100%	35,00
Intersection Identification	100%	30,00
Error Detection and Correction	100%	0,70
Cell Extraction (Exact number = 93)	100%	0,01

Figure 5: Numerical results in the case of standard blank forms.

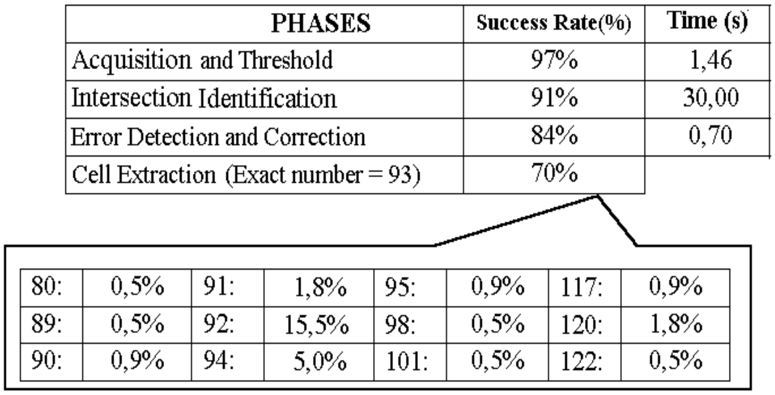
6.2 Filled-in forms

A second database made up of 263 grayscale form images was generated. The same type of forms was used in this second base as in the first, the images having the same dimension (3800 × 3000) always digitized in 300 dpi, but this time filled in with information.

To check the soundness of the methodology, those form images were again artificially skewed from 4 to 8 degrees to be later corrected by [3]'s algorithm. This type of form therefore still contains a total of 93 cells. The results obtained (Figure 6) by the approach proposed are 70% good segmentation. The errors encompass problems with binarization (3%), intersection identification (9%), identification error detection and correction (16%), and extraction of the right number of cells (30%). Concerning this last point, the numerical results obtained show that the errors arise from handwritten information added to those forms. Errors appear because of the difficulty in correcting certain intersection identification errors when the handwritten information touches or overlaps horizontal and/or vertical lines of the form.

The methodology employed tends to eliminate a correct intersection and merge two or more others in a single cell (in 24.2% of the cases). The interval of the number of cells extracted ranges from 80 to 122, peaking at 92 cells, therefore very close to the theoretical result of 93 cells. In terms of processing time, the Achilles' heel of this approach resides in the slowness of the intersection location and identification phase, while the identification error detection and correction phase is extremely fast.

Figure 8 depicts the results obtained from the segmentation of cells of a skewed form filled in with handwritten information by means of the proposed approach.



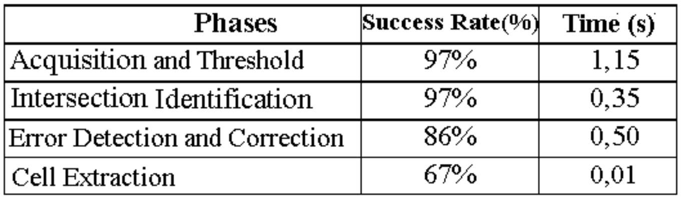
PHASES	Success Rate(%)	Time (s)
Acquisition and Threshold	97%	1,46
Intersection Identification	91%	30,00
Error Detection and Correction	84%	0,70
Cell Extraction (Exact number = 93)	70%	

80:	0,5%	91:	1,8%	95:	0,9%	117:	0,9%
89:	0,5%	92:	15,5%	98:	0,5%	120:	1,8%
90:	0,9%	94:	5,0%	101:	0,5%	122:	0,5%

Figure 6: Numerical results with filled-in standard forms.

6.3 Miscellaneous forms

A third database made up of 108 grayscale images of filled-in miscellaneous forms (varying sizes, number of cells, line thickness) digitized in 300 dpi was generated. The results obtained by the approach proposed are 67% good segmentation (Figure 7). The errors are distributed to problems with binarization (3%), intersection identification (3%), identification error detection and correction (14%), and extraction of the right number of cells (33%).



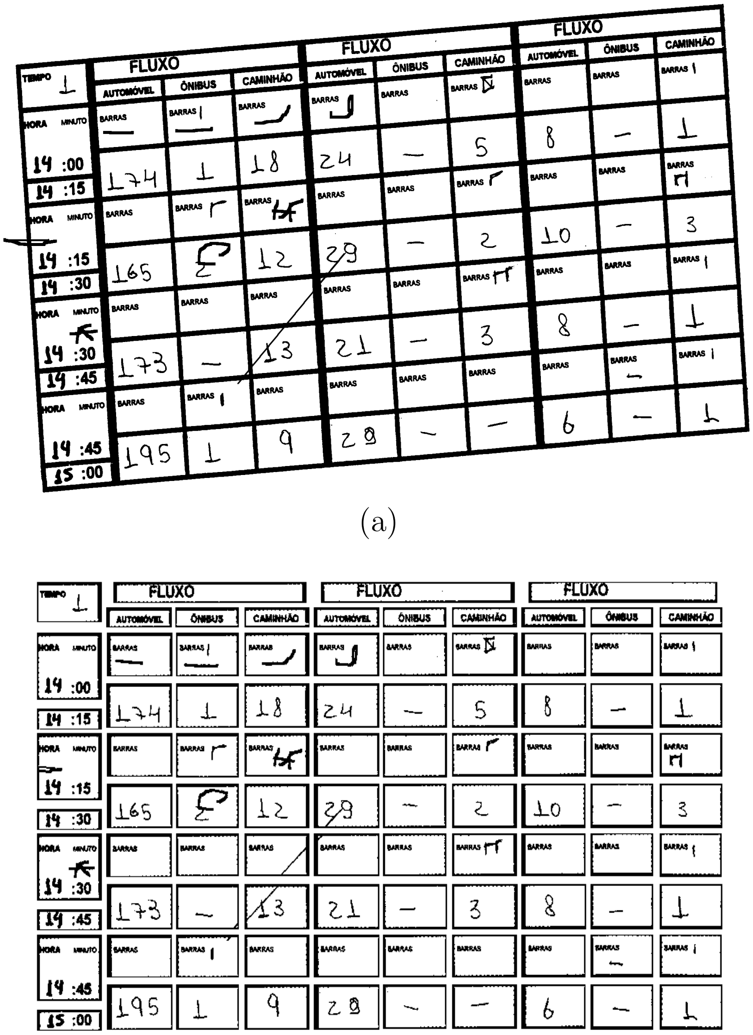
Phases	Success Rate(%)	Time (s)
Acquisition and Threshold	97%	1,15
Intersection Identification	97%	0,35
Error Detection and Correction	86%	0,50
Cell Extraction	67%	0,01

Figure 7: Numerical results for miscellaneous filled-in forms.

7 Conclusions

This article presented a methodology for the extraction of the physical structure and cells of binary images of forms. This methodology, based on few a priori knowledge of the form (the form must be closed) has three modules: the first is the intersection identification and location, the second is the detection of the errors originated in the previous module and the third is the analysis and correction of such errors. In section 6, presenting the tests of the different form databases, the trial results show that this methodology is extremely efficient for blank forms and that, for filled-in forms, the correctness percentage is 70%.

It is possible to conclude that the errors derive mainly from the handwritten information added. This will force the improvement of the analysis and correction of intersection identification errors, to prevent many cells from being merged into a single one. The optimization of the intersection location and identification processing time is also planned to better explore the promising aspects of this methodology. The study of the influence of significant skewness is also envisaged.



TEMPO 1	FLUXO			FLUXO			FLUXO		
	AUTOMÓVEL	ÔNIBUS	CAMINHÃO	AUTOMÓVEL	ÔNIBUS	CAMINHÃO	AUTOMÓVEL	ÔNIBUS	CAMINHÃO
HORA MINUTO 14 :00	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS
14 :15	174	1	18	24	–	5	8	–	1
HORA MINUTO 14 :15	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS
14 :30	165	[illegible]	12	29	–	2	10	–	3
HORA MINUTO 14 :30	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS
14 :45	173	–	13	21	–	3	8	–	1
HORA MINUTO 14 :45	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS	BARRAS
15 :00	195	1	9	28	–	–	6	–	1

(a)

(b)

Figure 8: Result of the segmentation of a filled-in form: (a) Initial Skewed image , (b) extracted cells

References

[1] Arias, Juan F. and Kasturi, Rangachar, "Efficient Extraction of Primitives From Line Drawings Composed of Horizontal and Vertical Lines", *Machine Vision and Applications*, v.10, pp.214-221, 1997.

[2] Facon Jacques, "*Mathematical Morphology: theory and examples*" - (in Portuguese) - Jacques Facon Editor, Brazil, 1996.

[3] Neves Luis. A P. and Facon Jacques, "Abordagem Morphológica de evaluação da inclinação de documentos contendo linhas", *XXV CLEI´99, Conferência Latino Americana de Informática, Asunción, Paraguai* v.1, pp.277–285, september 1999.

[4] Taylor, Suzanne Liebowitz and Fritzson Richard and Pastor Jon A., "Extraction of Data from Preprinted Forms", *Machine Vision and Applications*, v.5, pp.211-222, 1992.

[5] Thom, Richard Tran Van, "*Modélisation de Tableaux pour le traitement Automatique de Formulaires*", Master thesis, Laboratoire PSI Université de Rouen, France, 1997.

[6] Vincent Luc, "*Mathematical Morphology in Image Processing*", E. Dougerthy, ed., Marcel Drekker, New York, pp 255-288, september 1992.

[7] Watanabe, Toyohide and Luo Qin and Surgie, Noboru, "Layout Recognition of Multi-Kinds of Table-Form Documents", *IEEE Transations on Pattern Analysis and Machine Intelligence*, v.17, no. 4, pp.432-445, april 1995.