A Markovian Approach for Handwritten Document Segmentation

Stéphane Nicolas Thierry Paquet Laurent Heutte
Laboratoire LITIS – Université de Rouen
UFR des Sciences et Techniques
Avenue de l'Université, Technopôle du Madrillet
76801 Saint Etienne du Rouvray cedex, FRANCE
{Stephane.Nicolas, Thierry.Paquet, Laurent.Heutte}@univ-rouen.fr

Abstract

We address in this paper the problem of segmenting complex handritten pages such as novelist drafts or authorial manuscripts. We propose to use stochastic and contextual models in order to cope with local spatial variability, and to take into account some prior knowledge about the global structure of the document image. The models we propose to use are Markov Random Field models.

1. Introduction

In a document image analysis process, the segmentation is an important task because it is the process that allows to locate and to extract the entities to be recognized. 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,... These applications have been mainly focused on word or phrase recognition [1], but unconstrained recognition of full handritten pages is still a challenging task. However with the advance of digital technologies, numerous institutions are moving towards the use of digital document images rather than traditionnal paper copies of the original documents. This situation raises new needs for indexing and accessing to these numerical sources [2].

A lot of methods for segmenting machine printed documents 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, without taking context into account and then sometimes fail to find the good solution. However it is well known that context can help to disambiguate some complex interpretations. Even if handwritten documents are less structured than printed ones, the segmentation process can benefit from the use of prior knowledge about the global structure of the document, and contextual information. Due to the local variability of handwritten documents, a formal description of the layout is not possible. Stochastic models are well adapted to cope with ambiguities. Markov models are usually used for sequential data segmentation and recognition. In the case of images, Markov Random Fields (MRF) are powerful stochastic models of contextual interactions for bidimensional data. MRF framework has been widely studied these last decades [3], and MRF models were applied to various tasks of image analysis [4], but to the best of our knowledge, never to handwritten documents. We propose to use Markov Random Fields to segment complex handwritten documents, such as authorial drafts or historical documents, and we present here an application consisting in the segmentation of manuscripts of the french writer Gustave Flaubert into their elementary parts, namely: text lines, erasures, ponctuation marks, inter-linear annotations, marginal annotations (just to mention the most important of them). In the MRF framework, segmentation is addressed as an image labelling problem. This problem can be solved using optimization techniques. In section 2 we describe the theoretical framework of MRF, then in section 3 we present our implementation for authorial manuscripts segmentation and we discuss the obtained results 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 formally justified, it is however experimentally verified by literature experts that Flaubert's manuscripts exhibit some typical layout rules characterized by a large 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.

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

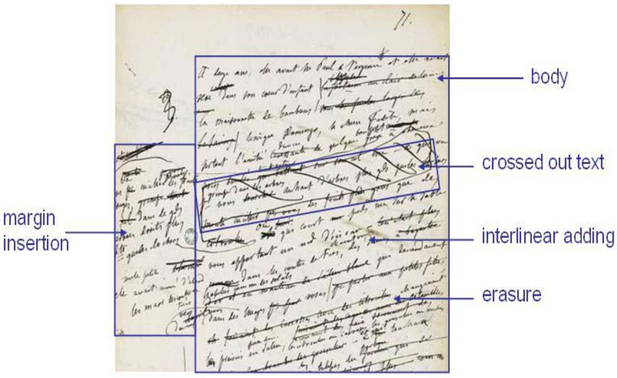

Figure 1. One example of Flaubert's manuscript layout.

Furthermore we deal with handwritten documents characterized by some local spatial variability in the layout, so a stochastic model is suited to cope with uncertainties in the disposal of layout elements.

According to MRF formalism [5], the image is associated with a rectangular grid G of size $n \times m$. Each image site s is associated to a cell on the grid defined by its coordinates over G and denoted $g(i,j)$, $1 \leq i \leq n$ $1 \leq j \leq m$. The site set is denoted $S = \{s\}$. Following the stochastic framework of Hidden Markov Random Fields, the image gives access to a set of observations denoted by $O=\{o(i,j), 1\leq i\leq n \quad 1\leq j\leq m\}$, on each site of the grid G. Furthermore, considering that each state X_s of the Markov Field X is associated to a label l corresponding to a particular layout rule or pattern class, the problem of layout extraction in the image can be formulated as that of finding among all the possible labellings or state configurations E of the field X that can be associated to the image, the most probable according to the model, i.e. finding:

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

which results in the following formula when applying Markovian hypothesis and independence assumption of observations: $\hat{X} = \arg\max_{X\in E}\left(\prod_g P(o_g|x_g)\prod_g P(x_g|x_h, h \in N_G(g))\right)$

where is $N_G(g)$ the neighborhood of the site g

While in this expression the term $\prod_g P(o_g|x_g)$ can be computed using Gaussian mixtures to model 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 sampling or Metropolis algorithm [6]. 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. According to the Hammersley-Clifford theorem, a MRF is equivalent to a Gibbs distribution [5], so that the prior model $P(X)$, can be rewritten as follows: $P(X) = \frac{1}{Z}\exp\left(-\sum_{c\in C} V_c(X)\right)$

where C is the set of all cliques over the image, defined according to the choosen neighboring system $N = \{N_S, s \in S\}$. A clique is a set of interdependent sites. V_c is a potential function associated to the clique c and Z is a normalization constant called partitionning function in the context of MRF framework. This allows to introduce the joint energy $U(X,O)$ of a configuration of the field, by calculating the negative logarithm of the joint probability: $U(X,O) = \sum_g -\log\big(P(o_g|x_g)\big) + \sum_{c\in C} V_c(X)$

Thus in the MRF-MAP framework, decoding or image labelling involves minimizing the joint energy function:

$$\hat{x} = \arg\min_x U(X,O)$$

It is a non trivial combinatorial problem, because the energy function may be non convex and exhibits many local minima. Various optimization techniques can be used to find the optimal configuration of the label field by minimizing the energy function [4].

3. Application of MRF labeling to handwritten document segmentation

When using MRF-MAP labelling framework to segment images, one has simply to make some choices concerning the modelling of the probability density function of observation emission, the clique potential function and the optimization method used to minimize the energy function. In this work we are interested in the segmentation of handwritten documents, such as drafts or authorial manuscripts, into their elementary parts using a prior MRF model. We describe here our implementation choices to solve this task.

- Probability densities

The probability densities are modelled 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[1] to learn the number of gaussian components and mixture parameters.

- Clique potential functions

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

$$
\begin{aligned}
&C = C_1 \cup C_2 \cup C_3\\
&\textit{where}\\
&C_1 = \{(i,j), 1 \le i \le n, 1 \le j \le m\}\\
&C_2 = \{((i,j),(i+1,j)), 1 \le i \le n, 1 \le j \le m\}\\
&\mathrm{C}_3 = \{((i,j),(i,j+1)), 1 \le i \le n, 1 \le j \le m\}
\end{aligned}
$$

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

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

where

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

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

As for the gaussian mixture parameters, these probabilities are learnt on few labelled examples, by counting the frequency of each possible clique configuration. If a configuration does not appear in the training examples, its probability is not set to zero but to a very low value, making it not impossible but very unlikely. Finally, the clique potential functions are defined as follows:

$$
V_c(w) = \begin{cases} -\log(P(w_k)) & \text{if } \mathrm{c} \in \mathrm{C}_1 \\ -\log(I_H(w_k, w_l)) & \text{if } c \in C_2 \\ -\log(I_V(w_k, w_l)) & \text{if } c \in C_3 \end{cases}
$$

In a similar way, according to these definitions, the use of 2-order cliques with 8-connected neighboring is very simple. One has only to take into account diagonal interactions too.

- Observations

Observations are features that are extracted on each site s at the position $g(i,j)$ on the grid G applied to the image. As we work on binary images, we have choosen to extract for each site s a bi-scale feature vector based on black pixel density measurement. This vector contains 18 features. The first 9 are the density of black pixels in the cell g(i,j) associated to the current site, 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 coarser scale level. Each cell at this scale corresponds to a 3×3 window at the previous scale. Note that the size of the cell $g(i,j)$ on the grid G must be adapted to the size of the smallest objects or layout elements we want to extract in the image. The choice of this size is necessarily the result of a compromise between the segmentation quality and the computationnal efforts. The smaller the cells are, the more labelling is fine, but more there will be sites, so more complicated will be the energy minimization process. On our images, depending on the considered segmentation task, we are using different cell sizes.

- Decoding strategy

To proceed to the decoding of the image by means of minimization of the energy function, we have implemented several of the methods described in the literature, mainly ICM, HCF, and 2D dynamic programming [7]. We have tested and compared these methods. The results are provided in the next section.

4. Results

First we present qualitative results obtained on some manuscripts of Gustave Flaubert and then we provide quantitative results obtained with various decoding strategies.

4.1 Qualitative results

Let us recall that Flaubert's manuscripts contain a lot of deletions and crossed out words or lines (see figure 1). 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" (Fig. 2.a.). Then we have defined a 5-state model by adding an "interword" state to the previous model (b), and finally a 6-state model by adding an "interline" state in order to model also the interlinear spacings (c). The knowledge of interlines allows to better isolate text lines, and to detect text blocks. The result obtained on a full page is shown on figure 2.d. Figure 2.e. 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, working at the pixel level allows us to segment different objects which are connected together. Figure 2.f. shows similar results on a fragment containing a word and an erasure connected by a descending loop. Both components are well separated. Note that our method simply label image areas at a pixel level, but no directly segment object of higher levels of abstraction such as text lines or blocks. However using the result of the labelling, these entities can be segmented by means of label

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

merging and connected components extraction, but we did not implement this part yet.

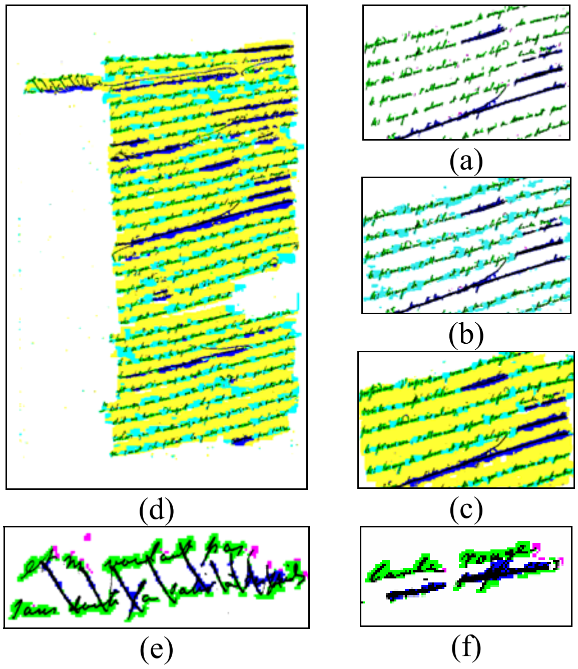

Figure 2. Labeling results

4.2 Quantitative results

In order to evaluate precisely the performance of our approach and to compare the decoding methods, we have considered an easier segmentation task allowing to easily and quickly label a database of Flaubert manuscript images manually for model learning and groundtruthing. We want to detect the main areas of the manuscripts such as text body, margins, header, footer, page number, and marginal annotations. Our model contains 6 labels. The database contains 69 manuscript images at 300 dpi. The average dimensions of the images are 2400×3700. All the images of the database were binarized and manually labelled according to the defined 6 labels. The database was divided into 3 parts: one for the learning of the model parameters (parameters of the gaussian mixtures, clique potential functions), an other for model setting, and the last one for testing. We use a regular and rectangular grid whose dimensions of each cell are 50×50 pixels. We compare the following decoding methods with the labelling results provided by the Gaussian Mixture model only: ICM, HCF, 2D dynamic programming (2D DP). For each decoding method we determine on the test database the global labelling rate (GLR) by counting the number of well-labelled sites and the normalized labelling rate (NLR) by counting the average number of well-labelled sites by label class. These results show that the use of a MRF model allows to increase the normalized labelling rate and that the HCF algorithm outperforms the other decoding methods. Furthermore HCF algorithm is faster than 2D dynamic programming method.

	local classifier	ICM	HCF	2D DP
GLR (%)	88,0	86,6	90,3	84,6
NLR (%)	83,7	87,5	88,2	87,4
time (s)	-	0,21	0,29	0,61

Tab 1. Labeling rates obtained with different decoding methods and decoding time

5. Conclusion

In this paper we have proposed to use Markov Random Field models to segment complex handwritten manuscripts into their elementary parts, such as text body, margins, header, footer, page numbers, deletions, ... by means of image labelling using various optimization techniques such as ICM, HCF and 2D dynamic programming. The proposed approach provides interesting results. The main advantages are the ability of Markov Random Fields to deal with local variability, to model prior knowledge and the learning possibilities which allow a possible adaptation to various types of documents, including machine-printed ones, on the condition of having some manually labelled examples and using eventually more adapted image features.

6. 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), Edinburgh, pp 448-459, 2003.

[2] H. Baird, Digital Libraries and Document Image Analysis, Proceeding of the seventh International Conference on Document Analysis and Recognition (ICDAR'03), Edinburgh, pp 2-14, 2003.

[3] Nagy, G, Twenty years of document image analysis in PAMI, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 22, n°1, pp 38-62, 2000.

[4] S.Z. Li, "Markov Random Field Modeling in Computer Vision", Springer, Tokyo, 1995.

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

[6] S. Geman, D. Geman, Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 6, pp 721-741, 1984.

[7] E. Geoffrois, Multi-dimensional Dynamic Programming for statistical image segmentation and recognition, International Conference on Image and Signal Processing, pp 397-403, 2003.