Handwritten Document Segmentation Using Hidden Markov Random Fields

Stéphane Nicolas, Yousri Kessentini, Thierry Paquet, Laurent Heutte
Laboratoire PSI FRE CNRS 2645 - Université de Rouen
Place E. Blondel, UFR des Sciences et Techniques
F-76821 Mont-Saint-Aignan cedex, France
{Stephane.Nicolas, Yousri.Kessentini,Thierry.Paquet, Laurent.Heutte}@univ-rouen.fr

Abstract

In this paper we present a method based on Hidden Markov Random Fields and 2D dynamic programming image decoding, for segmenting pages of complex handwritten manuscripts such as novelist drafts. After a formal description of the theoretical framework and the principles of the decoding method, we describe the implementation of the model and the decoding method. Then we discuss the results obtained with this approach on the drafts of the French novelist Gustave Flaubert.

1. Introduction

These last years improvements have been made in the field of handwriting recognition, especially in the context of industrial applications such as check reading, postal address recognition, form processing,... and have been mainly focused on word or phrase recognition [1]. However with the advance of digital technologies, numerous institutions are moving towards the use of digital documents images rather than traditionnal paper copies of the original documents. This situation raises new needs for indexing and accessing to these numerical sources [2].
In this context of shared access to our cultural and historical heritage, the Bovary Project, a digitization program of the manuscripts of the famous novel "Madame Bovary" of Gustave FLAUBERT, aims at providing a numerical web edition[1] of the genesis of this novel, to browse the original manuscripts of Flaubert associated with diplomatic textual transcriptions respecting as much as possible the layout of the original manuscripts. Such a numerical edition will be of great interest for researcher in literary.
However the production of the textual transcriptions of the 4127 manuscripts that constitute the Bovary directory is a challenging task. Considering the state of the art of document image analysis techniques, as well as the extreme variability of Flaubert's drafts, full automation of the process cannot be envisaged.

[1] http://www.univ-rouen.fr/psi/BOVARY

For this reason a network of volunteers has been recruited. However, it is assumed that their work could be greatly facilitated thanks to the use of automatic document analysis techniques. This is why we investigate the spectrum of such methods with the aim to apply on archived handwritten documents. The aim is not to recognize a full page of handwriting but to identify the regions of interest by extracting the layout of the manuscripts.
A lot of methods for machine printed document segmentation have been proposed [3], but these methods cannot be directly applied to handwritten documents because of the spatial variability of handwriting. The few existing methods dedicated to handwritten documents focus on a particular type of documents or a particular task of segmentation (word or line extraction only). Furthermore these methods are based on a local analysis, and sometimes fail to find the good solution. It is the reason why we propose to use a general formalism that could be adapted to different types of documents, and which takes into account some contextual information. Hidden Markov Random Field formalism has been retained for this purpose. It is used to proceed to the segmentation of Flaubert's manuscripts into their elementary parts, namely: text lines, erasures, ponctuation marks, inter-linear annotations, marginal annotations (just to mention the most important of them).
The paper is thus organized as follows. We first present in section 2 the theoretical framework of Hidden Markov Random Fields, and the image decoding method. In section 3 we explain our choices for an implementation of the formalism dedicated to Flaubert manuscript segmentation. The preliminary results obtained with this method are discussed in section 4.

2. Theoretical framework

Each document image is considered to be produced by implicit layout rules used by the author. While these rules cannot be formaly justified, it is however experimentaly verified by literacy experts that Flaubert's manuscripts exhibit some typical layout rules characterized by a an important text body

occupying two thirds of the page and containing a lot of erasures; and a marginal area with some text annotations as it can be seen in figure 1.

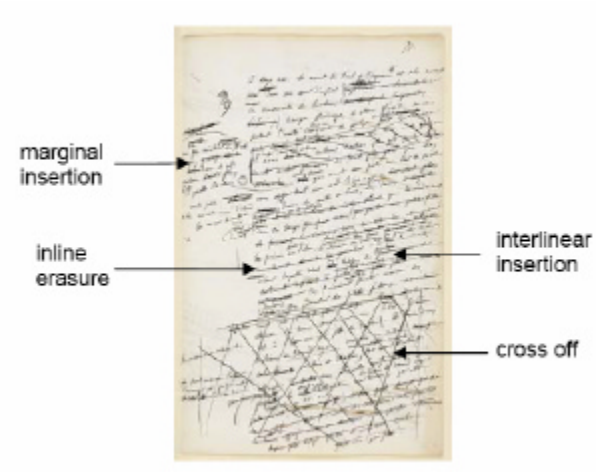


Figure 1. One example of Flaubert's manuscript layout.

As there exist some local interactions between these layout rules, a Markov Random Field (MRF) seems to be adapted to model the layout of a manuscript.

According to MRF formalism [4], the image is associated with a rectangular grid G of size $n \times m$. Each site on the grid is defined by its coordinates over G and is denoted $g(i,j), \quad 1 \le i \le n \quad 1 \le j \le m$.

Following the stochastic framework of Hidden Markov Random Fields, the image gives access to a set of observations on each site of the grid G denoted by $O = \{o(i,j), \quad 1 \le i \le n \quad 1 \le j \le m\}$. Furthermore, considering that each state of the Markov Field is associated to a particular layout rule, the problem of layout extraction in the image can be formulated as that of finding among all the possible state configurations X that can be associated to the image, the most probable according to the model, i.e. finding:

$$\hat{X} = \arg\max_{X \in E}\left(P(X,O)\right) = \arg\max_{X \in E}\left(P(O|X)P(X)\right)$$

which results in the following formula when applying Markovian hypothesis and independence assumption of observations.

Equation 1:

$$\hat{X} = \arg\max_{X \in E}\left(\prod_g P(o_g|x_g)\prod_g P(x_g|x_h, h \in N_G(g))\right)$$

While in this expression the term $\prod_g P(o_g|x_g)$ can be computed using Gaussian mixtures to modelize the conditional probability densities of the observations, the calculation of the second term (i.e. $\prod_g P(x_g|x_h, h \in N_G(g))$), which represents the contextual knowledge introduced by the model, appears to be intractable due to its non causal expression i.e. interdependance between neighboring states. To overcome this difficulty, one generally uses simulation methods such as Gibbs sampler or Metropolis algorithm [5]. Another possibility is to restrict the expression to a causal neighboring system. In any case however, finding the optimal segmentation solution requires a huge exploration of the configuration set E. This consideration is especially important because handwritten document images are particularly large. For this reason, we are currently using one intermediate suboptimal strategy based on the principle of the dynamic programming that exploits efficiently the grid structure of random fields [6].

This method relies on the fact that according to the Hammersley-Clifford theorem, a MRF is equivalent to a Gibbs distribution [4], so that the second term of equation 1 can be rewritten as follows:

$$P(X) = \frac{1}{Z}\exp\left(-\sum_{c \in C} V_c(X)\right)$$

where C is the clique set defined according to the choosen neighboring system, V_c the potential of the clique c and Z a normalization term. This allows to introduce the potential function U(X) of a configuration of the field, by calculating the logarithm:

$$U(X) = \sum_g -\log\left(P(o_g|x_g)\right) + \sum_{c \in C} V_c(X)$$

Thus, decoding becomes a minimization problem of the potential function of the label configuration:

$$\hat{X} = \arg\min_{X \in E} U(X)$$

The algorithm proposed in [6] allows to solve this problem efficiently, by recursively merging the N best configurations of each different regions (subset) of the label field X using dynamic programming.

Assume that two neighboring rectanglar regions O_1 and O_2 are associated to their respective state configuration $X_1(i,j)$ and $X_2(i,j)$. Then, the joint probability of region $O=O_1 U O_2$ and its associated to the state configuration $X(u,v)$ defined by:

$$X(u,v) = \begin{cases} X_1(i,j) \ \textit{if} \ (u,v)=(i,j) \\ X_2(k,l) \ \textit{if} \ (u,v)=(k,l) \end{cases}$$

can be derived as follows:

$P(X,O) = P(X_1,O_1)P(X_2,O_2)I(X_1,X_2)$

where the expression

$$I(X_1, X_2) = \prod_{g \in G_1, h \in G_2} P(x_g|x_h, h \in N_{G_1}(g))$$

denotes the interactions between the two state configurations. These interactions are evaluated on the sites at the frontiers of the two regions by the horizontal cliques. As a consequence, a particular state configuration can be evaluated for the entire image by iteratively merging neigboring regions and evaluating the interaction term at the frontiers of the two regions.
This simple principle allows to evaluate the probability of a particular configuration associated to one image block. However, we are looking for the optimal configuration that can be associated to a document image. Ideally, this would require to compute all the possible configurations associated to each region when proceeding to a merge, and to retain the optimal configuration at the end of the process. But this optimal strategy becomes rapidly intractable as soon as the size of the image exceeds a small size. For this reason we have called upon a sub-optimal strategy that takes into account only the N best configurations when proceeding to a merge of two regions, as suggested in [6].
Various region merging strategies could be used. The one we have retained is to start with all the single sites and then to merge regions 2 by 2 until the whole image is covered. We are currently using an alternate strategy that consists in merging regions horizontally and vertically successively, in order to not support a particular direction.

3. Application to handwritten document segmentation

As explained in [6], the 2D dynamic programming algorithm is general enough to be applied to different recognition and segmentation problems. One has simply to make some choices concerning the modelling of the probability density function of observation emission, the clique potentials and the merging strategy.

- Probability densities

The probability densities are modeled by gaussian mixtures. The parameters of the mixtures are learnt on manually labelled images, using the EM algorithm. The number of gaussians is determined automatically using the Rissanen criterion. We use Bouman's CLUSTER software[2] to learn the Gaussian Mixtures.

- Clique potentials

We consider the second order cliques associated to a 4-connected neighboring:

$$C = C_1 \cup C_2 \cup C_3$$

where

$$C_1 = \{(i, j), 1 \le i \le n, 1 \le j \le m\}$$
$$C_2 = \{((i, j), (i+1, j)), 1 \le i \le n, 1 \le j \le m\}$$
$$C_3 = \{((i, j), (i, j+1)), 1 \le i \le n, 1 \le j \le m\}$$

The interaction terms are defined as mutual information terms taking into account the only the horizontal and vertical directions (4-connectivity):

$$I_H = \frac{P(w_k | w_l)}{P(w_k)P(w_l)} \qquad I_V = \frac{P\left(\frac{w_k}{w_l}\right)}{P(w_k)P(wl)}$$

where

$P(w_k | w_l) = P(w_{(i,j)} = w_k, w_{(i+1,j)} = w_l)$ and

$P\left(\frac{w_k}{w_l}\right) = P(w_{(i,j)} = w_k, w_{(i,j+1)} = w_l)$

As for the gaussian mixture parameters, these probabilities are learnt on few labeled examples, by counting the frequency of each possible transition. If a rule transition doesn't appear in the learning examples, its probability is not set to zero but to a very low value, making it not impossible but very unlikely.
Finally, the clique potentials are defined as follows:

$$V_c(w) = \begin{cases} -\log(P(w_k)) & \text{if } c \in C_1 \\ -\log(I_H(w_k, w_l)) & \text{if } c \in C_2 \\ -\log(I_V(w_k, w_l)) & \text{if } c \in C_3 \end{cases}$$

- Observations

Observations are features that are extracted on each site g(i,j) of the grid G applied on the image. As we work on binary images, we have choosen to extract for each g(i,j) a bi-scale feature vector based on pixel density measurement. This vector contains 18 features. The first 9 are the density of black pixels in the current site g(i,j) and its 8-connected neighbors at the first scale level. Based on the same principle, the remaining 9 features are the density of black pixels extracted at the second scale level (see figure 2). Note that the size of the site g(i,j) must be adapted to the size of the objects we want to extract. On our images, we are currently using a 5x5 pixels site which has been found to be a good compromise both for extracting small objects like diacritics for example but also for coping with computational issues.

[2] http://dynamo.ecn.purdue.edu/~bouman/software/cluster

Figure 2. Multiresolution pixel density feature extraction

4. Preliminary results and discussion

The analysis of the results of a document image segmentation algorithm is a difficult and not always well defined task, since there exist very few protocols and image databases for performance evaluation [7]. The few existing ones are only designed for machine printed documents for which the proposed methodologies and metrics used to compare the algorithms are dedicated to well defined classes of methods or documents (newspaper, mail, form, postal address). To the best of our knowledge, there do not exist such methodologies and metrics in the fied of handwritten documents or historical documents. Consequently, the results we present here are preliminary qualitative results obtained on few images of full page of handwriting or parts of pages from the Bovary database. As we cannot at present provide quantitative results in terms of correct segmentation rates, we discuss only the results obtained on our test images. Nevertheless the results obtained on complex manuscripts such as those of Flaubert are very encouraging and tend to prove the potential of our method.

Let us recall that Flaubert's manuscripts contain a lot of deletions and crossed out words or lines (see figure 3).Therefore, in a first experiment, we have tried to evaluate the capabilities of our method on a specific task which consists in separating words (or parts of words) and deletions. For this purpose, we have defined a model made up of 4 states: "pseudo-word", "deletion", "diacritic" and "background". Figure 4 presents the results obtained with different models on a page fragment. Figure 5.a. shows a zoom on a deletion area where word and deletion strokes are completely connected. One can see on this result that the deletion lines are well separated from the strokes below. This result highlights the superiority of this method on the approaches working at the connected component level. Indeed, the fact of working at the pixel level allows us to segment different objects which are connected together. Figure 5.b. shows similar results on a fragment containing a word and an erasure connected by a descending loop. Both components are well separated.

In a second experiment, we have introduced an additionnal "inter pseudo-word space" state to the previous model. Firstly this enables us to study the behavior of the method when states are added to the model, and secondly the addition of this state makes it possible thereafter to extract the text lines because one can define a text line as a sequence of "pseudo-words" separated by "inter-word space". Thus from the results returned by the method, it is possible to extract text lines or other objects of higher level (such as text blocks for example), by applying label merging rules. Globally the results are promising, the inter-word spaces are well segmented (see figure 5.b).

Finally in the same way, we have defined a third model with 6 states by adding an "interlines" state to the previous model, in order to model also the interlinear spacings. The knowledge of interlines allows to better segment text lines, and to detect text blocks. The result obtained with this model on the same page fragment is shown on figure 5.c and the result obtained on a full page is shown on figure 4.

For these three models the results are globally satisfactory. However if we look locally at the results, we can see that some pixels are misclassified. One has to keep in mind that the 2D dynamic progamming algorithm with pruning procedure is a sub-optimal decoding algorithm. It means that the final segmentation obtained is not the optimal one. Some configurations of the label field can be locally less probable and thus be pruned during the merging procedure, whereas they could be globally the optimal ones. If the size of the image is large and if there are a lot of states in the model, the number of possible configurations of the label field is very large. In this case, it is not possible to store all the possible intermediate solutions, so the pruning threshold should not be too high. On the other hand, if this threshold is too low, the final configuration retained may be one of the least probable ones (because involving not probable transitions during the region merging). The choice of the merging strategy is important for the final segmentation result, but we think that the choice of features extracted on the observations is important too. Future works will concern these two open issues.

6. Conclusion

Despite these encouraging results, further developments are required in order to assess the method more significantly. They may concern the definition of the feature set, and the influence of the various parameters of the approach. Another difficulty of the approach is related to the desing of a feature vector able to describe each of the possible objects we want to label and to cope with size or scale efects. For this purpose, a multi-scale strategy would be probably of interest. In anay case however, this formalism is general enough to be adapted to different type of documents and layouts.

7. References

[1] H. Bunke, Recognition of Cursive Roman Handwriting, Past, Present and Future, Proceeding of the seventh International Conference on Document Analysis and Recognition (ICDAR'03), Endinburgh, 2003.

[2] H. Baird, Digital Libraries and Document Image Analysis, Proceeding of the seventh International Conference on Document Analysis and Recognition (ICDAR'03), Endinburgh, 2003.

[3] Nagy, G, Twenty years of document image analysis in pamiTwenty Years, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 22, n°1, 2000.

[4] Chellappa, R. et Jain, A., editors (1993). Markov Random Fields - Theory and application. Academic Press.

[5] S. Geman, D. Geman, Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 6, 1984.

[6] E. Geoffrois, Multi-dimensional Dynamic Programming for statistical image segmentation and recognition, International Conference on Image and Signal Processing, 2003.

[7] "Performance Evaluation: Theory, Practice, and Impact", T. Kanungo, H. S. Baird, R.M. Haralick, Guest Editors, special issue of International Journal on Document Analysis and Recognition, vol. 4, n°3, march 2002.

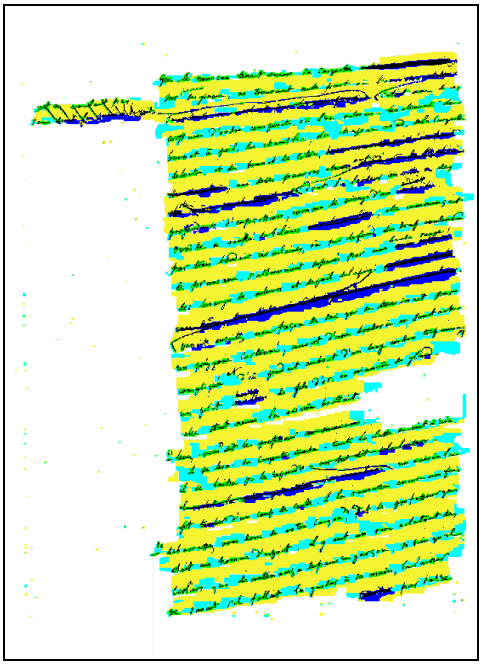
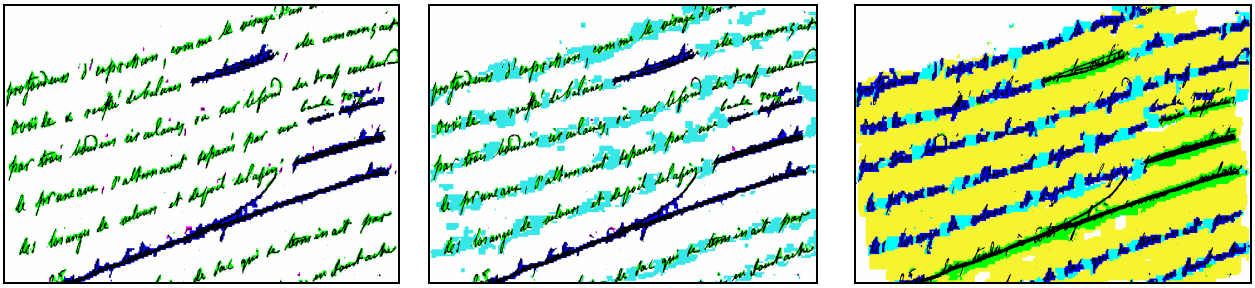
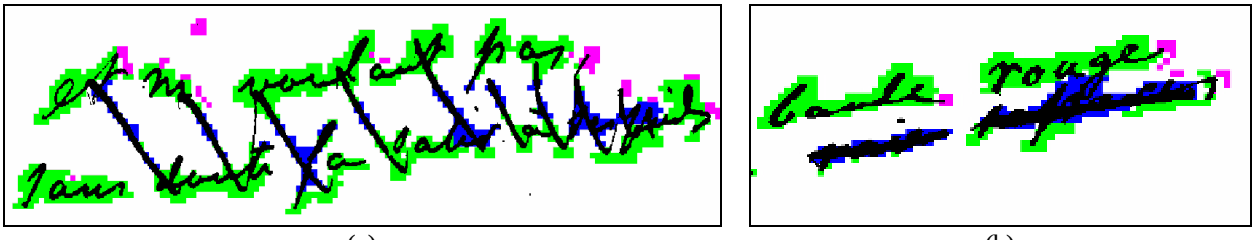


Figure 3. Segmentation results obtained on a complete Flaubert manuscript page (img1) using a 6-state model with the following color/label convention: white = background, green = textual component, blue = erasure, pink = diacritic, cyan = interwords spacing, yellow = interline.



(a) (b) (c)

Figure 4. Segmentation results obtained on a page fragment: (a) using a 4-state model; (b) using a 5-state model; (c) using a 6-state model, with the following color/label convention: white = background, green = textual component, blue = erasure, pink = diacritic, cyan = interwords spacing, yellow = interline.



(a) (b)

Figure 5. Segmentation results obtained on some complex page fragments using the 4-state model.