HAL
archives-ouvertes.fr

A generic method for structure recognition of handwritten mail documents

Aurélie Lemaitre, Jean Camillerapp, Bertrand Coüasnon

▶ **To cite this version:**

Aurélie Lemaitre, Jean Camillerapp, Bertrand Coüasnon. A generic method for structure recognition of handwritten mail documents. Document Recognition and Retrieval DRR XV, Jan 2008, San Jose, United States. 2008. <inria-00308565>

HAL Id: inria-00308565
https://hal.inria.fr/inria-00308565

Submitted on 2 Dec 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A generic method for structure recognition of handwritten mail documents

Aurélie Lemaitre[a], Jean Camillerapp[a] and Bertrand Coüasnon[a]

[a]IRISA/INSA, Campus de Beaulieu, 35042 Rennes Cedex, France

ABSTRACT

This paper presents a system to extract the logical structure of handwritten mail documents. It consists in two joined tasks: the segmentation of documents into blocks and the labeling of such blocks. The main considered label classes are: addressee details, sender details, date, subject, text body, signature. This work has to face with difficulties of unconstrained handwritten documents: variable structure and writing.

We propose a method based on a geometric analysis of the arrangement of elements in the document. We give a description of the document using a two-dimension grammatical formalism, which makes it possible to easily introduce knowledge on mail into a generic parser. Our grammatical parser is LL(k), which means several combinations are tried before extracting the good one. The main interest of this approach is that we can deal with low structured documents. Moreover, as the segmentation into blocks often depends on the associated classes, our method is able to retry a different segmentation until labeling succeeds.

We validated this method in the context of the French national project RIMES, which proposed a contest on a large base of documents. We obtain a recognition rate of 91.7% on 1150 images.

Keywords: mail, structure recognition, grammar, multiresolution, perceptive vision

1. INTRODUCTION

Nowadays, firms has to deal with the problem of incoming mail processing. Indeed, tasks such as mail subject identification (complaint, order, details modification ...), exact addressee determination are required before the final mail processing by the appropriate person. The fields of interest in a letter are different according to the step of treatment. Automation of such treatment would improve productivity inside of companies.

This paper presents our method for the extraction of the logical structure of French handwritten mail documents, but could be easily applied to other languages. It consists in two joined tasks: the segmentation of documents into blocks and the labeling of such blocks.

This work has to deal with difficulties of unconstrained handwritten documents: variable structure, adaptation to writer. Indeed, even if French language gives some conventions for writing a mail, the structuring is not always respected in handwritten documents. Thus, we are faced with very variable documents: some examples are given in the figure 1. Our goal is to detect elements like addressee details, sender details, date, subject, text body, signature, opening.

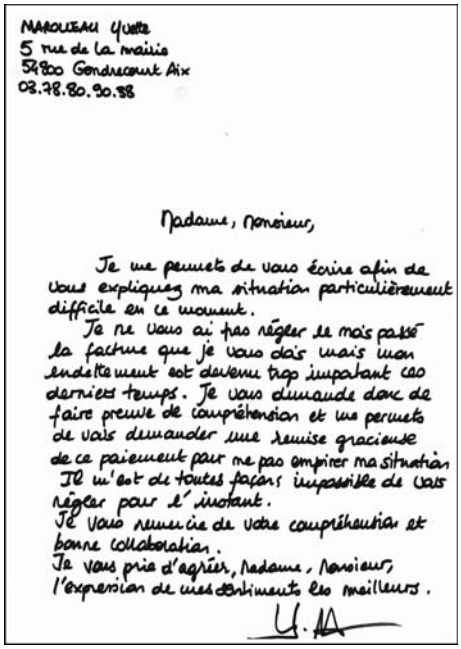
The first difficulty is to know how to segment the page into blocks. For example, consider the four text lines on the top right of mails 1(a) and 1(c). In a bottom-up approach, those text lines would probably be grouped in a single block. But in the case of mail 1(a), there are three text lines of addressee details and one giving date and place, whereas in mail 1(c), the four lines give addressee details. Consequently, we have to propose a method able to adapt the document segmentation into blocks according to the labeling.

Moreover, each element is not always present in every mail. Thus, date and place are only present in figures 1(a) and 1(d); addressee details are absent in figure 1(b). Then, we need to take into account several possible configurations of mail pages.

In the next section, we present a summary of approaches found in the literature. Then, we explain the generic method we are based on, and how we adapt it to mail documents structure recognition. In the fifth section, we present the evaluation of our work into the French national project RIMES. Results are then discussed.

Further author information: (Send correspondence to Aurélie Lemaitre)
Aurélie Lemaitre: E-mail: aurelie.lemaitre@irisa.fr, Telephone: +33 (0)2 99 84 75 39

(a)



(b)

(c)

(d)

Figure 1. Examples of handwritten mail documents

2. RELATED WORK

Several works have been proposed in the field of printed commercial document recognition, that includes mail document recognition. In the case of those printed documents, OCR are often applied, which makes possible the detection of keywords that are used for block classification. Thus, Lim *et al.*[1] propose an analysis of low resolution fax cover pages. Their work consists in classifying pages according to their global aspect, to segment them into blocks and then to use keywords to label these blocks. An approach based on keywords is also proposed by Likforman.[2]

In the case of handwritten documents, from various writers, difficulties overcome for applying an OCR on the whole text. This problem encouraged people to restrict their analysis to specific goals. Thus, Koch *et al.*[3] propose a HMM based analyzer to extract numerical fields in handwritten incoming mail documents. Their work aims at detecting fields like phone numbers, customer reference, zip code .... However, the extraction of numerical fields is not sufficient for labeling a whole mail document.

In order to label whole mails, some works have been proposed using only geometric information. Thus, Sainz Palmero *et al.*[4] propose a complete analysis of mail documents based on a recurrent neuro-fuzzy architecture. To classify blocks, they use structural and geometric data, such as x-y positions, number of lines, and contextual data like the relative positions of blocks. Matrakas *et al.*[5] propose another method based on geometric information, that uses the nearest neighbor rule technique. The use of geometric elements for labeling gives interesting results.

However, those two methods assume that the first step of document segmentation into blocks has been properly realized. RLSA (Run Length Smoothing Algorithm) is used for example in Ref. 5. This is convenient for printed document. But in the case of low structured handwritten documents, we have shown in introduction that the segmentation into blocks is strongly linked with labeling. So the correct segmentation of document can not be found without any information about future labeling.

We propose a method based on structural and contextual geometric information. We introduce our work in an existing generic system, the DMOS method, that makes it possible to try successive page segmentations (contrary to Refs. 4 or 5) until labeling succeeds for each block. Thus, we can deal with unconstrained low structured handwritten mail documents.

3. GENERIC DMOS METHOD

3.1 Principle of the method

DMOS method[6](Description and MOdification of Segmentation) is a generic method for structured document recognition.

It is based on two levels. Digital level is made of image tools such as a line segment detector based on Kalman filtering[7] or a connected component extractor. Over these digital data is the symbolic level based on a grammatical language EPF (Enhanced Position Formalism). This language makes it possible a structural, syntactic and semantic description of any studied kind of document, and a control of digital level.

Once a description of a kind of document has been realized in EPF language, the associated parser is automatically produced by the system, after a compilation stage. Thus, this method is generic because the knowledge is grouped in EPF description and separated from the system. This method has been applied for several kinds of documents:[8] music scores, mathematical formulas, table structures[6] or archive documents,[9] which shows its genericity. It has been validated at a large scale, on more than 500,000 document pages.

3.2 EPF language

EPF language is a two dimension grammatical formalism. It makes it possible the description of the structure of various kinds of document.

An example of a simplified grammar is given for the description of the kind of document like figure 1(b).

```
mailPage::=
    AT(topLeftPage) &&
        senderDetails &&
    AT(middlePage) &&
        opening &&
    AT(underOpening)&&
        textBody &&
    AT(underBody) &&
        signature.
```

It means that a mail page is made of four elements: sender details, opening, body and signature, that are disposed in the page at special relative positions. Each element is detailed in a specific rule that independently describes it.

Symbol `&&` is a concatenation operator. Operator `AT` makes it possible to describe a search zone. For example `topLeftPage` is a search zone. In this zone, we apply the rule `senderDetails` that looks for sender details, and so on to find the described elements in the page.

3.3 Use of perceptive vision

We introduced in DMOS method a notion of perceptive vision. Indeed, we shown[10] how a text-line can be described at low resolution as a line segment, or at high resolution as a succession of connected components.

Consequently, we updated DMOS method with the operator `FOCUSING` that gives the possibility to describe a document using different resolution levels for the same image. More details are given in Ref.9. A detailed example of application is presented for the case of mail documents in part 4.2.

4. STRUCTURE EXTRACTION OF MAIL DOCUMENTS

We study handwritten French mail documents (see figure 1 for examples). We propose to classify blocks into height classes: sender details, addressee details, subject, opening, date and place, text body, signature, attachment or post scriptum. We must notice that those elements are not always all present.

We apply DMOS method to this kind of documents. It consists in two steps:

- Choose the number of resolutions and the associated features required for perceptive vision.
- Give a symbolic grammatical description of document in EPF language.

4.1 Digital data

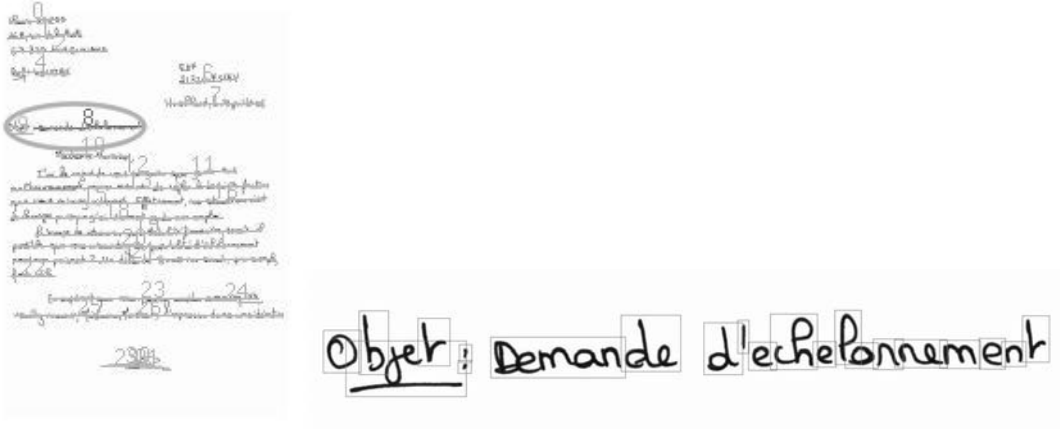
In the case of mail documents, our digital data is made of two resolution levels of image and our analysis is based on two kinds of features (terminals):

- Line segments extracted in low resolution image (dimensions divided by 16), that give an idea of text line positions.
- Connected components extracted in initial resolution image, that are made of full characters, parts and groups of characters.

Those elements are extracted using existing image tools. Then, the main work consists in describing the relations between those elements.

4.2 Symbolic description

The symbolic description consists in determining relation between elements for the different existing configurations.



(a) Detection of *lineSegment* number 8 at low resolution

(b) Focusing at initial resolution on this zone and detection of *groupOfComponents*

Figure 2. Application of rule *textLine*

4.2.1 Text line description

We base our analysis on the "text line" entity. Consequently, the first step consists in describing in EPF language the notion of "text line".

A text line can be seen as a succession of aligned connected components, supplemented by small components like punctuation or accents, which position varies around the line. Thus, in the case of varying handwritten documents, it is quite complex to determine all the elements that belong to a text line.

We propose a method based on perceptive vision. We extract each text line in two steps: first find a line segment at low resolution, and then focus attention in the associated zone to detail connected components. This concept is translated in EPF language by the following rule. The analysis begins at low resolution.

```
textLine::=
        lineSegment &&
    AT(lineSegmentZone) &&
        FOCUSING ON(upperResolution) FOR(groupOfComponents)
```

This rule is applied on figure 2.

Thanks to this line definition, document description consists in the description of text line arrangements.

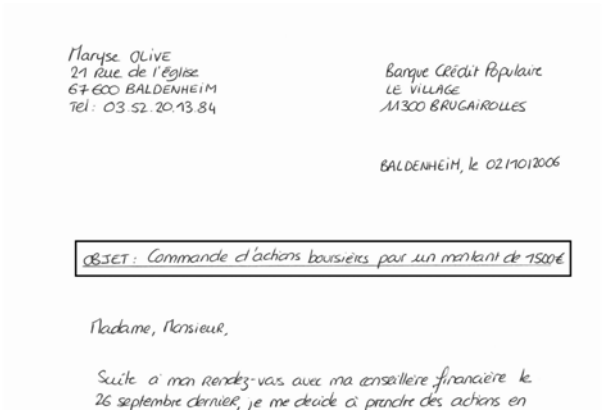
4.2.2 Full document description

This part of the system is a description of all the knowledge included for mail description.

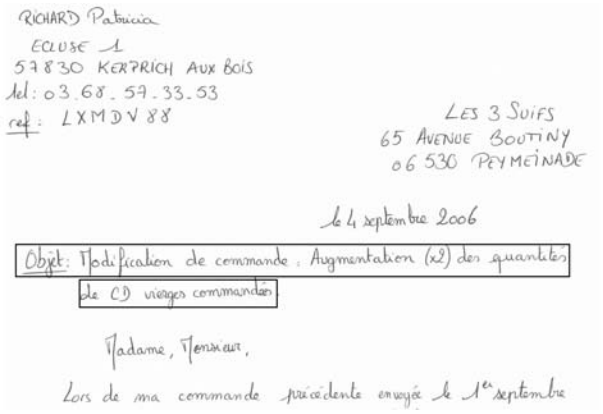
We base our analysis on rules commonly used for French mail: sender details are on the top left of the page, recipient details, date and place are on the top right, subject is before opening, and text body ends with a signature. However, we have to deal with a large variety of configurations, and the possible absence of each element.

In order to limit ambiguities in analysis, we chose a special order of analysis. Indeed, signature, text body and opening are almost always present, so we first begin analysis from the bottom of the document. We find successively a signature, the text body and an opening. Then, we can concentrate on the upper part of the document to find remaining elements.

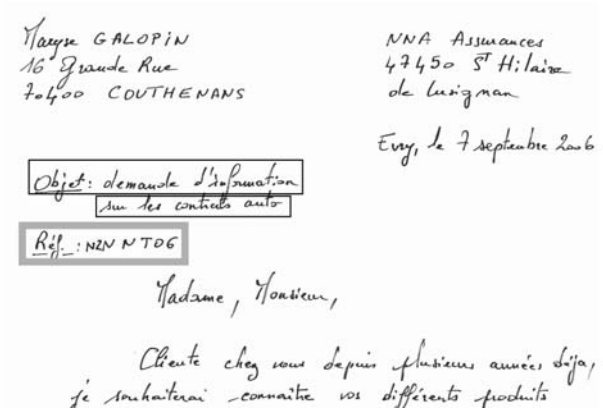
We recall that all this knowledge is easily described using EPF formalism.



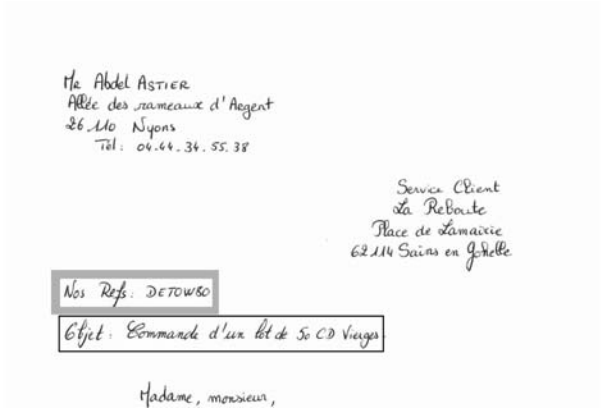
(a) Success on rule 6



(b) Success on rule 3



(c) Success on rule 2



(d) Success on rule 5

Figure 3. Recognition of subject (black bounding box) and identifier (grey bounding box): application of a different rule according to geometric data.

4.2.3 Precise example

In order to show the interest of our method in the case of complex documents, we propose to detail a small part of the analysis.

We take the example of two elements: subject and customer identifier (a part of sender details), that are sometimes very close to each other. They are both located in the left part of the document, just over opening. Their presence is facultative, and their order is not fixed. Various combinations are proposed. See an example in figure 3. Subject is represented with a thin black bounding box, identifier with a large grey bounding box. In mails 3(a) and 3(b), we can find a subject on one or two lines but no identifier. A customer identifier is present in mails 3(c) and 3(d), respectively under and over subject fields, made of two and one text-lines. This example shows a precise case in which the choice of segmentation is linked with labeling.

We describe this different combinations thanks EPF language.

We call `SubId` the rule made to search `Sub`ject and/or customer `Id`entifier.

A `SubId` can be made of three text lines, beginning either with an identifier or with the subject. We deduce two alternative rules:

```
(1) SubId::=
           shortIdentifierLine &&
       AT(underLine)&&
           longSubjectLine &&
       AT(underLine)&&
           subjectComplementLine.
(2) SubId::=
           longSubjectLine &&
       AT(underLine)&&
           subjectComplementLine &&
       AT(underLine)&&
           shortIdentifierLine.
```

A `SubId` can also be made of two text lines, so we complete the description with three new rules:

```
(3) SubId::=
           longSubjectLine &&
       AT(underLine)&&
           subjectComplementLine.
(4) SubId::=
           longSubjectLine &&
       AT(underLine)&&
           shortIdentifierLine.
(5) SubId::=
           shortIdentifierLine &&
       AT(underLine)&&
           longSubjectLine.
```

Else, if there is just one line, we write the rules:

```
(6) SubId::=
           longSubjectLine.
(7) SubId::=
           shortIdentifierLine.
```

The last rule describes the absence of subject or identifier.

```
(8) SubId::=
           noLine.
```

In those rules, `shortIdentifierLine` will produce a zone labeled "Sender details". `longSubjectLine` and `subjectComplementLine` are "Subject". Each of these elements is detailed in a specific rule. The difference between `shortIdentifierLine` and `subjectComplementLine` is the presence of indentation at the beginning of `subjectComplementLine`. The parser will succeed on the rule `(8)` even if nothing is recognized.

The parser is based on these rules: during an analysis, each rule will be tried in the defined order, until one succeeds. Thus, we give in figure 3 the number of the first succeeding rule.

This example shows the importance of an adaptable system, and how our method makes it easy to add another configuration. We notice that, in that case, the final segmentation into blocks (grouping or not text lines into the same class) is decided after labeling. This is not the case with the methods we found in the literature.

5. APPLICATION ON A LARGE BASE

5.1 RIMES contest

We introduced our work in the context of the national French project RIMES.[11] RIMES stands for "Reconnaissance et Indexation de données Manuscrites et de fac similES", which means "handwritten data and fax recognition and indexing". This project is financed by the French Ministries of Research and Defence. Its goal is to build a large handwritten document database (more than 5000 mail documents), to define criteria and metrics for the evaluation campaign using those data, and to drive this campaign. Nine French research teams took part in this project as participants on different tasks linked with document recognition and retrieving.

For the first campaign, the base supplied by RIMES is made of 1050 French mail images for learning systems, and 100 images for the first competition. Other images will be provided in future competitions. Images are stored in PNG and have a size around 2500*3500 pixels. They have been manually annotated. Indeed, each zone to label is represented by the coordinates of a rectangle and its associated class. Annotations are stored as a list of boxes in a XML file for each image, called *validation*.

The objective is to produce as a result a list of labeled box describing the document, in an XML file called *hypothesis*. Results are computed by comparison of *validation* and *hypothesis* files. In the metric proposed by RIMES project, we count each pixel of the image that has been wrongly labeled, balanced by its grey-level. Indeed, a white incorrect pixel has a small impact, whereas a dark incorrect pixel is a big mistake.

We produced the rules of our system by studying 300 of the 1050 given images. Then, we applied our system on the competition test set made of 100 images. RIMES metric gives one score of global error rate: we obtained 9.27% of error for our system, which correspond of the percentage of pixels wrongly labeled. As global results of RIMES contest have not been published yet, we can not compare with other competitors.

5.2 Results on a large base

The metric used by RIMES is not very significant according to us. Indeed, using grey level causes mistakes when the background is not white but slightly grey. So we decided to base our metric on a binary image. Then, our recognition rate correspond to the number of black pixels that has been properly labeled. We can notice that this level of precision is necessary because, contrary to other method like,[4] we have to evaluate classification but also segmentation into blocks.

Moreover, we think it is necessary to have a separate result for each class. We detail results obtained with our metric based on binary images, in table 1. Results are presented for three bases. Indeed, we used 300 images as a kind of learning base, as we partly watched them to produce the rules. *Recall* and *Precision* are defined below, in our case:

$$Recall = \frac{NbCorrectPixelForTheClass}{NbExpectedPixelForTheClass}$$

$$Precision = \frac{NbCorrectPixelForTheClass}{NbFoundPixelForTheClass}$$

We obtain a global recognition rate (recall) of 91.7% for the base of 1150 images, which correspond of the percentage of pixels wrongly labeled, out of the 254,207,489 black pixels we have to label. The global precision is 92.6%.

The third column gives the importance, in number of pixels, of the classes. "Text body" represents the main part, 61.5% of pixels. We can see that recognition rate is globally proportional to the importance of the class. "Subject" recognition is difficult because it can be easily confused with several other classes: "Date, Place"', "Sender details" or "Opening". Improving these three classes would automatically improve "Subject" recognition rate.

It is important to see that the results on the test base are close to the results on the learning base.

Class	Involved pixels	Learning base 300 images		Test base 850 images		Whole base 1150 images	
		Recall	Precision	Recall	Precision	Recall	Precision
Text body	61.5%	98.0%	96.8%	97.0%	97.0%	97.2%	96.9%
Sender details	15.1%	93.4%	93.0%	91.5%	91.7%	92.0%	92.1%
Addressee details	9.0%	87.3%	85.5%	83.0%	83.3%	84.1%	83.9%
Signature	4.2%	84.4%	90.2%	90.6%	90.9%	88.9%	90.7%
Subject	4.0%	68.4%	70.5%	65.7%	72.7%	66.4%	72.1%
Date, Place	3.2%	53.1%	79.1%	54.3%	78.2%	54.0%	78.4%
Opening	2.9%	85.8%	80.4%	79.5%	74.5%	81.1%	76.0%
Global	-	92.6%	92.7%	91.4%	92.6%	91.7%	92.6%

Table 1. Results on 1150 handwritten mail documents

5.3 Discussion

Our analysis is only based on the structure so it is independent from the writer, which is convenient for handwritten documents. However, we are sometimes faced with cases that cannot be decided: for example when sender and addressee details are reversed. In that case, structural analysis is insufficient.

The structure is simply expressed using EPF formalism: rules describe the relations between text line entities. This work has been applied for French mail documents but could be applied for any language, assuming rules are given.

A way to improve results would be to use locally a keyword or numerical field extractor. Thus, knowing an hypothesis for labeling a block can reduce the vocabulary required to recognize words. For example, if a block is supposed to contain the date, we could try to precisely detect a year or a month, which presence could confirm our hypothesis. The presence of numerical fields could also decrease ambiguities between classes. The introduction of a field extractor would be easy because it can be called directly from EPF rules, and depending on the context of analysis.

6. CONCLUSION

We present a method for handwritten mail document segmentation. Our work is based on the generic method DMOS, updated by the notion of perceptive vision. Our analysis combines two image levels: low and high resolutions, that compose the digital data.

The knowledge is separated from the system and easily expressed with EPF language. Document structure is directly and intuitively described. This is very convenient for this low structured handwritten mail documents because it makes it possible an easy definition of each possible arrangement of elements. If a new configuration is found in a document, we just have to add a rule which does not cause any trouble on the existing system.

Moreover, contrary to other methods, the document segmentation and the labeling are realized together, which is necessary in the case of unstructured handwritten mail documents, where blocks are not always well defined. Indeed, rules are applied successively until one succeeds. If none succeeds, it means that the searched element is not present, which is very common in our documents.

Our method has been evaluated in the context of the national RIMES campaign and obtains a recall of 91.7% and a precision of 92.6% on 1150 images. Our participation to RIMES project will permit us to apply our work on largest bases as soon as they are provided for RIMES competitions. Next improving will be the introduction of a classifier to detect numerical fields or keywords.

REFERENCES

1. Y.-K. Lim, H.-J. Kang, C. Ahn, and S.-W. Lee, "Structure analysis of low resolution fax cover pages," in *Document Analysis Systems*, S.-W. Lee and Y. Nakano, eds., *Lecture Notes in Computer Science* **1655**, pp. 99–113, Springer, 1998.

2. L. Likforman Sulem, "Name block location in facsimile images using spatial/visual cues," in *Proceedings of International Conference on Document Analysis and Recognition (ICDAR'01)*, pp. 680–684, 2001.
3. G. Koch, L. Heutte, and T. Paquet, "Numerical sequence extraction in handwritten incoming mail document," in *Proceedings of ICDAR'03*, pp. 369–373, 2003.
4. G. I. S. Palmero and Y. A. Dimitriadis, "Structured document labeling and rule extraction using a new recurrent fuzzy-neural system.," in *Proceedings of International Conference on Document Analysis and Recognition (ICDAR'99)*, pp. 181–184, 1999.
5. M. D. Matrakas and F. Bortolozzi, "Segmentation and validation of commercial documents logical structure," in *ITCC*, pp. 242–246, IEEE Computer Society, 2000.
6. B. Coüasnon, "DMOS: A generic document recognition method to application to an automatic generator of musical scores, mathematical formulae and table structures recognition systems," in *Proceedings of International Conference on Document Analysis and Recognition (ICDAR'01)*, pp. 215–220, 2001.
7. I. Leplumey, J. Camillerapp, and C. Queguiner, "Kalman filter contributions towards document segmentation," in *Proceedings of International Conference on Document Analysis and Recognition (ICDAR'95)*, pp. 765–769, 1995.
8. B. Coüasnon, "DMOS, a generic document recognition method: Application to table structure analysis in a general and in a specific way," *International Journal on Document Analysis and Recognition, IJDAR* **8(2)**, pp. 111–122, 2006.
9. A. Lemaitre, J. Camillerapp, and B. Coüasnon, "Contribution of multiresolution description for archive document structure recognition," in *Proceedings of International Conference on Document Analysis and Recognition (ICDAR'07)*, pp. 247–251, 2007.
10. A. Lemaitre and J. Camillerapp, "Text line extraction in handwritten document with kalman filter applied on low resolution image," in *Document Image Analysis for Libraries (DIAL'06)*, pp. 38–45, 2006.
11. RIMES Project, "http://www.int-evry.fr/rimes/."