Adaptive Technology for Mail-Order Form Segmentation

A.Belaïd[1], Y. Belaïd[2], Late N. Valverde[3] and S. Kébairi[3]

[1]*LORIA-CNRS, Campus scientifique, B.P. 239, 54506 Vandoeuvre-Lès-Nancy France*
[2]*LORIA-University of Nancy 2, Campus scientifique, B.P. 239, 54506 Vandoeuvre-Lès-Nancy France*
[3]*ITESOFT, Parc d'Andron - Le Séquoia - 30470 Aimargues - France*

Abstract

In this paper, an approach for adaptive region segmentation of mail-order forms for high volume application is described. Regions are first identified through a selection of their anchor points described by a constraint graph, illustrating their typographic aspects in the nodes, and their topographical relationships in the arcs. Then the identification of the actual anchor points is performed from a list of textual candidates, using the Arc Consistency Algorithm (AC4). Finally, some contextual heuristics are investigated for properly delimiting the regions. The originality of this approach lies mainly in the absence of a rigid a priori model, replaced by a simply and reliable association of anchor points. The constraint graph used for their description can be easily derived from a general logical definition of their content. Experimental results are overall encouraging and the methodology integration is under execution for commercialization.

1. Introduction

Nowadays, forms are among the most widespread documents circulating in the administrations and organizations, used for business transactions, insurance declarations, mail-ordering sales, etc. They vehicle a concise information related to different aspects : medical, financial or commercial. Although the information is often strictly organized, there are a number of limitations with automatic processing. These include the complexity of multi-modal writing recognition (faced to a mixture of printed and handwritten characters), difficulty of region location, vulnerability of layout interpretation (always changing), quality of forms, noise, etc .

Many pieces of work have been done on form recognition in the last two decades. The methods investigated in these pieces are based on some criteria depending on the complexity of the information organization and structure. These criteria denote the gradual complexity in form analysis which grows up with the structure variability and information grouping. It seems obvious that the most favorable case is when the information is grouped into cells [2] within identified regions in a stable structure [1,5,6]. At the opposite, more investigations will be needed for distributed information, not organized in any structure and where the structure is always changing [3,4,7,9].

The mail–order forms considered here fall in this last category. This is because the sailors consider a mail-order as a shopping window where they can include advertising, some attractive announcements and special offers. This leads to a deep modification of the layout, a damage of region limits, an introduction of some irregularities and suppression of helpful writing boxes, etc. Addition to this structure damage, some other drawbacks can come from the manually filling of the form, introducing fancy writings and leading to some line overlapping.

Figure 1 shows (at the top) an example of such form. The main regions, reproduced at the bottom, correspond to client address (ZA), correspondence address (ZC), references (ZR), amounts (ZM) and payment (ZP).

2. Adaptive technology

2.1. The point of view

Faced to this chronicle structure variation within mail-order forms, we have proposed an adaptive technology for form region location, enabling the use of an a priori fixed model. Its properties can be summarized in the following points:

- To continuously suit the layout changing without modifying the region definition and the segmentation strategy;
- To be less sensitive to the typography alterations and modifications;
- To break free from mail-order modeling. This opposes the sailor expertise to a fixed model for region description. The former gives indication on the logical content without fixing a rigid frame which can be altered by the layout changing.

0-7695-1263-1/01/$10.00 © 2001 IEEE

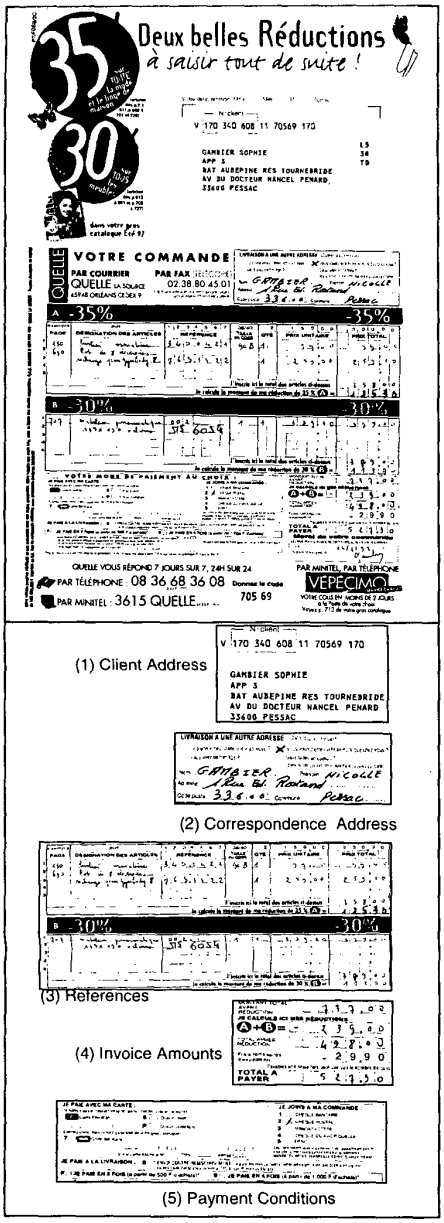


Figure 1 : Example of a mail-order form and its main regions.

2.2 . Sailor expertise vs. fixed model

The logic of the order form composition is expressed by some rules giving the relative position and alignment of its main regions.

- Rule 1: The reference region has a central position in the form;
- Rule 2 : The amount region is located below the reference region, on its right side, as an extension of the amount column;
- Rule 3 : The payment condition region is located below the reference region, on the left of the invoice amount region;
- Rule 4 : The client and correspondence address are situated above the reference region.

Figure 2 illustrates these rules.

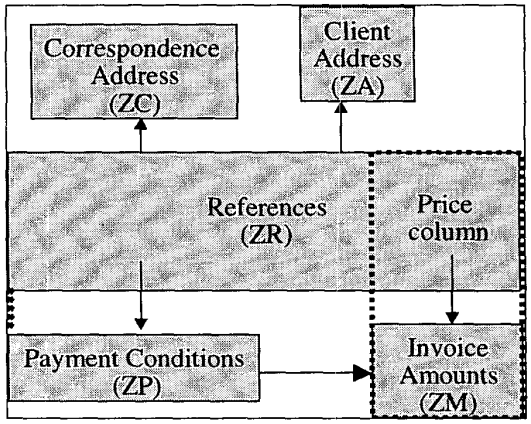


Figure 2: Physical interpretation of the mail-order logical structure

→ Relative position.

..... Relative alignment.

2.3. Region signature

For the adaptive technology, a region is characterized by a "signature" or a "print" representing the group of some specific information elements, characterized by either the same structure or the same nature. Furthermore, these elements are often highlighted by the use of some specific key-words or headings. This is very usual in forms where the content is always distributed in very specific regions. Contrary to the element contents, the headings are less altered by the noise and constitute a good support for the recognition of noisy images. For the mail-orders this is particularly the case of all the five regions mentioned at the bottom of the Figure 1. Each region is then characterized by a set of a stable and relevant structural elements.

2.4. Region representation

The previous remarks lead to base the region representation mainly on its signature description. This amounts to dispose of a data structure for each region revealing its specific keywords (named anchor points), individually and globally by giving the relationships

between them, within the same structure. Furthermore, as the anchor points should be insensitive to noise and layout variations, their definitions can be enough general to absorb these defects. The solution employed in this system is to use a constraint graph where the nodes represent the anchor points, and the arcs the relationships between them.

3. Related work

The form segmentation problem is transformed into a graph matching problem. The matching is realized by making consistent each constraint graph representing a region. For this, we have used the Arc Consistency Algorithm AC4, as described in [8]. The connected component analysis is used to cluster the components into groups that should be employed to locate the anchor points set in a given region. Thus, an anchor point is a group of connected components defined by computing a set of features which integrate pertinent geometric information. These measurements are then used for selecting character groupings as the possible candidates for the anchor point nodes.

3.1. Anchor point determination

Because of noisy images, we avoided the use of OCR to identify the text and based the anchor point extraction mainly on connected components groupings. So, from an image processing point of view, an anchor point is defined by a properly grouping of close connected components (CCG).

3.2. Connected component grouping.

Connected components (CCs) are first extracted from the form image and then grouped together by considering their vertical and horizontal proximity. Because of noise and font style spacing variability, the proximity searching is obtained by defining a tolerance surface whom the horizontal and vertical values are deduced from statistics performed during a primary step on word spaces within anchor points. This tolerance allows the system to dispose of thin or wide margins according to the font style used for the anchor point. Figure 3(a) shows a group (CCG) containing superposed CCs. The CC grouping procedure is primordial in this approach. The corresponding algorithm is presented in Figure 4.

From CCG_i groups the analysis is refined in order to detect more strict horizontal and vertical alignments

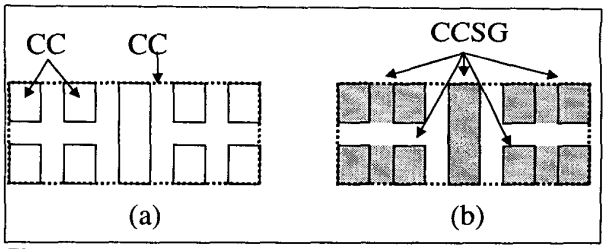


Figure 3 : Grouping definitions.

```
Algorithm

Initialization :
CCG_i = {CC_i} ∀i ∈ [1..n]
ListCCG = {CCG_1,....,CCG_n}
Modified_ccg = true;

Body:
While (Modified_ccg) do
        For each ccg of ListCCG do
                Modified_ccg = false;
                ccgSel = first element ofListCCG noticed ListCCG (0)
                If (ccg = ccgSel) do
                        ccgSel = ListCCG(1)
                Endif
                J=0;
                // We are looking for the first group which verifies the
                // proximity criterion with ccg
                While no (close(ccg, ccgSel)) and J<n do
                        J=J+1;
                        ccgSel = ListCCG(j)
                        If (ccg = ccgSel) do
                                J=J+1;
                        Endif
                End while
                If ccgSel is found do
                        ccgSel = ccgSel ∪ccg
                        Delete ccg from ListCCG
                        Modified ccg = true;
                Endif
        End do
End do
```

Figure 4 : CC grouping algorithm.

creating some sub-groups CCSGs in each CCG_i. In the horizontal case, we look for a succession of CCs horizontally close and without superposition. In the vertical case, only superposed and not consecutive CCs are considered. In both cases, the proximity criterion is more strict than for the groups (see Figure 3(b)).

In the horizontal detection of subgroups, one problem can occur due to the proximity of successive lines (accentuated by the presence of noise, or touching characters). This leads to find words distributed in several sub-groups (see Figure 3(b)).To overcome this defect, we try to reassemble these sub-groups in order to constitute a more logical element containing the two lines.

The objective of the grouping optimization step is to reduce the number of its candidate anchor points for each region. This is based on the consideration of some grouping properties of the region. Practically, one or more anchor points are located in a CCG_i. All the CCG_i containing the anchor points of the same region should

respect a given criterion of that region. Consequently, these groups can be reassembled in grouping sets according some horizontal and vertical alignment criteria ($SCCG_k$).

3.3. Anchor point selection

The labeling procedure is based on the set of groups ($SCCG_i$) and on the set of sub-groups derived from $SCCG_i$. Groups and sub-groups are performed according to specific proximity tolerance chosen for each region.

Once defining the constraint propagation method basis, the next procedure consists in a primary labeling of the graph, then in propagating the constraints in order to erase the erroneous labels.

3.4. Region extraction

Once the anchor points are selected, each region is delimited by the extension of its anchor point areas and also by some neighboring rules between the regions. The anchor point area is different for different region. For example, considering the reference region structure composed of product lines, the amount area corresponds to the product price column. Figure 5 shows an example of reference region area delimitation where the column limits have been rectified in order to encapsulate the filled zones. This is done by taking into account the width of some regular elements.

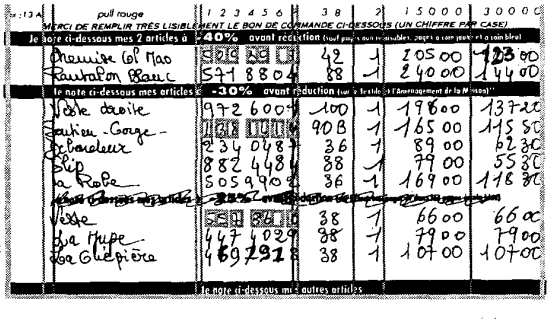


Figure 5 : Reference region coulum delimitation.

4. Experiments and results

We have thoroughly tested our application on two French mail-order business sellers form classes. The first database is composed of 800 images issued from 8 different lots. The second database is composed of 950 images. The images come from a real line production work.

4.1 First database results

Table 1 illustrates the total success rate on the global database and individually on each lot.

Table 1 : Segmentation rates for the first database.

	Number of images	Success %	success / zone %				
			ZR	ZM	ZC	ZA	ZP
Database	800	52	83,13	67,63	99,5	95,38	78,75
M072	25	72	100	96	96		92
M769	55	89	90,91	90,91	100	98,18	87,27
M511	500	68	92	92	99,4	97,8	88
M750	27	0	88,89	0	100	100	88,89
M733	24	13	91,67	12,5	100	100	41,67
M745	101	0	27,72	1,98	100	100	31,68
M251	37	0	75,68	8,11	100	100	62,16
M865	31	0	90,32	0	100	100	96,77

The main failure reasons are due to :

A. Splitting of zones ZR and ZM (lower border of ZR = upper border of ZM).
B. Location of the reference zone (see Figure 6).
C. Internal splitting of ZM (see Figure 7) .

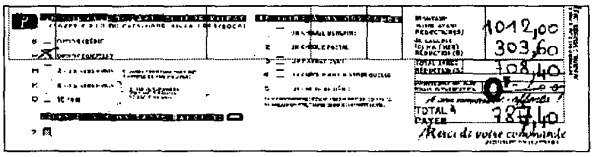


Figure 6 : Failure cause of type B.

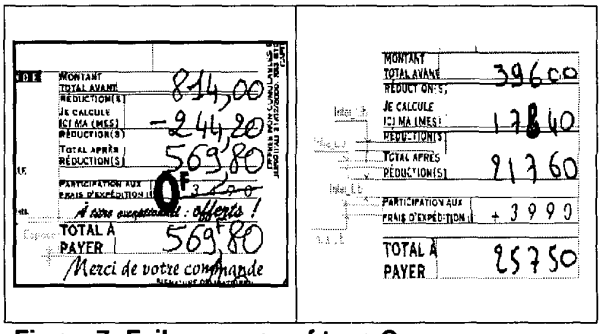


Figure 7: Failure cause of type C.

The detection can be improved by taking into account more relationships between reference, amount and payment regions, and by modifying the region detection order. The real algorithm doesn't take into account neither the fact that the low border of the reference region corresponds to the upper border of the payment region nor the high border of the payment region is aligned with those of the amount region.

A detailed study allowed the identification of the most frequent failure causes zone by zone, as synthesized in Table 2.

Table 2 : Main failure causes for region detection.

ZR	- No detection of anchor points because of noise, - Slant, - Bad splitting of the low border
ZC	- Noise on the wordings, - Handwriting overlapping of the printed text located above the zone or near the reference article wordings
ZA	- Not detection of the anchor point (because it is rare), - Bad delimitation of the zone.
ZM	- Bad splitting in the zone, - Bad detection of the high border. - Presence of an abundant advertising information in the proximity of the useful data. - Non regularity of the space items.
ZP	- Detection error of the amount zone low border: this border is used to delimit the search zone anchor points, - Big marks in the boxes.

Different improvements are planed essentially for the reference and amount zones.

4.2 Second database results

Table 3 illustrates the total success rate and failure rate on the global database. The success rate is very interesting for this form class. Indeed, during the conception phase of the form, the logical structure is slightly modified. The advertisement information are added around the interested regions. Is not the case for the first class.

Table 3 : Segmentation rates for the second database.

DataBase 950 images	Image in full	ZR	ZM	ZA	ZC
success	88,60%	93,0%	98,5%	97,8%	97,7%

5. Conclusion and perspectives

In this paper we have presented an adaptive technology for automatic segmentation of mail-order forms into peculiar information regions. This technology is based on a specific zone modeling in terms of an association of their anchor points, and on some contextual rules for border delimitation. In the preliminary experiments on some selected mail-order forms, very encouraging results have been obtained, in particular, anchor point extraction accuracy is satisfactory. Tests made by the ITESOFT company for high volume documents have lead to refine the initial technique by taking into account the topographical relationships between the regions. In the secondary experiments, the technique yields an acceptable segmentation accuracy allowing its integration into the production lines.

Future work will aim at improving the overall system performance and accuracy for others kinds of order-mail forms and invoices. In particular, the generic construction of the graph constraints from the sailor indications will also help to adapt the technology to the layout changing. Another line of improvement concerns the introducing OCR in the anchor point recognition. It seems usually rather easy to tune commercial OCR software to read forms essentially where the information can be located precisely.

6. References

[1] H. Arai and K. Okada , "Form Processing Based on Background Region Analysis", ICDAR'97, Ulm , Germany, pp. 164-169, 1997.

[2] Y. Belaïd and A. Belaïd, "Form Analysis by Neural Classification of Cells", International Workshop on Document Analysis Systems, Nagano, Japan, November 4-6, 1998.

[3] F. Cesarini, M. Gori, S. Mariani and G. Soda, "INFORMys : A Flexible Invoice-like Form Reader System". IEEE Trans. PAMI, 20(7):730-745, July 1998.

[4] Y. Ishitani , "Model Matching Based on Association Graph for Form Image Understanding". Proceedings of the IEEE ICDAR, Montréal, Canada, 1995, pp. 287-292.

[5] S. Kebairi and B. Taconet, A.Zahour, S. Ramdane "A Statistical Method For an Automatic Detection of Form types", Proceedings of the DAS'98, Nagano, Japan, November 4-6, 1998,pp.109-118.

[6] S. W. Lam, L. Javanbakht and S. N. Srihari. "Anatomy of a Form Reader". IEEE ICDAR, pp. 579-582, 1995.

[7] J. Lii and S. Srihari, "Location of Name and Address Cover Pages", ICDAR'95, Montreal, Canada, 1995, pp. 756-759.

[8] R. Mohr and T. C. Handerson. "Arc and Path Consistency Revisited". Artificial Intelligence, 28:225-233, 1986.

[9] J. J. Yuan , Y. Y. Tang and C. Y. Suen, "Four Directional Adjacency Graphs (FDAG) and their Application in Locating Fields in Forms". Proceedings of the IEEE ICDAR Montréal, Canada, 1995, pp. 752-755.