

Form Analysis by Neural Classification of Cells

Y. Belaïd and A. Belaïd

LORIA-CNRS
Campus scientifique B.P. 239
54506 Vandœuvre-lès-Nancy CEDEX France
{ybelaid,abelaid}@loria.fr

Abstract. Our aim in this paper is to present a generic approach for linearly combining multi neural classifier for cell analysis of forms. This approach can be applied in a preprocessing step in order to highlight the different kind of information filled in the form and to determine the appropriate treatment. Features used for the classification are relative to the text orientation and to its character morphology. Eight classes are extracted among numeric, alphabetic, vertical, horizontal, capitals, etc. Classifiers are multi-layered perceptrons considering firstly global features and refining the classification at each step by looking for more precise features. The recognition rate of the classifiers for 3. 500 cells issued from 19 forms is about 91%.

1 Introduction

Form analysis becomes with the success of OCR/ICR techniques a very promising domain with different issues and applications. Several administrations and companies are today faced to a fast treatment of their forms in different domains such as order lecture, revenue form capture or multiple choice question paper analysis. Systems designed this last decade for form analysis are numerous and themes are varied. However, all of these systems are oriented towards a full form recognition without a real separation between the different phases. This makes difficult the reuse of systems and leads sometimes, for a new application, to the complete rewriting of the techniques. So, we have considered that for some classes of forms such as the tax forms, cells are the base of the form analysis and cell classification can constitute a generic part of a form analysis system.

Considering cell detection and extraction, the literature mentions mainly two approaches. The most common one deals with known forms and uses a detailed model for each class of forms [2,3,4,8]. Although the systems are efficient on specific forms, they can be hardly applied to other kind of forms. In opposite, the systems, in the other approach, ignore any a priori knowledge on the form and base the analysis mainly on cell analysis [5,6]. Although they are more general than the first ones and can be applied on a wide range of forms, their performance is limited because of their lack of knowledge.

Our aim in this paper is to propose an intermediate solution for unknown form analysis based on cell classification. Cells are first extracted from the form

and classified according to different criteria based more on the content aspect than on its semantic interpretation.

The cell extraction and classification is a very important step in a form analysis process for several reasons:

- The information contained in the form is mainly located in the cells.
- The cell extraction allows to locate the information and to situate it according to the rows and columns. This leads to find at the same time the layout and the logical structure of the form (correspondence between rows and columns, number of dimensions, etc.).
- The exam of the content of each cell can help to the content classification, first by separating cells containing the information from empty ones, second, by analyzing the type of content (text, digits, etc.) in order to apply on it the adequate treatment.
- At last, the cell extraction and content classification can help to the pre-classification of forms with a modest investigation.

The outline of this paper is as follows. After a brief description of the approach used for the cell location in section 2, we present in section 3 the different classes retained for the classification and give in section 4 the classification schema. Details dealing with the main classification steps will be then exposed in this section. So, we show the different features used for the different classes and the hierarchy of neural architectures. At last, before concluding, some experiments and results will be discussed in section 5.

2 Cell Extraction

As mentioned in [7] cell location and extraction is operated in three steps.

In the first step, lines are detected in the image by applying Hough Transform. This technique was used for its robustness and reliability. It transforms the following line problem in a counting point problem [1]. In order to avoid a multitude of line candidates, voting points are limited to only those belonging either to the contours or to the black or gray areas. A recursive cut of the polar parameter space of lines and a fusion of close cells allow to fast locate the accumulation areas.

In the second step, segments associated to the lines, are extracted from the image. The line following is operated by advantaging the closest black pixels of the Hough lines. The lines detected can be simple, double, continuous or discontinuous, contours of black and gray areas, or vertical alignments of parentheses.

The cells are located at the third step. They are represented by a graph which arcs are the horizontal and vertical segments and which nodes are the intersection points between horizontal and vertical lines. Cells are given by the research for minimum circuits of the graph.

This first part of the system has been tested with success on French tax forms as well as on tables. The line extraction takes about 30" per image.

3 Cell Classes

A detailed study of French tax forms leads us to define eight classes for cells described below:

- DIGI: it regroups the set of cells containing only digits. These digits generally correspond to amounts and can be preceded by the sign '+' or '-'.
- GRAY: it corresponds to gray areas which cannot be filled by any kind of data.
- HLET (for horizontal letters): all cells which text is horizontally aligned and which are constituted by alphanumeric chains containing lower-case letters and probably higher-case letters are affected to this class. They correspond essentially to form wordings.
- VLET (for vertical letters): it reassembles the same kind of cells within the class HLET but with text vertically aligned.
- HHCL (for horizontal higher-case letters): cells containing higher-case letters horizontally aligned and some digits or symbols such as parentheses are attributed to this class. These cells often correspond to wordings representing amounts.
- VHCL (for vertical higher-case letters): it corresponds to the class HHCL but considering only cells which text is vertically aligned.
- BLAC: it regroups cells in inverse video (with black foreground). These cells play a particular role in our forms.
- EMPT (for empty classes): this class regroups all the cells containing any data.

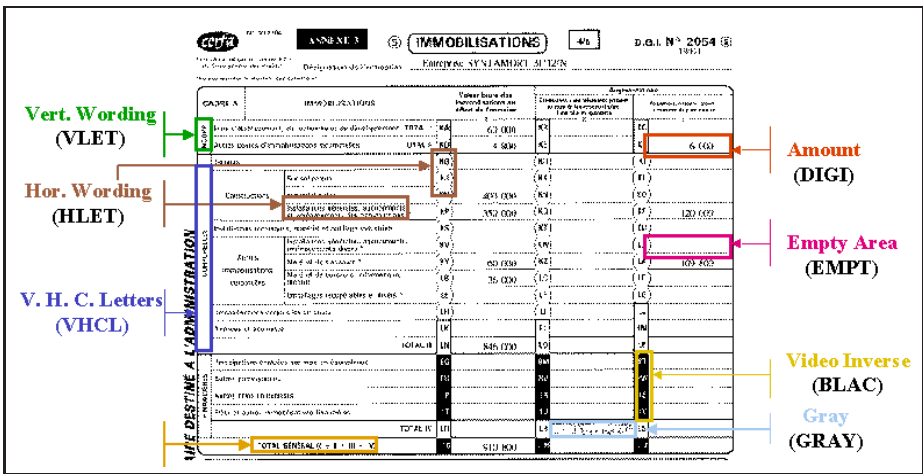


Fig. 1. Example of Classes.

The choice of these classes depends of course on our application but can be adapted on other kind of forms. Figure 1 gives an example of a form highlighting these different classes.

4 Analysis Phases

The analysis schema consists of three main phases (cf. fig. 2):

- In the first phase, cells are extracted by line searching, segment reconstruction and cycle delimitation within the graph of line crossings
- In the second phase, some features are extracted from the cells either globally on the hole data or more precisely on the different connected components within each cell
- In the third phase, these features are analyzed by a list of classifiers which decide for one belonging class among the eight classes defined below

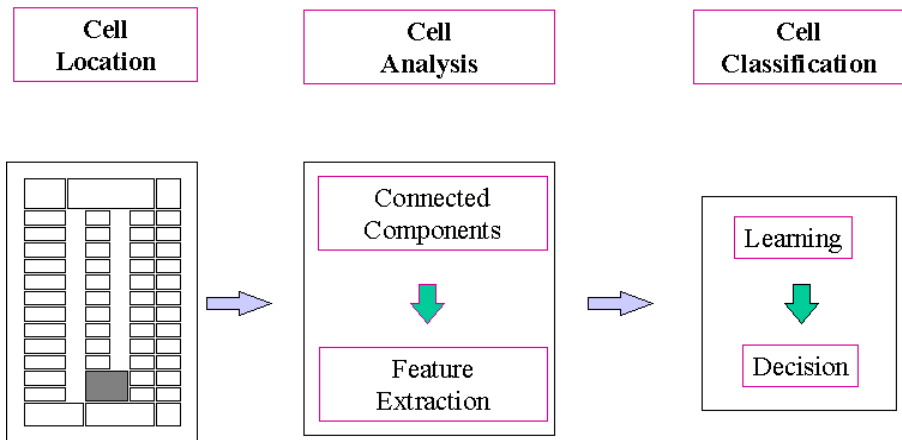


Fig. 2. Analysis Phases.

4.1 Cell Extraction

For line extraction, we used Hough transform on polar coordinates. Because of its slowness, the original method was improved by a filtering of voting points and a clever clustering of the accumulator cells in the Hough space (depending on line direction).

For filtering, only points belonging to contours vote. Thanks to this filtering, the system is not disturbed by the black strips and can detect accurately their contour. For thick lines, the two edges are merged.

The clustering in the polar coordinates is made up by a recursive cutting up algorithm.

Figure 3.a shows the hough straight line, the voting pixel and the current segment followed. Figure 3.b shows the line extraction by Hough. We can notice that lines made up by brackets and black strips are also found. Finally, Figure 3.c synthesises the different extraction rates, for well delimitation, under- and over-segmentation, and for omission.

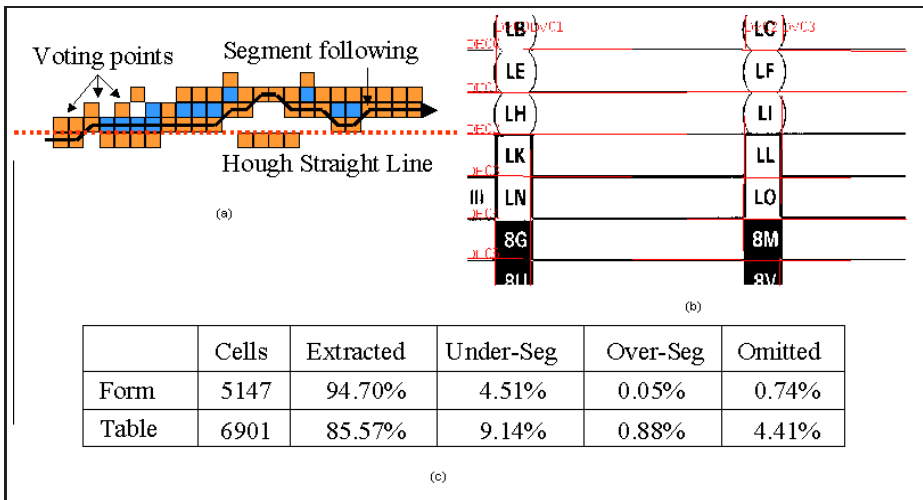


Fig. 3. Cell Extraction.

The method tolerates a reasonable orientation. Each item is depicted by a box in a specific colour. Even though the method gives a very indifferent quality, it provides a good base for a global classification of the cells. Joined characters and overlapping characters with horizontal lines are usually disturbing the classification(cf. fig. 4).

4.2 Cell Classification

The cell classification schema can be divided into three steps as shown if fig. 5.

The classification is performed progressively in order to discard rapidly the obvious cases and to refine the others.

The role of the first step is to give a rapid decision on the simple cases such as GRAY, EMPT and BLAC and to separate the remaining classes into two main

(Ne pas reporter le montant des centimes)*		Exercice N 1	Exercice N-1 2
PRODUITS EXCEPTIONNELS	Produits exceptionnels sur opérations de gestion	HA	
	Produits exceptionnels sur opérations en capital *	HB	219 800
	Reprises sur provisions et transferts de charges	HC	
	Total des produits exceptionnels (7) (VII)	HD	219 800
CHARGES EXCEPTIONNELS	Charges exceptionnelles sur opérations de gestion	HE	5 416
	Charges exceptionnelles sur opérations en capital *	HF	125 987
	Dotations exceptionnelles aux amortissements et provisions	HG	5 200
	Total des charges exceptionnelles (7) (VIII)	HI	136 603
3 - RESULTAT EXCEPTIONNEL (VII - VIII)	II	83 196	
Participation des salariés aux résultats de l'entreprise (IX)	IJ		
Impôts sur les bénéfices *	IK		
TOTAL DES PRODUITS (I + III + V + VII)	HL	3 336 905	
TOTAL DES CHARGES (II + IV + VI + VIII + IX + X)	HM	3 273 685	
4 - BENEFICE OU PERTE (total des produits - total des charges)	HN	63 220	
	HO		

Fig. 4. Cell Extraction.

categories depending on the alignment: horizontal or vertical. Because of the ill quality of the image, this classification is operated globally on the cell images by using some numeric features.

The objective of the second step is to refine the classification within both categories. The classification is operated directly on the image of the CCs of the cell. Two specific neuronal nets are used at this level.

The third stage is used for some disturbing cases corresponding to cut characters such as the zero which seriously handicaps the straightforward classification of the digits. A specific classifier is used for a possible merging of cut digits.

4.3 Classification from CC Parameters

12 simple numerical parameters have been experimentally determined for the classification in the first step.

1. Number of connected components: this is related to the number of CCs after a merging step introduced because of the presence of numerous cut characters (cf. fig. 6). The merging is made within a cell, line by line. Two CCs are merged if they respect the following constraints:

- they belong to the same text line,
- they are consecutive in the line,
- they are superposed (cf. fig. 7.a) or overlapped with an important intersection area (cf. fig. 7.b) or overlapped with a small intersection and

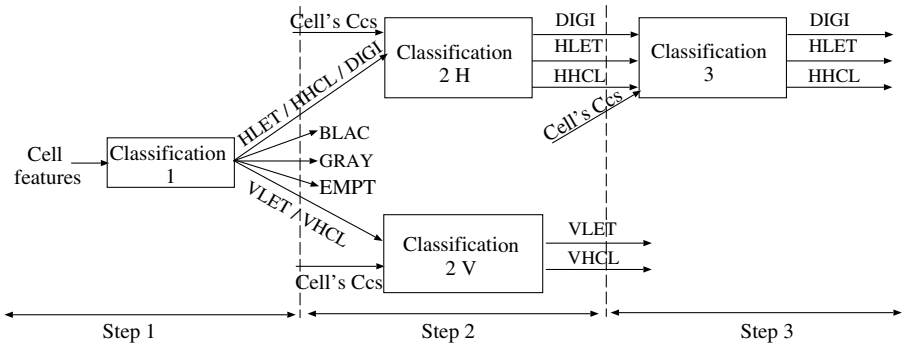


Fig. 5. Cell Classification.



Fig. 6. Connected Components Before Merging.

where one of the CC is very small compared to the average size of the Cell's CCs studied.

2. Text alignment: we have observed in the forms studied that the cells which are more wide than tall, contain text horizontally aligned. In opposite, when the cells are more tall then wide, the text can be aligned horizontally or vertically. An analysis of the text is then necessary in this case. Three cases are considered:

- the number of horizontal CCs , i.e. those which the height is greater than the width,
- the homogeneity of the height of the text lines,
- the height of the greater CCs of each line.



Fig. 7. Different configurations of Connected Components.

If the number of horizontal CCs is important, text lines are homogeneous and if there are CCs in each line which height is similar to the line height, the text is considered as horizontally aligned. Otherwise, it is considered as vertically aligned.

3. Number of text lines: the text lines are detected by analyzing the histogram obtained by horizontal or vertical projection of the image.

This analysis is performed in three steps:

- In the first step, the black areas of the histogram are delimited. When the text is of a good quality and lines are not overlapped, each black area corresponds to a text line. To avoid to take into account the noise, only black areas are considered with a size experimentally fixed to 3 pixels for the height and 5 pixels for the width. It is important to combine the height and the width in order to avoid to consider as noise, lines containing only one character (cf. fig 8).



Fig. 8. Example of a Line containing only one Character.

- In the second step, picks found are merged where they are close (separated with less than 3 pixels). In fact, in some cases, a line can be represented by two picks. This is the case for the line of 9 because the letter 'g' of the word 'outillage' is composed of two CCs.

**Installations techniques, matériel et outillage
industriels**

Fig. 9. Bad Segmented Characters.

- Lines previously found are examined in order to separate couples of consecutive lines connected by the down-strokes of the ones or the stems of the others (cf. fig 10). For this, each black area is analyzed so that picks separated by a valley which height is less than a threshold 's' which value

has been fixed at 28% of the greatest height (cf. fig 11) are extracted. These picks are compared if they are neighbours. We begin by analyzing the picks having the deepest valleys. If the distance between a pair of picks is less than the sum of the width of the two picks, this probably indicates the presence of two close text lines. If the widths of these two potential lines are comparable, then two separated lines are considered. In the other cases, we merge the two picks which forms a single line. When all the picks have been treated by pairs, we compare the size of the lines obtained during this step. If their size is homogeneous, the lines are selected, else only lines extracted in the second step are preserved.



Fig. 10. Very Close Text Lines.

4. Number of classes of CC heights. It is obtained from the analysis of height histogram of the CCs.

A pick is indicated by a high value of the histogram. The class searching is determined as follows:

Begin

Create a class with the biggest pick.

Examine the others picks in a decreasing order.

Let P_c be the current pick, search P_g a bigger pick than P_c and the closest of P_c .

If P_c is enough close to P_g

then P_c belongs to the same class than P_g

else create a new class containing P_c .

End

5. Number of CC width classes.
 6. Number of width classes of the spaces between CCs.
 7. Average number of black segments per line in a CC.
 8. Average density of black pixels per CC.
 9. Density of black pixels in a cell.
 10. Average height of CCs.
 11. Number of CCs deleted: this value is determined during the CC extraction step; it corresponds to the number of CCs assimilated to the noise.
 12. Ratio between the number of CCs deleted and the total number of CCs.
- The choice of these 12 parameters results from a series of observations and tests realized on a database built up for this problem and from which we have

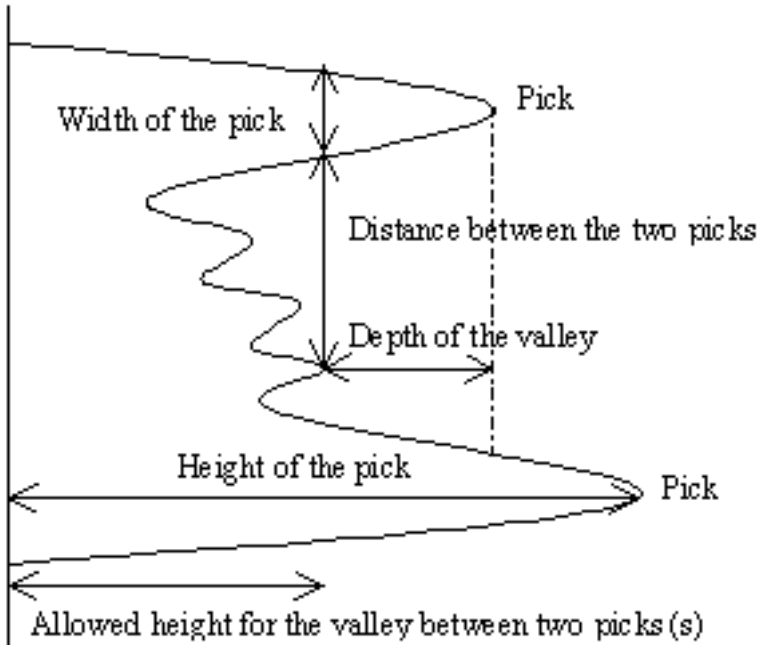


Fig. 11. Example of two Consecutive Picks Used for Line Detection.

verified the contribution of each one of these parameters in the classification process.

The classifier uses a mono-layer perceptron with 12 neurons on the entry layer (for the 12 parameters) and 5 neurons on the exit layer (for the 5 classes retained).

4.4 Classification from Cell CCs

This phase including the steps 2 and 3 of the classification process, tends to classify cells containing text into three classes: DIGI, HLET and HHCL for cells with horizontal alignment or into two classes: VLET and VHCL for cells with vertical alignment. In the tax forms, there was not amounts vertically aligned. This explains the class difference in relation with the alignment.

In the second step, the entry data of the classifier is the size normalized image of a CC. The classifier retained is a perceptron with a hidden layer. It contains 64 neurons on the entry layer (the 64 pixels of the normalized image of a CC), 12 neurons on the hidden layer and 2 or 3 neurons on the output layer according to the alignment of the studied cell. A value is associated to each output. For every output, we compute the product of the values of each CC of the considered cell. The output having the higher product is attributed to the cell.

The results obtained at the end of this step are satisfactory except for the DIGI class. Errors come essentially from the '0' often bad segmented and cut in two parts. So, cells containing a majority of '0' are bad classified. The solution for this problem is given in the third step.

In this step, CCs presented to the classifier are normalized and grouped by pairs. This treatment is realized for cells which number of height class of CCs is 1 or 2. The classifier used is a perceptron having 128 neurons on the entry layer (the 128 pixels of the image normalized and merged into 2 CCs), 18 neurons on the hidden layer and 2 neurons on the output layer: one for the '0' class and one for the other characters.

These results are compared with those of the step 2 and a decision is taken for the belonging or not of the cell to the classes DIGI, HLET or HHCL.

A score is determined for each one of the three classes and we retain the one which score is the highest. Let notice

ScDIGI: the score of the class DIGI for the cell considered,

ScHLET: the one of HLET,

ScHHCL: the one of HHCL,

SjC2Hcci: the output value j of the classifier 2H for the CC i,

SjC3cci: the output value j of the classifier 3 for the CC i merged with the CC i+1.

The computations are made as follows:

Begin

ScDIGI, ScHLET, ScHHCL = 1

i = 1

For every CC i of the current cell do

 If S1C3cc i ; S2C3cc i then

 ScDIGI = ScDIGI * S1C2Hcci

 ScHLET = ScHLET * S2C2Hcci

 ScHHCL = ScHHCL * S3C2Hcci

 i = i + 1

 Else

 ScDIGI = ScDIGI * S1C3cci

 ScHLET = ScHLET * S2C2Hcci * S2C2Hcci + 1

 ScHHCL = ScHHCL * S2C2Hcci * S2C2Hcci + 1

 i = i + 2

 Endif

Endfor

End

5 Results and Discussion

The classification process was tested on 19 French tax forms belonging to the General Direction of French Revenue. The classification time for one form is about 1'45" on a SUN Ultra Spark station, model 140 MHz.

The results obtained are detailed in the table 1 and the classification rates are presented in table 2. A classification example of a form is given in 12. The color attributed to each cell indicates the class found (cf. fig 3). A classification error is materialized by the presence of a little square on the bottom left. It has the colour of the wanted class.

	DIGI	HLET	HHCL	VLET	VHCL	GRAY	BLAC	EMPT	Total Number of Cells
DIGI	638	9	6	0	0	0	0	0	653
HLET	5	587	233	0	0	4	0	0	829
HHCL	4	22	308	4	2	2	0	0	342
VLET	0	0	0	23	0	0	0	0	23
VHCL	0	0	0	0	72	0	0	0	72
GRAY	0	0	1	0	0	23	1	0	25
BLAC	0	0	0	0	0	0	72	0	72
EMPT	0	0	2	0	0	7	0	1479	1488

Table 1. Form Classification Results.

DIGI	HLET	HHCL	VLET	VHCL	GRAY	BLAC	EMPT	TOTAL
97.70%	70.81 %	90.06 %	100.00 %	100.00 %	92.00 %	100.00%	99.40%	91.38%

Table 2. Classification Rates.

We can remark that the scores are very good for the classes DIGI, VLET, VHCL, GRAY, BLAC and EMPT, but are less good for the classes HLET and HHCL. There are several reasons explaining the confusions:

- The bad quality of images can produce an over-segmentation (cf. fig. 11) or a under-segmentation (cf. fig. 12).
- Some characters have the same morphology in lower-case and higher-case and cannot be differentiated after normalization. It is the case of characters as 'c', 'o', 's', 'u', 'v', 'x' and 'z'.

Several solutions can be considered in order to resolve these problems:

- creation of a reject class for cells for which it is difficult to make a choice between the classes HLET and HHCL,
- fusion of the classes HLET and HHCL: this fusion gives a general classification rate equal to 98,32%,
- the consideration of the CC height before normalization.

DIGI	HLET	HHCL	VLET	VHCL	GRAY	BLAC	EMPTY
red	brown	orange	clear green	dark green	cyan	yellow	mauve

Table 3. Colors of the eigh Cell Classes.

6 Conclusion

This paper outlines a feasibility study for the classification of form cells into eight classes depending on the presence of information or not, on the text alignment: horizontal or vertical and the character modes higher-case or lower-case. Few systems have been developed in this sense. Most of the classification methods developed try to differentiate between text and non text areas. We used a perceptron for the classification. It is mono-layer for the first step which realizes a pre-classification by using numerical parameters. It contains one hidden layer for the steps 2 and 3 which analyze the text areas from normalized images of CCs. The results obtained are acceptable. Improvements and adaptations remain are possible. Acknowledgements: The authors wish to thank N. Pican for providing us with his implementation of the perceptron algorithm.

References

1. RISSE T., Hough Transform for Line Recognition : Complexity of Evidence Accumulation and Cluster Detection, *Computer Vision, Graphics, and Image Processing*, Vol. 46, 1989, pp. 327-345. 59
2. CASEY, FERGUSON D., MOHIUDDIN K. and WALACH E., Intelligent Form s Processing System, *Machine Vision and Applications*, Vol. 5, no 3, 1992, pp. 14 4-155. 58
3. ISHITANI Y., Model Matching Based on Association Graph for Form Image Understanding, in *Proceedings of ICDAR'95 : 3rd International Conference on Document Analysis and Recognition*, Montréal, Canada, 1995, pp. 287-292. 58
4. YUAN J., TANG Y. Y. and SUEN C. Y., Four Directional Adjacency Graphs (FDAG) and Their Application in Locating Fields in Forms, in *Proceedings of ICDAR'95 : 3rd International Conference on Document Analysis and Recognition*, Montréal, Canada, 1995, pp. 752-755. 58
5. HIRAYAMA Y., Analysing Form Images by Using Line-Shared-Adjacent Cell Relations, in *Proceedings of ICPR'96 : 13th International Conference on Pattern Recognition*, 1996, pp. 768-772. 58
6. SHIMOTSUJI S. and ASANO M., Form Identification based on Cell Structure, in *Proceedings of ICPR'96 : 13th International Conference on Pattern Recognition*, 1996, pp. 793-797. 58
7. TUROLLA E. BELAÏD Y. and BELAÏD A., Form item extraction based on line searching, in *Graphics Recognition : Methods and Applications*, Lecture Notes in Computer Science, Vol. 1072, 1996, pp. 69-79. 59
8. ARAI H. et ODAKA K., From Processing Based on Background Region Analysis, in *Proceedings of ICDAR'97 : 4th International Conference on Document Analysis and Recognition*, Ulm, Allemagne, Vol. 1, 1997, pp. 164-169. 58

cerfa N° 06-0184
Formulaire fiscal 2004
à compléter par les contribuables

② **BILAN — PASSIF avant répartition**

D.G.F. N° 2051
2004

1^{er} EXEMPLAIRE DESTINÉ À L'ADMINISTRATION

Désignation de l'entreprise		N° par secteur le moyen des comptes *		Exercice N		Exercice N - 1	
				1		2	
CAPITAUX PROPRES	Capital social ou individuel (1) ¹ (dont verse	J	937 200	3 927 500	3 927 500		
	Primes d'émission de titres d'apport			166 992	166 992		
	Avants de réévaluation (2) ² (dont écart d'équivalences	IX		54 459	54 459		
	Réserve légale (3)						
	Réserves statutaires ou conventionnelles						
	Réserves réglementées (4) (a)			67 150	67 150		
	Autres réserves			17 740	17 740		
	Rapport à nouveau					-599 479	
	RÉSULTAT DE L'EXERCICE (bénéfice ou perte)			772 248	636 457		
	Subventions d'investissement					201 560	
Provisions réglementées *							
TOTAL (II)			5 016 091	4 732 911			
Autres titres immobiliers	Produit des émissions de titres participatifs						
	Autres immobilisations						
	TOTAL (III)						
Provisions pour risques ou charges	Provisions pour risques			18 000			
	Provisions pour charges			4062	9 376		
	TOTAL (III)			22 062	9 376		
DETTES (5)	Emprunts obligataires convertibles						
	Autres emprunts obligataires						
	Emprunts et dettes auprès des établissements de crédit (a)			13 752 083	16 507 934		
	Emprunts et dettes financières divers (7)			710 146	806 780		
	Avances et acomptes reçus sur commandes en cours						
	Dettes fournisseurs et comptes rattachés			3 015 202	3 623 572		
	Dettes fiscales et sociales			2 854 984	2 627 870		
	Dettes sur immobilisations et comptes rattachés			349 870	2 301 750		
	Autres dettes			16 854	18 924		
	Produits constatés d'avance (8)			20 699 141	25 886 832		
TOTAL (IV)							
Autres titres	Avance de conversion passif *						
	TOTAL GÉNÉRAL (II + VI)			25 737 294	30 629 120		
Total du bilan de l'exercice N en francs et centimes *				25 737 294		30 629 120	
RÉSERVES	(11) Avant de réévaluation incorporé au capital						
	(21) Dont	Réserve spéciale de réévaluation (1994)					
		Écart de réévaluation libre					
		Réserve de réévaluation (1974)					
	(31) Dont réserve réglementée des plus-values à long terme *			67 150	67 150		
	(41) Dont réserve relative à l'achat d'immobilisations corporelles d'actifs étrangers *						
	(51) Dettes et produits constatés d'avance à moins d'un an			11 393 937	15 315 104		
(61) Dont concours bancaires courants et soldes créditeurs de banques et CCP			1 115 800	679 637			
(71) Dont emprunts participatifs			710 146	806 780			

* Des explications détaillées de cette rubrique sont données dans le mode n° 2052.

Fig. 12. An Example of Form Classification.