

Utilisation de modèles markoviens 2D pour la segmentation d'images de documents

S. Nicolas* J. Dardenne[†] T. Paquet* L. Heutte*

* Laboratoire LITIS EA 4108, Université de Rouen, France
{Stephane.Nicolas, Thierry.Paquet, Laurent.Heutte}.univ-rouen.fr

[†] CREATIS, INSA Lyon, France
julien.dardenne@creatis.insa-lyon.fr

Résumé

Nous nous intéressons dans ces travaux à l'analyse d'images par le biais de modèles markoviens génératifs et discriminants : les champs de Markov et les champs aléatoires conditionnels. Pour les champs conditionnels nous présentons une implantation 2D de ces modèles principalement utilisés jusqu'à présent pour l'analyse de données monodimensionnelles. Le modèle que nous proposons repose sur une approche de type combinaison de classifieurs discriminants. Nous illustrons et comparons les capacités de cette approche par rapport aux modèles de Champs de Markov Cachés au travers d'un exemple d'application à l'analyse de la structure de documents manuscrits complexes et dégradés : les manuscrits de Flaubert.

Abstract

In this work we are interested in image analysis thanks to the use of Markovian generative and discriminative models : Hidden Markov Random Fields and Conditional Random Fields. We present a 2D implementation of conditional random fields which have been used mainly for on dimensional data until now. The proposed approach is based on a discriminative classifiers combination scheme. We illustrate and compare the performance of this approach with Hidden Markov random fields through its application to document structure analysis of complex degraded Handwritten Documents : Flaubert's manuscripts.

1. Introduction

La valorisation de notre patrimoine culturel par le biais de technologies numériques requiert des méthodes d'analyse d'images de documents robustes permettant de reconnaître les structures du document. Il s'agit généralement de détecter les régions d'intérêt pour faciliter (éventuellement automatiser) l'indexation de vastes corpus de documents anciens et contribuer ainsi à l'enrichissement des bibliothèques numériques.

Les modèles stochastiques ont prouvé leur capacité à prendre en compte la variabilité et à s'affranchir des ambiguïtés de l'écriture manuscrite. Par exemple les modèles de Markov cachés sont communément utilisés avec un certain succès pour la segmentation et la reconnaissance de données séquentielles. De manière similaire, dans le cas de l'analyse d'image, les champs de Markov (Markov Random Fields ou MRF) ont montré qu'il s'agit de modèles stochastiques puissants pour l'analyse de données bidimensionnelles. Dans cet article nous présentons une application de ces modèles markoviens à la segmentation de manuscrits de Gustave Flaubert, en leurs parties élémentaires, et à la détection de régions d'intérêt (Figure 1).

Dans le cadre théorique des champs de Markov, la segmentation est appréhendée comme un problème d'étiquetage d'image. Ce problème peut être résolu en utilisant des techniques d'optimisation, et est généralement appelé décodage de l'image. Nous décrivons dans la section 2 comment les modèles de champs de Markov cachés peuvent être utilisés dans le contexte de l'analyse d'images de documents complexes. Nous décrivons le modèle proposé, ainsi que les procédures utilisées pour l'apprentissage des paramètres et pour l'étiquetage de l'image en utilisant différentes techniques de décodage.

Nous explorons ensuite dans la section 3 l'utilisation des champs conditionnels qui s'avèrent être des modèles discriminants plus puissants pour de telles tâches de segmentation. Cette section traite du cadre théorique de ces modèles et décrit le modèle conditionnel que nous proposons. Nous comparons ensuite les deux modèles sur la même tâche de segmentation d'image de document.

2. Modélisation par champs de Markov

On peut considérer que les manuscrits d'auteurs, comme par ailleurs n'importe quel document, sont régis par des règles de mise en forme utilisées implicitement par l'auteur du document. La différence avec les documents imprimés réside notamment dans le fait que ces règles de mise en forme sont purement implicites dans le cas des documents manuscrits car généralement pas explicitées par l'auteur. Ainsi par exemple dans le cas des manuscrits de Flaubert, même si ces règles ne peuvent pas être formellement justifiées, il est pourtant admis et vérifié de manière empirique par les chercheurs en génétique littéraire que les manuscrits de Flaubert présentent certaines règles de mise en page caractéristiques. Dans les manuscrits de Flaubert, elles se traduisent notamment par un important corps de texte occupant les deux tiers de la page et présentant un nombre important de ratures, et par une marge bien définie occupant le tiers restant de la page et comportant de nombreuses annotations textuelles et corrections comme on peut le voir sur la figure 1.

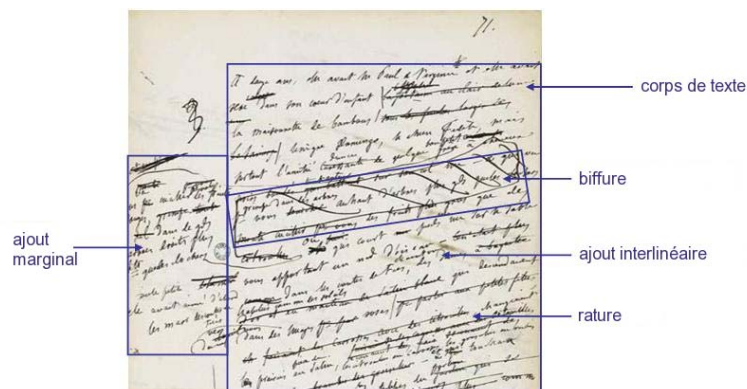


FIG. 1 - Un exemple de mise en page typique des manuscrits de Flaubert

Dans la mesure où la mise en page finale du document est le résultat de la combinaison de ces différentes règles implicites, les champs de Markov apparaissent être adaptés à modéliser de telles mises en page en utilisant des règles de dépendances locales. De plus, de tels modèles stochastiques semblent adaptés à prendre en compte la variabilité spatiale locale dans la disposition des différents éléments de la mise en page.

2.1. Fondements théoriques

2.1.2 Modélisation par champs de Markov

Selon le formalisme des champs de Markov, l'image est associée à une grille rectangulaire G de taille $n \times m$. Nous appelons "site", chaque cellule sur la grille G . Un site s est donc défini par ses coordonnées sur la grille G , et est noté $s(i, j)$, $1 \leq i \leq n$ $1 \leq j \leq m$. L'ensemble des sites est noté $S = \{s / s(i, j), 1 \leq i \leq n$ $1 \leq j \leq m\}$.

Conformément au cadre stochastique des champs de Markov cachés, l'image donne accès à un ensemble d'observations sur chaque site de S , noté $O = \{o(i, j), 1 \leq i \leq n$ $1 \leq j \leq m\}$. De manière similaire chaque site est associé à une variable aléatoire notée X_s . L'ensemble des variables aléatoires sur toute la grille définit le champ de Markov caché X .

Chaque variable aléatoire X_s prend alors une valeur dans un ensemble discret et fini de q étiquettes ou états noté $L = \{l_i\}$, $1 < i < q$, et correspondant à une règle de mise en page particulière ou à un certain motif ou élément de la mise en page. Le problème de l'extraction de la mise en page de l'image du document peut alors être formulé comme le problème consistant à trouver la configuration d'états ou d'étiquettes la plus probable parmi tous les étiquetages possibles E sur le champ des étiquettes X , pouvant être associés à l'image. Dans un cadre stochastique ou probabiliste cela revient à trouver l'étiquetage qui maximise la probabilité *a posteriori* de X étant données les observations O :

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

D'après la règle de Bayes nous avons :

$$P(X / O) = \frac{P(O / X)P(X)}{P(O)}$$

La probabilité *a priori* $P(O)$ est une constante qui est indépendante de la configuration E du champ d'étiquettes X , nous pouvons donc écrire que :

$$P(X / O) \propto P(X, O) = P(O / X)P(X)$$

Ainsi trouver la segmentation la plus probable de l'image est équivalent à trouver l'étiquetage qui maximise la probabilité jointe :

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

ce qui aboutit à la formule suivante lorsque l'on applique les hypothèses markoviennes et les hypothèses d'indépendance des observations :

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

Dans cette formule $N_G(s)$ désigne le voisinage du site s .

Alors que dans cette expression le terme $P(o_s | x_s)$ peut être calculé en utilisant des mélanges de gaussiennes pour modéliser les densités de probabilité des observations, la détermination du second terme (le terme $\prod_s P(x_s | x_{s'}, s' \in N_G(s))$), qui représente la connaissance contextuelle

introduite par le modèle, ou modèle *a priori*, n'est pas calculable directement à cause de son expression non causale, c'est-à-dire à cause des interdépendances qui existent entre les états voisins.

Pour contourner cette difficulté on a généralement recours à des méthodes de simulation telles que l'échantillonnage de Gibbs ou l'algorithme de Metropolis [1]. Une autre possibilité est de restreindre l'expression à un système de voisinage causal. Dans tous les cas cependant la recherche de la solution de segmentation optimale nécessite l'exploration de l'ensemble des configurations E , ce qui induit une combinatoire importante. Cette considération n'est pas négligeable car les images des documents que nous considérons sont particulièrement grandes et impliquent donc un nombre de sites important et donc un nombre de configurations possibles également très important.

La recherche de la configuration optimale se rapporte au cadre théorique MRF-MAP. Selon le théorème d'Hammersley-Clifford, un champ de Markov suit une distribution de Gibbs [1], ainsi le modèle *a priori* $P(X)$ peut être réécrit de la manière suivante :

$$P(X) = \frac{1}{Z} \exp \left(- \sum_{c \in C} V_c(X) \right)$$

Dans cette expression C désigne l'ensemble des cliques sur la grille G associée à l'image, selon le système de voisinage

choisi $N_G = \{N_G(s), s \in S\}$. Une clique est définie par un ensemble de sites mutuellement voisins selon le système de voisinage défini. V_c désigne une fonction appelée fonction de potentiel associée à la clique c , et Z est une constante de normalisation appelée fonction de partition dans le cadre des champs de Markov. Grâce à ce théorème il est possible de calculer les probabilités locales sur chaque site et ainsi le calcul de la probabilité sur l'ensemble de la configuration devient possible. La fonction de potentiel locale peut être dérivée des probabilités jointes locales des cliques d'ordre n définies par le système de voisinage choisi. Cela nous permet d'introduire l'énergie jointe $U(X, O)$ d'une configuration sur le champ d'étiquettes, en calculant le logarithme négatif de la probabilité jointe :

$$\begin{aligned}
 -\log(P(O|X)P(X)) &= \sum_s -\log(P(o_s|x_s)) - \log(Z) + \sum_{c \in C} V_c(X) \\
 -\log(P(O|X)P(X)) &= U(X, O) - \log(Z) \\
 \text{avec } U(X, O) &= \sum_s -\log(P(o_s|x_s)) + \sum_{c \in C} V_c(X)
 \end{aligned}$$

Cette formulation fait apparaître deux termes : un terme d'attache aux données représenté par l'énergie $U(O|X) = -\log(P(o_s|x_s))$, et un terme de régularisation sur le champ des étiquettes (ou modélisation de la connaissance *a priori*) défini par la somme des potentiels sur les cliques du champ des étiquettes $U(X) = \sum_{c \in C} V_c(X)$

Ainsi dans le cadre MRF-MAP, décoder l'image revient à maximiser la probabilité jointe globale $P(X, O)$ ce qui est équivalent à minimiser la fonction d'énergie jointe puisque Z est une constante pour une image donnée :

$$\hat{x} = \arg \min_x U(X, O)$$

Il s'agit d'un problème combinatoire non trivial car la fonction d'énergie peut ne pas être convexe et présenter des minima locaux. Différentes techniques d'optimisation peuvent être utilisées pour trouver la configuration optimale du champ d'étiquettes par minimisation de la fonction d'énergie.

2.1.2 Décodage

Pour procéder au décodage de l'image par minimisation de la fonction d'énergie plusieurs méthodes sont envisageables. Nous avons choisi dans ces travaux d'utiliser des méthodes de relaxation, en l'occurrence les algorithmes ICM (Iterated Conditional Modes) et HCF (Highest Confidence First), car si elles ne garantissent pas de trouver l'étiquetage optimal, elles fournissent néanmoins en pratique un résultat très satisfaisant avec un temps de traitement raisonnable comparativement aux méthodes dites optimales comme le recuit simulé ou les algorithmes génétiques. Ce critère est important pour la tâche de segmentation que nous considérons car les images sur lesquelles nous travaillons sont de grandes dimensions (environ 2500 par 3500 pixels) et impliquent donc un nombre de sites à étiqueter important et donc une complexité combinatoire accrue.

2.2 Application à la segmentation d'images de documents

Pour appliquer le cadre théorique de l'étiquetage par champs de Markov à la segmentation d'images, nous devons faire des choix concernant la modélisation de la fonction de densité de probabilité d'émission des observations, des fonctions de potentiel des cliques et le choix de la méthode d'optimisation utilisée pour minimiser la fonction d'énergie. Dans ces travaux nous nous intéressons à la segmentation de documents manuscrits tels que des brouillons ou des manuscrits d'auteurs en leurs parties élémentaires telles que les zones de corps de texte, d'annotation,... Nous utilisons pour cela un modèle de champ de Markov pour modéliser la connaissance *a priori* que nous avons du problème à traiter. Nous décrivons dans la suite les choix que nous avons effectués dans la mise en oeuvre de notre modèle afin de résoudre cette tâche.

2.2.1. Modèle d'attache aux données

Dans un modèle de champ de Markov caché le modèle d'attache aux données est défini par les densités de probabilité d'émission des observations. Dans notre modèle nous avons choisi de modéliser ces densités par des mélanges de gaussiennes. Les paramètres de ces mélanges sont appris de manière supervisée par l'algorithme EM sur des images de documents manuellement étiquetées. Le nombre de

composantes gaussiennes dans chaque modèle de mélange est déterminé automatiquement en utilisant le critère de Rissanen. Pour déterminer le nombre de composantes gaussiennes et apprendre les paramètres de ces mélanges nous utilisons le module CLUSTER de Bouman¹. Les observations que nous prenons en compte sont des caractéristiques extraites en chaque site $s(i, j)$ sur la grille G appliquée sur l'image. Comme nous travaillons sur des images binaires nous avons simplement choisi d'extraire en chaque site s un vecteur de caractéristiques de densités de pixels sur deux niveaux de résolution. Ce vecteur contient 18 caractéristiques. Les 9 premières correspondent aux densités de pixels noirs dans le site $s(i, j)$ et dans ses 8 voisins (selon un voisinage 8-connexe) sur le premier niveau de résolution (pleine résolution). En se basant sur le même principe les 9 caractéristiques suivantes correspondent aux densités de pixels noirs extraites sur le second niveau de résolution, c'est-à-dire à un niveau de résolution plus faible (on utilise pour cela une grille plus grande) comme on peut le voir sur la figure 2. Sur ce second niveau chaque site correspond en réalité à une fenêtre d'analyse de 3×3 sites sur le niveau précédent. Il est important de noter que la taille des sites $s(i, j)$ de la grille initiale G doit être adaptée à la taille des plus petits objets ou éléments de mise en page que l'on souhaite extraire de l'image. Le choix de cette taille résulte nécessairement d'un compromis entre la qualité de la segmentation que l'on souhaite obtenir et la complexité combinatoire que la résolution du problème à résoudre implique. Plus la grille sera fine plus précis sera l'étiquetage mais s'il y a beaucoup de sites à étiqueter le processus de minimisation de l'énergie sera beaucoup plus complexe et coûteux à résoudre. Il est donc important d'adapter la taille de la grille en fonction de la tâche de segmentation considérée et du compromis voulu.

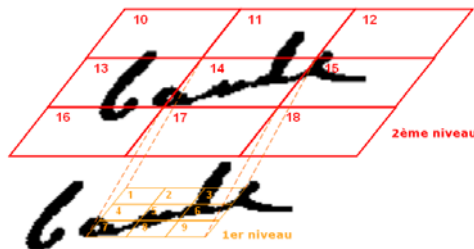


FIG. 2 - Principe d'extraction des caractéristiques de densité multirésolution

¹ <http://dynamo.ecn.purdue.edu/~bouman/software/cluster>

2.2.2. Modèle de régularisation

Dans un modèle de champ de Markov caché le modèle de régularisation est relatif aux fonctions de potentiel sur les cliques du graphe des dépendances associé au champ.

L'énergie jointe sur l'image est donnée par :

$$U(X, O) = -\sum_s \log(P(o_s / x_s)) + \sum_c V_c(X)$$

Les V_c sont les potentiels (ou énergies) associés aux cliques du champ.

$$U(X) = -\log P(X) = \sum_c -\log P_c(X) = \sum_c V_c(X)$$

Si nous considérons les cliques d'ordre 2 associées à un voisinage 4-connexe nous avons alors :

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

où

$$C_1 = \{(i, j), 1 \leq i \leq n, 1 \leq j \leq m\}$$

$$C_2 = \{((i, j), (i+1, j)), 1 \leq i \leq n, 1 \leq j \leq m\}$$

$$C_3 = \{((i, j), (i, j+1)), 1 \leq i \leq n, 1 \leq j \leq m\}$$

D'où finalement :

$$V_c(X) = \sum_{c=\{g,h\}, h \in N(g), g \in N(h)} -\log P(x_g, x_h)$$

La probabilité jointe $P(x_g, x_h)$ peut être exprimée de la manière suivante :

$$P(x_g, x_h) = P(x_g) I(x_g, x_h) P(x_h)$$

où $I(x_g, x_h)$ désigne un terme d'interaction entre les étiquettes sur les sites adjacents. Ainsi l'énergie *a priori* $U(X)$ peut être définie par :

$$U(X) = \sum_{c \in C} V_c(X) = -\sum_{s \in S} \log(P(x_s)) - \sum \log(I(x_g, x_h))$$

Les termes d'interaction sont définis comme des termes d'information mutuelle prenant en compte seulement les directions horizontales et verticales (avec un voisinage 4-connexe) comme suit :

$$I_H = \frac{P(w_k | w_l)}{P(w_k)P(w_l)} \quad I_V = \frac{P\left(\frac{w_k}{w_l}\right)}{P(w_k)P(w_l)}$$

où

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

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

Comme pour les paramètres des mélanges de gaussiennes, ces probabilités sont apprises de manière supervisée sur des exemples étiquetés manuellement, en déterminant par comptage la fréquence de chaque transition possible entre les états sur chacune des directions. Si une règle de transition n'apparaît pas dans les exemples d'apprentissage, sa probabilité n'est pas fixée à zéro mais à une valeur non nulle très faible cependant, mais rendant cette transition non pas impossible mais très peu probable.

Au final les fonctions de potentiel sur les cliques sont définies comme suit :

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

De manière similaire, d'après ces définitions, l'utilisation de cliques d'ordre 2 avec un système de voisinage en 8-connexité est très simple. On a simplement à prendre en compte les interactions diagonales également.

3. Modélisation par champs aléatoires conditionnels

Les CRF ont été initialement introduits dans le domaine de l'extraction d'information par Lafferty et al. [2] notamment pour des tâches d'analyse lexicales et syntaxiques. Jusqu'à présent ils ont principalement été utilisés pour la modélisation et l'analyse de données séquentielles (1D). Quelques travaux concernant des modèles CRF 2D pour l'analyse d'images de scènes ont été proposés récemment [3][4][5], mais à notre connaissance, à part les travaux présentés dans [6] sur la reconnaissance de diagrammes, ceux présentés dans [7] sur la reconnaissance de mots manuscrits et tout récemment les travaux de Shetty et al. [8] sur la discrimination

imprimé/manuscrit, peu de modèles CRF 2D et peu d'applications des CRF à l'analyse d'image de documents ont été proposés, notamment pour des tâches de segmentation/reconnaissance de la structure de documents complexes et bruités.

Les CRF ont montré leur supériorité sur les MRF dans d'autres domaines [1][3][4] en fournissant des résultats bien meilleurs que les modèles génératifs sur des tâches clairement discriminantes comme la segmentation. Cependant ils ont été peu utilisés encore en analyse d'images de documents, c'est pourquoi il nous semble intéressant d'adapter ces modèles et de les appliquer à ce domaine.

Le modèle que nous proposons consiste en une combinaison de classifieurs discriminants prenant des décisions d'étiquetage en chaque site en fonction de caractéristiques intrinsèques extraites de l'image et des scores de classification des sites voisins. Le décodage de l'image à l'aide d'un tel modèle étant non causal, nous avons recours à des techniques d'inférence itératives adaptées des méthodes de relaxation telles que l'algorithme ICM ou l'algorithme HCF classiquement utilisés avec les champs de Markov. L'un des principaux avantages du modèle proposé réside dans l'utilisation de procédures d'apprentissage, permettant ainsi une adaptation relativement rapide et aisée à différentes problématiques d'analyse. L'apprentissage du modèle consiste à entraîner les classifieurs et à déterminer les paramètres de combinaison de manière supervisée sur des données étiquetées. Nous avons effectué une évaluation sur des brouillons manuscrits en considérant une tâche consistant à localiser les principales zones d'intérêts de ces documents. Les résultats que nous obtenons montrent que le modèle proposé fournit de meilleurs résultats d'étiquetage qu'un modèle de champ de Markov. Ces résultats sont en adéquation avec les travaux récents sur les champs conditionnels effectués dans d'autres domaines d'application.

3.1. Fondements théoriques

Alors qu'un modèle de champ de Markov donne accès de manière indirecte à la probabilité *a posteriori* de l'étiquetage sachant les observations (ce que l'on cherche lorsque l'on résout un problème de segmentation) par le biais de la règle de décomposition de Bayes $P(X/O) \propto P(O/X)P(X)$, un modèle de champ aléatoire conditionnel modélise directement la probabilité *a posteriori*. Un CRF fournit donc une formulation directe de la tâche de discrimination entre les étiquettes. Un tel modèle cherche en effet à définir une distribution de

probabilité sur les configurations d'étiquettes possibles x étant donnée une observation o .

La forme générale d'un modèle CRF est donnée par la formule suivante :

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

Un modèle CRF est donc défini comme un produit sur l'ensemble des sites de l'image, d'exponentielles d'une combinaison non linéaire de k fonctions appelées fonctions de caractéristiques, dont l'évaluation dépend à la fois des observations o et de la configuration d'étiquettes x .

Z est un facteur de normalisation traditionnellement appelé fonction de partition comme pour les champs de Markov, et qui permet d'avoir une quantité qui soit bien une probabilité.

En considérant le logarithme négatif de cette probabilité *a posteriori* on peut introduire la notion d'énergie comme dans le cadre des champs de Markov. Cependant cette énergie est cette fois directement une énergie conditionnelle (ou *a posteriori*) globale, et non plus une énergie jointe :

$$\begin{aligned} U(X = x/O = o) &= -\log(P(X = x/O = o)) \\ &= -\left(\sum_{s \in S} U_s(X_s = x_s / X^s = x^s, O = o) \right) - \log Z \end{aligned}$$

Dans cette expression X^s désigne la configuration sur le champ d'étiquettes X exceptée au site s , et $U_s(X_s = x_s / X^s = x^s, O = o)$ est l'énergie conditionnelle locale de l'étiquette x_s au site s étant donnée la configuration d'étiquettes sur le reste du champ d'étiquettes, et étant données les observations.

Cette énergie locale (ou potentiel local) U_s est définie comme une combinaison de fonctions de caractéristiques :

$$U_s(X_s = x_s / X^s = x^s, O = o) = \sum_k \lambda_k f_k(x, o, s)$$

Le principal avantage des modèles conditionnels sur les modèles génératifs comme les champs de Markov cachés est qu'ils ne décomposent pas la probabilité *a posteriori* en un modèle d'attache aux données et un modèle de régularisation (ou modèle *a priori*). Or dans le contexte des champs de Markov cachés ces deux modèles sont connus pour être difficiles à estimer correctement. De plus il est bien connu que les

modèles génératifs ne sont réellement efficaces qu'avec des espaces d'observations de faibles dimensions généralement modélisés par des modèles de mélanges gaussiens, et nécessitent de poser des hypothèses d'indépendance des observations qui ne se vérifient pas en pratique. Les modèles conditionnels ne présentent pas ces inconvénients.

La modélisation et la résolution de problèmes à l'aide de champs aléatoires conditionnels nécessitent simplement de définir les fonctions de caractéristiques du modèle et de choisir une méthode d'apprentissage des paramètres du modèle ainsi qu'une méthode d'inférence (décodage). Nous explicitons dans la suite le modèle conditionnel que nous proposons pour procéder à la segmentation des manuscrits de Flaubert, ainsi que les choix que nous avons effectués concernant ces différents points.

3.2 Un modèle conditionnel à deux niveaux pour la segmentation d'image

Les fonctions de caractéristiques sont des fonctions à valeurs réelles. C'est par ces fonctions que l'on peut intégrer dans le modèle la connaissance *a priori* que l'on a sur le problème considéré à résoudre. Dans le modèle CRF initial de Lafferty et al. les observations sont discrètes et les fonctions de caractéristiques sont donc des fonctions binaires retournant une valeur 1 si un phénomène donné est observé (par exemple la présence d'un certain mot ou d'une certaine étiquette) et 0 sinon. Dans le problème de segmentation d'images que nous considérons les observations sont continues puisqu'il s'agit de mesures effectuées dans l'image, c'est pourquoi comme Kumar et al. [4], ou encore plus récemment He et al.[3], nous avons choisi de modéliser les fonctions de caractéristiques $f(x, o, s)$ à l'aide de classifieurs discriminants. Nous utilisons pour cela des classifieurs de type Perceptron Multicouche (PMC) car ils sont rapides en décision et ont de bonnes propriétés de généralisation même avec des espaces de caractéristiques de grandes dimensions. Cependant il est possible de considérer d'autres classifieurs discriminants comme des SVM ou des classifieurs linéaires par exemple. Notre modèle CRF peut alors être vu comme un réseau de classifieurs interdépendants prenant leurs décisions en se basant à la fois sur les caractéristiques extraites de l'image ainsi que sur l'information contextuelle apportée par les décisions des classifieurs voisins.

L'énergie conditionnelle locale est définie selon la relation suivante :

$$U_s(X_s = x_s / X^s = x^s, O = o) = \sum_k \lambda_k f_k(x, o, s)$$

Dans cette relation k désigne le nombre de composantes discriminantes prises en compte dans le modèle, les paramètres λ_k sont les pondérations associées à ces composantes et $f_k(x, o, s)$ est le score fourni par le $k^{\text{ième}}$ classifieur discriminant de la combinaison. Comme nous utilisons des perceptrons Multicouche nous pouvons assimiler ces scores à des probabilités conditionnelles locales $P(x_s/x, o, s)$. Il s'agit des probabilités d'attribuer l'étiquette x_s au site considéré s étant donné un certain jeu de caractéristiques sur la configuration d'étiquettes et les observations, éventuellement dans un certain voisinage plus ou moins restreint autour du site s , ou bien même sur tout le champ.

Dans un premier temps nous avons considéré seulement deux niveaux d'analyse dans notre modèle ($k = 2$), et nous avons donc défini deux fonctions de caractéristiques : une fonction de caractéristiques sur les observations f_o et une fonction de caractéristiques sur les étiquettes f_L .

La fonction de caractéristiques f_o tient compte des observations extraites sur une fenêtre d'analyse locale centrée sur le site s alors que la fonction de caractéristiques f_L considère l'information contextuelle disponible, c'est-à-dire la contribution des étiquettes sur les sites voisins dans l'étiquetage du site s . Pour cela les énergies (c'est-à-dire les scores retournés par les classifieurs) des sites sur un voisinage défini par une fenêtre d'analyse ont considérés. La fonction d'énergie conditionnelle locale $U_s(X_s/X^s, O)$ en chaque site s est alors définie par une combinaison linéaire (figure 3) de ces deux fonctions de caractéristiques f_o et f_L :

$$U_s(X_s/X^s, O) = \lambda_o f_o + \lambda_L f_L$$

Comme nous l'avons dit, nous utilisons des Perceptron Multicouche pour modéliser ces fonctions de caractéristiques. Les scores fournis par de tels classifieurs peuvent être considérés comme des probabilités *a posteriori*. Nous pouvons donc exprimer l'énergie locale de la manière suivante :

$$U_s(X_s/X^s, O) = \lambda_o P_o(X_s/O) + \lambda_L P_L(X_s/X^s)$$

Cette formulation combine un modèle discriminant local sur les observations et un modèle discriminant local sur les étiquettes, ce qui permet de capturer l'information issue du champ des observations et

l'information issue du champ des étiquettes X dans un certain voisinage. Ce modèle permet de prendre en compte un contexte plus riche et permet une meilleure homogénéisation du champ des étiquettes X , tout en prenant en compte l'information observée. De plus en considérant un cadre réellement discriminant il est possible de relâcher l'hypothèse d'indépendance des observations si pénalisante dans les modèles génératifs. Cela permet de prendre en compte des caractéristiques éventuellement corrélées sur un voisinage plus large.

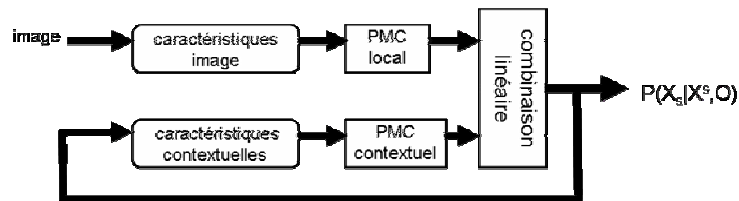


FIG. 3 - Combinaison linéaire de l'information locale image et de l'information contextuelle sur les étiquettes

Dans ce modèle, au fur et à mesure du décodage de l'image, le champ des étiquettes évolue, et donc les caractéristiques contextuelles déterminées sur ce champ d'étiquettes évoluent également, et doivent être constamment remises à jour. C'est ce processus que représente le rebouclage sur la figure 3. Les caractéristiques image elles n'évoluent pas et peuvent donc être évaluées une fois pour toutes.

3.2.1. Caractéristiques sur les observations

La fonction de caractéristiques sur les observations prend seulement en compte des caractéristiques extraites sur l'image dans une fenêtre d'analyse centrée sur un site donné (caractéristiques image locales). Cette fonction modélise l'attache aux données, c'est-à-dire la relation entre l'étiquette d'un site donné s et les observations locales sur l'image en ce site. Nous prenons en compte exactement les mêmes caractéristiques image que celles que nous avons utilisées avec le modèle de champ de Markov que nous avons présenté précédemment, c'est-à-dire des caractéristiques de densités de pixels sur deux niveaux de résolution. Ces caractéristiques sont extraites en chaque site et forment un vecteur de caractéristiques appliqué en entrée du PMC qui modélise la fonction de caractéristiques sur les observations. Les scores fournis en sortie du PMC

représentent les valeurs de la fonction de caractéristiques pour les différentes étiquettes possibles $l_i \in L = \{l_1, \dots, l_q\}$ que l'on peut associer au site courant considéré s .

3.2.2 Caractéristiques sur les étiquettes

La fonction de caractéristiques sur les étiquettes prend en compte seulement les énergies conditionnelles locales $U(X_s = l_i, i = 1, \dots, q / X_N, O)$ sur le champ des étiquettes X dans un certain voisinage N plus ou moins restreint autour du site courant. Ce voisinage est défini par une fenêtre d'analyse glissante centrée sur le site courant considéré, et dont la taille dépend de la quantité d'information contextuelle que l'on souhaite intégrer dans le modèle. Par exemple en utilisant une fenêtre de taille 3×3 et en considérant un alphabet d'étiquettes de taille $q = 3$, nous pouvons définir 27 probabilités conditionnelles *a posteriori*, ce qui nous donne un vecteur de 27 caractéristiques contextuelles appliqué en entrée du PMC modélisant la fonction de caractéristiques contextuelle sur les étiquettes. Le déplacement de la fenêtre d'analyse est contrôlé par les stratégies de décodage utilisées que nous présentons dans la section 3.1.4.

3.1.3 Apprentissage des paramètres

L'apprentissage des paramètres du modèle consiste simplement à entraîner les PMC et à déterminer les pondérations λ de la combinaison linéaire. Comme nous disposons de données manuellement étiquetées nous pouvons utiliser une approche supervisée. Pour chaque image de la base d'apprentissage nous disposons d'un étiquetage vérité-terrain. Dans ces travaux, cette vérité-terrain a été obtenue manuellement en utilisant un simple éditeur d'image et en définissant une table d'association particulière afin d'associer une étiquette à chaque couleur de l'image étiquetée. Cette base d'apprentissage a ensuite été divisée en deux parties. Tous les PMC sont entraînés en utilisant l'algorithme de rétropropagation du gradient. Le PMC sur les observations est d'abord entraîné sur la première partie de la base d'apprentissage, en ne considérant seulement que les caractéristiques locales extraites de l'image. Une fois entraîné, ce PMC est utilisé en classification sur la seconde partie de la base d'apprentissage, et les sorties de ce PMC sont utilisées pour estimer les énergies conditionnelles d'attache aux données $U(X_s / O)$ en chaque

site s de chaque image de cette seconde partie de la base d'apprentissage. Ces énergies conditionnelles sont ensuite utilisées comme données d'entrée pour l'apprentissage du second classifieur, c'est-à-dire le PMC contextuel sur les étiquettes.

Enfin les pondérations de la combinaison linéaire sont déterminées en utilisant une méthode de descente de gradient de manière à minimiser l'erreur d'étiquetage en chaque site sur la base d'apprentissage.

3.1.4 Décodage de l'image

Les techniques utilisées pour réaliser l'inférence sur des modèles de champs aléatoires conditionnels (décodage) sont quasiment les mêmes que celles proposées pour l'inférence avec les modèles de champs de Markov. Dans le cas bidimensionnel le modèle a une structure générale de graphe. Hors il n'existe pas de méthode d'inférence exacte pour les graphes, il est donc nécessaire d'avoir recours à des solutions approchées. Les techniques d'inférence approchée les plus utilisées pour les modèles de champs aléatoires sont l'algorithme *Belief Propagation* (propagation de croyances) et les techniques d'échantillonnage telles que l'échantillonneur de Gibbs et l'échantillonneur de Metropolis. Les méthodes de relaxation probabiliste telles que le recuit simulé ou l'algorithme des modes conditionnels itérés (ICM) sont également utilisées. Ces algorithmes permettent de trouver une solution approchée à l'étiquetage optimal du champ, en utilisant le critère du Maximum A Posteriori (MAP) :

$$\hat{x} = \arg \max_x P(X = x / O = o)$$

Comme pour le modèle de champ de Markov caché que nous avons présenté dans la section précédente, nous avons choisi d'utiliser et d'adapter les algorithmes ICM et HCF pour réaliser l'inférence avec le modèle de champ aléatoire conditionnel que nous proposons, car il s'agit d'algorithmes rapides et efficaces.

Lors du décodage les caractéristiques image n'évoluent pas et elles sont donc évaluées une fois pour toutes dès le début, et mémorisées. La configuration d'étiquettes par contre évolue lors du décodage. Les caractéristiques contextuelles locales et globales portant sur cette configuration d'étiquettes, elles doivent être remise à jour régulièrement. Les caractéristiques contextuelles locales sont remises systématiquement à jour dès lors que la probabilité conditionnelle locale en un site est évaluée. Les caractéristiques globales elles sont réévaluées à chaque itération complète sur l'image, c'est-à-dire sur l'ensemble des sites.

3.3 Intégration d'informations contextuelles dans le modèle

La formulation du modèle CRF que nous proposons permet d'intégrer facilement d'autres niveaux d'analyse. Nous le montrons en intégrant une troisième source d'information opérant à un niveau plus global. Cette analyse globale est réalisée en utilisant un troisième classifieur PMC estimant les probabilités *a posteriori* $P(X_s = l_i, \forall l_i \in L / F_G(X))$ d'associer l'étiquette l_i au site courant s sachant un ensemble F_G de caractéristiques statistiques globales extraites sur la configuration globale d'étiquettes dans un voisinage plus important que celui considéré par la fonction de caractéristiques contextuelles. Ces caractéristiques sont des paramètres d'Haralick calculés à partir des matrices de co-occurrences des configurations d'étiquettes.

L'originalité de notre approche réside dans le fait que nous déterminons ces caractéristiques non pas directement dans l'image, mais sur la configuration d'étiquettes. En effet, si chacune des $N+1$ classes est associée à un indice numérique de 0 à N , la configuration des étiquettes forme alors une image de profondeur $N+1$ niveaux de gris, et il est possible sur cette image de calculer des matrices de co-occurrence pour caractériser des textures et des motifs formés par des configurations locales ou plus globales d'étiquettes.

En ce qui concerne la combinaison des différents niveaux d'analyse, plusieurs solutions sont envisageables. Partant du modèle précédent, la solution la plus immédiate est de combiner linéairement cette troisième source d'information. Mais nous proposons également une autre solution de combinaison consistant à utiliser un autre classifieur PMC pour combiner de manière non linéaire les trois niveaux d'interprétation. Les résultats obtenus avec ces différentes solutions, notées respectivement impl.1, impl.2 et impl.3, sont présentés et discutés dans la section suivante.

3.3.1 Fonction de caractéristiques globales

Une analyse du champ d'étiquettes X à un niveau plus global est prise en compte par une troisième fonction de caractéristiques appelée fonction de caractéristiques globale notée $f_G = (x_s / F_G(x))$. Cette analyse globale est menée à l'aide d'un troisième PMC. Ce classifieur estime les probabilités *a posteriori* $P(X_s = l_i, \forall l_i \in L / F_G(x))$ d'associer

l'étiquette l_i au site courant s étant donné un ensemble F_G de caractéristiques statistiques globales extraites sur la configuration globale d'étiquettes x dans un voisinage plus large que celui pris en compte dans la fonction de caractéristiques contextuelles. Ce classifieur global est également un classifieur contextuel mais qui prend en compte la configuration d'étiquettes à une résolution plus basse. Typiquement pour cela la configuration est échantillonnée avec une grille plus large que la grille appliquée sur l'image. A cette résolution le champ d'étiquettes est donc divisé en plusieurs zones en superposant une grille H plus large que la grille initiale G . Chaque maille de cette nouvelle grille H donne accès à un ensemble de sites sur la grille initiale G , eux-mêmes donnant accès à un ensemble de pixels. Des paramètres statistiques sont calculés sur chacune des mailles de la grille H . Pour cela nous construisons les matrices de co-occurrence des étiquettes dans chacune de ces mailles, pour différentes orientations. Plus précisément quatre matrices de co-occurrences sont calculées pour les orientations 0° , 45° , 90° et 135° . A partir de ces quatre matrices de co-occurrences, cinq paramètres d'Haralick sont déterminés : l'homogénéité, l'homogénéité locale, la corrélation, l'entropie et le contraste. Il existe en réalité 14 paramètres d'Haralick, mais les 5 paramètres précédents sont reconnus comme étant discriminants [9]. L'originalité de notre approche réside dans le fait que nous ne déterminons pas seulement des caractéristiques dans l'image, mais également sur la configuration des étiquettes, c'est-à-dire sur la manière dont les sites voisins sont étiquetés. Les matrices de co-occurrences et les paramètres d'Haralick sont traditionnellement plutôt utilisés pour caractériser des textures dans des images en niveaux de gris, mais nous les utilisons ici pour caractériser des motifs dans une configuration d'étiquettes. L'idée est de caractériser des configurations particulières d'étiquettes à l'aide d'indicateurs numériques formant le vecteur de caractéristiques appliqué en entrée du PMC global. En plus de ces 5 paramètres d'Haralick on ajoute également la position de la zone H_s à laquelle appartient le site s considéré, ainsi que les probabilités conditionnelles locales des étiquettes en ce site. A partir de ces caractéristiques sur le champ d'étiquettes le PMC global estime la probabilité d'appartenance du site aux différentes classes. La sortie de ce classifieur doit ensuite être combinée avec les sorties des classifieurs locaux et contextuels. En ce qui concerne l'apprentissage de ce troisième classifieur, il est réalisé séquentiellement après l'apprentissage des deux autres.

3.3.2 Combinaison des sources d'information

Ce second modèle intègre trois sources d'information opérant à trois niveaux d'analyse différents. Chacune des fonctions de caractéristiques associées à ces sources d'information est modélisée par un classifieur discriminant de type Perceptron Multicouche (PMC). En chaque site de la grille appliquée sur l'image, et pour chacune de ces trois sources d'information, le PMC associé prend une décision d'étiquetage en fonction de la source d'information prise en compte. Ainsi, l'un de ces classifieurs prendra sa décision en fonction de caractéristiques extraites de l'image, un autre en fonction des scores attribués à chaque étiquette dans un voisinage plus ou moins large, et le dernier en fonction de caractéristiques extraites de la configuration d'étiquettes à un niveau global.

La fonction de caractéristique totale $f(x, o, s)$ en chaque site s , pouvant être interprétée comme la probabilité conditionnelle locale $P(X_s / X^s, O)$ étant donné les observations sur l'image O et le reste du champ d'étiquettes X , est donc maintenant définie par $f(x, o, s) = h(f_L, f_C, f_G)$ où h désigne une fonction de combinaison des trois sources d'information : locale, contextuelle et globale.

En pratique il y a plusieurs manières de réaliser cette fonction de combinaison h . Par la suite nous proposons deux solutions de combinaison.

3.3.2.1 Combinaison linéaire des sources d'information

La manière la plus simple et la plus intuitive d'intégrer la fonction de caractéristiques globales à notre modèle CRF, est de la combiner directement avec la sortie du modèle proposé précédemment. Dans ce cas le modèle intègre de manière séquentielle deux niveaux de combinaison. La sortie de ce modèle est donc une combinaison non linéaire (figure 4). L'apprentissage des pondérations de ces deux combinaisons est réalisé en deux temps. Tout d'abord les pondérations λ_L et λ_C de la première combinaison sont déterminées sur les données d'apprentissage de manière à maximiser le taux d'étiquetage moyen sans tenir compte l'information globale sur les étiquettes. Une fois ces pondérations optimisées, celles de la seconde combinaison, à savoir λ_{L-C} et λ_G , sont déterminées en utilisant la même procédure, mais intégrant cette fois l'information globale.

Ce type de combinaison permet de contrôler facilement la quantité d'information et la confiance apportées par la fonction de caractéristiques globales par rapport à l'information apportée par le modèle local. La première combinaison effectuée en quelque sorte la synthèse de l'information locale alors que la seconde permet d'intégrer l'information globale.

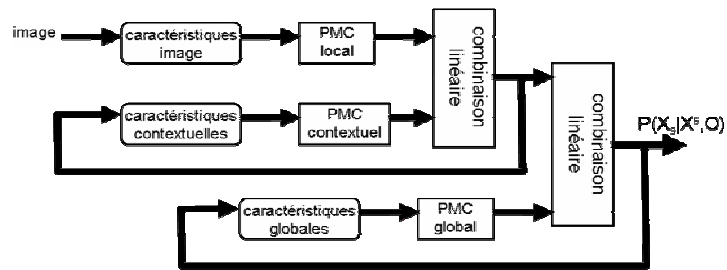


FIG. 4 - Double combinaison en cascade des trois niveaux d'analyse

3.3.2.2 Combinaison des sources d'information à l'aide d'un PMC

Nous avons également expérimenté l'utilisation d'un Perceptron Multicouche comme fonction de combinaison des différentes sources d'information (figure 5). En effet, la définition théorique et mathématique d'un PMC est très proche de celle d'un champ conditionnel puisque qu'un PMC agit comme une combinaison non linéaire de caractéristiques. Les valeurs des sorties des trois fonctions de caractéristiques pour les différentes étiquettes possibles permettent d'alimenter un seul classifieur PMC. Ainsi, si l'alphabet d'étiquettes comporte q étiquettes, la dimension du vecteur de caractéristiques appliqué en entrée du PMC de combinaison sera de $3q$. En utilisant cette solution de combinaison, les différentes sources d'information sont combinées en parallèle et les confiances à attribuer aux différentes sources d'information sont déterminées par les pondérations synaptiques du PMC de manière complètement transparente pour l'utilisateur.

La quantité d'information apportée par chaque source dépend donc des poids du PMC, mais ne peut pas être connue explicitement, contrairement à ce qui est le cas lorsque l'on utilise une combinaison linéaire.

L'apprentissage de ce modèle consiste à entraîner indépendamment les différents PMC modélisant les fonctions de caractéristiques, sur la base d'apprentissage, puis à entraîner le PMC de combinaison à partir des valeurs des fonctions de caractéristiques estimées par les PMC

précédemment appris. L'apprentissage de tous les PMC est réalisé en utilisant l'algorithme de rétro-propagation de l'erreur.

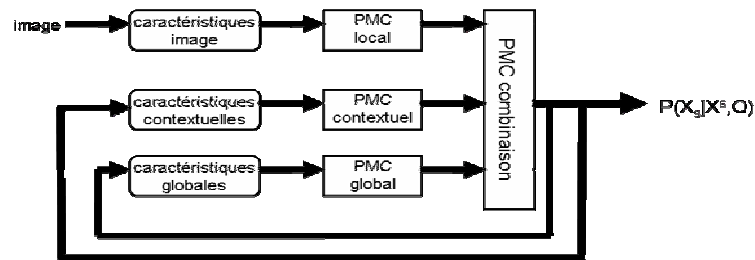


FIG. 5 - Combinaison des trois niveaux d'analyse à l'aide d'un PMC

L'avantage de cette solution de combinaison est qu'elle ne comporte qu'un seul niveau de combinaison, ainsi il est vraiment très facile de combiner plusieurs sources d'information et de déterminer par apprentissage la part d'information pertinente apportée par chacune de ces sources dans la décision finale d'étiquetage d'un site donné. Son principal inconvénient est de rajouter un PMC de plus dans le modèle, ce qui fait un apprentissage de plus à réaliser, or il est bien connu que s'ils sont rapides en décision, les réseaux de neurones nécessitent un temps d'apprentissage plus ou moins important en fonction de leur topologie (nombre d'entrées, nombre de couches cachées et nombre de neurones sur la couche cachée, et nombre de sorties). Dans ces travaux nous avons utilisé de manière très classique des PMC avec une seule couche cachée, et un nombre de neurones sur la couche cachée égal à la moitié du nombre d'entrées additionné au nombre de sorties..

Cependant, malgré tout avec cette solution, les classifieurs modélisant les différentes fonctions de caractéristiques peuvent être entraînés en parallèle, indépendamment les uns des autres ce qui permet de réduire un peu le temps nécessaire à l'apprentissage. De plus les PMC sont des classifieurs rapides et performants en décision.

4. Comparaison des deux modèles

Les expérimentations ont été menées sur une base de 69 images de manuscrits de Flaubert, en considérant une tâche d'analyse à un niveau bloc consistant à détecter de larges zones d'intérêt. Pour cela nous avons défini un modèle à six états et nous apprenons les paramètres du modèle sur des exemples manuellement étiquetés. Les états du modèle sont :

"corps de texte", "marge", "en-tête", "pied de page", "bloc de texte", "numéro de page". Les 69 images de la base ont été étiquetées manuellement en utilisant la procédure d'annotation expliquée en section 3, et elle a été divisée en trois sous-bases : une de 23 images pour l'apprentissage, une deuxième de 23 autres images pour la validation du modèle et une dernière avec les 23 images restantes pour les tests. Ces bases sont de petite taille si on considère le nombre d'images qu'elles comportent, mais si on raisonne en termes de nombre de sites, elles sont de taille raisonnable pour l'apprentissage des classifieurs PMC puisque chacune des trois bases comportent environ 86000 exemples pour des vecteurs de 20 caractéristiques.

En ce concerne la taille de la grille G , elle a été fixée empiriquement de manière à obtenir un bon compromis entre réduction de la complexité combinatoire (en réduisant le nombre de sites à étiqueter grâce à une grille plus grande) et finesse de l'étiquetage produit (plus la grille est fine et plus l'étiquetage sera précis) pour la tâche considérée. Nous avons choisi une taille de 50*50 pixels qui correspond à peu près à la largeur des espaces inter-mots et à la hauteur des lettres ascendantes ou descendantes.

Pour évaluer les résultats obtenus, nous déterminons le taux moyen d'étiquetage correct de la manière suivante :

$$TEM = \frac{\sum_{i=0}^{q-1} \left(\frac{\text{nombre de pixels correctement étiquetés}_i}{\text{nombre total de pixels d'étiquette}_i} \right)}{q}$$

où q est le nombre d'étiquettes

Le tableau 1 donne les taux d'étiquetage corrects des pixels obtenus avec les trois modèles CRF que nous avons présentés précédemment, en utilisant différentes tailles de fenêtre contextuelle sur les étiquettes comparativement aux résultats obtenus avec un classifieur local appliqué en chaque site. Impl.1 se réfère au modèle intégrant uniquement le niveau local et le niveau contextuel. Impl.2 se réfère au modèle combinant linéairement les trois niveaux : local, contextuel et global. Enfin impl.3 se réfère au modèle combinant de manière non linéaire à l'aide d'un MLP ces trois mêmes niveaux. Le jeu n°1 correspond à une fenêtre contextuelle de taille 3*3 et le jeu n°2 à une fenêtre de taille 5*5. Ces tailles de fenêtres ont également été fixées empiriquement compte tenu de la taille de la maille. Une fenêtre contextuelle de 5*5 avec une maille de 50*50 pixels permet de couvrir des zones de l'image de 250*250 pixels, ce qui constitue un contexte suffisant compte tenu de la tâche de segmentation considérée. Les résultats obtenus montrent que l'ajout de contexte et l'ajout de niveaux d'analyse (impl. 3 comparativement à

impl.1) permet d'obtenir de meilleurs résultats qu'un classifieur local. Parmi les trois implantations que nous proposons, la troisième qui exploite une combinaison de l'information locale, contextuelle et globale à l'aide d'un PMC semble être la meilleure.

	PMC local	impl. 1	impl. 2	impl. 3
jeu 1	90,56	92,55	93,90	94,04
jeu 2	90,56	93,91	93,93	94,16

Tab. 1 Taux d'étiquetage moyens obtenus avec différentes mise en oeuvre notre modèle CRF et en utilisant ICM en inférence

	Mélanges gaussiens	PMC	MRF	CRF
TEM (%)	83.70	87.50	90.56	93.91

Tab. 2 Comparaison des taux d'étiquetage obtenus avec différents modèles

Le Tableau 2 compare le taux d'étiquetage moyen (TEM) des pixels fourni par notre modèle CRF avec ceux fournis par notre modèle MRF, ainsi que par des classifieurs locaux génératifs (mélanges de gaussiennes) et discriminants (PMC). Ces classifieurs locaux utilisent les mêmes caractéristiques locales extraites de l'image que nos modèles MRF et CRF, à savoir les caractéristiques de densité de pixels noirs décrites dans la section 2.2.1. Pour le modèle génératif il s'agit d'un modèle de mélange de gaussiennes appris dans les mêmes conditions que le modèle d'attache aux données de notre modèle MRF, et le classifieur discriminant est un PMC identique à celui utilisé pour modéliser le terme local dans notre modèle CRF.

Notre modèle CRF est un modèle à la fois discriminant et contextuel qui permet d'améliorer très nettement les résultats d'étiquetage obtenus en intégrant dans un cadre discriminant des caractéristiques intrinsèques et contextuelles sur l'image et sur la configuration d'étiquettes. On peut voir sur la figure 6 un exemple de résultat d'étiquetage des zones d'intérêt dans un manuscrit de Flaubert. On peut voir notamment que toutes les zones sont correctement étiquetées.

5. Conclusion et perspectives

Nous avons proposé et présenté deux modèles markoviens pour l'étiquetage de données 2D en particulier pour la segmentation d'images

de documents : un modèle de champ de Markov 2D et un modèle de champ aléatoire conditionnel. Le premier est un modèle génératif qui en posant une hypothèse d'indépendance des observations peut se décomposer en un modèle d'attache aux données (par le biais de vraisemblances locales modélisées par des mélanges de gaussiennes) et un modèle contextuel *a priori* défini sur les cliques du champ des étiquettes.

Notre modèle de champ aléatoire conditionnel repose quant à lui sur une combinaison de classifieurs discriminants tels que des réseaux de neurones de type Perceptron Multicouche prenant en compte des caractéristiques sur les observations extraites de l'image et des caractéristiques sur les configurations d'étiquettes possibles, cela à différents niveaux d'analyse. Des solutions de combinaison simples telles que des combinaisons linéaires sont utilisées dans ce modèle, mais cependant des solutions potentiellement plus efficaces telles que des solutions de fusion d'information reposant sur l'utilisation des fonctions de croyance seront envisagées par la suite. Notre but premier était de montrer que même des solutions simples de combinaison de différentes sources d'information opérant à différents niveaux d'analyse sont intéressantes et efficaces.

Un des avantages principaux des deux modèles que nous avons présentés (MRF et CRF) réside dans le fait que les paramètres de ces modèles peuvent être déterminés automatiquement à l'aide de procédures d'apprentissage, ce qui facilite leur adaptation à différents types de données et à différentes tâches.

Les expérimentations que nous avons menées sur la segmentation en zones d'intérêt des manuscrits de Flaubert, montrent que notre modèle CRF fournit pour cette tâche de meilleurs résultats que le modèle de champ de Markov. Ces résultats sont en adéquation avec d'autres résultats présentés dans des travaux récents sur les champs markoviens qui montrent la supériorité des modèles discriminants sur les modèles génératifs pour des tâches de segmentation [2][3]. En effet, en modélisant directement la probabilité *a posteriori* de la configuration d'étiquettes étant donnés les observations, ce qui permet de prendre en compte des caractéristiques discriminants sans hypothèse d'indépendance, les modèles conditionnels sont plus efficaces que les modèles MRF qui reposent sur des hypothèses simplificatrices et sur des modèles génératifs comme les mélanges de gaussiennes qui s'avèrent être moins efficaces pour des tâches de segmentation et de classification.

Les perspectives à ce travail concernent principalement l'intégration de plus de caractéristiques intrinsèques et discriminantes dans notre

modélisation CRF, ainsi que l'étude de solutions de combinaison de l'information plus efficaces.

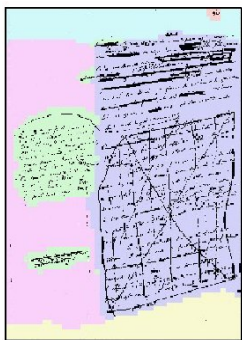


FIG. 6 - Exemple de résultat d'étiquetage en zones d'intérêt obtenu avec notre modèle CRF sur un manuscrit de Flaubert

6. REFERENCES

- [1] Chellappa, R. et Jain, A., editors (1993). *Markov Random Fields - Theory and application*, Academic Press.
- [2] J. Lafferty, A. McCallum, F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data", *18th International Conference on Machine Learning*, pp 282-289, Williamstown, USA, 2001.
- [3] X. He, R.S. Zemel, M. A. Carreira-Perpinan, "Multiscale Conditional Random Fields for Image Labeling", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04)*, pp. 695-702, Washington DC, USA, 2004.
- [4] S. Kumar and M. Hebert, "Discriminative Random Fields", *International Journal of Computer Vision (IJCV)*, 68(2), pages 179-201, 2006.
- [5] A. Quattoni, S. Wang, L. Morency, M. Collins, and T. Darrell, "Hidden Conditional Random Fields", *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 10 (Oct. 2007), pages 1848-1852, 2007.
- [6] M. Szummer and Y. Qi, "Contextual Recognition of Hand-drawn Diagrams with Conditional Random Fields", In *9th International Workshop on Frontiers in Handwriting Recognition (IWFHR'04)*, pages 32-37, Tokyo, Japon, 2004.
- [7] 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 2nd International Conference on Document Image Analysis for Libraries (DIAL'06)*, pages 30-37, Lyon, France, 2006.
- [8] S. Shetty, H. Srinivasan, S.N. Srihari and M. Beal, "Use of Conditional Random Fields in Document Image Retrieval", *proceedings of SPIE*,

Document Recognition and Retrieval IV, pages 6500U-1-11, San José, CA, USA, 2007.

- [9] S. Aksoy, R. M. Haralick, "Textural features for image database retrieval", proceedings of IEEE Workshop on Content-Based Access of Image and Video Libraries, in conjunction with CVPR'98, pages 45-49, 1998.