Unconstrained Handwritten Document Layout Extraction using 2D Conditional Random Fields

Florent Montreuil
DGA/Centre d'Expertise Parisien
16, bis avenue Prieur de la Côte d'or
94114 Arcueil cedex, FRANCE
Florent.Montreuil@gmail.com

Emmanuèle Grosicki
DGA/Centre d'Expertise Parisien
16, bis avenue Prieur de la Côte d'or
94114 Arcueil cedex, FRANCE
Emmanuele.Grosicki@etca.fr

Laurent Heutte
Université de Rouen, LITIS EA 4108
BP 12 - 76801 Saint-Etienne
du Rouvray, FRANCE
Laurent.Heutte@univ-rouen.fr

Stéphane Nicolas
Université de Rouen, LITIS EA 4108
BP 12 - 76801 Saint-Etienne
du Rouvray, FRANCE
Stephane.Nicolas@univ-rouen.fr

Abstract

The paper describes a new approach using a Conditional Random Fields (CRFs) to extract physical and logical layouts in unconstrained handwritten letters such as those sent by individuals to companies. In this approach, the extraction of the layouts is considered as a labeling task consisting in assigning a label to each pixel of the document image. This label is chosen among a set of labels depicting the layout elements. The CRF-based method models two stochastic processes : the first one corresponds to the association between pixels and labels, the second one to the relationship of one label with respect to its neighboring labels. The CRF model gives access to the global conditional probability of a given labeling of the image according to image features and some prior knowledge about the structure of the document. This global probability is computed by means of local conditional probabilities at each pixel. To find the best label field, a key point of our model is the implementation of the optimal inference 2D Dynamic Programming method. Experiments have been performed on 1250 handwritten letters of the RIMES database. Good results have been reported showing the capacity of our approach to extract simultaneously the physical and logical layouts.

1. Introduction

Document layout analysis is a key concept in document processing [14] and a crucial step in many applications related to document images, like text extraction using optical character recognition (OCR), document reflowing, and layout-based document retrieval.

Layout analysis is the process of identifying layout structures by analyzing page images. The process is generally subdivided into several sequential stages. Each of them considering the output result of the previous one. In particular, two key steps are considered: the physical structure extraction (or segmentation) which consists in segmenting the document into pages, blocks, lines, words and the logical structure extraction (or labeling) which groups these differents segments to form logical units which are assigned meaningful labels.

The majority of the state-of-the-art methods in the litterature deals with machine-printed documents [13], [14] and are based on rules. These methods allow to obtain sequentially the physical and logical structures [5]. Few methods are really dedicated to unconstrained handwritten documents because of the variability of such documents: skew lines, slant words, gaps between words or lines. Due to this variability the analysis of the structures is very difficult. As a consequence, the existing methods are often based on statistical approaches as [9] and [12]. Moreover, it appears that it is better to extract simultaneously the physical and logical layout structures rather than sequentially as they are particularly linked in this kind of documents [13].

This article focuses on unconstrained handwritten letters such as those sent by individuals to companies or administrations. Our aim here is to extract relevant information such as date, place, sender details which are useful for an automatic processing of these letters [3] (see Figure 1). To achieve this task, we have chosen a statistical approach.

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.132

The layout extraction is then viewed as a labeling problem whose goal is to classify each pixel of the image into one of a set of predefined labels. Among existing statistical methods, we have chosen a Markovian discriminant approach based on CRF models which are known to be more suitable to this kind of labelling problem than generative models like MRFs (Markov Random Fields) [4], [6] and [12]. CRFs possessed advantage do not make any assumption about the image data, they do not require a large amount of labeled data for the estimation of their parameters and can more easily integrate different levels of contextual information (global and local). Moreover, an other advantage of CRFs is the possibility to use a large number of various input features without changing the structure of the model.

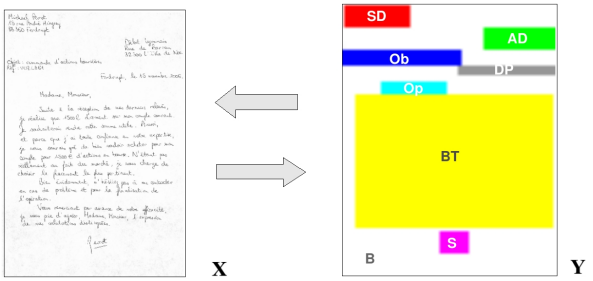

Figure 1. Example of a letter (left) with its layout (right). The letter is segmented into blocks : B: Background, SD: Sender Details, DP: Date Place, AD: Addressee Details, Ob: Object, Op: Open, BT: Body Text, S: Signature. [3]

In this paper, we consider low level features. We combine features on the physical and logical structures of the document respectively textural information and spatial position information. We show how it is possible to consider higher level features based on textual information to increase the knowledge of our system. The proposed method combines these features with Markovian contextual information by using SVM classifiers hence giving access to the conditional probabilty of a label given these features. To obtain the optimal configuration of labels given observations, we propose to use the optimal 2D Dynamic Programming method proposed by [2]. We have tested our model on 1250 handwritten letters belonging to the RIMES database and we have obtained good performance compared to the systems evaluated during the second RIMES campaign.

The rest of the paper is organised as follows. Section 2 describes the proposed CRF model. Experiments and results are described in section 3 followed by a discussion about the future works we propose to improve the performances of our system.

2. Conditional Random Field model

In a CRF approach, the document layout is supposed to be produced by a hidden state field noted Y taken values in a finite set of labels L (see Figure 1). The configuration of these labels depicts the document layout. According to this CRF formalism, we assume that the state field is Markovian what means that each state depends only on the states belonging to its neighbouring. Labels of these underlying states are estimated from observations which are extracted on the entire image X contrary to MRF models. In this kind of approach, the image is segmented in $n \times m$ patches forming a rectangular grid G. Each patch is associated with a state y_j allowing to affect a label to the pixels in this patch. The pair (X, Y) thus defined corresponds to a CRF model as proposed in [7].

In the CRF approach, the conditional probability of a configuration y of the state field Y given a set of observations x is directly accessed. To obtain the optimal layout $\hat{Y}$, we just have to calculate the configuration of states y among $\mathcal{Y}$ configurations that maximizes this conditional probability:

$$\hat{Y} = \underset{y \in \mathcal{Y}}{arg\,max}\ P(Y = y|X = x) \tag{1}$$

In these CRF models, the conditional probability of Y given the input image X can be decomposed in two potential functions as proposed in [6]:

$$P(Y = y|X = x) = \frac{1}{Z} \prod_{j=1}^{n \times m} \left(exp(A(j, y_j, x) + \sum_{j,k \in E} I(j, k, y_j, y_k, x) \right) \tag{2}$$

where Z is a normalization factor that sums this product on all labels, E the set of edges that show the dependence between the different states and k the index of the different states located in the neighbourhood of the state y_j.

A is called the association potential linking the observed data to states and I is the interaction potential, which gives the neighbour/contextual dependencies by associating pair wise the interaction of the neighboring labels to the observed data.

These potential functions can be calculated as a sum of feature functions f_c weighted by a parameter θ_c:

$$A(j, y_j, x; \theta^A) = \sum_c (f_c^A(j, y_j, x)\theta_{cj}^A)$$
$$I(j, k, y_j, y_k, x; \theta^I) = \sum_c (f_c^I(j, k, y_j, y_k, x)\theta_{cjk}^I) \tag{3}$$

We have chosen to use discriminant classifiers to estimate these potential functions as proposed in [4] and in [12]. Among existing classifiers, we have chosen SVM classifiers (Support Vector Machine) as they are known to have good generalization properties compared to conventional classifiers [10]. Moreover the association SVM/CRF is highly efficient as it benefits from maximum-margin nature of SVMs and also from the ability of CRFs to model correlations between neighboring output labels. These SVM classifiers allow us to obtain an estimation of the posterior probabilities of the labels according to the features $F(x, y)$ discussed in section 2.1 and 2.2. We can rewrite the global conditional probability as follows:

$$P(Y|X) = \frac{1}{Z} \prod_{j}^{n \times m} \left(exp(\sum_{l \in \{A,I\}} \theta_l P_l(y_j|F(y, x, j))) \right) \quad (4)$$

So the CRF is built from two models: an association model and an interraction model. In our approach, we have decided to use a linear combination to link these models as proposed in [12]. Figure 2 presents our CRF model.

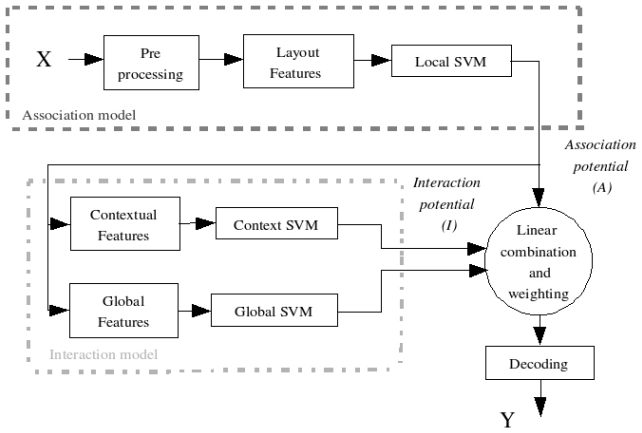

Figure 2. Description of our model.

For the experiments, the LIBSVM package was employed [1]. It offers an efficient multi-class support using internally a one-against-one approach. In order to choose SVMs parameters, we run a cross-validation on a data set independent from the training and testing sets.

2.1. Association Model

The association model allows us to estimate the probability $P_L(y_j|X)$ of a state y_j given a set of continuous features extracted from the image. As said previously, the image is segmented in patches and each state is associated with a patch. All pixels of a patch will be associated with the label l_i of this patch. The size of these patches is an important parameter which is directly related to the accuracy of the model. It has been chosen empirically. To define these states, we have decided to extract textural information describing the physical layout and spatial position information describing the logical layout:

- the normalized x-coordinate and y-coordinate of the center of the patch in the image [9].
- the mean grey level of three windows centered on the considered patch. Corresponding in size to one patch, five patches and nine patches giving a multiscale representation of the grey level (see Figure 3).

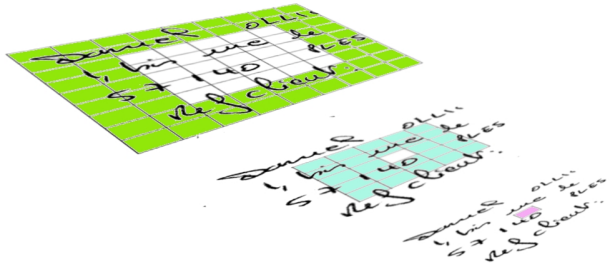

Figure 3. Multi scale analysis.

2.2. Interaction model

The interaction model allows us to regularize the association probability by taking into account the neighbouring labels (Markov dependence). We have decided to consider two levels of neighbourhood called respectively "contextual level" and "global level", as proposed in [12] (see Figure 2):

- The contextual model use near neighbours noted Y_c to regularize the association model. For each state, we take as features the probabilities obtained in output of the local SVM. We take into account a neighbourhood of 4 states for "computional reasons". So we obtain a contextual probability: $P_C(y_j|X, Y_c)$
- The global model allows us to consider a larger neighbourhood noted Y_g. For each state, we take as features the probabilities of occurence of each label in a window centered around the considered state. So we obtain a probability: $P_G(y_j|X, Y_g)$

2.3. Inference

These different levels described by P_L, P_C and P_G are linearly combined in order to obtain the conditional probability $P(y_j|X, Y_c, Y_g)$ for each state. To determine the optimal configuration of states corresponding to the document layout, the two-dimensional dynamic programming

algorithm is used. This inference algorithm [2] has the advantage to be optimal compared to classical inference algorithm like ICM (Iterated Conditional Modes) [12] or descent algorithm. It consists in merging region of states in order to obtain one region whose configuration of labels gives the layout document. Each region R is attached to a list of possible configurations associated to a corresponding conditional probability $P(y_j|X, Y_c, Y_g, y_k = l_i)$. The merging process produces a new region R which is attached to a new list of configurations associated to a conditional probability $P(R = c|X)_{c \in C}$ (where C are the set of the possible configurations for the region R) product of the conditional probability of both regions. In our model, to take into account the different state configurations at each step of the 2D dynamic programming, each conditional probabilty $P(y_j|X, Y_c, Y_g)$ is weighted by the conditional probability given a particular configuration of its neihgbours. It results a new probability: $P(y_j|X, Y_c, Y_g, y_k = l_i)$ for each state.

3. Experiments

We have tested our model on the database of the second RIMES evaluation campaign composed of 1250 letters (1050 for the learning database, 100 letters for the validation database and 100 for the test database). Each letter has been scanned in 300 dpi which corresponds approximatively to a size of 2500x3500 pixels. The size of each patch has been chosen empirically equal to 65x65 pixels which corresponds to a good trade-off between computational complexity and accuracy.

To compare the layout extraction results obtained by automatic systems to ground-truths, we use the metric proposed by the RIMES evaluation based on grey level [3]. It consists in comparing labels of each pixel in both hypothesis and ground-truth. This metric called Err corresponds to pixel error rate defined by the sum of grey levels of miss classified pixels normalized by the sum of grey levels of all pixels. The results presented in table 1 show the error rates obtained by 3 systems during the second RIMES campaign and by our system. The three systems which have been presented to the second RIMES evaluation campaign are:

- MRF approach with low level features and post processing [9] (lab1).
- DMOS system (Description and Modification of Segmentation), grammar based method [8] (lab2).
- MRF approach with low level features [11] (lab3).

Lab 1 and lab3 proposed a statistical approach based on MRF whereas lab2 proposed a system based on rules. The difference between performance obtained by lab 1 and lab3

Table 1. Results obtained on the RIMES database

	lab1	lab2	lab3	Our model
Err (%)	8,53	8,97	12,62	11,55

can be explained by the addition in lab1 solution of a post-processing based on rules to correct errors generated by its proposed MRF solution. Concerning lab2, its solution gives good results but as we said before, the use of rules is a too rigid solution, not easily adaptable to other kinds of documents. Besides, the analysis of errors obtained by this method show high error rate on some images whose layout are not classical and so do not correspond to the predefined grammar rules.

The results obtained by our method (see Figure 4) can be compared to those obtained by lab3 as we use the same kind of layout features. The difference of results can be explained by the use of different contextual information and the 2D programming inferring method. To improve the performance of our model, we propose to take into account high level features as we will see in next section.

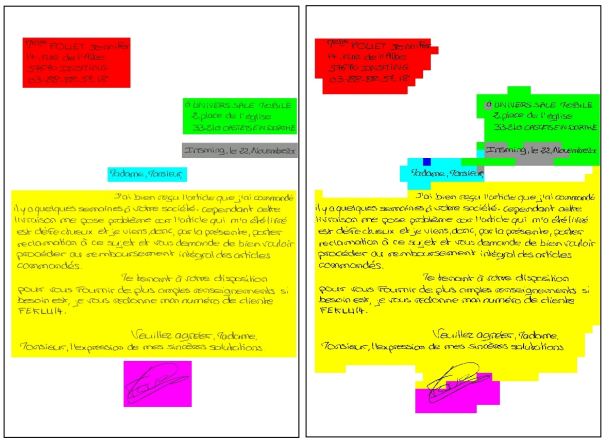

Figure 4. Ground truth (left) and Layout extraction (right).

4. Error analysis and discussion on future works

By analysing the different errors generated by our algorithm, the chosen features based on spatial and textural information appear to be not enough discriminant to separate close blocks (see Figure 5). Table 2 gives the confusion matrix for the different labels. Most of the confusions are done

on the labels Object (Ob), Opening (Op), Signature (S) and Date-place (DP) which correspond to blocks difficult to delimitate (All these confusion are in bold in the Table 2). For example, there is often no blank space between the "opening" block (Op) and the "text body" block (BT).

Table 2. Label confusion matrix

	B	SD	AD	Ob	Op	S	BT	DP
B	86,0	1,3	1,9	0,5	0,5	1,6	7,8	0,4
SD	8,2	84,8	1,5	3,7	0,3	0,0	0,6	0,9
AD	10	1,8	79,1	2,5	0,2	0,8	1,9	3,3
Ob	16,8	**11,2**	**12,7**	**37,1**	**13,3**	0,0	2,6	6,3
Op	13,7	0,9	7,8	7,7	**40,6**	0,0	**27,8**	1,4
S	7,2	0,0	0,0	0,0	0,0	**55,6**	**37,2**	0,0
BT	3,8	0,0	1,2	0,5	0,9	0,1	93,2	0,2
DP	8,5	2,6	**24,6**	5,7	2,5	1,7	7,7	**46,8**

To improve the knowledge of our model, we propose to add high level features based on textual content as proposed in [5]. The proposed idea consists in detecting some key words to improve the letter segmentation. For example, in figure 5, we can notice some confusions between the block "Date, Place" (DP) and the block "Addressee Details" (AD). The detection of the word "*Novembre*" (November) referring to the block "Date,Place" would allow us to correct some label errors.

Figure 5. Zoom on an error. (top: model result ; bottom: ground-truth)

5. Summary and Conclusion

A Conditional Random Field model for layout extraction in unconstrained handwritten letters has been proposed and discussed in this paper. This approach show that if we have ground-truths, it is possible to learn and extract any document layout. So it is easly to extend this model to another document changing the features. In this approach, we have used simple low level features: multiscale textural information and position information. An advantage of this approach is to be able to combine features of different nature that describe physical and logical layout structure. The first experiments on the RIMES database ehibit good results even with few features. Some confusions appear in regions where the low level features are not sufficient. One solution to improve these results may be to take into account textual information. These information can be easily incorporated in our model, so we will consider this solution in future works.

References

[1] C. C. Chang. Libsvm: a library for support vector machines. *Software available at: http://www.csie.ntu.edu.tw*, 2006.

[2] E. Geoffrois. Multi-dimensional dynamic programming for statistical image segmentation and recognition. In *the Conference on Image and Signal Processing (ICISP03)*, 2003.

[3] E. Grosicki. Rimes evaluation campaign for handwritten mail processing. *the 11th International Conference on Frontiers in Handwritting Recognition (ICFHR08)*, 2008.

[4] X. He. Multiscale conditional random fields for image labelling. In *the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR04)*, volume 2, pages 695–702, 2004.

[5] S. Klink. Document structure analysis based on layout and textual features. In *the International Workshop on Document Analysis Systems, DAS00*, pages 99–111. IAPR, 2000.

[6] S. Kumar. Discriminative random fields. *International Journal of Computer Vision*, 68(2):179–202, 2006.

[7] J. D. Lafferty. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *the 18th International Conference on Machine Learning (ICML01)*, pages 282–289. Morgan Kaufmann, San Francisco, CA, 2001.

[8] A. Lemaitre. Multiresolution cooperation makes easier document structure recognition. *IJDAR*, 11(2), November 2008.

[9] M. Lemaitre. Approche markovienne bidimensionnelle d'analyse et de reconnaissance de documents manuscrits. Technical report, France, 2007.

[10] V. Malathii. Support vector machine for discrimination between fault and magnetizing inrush current in power transformer. *Journal of Computer Science*, 11(3):894–897, 2008.

[11] S. Nicolas. Handwritten document segmentation using hidden markov random fields. In *the 8th International Conference on Document Analysis and Recognition (ICDAR05)*, pages 212–216, Washington, DC, USA, 2005. IEEE Computer Society.

[12] S. Nicolas. 2d markovian models for document structure analysis. In *the 11th International Conference on Frontiers in Handwriting Recognition (ICWFHR08)*, pages 658–663, Montreal, Quebec, Canada, 2008.

[13] S. Souafi. Contribution a la reconnaissance des structures des documents ecrits : Approche probabilistes. Technical report, France, 2002.

[14] Y. Y. Tang. Document analysis and understanding : A brief survey. In *the 1st International Conference on Document Analysis and Recognition (ICDAR91)*, pages 17–31, 1991.