Text Line Extraction in Handwritten Document with Kalman Filter Applied on Low Resolution Image

Aurélie Lemaitre
IRISA / INSA
Campus universitaire de Beaulieu
F-35042 Rennes Cedex, France
aurelie.lemaitre@irisa.fr

Jean Camillerapp
IRISA / INSA
Campus universitaire de Beaulieu
F-35042 Rennes Cedex, France
jean.camillerapp@irisa.fr

Abstract

In this paper we present a method to extract text lines in handwritten documents. Indeed, line extraction is a first interesting step in document structure recognition. Our method is based on a notion of perceptive vision: at a certain distance, text lines of documents can be seen as line segments. Therefore, we propose to detect text line using a line segment extractor on low resolution images. We present our extractor based on the theory of Kalman filtering. Our method makes it possible to deal with difficulties met in ancient damaged documents: skew, curved lines, overlapping text lines... We present results on archive documents from the 18th and 19th century.

1. Introduction

Text line extraction is an important part of document layout analysis since a large number of documents are made of lines. It represents a first step in extracting the structure of a document. Ancient handwritten documents require efficient methods. Indeed, they present the difficulty of handwritten documents: irregular spacing between letters, words, and lines, words inserted between lines, various line directions within a same page. Ancient documents are also sometimes damaged or distorted by digitalization process, which represent an important deal for analysis.

We present a method based on a notion of perceptive vision: at a certain distance, text lines of a document can be seen as line segments. The great interest of this method is to reduce details in the image during the analysis. That is why we propose to apply a line segment extractor on low resolution picture. This segmentation tool is based on Kalman filtering.

First, we shall see related work on text line extraction. Afterwards, we will present our method to extract text lines and the segmentation tool based on Kalman filtering. We will end with an application on handwritten archive documents.

2. Related work on text line extraction

Among the problem of document structure extraction, that has been widely studied, numerous methods have been proposed for text line detection. We propose to classify them according to the kind of document to which they refer.

The first and most important class, according to Kise *et al.* in [4], is document with rectangular layout. A rectangular layout means that each element can be enclosed into a rectangular shape. In that case, text lines are regular and parallel to a given direction. They can be extracted thanks to methods as projections or smearing. Both methods are examined in [8]. Indeed, projection based methods are sensitive to curvature and line skew. Smearing based methods require an appropriate threshold like inter-word gap.

In order to deal with non rectangular printed documents, Kise *et al.* propose in [4] a method based on Voronoï tessellation that can detect lines whatever their direction. However, this method is based on area or distance estimations and does not seem convenient for handwritten document with irregular distances between words and characters, as we studied for word segmentation in [6].

Recent researches have been proposed to treat handwritten documents. Nicolas *et al.* present in [10] an interesting overview. Globally speaking, methods are mainly based on ”bottom-up” approaches and consist in regrouping aligned connected components into lines. Likforman-Sulem *et al.* proposes in [9] a method based on Hough transform that can detect lines in handwritten documents. Another method is presented in [10], based on knowledge introduction to detail relation between connected components.

According to Likforman-Sulem *et al.* in [8], extracting text lines does not need any recognition of text. At a dis-

0-7695-2531-8/06 $20.00 © 2006 IEEE

tance, text appears as lines. Consequently, the authors propose a method based on the extraction of aligned connected components. However, even if this method is based on the global notion of alignments, the analysis is a local regrouping of connected components.

The problem of all these approaches based on regrouping connected components is to deal with overlapping lines. Thus, in handwritten documents, high letters sometimes overlap letters of other lines, constituting a connected component based on two or more different lines. Then, the difficulty is to choose to which line the component belongs.

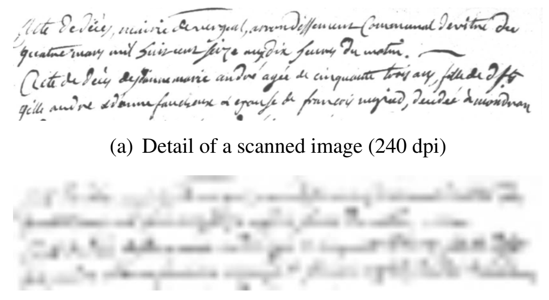
Déforges *et al.* in [3] also propose a local detection of linear objects, at low resolution, but they do not use a global vision of the picture.

Consequently, we propose an approach based on line segment detection and a global vision of document, that does not use the notion of connected component, and thus able to deal with overlapping letters.

3. Text line detection with a segment extractor

We propose to exploit the same observation as Likforman-Sulem *et al.* in [8]: at a certain distance, text lines can be seen as line segments. This idea is also used by Déforges *et al.* in [3] that looks for the resolution into which the text appears as a regular stroke.

Our method is based on this idea that working at low resolution makes it possible to extract text line as segment. It works on a global vision, without any extraction of connected components or knowledge required. Indeed, the global vision makes it possible to decrease noise around lines, which is especially convenient for handwritten documents. A visual example is proposed in figure 1. Figure 1(a) is a detail of a 240 dpi image: we can see clearly each word and letter. Figure 1(b) is the same global image at low resolution: text lines are visible as linear drawings.



(a) Detail of a scanned image (240 dpi)

(b) Detail of image at work resolution (15 dpi)

Figure 1. Example of perceptive vision

The choice of the resolution depends on the kind of documents : for each given kind of documents, we can find a convenient resolution that enables the text line extraction as segments. As we work on large document bases, digitized in the same conditions, we choose manually on few examples the good resolution and apply it for the kind of document. Examples on several kinds of document will be given in section 5.

4. Segment extraction

We present a method based on Kalman prediction and verification techniques. We propose to work on grey level image.

We first present the general context of line segment extraction. Then, we recall Kalman filtering theory and present our application for segment extraction. In these next parts, we will present only the extraction of line segments with horizontal tendency. However, the extractor is easily transposable to other direction in order to extract vertical or diagonal tendency lines.

4.1. Generalities on line segment extraction

A ”line drawing” or ”segment” is defined as a long and thin alignment of pixels. An ideal segment could be defined as a succession of connected run-lengths (a run-length is a set of connected black pixels within a column, thus approximately orthogonal with the segment direction), which have approximately the same thickness and from which the middle points of the run-lengths are on a line segment (figure 2). In our case, each run-length contains few pixels (commonly from 3 to 7).

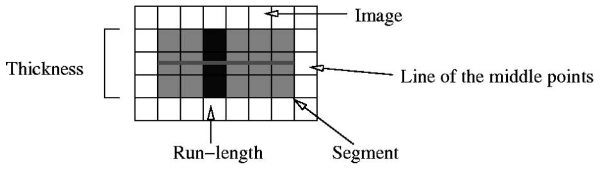
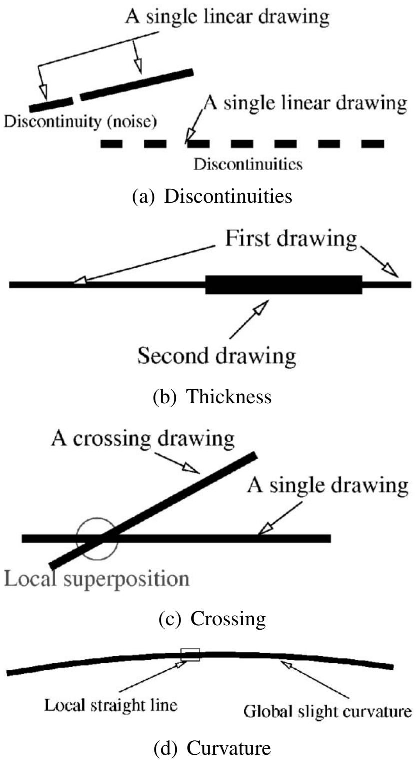


Figure 2. Example of segment

In real cases, segments take different forms. In order to show the interest of our method, we present interesting properties for a good line segment extractor.

1. Importance of the neighbourhood: an alignment of isolated points is not considered as a line segment.
2. Possible existence of discontinuities: it is useful to allow locally an absence of points, due to the quality or the nature of the extracted object (dotted line, binarization default, noise or inter-word gap in our case) (figure 3(a)).

3. Use of thickness of observation corresponding to each representative point (figure 3(b)).
4. Variable size of segments, ranging from a few representative points to several hundred.
5. Possibility of crossing segments (figure 3(c)).
6. Possibility to include a ”curvature” as a ”straight segment” although it may seem contradictory. Thus, when we observe locally a straight line which is indeed a slightly curved segment, we can decide to keep it (figure 3(d)).
7. Possibility to deal with skew, up to 30 degree.



(a) Discontinuities

(b) Thickness

(c) Crossing

(d) Curvature

Figure 3. Illustration of extractor required properties

The methods presented in section 2, like projections or Hough transform, are not suitable for line segment extraction because they do not respect all the principles. For example, Hough transform does not respect 1 or 6.

The method we propose respects all the seven previously described properties. It locally works on segment hypotheses which are confirmed or not during image analysis. The used measures do not only concern the spatial location of the drawing, but also other characteristics (thickness and/or grey level).

4.2. Kalman filtering

We present the principle of a prediction/verification approach and Kalman filter equations that are based on this principle. This approach has already been used in document processing by Leplumey *et al.* in [7] for line detection or by Lallican *et al.* in [5] for drawing extraction.

4.2.1 General principle of prediction/verification approach

The basic hypothesis is that any object can be characterised by a state e_t varying in relation to a scale, usually time. The approach allows following the evolution of the state, to estimate it, and also to predict it. That is why a recursive linear estimator is used that can predict the state e_t according to its past:

$$e_t = linear\_estimation(e_{t-1}, e_{t-2}, e_{t-3}, ..., e_0)$$

After the prediction of the state e_t, the next step is to check it by observing the current state of the object, and to correct the estimator by taking into account the error made between prediction and measured observation.

Accepting errors and correcting the estimator by taking those errors into account is one interest of this approach. However, it is necessary to know how the state of the object can vary with time.

4.2.2 Kalman filter equations

The Kalman filter is a formalization of the prediction/verification approach. It is a recursive linear estimator that is able to take into account calculated errors between estimations and measurements in order to improve future estimations.

Model definition

This method is based on two models that are commonly used in Kalman applications, see [11] for more details. We briefly recall the involved models:

- a state model S, varying with time;
- a model of measure X, related to S.

Both models S and X are represented by a vector.

The predicted state $\hat{S}$ is estimated thanks to this evolution equation:

$$\hat{S}(t) = A(t-1).S(t-1) + W(t-1)$$

where A is the evolution matrix of the vector S. W stands for the error model, characterized by a zero mean and a variance.

0-7695-2531-8/06 $20.00 © 2006 IEEE

The measure prediction $\hat{X}$ is given by:

$$\hat{X}(t) = C(t).S(t) + N(t)$$

where C is a matrix used to deduce the measure vector X from the state vector S, and where N stands for the noise measure with a zero mean and a variance. Besides, it is important to note that W and N are supposed to be uncorrelated.

We also calculate the co-variance matrix H of error prediction.

Measure integration

When a real observation $X(t)$ has been realized, the state S has to be updated considering the predicted state $\hat{S}$ and the predicted measure $\hat{X}$.

$$S(t) = \hat{S}(t) + G(t).(X(t) - \hat{X}(t))$$

The gain G balances the importance of the measure relatively to the previous state. G has to be updated after each measure. At last, the co-variance matrix H of error prediction has to be updated.

4.3. Application of Kalman filtering to segment extraction

To apply the theory presented in section 4.2, we have to determinate the kind of object that is studied, its possible states and its evolution model in the course of time. In our case, an observation is a run-length approximately orthogonal with the segment direction. The index of evolution t corresponds to a column index k.

Our method includes two levels for the treatment of observation from images. The first level contains a prediction/verification tool: the Kalman filter. The second level is a control layer permitting to interpret the expected cases (discontinuities, crossings).

4.3.1 Observation extracted from an image

We are trying to extract horizontal lines in grey level images. Those lines can be considered as dark thin objects on a clear background.

The choice of the next observation is realized among the set of run-length of the next column. The next run-length must be consistent with estimated state, according to possible error calculated in co-variance matrix H.

For the chosen observation, we finally extract the thickness of the run-length and the middle point. These parameters are based on the extraction of the run-length and they are given with sub-pixel values as we work on grey level images.

4.3.2 Used filters

The line segment extraction is realized thanks to two Kalman filters that describes the estimated values.

The first filter stands for the thickness T of the line, which is supposed to be nearly constant. The associated estimation equation is:

$$\hat{T}(k) = T(k-1)$$

The second filter represents the position Y of the line and its slope $\dot{Y}$. As we are looking for line segments, we suppose that the slope is nearly constant. Consequently, the equation is given by:

$$\begin{bmatrix} Y(k) \\ \dot{Y}(k) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} Y(k-1) \\ \dot{Y}(k-1) \end{bmatrix}$$

It means that the ordinate in column k depends on the ordinate in column $k-1$ and slope in column $k-1$. Slope is supposed to be the same in columns k and $k-1$.

Even though prediction models are supposed to be respectively a constant thickness and a constant slope, they can evolve slowly during the analysis, according to observations.

4.3.3 Interpretation

The interpretation phase consists in finding relations between successive observations. These relations are based on thickness and position. Three main cases are possible when making a new observation:

- The predicted point is dark with the correct position of the middle point and correct thickness, where ”correct” is defined thanks co-variance matrix. In that case, the observation is integrated, which means estimation state is updated, and the analysis goes on.

- The predicted point is dark with a too large thickness: the line segment is supposed to get through a larger object (figures 3(b) and 3(c)), the state is not updated but the analysis goes on, hoping to find later an other column with correct observation to integrate.

- The predicted point is white: it might be the end of the line segment but in order to deal with dotted lines (figure 3(a)); the analysis can go on for a given number of columns.

The segment detection stops when the observation has been white for a too long time.

5. Application to archive documents

We applied our method on several kinds of archive documents: registers of birth, marriages and deaths, parish registers, naturalization decree registers. This examples show various interests of our method:

- extracting both handwritten and printed text;
- extracting text in horizontal and vertical direction;
- extracting overlapping text lines;
- working on ancient damaged document;
- dealing with skew or curved lines.

We will present these aspects on concrete examples.

5.1. Example of register of birth, marriages and deaths

A first example is presented on a register of birth, marriages and deaths from 1872 (figure 4).

In this kind of document, complementary information is present in the margins, sometimes in a vertical direction. The segment extractor makes it possible to detect both horizontal and vertical tendency line segments. Moreover, text lines are overlapping, which means that a method based on grouping connected components would be difficult to apply as we explained in section 2.

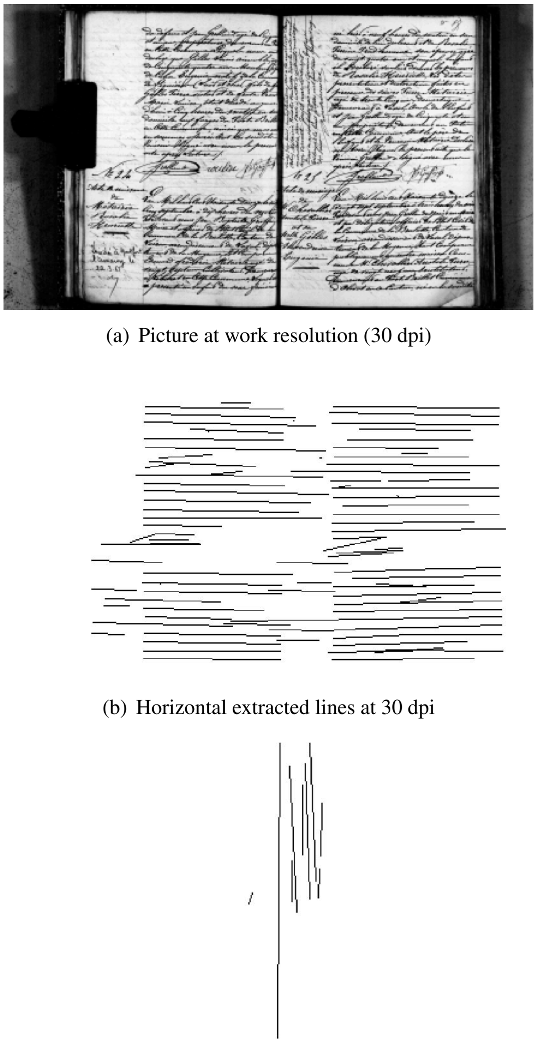
Initial images are digitalized at 240 dpi (image size around 3500*2000 pixels) and stored in JPEG. We worked at a resolution of 30 dpi (440*250 pixels). From the initial image presented in figure 4(a), we apply the segment extractor and manage to extract horizontal and vertical lines (figures 4(b) and 4(c)).

We can see that text lines are extracted even though letters sometimes overlap. Moreover, segment direction corresponds to text direction. This results will enable us to detect margins of documents and to classify marginal mentions as horizontal or vertical text.

A limit appears in that case : horizontal text lines are sometimes extended on the left, due to the presence of vertical ones. This is due to a Kalman filter parameter, that makes it possible to deal with white spaces between words. In that case, this blank is similar to the space between vertical and horizontal text. This should not be too annoying because we can work on the global set of lines to extract the margin position.

5.2. Analysis of parish register

In order to show the adaptability of the method, we extracted lines in a parish register from 1763 in the same conditions as register of births, marriages and deaths. Extraction results are given in figure 5.

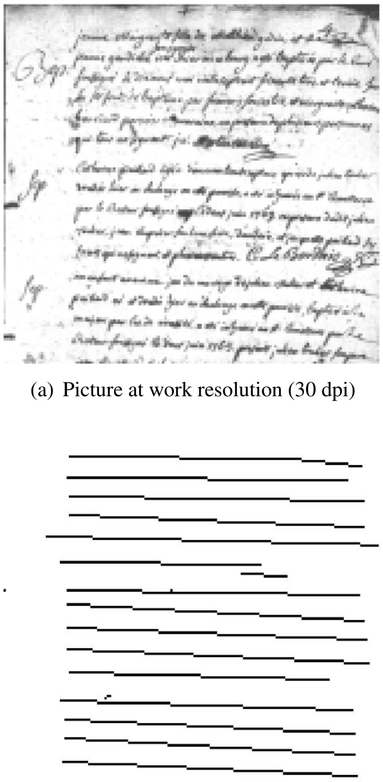


(a) Picture at work resolution (30 dpi)

(b) Horizontal extracted lines at 30 dpi

(c) Vertical extracted lines at 30 dpi

Figure 4. Line extraction in register of births, marriages and deaths

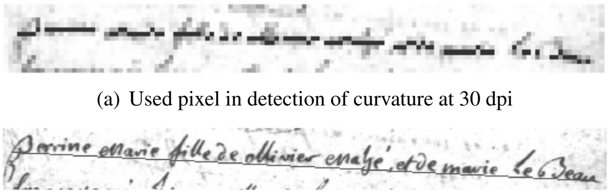


(a) Picture at work resolution (30 dpi)

(b) Extracted segments at 30 dpi

Figure 5. Line extraction in a 1763 parish register

We can notice that the method is available on ancient damaged document. Moreover, lines are detected whatever their skew and even when they are curved. A detail of extraction is presented in figure 6 to show the analysis of curvatures. Figure 6(a) presents the successive points that have been considered along the curvature. Figure 6(b) presents the final line drawn from its extremities.



(a) Used pixel in detection of curvature at 30 dpi

(b) Extracted line, taken back at 120 dpi, drawn from its extremities

Figure 6. Extraction of curved line

This result is very important for document that could be digitized with skew and for irregular handwritten text lines. Our method is able to deal with this cases whereas traditional methods based on projection for example must be adapted to treat such documents.

5.3. Analysis of naturalization decree registers

We propose an example of text line extraction on register of naturalization decree from the end of 19th century and the beginning of 20th century. In this kind of document, we are interested in extracting text lines in order to detect surnames located at the beginning of each paragraph. We can apply the same mechanism for this kind of documents, whatever it is handwritten or printed.

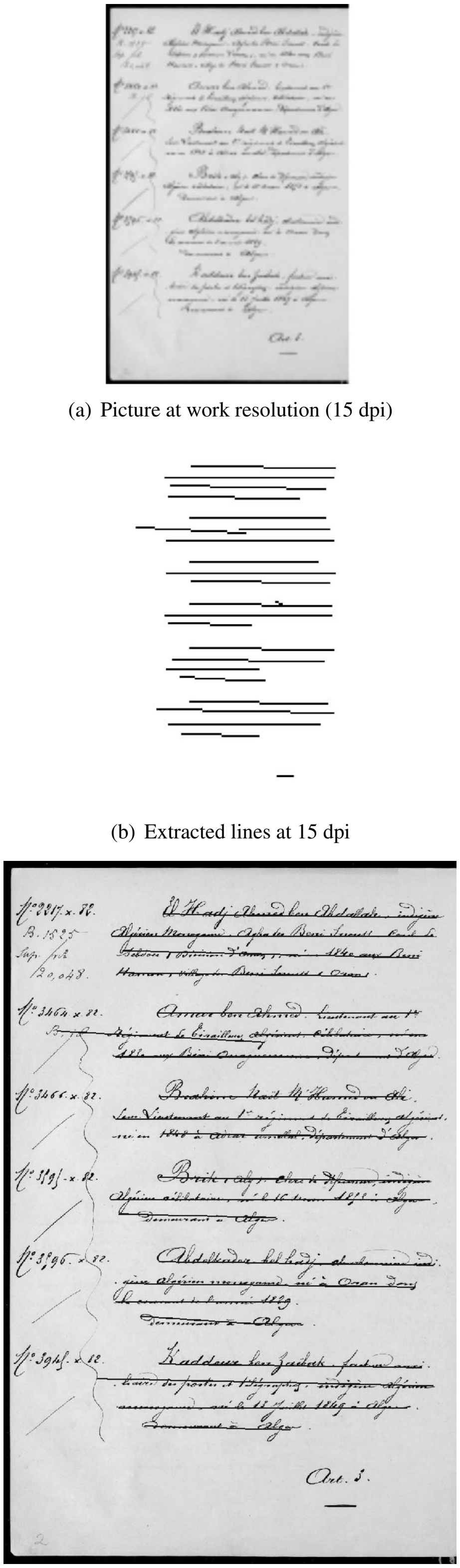
Initial images are digitalized at 240 dpi (image size around 2000*3100 pixels) and stored in JPEG; we work at a resolution of 15 dpi (120*190 pixels).

From the initial image presented in figure 7(a), we apply Kalman segment extractor and obtain lines presented in figure 7(b). For a better visualisation, we take back extracted lines at a resolution of 120 dpi in figure 7(c). The same results are possible for printed document presented in figure 8.

These results illustrate the possibility to extract text on both handwritten and printed documents. The extraction of text lines in such document will make us possible to describe paragraphs and to localize surnames.

5.4. Statistical results

We realized a first statistical study on naturalization decree registers that have been presented previously. For this

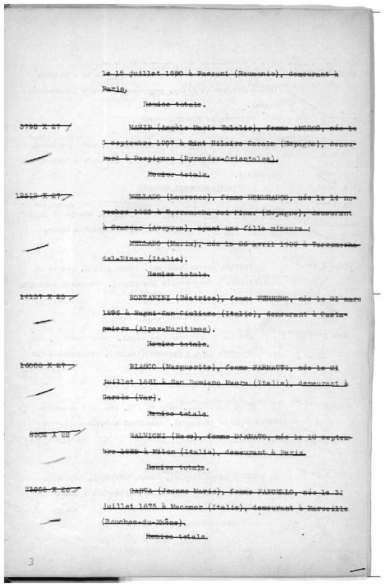


(a) Picture at work resolution (15 dpi)

(b) Extracted lines at 15 dpi

(c) Lines taken back at 120 dpi

Figure 7. Line extraction in handwritten naturalization decree register



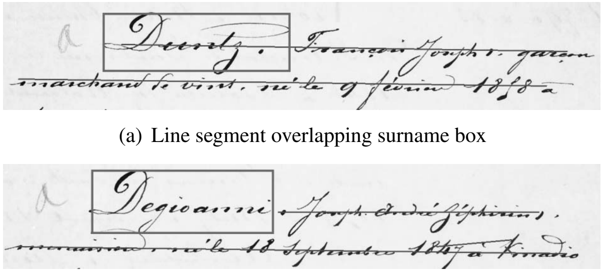
(a) Lines taken back at 120 dpi

Figure 8. Line extraction in printed naturalization decree register

document, the purpose is to detect surnames present at the beginning of text lines, in front of each registration numbers.

These document have been studied in [6]. For each one of the 1124 pages, names have been roughly detected thanks to connected components and adjusted manually to constitute a validation base.

Consequently, we studied for each name to find whether the part of the line corresponding to the surname was detected as a line segment. An example is given in figure 9(a).



(a) Line segment overlapping surname box

(b) Example of 2.9% mistake cases

Figure 9. Application on surname recognition

We studied 1124 pages of naturalization decree, which represent 3780 surnames to find. In 97.1% cases, a line segment is detected over the surname.

In remaining cases, the line segment is often detected at the end of the line, but not over the surname, due to a too large white space between surname and rest of text (figure 9(b)). However, we will be able to deal with these surnames as the rest of the line is detected.

The main interest of that study is a simplification of the document description : a text line that was previously described as a succession of globally aligned bounding boxes will now be described as a line segment in low resolution image.

We applied our method on several kinds of handwritten archive documents. The first results are really encouraging: text line extraction is one of the bases in structured document analysis and our method makes it possible to deal with various kind of documents, even noisy or damaged.

6. Future work

The next step in our work will be to treat the extracted text lines in order to analyse the structure of document: grouping lines into paragraphs, extracting parts of lines containing interesting information, determining margins position.

This analysis may be realized with DMOS method (Description and Modification of Segmentation) that has been presented in several papers [1], [2]. It makes it possible to give a grammatical description of a kind of document thanks to an appropriated formalism, EPF (Enhanced Position Formalism). Then, the associated analyser of a kind of document is automatically computed from the grammatical description. Consequently, this method is generic and has been validated on several kinds of documents: musical scores, mathematical formulae, recursive table structures, and various kinds of archive documents. It has been applied on larges bases, as presented in [2], such as 165,000 military form pages from the 19th century.

At the present time, the description of documents, like those presented in part 5, with EPF considers the relative position of connected components or line segment in high resolution image. We could introduce low resolution extracted line segments as other numerical data in DMOS analyser, which could be described as text line in document structure. The introduction of a low resolution grammar will extend the capacity of expression of our system and simplify detection.

Applications on large document bases will be realised.

7. Conclusion

The method we presented here aims at extracting text lines in documents, and most particularly in handwritten archive ones. We based our work on a perceptual vision mechanism: at a certain distance from the document, text lines appear as line segments. Consequently, we used a line segment extractor based on the theory of Kalman filter, which was used on low resolution document images.

We applied our method on several ancient archive documents. The first results are interesting and we will exploit it with DMOS method in order to simplify the grammatical description of documents.

References

[1] 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.

[2] 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.

[3] 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.

[4] K. Kise, M. Iwata, A. Dengel, and K. Matsumoto. Text-line extraction as selection of paths in the neighbor graph. In *Document Analysis Systems*, pages 225–239, 1998.

[5] P. M. Lallican and C. Viard-Gaudin. A Kalman approach for stroke order recovering from off-line handwriting. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 519–523, 1997.

[6] A. Lemaitre, B. Coüasnon, and I. Leplumey. Using a neighbourhood graph based on Vorono tessellation with DMOS, a generic method for structured document recognition. In *Proceedings of GREC, Sixth IAPR International Workshop on Graphics Recognition*, pages 260–271, Hong Kong, China, August 2005.

[7] I. Leplumey, J. Camillerapp, and C. Queguiner. Kalman filter contributions towards document segmentation. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 765–769, 1995.

[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] L. Likforman-Sulem, A. Hanimyan, and C. Faure. A Hough based algorithm for extracting text lines in handwritten documents. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 774–777, 1995.

[10] S. Nicolas, T. Paquet, and L. Heutte. Text line segmentation in handwritten document using a production system. In *9th International Workshop on Frontiers in Handwriting Recongnition (IWFHR)*, pages 245–250. IAPR, 2004.

[11] H. W. Sorenson. *Kalman Filtering: Theory and application*. IEEE Press, New York, 1985.