# Extraction de structure de documents manuscrits non contraints par Champs Aléatoires Conditionnels 2D

Florent MONTREUIL[1], Emmanuèle GROSICKI[1], Stéphane NICOLAS[2], Laurent HEUTTE[2]

[1]DGA/Centre d'Expertise Parisien 16, bis avenue Prieur de la Côte d'or 94114, Arcueil cedex, France

[2]Université de Rouen, LITIS EA 4108 BP 12 - 76801, Saint-Étienne du Rouvray, France

Florent.Montreuil@gmail.com, Emmanuele.Grosicki@etca.fr
Stephane.Nicolas@univ-rouen.fr, Laurent.Heutte@univ-rouen.fr

**Résumé –** Cette article décrit une nouvelle approche utilisant des Champs Aléatoires Conditionnels (CACs) pour extraire la mise en page de documents manuscrits non contraints. Dans cette approche, l'extraction de la mise en page est considérée comme une tâche d'étiquetage consistant à assigner une étiquette à chaque pixel de l'image du document. Le modèle CAC donne directement accès à la probabilité conditionnelle globale d'un étiquetage de l'image sachant des caractéristiques image et des connaissances *a priori* sur la structure du document modélisées. Pour déterminer l'étiquetage optimal, un point clé de notre modèle est l'implémentation de l'algorithme d'inférence optimal de Programmation Dynamique 2D. Ce modèle a été testé sur 1250 lettres manuscrites de la base RIMES. De bons résultats ont été obtenus montrant la capacité de cette approche à extraire la mise en page d'un document complexe à partir d'informations de différentes natures. (morphologiques, spatiales, ...)

**Abstract –** The paper describes a new approach using Conditional Random Fields (CRFs) to extract layouts in unconstrained handwritten documents. 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. The CRF model gives access to the global conditional probability of a given labeling of the image according to some image features and some prior knowledge about the structure of the document. 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 layout on a complex document from informations of different natures. (morphological, spatial, ...)

## 1 Introduction

De nombreuses applications telles que le traitement de courriers par des entreprises ou des administrations nécessitent de pouvoir trier automatiquement des grands volumes de données contenant des écritures manuscrites. Il s'agit d'une tâche difficile et non encore résolue car elle nécessite non seulement le développement de systèmes fiables de reconnaissance de l'écriture manuscrite mais également le développement de méthodes robustes d'analyse automatique de la mise en page. Celle-ci est usuellement faite en deux étapes séquentielles : extraction de la structure physique (segmentation) puis extraction de la structure logique (étiquetage ou reconnaissance des entités segmentées). La première étape (segmentation) consiste à découper l'image du document en blocs, lignes, mots, alors que la seconde étape (étiquetage) vise à regrouper ces différents segments pour former des unités logiques auxquelles on associe des étiquettes donnant la fonction de ces unités dans le document (par exemple le bloc date ou objet dans les courriers manuscrits). Toutefois, dans le cas où le document est difficile à segmenter (par exemple les documents manuscrits non contraints), il apparait que la structure logique contient des informations pouvant améliorer l'extraction de la structure physique. Dans cet objectif, il semble plus optimal d'extraire ces deux structures conjointement plutôt que sequentiellement. Pour cela, nous proposons une approche statistique exploitant des modèles de type Champs Aléatoires Conditionnels (CAC) permettant de combiner des caractéristiques décrivant à la fois la structure physique et la structure logique.

Dans l'état de l'art, il existe peu de travaux sur l'analyse de la mise en page des documents manuscrits. On peut néanmoins distinguer deux grandes catégories de modèles pour résoudre ce problème : les modèles à base de règles et les modèles statistiques. Les modèles de la première catégorie définissent un grand nombre de règles pour englober toutes les structurations possibles [5]. Ces règles bien que nombreuses ne permettent pas de contrôler toute la variabilité de ces documents. Elles créent donc des exceptions qui provoquent des erreurs de segmentation. La deuxième catégorie correspond aux modèles statistiques. On peut citer deux types de modélisation : les modèles génératifs avec les modélisations par Champs Aléatoires Markovien (CAM) [4] et les modèles discriminants avec les modélisations par CAC [2]. On se propose ici de travailler sur un modèle CAC car ces derniers semblent plus appropriés à une tâche d'étiquetage, de part sa nature discriminante. En effet, le problème d'étiquetage se rapproche d'un problème de

discrimination entre classes et les CAC modélisent directement la probabilité *a posteriori* d'un étiquetage sachant des observations. De plus, les CAC possèdent de nombreux avantages comme celui d'intégrer plus facilement différents niveaux de contexte (voir 2) ou celui de tenir compte de l'ensemble des observations faites sur l'image sans hypothèse d'indépendance entre ces dernières.

# 2 Modélisation par Champs Aléatoires Conditionnels (CAC)

Dans ces approches, la mise en page du document est supposée produite par un champ d'états cachés noté $Y$ prenant des valeurs dans un ensemble fini d'états $L$. Ce champ est supposé Markovien ce qui signifie qu'il y a une dépendance conditionnelle au voisinage. Chaque état du champ est associé à un site $s$ (ensemble de pixels) de l'image auquel sera affecté l'étiquette de l'état correspondant. Chaque étiquette de ces états est estimée conditionnellement aux états voisins mais aussi aux observations $X$ extraites dans l'image entière.

Dans une approche CAC, la probabilité d'une configuration $y$ du champ d'états sachant un ensemble d'observation $x$ est directement donnée par le modèle (pas de transformation par le théorème de Bayes). Donc, pour obtenir la configuration d'états optimale (mise en page) $\hat{y}$, nous cherchons la configuration $y$ dans l'ensemble des configurations possibles $\mathcal{Y}$ qui maximise la probabilité conditionnelle :

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

Cette probabilité globale *a posteriori* est définie classiquement dans les modèles CACs comme le produit sur un ensemble de sites $s$, de l'exponentielle d'une somme pondérée de $k$ fonctions appelées fonctions de caractéristiques $f_k$ :

$$P(Y = y|X = x) = \frac{1}{Z} \prod_s (exp(\sum_k \theta_k f_k(x, y, s))) \tag{2}$$

Où $Z = \sum_L \prod_s (exp(\sum_k \theta_k f_k))$ est un coefficient de normalisation sur l'ensemble des étiquettes $L$ possibles.

Ces fonctions de caractéristiques dépendent des observations $x$, de la configuration d'étiquettes $y$ et du site courant. Ce sont des fonctions à valeurs réelles à travers lesquelles toutes les connaissances du modèle sont intégrées. Ces fonctions sont pondérées par des paramètres $\theta_k$ (paramètres du modèle CAC) permettant de régler l'importance des connaissances introduites par ces fonctions dans le modèle. Ces fonctions sont extraites à différents niveaux d'analyses $k$. Chaque niveau d'analyse défini un contexte d'informations différent en fonction : du type d'observations (image ou étiquette), de l'échelle ou du voisinage, du type de caractéristiques (spatiales, textuelles, morphologiques ...).

Dans notre approche, nous avons choisi de modéliser ces fonctions de caractéristiques par des classifieurs discriminants comme proposé dans [2]. Le modèle CAC peut alors être vu comme un réseau de classifieurs inter-connectés prenant leurs décisions en fonction de caractéristiques extraites sur le niveau d'analyse $k$ considéré. La sortie de chaque classifieur nous permet d'obtenir une estimation de la probabilité locale *a posteriori* des étiquettes sachant les caractéristiques d'entrée. La probabilité conditionnelle globale du modèle s'écrit alors :

$$P(Y = y|X = x) = \frac{1}{Z} \prod_s \left( exp(\sum_k \theta_k P_k(y_j|F(y, x, j))) \right) \tag{3}$$

Où $F(y, x, j)$ sont les caractéristiques prises en entrée des classifieurs définies pour chaque niveau d'analyse. (voir 2.2)

Parmi les classifieurs existants, nous avons choisi des SVMs (Séparateurs à Vastes Marges) car ils possèdent des bonnes propriétés de généralisation comparés aux classifieurs conventionnels. De plus, la combinaison SVM/CAC est très précise car elle bénéficie de la nature des SVMs à rechercher des hyperplans séparateurs de marges maximums et aussi de la nature des CACs à modéliser la corrélation entre étiquettes voisines [6].

On définit dans notre modèle 3 niveaux de contexte (voir Fig. 1). Un premier niveau local qui nous permet d'intégrer les observations image. Puis deux niveaux qui régularisent la probabilité locale en prenant en compte les étiquettes voisines (dépendance Markovienne) comme proposé dans [2] :

- **niveau local** : estime une probabilité local $P_L(y_j|X)$ d'un état $y_j$ sachant un ensemble de caractéristiques continues extraites de l'image. (voir Fig. 1).
- **niveau contextuel** : estime la probabilité des étiquettes en fonction des étiquettes mitoyennes probables noté $Y_c$, on obtient une probabilité $P_C(y_j|X, Y_c)$
- **niveau global** : estime la probabilité des étiquettes en fonction des étiquettes comprises dans voisinage plus large noté $Y_g$, on obtient une probabilité $P_G(y_j|X, Y_g)$.

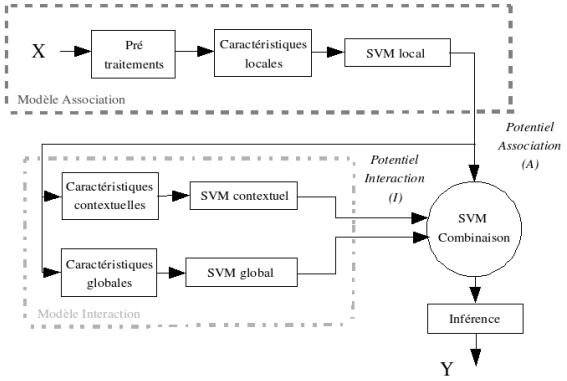


FIG. 1 – Notre modèle CAC

## 2.1 L'apprentissage

L'un des principaux avantages du modèle CAC que nous proposons réside dans l'utilisation de procédures d'apprentissage,

permettant ainsi une adaptation relativement rapide et aisée à différents types de documents et à différentes tâches d'analyse. L'apprentissage est réalisé de manière supervisée, il consiste d'une part à entraîner les différents SVMs pour chaque niveau puis dans un deuxième temps à déterminer les paramètres de la combinaison des différents classifieurs [6]. Pour déterminer ces paramètres, un classifieur SVM supplémentaire est utilisé prenant en entré les sorties des classifieurs de chaque niveau $k$ (voir Fig. 1). Les poids $\theta_k$ associés à chaque niveau sont réglés par la procédure d'apprentissage et ne sont pas connus explicitement. Cette méthode possède l'avantage de ne demander qu'un ré-apprentissage de ce dernier classifieur si l'on souhaite ajouter des niveaux de caractérisations supplémentaires. Cependant, il est difficile de choisir les paramètres du SVM (en particulier le paramètre $C$ de pénalité) et par conséquent la procédure d'apprentissage peut être lourde. Pour fixer ces paramètres, la méthode la plus commune (réalisée pour les $k+1$ classifieurs) est d'effectuer une validation croisée sur une base de données indépendante des bases d'apprentissage et de test.

### 2.2 Choix des caractéristiques

On choisit pour le niveau local un classifieur SVM qui prend ses décisions en fonction de caractéristiques spatiales et morphologiques observées sur l'image. L'un des atouts de notre modèle est de pouvoir combiner des caractéristiques décrivant différents niveaux d'information : la structure physique (caractéristiques morphologiques) et la structure logique (caractéristiques spatiales). On obtient un vecteur de caractéristiques comprenant :

- Les coordonnées normalisées en abscisse et en ordonnée du centre de chaque site dans l'image [4].
- Les densités de pixels de 27 fenêtres réparties sur trois échelles, soit 3x9 fenêtres de tailles respectivement égales à 1, 5 et 9 sites. Pour chaque échelle les fenêtres sont regroupées sous forme d'un masque 3x3 centré sur le site courant (voir Fig. 2). Cela permet d'obtenir une représentation multi-échelle des niveaux de gris.

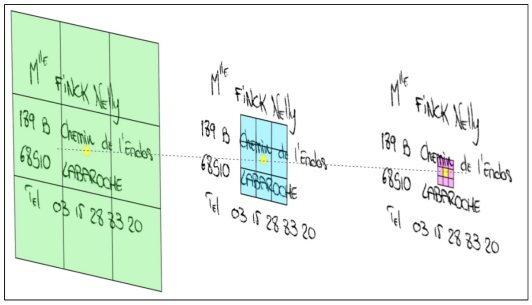

FIG. 2 – Caractéristiques de densité de niveau de gris extraites sur 3 échelles.

Pour le niveau contextuel, on cherche à utiliser les informations comprises dans un voisinage proche pour régulariser le modèle d'association. Pour chaque état, nous prenons comme caractéristiques les probabilités obtenues en sortie du classifieur local. Nous nous limitons à un voisinage 4-connexes pour des raisons de complexité de calcul. Le modèle global nous permet de considérer un voisinage plus large. Pour chaque état, nous prenons comme caractéristiques les probabilités d'occurrence de chaque étiquette dans une fenêtre centrée sur l'état considéré.

### 2.3 Inférence

Ces différents niveaux décrits par $P_L$, $P_C$ et $P_G$ sont combinés par l'intermédiaire d'un classifieur SVM pour obtenir la probabilité conditionnelle $P(y_j|X,\ Y_c, Y_g)$ pour chaque état. Pour déterminer la configuration optimale d'états correspondant à la mise en page du document, nous adaptons l'algorithme de programmation dynamique 2D proposé dans [4] pour les CAM. Cet algorithme d'inférence est une extension naturelle de l'algorithme classique 1D. Il possède l'avantage d'être optimal et rapide comparé aux algorithmes d'inférence classiques [3]. Il reprend la stratégie de « Diviser pour mieux régner ». Chaque état de la grille est fusionné à une région pour laquelle une liste de configurations possibles est calculée correspondant au produit des probabilités conditionnelles de chaque état de la région. L'opération est ré-itérée jusqu'à ce qu'il n'y ait plus qu'une région $R$ recouvrant tout le document. La configuration de la région $R$ dont la probabilité est maximale correspond à la mise en page du document.

## 3 Expérimentation

Nous avons testé notre modèle sur 1250 lettres manuscrites de la base de lettres RIMES. La tâche consiste à étiqueter les différentes parties (blocs) du document correspondant à l'une des étiquettes suivantes : Coordonnées Expéditeur, Date Lieu (DL), Coordonnées Destinataire (CD), Objet, Ouverture, Corps de Texte, Signature (voir Fig. 3).

Nous comparons nos résultats à ceux obtenus durant la seconde campagne d'évaluation RIMES de Juin 2008. La métrique $Err$ utilisée lors de cette campagne d'évaluation correspond à un taux d'erreur de classification défini par la somme des pixels noirs mal classés normalisés par la somme de tous les pixels noirs. Les résultats présentés dans le tableau 1 montre les taux d'erreurs obtenus par 3 systèmes au cours de la deuxième campagne d'évaluation RIMES qui sont comparés à ceux obtenus par notre modèle. Lab1 et lab3 propose une approche statistique basée sur des CAM tandis que le lab2 propose un système fondé sur des règles. L'approche proposée par le lab2 donne de bons résultats mais comme explicité précédemment,

TAB. 1 – Taux d'erreur obtenus à la seconde campagne d'évaluation RIMES et notre modèle.

| | lab1 [4] | lab2 [5] | lab3 | Notre modèle |
|---|---|---|---|---|
| $Err$ (%) | 8,53 | 8,97 | 12,62 | 9,51 |

l'utilisation de règles est une solution très rigide pas facilement adaptable à d'autres types de documents sinon par la définition de nouvelles règles. De plus, l'analyse des erreurs obtenus par cette approche montre un fort taux d'erreur sur certaines images dont la mise en page ne correspond pas aux règles de grammaire prédéfinies. Concernant notre modèle, les résultats obtenus (voir Fig. 3) peuvent être comparés à ceux du lab3 qui utilise aussi une modélisation statistique avec le même type de caractéristiques. Les deux méthodes diffèrent dans le choix de la modélisation (CAC pour notre modèle et CAM pour le lab3) et de l'utilisation de l'algorithme d'inférence optimal de programmation dynamique 2D pour notre approche. Le modèle proposé par le lab1 est lui aussi un modèle CAM. La différence entre les résultats obtenus par le lab1 et le lab3 s'explique par l'ajout dans le modèle du lab1 d'un post-traitement permettant de corriger les erreurs générées par le modèle CAM. Ce post-traitement est réalisé à base de règles dans lesquelles sont introduites des informations de plus haut niveau que celles utilisées dans notre modèle et dans celui du lab3.

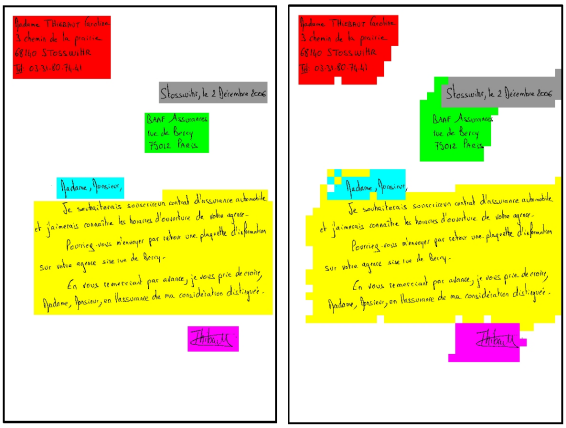

FIG. 3 – Résultat de mise en page, à gauche la vérité terrain à droite la mise en page automatique.

## 4 Conclusion et perspectives

Un modèle CAC 2D pour l'extraction de la mise en page de documents manuscrits non contraints a été proposé et discuté dans le présent document. Cette approche montre que si nous disposons de données étiquetées, il est possible d'apprendre et d'extraire les blocs d'un document. Il est possible facilement d'étendre ce modèle à d'autres types de documents en effectuant un nouvel apprentissage des classifieurs. Dans cette approche, nous avons utilisé de simples caractéristiques de textures et de positions. Un avantage de cette modélisation est de pouvoir combiner des informations de natures différentes décrivant la structure physique et logique de la mise en page. Les premières expériences sur la base de données RIMES montrent de bons résultats même avec peu de caractéristiques. Il reste cependant des erreurs d'étiquetage dans les régions où le niveau de caractérisation n'est plus suffisant. En effet, des confusions apparaissent aux frontières des blocs mitoyens lorsque les caractéristiques de position ne sont plus suffisantes pour discriminer les étiquettes.

Nous avons utilisé dans notre approche des informations spatiales et graphiques. Ces informations sont liées dans un document manuscrit aux informations textuelles « de manière entrelacées et inséparables », [1]. Pour cela, nous nous proposons d'ajouter dans notre modèle, des caractéristiques de haut niveau basées sur le contenu textuel. L'idée proposée consiste en la détection de certains mots clés pour améliorer la segmentation du document. Par exemple sur la Fig. 4, certains sites du bloc DL se trouve confondu avec des sites CD. La détection du mot « Novembre » se référant au bloc DL nous permettrait de corriger certaines erreurs d'étiquetage. Ces informations peuvent être facilement incorporées dans notre modèle, nous allons étudier cette solution dans nos futurs travaux.

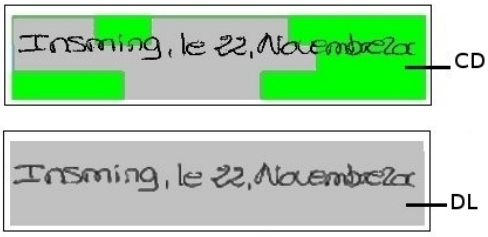


FIG. 4 – Confusion entre les blocs DL et CD, en bas la vérité terrain, en haut la segmentation automatique.

## Références


[1] A. Crasson, J.D. Fekete. *Structuration des manuscrits : Du corpus à la région.* Colloque International Francophone sur l'Ecrit et le Document (CIFED04), pages 162-168, 2004.

[2] S. Nicolas, J. Dardenne, T. Paquet, L. Heutte. *Un modèle de champ aléatoire conditionnel 2D appliqué à la segmentation d'images de documents.* 6e congrès francophone AFRIF-AFIA, Reconnaissance des Formes et Intelligence Artificielle, (RFIA08), 2008.

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

[4] M. Lemaitre. *Approche markovienne bidimensionnelle d'analyse et de reconnaissance de documents manuscrits.* Université René Descartes, Paris 5, 2007.

[5] A. Lemaitre, J. Camillerapp, B. Coüasnon. *Multiresolution cooperation makes easier document structure recognition.* International Journal on Document Analysis and Recognition, volume 11, numéro 2, 2008.

[6] G. Hoefel, C. Elkan. *Learning a two-stage SVM/CRF sequence classifier.* CIKM '08 : Proceeding of the 17th ACM conference on Information and knowledge management, pages 271-278, 2008.