Contribution of Multiresolution Description for Archive Document Structure Recognition

Aurélie Lemaitre, Jean Camillerapp, Bertrand Coüasnon

Contribution of Multiresolution Description for Archive Document Structure Recognition

Aurélie Lemaitre Jean Camillerapp Bertrand Coüasnon
IRISA - INSA
Campus universitaire de Beaulieu
35042 RENNES CEDEX
aurelie.lemaitre@irisa.fr

Abstract

When reading a document, we intuitively have a first global approach in order to determine the whole structure, before reading parts in details. We propose to apply the same kind of mechanism by introducing the concept of multiresolution in an existing generic method for structured document recognition. This new combination of different vision levels makes it possible to recognize low structured documents.

We present our work on an example: the multiresolution description of archive documents that are naturalization decree registers from the 19th and 20th century. The validation has been made on 85,088 images. Integrated in a platform for archive documents, the located elements offers to users a fast leaf through naturalization decrees.

1. Introduction

In the field of document structure analysis, we presented, in various papers ([3][4]), DMOS (Description and MOdification of Segmentation), a generic method for structured document recognition. This method is made of a formalism that can be seen as a description language for document structure. It is particularly convenient for ancient documents because it can deal with noise and text variations.

In the case of low structured documents, like a tabular structure without rulings for example, it can be easier to study, in a first step, the document at low resolution in order to extract a global structure, and then to focus on interesting parts. In this paper, we propose to express this mechanism by introducing multiresolution analysis in our generic description method.

We start this paper with related work on multiresolution vision in recognition systems. Then, we present our generic method and the introduction of multiresolution. In sections 4 and 5, we show the interest for low structured archive document recognition and present results.

2. Related work on multiresolution in document analysis systems

Mao *et al.* present in [9] a good survey of document structure analysis algorithms. They explain that many systems use grammars to describe hierarchical document physical layouts. On another hand, they identify top-down approaches that start with a large vision of the document and split it recursively into regions of interest. Such methods are developed for example by Ha *et al.* in [6]. However, this notion of perceptive vision remains conceptual as all the analysis is realized at the same image resolution.

The idea of working at low resolution has been developed mainly for line detection. Indeed, Likforman-Sulem *et al.* in [8] observe that, at a certain distance, text lines can be seen as line-segments. The idea is also used by Déforges *et al.* in [5] that look for the resolution into which the text appears as a regular stroke. We have developed this idea in [7] for text line extraction based on Kalman filtering.

The idea of combining different resolution levels for document segmentation has been proposed by Cinque *et al.* in [2] and [1]. They propose to combine different levels of numerical data with a given mechanism. Nevertheless, their work is dedicated to newspapers and the knowledge is really linked with each resolution level, without a global vision.

In our work, we propose to combine visions from different resolutions: low resolution is interesting for extracting text line, and high resolution is interesting for details. Our particularity is to introduce this cooperation between resolutions in a generic recognition system based on a grammatical description. Consequently, this cooperation is directed by a global knowledge associated to each kind of document.

Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)
0-7695-2822-8/07 $25.00

3. Grammar based description method

In [4], we presented DMOS, a generic recognition method for structured documents. This method is made of the grammatical formalism EPF (Enhanced Position Formalism) which can be seen as a description language for structured documents. EPF makes it possible at the same time a graphical, a syntactical, or even a semantical description of a document. This grammatical description is then used to compute automatically the associated document analyser.

We first recall the basic principle of analysis and then present how we introduced a new multiresolution mechanism for document description.

3.1. Grammatical description

The document analysis is based on information that come from the image (numerical level): a list of line-segments obtained by a Kalman filter-based extractor (see [7] for more details), and a list of connected components. Those line-segments and connected components represent the terminals for the grammar.

The knowledge is then described using the EPF formalism and aims at describing the relative positions between terminals. Therefore, a grammatical description is mainly based on terminal extractors and position operators.

The *terminal operators* used to detect a connected component `Cmp` or a segment `Seg` are:

```
TERM_CMP PreCond PostCond Cmp
TERM_SEG PreCond PostCond Seg
```

They can accept pre or post conditions on the searched terminal, `Cmp` or `Seg`.

The *position operators* are defined as below: let `A` and `B` be terminals or non-terminals, and `&&` the concatenation operator.

```
A && AT(pos) && B
```

means that we have `A`, and at the position `pos` in relation to `A`, we find `B`. The writer can define as many position operators (like `pos`) as necessary.

Many other operators are available in EPF language, for example in order to deal with noises (see [3] [4]).

3.2. Introduction of multiresolution

As presented previously, the use of several image resolutions can be complementary. In order to combine those vision levels, we propose to introduce a new operator in the grammar. Our objective is to keep the possibility to change the resolution level of our analysis whenever we want in document description.

Now, for the numerical level, the analysis is based on as many {segment list, components list} couples as resolutions used. Indeed, for each used resolution, we can extract one list of line-segments and one list of connected components.

When describing the document, the analysis must begin at the lowest resolution and we have the possibility to focus on an interesting zone in order to detail it. The associated operator is:

```
A && FOCUSING ON(resol) FOR(B)
```

It means that, at a lower resolution, we have found `A`, which is an element we want to detail. Then, we will focus on resolution `resol`, relatively to `A` in order to detect `B`. `A` and `B` can be terminals or non-terminals. `A` is based on elements from the image at low resolution. `B` is based on elements from the image at resolution `resol`. The description of `B` can also be recursively based on another focus.

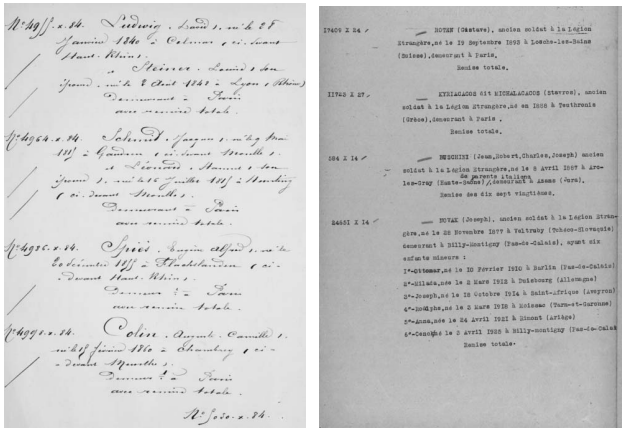
We show in section 4 an example on archive documents.

4. Multiresolution description of naturalization decrees

4.1. Naturalization decree registers

Those naturalization registers date between 1883 and 1930. They are archive documents that can be, for some people, the only way to justify their French nationality.

A decree is usually composed of about ten handwritten or typewritten pages. Each page is composed of a succession of acts in two columns: a margin and a body. For each act, the margin contains a registration number. The body contains a paragraph beginning with the name of the interested person (figure 1).



(a) 1884 handwritten page (b) 1928 typewritten page

Figure 1. Example of decree pages

These decrees are organized relatively to their date. However, inside of a decree, names are neither sorted alphabetically nor relatively to the registration number. Thus, it is very tedious to retrieve the act concerning a given person because the reader has to move through all the pages

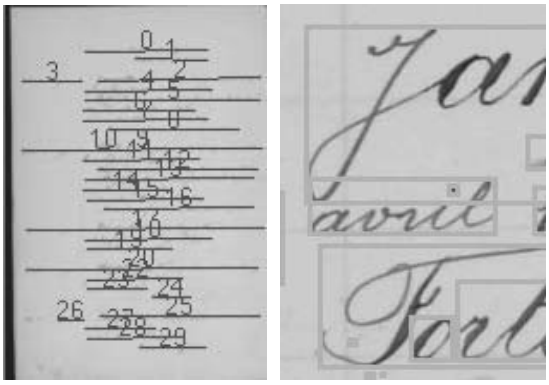
of all the decrees. So, our first goal is to extract interesting parts in the document: surnames and registration numbers. An example of final result is given figure 3(d).

Even if the documents are always based on the same logical structure, images can be very various: large sized letters in handwritten documents (figure 1(a)), tiny script on typewritten documents (figure 1(b)), first and last register pages with a slightly different structure. Our description has to be generic enough to deal with all these cases, and must be only based on structure agency without adapted threshold nor dimension. A method based on a grammatical description is well adapted for such documents.

Compared to other old forms made with rulings [4], the structure is here very weak because merely based on the organization of the text into "virtual" columns and paragraphs. In that case of low structured document, having a global view for detecting structure seems interesting, that is why we propose to use a multiresolution approach.

4.2. Cooperative multiresolution analysis

Our description is based on two resolutions: the *original* is the image at initial size (240 dpi); the *low* is obtained with a low pass filtering and under-sampling(15 dpi). Our analysis is realized on two sets of elements: line-segments extracted from low resolution (figure 2(a)), and connected components from original resolution (figure 2(b)).



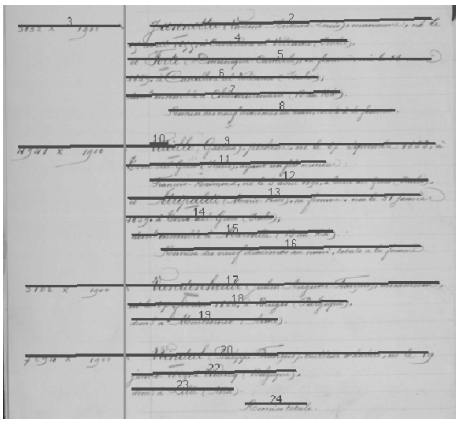
(a) Line-segments in low resolution image

(b) Connected components in original resolution image

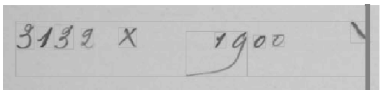
Figure 2. Basic elements

The global mechanism for each page is described below:

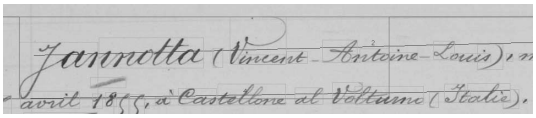
1. Margin detection at low resolution:
   (a) Detect text lines as line-segments.
   (b) Deduce the margin position (figure 3(a)).
2. Recognition of all the acts; for each act:
   (a) Focus at original resolution in the margin, find a registration number (figure 3(b)) as a succession of aligned connected components.
   (b) Go back to low resolution, find the text line as line-segment in the body, in front of the previous number.
   (c) Focus at original resolution, over the previous text line, detail connected components and compute the search surname (figure 3(c)).



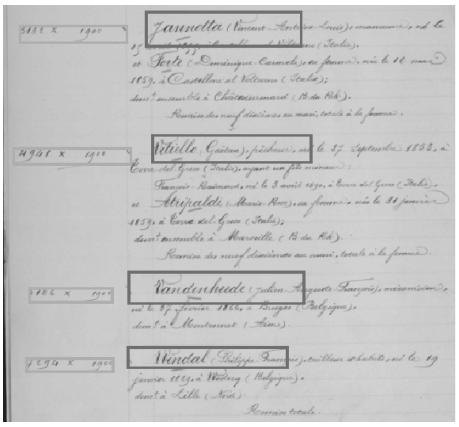
(a) Text lines at low resolution and computed margin



(b) Focusing on margin for number



(c) Focusing on text line for name



(d) Final result: numbers and names

Figure 3. Analysis mechanism

The particularity of our analysis is that we do not restrict to two successive researches at two resolution levels, but we really combine both resolutions and make possible a dialogue, guided by the knowledge, between data extracted at different levels. Thus, knowledge is really defined in relation to the searched structure and not to resolution levels.

4.3. Easiness of knowledge expression

In order to introduce the previous mechanism in the DMOS method, we have to translate it into EPF language. This can be done very easily thanks to the new multiresolution operator presented previously, and this is one of the strength of our method. Indeed, from a beginning resolution, we can focus, when necessary, on other resolutions. The local results are automatically translated, whatever their resolution, and usable later in analysis. For reading reasons, some attributes of the grammar are not presented here.

The analysis of the page consists in finding a margin and then extracting acts:

```
pageOfDecree ::=
    margin && AT(topPage) && setOfActs.
```

Finding the margin requires to recognize text lines as line-segments. The analysis begins at low resolution.

```
margin ::=
    AT(topPage) && setOfTextLine &&
    computeAverageMargin.
```

The set of text lines is extracted recursively with:

```
setOfTextLine ::=
    TERM_SEG noCond noCond FoundSeg &&
    AT(underSeg FoundSed) &&
    setOfTextLine.
```

The set of acts `setOfActs` is found recursively and each act is described as below:

```
act ::=
    AT(marginZone) &&
    FOCUSING ON(originalResolution)
        FOR(numberDetail Nb) &&
    AT(inFrontOfNumber Nb) &&
    TERM_SEG noCond noCond NameLine &&
    AT(nameLineZone NameLine) &&
    FOCUSING ON(originalResolution)
        FOR(nameDetail).
```

It is important to see that the analysis is realized at low resolution except for predicates included in the operator `FOCUSING`.

Once written, this EPF code is compiled using the DMOS method in order to produce the convenient analyser for naturalization decree pages recognition.

5. Results

5.1. Large base of documents

We applied our method on 15,699 registers, dated between 1883 and 1930, which represents 85,088 pages. Initial images were at a resolution of 240 dpi (image size around 2000*3000 pixels) and stored in JPEG. The so called *low resolution* was 15 dpi (about 120*190 pixels).

The initial images were very various as presented on figure 1 and we had to deal with all the difficulties linked with archive documents: damaged pages, presence of noise, ink from the other side visible.

We detected on this base 433,230 acts {number, surname} (5 per page on the average). 106 pages were wrongly detected empty, which represents an omission rate of 0.1%.

5.2. Results evaluation

In order to estimate more precisely our recognition rate, we set up manually three ground-truth validation bases. The *handwritten* base is composed of handwritten images only, dated between 1883 and 1884. The *various* base is composed of both handwritten and printed pages, taken at random from 47 years, with approximately the same number of images for each year. The last *representative* base is build with taking in the chronological order one image out of 250. Thus, it is representative of the ratio handwritten/printed and the different problems of the whole base, and presents the most relevant results.

Initial version without multiresolution In a previous work [4], we made a recognition system for naturalization decree registers *without* multiresolution. For example, the margin research was based on the detection of globally aligned connected components, which was sometimes vague and very sensitive to noise. Moreover, this method has been validated only on handwritten documents from 1883 and 1884. The results are given in table 1.

This previous method was adapted to handwritten documents and thus obtain good results for the *handwritten* base: 99.06% recognition. But when we applied it to the *representative* base, we only obtained a rate of 92.69% recognition. Indeed, this method could not be generic enough to deal with both handwritten and printed documents.

Base	Page/ Act number	initial version	multi resolution
Handwritten	1,999/13,896	99.06%	**98.63%**
Various	320/2,706	86.66%	**98.93%**
Representative	347/3,186	92.69%	**98.31%**

Table 1. Recognition rate on validation bases with two versions

New version using multiresolution We obtained results presented in table 1. We can see here that we really improved the recognition, mainly thanks to the generic aspect of our description that is not restrictive to handwritten documents. Thus, we obtain a recognition rate of 98.31% for the *representative* base, instead of 92.69% with the previous version. For this base, we obtain 4.33% false recognition, which is not problematic in the case of our fast leaf through application (figure 4).

5.3. Interests of multiresolution

In the case of naturalization decree registers, the introduction of multiresolution made it possible to set up a generic grammatical description. Thus, we obtain good results on the whole base, whatever pages are handwritten or typewritten. This is possible due to the ability to extract equally handwritten or printed text lines as line-segments at low resolution, with a better vision than at initial resolution. Moreover, working at low resolution decreases noise and a global structure can be easily extracted. Then, this extracted structure is a strong support for high resolution analysis.

More generally, the introduction of multiresolution gives a bigger power of expressibility for document description: it makes it possible to describe the different things you can see depending on the resolution. In our case, we only use two resolutions but we can imagine to use more levels, which is possible thanks to the generic aspect of our method.

Using EPF formalism, a multiresolution description of a new kind of document is easy to produce. In any case, the cooperation between resolutions is ordered by the operator in the grammar, that is to say guided by the knowledge.

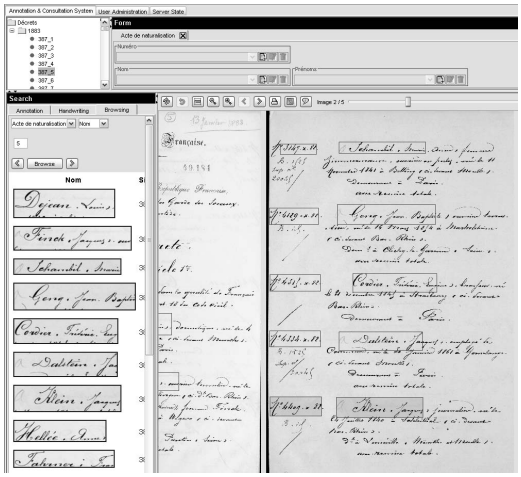


Figure 4. Consultation platform: fast leaf through by names (left); whole page (right).

6. Conclusion

We presented in this paper the introduction of a multiresolution collaboration mechanism inside a generic recognition system based on a grammatical formalism. The interest of using various levels of resolutions comes from the notion of perceptive vision: you will not see the same thing at different distances from an object.

By introducing this mechanism in a grammatical formalism, we obtain the possibility to easily give a description of any kind of document and to control resolution usage by knowledge. Furthermore, resolutions are used simultaneously and cooperatively, and not only successively.

We applied this work to the analysis of archive documents: naturalization decree registers. The system has been used for the detection of 433,230 acts in 85,088 images and validated. The use of multiresolution improved the recognition rate for our representative validation base.

An access to those results will be set up on a consultation platform ([4]) at the CHAN, *Centre Historique des Archives Nationales* (national French archives). Thanks to our extraction, the registers will be easily accessible by name or by number for a fast leaf through (figure 4). The next step will be a work on handwriting for indexing.

7. Acknowledgements

This work has been done in cooperation with the *Centre Historique des Archives Nationales* (CHAN), in France.

References

[1] L. Cinque, L. Forino, S. Levialdi, L. Lombardi, and S. L. Tanimoto. Understanding the page logical structure. In *1oth International Conference on Image Analysis and Processing (ICIAP 1999)*, pages 1003–1008, 1999.

[2] L. Cinque, L. Lombardi, and G. Manzini. A multiresolution approach for page segmentation. *Pattern Recognition Letters*, 19(2):217–225, 1998.

[3] 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 *International Conference on Document Analysis and Recognition (ICDAR)*, pages 215–220, 2001.

[4] B. Coüasnon, J. Camillerapp, and I. Leplumey. Making handwritten archives documents accessible to public with a generic system of document image analysis. In *International Conference on Document Image Analysis for Libraries (DIAL)*, pages 270–277, 2004.

[5] O. Déforges and D. Barba. A fast multiresolution text-line and non text-line structures extraction. In *International Conference on Image Processing (ICIP)*, pages 134–138, 1994.

[6] J. Ha, R. M. Haralick, and I. T. Phillips. Recursive X-Y cut using bounding boxes of connected components. In *ICDAR*, pages 952–955, 1995.

[7] 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)*, pages 38–45, 2006.

[8] L. Likforman-Sulem and C. Faure. Extracting text lines in handwritten documents by perceptual grouping. In *Advances in handwriting and drawing : a multidisciplinary approach*, pages 117–135. C. Faure, P. Keuss, G. Lorette, A. Winter, Europia, Paris, 1994.

[9] S. Mao, A. Rosenfeld, and T. Kanungo. Document structure analysis algorithms: a literature survey. In *Document Recognition and Retreval X, (Proceedings of SPIE/IST)*, volume 5010, Santa Clara, California, Jan. 2003.