Extraction de la structure de documents manuscrits complexes à l'aide de champs Markoviens

Stéphane Nicolas[1] – Thierry Paquet[1] – Laurent Heutte[1]

[1] Laboratoire LITIS
UFR des Sciences et Techniques
Avenue de l'Université, Technopôle du Madrillet
76801 Saint Etienne du Rouvray cedex

{Stephane.Nicolas,Thierry.Paquet,Laurent.Heutte}@univ-rouen.fr

Résumé *: Nous abordons dans ces travaux le problème de l'extraction de la structure physique de documents manuscrits non contraints possédant une mise en page plus ou moins complexe comme les manuscrits d'auteurs. Nous proposons une méthode de segmentation basée sur une modélisation a priori de la structure de la page. Nous avons opté pour des modèles statistiques, les champs Markoviens. Dans ce cadre, la segmentation est vue comme un problème d'étiquetage ou de décodage d'image. Nous avons appliqué ces approches à la segmentation de manuscrits d'auteurs, et nous discutons les résultats obtenus sur des manuscrits de l'écrivain Gustave FLAUBERT avec différentes méthodes de décodage.*

Mots-clés *: Analyse de documents manuscrits, analyse de la mise en page, segmentation, étiquetage par champs Markoviens, décodage d'image, manuscrits d'auteurs*.

1 Introduction

Dans la chaîne d'analyse d'images de documents, l'étape de segmentation est importante car elle permet d'extraire et de séparer les entités à reconnaître. De nombreux progrès ont été réalisés ces dernières années dans le domaine de la reconnaissance de l'écrit, en ce qui concerne la reconnaissance de mots ou de phrases [BUN 03], néanmoins l'analyse de pages complètes d'écriture reste encore un problème non résolu de manière satisfaisante. Avec le développement des technologies numériques et l'explosion récente des projets de valorisation et préservation du patrimoine culturel et technologique, apparaissent de nouveaux besoins, notamment en ce qui concerne l'indexation et l'accès aux documents numérisés [BAI 03]. Ce contexte encourage le développement de méthodes robustes pour l'analyse d'images de documents. De nombreuses méthodes ont été proposées ces 20 dernières années pour la segmentation de documents imprimés [NAG 00], mais ces méthodes ne peuvent être appliquées directement aux documents manuscrits du fait de la forte variabilité spatiale de l'écriture manuscrite. Les quelques méthodes proposées dans le domaine du manuscrit sont dédiées à un type particulier de document, ou une tâche de segmentation précise (extraction de mots ou de lignes seulement) [BRU 99]. De plus beaucoup de ces méthodes sont basées sur une analyse locale, et ne prennent pas ou peu en compte le contexte ou d'éventuelles connaissances a priori, et de ce fait ne permettent pas toujours de trouver la meilleure solution. Pourtant l'introduction du contexte permet souvent de soulever certaines ambiguïtés d'interprétation. Même si les documents manuscrits ont souvent une structure moins rigide que les document imprimés nous pensons que la procédure de segmentation peut être guidée efficacement par l'introduction de connaissances a priori sur la structure globale de la page, et d'information contextuelle. Du fait de la forte variabilité inhérente aux documents manuscrits, une modélisation formelle de la structure est difficilement envisageable, car trop rigide. Les modèles stochastiques s'accommodent mieux des incertitudes. La nature de l'image étant fondamentalement bidimensionnelle, une modélisation par champs de Markov nous est apparue la plus adaptée. La théorie des champs Markoviens a été largement développée ces dernières décennies [CHE 93], et appliquée dans différents domaines du traitement d'image [LI 95], mais à notre connaissance elle n'a jamais été utilisée dans le cadre de l'analyse de documents manuscrits.

Dans ce papier nous proposons une modélisation par champs de Markov pour la segmentation de documents manuscrits, tels que les manuscrits d'auteurs, et nous présentons une application consistant à extraire les principales zones d'intérêt dans des manuscrits de l'écrivain Gustave Flaubert, à savoir principalement les lignes de texte, les éléments raturés, ou encore les ajouts en marge. Nous expliquons le formalisme utilisé en section 2, puis dans la section 3 nous explicitons les choix que nous avons effectués pour la modélisation des différents paramètres. Dans la section 4 nous présentons les résultats de différentes tâches de segmentation, obtenus sur des manuscrits de Flaubert. Nous terminons par une conclusion en section 5.

2 Segmentation d'images de documents à l'aide de champs de Markov

Nous nous plaçons dans un cadre Bayésien de reconnaissance d'image, pour lequel l'image observée résulte d'une configuration ou champ d'états sous-jacents. Dans le cas où l'on cherche à extraire la structure d'un document, cette configuration d'états cachés correspond à la structure que l'on souhaite extraire. En effet, un document vérifie généralement des règles de structuration utilisées implicitement par son auteur, même si ces règles ne peuvent pas toujours être formellement identifiées. Ainsi par exemple il est vérifié par les chercheurs que les manuscrits de Flaubert présentent une mise en page spécifique et particulière caractérisée notamment par un corps de texte occupant environ 2/3 de la page, et par la présence de nombreuses ratures et annotations en marge (figure 2.a). Une image étant de nature bidimensionnelle, on considère que cette configuration d'états sous-jacents, constitue un champ X de variables aléatoires, ou états, pouvant prendre des valeurs dans un ensemble fini $L = \{l_1, l_2, ..., l_q\}$ de $q = |L|$ étiquettes qui désignent les objets de la structure que l'on souhaite extraire. On fait l'hypothèse que ce champ de variables aléatoires est un champ Markovien, c'est à dire que la valeur d'une variable quelconque du champ n'est conditionnée que par les valeurs des variables aléatoires appartenant à un voisinage restreint de cette variable. On ne peut estimer les états sous-jacents de ce champ qu'au travers d'observations ou mesures effectuées sur l'image. On est donc en présence de deux champs aléatoires, l'un noté Y est directement observable, c'est le champ des observations ou mesures, et l'autre, le champ Markovien des états noté X, ne peut être déduit qu'à partir du premier. Une grille rectangulaire G de taille $n \times m$ est associée à l'image. Les deux champs X et Y sont définis sur cette grille. Les éléments sur cette grille sont appelés sites. Chaque site, repéré par ses coordonnées sur la grille G est noté $s(i, j)$, $1 \le i \le n$ $1 \le j \le m$. Dans ce contexte, le problème de la segmentation est vu comme un problème d'étiquetage de pixels. Il s'agit de déterminer à partir du champ des observations, le champ d'états sous-jacents ou la configuration d'étiquettes correspondante. C'est à dire trouver la meilleure configuration d'étiquettes parmi l'ensemble E des configurations possibles, selon le modèle défini. C'est un problème de recherche de maximum a posteriori (MAP). Il s'agit d'estimer la configuration optimale $\hat{x}$ du champ d'étiquettes X qui maximise la probabilité a posteriori $P(X/Y)$. Ce qui peut s'énoncer de la manière suivante: $\hat{x} = \arg\max_x P(X = x / Y = y)$. D'après le théorème de Bayes on a $P(X/Y) = \frac{P(Y/X)P(X)}{P(Y)} \propto P(Y/X)P(X)$ soit $\hat{x} = \arg\max_{x \in E}\left(P(X = x, Y = y)\right) = \arg\max_{x \in E}\left(P(Y = y | X = x)P(X = x)\right)$. En appliquant les hypothèses Markoviennes et d'indépendance entre les observations on obtient: $\hat{x} = \arg\max_{x \in E}\left(\prod_s P(y_s/x_s)\prod_s P(x_s/x_h, h \in N_G(s))\right)$ où $N_G(s)$ désigne le voisinage du site s. Dans cette expression, le terme $\prod_s P(y_s/x_s)$, appelé terme d'attache aux données, représente les densités de probabilités conditionnelles des observations. Ces densités peuvent être modélisées par des distributions gaussiennes ou des modèles de mélange. Le second terme, $\prod_s P(x_s/x_h, h \in N_G(s))$, représente la connaissance contextuelle introduite par le modèle. Il s'agit de la probabilité a priori de la configuration x du champ X. Sous cette forme cette probabilité est difficile à estimer du fait de sa forme causale liée à l'interdépendance entre les sites voisins. Pour s'affranchir de ce problème on préfère utiliser des modèles de champs de Markov non-causaux et dans ce cas raisonner plutôt en terme d'énergie. En effet, d'après le théorème de Hammersley-Clifford un champ aléatoire de Markov suit une distribution de Gibbs [GEM 84]. Le second terme peut donc être réécrit de la manière suivante $P(X) = \frac{1}{Z}\exp\left(-\sum_{c \in C} V_c(X)\right)$ où C est un ensemble de cliques associé au système de voisinage N_G défini sur la grille G. Une clique étant un sous-ensemble de sites mutuellement voisins (sous-graphes complets de G). Une clique de cet ensemble C est notée c. V_c est la fonction de potentiel de la clique c. Z est un terme de normalisation. On peut ainsi introduire la notion d'énergie en calculant le logarithme négatif de la probabilité jointe $P(X,Y)$ (à une constante $\log Z$ près), soit $U(X,Y) = \sum_s -\log\left(P(y_s/x_s)\right) + \sum_{c \in C} V_c(X)$

Finalement le problème de la segmentation (ou décodage) revient à un problème de minimisation de la fonction d'énergie: $\hat{x} = \arg\min_{x \in E} U(x, y) = \arg\min_{x \in E} \sum_s -\log(P(y_s / x_s)) + \sum_{c \in C} V_c$. Cette fonction d'énergie peut être minimisée en utilisant un algorithme d'optimisation. La configuration d'états de plus faible énergie ainsi trouvée correspond dans ce cas à la meilleure segmentation de l'image selon le modèle défini. La modélisation de la connaissance a priori sur la structure d'une classe de document donnée consiste à définir la forme des fonctions de potentiels des cliques et à fournir une estimation des densités de probabilités conditionnelles des observations. L'estimation des paramètres peut ensuite s'effectuer de manière supervisée par apprentissage. La minimisation de la fonction d'énergie est un problème combinatoire qui n'est pas trivial, puisque la fonction d'énergie peut ne pas être convexe, et présenter de nombreux minima locaux. Différentes techniques d'optimisation peuvent être utilisées pour trouver la configuration optimale du

champ d'étiquettes, en minimisant la fonction d'énergie [LI 95]. Parmi les nombreuses techniques d'optimisation existantes, nous pouvons citer les méthodes suivantes qui ont été directement appliquées à la minimisation de la fonction d'énergie de champs markoviens: le recuit simulé stochastique (SA), les modes conditionnels itérés (ICM), la méthode HCF (Highest Confidence First), la programmation dynamique 2D [GEO 03], les algorithmes génétiques, ou encore les systèmes à base de colonies de fourmis.

3 Application à la segmentation de manuscrits d'auteurs

Dans ce travail nous nous intéressons à la segmentation de documents manuscrits tels que des brouillons ou des manuscrits d'auteurs. On souhaite en extraire les principales zones d'intérêt, en utilisant une modélisation par champs de Markov. Nous décrivons dans cette partie les paramètres du modèle que nous avons défini pour résoudre ce problème.

3.1 Observations et modélisation des densités de probabilité

Les observations que nous considérons sont des caractéristiques extraites en chaque site. Dans la mesure où nous travaillons sur des images binaires, nous avons choisi d'extraire un vecteur de caractéristiques bi-résolution, basé sur une mesure de densités de pixels. Ce vecteur comporte 18 caractéristiques. Les 9 premières sont les densités de pixels noirs dans une fenêtre centrée sur le site courant, et sur ses 8 voisins. En se basant sur le même principe, les 9 caractéristiques suivantes correspondent aux mesures de densités de pixels noirs extraites à une résolution inférieure.

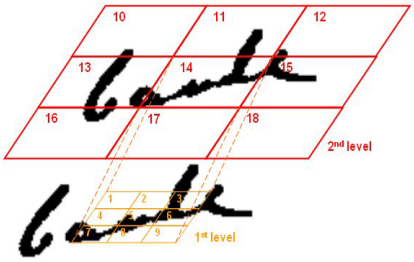


FIG. 1 - extraction des caractéristiques multirésolution de densité de pixels

Les densités de probabilité des observations sont modélisées par des modèles de mélanges de gaussiennes. Les paramètres des mélanges, c'est dire le nombre de composantes, les moyennes et variances des composantes gaussiennes, sont appris sur une base d'images étiquetées manuellement, en utilisant l'algorithme EM. Nous avons pour cela utilisé le module CLUSTER de Bouman [BOU 00], essentiellement pour des raisons pratiques. Nous renvoyons à ses travaux pour plus de précisions sur l'algorithme EM et sur le critère de Rissanen utilisé dans ce module pour déterminer le nombre de composantes gaussiennes.

3.2 Potentiels des cliques

Nous considérons des cliques d'ordre 2 associées à un voisinage 4-connexe:

$$C = C_1 \cup C_2 \cup C_3$$
$$où$$
$$C_1 = \{(i,j), 1 \le i \le n, 1 \le j \le m\}$$
$$C_2 = \{((i,j),(i+1,j)), 1 \le i \le n, 1 \le j \le m\}$$
$$C_3 = \{((i,j),(i,j+1)), 1 \le i \le n, 1 \le j \le m\}$$

On définit des potentiels d'interactions sous forme de termes d'information mutuelle prenant en compte simplement les directions horizontales et verticales. Comme pour les paramètres des modèles de mélange, ces potentiels sont appris statistiquement sur des exemples étiquetés, en estimant les fréquences d'apparition de chaque couple d'étiquettes selon les deux directions:

$$I_H = \frac{P(x_k|x_l)}{P(x_k)P(x_l)} \qquad I_V = \frac{P\left(\frac{x_k}{x_l}\right)}{P(x_k)P(x_l)}$$
$$\text{où } P(x_k|x_l) = P(x_{(i,j)} = x_k, x_{(i+1,j)} = x_l)$$
$$\text{et } P\left(\frac{x_k}{x_l}\right) = P(x_{(i,j)} = x_k, x_{(i,j+1)} = x_l)$$

Les fonctions de potentiel des cliques sont ensuite simplement déterminées en calculant le logarithme négatif de ces potentiels:

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

3.3 Décodage

Le décodage de l'image est réalisé par minimisation de la fonction d'énergie du champ. Nous avons implémenté et testé deux algorithmes couramment utilisés dans le cadre des champs de Markov, l'algorithme ICM et l'algorithme HCF. Nous avons également implémenté une méthode plus récente, la méthode par fusion de régions de [GEO 03], qui est basée sur le principe de la programmation dynamique 2D (2D DP). Les résultats obtenus avec ces trois algorithmes sont comparés dans la section suivante.

4 Résultats

L'analyse des résultats d'un algorithme de segmentation d'images de documents est un problème difficile et pas toujours bien défini, dans la mesure où il existe très peu de protocoles et de bases de d'images pour l'évaluation des performances [KAN 02]. Les quelques bases existantes concernent seulement les documents imprimés pour lesquels les méthodologies et métriques proposées pour comparer les algorithmes sont souvent dédiées à des problèmes particuliers ou à des classes bien définies de méthodes et de documents (articles de journaux, lettres, formulaires, courriers postaux). A notre connaissance il n'existe pas de telles méthodologies et métriques dédiées aux documents manuscrits ou aux documents historiques. Dans la mesure où notre approche permet de produire un

étiquetage de l'image à différents niveaux d'analyse, simplement en utilisant différentes tailles de grilles d'analyse, nous présentons des résultats obtenus sur deux tâches de segmentation à deux niveaux différents. La première consiste en un étiquetage de grandes zones d'intérêt, telles que les zones de corps de texte, de marges, de blocs textuels, d'en-tête et de pied de page, dans des manuscrits d'auteurs, en travaillant à une résolution assez faible. Les résultats obtenus sont illustrés visuellement et discutés. Avec la seconde tâche qui consiste en un étiquetage des lignes de texte, nous illustrons les capacités de cette approche à analyser l'image à un niveau plus fin, afin d'extraire et de séparer de petites entités telles que des mots ou des ratures.

4.1 Etiquetage de zones d'intérêts

De manière à pouvoir évaluer précisément les performances de notre approche, et pouvoir comparer les méthodes de décodage entre elles, selon des critères tels que le taux d'étiquetage correct ou le temps de traitement, nous avons tout d'abord considéré une tâche de segmentation pour laquelle un étiquetage simple à une résolution relativement faible est possible. En fait dans le cadre d'une telle tâche, il est simple et rapide d'étiqueter manuellement une base d'images de manuscrits, afin de pouvoir établir une vérité terrain, et de pouvoir générer des données pour l'apprentissage des modèles. Le problème que nous considérons consiste en l'étiquetage des principales zones d'intérêt de manuscrits de Gustave Flaubert, à savoir, le corps de texte, les marges, les zones d'en-tête et de pied de page, les numéros de folios, et les annotations en marge. Le modèle que nous avons défini pour cette tâche contient 6 étiquettes. La base d'images contient au total 69 images de manuscrits numérisés à une résolution de 300 dpi, et de dimensions moyennes 2400×3700. Toutes les images de la base ont été binarisées et étiquetée manuellement selon le modèle à 6 états défini (figure 2.b). Nous avons divisé la base d'images, en trois sous-bases: une pour l'apprentissage des paramètres du modèle (à savoir les paramètres des mélanges de gaussiennes, et les fonctions de potentiel), une autre pour la validation des paramètres du modèle, et la dernière pour l'évaluation des performances de la méthode. Pour cette tâche nous utilisons une grille régulière de dimensions 50×50 pixels. Cette taille de grille a été fixée empiriquement en fonction de la taille des zones à extraire. Nous comparons les résultats obtenus avec un modèle de mélange, en utilisant le critère de Maximum de Vraisemblance, et les résultats obtenus avec les algorithmes ICM, HCF et Programmation Dynamique 2D (2D DP), avec la vérité terrain produite par l'étiquetage manuel des images de la base. Pour chaque méthode de décodage, nous évaluons, le taux global d'étiquetage correct des sites de l'image (TEG), en comptabilisant le nombre de sites correctement étiquetés, et le taux d'étiquetage normalisé (TEN), en comptabilisant le nombre de sites correctement étiquetées par classe:

$$TEG = \frac{nombre\ de\ sites\ correctement\ étiquetés}{nombre\ total\ de\ sites}$$

$$TEN = \frac{\sum_{i=0}^{q-1}\left(\frac{nombre\ de\ sites\ correctement\ étiquetés\ l_i}{nombre\ total\ de\ sites\ d'étiquette\ l_i}\right)}{q}$$

où q est le nombre d'étiquettes

Pour chaque méthode de décodage nous donnons également le temps de traitement moyen en secondes, pour le décodage d'une page. Ce temps ne prend en compte que le décodage de l'image, le temps nécessaire à l'estimation des distributions de probabilités n'est notamment pas indiqué, puisque qu'il est commun, quelle que soit la méthode de décodage utilisée ensuite.

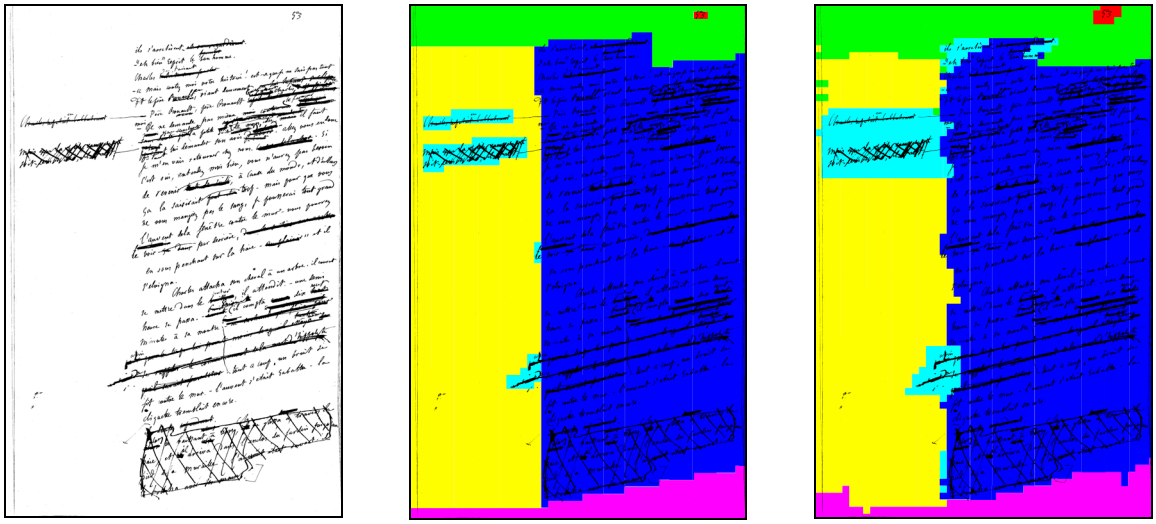
	Modèle de mélange	ICM	HCF	2D DP
TEG (%)	88,0	86,6	90,3	84,6
TEN (%)	83,7	87,5	88,2	87,4
temps (s)	-	0,21	0,29	0,61

TAB. 1 - Taux d'étiquetage obtenus avec différents algorithmes de décodage

Ces résultats montrent que l'utilisation de l'algorithme HCF permet d'augmenter à la fois le taux global et le taux normalisé, et que cet algorithme surpasse les autres. De plus il est relativement rapide comparé à la méthode de Programmation Dynamique 2D. Les différences que l'on peut observer dans le tableau de résultat entre le taux global et le taux normalisé s'expliquent par une répartition non homogène des classes. La figure 2.c illustre un exemple de segmentation obtenue.

4.2 Etiquetage des lignes de texte

Les manuscrits de Flaubert contiennent de nombreuses ratures et passages biffés (voir figure 2.a), c'est pourquoi nous avons également testé les possibilités de l'approche que nous proposons sur une deuxième tâche de segmentation nécessitant une analyse plus fine de l'image, afin de séparer les ratures des mots, et d'extraire les lignes de texte, en utilisant une modélisation a priori intégrant plusieurs états. Pour cela nous avons tout d'abord défini un modèle comportant 4 états: "pseudo-mot", "rature", "symbole diacritique et ponctuation", et "fond". L'analyse s'effectue au niveau le plus fin, c'est à dire au niveau du pixel, en utilisant une grille régulière 1×1. Pour obtenir les résultats que nous présentons, nous avons utilisé la méthode de Programmation Dynamique 2D lors du décodage de l'image. Nous ne fournissons ici que des résultats qualitatifs, dans la mesure où il est très difficile et fastidieux sur ce type de tâche de segmentation, d'étiqueter manuellement suffisamment d'images, à un niveau aussi fin, pour produire une vérité terrain permettant l'apprentissage des paramètres et l'évaluation des résultats de segmentation. Sur la figure 3.a, on peut voir les résultats obtenus avec ce modèle sur un fragment de manuscrit.



(a) (b) (c)

FIG. 2 – Etiquetage de zones d'intérêt: (a) manuscrit (b) vérité terrain (c) résultats de la segmentation avec conventions de couleurs suivantes: bleu = corps de texte, jaune = marge, cyan = bloc annotation, vert = en-tête, rose = pied de page, rouge = numéro de page

La figure 3.b montre précisément les résultats obtenus sur une zone de suppression, où les mots et les traits de ratures sont complètement connectés. On peut voir sur cette image que la méthode arrive à bien séparer l'écriture des traits de rature. Ce résultat montre l'intérêt de la méthode par rapport aux approches basées sur l'analyse des composantes connexes de l'image couramment utilisées pour l'analyse des documents manuscrits. En effet, le fait d'analyser l'image et non pas les composantes connexes, permet de segmenter des éléments qui pourraient être connectés. Ce cas de figure se présente souvent dans les manuscrits d'auteurs, car la densité d'information est telle que souvent tous les tracés sont connectés. La figure 3.c, montre un résultat similaire obtenu sur un fragment contenant un mot et une rature connectés ensemble. On peut remarquer que la méthode est capable de les séparer correctement. Ce modèle tel que nous l'avons défini précédemment permet d'extraire les mot ou pseudo-mots (fragments de mots), ainsi que les ratures, mais ne modélise pas réellement les lignes de texte. Or nous souhaitons être capable de détecter les lignes. Une ligne peut être considérée d'un point de vue structurel, comme une succession de mots et éventuellement de ratures, séparés par des espaces. Afin de pouvoir modéliser les lignes, nous avons donc ajouté un état supplémentaire "espace inter-mot" à notre modèle. Une fois ces espaces étiquetés par la méthode il est aisé d'extraire les lignes de texte, ou des éléments de la structure physique de la page de plus haut niveau, tels que les blocs, en appliquant des règles de regroupement des régions étiquetées. Les résultats obtenus sont globalement prometteurs, puisque les espaces inter-mots semblent bien segmentés (voir figure 3.d). De manière à mieux modéliser les lignes de texte, nous avons ensuite défini un troisième modèle à 6 états, en ajoutant au modèle précédent un état "interligne", modélisant les espaces entre les lignes de texte. En effet la connaissance des interlignes permet de mieux segmenter les lignes de textes, et de détecter les blocs. On peut voir sur la figure 3.e le type de résultat obtenu sur le même fragment que précédemment, et sur la figure 3.f le résultat obtenu sur une page complète. Il est important de noter que la méthode que nous proposons fournit simplement un étiquetage de l'image au niveau pixel (à l'instar des travaux présentés dans [BAI 06]), mais ne segmente pas directement les objets de plus hauts niveaux d'abstraction tels que les lignes de texte ou les blocs. Cependant à partir de cet étiquetage bas niveau, il possible de segmenter ces entités de plus hauts niveaux, en utilisant des règles de fusion d'étiquettes et des techniques d'extraction de composantes connexes. Nous n'avons toutefois pas encore développé cette partie du système.

5 Conclusion

Nous avons présenté une méthode de segmentation d'images de documents manuscrits non contraints présentant une mise en page plus ou moins complexe, tels que les manuscrits d'auteurs notamment. Cette méthode est basée sur une modélisation statistique par champs de Markov, des relations de dépendance contextuelle entre les différentes zones d'intérêt de l'image, et sur un décodage par des algorithmes d'optimisation visant à minimiser la fonction d'énergie du champ caché représentant la segmentation idéale. Cette méthode a été appliquée à la segmentation d'images de manuscrits de l'écrivain Gustave Flaubert, afin de déterminer les zones de ratures, les mots, les symboles diacritiques et le fond de la page. Les résultats obtenus, bien que préliminaires, sont encourageants et

montrent le potentiel de cette méthode. Plusieurs algorithmes de décodage par minimisation de la fonction d'énergie ont également été testés. Sur les tests que nous avons effectués, l'algorithme HCF donne de meilleurs résultats. Les principaux avantages de la méthode que nous proposons résident dans la capacité des champs de Markov à tenir compte à la fois de la variabilité locale et à modéliser de manière globale la connaissance a priori sur la structure du document, ainsi que dans ses facultés d'adaptation à différents classes de documents, puisque les paramètres du modèle peuvent être déterminés automatiquement par apprentissage. Cette méthode nous semble bien adaptée aux documents présentant une forte variabilité ou dégradés, tels que les documents manuscrits ou les documents d'archive, mais a priori elle peut également s'appliquer aux documents imprimés. La condition est de disposer de quelques exemples étiquetés et éventuellement d'enrichir le jeu de caractéristiques image utilisé. Dans les prochains mois nous prévoyons de tester cette méthode sur différentes catégories de documents.

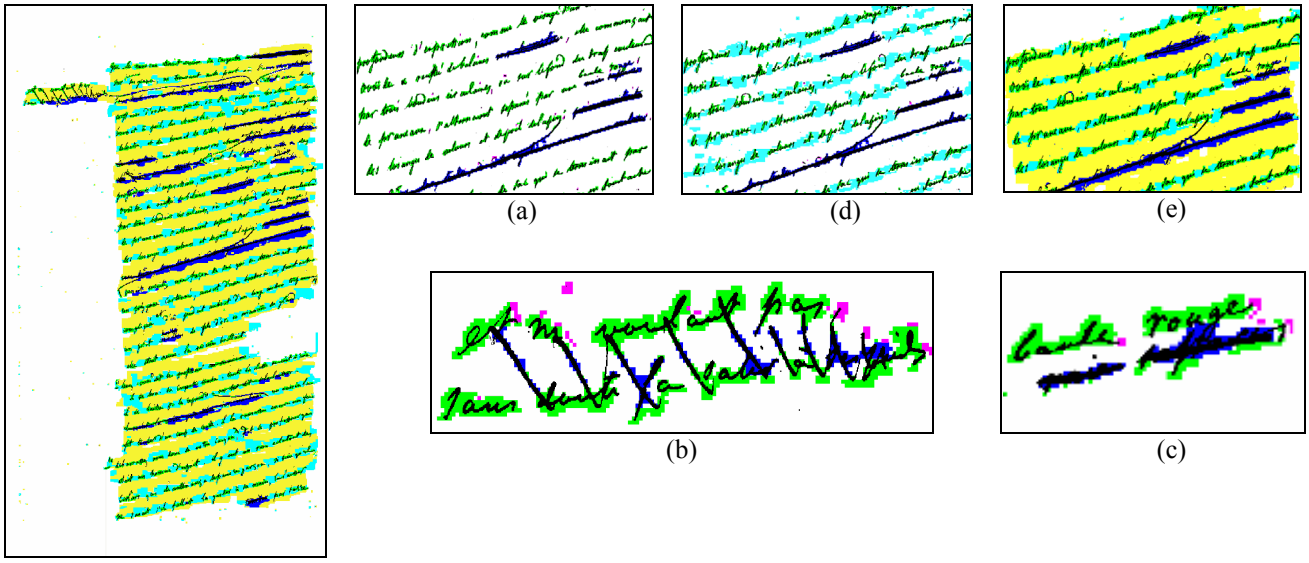


FIG. 3 – résultats de segmentation obtenus avec les conventions de couleurs suivantes: blanc = fond, vert = pseudo-mot, bleu = rature, cyan = espace inter-mots, jaune = interligne, rose = diacritique

Références

[BAI 03] H. Baird, "Digital Libraries and Document Image Analysis", Proc. of the seventh Int. Conference on Document Analysis and Recognition (ICDAR'03), Edinburgh, pp 2-14, 2003.

[BAI 06] H. Baird, M. A. Moll, J. Nonnemaker, M. R. Casey and D. L. Delorenzo, "Versatile Document Content Extraction", Proc of SPIE/IS&T Document Recognition & Retrieval XIII Conf., San Jose, CA, January 18-19, 2006.

[BOU 00] C.A. Bouman, "CLUSTER: An unsupervised algorithm for modeling Gaussian mixtures", Technical Report, Purdue University, 2000, http://dynamo.ecn.purdue.edu/~bouman/software/cluster/

[BUN 03] H. Bunke, "Recognition of Cursive Roman Handwriting, Past, Present and Future", Proc. of the seventh International Conference on Document Analysis and Recognition (ICDAR'03), Edinburgh, pp 448-459, 2003.

[BRU 99] Bruzzone, E., Coffetti, M.C, "An algorithm for extracting cursive text lines", Fifth International Conference on Document Analysis and Recognition, pp. 749-752, September 1999.

[CHE 93] Chellappa, R. et Jain, A., editors (1993). "Markov Random Fields - Theory and application", Academic Press.

[GEM 84] S. Geman, D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 6, n°6, , pp. 721-741, novembre 1984.

[GEO 03] E. Geoffrois, "Multi-dimensional Dynamic Programming for statistical image segmentation and recognition", Int. Conference on Image and Signal Processing (ICISP'03), 2003.

[KAN 02] T. Kanungo, H. S. Baird, R.M. Haralick, Guest Editors, "Performance Evaluation: Theory, Practice, and Impact", special issue of *International Journal on Document Analysis and Recognition*, vol. 4, n°3, march 2002.

[LI 95] S.Z. Li, "Markov Random Field Modeling in Computer Vision", Springer, Tokyo, 1995.

[NAG 00] Nagy, G, "Twenty years of document image analysis in PAMI", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 22, n°1, pp. 38-62, 2000.