Table detection in handwritten chemistry documents using conditional random fields

Nabil Ghanmi, Belaïd Abdel

Table detection in handwritten chemistry documents using conditional random fields

Nabil Ghanmi*†
*LORIA, Nancy, France
nabil.ghanmi@loria.fr
†eNovalys, Illkirch, France
http://www.enovalys.com

Abdel Belaïd
Universit de Lorraine - LORIA
Nancy, France
abdel.belaid@loria.fr

***Abstract*—In this paper, we present a new approach using conditional random fields (CRFs) to localize tabular components in an unconstrained handwritten compound document. Given a line-segmented document, the extraction of table is considered as a labeling task that consists in assigning a label to each line: TableRow label for a line which belongs to a table and LineText label for a line which belongs to a text block. To perform the labeling task, we use a CRF model to combine two classifiers: a local classifier which assigns a label to the line based on local features and a contextual classifier which uses features taking into account the neighborhood. The CRF model gives the global conditional probability of a given labeling of the line considering the outputs of the two classifiers. A set of chemistry documents is used for the evaluation of this approach. The obtained results are around 88% of table lines correctly detected.**

***Keywords*-table detection; conditional random fields; feature functions; labeling; local features; contextual features.**

I. Introduction

When talking about table understanding in the document image, one realizes two different sub-problems [1]: table detection and table recognition. Table detection deals with the problem of finding boundaries of tables in a document image. As for table recognition, it focuses on analyzing the detected table by finding its rows and columns and tries to extract the logical structure of the table.

Many works on table understanding topic assume that the region containing the table is already known and mainly focus on the extraction of its physical and logical structure. On the other hand, some researches are conducted to detect tables in the document images but most of them are dedicated to some specific table structures or they make a priori assumptions on the position and the layout of the table to deal with some difficulties.

When the problem of table detection is treated in handwritten documents, challenges are getting bigger because of the huge variability of the handwriting styles and the imperfections affecting the tables. The document dataset that we are interested in, is a collection of handwritten chemistry documents containing hand drawings, tables and text blocks. These documents are different from most in the literature because there is no constraint neither on the structure nor on the size of the table. The examination of the existing table structures (see example in Figure 1) in our dataset discloses many imperfections such as:

- Missed ruling lines separating cells.
- Missed cells.
- Imperfect vertical alignment of cells.
- Irregular horizontal spacing between cells.
- Presence of fields spread over two (or more) rows and/or columns.

Despite the inherent difficulties, we propose in this paper a technique for table detection without making any assumption about the start and the end of the table in the document and we do not use any a priori knowledge about its structure. Our approach relies on labeling lines to know if they belong to a table or not. We start from a line-segmented document and from each line we extract some selected features that will be used to perform a first classification. Due to the line imperfections previously mentioned, some errors may occur. Hence, we perform a second classification using contextual features taking into account the inter-correlation between the neighboring lines. These two classifiers are then combined using CRF.

The remainder of this paper is organized as follows: in section II, we explore the most important works conducted on the table detection field. In section III, we explain our proposed approach. We describe the selected local and contextual features and we expose the line labeling process using CRF. We present our experimental setup and preliminary results in section IV and conclude in section V.

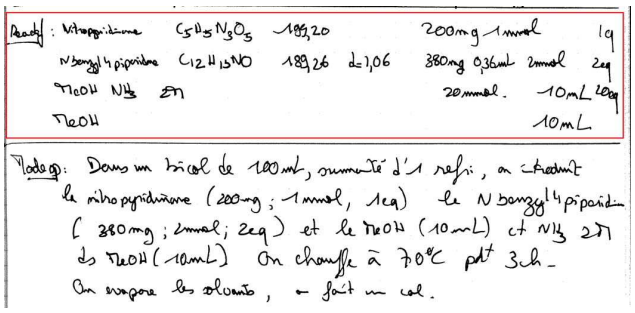


Figure 1. A sample document containing table (outlined in red)

2167-6445/14 $31.00 © 2014 IEEE
DOI 10.1109/ICFHR.2014.32

II. Related Works

A survey of existing table detection approaches discloses that the main works are based on three principal aspects. Thus, we can distinguish three groups of approaches:

(a) Ruling line detection based approach [2] [3] [6] [7].
(b) White space analysis based approach [1] [4] [5].
(c) Vertical alignment of text blobs based approach [1] [8].

The works of the group (a) use ruling lines as initial indicator of table regions and then further refine this decision by a measure based on some features. Chen and Lopresti [2] use a probabilistic alternative of Hough transform to detect lines in the document. In order to ensure high recall of table rulings, some lines are excluded based on the fact that the table ruling lines are parallel or orthogonal. Spatial displacement of text is also used to remove other false-alarms. Then, they detect key points by computing the intersections of horizontal and vertical rulings. Among these key points, the most probable subset which constitutes the table structure is selected using an optimization procedure. Kasar and al. [3] proposed a method to locate table regions in a heterogeneous corpus of French, English and Arabic documents by detecting the line separators in the table. They use a run-length based technique to extract the horizontal and vertical ruling lines. A set of 26 features is computed from each group of intersecting lines. These features describe the line positions and lengths as well as the regularity in the arrangement and spacing between two adjacent lines. An SVM classifier is used to check if these lines belong to a table or not.

In the group (b), Chen and Lopresti [4] propose a method for simultaneous detection and recognition of tabular structure in noisy handwritten documents. The detection method is based on the location of key-points defined as the intersections of white streams within text lines (inside-space) and between text lines (interline space). Then, grids of key points are built using clustering and horizontal projection techniques. The Min-cut/Max-Flow algorithm, based on some structural features, is used to validate key points in the grid. Hu and al. [1] propose a medium independent table detection method. They present a high-level framework that determines the optimization problem and an algorithm for its solution. The authors do not make any assumption about the position and the structure of the table but they calculate probabilities for all possible start or end positions of the table. The high-level detection algorithm is independent of any particular table quality measure. In order to apply the proposed general solution, the authors propose two quality measures. The first measure is based on the inside-space and describes the correlation between the white space streams in two lines. The second measure relies on vertical connected component analysis (VCCA) to describe the vertical alignment of words. Words are vertically aligned if they overlap significantly and have similar lengths. Based on this measure, vertical connected components, which are somehow equivalent to table columns, are constructed.

One of the most important works lying in (c) is that presented by Kieninger [8]. The author proposed a method of table extraction based on block segmentation of the document. The method uses the word bounding boxes and recursively groups them into blocks based on the horizontal overlapping with their vertical neighbors in the previous and next lines. Admitting the existence of an horizontal spacing between table columns, this segmentation allows identifying and isolating these columns. One problem with this method is that the defined column block is broken up by occasional inconsistent lines (blank or single word line for example).

Apart from the used classification, another work on table detection is that presented by Pinto and al. [9]. The authors proposed a CRF based approach to extract table in plain-text government statistical reports. They start by extracting a set of features to identify the line types. They use a CRF model for labeling each line of a document with a tag that describes the line function. Two goals are simultaneously accomplished: the performed labeling marks the boundaries of the table (table location) and identifies the row types and roles in the table (table recognition).

The above methods can not effectively deal with a large variety of table structures in handwritten documents, like those in which we are interested. In fact, the approaches based on space analysis or text blobs alignment assume that the table is well structured. While the tables in question are unruled and their structures present many imperfections.

III. Proposed Method

In the context of the present work, we consider the table detection problem as a labeling task. An image document is seen as a line sequence $L = \{l\}$. A label is associated to each line. The line labels are supposed to be produced by a field of hidden states denoted X taking values in a finite set of states T. This field is assumed Markov which means that there is a conditional dependence on the neighbor lines. In this paper we focus on the binary case $T = \{0, 1\}$. Each state of the field is associated to an image line which will be assigned to the corresponding label. The problem can be formulated as follows: given a set of observations Y, it is to find the most probable label configuration of the field X among all the possible labeling E that can be associated to the image, i.e. finding:

$$\hat{X} = \arg\max_{X \in E}(P(X/Y)) \tag{1}$$

To find this posterior probability, CRF model has been proven to be an interesting tool. CRFs lie in a probabilistic framework and are based on a conditional approach for labeling data sequence. These models consider the conditional probability $P(X/Y)$ rather than the joint probability $P(X, Y)$. Therefore, they give the probabilities of the possible label sequence given an observation sequence. Unlike

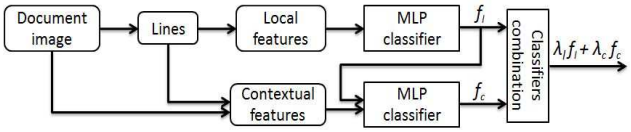


Figure 2. Local and contextual classifier combination

generative models (Markov Random Fields for example), CRFs do not model the observations. The discrimintaive task is therefore directly formulated by:

$$P(X = x/Y = y) = \frac{1}{Z} \prod_{l \in L} exp(\sum_{k} \lambda_k f_k(x, y, l)) \quad (2)$$

where Z is a normalization factor over all state sequences, f_k is an arbitrary feature function over its arguments, and λ_k is a learned weight for each feature function.

A. Feature functions

The feature functions assess the compatibility of labels according to the observation. In this work, we have opted for discriminative classifiers to model feature functions [10]. Several classifiers can be used for such task. We have chosen Multilayer Perceptron (MLP) because it does not make any assumption regarding the probabilistic information about the classes under consideration in comparison to other probability based models. They are also simple and fast.

In our model, we consider two levels of line classification: an individual classification based on local features Y_l and a contextual classification taking into account the line neighborhood information Y_c. Two feature functions f_l and f_c are modeled by these classifiers. Our model can be seen as a combination of two feature functions (see Figure 2) and the conditional probability can be written as:

$$P(X/Y_l, Y_c) = \lambda_l f_l(X, Y_l) + \lambda_c f_c(X, Y_c) \quad (3)$$

B. Feature set

An important advantage of CRF on generative models is that dependencies among the observed variables Y do not need to be explicitly represented, affording thus a use of rich, global features of the input. In this work we extract two sets of features: local features and contextual features.

We start from a line-segmented document. We perform a segmentation of each line into patches (see Figure 3). We used a segmentation method based on the histogram of the distances between connected components in the line [11]. The distance histogram has two peaks: the first is the most frequent distance which corresponds to the distance between the connected components of the same word and the second most frequent peak corresponds to the inter-patches distance.

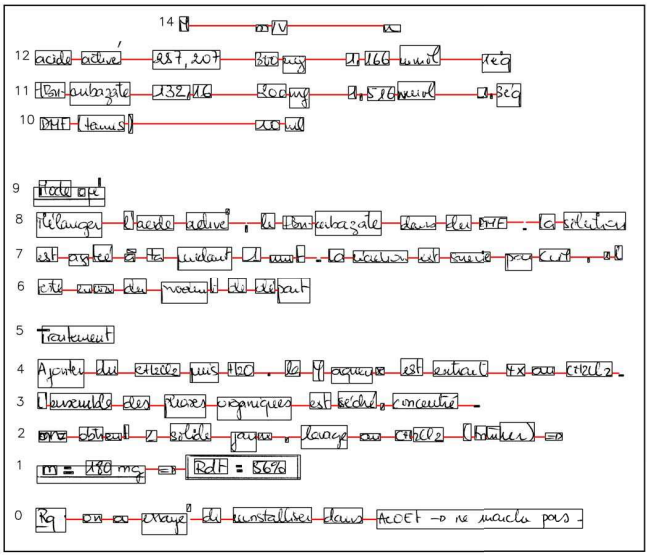


Figure 3. A sample document segmented into patches

Table I
DESCRIPTION OF THE LOCAL FEATURES

Percentage of white space	The sum of white space lengths divided by the line length
Avg white space length	The mean length of the white spaces within a line
Variance white space length	The variance of the white space lengths within a line
Number of patches	The number of the patches within a line
Avg patch width	The mean width of the patches within a line
Avg connected component width	The mean width of the connected component within a line

Local features

Local features are used by the classifier in order to associate a label to each line using the characteristics of that line alone. The selected features are expected to describe both white space and ink in the line. Six features are extracted from each line as described in Table I.

Contextual features

The contextual features take into account the line neighborhood information. We opted for contextual features that measure the correlation [1] of the current line with its two neighbors: the south and north lines. The following features are extracted:

- White space based inter-correlation: this feature is based on the horizontal overlapping of the spaces within two lines. The space between word bounding boxes is considered. We define the horizontal overlapping rate in the following way (see Figure 4). Suppose the horizontal extents of two spaces S_1 and S_2 in two adjacent lines L_1 and L_2 are respectively (x_{11}, x_{12}) and (x_{21}, x_{22}). Without loss of generality, we assume that

$x_{11} < x_{21}$ and $x_{12} < x_{22}$. The horizontal overlapping rate τ_o between S_1 and S_2 is defined as:

$$\tau_o(S_1, S_2) = \frac{x_{12} - x_{21}}{\min(x_{12} - x_{11}, x_{22} - x_{21})} \quad (4)$$

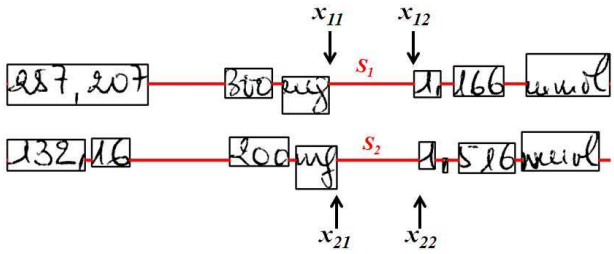


Figure 4. Space-based inter-correlation between two consecutive lines

This formula is used to compute the space-based inter-correlation on the entire lines L_1 and L_2. Let $\{S_{1i}, i \leq N_1\}$ and $\{S_{2j}, i \leq N_2\}$ the set of spaces respectively within L_1 and L_2

$$Incorr_{space}(L_1, L_2) = \frac{\sum_{i \leq N_1, j \leq N_2} \tau_o(S_{1i}, S_{2j})}{\min(N_1, N_2)} \quad (5)$$

where N_1 and N_2 are the number of spaces respectively within L_1 and L_2.

- Patch bounding boxes-based inter-correlation: this feature describes the vertical alignment of the patch bounding boxes within two adjacent lines. Two patches on adjacent lines are considered vertically aligned if their bounding boxes overlap significantly. In the same way that defines the horizontal overlapping between spaces, we define the horizontal overlapping rate between patch bounding boxes and we compute the bounding boxes-based inter-correlation of the two lines.
- Patches number-based similarity: this is a binary feature that takes the value 1 if two adjacent lines have the same number of patches, 0 otherwise.
- In addition to these inter-correlation features, we consider the local conditional probabilities on the label field X in the two adjacent lines. These probabilities are already determined by the local classifier.

IV. Experimentations

A. Data preparation

Our approach lies in a supervised framework. Therefore, we prepared for the experiments, a ground truth composed of 117 documents containing a total of 1785 lines. A line level labeling is performed manually using a simple image editor. The documents are taken from chemistry manuscripts in an unconstrained industrial framework. They are heterogeneous and multi-writer documents. They contain three main regions: hand-drawn chemical formula, table and text blocks. The chemical formula extraction has been the subject of an earlier work [12]. For the experiments of the present work, we assume that the hand-drawn formula was correctly extracted and we limit the search to the zone of table and text blocks. To train each of the both MLP classifiers, a subset of 66 labeled documents is used for the learning. The 51 remaining documents are used for the test. These documents contain a total of 799 lines including 200 that belong to tables (49 tables).

B. Model learning and inference

We used FANN[1] library for both learning and inference of the model. We use two MLPs with one input layer (composed of 6 and 8 neurons respectively for local and contextual MLP), one hidden layer of 30 neurons and one output layer of 2 neurons (TableRow and TextLine).

Model parameters learning consists in training the two MLP classifiers and determining their corresponding weights used for the combination. To train both classifiers, we used a labeled data set. Firstly, the local classifier is trained using only the local features. The output of this classifier is used to estimate the conditional probabilities of the label association to the line in question. Being in the case of binary classification, we used the following output transformation to obtain the conditional probabilities:

$$p(X = i/Y) = \frac{o_i}{o_1 + o_2}\ ,\ i = 1, 2 \quad (6)$$

where (o_1,o_2) are the MLP outputs.

As shown in Figure 2, the input of the contextual MLP is constituted by the probabilities estimated by the local MLP and the contextual features. This MLP is trained using the same labeled data set as the local one. Both MLPs are trained using the standard back-propagation algorithm. The weights λ_l and λ_c used to combine the two classifiers are determined experimentally. These parameters take values in the interval $[0, 1]$ such that $\lambda_l + \lambda_c = 1$. To choose the value λ_l, several experiments of the model are performed, using all the possible values between 0 and 1, with step 0.1. The value which maximizes the recall and the precision of the system on the learning data set is selected.

The inference in the model aims at finding a solution of the optimal field labeling X, i.e. resolving the equation (1), based on maximum a posteriori criterion. The inference process in our system can be described by the following steps:

- First labeling: it is performed by the local classifier. Only local features are taken into account.
- Second labeling: it is performed by the contextual classifier. Both contextual features and local classifier outputs (after being transformed in probabilities) are taken as input for the contextual classifier.

[1]Fast Artificial Neural Network is a free open source neural network library, which implements multilayer artificial neural networks in C .

- Combined labeling: for each line l, the outputs of the two above classifiers are combined to evaluate the score of the potential association of the line to each possible label l_i. The obtained score can be considered as probability and the label maximizing this score is assigned to the line in question.

C. Results and Interpretation

The performance of our system is firstly evaluated at the line level. Each line is labeled as either $TableRow$ or $TextLine$. Table II shows both precision and recall obtained firstly using the local classifier only ($\lambda_l = 1$ and $\lambda_c = 0$), next using the contextual classifier only ($\lambda_l = 0$ and $\lambda_c = 1$) and finally using the combination of the two classifiers (with the experimental values of λ_l and λ_c).

These results show that, by considering both local and contextual levels of analysis, we obtain better results than using the local or the contextual classifier only. We also notice that the system as organized in 2 outperforms the case of using all features (local and contextual) in one vector together with one MLP, especially for labeling table lines which is the main objective. This improvement of table lines detection is due to the fact that the second level of classification regulates the labeling probabilities taking into account the neighboring labels.

An evaluation of the system performance at the table level is also performed. This evaluation is based on metrics employed in [13] [14]. For clarity and completeness, these metrics are described here and adapted depending on the tables in study. Let T_G and T_D be elements representing respectively the ground truth and the detected table in each document. The amount of the overlapping between both elements is defined as:

$$O(T_G, T_D) = \frac{2|T_G \cap T_D|}{|T_G| + |T_D|} \quad (7)$$

where $|T_G \cap T_D|$ represents the number of lines of the intersection of the two tables, $|T_G|$ and $|T_D|$ denote the number of lines of the ground-truth and the detected table respectively. It is clear that the overlapping amount vary between 0 and 1. It measures the "correctness" of the detected table in comparison with the ground truth. Figure 6 shows the percentage of detected tables with an overlapping $O \geq s$; where s varies over the range [0,1]. Using this amount of the overlapping , the following metrics are defined:

- Correct: the number of detected tables that have an overlapping $O \geq 0.85$ with the corresponding ground-truth (see example in Figure 5(a)).
- Partial: the number of detected tables that have an overlapping $0.2 < O < 0.85$ with the corresponding ground-truth (see example in Figure 5(b)).
- False: the number of detected tables that do not have significant overlapping ($O < 0.2$) with any of the ground-truth tables (see example in Figure 5(c)).

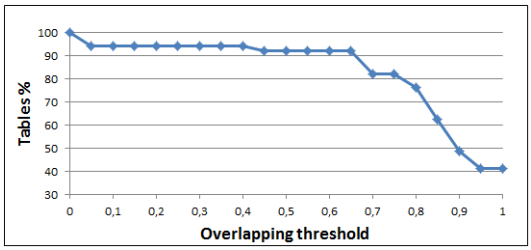


Figure 6. Percentage of detected tables for different values of overlapping threshold

- Missed: the number of ground-truth tables that do not have significant overlapping ($O < 0.2$) with the detected tables (see example in Figure 5(d)).

Table III shows the average error rates at the table level.

Table III
BLOCK LEVEL ERROR RATE

Total of documents	Correct	Partial	Missed	False
51 (containing 49 tables)	30	16	1	11

This is due to the rich set of features used in our approach and the high discriminative capability of the CRF. Whereas, the method proposed in [14] is based on the line correlation alone which, we think, is not very efficient to localize table in handwritten noisy documents.

V. Conclusion

In this work we have proposed a CRF model for line labeling in order to detect table lines. The model exploits contextual information by using features related to neighboring lines in addition to features specific to the current line. The presented model is a general framework which allows flexible use of many arbitrary, non-independent features and can be applied in many other labeling and recognition tasks. We have also presented robust line features for table detection. We have evaluated the efficiency of our method in real-world documents and the obtained results are promising.

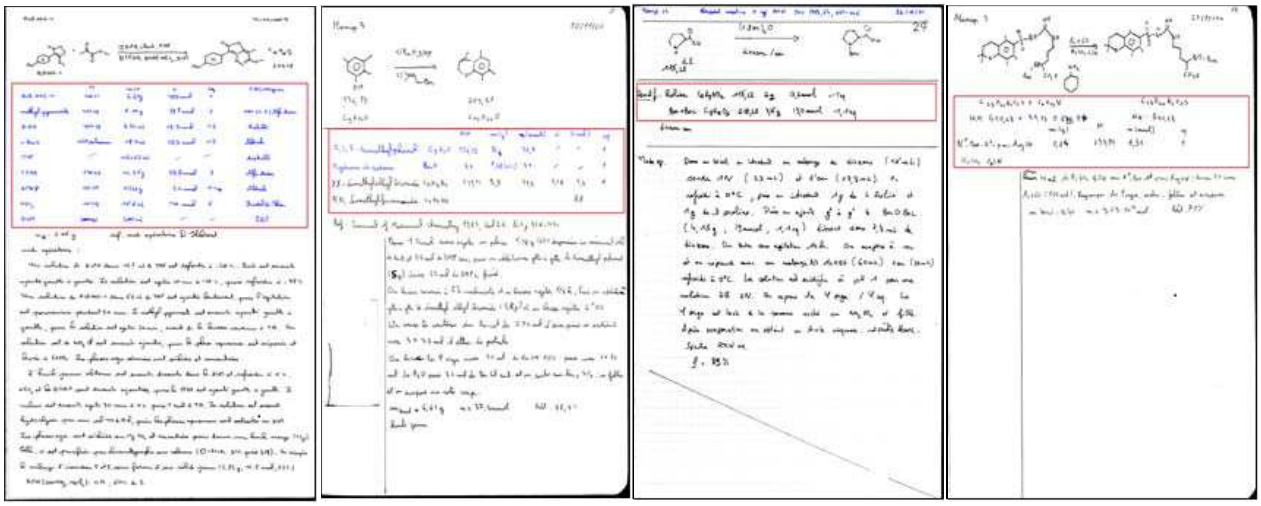
Future work concerns, in short-term, the automatic determination of the feature function weights which are fixed experimentally in the present work. Thus, no manual parameter setting is necessary. This allows an easy adaptation to different types of documents and different analysis tasks. Our future works include also the widening of the document database in order to test the approach in a big variety of documents. In addition, we plan to design more discriminative structural features for training and testing the model.

Acknowledgment

This work is carried out in the framework of **CIFRE**. We would like to thank **eNovalys** and **ANRT** for financing the work. We also thank eNovalys team for providing the document dataset.

Table II
AVERAGE LINE-LABELING RATES

	$\lambda_l = 1$ and $\lambda_c = 0$		$\lambda_l = 0$ and $\lambda_c = 1$		Using all features with one MLP		$\lambda_l = 0,7$ and $\lambda_c = 0,3$	
	Precision	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**
TextLine	93, 37	96, 33	95, 38	89, 65	94, 9	95, 5	95, 98	95, 66
TableRow	87, 85	79, 50	74, 15	87, 50	86, 5	85, 5	87, 13	88, 00
Weighted Avg	91, 98	92, 12	90, 07	89, 11	92, 8	92, 9	93, 76	93, 74



(a) Correct (b) Partial (c) False (d) Missed

Figure 5. An illustration of different performance measures. The ground truth is outlined in red and the detected table lines are colored in blue

REFERENCES

[1] J. Hu, R. Kashi, D. Lopresti, and G. Wilfong. *Medium-independent table detection*. in Document Recognition and Retrieval, pp. 44-55, 2000.

[2] J. Chen and D. Lopresti, *Ruling-based Table Analysis for Noisy Handwritten Documents*, in International Workshop on Multilingual OCR, pp. 1-5, 2013.

[3] T. Kasar, P. Barlas, S. Adam, C. Chatelain and T. Paquet, *Learning to Detect Tables in Scanned Document Images using Line Information*, in International Conference on Document Analysis and Recognition, pp. 1185-1189, 2013

[4] J. Chen and D. Lopresti, *Model-based Tabular Structure Detection and Recognition in Noisy Handwritten Documents*, in International Conference on Frontiers in Handwriting Recognition, pp. 75-80, 2012

[5] A. Laurentini and P. Viada. *Identifying and understanding tabular material in compound documents*. in International Conference on Pattern Recognition, pp. 405-409, 1992.

[6] L. A. Neves, J. M. De Carvalho, J. Facon, F. Bortolozzi, *A New Table Extraction and Recovery Methodology with Little Use of Previous Knowledge*, in International Workshop on Frontiers in Handwriting Recognition, 2006

[7] B. Gatos, D. Danatsas, I. Pratikakis and S. Perantonis, *Automatic table detection in document images*. in International Conference on Advances in Pattern Recognition, pp. 609-618, 2005.

[8] T. G. Kieninger, *Table Structure Recognition Based On Robust Block Segmentation*, in SPIE Conference on Document Recognition V, pp. 22-32, 1998

[9] D. Pinto, A. McCallum, X. Wei and W. B. Croft, *Table Extraction Using Conditional Random Fields*, in International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 235-242, 2003

[10] S. Nicolas, J. Dardenne, T. Paquet, and L. Heutte, *Document Image Segmentation Using a 2D Conditional Random Field Model*, in International Conference on Document Analysis and Recognition, pp. 407-411, 2007

[11] Y. Zheng , H. Li and D. Doermann, *The segmentation and identification of handwriting in noisy document images*, in Document Analysis System, pp. 95-105, 2002.

[12] N. Ghanmi and A. Belaid, *Extraction de formules chimiques dans des documents manuscrits composites*, accepted in Colloque International Francophone sur l'Ecrit et le Document, 2014.

[13] F. Shafait and R. Smith, *Table Detection in Heterogeneous Documents*, in International Workshop on Document Analysis System, pp. 65-72, 2010.

[14] J. Chen and D. Lopresti, *Table detection in noisy offline handwritten documents*, in International Conference on Document Analysis and Recognition, pp. 399-403, 2011.