

# Line and Cell Searching in Tables or Forms

E. TUROLLA, Y. BELAÏD, A. BELAÏD

CRIN-CNRS, Bât. Loria, BP 239,  
54506 Vandoeuvre-lès-Nancy Cedex, France  
Tel: (33) 83.59.20.83 - Fax: (33) 83.41.30.79  
E-mail: {turolla, ybelaid, abelaid}@loria.fr

**Abstract.** Item searching is an important step in form analysis. The goal of this paper is to describe a robust method to locate the items whose boundaries are black lines, and an algorithm to classify those items. Our method is based on detection of lines by Hough transform and on searching of cycles in a graph that represents the cells (boundaries of the items). We will show that our approach is robust, skew independent and can be applied to several kind of lines such as continuous, dashed, doubled, etc. We classify the items in four categories : blank, black, gray and meaningful.

**Keywords :** Hough Transform, French Tax Forms, Line Searching, Item Extraction, Item Classification

## 1. Introduction

According to Doermann *et al* ([1]), there are three kinds of form processing systems, depending on the a priori knowledge they use. The most general system treats *unknown forms*, by using general knowledge about the form domain; their performance are limited by their lack of knowledge. In opposite, the systems that deal with *known forms* use a detailed model for each member of a small group of specific forms; they are often efficient in their field and are widely used ([2]); but they can be hardly applied to other kind of forms. In half way between these two classes, the systems which deal with *known classes of forms* have logical information but no information about the exact form layout; they seem to be the good compromise between reliability and flexibility ([3]).

In our case, we treat unknown forms, in such a way that results can be used for systems designed for known forms or known classes of forms. The physical and logical structure of a form is based on the neighboring relationships between its items (alignment, juxtaposition, etc.). We want to prove that is possible to approach the structure of a specific form by using general knowledge such as the existence of boundaries between items, the meaning of alignments, etc. We limited this paper to expose the item extraction, based on the search of the horizontal and vertical segments, and the item classification.

We have applied our system on French tax forms (see Figure 1) and on tables. The items are separated by continuous lines, vertical alignments of brackets or edges of black strips. The tables we have tested, come from scientific publications. Their items are separated by continuous, dashed or double lines.

In the two following sections, we describe our line and cell searching. In the third

cerfa N° 30 7190

Formulaire obligatoire annexé 53 à  
du Code général des impôts  
(ne pas reporter les totaux)

(11)

**TABLEAU D'AFFECTATION DU RÉSULTAT ET RENSEIGNEMENTS DIVERS**

FINAL

Désignation de l'entreprise : \_\_\_\_\_

**TABLEAU D'AFFECTATION DU RÉSULTAT DE L'EXERCICE PRÉCÉDENT** (Entreprises soumises à l'impôt sur les sociétés (1))

Report à nouveau figurant au bilan de l'exercice antérieur à celui pour lequel la déclaration est établie

Résultat de l'exercice précédent celui pour lequel la déclaration est établie

Prélèvements sur les réserves (à détailler)

Sous-total (à reporter dans la colonne de droite)

**TOTAL I** (28)

Affectations sur réserves

- Réserve légale (29)
- Réserve spéciale des plus-values à long terme (30)
- Autres réserves (31)

Dividendes (32)

Autres répartitions (33)

**TOTAL II** (34)

Report à nouveau

IN 8. Le total I (qui doit nécessairement être égal au total II)

Exercice N : Exercice N - 1 :

Figure 1. An extract of a French tax form page.

section, we expose our classification algorithm. We discuss the results in the last section.

## 2. Line searching

Lines are diverse and their quality can be poor (cut by bad scannerization, closed to text, skewed, etc.). Projective methods are unsuited when the lines are slant or too close to the text. Various line following methods have been developed to analyze mechanical drawings ([4]). But they need a lot of thresholds which closely depend on the kind of the document and they have difficulties with crosses. So we have chosen the *Hough transform*. It detects globally all the straight lines that exist in an image. Although this method is time consuming, it is insensitive to noise and slope, is not disturbed by crosses and can detect all kind of lines among text blocks.

In order to obtain a regular distribution probability of the voting points, we have chosen the Hough parameter space  $(\alpha, \rho)$  ([5]).  $\alpha$  is the slope of the line with respect to the x-axis, and  $\rho$  is the distance between the line and the origin of the cartesian space. The Hough space is represented by two dimensions matrix (one for horizontal lines, one for vertical lines).

To avoid finding several straight lines within one thick line and to increase the system speed, only the points that can belong to the edges of a horizontal (resp. vertical) line, vote. The black pixels that vote in the vertical Hough matrix, must have the configuration «001» (white, white, black) for the left side and "100" (black, white, white) for the right side. The transpositions of those configurations are used to fill the horizontal Hough matrix. These filters are enough to get rid of points inside black strips and keep working on slanted lines

We adapted the algorithm described in [6] to look for the significant clusters of the filled Hough space. After removing the accumulator cells whose value is very low, the system makes a recursive cutting up of each matrix; it stops when the size of the cluster is insufficient or its voting counter becomes smaller than a threshold  $T$ . This threshold is calculated to avoid to find the very short line or the alignments of letters in the image;  $T$  is the multiplication of the average of the value of the accumulator items, with a given weight ( $> 1$ ).

In the chosen parameter space, the shape of the area which corresponds to a line, is like a butterfly. So, the clusters which are adjacent, or not too far and in the good direction, are grouped together. Finally, the Hough lines are the barycenters of each remaining group of clusters

Then, the searching of black segments consists in following each Hough line on the image and merging together into a segment, the black consecutive points which are the nearest of the Hough line. The very short segments are suppressed. By this way, the continuous line and the brackets are found. To take into account the dashed lines, the system links the segments that belongs to the same Hough line and are not too far.

### 3. Cell searching

We use a graph to extract the form items. The nodes correspond to the intersection points between the horizontal and vertical lines of the image. Arcs correspond to segments joining two neighboring nodes. The cells are the minima cycles of this graph. The graph is represented by a two-dimensional table, whose rows correspond to the horizontal Hough lines and columns to the vertical ones. The horizontal (resp. vertical) Hough lines are sorted in an ascending vertical (resp. horizontal) order. Each element  $(i, j)$  of the table represents the intersection point between the  $i^{\text{th}}$  horizontal line and the  $j^{\text{th}}$  vertical line. A node  $(i, j)$  can be linked to his four neighboring nodes  $(i-1, j)$ ,  $(i+1, j)$ ,  $(i, j+1)$  and  $(i, j-1)$ .

The graph analysis operates as follows : after suppressing recursively the nodes that have only one arc, it obtains the set  $E_{ca}$  by extracting all the nodes that may be a left up vertex. It calculates  $\text{Min}_\angle(E_{ca})$ , which is the minimum element of  $E_{ca}$  for the relation  $\angle$  :

$$(i_1, j_1) \angle (i_2, j_2) \Leftrightarrow (i_1 \leq i_2) \vee ((i_1 = i_2) \wedge (j_1 < j_2))$$

The search starts from the node  $N_0 = \text{Min}_\angle(E_{ca})$ , then proceeds to the right. When a node  $N$  is reached from the direction  $d$ , it proceeds by the direction where an arc exists and whose priority is the lower (see Figure 2). A partial failure is met when the current node is prior to the cycle or above  $N_0$ ; then the search makes a back tracking. If it backtracks till  $N_0$ , then it is a complete failure and it choose a new start node. To avoid finding several times the same cycle, the start node  $N_0$  and all the nodes that are reached from bottom ( $\uparrow$ ) and are left by the right ( $\rightarrow$ ) are suppressed from  $E_{ca}$ . All the cycles of the graph are found when  $E_{ca}$  is empty. To accelerate the algorithm, the nodes that have lead to a partial failure for the current cycle are stored; if any of these nodes is reached again, then it is a complete failure.

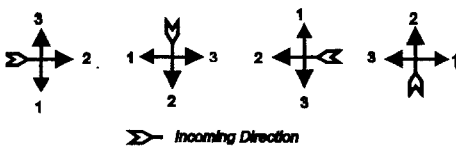


Figure 2. Priority of the direction, according to the incoming direction.

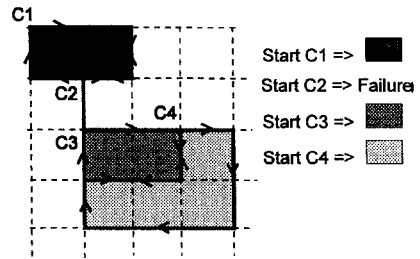


Figure 3. Examples of cycle searching.

#### 4. Item classification

The classification identifies four classes: *blank* (this item contains no information), *black* (its background is black and its foreground is white), *gray* and « *meaningful* » item (it contains information : preprinted or filled by the user).

First, the density  $D_1$  of black pixels is calculated. A high density ( $D_1 > T_{1b}$ ) means a black item, whereas a low density ( $D_1 < T_{1e}$ ) denotes a blank item. An average density implies a gray or a meaningful item. In this case, the cell is divided into little squares of five pixels length and the density  $D_2$  of « *black squares* » (they contain at least one black pixel) is calculated. A very low density ( $D_2 < T_{2m}$ ) denotes a meaningful item, with few text. In order to definitely discriminate gray between meaningful item, the regularity of black squares is studied by calculating the standard-deviation SD of black squares. If SD is low ( $SD < T_{3g}$ ), then the item is classified as gray, else it is classified as meaningful. The best thresholds we have fixed by experiments are :  $T_{1b} = 75\%$ ,  $T_{1e} = 1\%$ ,  $T_{2m} = 50\%$  and  $T_{3g} = 0.35$ .

#### 5. Experiments and discussion

We have applied this item searching method on 41 images of French tax forms and 92 tables, scanned at a resolution of 300 dpi. All these images have been analyzed with the same set of parameters. For skewed documents, it was only necessary to increase the size of the Hough matrices. The treatment of an image of a French tax form takes about 2 minutes on a SUN SPARC station IPX.

The results of line detection are described in Table 1 and Table 2. The segments whose length is smaller than 8 % of the longest line in the image and very thick lines are often ignored. It is due to the way the threshold T is calculated for Hough analysis.

Line type	Quality	Total number	Whole found lines	Partially found lines	Missed lines
Continuous	Unskewed	4553	95.02 %	0.02 %	4.96 %
	Skewed	814	96.2 %	0.12 %	3.68 %
	Merged characters	22	100 %	0 %	0 %
	Total	5389	95.21 %	0.037 %	4.75 %
Boundaries of black strips	Unskewed	544	99.3 %	0 %	0.7 %
	Skewed	92	100 %	0 %	0 %
	Total	636	99.37 %	0 %	0.63 %
Vertical alignment of brackets	Unskewed	144	81.7 %	0 %	8.3 %
	Skewed	38	89.4 %	0 %	10.6 %
	Total	182	91.2 %	0 %	8.8 %





**Table 1.** Results of line detection of French tax forms.

Line type	Quality	Total number	Whole found lines
Continuous	Unskewed	656	96.95%
	Skewed	143	100%
	Total	799	97.49%
Dashed	Unskewed	575	97.73%
Double	Unskewed	266	100%

**Table 2.** Results of line detection of tables.

Cells	Total	Well found cell	Under-segmented cells	Over-segmented cells	Forgotten cells
French Tax Forms	7377	88.59%	4.58%	2.72%	4.11%
Tables	3915	92.87%	4.90%	0.18%	2.04%

**Table 3.** Results of cell extraction.

Item type	Total number	Well classified items	Bad classified items			
			Blank	Black	Gray	Meaningful
Blank	2281	92.9 %		0 %	0 %	7.1 %
Black	371	99.7 %	0 %		0 %	0.3 %
Gray	80	41.3 %	13.7 %	3.75 %		41.2 %
Meaningful	3582	97.5 %	2.48 %	0 %	0 %	
Total	6314	95.2 %	1.58 %	0.475 %	0 %	3.10 %

**Table 4.** Results of item classification.

The results of cell extraction are described in Table 3. Under-segmentation comes from missed lines. Over-segmentation comes from black items crossed by a Hough line or from characters inside an item which are lined up with a Hough line and very close to two other lines. Forgotten cells are due to rounded vertices which are bad taken into account. All these errors can be corrected by specific treatments.

We have classified the items of the French tax forms. The results of item classification are described in Table 4. The classification of blank or meaningful items can be easily improved by removing the black pixels of their boundaries. But the classification of gray items appears to be a complex problem. Indeed, it is difficult to model the binary image of a gray area; this image is very dependent of the scanner and the quality of the original document. We have realized that the regularity of a gray area is very theoretical !.

## 6. Conclusion

We have exposed a general method for item extraction by line searching. Based on Hough transform, it can deal with medium quality documents, even skewed, and can extract several kind of lines (continuous, dashed, vertical alignment of brackets, boundaries of blacks strips). The boundaries of items are represented in a graph. The minima cycles of this graph represent the items. We have proved the robustness of our

method by applying it on French tax forms and on tables.

We will improve our line following module by dealing with pointed lines and classifying the found lines. We will complete the line searching by a module that searches the white bands.

We have begun to classify the found items in French tax forms; the results are promising but need to be improved for gray items.

The next step consists in building a specific structure of the document, that contains the physical and logical relationship we can extract from the items.

## References

- 
- [1] D. S. DOERMANN, A. ROSENFELD, "the Processing of Form Documents", ICDAR 1993, pp. 497-501
  - [2] J. YUAN, L. XU, C.Y. SUEN, "Form Items Extraction by Model matching", ICDAR 1991, pp.210-218.
  - [3] G. MADERLECHNER, "'Symbolic subtraction' of fixed formatted graphics and text from filled in forms", MVA'90, IAPR Workshop on Machine Vision Applications, Nov 28-30 1990, Tokyo
  - [4] R. KASTURI, R. RAMAN, C. CHENNUBHOTLA, L. O'GORMAN "Document image analysis, an overview of techniques for graphics recognition", SSPR 1990.
  - [5] T. RISSE, "Hough Transform for Line Recognition: Complexity of Evidence Accumulation and Cluster Detection", Computer Vision, Graphics, and Image Processing, vol. 46, pp. 327-345, 1989
  - [6] Y. MULLER-BELAÏD, R. MOHR "Planes and quadrics detection using Hough transform", 7th International Conference on Pattern Recognition; Montreal, Canada, July 30 - August 2, 1984

Figure 4. Extract of the found cells in a form.

A 2	A 3	N 7	B 8
A 6		N 10	B 11
A 9		G 14	
A 12	B 13		
	B 15		
	B 16		
	B 17	N 18	B 19
		N 21	B 22

Figure 5. An extract of the found and classified items of Figure 1; 'A' indicates a meaningful item, 'G', a gray item, 'N', a black item and 'B', a blank item.