

# Document Image Segmentation Using a 2D Conditional Random Field Model

Stéphane Nicolas<sup>1</sup>, Julien Dardenne<sup>2</sup>, Thierry Paquet<sup>1</sup>, Laurent Heutte<sup>1</sup>

<sup>1</sup> Laboratoire LITIS EA 4108, Université de Rouen, France

<sup>2</sup> CREATIS, INSA Lyon, France

## Abstract

*This work relates to the implementation of a 2D conditional random field model in the context of document image analysis. Our model makes it possible to take variability into account and to integrate contextual knowledge, while taking benefit from machine learning techniques. Experiments on handwritten drafts of Flaubert show that these models provide interesting solutions.*

## 1. Introduction

The valorization of our cultural heritage using digital technologies requires robust document image analysis methods allowing to recognize the structure and to detect areas of interest facilitating the indexing of large corpora of ancient documents and the production of digital libraries.

Due to the variability of ancient and handwritten documents, traditional analysis methods are not adapted and 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. In our previous work we have used MRF for document image labelling [1]. This approach has given interesting results but MRF models exhibit some limitations. Recently, Conditional Random Fields (CRF) have been proposed in order to avoid the limitations of the generative models.

The CRF were initially introduced in the field of information extraction by Lafferty and others [2] for part-of-speech tagging and syntactical analysis. Up to now they have been mainly used for sequential data modeling. Few works concerning 2D CRF models for image analysis have been proposed very recently [3][4], but to the best of our knowledge, except the work in [5] on diagram recognition and the work in [6] on handwritten word recognition, no application of the CRF to document image analysis have been proposed. The superiority of CRF models compared to MRF models has been generally reported for sequence

modeling. Contrary to MRF and other generative models which define a joint probability over observation and label configurations, what requires in theory the enumeration of all possible observation configurations for the calculation of the normalization constant, discriminative models like CRF directly model the conditional probabilities of label configurations given observations. Furthermore CRF models don't require any independence hypothesis about the observations and are particularly efficient for discriminative tasks like segmentation, labelling or recognition.

In this paper we propose to adapt and apply CRF to document image analysis. In section 2 we present the theoretical framework of CRF and we present our 2D model. In section 3 we discuss the learning of the model whereas section 4 is dedicated to the inference. Improvements of the model are described in section 5 where we investigate the integration of multiple contextual information sources. The results obtained with CRF model are given and discussed in section 6. It is shown that CRF compare favorably with traditional MRF. We conclude in section 7.

## 2. Proposed model

We consider the document image analysis as a labelling problem. Each document image is considered to be produced by implicit layout rules used by the author. As there are some local interactions between these rules, Random Fields appear to be adapted to model the layout of a document. 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 is denoted  $g(i, j), 1 \leq i \leq n \quad 1 \leq j \leq m$ . The set of sites 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  $Y = \{y(i, j), 1 \leq i \leq n \quad 1 \leq j \leq m\}$ .

Furthermore, considering that each state  $X_s$  of the 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 the most probable label configuration of the

field  $X$  among all the possible labelling  $E$  that can be associated to the image, i.e. finding:

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

Whereas the MRF model gives access to the posterior probability indirectly using the Bayes rule decomposition  $P(X/Y) \propto P(Y/X)P(X)$ , a CRF model does not use this decomposition and therefore provides a direct formulation of the discriminative task i.e. discrimination between the labels. The general form of CRF model is given by the following formula:

$$P(X = x/Y = y) = \frac{1}{Z} \prod_{s \in S} \exp \left( \sum_k \lambda_k f_k(x, y, s) \right)$$

A CRF model is defined as the product on a set of sites of the exponential of a linear combination of  $k$  functions called feature functions, depending on the observations  $y$ , the label configuration  $x$  and the current site  $s$ .  $Z$  is a normalization factor traditionally called the partition function.

As one can see, the main drawback of MRF models as opposed to CRF models is that they introduce two intermediate models (the data model via the likelihood function and the document model via the probability function of the label field) that are themselves difficult to estimate. Furthermore, generative models are known to be limited to using low dimension observations that are generally modeled using Gaussian Mixtures.

The modeling and the solving of problems using conditional random fields requires one to define the feature functions and to choose a parameter learning method and an inference method. We explain now the model we propose and our choices for these points.

### Feature functions

Like Kumar and others [3], and more recently He and others [4], we have chosen to model the feature functions  $f(x, y, s)$  by discriminative classifiers. We use Multilayer Perceptron (MLP) for this task because they are fast and provide good generalization properties even in high dimensional spaces. However we could consider other classifiers such as SVM or logistic classifiers. Our CRF model can be seen as a network of interconnected classifiers taking their decision using image features as well as contextual information by incorporating the decisions of the neighboring classifiers. The conditional probability of the label field is defined according to the following relation:

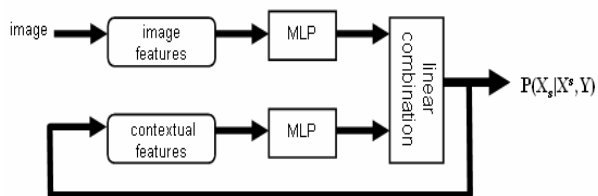
$$P(X/Y) = \frac{1}{Z} \prod \left[ \exp \left( \sum_k \lambda_k P(x_s / \mathcal{F}(x, y, s)) \right) \right]$$

Where  $k$  stands for discriminative components taken into account in the model, and the  $\lambda_k$  are the weights associated to these components. In a first time we have

considered only two levels of analysis ( $k = 2$ ) and we have defined two feature functions: a local feature function  $f_L$  and a contextual one  $f_C$ . The  $f_L$  function takes only into account the features extracted from the observations  $y$  on a local window of analysis. The  $f_C$  feature function takes into account contextual information i.e. the label probabilities over a neighborhood defined by a window of analysis. The local conditional probability  $P(X_s / X^s, Y)$  at each site  $s$  is defined by a linear combination (Figure 1) of the two feature functions  $f_L$  and  $f_C$ :

$$P(X_s / X^s, Y) = \lambda_L f_L + \lambda_C f_C$$

This formulation combines a local discriminative model and a contextual discriminative model what allows one to capture the informations issued from the observation field and the information issued from the label field  $X$  in a limited neighborhood. This model makes it possible to capture a rich context and thus allows a good regularization and homogenization of the label field  $X$ , while taking into account the observed information. Furthermore by considering a discriminative framework, it is possible to relax the conditional independence hypothesis of the observations. Then we can take into account correlated features on a wide neighborhood.



**Figure 1. Linear combination of local and contextual information**

### Local features

The local feature function takes only into account features extracted on the observed image, at a given site (local image features). This function models the data association that is the adequation between the label associated to a given site and the local observation at this site  $s$ . We take into account the same feature sets as those that we used with the markov random field model we have presented in [1], that is multiresolution pixel density features and site relative position. These features are extracted on each site and form a feature vector which feeds the input of the MLP which models the local feature function. The scores returned by the MLP are the values of the feature function for the different possible labels  $l_i \in L = \{l_1, \dots, l_q\}$  which can be associated to the current site  $s$ .

## Contextual features

The contextual feature function takes only into account the local conditional probability densities  $P(X_s = l_i, i = 1, \dots, q / X_N, Y)$  on the label field  $X$  in a neighborhood  $N$  around the current site. This neighborhood is determined by defining a sliding window the size of which depends on the quantity of contextual information we wish to integrate. For example, using a window of size  $3 \times 3$  and considering a label set of size  $q = 3$  we can define 27 conditional posterior probabilities, that is a vector of 27 contextual features applied at the input of the contextual MLP.

### 3. Parameter learning

To learn the parameters of the model consists in training the two MLP and determining the weights  $\lambda_L$  and  $\lambda_C$  of the linear combination. We use a supervised approach considering we have complete manually labeled data as is the case for the CRF framework: for each image of the training database, we have the corresponding groundtruth labelling. In this study, the labelling has been entered manually using a simple image editor and using a particular lookup table so as to associate a particular label to each color. Both MLP are trained using the backpropagation algorithm. The local MLP is trained first considering only the features extracted from the image. The output of the MLP is used to estimate the data association conditional probabilities  $P(X_s / Y_s)$  at each site  $s$  of the image. Then these conditional probabilities are used as input features for the training of the contextual MLP. The weights of the linear combination are determined simply using a gradient descent method in order to minimize the labelling error at each site.

### 4. Model inference

The techniques used for conditional random field model inference are similar to those proposed for Markov random field inference. In the 2D case, the model has a general graph structure. As a consequence, there is no exact inference method for such structure, and the search for a sub-optimal labelling solution is only possible. The most popular techniques for inexact inference on random field models are the Belief Propagation algorithm and the sampling techniques such as Gibbs or Metropolis sampler. Stochastic relaxation methods such as simulated annealing or ICM algorithm can either be used. These algorithms allow finding an approximate solution of the optimal field labelling using MAP criterion:

$$\hat{x} = \underset{x}{\operatorname{argmax}} P(X = x / Y = y)$$

As for markov random field models, we have chosen to use ICM (Iterated Conditional Modes) and HCF (Highest Confidence First) algorithm for the inference because they are known to be fast and efficient. The principle of the inference is the following. We proceed to a first labelling using only the local classifier and the intrinsic image features. During this first process, the contextual information about the labelling on the neighboring sites is not taken into account. This process allows one to initialize the label field and to compute at each site the values of the local feature function. These features are then used as inputs for the contextual classifier. For the next iterations the contextual feature function is also taken into account to evaluate the local potential function (log of the conditional probability) at each site of the image. The inference then consists in visiting all the sites and to evaluate for each of them the score of the potential function for each possible label  $l_i$  of the set  $L$ , by combining the outputs of the local classifier and the outputs of the contextual classifier. This score can be seen as the probability of assigning the label  $l_i$  to the site  $s$  given the local observations and the probabilities of the labels on the vicinity. The label providing the highest score is assigned to the current site, but all the others conditional probabilities are memorized. These probabilities are iteratively updated during the inference of the label field. This updating process is repeated until convergence of the label configuration.

### 5. Integration of more contextual information

We have first considered only two analysis levels, and we present the results obtained with this implementation in the section 6, but the global formulation of our model allows one to integrate easily more information sources in the decision. So a second model is also proposed by integrating a third feature function called global feature function.

#### Global feature function

An analysis of the label field  $X$  at a more global level is taken into account by a third feature function called global feature function. This global analysis is carried out using a third MLP. This classifier estimates the posterior probabilities  $P(X_s = l_i, \forall l_i \in L / \mathcal{F}_G(X))$  of associating the label  $l_i$  to the current site  $s$  given a set  $\mathcal{F}_G$  of global statistical features extracted on the global label configuration over a larger neighborhood than that taken into account by the contextual feature function. The global classifier is also a contextual

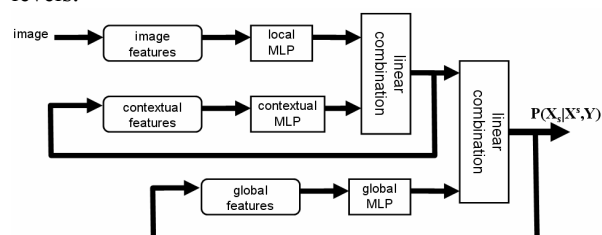
classifier that takes into account the label configuration at a coarse resolution. At this resolution, the label field is divided into several zones by superposing a grid  $H$  larger than the initial grid  $G$ . Each cell of this grid gives access to a set of sites. Statistical parameters are computed on these cells. We construct the co-occurrence matrix of the labels on each cell for different orientations. More precisely, four co-occurrence matrices are calculated for the orientations  $0^\circ$ ,  $45^\circ$ ,  $90^\circ$  and  $135^\circ$ . From these four co-occurrence matrices, five Haralick parameters are computed. Here the originality of our approach is that we determine these features not directly on the image, but on the label configuration.

### Combination of the information sources

This second model integrates three information sources operating at three different levels of analysis. Each of them is modelled by a MLP. At each image site and for each of these three information sources the MLP classifier takes a decision according to the source taken into account. The local conditional probability  $P(X_s/X^s, Y)$  in each site  $s$ , given the observations  $Y$  and the remainder of the field  $X$  is now defined by  $P(X_s/X^s, Y) = h(f_L, f_C, f_G)$  where  $h$  is a combination function of the three information sources: local, contextual and global. In practice there are several manners to implement this combination function  $h$ . We propose two combination solutions.

#### Linear combination of the information sources

The simplest way to integrate the global feature function to our CRF model is to combine it with the output of the previous model. In this case the model integrates in a sequential manner two combination levels.



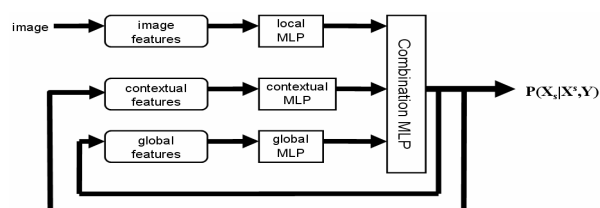
**Figure 2. Cascading combination of the three levels of analysis**

The output is then a non-linear combination (Figure 2). The training of the weights of these two combinations is performed in two steps. First the weights  $\lambda_L$  and  $\lambda_C$  of the first combination are trained on the training set in order to maximize the average labelling rate without taking into account the global information. These weights being optimized, those of the second combination  $\lambda_{L-C}$  and  $\lambda_G$  are determined using the same procedure but taking the global information into

account. This type of combination allows one to control easily the quantity of information given by the global feature function compared to the information given by the local model. The first combination summarizes local information whereas the second one integrates global information.

#### Combination of the information sources using a MLP

We have also investigated the use of a Multilayer Perceptron as a combination function of the different information sources (Figure 3). Indeed the theoretical and mathematical definition of a MLP is rather close to that of a conditional random field since it acts as a non-linear combination of features. The values of the output of the three feature functions for the different possible labels, feed the input of one single MLP. If the label set contains  $q$  labels, the dimension of the feature vector applied at the input of the combining MLP will be  $3q$ . Using this type of combination the three information sources are combined in parallel.



**Figure 3. Combination of the three levels of analysis using a MLP**

The quantity of information provided by each component depends on the weights of the MLP, but cannot be known explicitly, as that is the case when using linear combination. The learning of this model consists in training the three MLP modeling the three feature functions, on the training dataset, hence in training the combining MLP. The training of all the MLP is performed using the backpropagation algorithm. The advantage of this solution is that there is only one combination level, so it is very simple to combine several information sources. Its main drawback is the prohibitive time required for training the MLP.

## 6. Experiments and results

For the experiments we have used a dataset of 69 images of Gustave Flaubert's manuscripts. We consider a labelling task at a block level which consists in detecting large area of interest. A model with six states is defined for this task, and the parameters are learned on manually labeled images. The states are "text body", "text blocks", "page numbers", "margins", "headers" and "footers". The size of the grid is fixed empirically to obtain a good compromise between

complexity reduction and quality of the labelling for the considered task. We choose a size of 50\*50 pixels which corresponds roughly to the width of the inter-word spaces and to the height of the ascending or descending letters. The Tab. 1 shows the labelling rates obtained at the pixel level with the three implementations of our CRF model, where impl.1 stands for the linear combination of local and contextual features, impl.2 for the linear combination of local, contextual and global features and impl.3 for the combination of these three levels with a MLP. Comparison with the results obtained using a local classifier applied at each site using different sizes of context (the set 1 corresponds to a 3\*3 contextual window and the set 2 to a 5\*5 window) is also given. The results show that by increasing the number of levels of analysis (impl.3 versus impl.1) we obtain better results than when using a single local classifier. Among the different implementations the MLP combination (impl.3) seems to be better.

	local MLP	impl. 1	impl. 2	impl. 3
set 1	90.56	92.55	93.90	94.04
set 2	90.56	93.91	93.93	94.16

**Tab. 1 Average labelling rates obtained with different implementations of our CRF model and using ICM inference**

The Tab. 2 compares the average pixel labelling rate (ALR) provided by our CRF model and the average labelling rate obtained when using our previous MRF model described in [1] and local generative (gaussian mixtures) and discriminative (MLP) classifiers. These results show that discriminative CRF models outperform the traditional generative MRF models considering a general image labelling task. We can see on Figure 4 an example of labelling result.

	Gaussian mixture	local MLP	MRF	CRF
ALR (%)	83.70	87.50	90.56	93.91

**Tab. 2 comparison of labelling rates obtained with different models**

## 7. Conclusion and future work

In this work we have proposed a Conditional Random Field model for 2D data labelling, in particular for document image segmentation. One of the main advantages of this model is that it can be learned automatically using machine learning procedures, so no manual parameter setting is necessary. This allows an easy adaptation to different types of documents and different analysis tasks. The results we have obtained on Flaubert's manuscripts show that the proposed model provides better results than MRF generative models. These results are similar to those presented in

other recent work on conditional random fields. Future work concerns the integration of more intrinsic and contextual features in the model, the replacement of MLP by logistic classifiers faster to train, and the definition of hierarchical CRF models for document image analysis.



**Figure 4. Example of labelling result on a manuscript of Flaubert**

## 7. References

- [1] S. Nicolas, T. Paquet and L. Heutte, "A markovian approach for handwritten document segmentation", in proceedings of the 18<sup>th</sup> IEEE International Conference on Pattern Recognition, (ICPR 2006), volume 3, pages 292-295, 2006.
- [2] J. Lafferty, F. Pereira and A. McCallum, "Conditional random fields: Probabilistic models for segmenting and labelling sequence data", in proceedings of the International Conference on Machine Learning (ICML'01), pages 282-289, 2001.
- [3] S. Kumar and M. Hebert, "Discriminative random fields: A discriminative framework for contextual interaction in classification", in proceedings of the 9<sup>th</sup> IEEE International Conference on Computer Vision (ICCV'03), volume 2, pages 1150-1159, 2003.
- [4] X. He, R.S. Zemel and M.A. Carreira-Perpinan, "Multiscale Conditional Random Fields for Image Labelling", in IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04), volume 2, pages 695-702, 2004.
- [5] M. Szummer and Y. Qi, "Contextual Recognition of Hand-drawn Diagrams with Conditional Random Fields", In 9<sup>th</sup> International Workshop on Frontiers in Handwriting Recognition (IWFHR'04), pages 32-37, Tokyo, Japan, 2004.
- [6] S. Feng, R. Manmatha and A. McCallum, "Exploring the Use of Conditional Random Field Models and HMMs for Historical Handwritten Document Recognition", Proceedings of the 2<sup>nd</sup> International Conference on Document Image Analysis for Libraries (DIAL'06), pages 30-37, Lyon, France, 2006.