Markov Random Field Models to Extract The Layout of Complex Handwritten Documents

Stéphane Nicolas *Thierry Paquet* *Laurent Heutte*

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

Abstract

We consider in this paper the problem of complex handwritten page segmentation 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. Using this model, the segmentation is performed using optimization techniques. Using the MRF framework, the segmentation is equivalent to an image labeling problem and is performed using optimization techniques.

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 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]. 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, without taking context into account and then sometimes fail to find the good solution. Or 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 of 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 in bidimensional data. MRF framework has been widely studied these last decades [4], and MRF models have been applied for different tasks in image analysis [5], but at the best of our knowledge, never for handwritten document.

We propose to use Markov Random Field 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 french writer Gustave Flaubert into their elementary parts, namely: text lines, erasures, punctuation marks, inter-linear annotations, marginal annotations (just to mention the most important of them). In the MRF framework, segmentation is adressed as an image labeling problem. This problem can be resolved 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. We conclude by some expected future works in section 5.

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. 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.

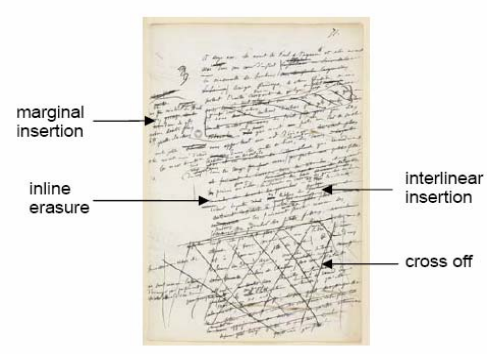


Figure 1. One example of Flaubert's manuscript layout.

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 [4], 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 $g(i,j)$, $1 \le i \le n$ $1 \le j \le m$. 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 on each site of the grid G denoted by $O = \{o(i,j),\ 1 \le i \le n\ \ 1 \le j \le m\}$. 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 class pattern, the problem of layout extraction in the image can be formulated as that of finding among all the possible labeling or state configurations of the field X that can be associated to the image, the most probable according to the model, i.e. finding:

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

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

$$\hat{X} = \arg\max_{X \in E}\left(\prod_s P(o_s|x_s)\prod_s P(x_s|x_{s'}, s' \in N_G(s))\right)$$

where $N_G(s)$ is the neighborhood of the site s.

While in this expression the term $\prod_s P(o_s|x_s)$ can be computed using Gaussian mixtures to modelize the conditional probability densities of the observations, the calculation of the second term (i.e. $\prod_s P(x_s|x_{s'}, s' \in N_G(s))$), which represents the contextual knowledge introduced by the model or prior 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. Image decoding using Markov Random Field models is an optimization problem. It consists in finding the realization $\hat{x}$ of the label field X which maximize the joint probability $P(X,O)$ of the observations set O and the label field X, or similarly an energy function. In fact according to the Hammersley-Clifford theorem, a MRF is equivalent to a Gibbs distribution [4], 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 $N = \{N_S, s \in S\}$. V_c is a potential function associated to the clique c and Z is a normalization constant called partionning 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:

$$U(X,O) = \sum_s -\log\left(P(o_s|x_s)\right) + \sum_{c \in C} V_c(X)$$

Thus in the MRF-MAP framework, decoding or image labeling 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. Different optimization techniques can be used to find the optimal configuration of the label field by minimizing the energy function [5]. Among them we can cite Simulated Annealing (SA) [6][7], Iterated Conditional Modes (ICM) [8], Highest Confidence First (HCF) [9], 2D Dynamic Programming Region Merging method [10][11], Genetic Algorithms (GA) [12], or Ant Colony System (ACS)[13].

3. Application of MRF labeling to handwritten document segmentation

When using MRF-MAP labeling 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 by 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 resolve this task.

- Probability densities

The probability densities are modeled by gaussian mixtures. The parameters of the mixtures are learned 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:

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

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

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

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

As for the gaussian mixture parameters, these probabilities are learned 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 potential functions are defined as follows:

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

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 on the image. As we work on binary images, we have choosen to extract for each site s 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 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 (see Figure 2).

Figure 2. Extracted multiscale image features

Note that the size of the cells $g(i,j)$ 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. More the cells are small, more the labeling is fine, but more sites there will be, 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 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. The results are provided in the next section.

4. Results

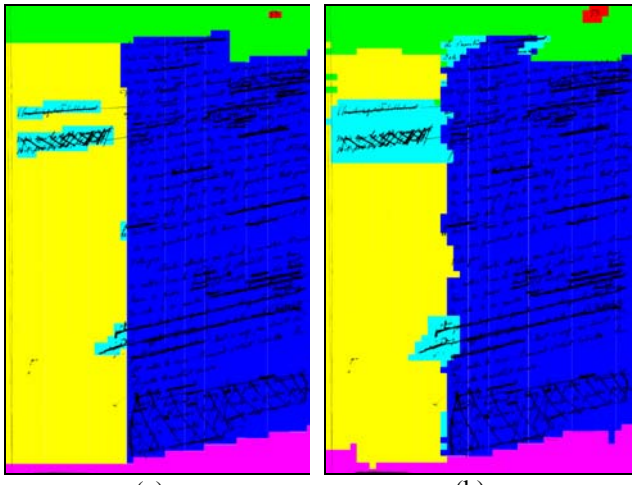
The analysis of the results of a document image segmentation algorithm is a difficult and not always a well defined task, since there exist very few protocols and image databases for performance evaluation [14]. The few existing ones are only designed for machine printed documents for which the proposed 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 field of handwritten documents or historical documents. As our approach is able to produce labelings at different analysis level using different grid sizes, we present here the results obtained on two different segmentation tasks working at two different scales. The first task consists on labeling large areas of interest in manuscript images, such as text body, margins or text blocks, working at a coarse resolution. For this task we provide quantitative results in term of labeling rates and processing times, for several decoding methods. The results obtained are also illustrated visually and discussed. With the second task, which consists on text line labeling, we show the ability of this approach to perform at finer level, in order to

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

extract and separate small entities such as words or word fragments and erasures. For several reasons we explain, we provide for this task qualitative results only obtained on few images of full page of handwriting or parts of pages from the Bovary database.

4.1. Zone Labeling

In order to evaluate precisely the performance of our approach and compare the decoding methods according to labeling rate and processing time, we have first considered a segmentation task where a simple coarser labeling is possible. In this case, it is easy and fast to label a database of Flaubert manuscript images manually for model learning and groundtruthing. The task we consider consists in labeling the main regions of the manuscripts such as text body, margins, header, footer, page number, and marginal annotations (see Figure 3.a). The 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 have been binarized and manually labeled according to the defined 6 labels.



(a) (b)

Figure 3. Zone labeling at a coarser scale: (a) groundtruth (b) result with Markovian labeling using the following color/label convention: red = page number, green = header, blue = text body, pink = footer, cyan = text block, yellow = margin.

The database has been 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 have a regular grid where the dimensions of each cell are 50×50 pixels. We compare the results obtained with a Mixture Model using Maximum Likelihood criterion and the results obtained with ICM, HCF and 2D Dynamic Programming (2D DP) decoding with the groundtruth labelings manually produced (Figure 3). For each decoding method we evaluate the global labeling rate (GLR) by counting the number of well-labeled sites according to the following formula:

$$GLR = \frac{number\ of\ correctly\ labeled\ sites}{total\ number\ of\ sites}$$

$$NLR = \frac{\sum_{i=0}^{q-1} \left(\frac{number\ of\ sites\ correctly\ labeled\ l_i}{total\ number\ of\ l_i\ sites\ in\ groundtruth} \right)}{q}$$

For each decoding methods we also give the average processing time in seconds for one page decoding. This time is only related the to decoding process, probability distribution estimation is not taken into account. Results are provided in Tab 1.

	Mixtures	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 method

These results show that the use of a MRF model allows to increase the normalized labeling rate and that the HCF algorithm outperforms the other decoding methods. Furthermore HCF algorithm is faster than 2D dynamic programming method. The difference between GLR and NLR are due to non homogeneous class repartition in dataset (see Figure 4).

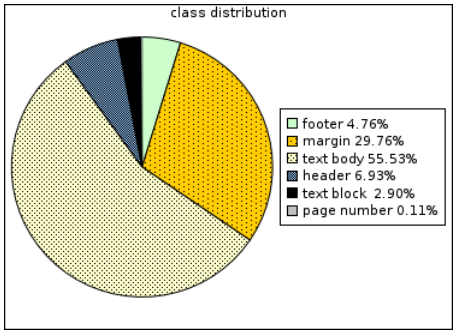


Figure 4. Class distribution in dataset

4.2. Text Line Labeling

Let us recall that Flaubert's manuscripts contain a lot of deletions and crossed out words or lines (see Figure 1). Therefore, in this second experiment, we have tried to evaluate the capabilities of our method to work at at finer analysis level, on a specific task which consists in separating words (or parts of words) and deletions, and to extract text lines using a prior model which integrates several states. For this purpose we have first defined a model made up of 4 states: "pseudo-word", "deletion", "diacritic" and "background". We work at the pixel level using a regular grid of 1×1 cells and we use the the 2D Dynamic Programming method of Geoffrois for decoding. For this task we provide qualitative results only because it is very hard to manually label images at a pixel level for groundtruthing. Figure 5.a. presents the results obtained with this model on a page fragment.

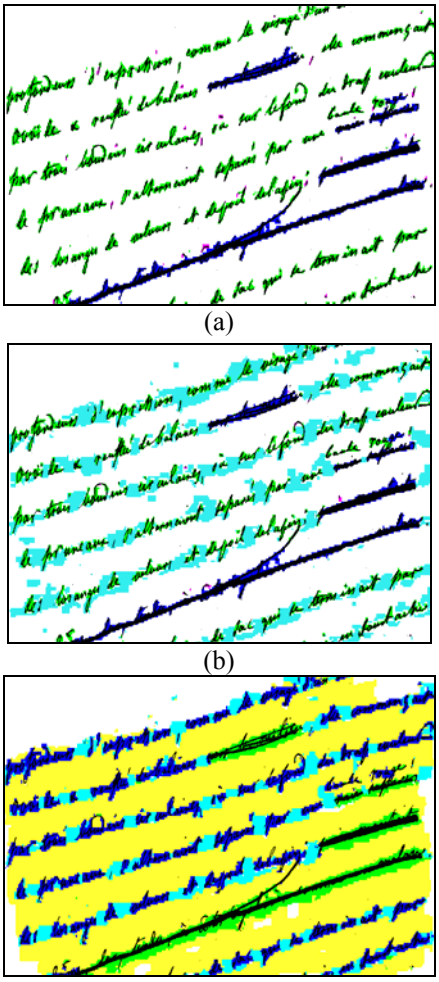


Figure 5. Segmentation results obtained on a page fragment: (a) using a 4-state model; (b) using a 5-state model; (c) using a 6-state model, with the following color/label convention: white = background, green = textual component, blue = erasure, pink = diacritic, cyan = interwords spacing, yellow =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

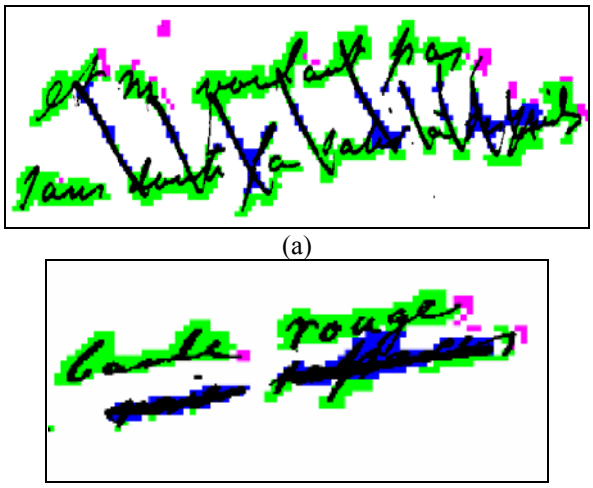


Figure 6. Segmentation results obtained on some complex page fragments using the 4-state model.

Figure 6.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 6.b. shows similar results on a fragment containing a word and an erasure connected by a descending loop. Both components are well separated. This model allows to extract word fragments and erasures, but does not model text lines, so we have refined it by introducing an additionnal "inter pseudo-word space" state. 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 7.

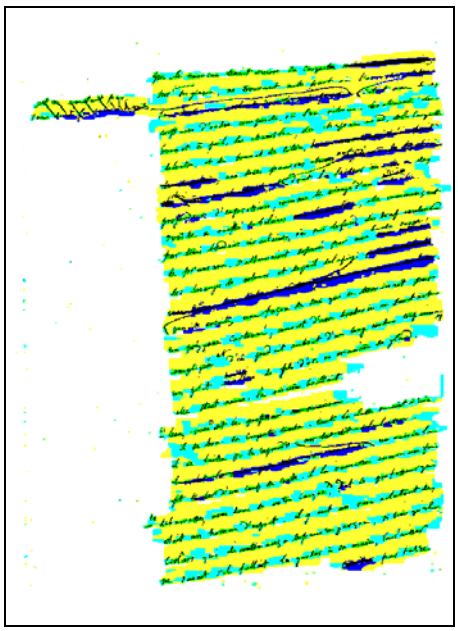


Figure 7. Segmentation results obtained on a complete Flaubert manuscript page using a 6-state model with the following color/label convention: white = background, green = textual component, blue = erasure, pink = diacritic, cyan = interwords spacing, yellow = interline.

5. Future works

As pointed out by He in [15], Markov Random Fields have two main drawbacks. First, they make hypothesis about the independance of observations, for inference tractability reasons. These hypothesis are too strong for image labeling. For these reasons only local relationships between neighboring nodes are incorporated into the model. The second one is their generative nature. Markov Random Field attempt to model the joint distribution of the observed image and the corresponding label field, so a great effort is spent to model the observation distribution. However during the decoding the problem is to estimate the conditional distribution of the label field according to the observed image, there is no need to try to model the joint distribution, which may be very complex, but only the conditional distribution of the label field given observations. In consequence, less training data are needed. It is the reason why discrimative models, such as Conditional Random Fields, have been proposed recently to directly model this conditional distribution. Conditional Random Field have been introduced first by Laferty and Mc Callum [16], for part-of-speech tagging, that is to segment one-dimensional sequences, and have been adapted to image segmentation. In [15], He and al. propose a MLP-based CRF implementation for image labeling, which aim to take into account and to learn features that operate at different scales of the image. Pointing the fact that labeling process needs contextual information because of the dependance of the labels accross the image, the authors propose a multiscale conditional random field model considering three analysis levels: a local analysis, a regional analysis and a global analysis.

Up to now Conditional Random Field have not yet been applied to document image segmentation. In future work, and starting from our MRF model, we propose to transform it to a discrimative conditional model.

6. 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 labeling using different optimization techniques such ICM, HCF and dynamic programming. We have tested the approach on a dataset of manuscripts of french writer Gustave Flaubert. The proposed approach provides interesting results especially with HCF algorithm. 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 an easier adaptation to different type of documents. However due to their generative nature, Markov models suffer from several limitations. For this reason we plan in future works to provide our system an evolution towards Conditional Random Fields which are discriminative models.

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)*, pp. 448-459, Edinburgh, Scotland, 2003.

[2] H. Baird, "Digital Libraries and Document Image Analysis", *Proceeding of the seventh International Conference on Document Analysis and Recognition (ICDAR'03)*, pp. 2-14, Edinburgh, Scotland, 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, pp. 38-62, 2000.

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

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

[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] S. Kirkpatrick, C. D. Gellatt and M. P. Vecchi, "Optimization by Simulated Annealing", *Science* (1983), no. 220, pp. 671-680, 1983.

[8] J. E. Besag, "On the statistical analysis of dirty pictures", *Journal of the Royal Statistical Society B*, vol. 48, no. 3, pp. 259-302, 1986.

[9] P. Chou, C. Brown, "The theory and practice of Bayesian image labeling", *International Journal of Computer Vision*, 4, pp.185-210, 1990.

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

[11] E. Geoffrois, S. Chevalier, F. Prêteux, "Programmation dynamique 2D pour la reconnaissance de caractères manuscrits par champs de Markov", *proceedings of RFIA, Reconnaissance de Formes et Intelligence Artificielle*, pp. 1143-1152, Toulouse, France, Janvier 2004.

[12] E.Y. Kim, S.H. Park, H.J. Kim, "A genetic algorithm-based segmentation of Markov Random Field modeled images", *IEEE Signal processing letters* 11(7): 301-303, 2000.

[13] S. Ouadfel, M. Batouche, "MRF-based image segmentation using Ant Colony System", *in Electronic Letters on Computer Vision and Image Analysis (LCVIA)*, vol. 2, n°1, pp. 12-24, August 2003.

[14] "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.

[15] X. He, R.S. Zemel, M. A. Carreira-Perpinan, "Multiscale Conditional Random Fields for Image Labeling", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04)*, pp. 695-702, Washington DC, USA, 2004.

[16] J. Lafferty, A. McCallum, F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data", *18th International Conference on Machine Learning*, pp 282-289, Williamstown, USA, 2001.