Form Analysis by Neural Classification of Cells

Yolande Belaïd, Jean Luc Panchèvre, Abdel Belaïd

HAL Id: inria-00098513

https://hal.inria.fr/inria-00098513

Submitted on 27 Feb 2007

Form Analysis by Neural Classification of Cells

Y. Belaïd[1], J. L. Panchvre[2] and A. Belaïd[3]

[1]LORIA-University of Nancy II
B.P. 239, 54506 Vandoeuvre-Lès-Nancy France
E-mail: ybelaid@loria.fr

[2]EDF - DEPT - Services et Ingénierie Nord-Ouest
59041 LILLE CEDEX FRANCE

[3]LORIA-CNRS
B.P. 239, 54506 Vandoeuvre-Lès-Nancy France

Our aim in this paper is to present a methodology for linearly combining multi neural classifier for cell analysis of forms. 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 form interpretation based on 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 [2,8,4,1]. Although the systems are efficient on specific forms, they can be hardly applied to others kinds 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 [?,3].

Although they are more general than the formers 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 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. 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. 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 performed 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 extracted by the search of minimum cycles in 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 '+' or '-'.

- GRAY: gray areas which cannot be filled by any kind of data,

- HLET: text horizontally aligned, constituted by alphanumeric chains containing lower-case letters and probably higher-case letters, corresponding to form wordings,

- VLET: similar to HLET cells but with text vertically aligned,

- HHCL: higher-case letters horizontally aligned or digits or symbols such as parentheses, often corresponding to wordings representing amounts.

- VHCL: similar to HHCL but with text vertically aligned,

- BLAC: inverse video

- EMPT: empty cells.

The choice of these classes depends of course on our application but can be adapted on other kind of forms.

4   System Overview

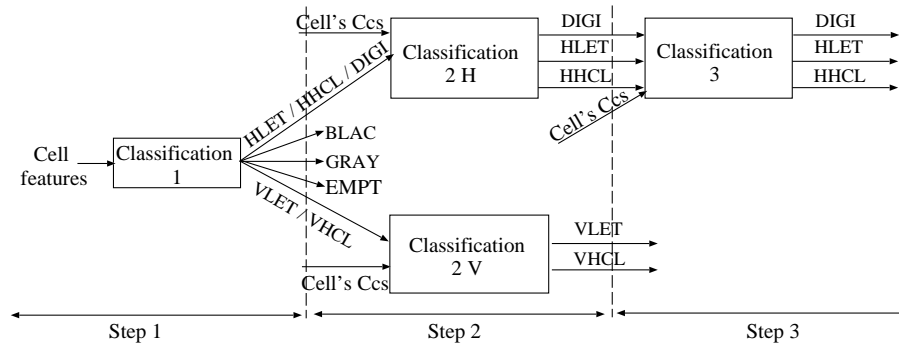The cell classification schema can be divided into three steps as shown if fig.4.

Figure 1: *Cell Classification Steps.*

In the first step, the classification is operated on numeric parameters extracted from the cells. These parameters which will be detailed further do not allow to distinguish neither between the classes: DIGI, HLET and HHCL for the text horizontally aligned, nor between the classes VLET and VHCL for the text vertically aligned. This is so for the bad quality of images and for the variability of fonts and character sizes.

3

The objective of the second step is to separate the classes DIGI, HLET and HHCL and the classes VLET and VHCL.

In theory, a CC must correspond to one character but this is false in practice because some characters are attached and some others are cut. This problem is very frequent for the digit '0' which is always cut into two parts. This is a source of error for the digit classification and explains the use of the third step which objective is to detect the '0' cut and so to improve the classification into DIGI, HLET and HHCL.

We used neural networks in the third step of the system, a one-layer perceptron in the first step and a multi-layer perceptron in the second and third steps.

4.1 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. The merging is made within a cell, line by line. Two CCs are merged if they belong to the same text line, are consecutive in the line, and are superposed or overlapped with an important intersection area or overlapped with a small intersection and where one of the CCs 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 taller than wider, the text can be horizontally or vertically aligned. 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.

   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.

4

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.

- 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.

- 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. 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 28the 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.

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 Pc be the current pick, search Pg a bigger pick than Pc and the closest of Pc.

If Pc is enough close to Pg

> then Pc belongs to the same class than Pg
>
> else create a new class containing Pc.

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 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 output layer (for the 5 classes retained).

4.2 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

6

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 to 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 the class havin the highest score is retained. 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 +
    Else
      ScDIGI = ScDIGI * S1C3cci
      ScHLET = ScHLET * S2C2Hcci * S2C2Hcci + 1
      ScHHCL = ScHHCL * S2C2Hcci * S2C2Hcci + 1; i= i+2
    Endif
Endfor
End

7

5 Results and Discussion

The classification process was performed 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 are detailed in the table 5 and the classification rates are presented in table 5. A classification example of a form is given in 5. A classification error is materialized by the presence of a little square on the bottom left. It has the color 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	EMPTY	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 note that the scores are great for DIGI, VLET, VHCL, GRAY, BLAC and EMPT classes, but are bad for HLET and HHCL classes. There are several reasons explaining the confusions:

- The bad quality of images can produce an over-segmentation or a under-segmentation.

- Some characters have the same morphology in lower-case and higher-case and cannot be differentiated after normalization. It is the case of '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.

8

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. Arai H. et Odaka K., From Processing Based on Background Region Analysis, in Proceedings of ICDAR'97, Ulm, Germany, Vol. 1, 1997, pp. 164-169.
2. Casey, Ferguson D., Mohiuddin K. and Walach E., Intelligent Form s Processing System, Machine Vision and Applications, Vol. 5, n 3, 1992, pp. 14 4-155.
3. Hirayama Y., Analysing Form Images by Using Line-Shared-Adjacent Cell Relations, in Proceedings of ICPR'96, pp. 768-772, Vienna, Austria, 1996.
4. Ishitani Y., Model Matching Based on Association Graph for Form Image Understanding, in Proceedings of ICDAR'95, Montral, Canada, 1995, pp. 287-292.
5. Risse T., Hough Transform for Line Recognition : Complexity of E vidence Accumulation and Cluster Detection, Computer Vision, Graphics, and Image Processing, Vol. 46, 1989, pp. 327-345.
6. Shimotsuji S. and Asano M., Form Identification based on Cell St ructure, in Proceedings of ICPR'96 : 13th International Conference on Pattern Re cognition, 1996, pp. 793-797.
7. Turolla E. Belaïd Y. et 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.
8. Yuan J., Tang Y. Y. and Suen C. Y., Four Directional Adjacency G raphs (FDAG) and Their Application in Locating Fields in Forms, ICDAR'95, Montral, Canada, 1995, pp. 752-755.

Figure 2: *An Example of Form Classification: DIGI(red), HLET(brown), HHCL(orange), VLET(clear green), VHCL(dark green), GRAY(cyan), BLAC(yellow), EMPTY(mauve).*