Laboratoire d'Informatique, de Traitement de l'Information et des Systèmes
UNIVERSITE DE ROUEN
U.F.R DES SCIENCES ET TECHNIQUES

THESE DE DOCTORAT

Pour obtenir le grade de
DOCTEUR DE L'UNIVERSITE DE ROUEN

Discipline: Sciences Appliquées
Spécialité: Informatique, Automatique, Systèmes

Segmentation par champs aléatoires pour l'indexation d'images de documents

Stéphane Nicolas

Soutenue le 4 décembre 2006 devant le jury composé de :

M	Laurent	HEUTTE	Université de Rouen	Examinateur
M	Jean-Marc	OGIER	Université de La Rochelle	Président du jury
M	Thierry	PAQUET	Université de Rouen	Directeur de thèse
Mme	Françoise	PRETEUX	INT Evry	Rapporteur
M	Karl	TOMBRE	Ecole des Mines de Nancy	Rapporteur

Résumé

Avec le développement des technologies numériques, la valorisation de notre patrimoine documentaire est devenue un enjeu majeur, qui pose des difficultés d'indexation et d'accès à l'information. L'analyse de documents peut apporter une solution, mais les méthodes classiques ne sont pas suffisamment souples pour s'adapter à la variabilité rencontrée. Notre contribution porte sur l'implémentation d'un modèle de champ de Markov 2D et d'un modèle de champ aléatoire conditionnel 2D, qui permettent de prendre en compte la variabilité et d'intégrer des connaissances contextuelles, en bénéficiant de techniques efficaces d'apprentissage. Les expérimentations effectuées sur des brouillons d'auteurs et sur des manuscrits de la Renaissance, montrent que ces modèles représentent une solution intéressante et que le modèle conditionnel, de par son caractère discriminant et sa capacité naturelle à intégrer plus de caractéristiques et d'information contextuelle, offre de meilleures performances.

Mots-clés : valorisation du patrimoine, analyse d'images de documents, étiquetage d'images, champs de Markov cachés champs aléatoires conditionnels, méthodes d'inférence, apprentissage.

Abstract

With the development of digital technologies, the valorization of our cultural heritage is becoming a major stake, which exhibits a lot of difficulties for information indexing and retrieval. Document image analysis can bring a solution, however traditionnal methods are not flexible enough to deal with the varability found in patrimonial documents. Our contribution relates to the implementation of a 2D Markov random field model and a 2D conditional random field model, which make it possible to take variability into account and to integrate contextual knowledge, while taking benefit from machine learning techniques. Experiments on handwritten drafts and manuscripts of the Renaissance, show that these models can provide interesting solutions. Furthermore, the conditional random field model provids better results, allowing to integrate more intrinsic and contextual features in a discriminative framework, using a classifier combination approach.

Keywords : patrimonial document valorization, document image analysis, image labelling, hidden markov random fields, conditional random fields, inference techniques, learning.

Table des matières

Introduction générale

De nombreux progrès ont été réalisés ces dernières années dans le domaine de l'analyse d'images de documents, que ce soit pour la reconnaissance de documents imprimés ou la reconnaissance de l'écriture manuscrite. Ces progrès ont été réalisés notamment dans le cadre d'applications industrielles spécifiques telles que la Gestion Électronique de Document (GED), le traitement automatique des formulaires et des chèques ou encore le tri du courrier.

En ce qui concerne l'analyse des documents imprimés les avancés réalisées touchent principalement l'analyse et la reconnaissance des structures. Nombre des techniques proposées dans ce domaine arrivent aujourd'hui à maturité, mais des difficultés subsistent encore en ce qui concerne l'analyse de documents hétérogènes comportant des informations de natures différentes (texte, images, graphiques) et des mises en page complexes. Ces difficultés sont liées à la variabilité des structures mais également au manque de généricité et de souplesse des solutions proposées, car il s'agit souvent de méthodes dédiées.

Dans le domaine de la reconnaissance de l'écrit, les efforts se sont principalement concentrés sur la reconnaissance de mots ou de phrases [Bunke 03], et de nombreux progrès ont été réalisés, notamment grâce à l'utilisation d'approches faisant coopérer les modules d'analyse et de reconnaissance. Cependant peu de travaux se sont intéressés à l'analyse des structures dans les documents manuscrits, et 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 du patrimoine au moyen de la numérisation, apparaissent encore de nouveaux besoins en matière d'analyse d'images de documents, notamment en ce qui concerne l'indexation et l'accès aux

documents numérisés [Baird 03]. Ce contexte encourage le développement de méthodes robustes pour l'analyse d'images de documents. En effet le traitement des documents patrimoniaux soulève de nombreuses difficultés. Ces difficultés sont liées à la grande variété des problèmes et des types de documents rencontrés, puisque les documents patrimoniaux regroupent à la fois des documents anciens et des documents plus récents, des documents imprimés et des documents manuscrits. De plus les documents considérés sont souvent des documents dégradés et fortement bruités.

Notre contribution à ce problème porte sur le développement de méthodes d'analyse d'images de documents basées sur l'utilisation de modèles markoviens , tels que les champs de Markov cachés et les champs aléatoires conditionnels. L'utilisation de ces modèles permet de considérer la segmentation et la reconnaissance de manière conjointe comme un problème d'étiquetage d'images, dans un cadre probabiliste, ce qui permet de mieux s'affranchir de la variabilité inhérente aux document patrimoniaux. De plus les techniques utilisées permettent une adaptation simple à différents contexte d'analyse grâce à des procédures d'apprentissage automatique.

Nous nous proposons dans le premier chapitre de faire un panaroma des différents problèmes que soulèvent la numérisation des documents patrimoniaux, et nous montrons au travers d'une étude bibliographique de différents travaux réalisés dans ce domaine ces dernières années, comment l'analyse d'images de documents peut apporter une solution à l'indexation des masses de données issues de la numérisation. Nous nous attachons à montrer dans ce chapitre qu'il existe de réels besoins de méthodes robustes et adaptables en matière d'analyse des structures des documents, afin de faciliter les tâches d'indexation.

Dans le deuxième chapitre, nous nous intéressons donc naturellement à la problématique de l'analyse d'images de documents, afin d'étudier les solutions apportées ces dernières années dans ce domaine, à la fois pour les documents imprimés, et pour les documents manuscrits. Nous montrons plus particulièrement dans ce chapitre que les problèmes considérés et les solutions apportées dépendent beaucoup des types de documents considérés, et ne sont pas facilement adaptables à différentes tâches et différents types de

documents. Nous mettons également en avant le fait que pour des documents très bruités comme les documents manuscrits ou les documents anciens, il est nécessaire de mettre en place des stratégies faisant coopérer les différents niveaux d'analyse, afin de mieux s'affranchir de la variabilité.
Partant du constat que dans d'autres domaines de l'analyse d'images, il existe des techniques de modélisation probabiliste efficaces, comme les modèles de Markov cachés, qui permettent de considérer conjointement l'analyse et la reconnaissance et d'intégrer de l'information contextuelle dans le processus, tout en bénéficiant de techniques efficaces d'apprentissage, nous proposons donc d'utiliser de tels modèles pour l'analyse d'images de documents. Nous proposons dans les chapitre 3 et 4 deux modèles Markoviens 2D : un modèle de champ de Markov caché puis un modèle de champ aléatoire conditionnel. Ces modèles permettent de considérer de manière conjointe la segmentation et la reconnaissance comme un problème d'étiquetage d'image.
Dans le troisième chapitre, nous commençons donc par fournir un état de l'art sur l'utilisation des champs de Markov dans différents domaines de l'analyse d'image, afin de montrer que ces modèles probabilistes ont été peu utilisés jusqu'à présent dans le domaine de l'analyse d'images de documents si ce n'est pour des tâches de reconnaissance d'écriture, telles que la reconnaissance de chiffres ou de mots manuscrits. Ces modèles sont pourtant utilisés depuis plusieurs années avec un certain succès, en analyse d'image, que ce soit pour des tâches de prétraitement de bas niveau comme le débruitage ou la segmentation en contours ou en régions, mais également pour des tâches d'interprétation de plus haut niveau comme l'analyse de scènes. La force de ces outils réside dans une modélisation probabiliste et contextuelle des problèmes d'analyse et dans le fait qu'ils bénéficient de l'avantage apporté par des techniques d'apprentissage automatique des modèles. Nous proposons donc d'appliquer ce type de modélisation à l'analyse des structures de documents, à différents niveaux, comme la structure physique ou la structure logique, afin de permettre l'extraction d'informations utiles à l'indexation des documents.
Après une description du cadre théorique des champs de Markov, et plus particulièrement des champs de Markov cachés, nous présentons le modèle que nous proposons pour procéder à l'analyse de documents de différentes natures. Notre objectif se concentre en particulier sur l'analyse des documents fortement bruités comme les documents manuscrits et les documents

anciens, pour lesquels les approches classiques d'analyse n'apportent pas de solutions viables. Cependant, la méthodologie que nous proposons ne se limite pas qu'au traitement de ces documents, puisqu' elle est suffisamment générique et adaptable pour permettre de prendre en compte différents types de documents, et différentes tâches d'analyse. Le pouvoir d'adaptation de la méthode réside dans l'utilisation de techniques d'apprentissage automatique. La méthode proposée est ensuite évaluée à la fois sur des manuscrits de Flaubert et sur des documents de la Renaissance en considérant différentes tâches d'étiquetage.
Il a été montré récemment dans la littérature que les modèles de Markov présentent un certain nombre de limitations pour la réalisation de tâches discriminantes, du fait de leur nature fondamentalement générative, et que des modèles purement discriminants permettent d'obtenir des meilleurs résultats. Dans le quatrième chapitre nous proposons donc un modèle discriminant 2D se basant sur la théorie des champs aléatoires conditionnels. Après une présentation du cadre théorique et un état de l'art sur les quelques travaux réalisés dans ce domaine, notamment en ce qui concerne la définition de modèles 2D, nous présentons le modèle que nous proposons. Ce modèle consiste en une combinaison de classifieurs discriminants permettant de considérer l'analyse d'images de documents dans un cadre discriminant. Comme pour le modèle de champ de Markov caché présenté au chapitre 3, nous proposons une évaluation de l'approche sur des manuscrits de Flaubert et sur des documents de la Renaissance, et nous montrons que cette modélisation discriminante permet d'améliorer nettement les résultats de l'analyse, et est capable de s'adapter à différentes tâches d'étiquetage.

Nous terminons ce mémoire par des conclusions sur les travaux que nous avons effectués et nous présentons des perspectives d'évolution des méthodes présentées.

Chapitre 1

Numérisation du patrimoine et l'indexation de masses de documents

1.1 Problématique et enjeux

Notre patrimoine culturel et scientifique constitue en quelque sorte la mémoire collective de nos sociétés. Tous les experts s'accordent à dire qu'il est important de préserver ce bien précieux et qu'il est urgent de mener à bien des actions fortes garantissant à l'avenir une conservation durable de ces ressources, afin de pouvoir transmettre aux générations futures le savoir accumulé depuis des siècles sous diverses formes. L'informatique qui connaît un développement croissant depuis plusieurs années, semble de plus en plus en mesure d'apporter des solutions réalistes à ces problèmes, en permettant la création d'artefacts numériques des documents et objets de notre patrimoine, qui facilitent l'accès et l'exploitation de ce savoir.

Avec les progrès réalisés dans le domaine des technologies numériques, et la diminution des prix des dispositifs de numérisation, la préservation et la valorisation du patrimoine documentaire par le biais de substituts numériques sont donc devenues des enjeux majeurs pour nombre d'institutions. L'objectif est de préserver les ouvrages et documents anciens constituant notre mémoire collective, en proposant des éditions électroniques consultables en ligne ou sur divers supports de diffusion (CD-ROM, DVD-ROM). Il s'agit donc de développer de véritables bibliothèques numériques consti-

tuées d'un ensemble de données numériques hétérogènes (images, textes, séquences vidéos, contenus interactifs, outils de recherche d'information, outils d'analyse, ...), qui soient plus que de simples bases de données, mais de véritables outils interactifs et pédagogiques qui permettent à tout un chacun d'enrichir ses connaissances sur différents domaines.

Ces dernières années ont donc vu l'apparition de nombreuses bibliothèques numériques, parfois dédiées à une thématique particulière ou bien généralistes, et proposant l'accès à de nombreuses références. Le nombre de bibliothèques numériques et catalogues documentaires accessibles aujourd'hui en ligne est trop important pour pouvoir en faire une liste exhaustive ici. Mais en ce qui concerne les principales bibliothèques numériques françaises nous pouvons citer le projet Gallica[1] de la Bibliothèque Nationale de France[2] qui est un précurseur dans le domaine des bibliothèques numériques en France. On trouvera sur le site du Ministère de la Culture une base de données recensant l'ensemble des opérations de numérisation (réalisations et projets) conduites par les institutions culturelles françaises (bibliothèques, services d'archives, musées, services de l'inventaire, grands établissements culturels, associations etc.), soit près de 1000 collections répertoriées, et on trouvera sur le site de l'IFLA[3] (International Federation of Library Associations and Institutions), une liste des principaux programmes de numérisation américains et européens. Au travers de ces recensements on peut appréhender l'ampleur du phénomène. Il s'agit d'un domaine en pleine expansion. Cependant l'exploitation de ces masses de données issues de la numérisation et la réalisation de telles bibliothèques numériques posent des problèmes relatifs à la structuration de l'information. En effet, pour être exploitables les bibliothèques numériques ne peuvent se résumer à de simples bases d'images numérisées, il faut que l'information soit structurée, et que diverses données d'indexation (transcriptions textuelles, annotations, métadonnées) soient associées aux documents pour en permettre la recherche et l'exploitation. Cela renvoie donc à un problème d'indexation. Ce problème ne peut pas être résolu uniquement de manière manuelle, car les quantités d'information à traiter sont trop importantes. Il est donc nécessaire de mettre en place des solutions de traitement automatique permettant une

[1] http ://gallica.bnf.fr/
[2] http ://www.bnf.fr/
[3] http ://www.ifla.org/II/diglib.htm

exploitation viable des corpus constitués.

Le développement d'éditions électroniques et de bibliothèques numériques est donc une problématique proche de celle de la Gestion Électronique de Documents (GED) rencontrée dans le contexte industriel pour l'exploitation des documents imprimés, et qui offre des solutions pour intégrer les documents numérisés dans des systèmes informatiques de gestion documentaire. Ces systèmes de GED bénéficient des énormes progrès réalisés depuis maintenant plusieurs décennies dans les domaines de la reconnaissance de formes et de l'analyse d'images de documents, que ce soit dans les prétraitements, l'analyse et la reconnaissance des structures, ou encore la lecture optique des caractères [Nagy 00]. Ces résultats spectaculaires tiennent en partie aux spécificités des documents imprimés. En effet contrairement aux documents manuscrits et aux documents anciens, les documents imprimés présentent moins de variabilité spatiale, sont généralement plus structurés et leur mise en page est codifiée. De plus ils présentent moins de dégradations. D'énormes progrès ont toutefois également été réalisés ces dernières années dans le domaine de l'analyse et de la reconnaissance de l'écriture manuscrite, en particulier en ce qui concerne la reconnaissance de mots et de phrases [Bunke 03], mais ceci uniquement dans le cadre d'applications industrielles très spécifiques pour lesquelles les problèmes sont contraints par divers facteurs permettant de réduire la variabilité et la complexité de l'analyse (écriture contrainte, style particulier, utilisation de connaissances fortes sur la structure, vocabulaire réduit, redondance de l'information, ...). Ces progrès concernent le tri automatique du courrier ou encore le traitement des chèques et des formulaires. Le traitement des documents manuscrits non contraints et l'analyse de pages complètes d'écriture restent cependant encore des problématiques de recherche très difficiles [Koch 06] et non encore résolues de manière satisfaisante.

Si pour les documents imprimés, il existe donc maintenant des solutions de GED matures qui intègrent des traitements efficaces pour ce type de documents, ce n'est pas le cas dans le domaine des documents patrimoniaux, car divers facteurs empêchent une gestion efficace et cohérente de ces vastes corpus numériques [MADONNE 06]. Parmi ces facteurs il y a notamment le coût important engendré par les campagnes de numérisation et les sur-coûts souvent occasionnés par un manque de concertation entre les institutions. Cela se traduit généralement par des problèmes de gaspillages de ressources,

d'efforts et d'investissements, ainsi que par la mise en place de solutions non adaptées à l'objectif considéré. De plus, si certaines précautions ne sont pas prises lors du processus de numérisation pour assurer la mise en place et le bon fonctionnement des systèmes de traitement automatique de documents nécessaires à l'indexation, les corpus numériques produits ne sont alors pas exploitables. Le processus de numérisation des documents patrimoniaux est donc un processus long et coûteux qui soulève des nouvelles problématiques de recherche. D'abord parce que la masse de documents à traiter est importante, et d'autre part à cause des spécificités que présentent les documents anciens par rapport aux documents imprimés modernes, et la grande variabilité d'information qu'ils contiennent.

En effet, les documents patrimoniaux, également parfois désignés abusivement sous le terme de documents anciens, englobent un ensemble de documents très variés issus de notre patrimoine historique, culturel, et scientifique, tels que des manuscrits médiévaux, mais également des manuscrits plus modernes tels que des brouillons d'écrivains, des partitions musicales, des rapports techniques et scientifiques, ou encore des documents historiques issus par exemple de registres (registres d'état civil, registres paroissiaux, registres militaires, ...). Ce qui caractérise les documents patrimoniaux, c'est donc une très grande hétérogénéité, à la fois dans les contenus, mais également dans la présentation et dans l'écriture (imprimé, manuscrit), du fait de la variation des styles et des scripteurs. Cette très grande hétérogénéité rend le traitement des documents patrimoniaux très difficile.

Comme cela est expliqué dans [MADONNE 06], l'indexation des documents patrimoniaux nécessite la coopération de différentes processus d'analyse d'images de documents : pré-traitement, analyse de documents structurés, analyse de documents graphiques, reconnaissance de l'écriture manuscrite, ... Le processus de numérisation des documents ne se réduit donc pas qu'au seul processus d'acquisition de facs-similés numériques des documents originaux, mais consiste en une chaîne complète de traitements partant effectivement de l'acquisition jusqu'à la production de contenus numériques structurés hétérogènes (images, transcriptions textuelles, métadonnées, termes d'indexation, hyperliens, ...) permettant la navigation dans les masses numérisées, en passant par un ensemble de prétraitements et de traitements visant à extraire, analyser et structurer l'information [Baird 03]. Pour pouvoir structurer de larges corpus de documents hétérogènes et

permettre l'interrogation de ces bases et la recherche d'information, il faut analyser le contenu des documents pour en extraire des métadonnées permettant de les indexer. Une indexation pertinente des corpus de documents patrimoniaux nécessite d'exploiter toutes les caractéristiques utiles pour les besoins de recherche, ce qui inclut le traitement de différents types d'informations rencontrées dans ces documents, tels que les illustrations, les textes, les styles, les symboles, les annotations manuscrites, ... [MADONNE 06]. Nous allons présenter chacune des étapes de la chaîne de production de documents numériques, en mettant en lumière les différents problèmes et les éventuels verrous technologiques soulevés par chacun des aspects de la numérisation, ainsi que les solutions que l'analyse d'image apporte.

- La numérisation physique ou acquisition de l'image numérique

Il est important lorsque l'on parle de numérisation, de distinguer la numérisation physique, du processus complet de numérisation. La numérisation physique est le processus d'acquisition numérique qui consiste à créer une image numérique du document. L'image numérique du document ainsi obtenue n'est alors qu'un fichier bitmap c'est à dire une simple matrice de points, ou pixels. L'information n'est donc pas structurée, et il ne s'agit que d'un fichier que l'on peut visualiser.

L'objectif de la numérisation des collections patrimoniales, est certes de produire de vastes bibliothèques virtuelles permettant de visualiser des fac-similés numériques des documents numérisés, mais cela implique nécessairement de structurer les données de manière à favoriser la navigation et la consultation au sein de ces vastes collections, ainsi que la recherche de documents pertinents. Numériser c'est donc offrir plus que la simple visualisation des images des documents, c'est aussi apporter des solutions pour structurer et rechercher l'information. Cette étape de numérisation physique, si elle est très importante car elle correspond au début de la chaîne de traitement, et qu'elle pose un certain nombre de choix qui conditionnent l'application finale et l'exploitation que l'on peut faire des documents numérisés, n'est pas l'unique objectif du processus de numérisation, car sinon ces grandes masses de données numérisées seraient inexploitables. C'est pourquoi numériser un document consiste également à associer aux données "image" des informations textuelles (termes d'indexation, transcription, métadonnées)

structurées sur lesquelles des recherches automatiques sont possibles.

Techniquement la numérisation physique des documents est réalisée au moyen d'un dispositif d'acquisition d'images numériques. Il peut s'agir d'un scanneur (ou numériseur) ou d'un appareil photographique numérique. On distingue différents types de scanneur (scanneur à plat, à tambour,...) plus ou moins adaptés à certaines tâches de numérisation. Le premier choix à effectuer lors d'une campagne de numérisation des collections est donc le choix du dispositif d'acquisition. Ce choix est important car l'acquisition d'un dispositif de numérisation de qualité représente un investissement économique conséquent. De plus, il s'agit de pouvoir traiter de grands corpus documentaires très hétérogènes comprenant parfois des documents très précieux et très fragiles. Il est donc primordial de choisir un dispositif d'acquisition adapté et de qualité. D'autant plus qu'une campagne de numérisation est un processus long qui s'échelonne sur de longues périodes, de plusieurs mois à plusieurs années, et qui par conséquent coûte cher. Il faut donc faire le bon choix dès le départ. En matière de préservation et de valorisation du patrimoine, la qualité est importante, et une haute résolution est souvent nécessaire pour restituer les éléments les plus fins, de l'écriture, des graphismes, mais également du support pour certains types de documents très anciens ou très précieux (papyrus, parchemin, papier de riz, ...). Une résolution suffisante est également nécessaire pour pouvoir appliquer les traitements d'analyse ultérieurs nécessaires à l'indexation automatique des documents.

- Prétraitement des images

Les images numériques des documents anciens présentent toujours plus ou moins des défauts visuels. Ces défauts sont liés soit à l'état de conservation du document physique original (tâches, trous, parties manquantes, papier froissé, encre qui traverse le papier, tracés ou écriture partiellement effacés), soit à l'étape de numérisation physique en elle-même (mauvaises conditions d'éclairage, mauvaise orientation du document ou de la caméra, qualité du capteur, présence d'éléments extérieurs, courbure du texte sur les bords due à l'épaisseur du livre, ...). Des traitements numériques permettent de corriger en partie ou totalement ces défauts de l'image, de manière à en améliorer la visualisation et l'analyse par des traitements d'indexation ulté-

rieurs. Ce processus d'amélioration des images est généralement désigné sous le terme de prétraitement. Il peut s'agir d'améliorer le contraste de l'image, de corriger l'orientation du document, de supprimer les bords sombres, de procéder à une séparation fond/forme ...

Pour corriger la luminosité, réduire les bruits et réhausser les constrastes, des opérations sur l'image telles que la modification d'histogramme ou le filtrage sont utilisées. Pour la correction de l'inclinaison ou de la courbure du document, sont utilisées des techniques de rééchantillonnage de l'image suivant un angle d'inclinaison déterminé notamment par des techniques de projection. Beaucoup de travaux ont été menés sur ces aspects de prétraitement des images de documents anciens

Cependant ce qui importe, comme Laurence Likforman l'explique dans [Likforman-Sulem 03], c'est que ces prétraitements doivent être maniés avec prudence, car ils peuvent corriger un défaut mais avoir des effets néfastes sur d'autres éléments de l'image. Ainsi par exemple un filtrage passe-bas permet d'éliminer certains bruits mais rend les tracés d'écriture flous. Cela peut avoir une incidence sur le bon fonctionnement des traitements ultérieurs.

Ces prétraitements peuvent être suivis d'une étape de binarisation, qui vise à séparer le premier plan de l'image comportant les tracés d'écriture et les éléments graphiques, et correspondant aux pixels sombres de l'image, de l'arrière plan du document qui correspond aux zones de pixels clairs. Ce processus produit une image binaire en noir et blanc. L'intérêt est de réduire l'information utile pour les traitements d'analyse ultérieurs, ou pour réduire la taille mémoire des images lorsque l'information couleur n'est pas nécessaire.

Les méthodes classiques de segmentation et de reconnaissance présentées dans la littérature opèrent généralement sur des images binaires, afin de réduire la complexité combinatoire lors de l'analyse, cependant de plus en plus de méthodes opérant directement sur des images en niveaux de gris, voire même sur des images couleur sont proposées. En effet, si pour de nombreux documents imprimés, la séparation du fond et des formes ne pose pas de difficulté, ce n'est pas le cas pour les documents anciens ou pour les documents texturés. Sachant que toute erreur produite lors des phases de prétraitement peut avoir des répercussions importantes sur le résultat final de l'analyse de la structure ou de la reconnaissance, plutôt que d'opter pour une séquentialisation des traitements, il est souvent plus prudent

de limiter les prétraitements et d'intégrer le plus d'information possible dans le processus d'analyse. Si ceci n'était pas possible il y a quelques années, du fait des limitations techniques des ordinateurs, la puissance des machines d'aujourd'hui permet de l'envisager. La couleur est assurément une information importante pour l'analyse des documents, en particulier pour les documents anciens.

- Compression des images

La numérisation des documents patrimoniaux soulève également des problèmes relatifs au stockage et à la diffusion de ces documents numérisés sur les réseaux d'information, locaux ou Internet. En effet, si on considère à titre d'exemple une image couleur RVB de dimensions 900×1400 pixels, sans compression l'encombrement mémoire d'une telle image est de l'ordre de $3,6$ Mo. La taille des documents numérisés devient donc un paramètre critique dès lors qu'il s'agit de stocker ou de permettre l'accès à de grandes bases d'images. Les techniques de compression permettent de réduire la taille mémoire des images en exploitant notamment la redondance de l'information. Au vu des masses de données considérées, de la taille des documents concernés et des débits sur les réseaux, seules des approches de compression avec perte permettent d'obtenir des gains suffisants. Cependant dans le domaine de la valorisation du patrimoine, plus encore que dans d'autres domaines, il est important que les pertes d'information engendrées ne dégradent pas de manière trop importante les images des documents. Il faut donc avoir recours à des méthodes particulièrement adaptées aux documents considérés, de manière à compresser efficacement l'information, sans la détériorer. De nombreuses méthodes de compression d'images avec perte ont été proposées et ont fait leurs preuves, comme le très populaire format JPEG. Cependant le format JPEG est adapté aux images photographiques mais pas aux images de documents, et des gains importants ne sont obtenus qu'au prix de dégradations importantes. Plus récemment de nouveaux formats de compression d'images couleur ont été proposés, comme les formats JPEG2000 [4], DjVu [Bottou 00] ou encore le format de compression adapté aux images de textes imprimés qui a été développé dans le cadre du projet DEBORA [Lebourgeois 03]. Ces deux derniers formats sont particulièrement adaptés

[4] http ://www.jpeg.org/jpeg2000

à la compression d'images de textes, alors que JPEG2000 est comme JPEG adapté aux images photographiques. En effet les images de textes sont des images particulières qui ne peuvent pas être traitées comme des images naturelles. Dernièrement une méthode de compression dédiée aux documents manuscrits a également été proposée par Abed et al. [Abed 05].

Ces méthodes de compression dédiées aux images de documents, telles DjVu ou Debora, permettent d'obtenir à la fois une bonne qualité d'image et des taux de compression importants, en décomposant l'image en objets élémentaires qui seront comprimés ensuite avec des techniques adaptées. Ainsi par exemple le format de compression DjVu [Bottou 00] commence par décomposer l'image du document en différents plans (avant-plan et arrière-plan) à l'aide d'une méthode de segmentation utilisant la combinaison d'un champ de Markov causal bidimensionnel et d'heuristiques fondées sur le principe de Minimum Description Length (MDL) [Niblack 92], puis applique ensuite des techniques de compression différentes pour chacun de ces plans. Le calque de premier plan comportant les caractères et les éléments graphiques est compressé en utilisant des techniques basées sur la redondance des formes (compression de type JBIG2), alors que le calque d'arrière-plan, ainsi débarrassé des différents tracés d'écriture, est compressé plus efficacement en utilisant un algorithme de compression par ondelettes (IW44). Dans la méthode de compression Debora les éléments textuels et les éléments graphiques sont séparés en deux plans distincts. Quatre plans sont donc utilisés au total : une image binaire des éléments textuels qui est compressée efficacement par appariement des formes de caractères sur un ouvrage entier en exploitant la redondance, une image binaire des éléments graphiques compressée sans perte du fait de la faible redondance, une image couleur d'arrière-plan compressée fortement en utilisant la compression JPEG et enfin un plan compensatoire compressé sans perte et contenant les différences entre l'image décompressée et l'image d'origine.

Nous voyons au travers de ces deux exemples, que la compression des images de documents peut bénéficier des résultats des techniques d'analyse et de reconnaissance de la structure du document pour procéder à la séparation des différentes couches informatives, et ainsi pouvoir compresser efficacement l'image du document. La norme définie par JPEG2000 permet également de définir des zones d'intérêt (Region Of Interest ou ROI) dans l'image, et d'appliquer une compression moins importante sur ces régions

et plus importante sur le reste du document, ce qui permet de conserver la qualité sur ces zones d'intérêt tout en ayant un taux de compression important. La norme JPEG prévoit pour l'instant de définir manuellement ces zones d'intérêt, cependant on peut imaginer également vouloir extraire automatiquement certaines zones d'intérêt prédéfinies par l'utilisateur (par exemple certains types d'illustration, ou encore des zones textuelles spécifiques). Là encore cela fait appel à des techniques d'analyse et de reconnaissance de la structure du document. Nous voyons donc que le processus d'analyse des structures n'est pas uniquement un préalable au processus de reconnaissance ou de rétro-conversion du document, mais peut également être exploité efficacement par d'autres traitements comme les traitements de compression d'image et les traitements d'indexation. De même les résultats de ces traitements d'analyse effectués lors du processus de compression peuvent être ré-exploités ultérieurement à des fins d'indexation. Les traitements appliqués lors d'un processus de numérisation sont donc étroitement liés et dépendants les uns des autres. La qualité des résultats de ces traitements est donc primordiale, et plutôt que d'adopter un schéma de traitement purement séquentiel qui présente des risques de propagation et d'amplification des erreurs, il est selon nous important de faire coopérer les traitements de manière à résoudre de manière conjointe les différents problèmes.

- L'analyse de la structure des documents

L'analyse de la structure est une étape importante pour l'indexation automatique des images de documents. Elle peut également bénéficier aux méthodes de compression des images comme nous l'avons vu précédemment. L'indexation des documents est généralement effectuée par le contenu textuel, cependant elle peut également s'effectuer sur les parties graphiques, dans la mesure où de nombreux documents patrimoniaux contiennent également des parties graphiques (figure 1.1). L'indexation peut également bénéficier d'informations sur les structures physiques et/ou logiques du document. Par exemple, certains documents possèdent une mise en page spécifique qui permet de les distinguer aisément au sein d'un corpus. L'analyse de la structure peut permettre également de localiser les zones graphiques afin de procéder à une indexation des images, ou à localiser des zones textuelles

particulières afin de procéder à une reconnaissance partielle du contenu et permettre ainsi les recherches plein texte ou la création d'index (tables des personnages, des lieux,...).

Fig. 1.1. Exemples de parties graphiques dans des documents patrimoniaux, d'après [MADONNE 06]

L'analyse de la structure (physique ou logique) des documents anciens, n'est pas un problème simple à résoudre de part la grande diversité de présentation et de style des documents dans les masses de données documentaires numérisées, des dégradations importantes de certains documents, et du fait de la forte variabilité spatiale de la structure et de l'écriture inhérente à certains types de documents comme les documents manuscrits notamment. De ce fait, les méthodes traditionnelles proposées dans la littérature pour l'analyse des documents structurés imprimés (que nous abordons dans le chapitre suivant), ne sont pas adaptées aux documents patrimoniaux, car trop spécifiques à certains types de structures et pas assez souples pour s'adapter à la forte variabilité en présence. Il est donc primordial de développer des

méthodes et des outils pouvant s'adapter à un grand nombre de structures et de cas d'usages. Dans le cadre du traitement des documents anciens, l'important n'est pas nécessairement d'avoir des performances exceptionnelles, mais de pouvoir traiter automatiquement et correctement le plus de tâches possibles en se basant sur des interactions avec l'utilisateur de manière à satisfaire au maximum ses besoins.

1.2 Travaux existants en matière d'aide à l'indexation de documents anciens

Un certain nombre de projets se sont intéressés à la spécification des besoins et usages suscités par la numérisation des documents patrimoniaux ainsi qu'à « l'utilisation de l'informatique pour le traitement des documents d'autrefois » comme l'expriment Jacques André et Marie-Anne Chabin en introduction du premier numéro spécial consacré aux documents anciens de la revue "Document Numérique" [André 99]. Ces projets, souvent pluridisciplinaires et impliquant la collaboration de personnes issues de domaines différents, bibliothécaires, archivistes, chercheurs en Sciences Humaines, chercheurs en informatique, se concentrent généralement sur certains aspects liés à la numérisation, tels que la compression des images, le couplage texte/image, ou encore la structuration et l'indexation de l'information. Ces problématiques liées aux traitements des documents anciens connaissent un regain d'intérêt dans la communauté scientifique ces dernières années. Ainsi de plus en plus de numéros spéciaux de revues scientifiques sont consacrés à ces problématiques [André 99][Coüasnon 03b], et des conférences scientifiques centrées sur ces thèmes sont mêmes apparues très récemment, comme par exemple la conférence internationale Document Image Analysis for Libraries (DIAL). Nous présentons dans ce qui suit quelques projets et réalisations importantes dans le domaine de l'aide à l'indexation de documents patrimoniaux.

- **Le Poste de Lecture Assistée de la Bibliothèque Nationale de France**

Le projet de développement d'un poste de lecture assistée par ordinateur (PLAO), mené par la Bibliothèque Nationale de France (BNF) à Paris de

1990 à 1993, est l'un des premiers projets portant sur la réalisation de postes de travail dans le domaine des hypermédias littéraires, à s'être intéressé aux problématique de l'accès, de la visualisation et de l'édition de sources issues de documents numérisés [Virbel 93]. L'objectif pour la BNF était de mettre en place une station de travail permettant l'accès, la lecture et l'annotation de contenus documentaires numérisés issus de ses collections, en offrant une aide à la lecture, ainsi que tous les outils nécessaires aux étudiants et aux chercheurs pour mener à bien leurs travaux de recherche et d'études documentaires. Il ne s'agissait donc pas de se limiter à la lecture, mais également de fournir des outils de travail et de traitement pour l'étude et l'annotation des sources numérisées. Malheureusement ce projet ambitieux n'a pas pu être mené à son terme et aboutir à cet objectif, mais il a permis d'ouvrir la réflexion et de poser un certain nombre de points fondamentaux sur la manipulation des documents numérisés. Il a de plus inspiré beaucoup de d'autres projets par la suite.

- **Le projet PHILECTRE**

Le projet PHILECTRE (PHILologie ELECTRoniquE) [Likforman-Sulem 97], mené de 1994 à 1997, est un projet pluridisciplinaire qui se base sur les travaux et les conclusions de l'expérience du Poste de Lecture Assistée de la BNF. Ce projet avait pour but d'explorer les techniques informatiques nécessaires aux chercheurs en sciences littéraires, notamment les généticiens et les médiévistes. L'idée était en quelque sorte de poursuivre le travail mené lors du projet de Poste de Lecture de la BnF et de reprendre là où ce projet s'était arrêté. Comme l'explique Eric Lecolinet dans [Lecolinet 99], pour les généticiens des textes une oeuvre littéraire ne se réduit pas à un "texte canonique", mais consiste en une collection de données hétérogènes, documents manuscrits originaux, transcriptions, commentaires, apparât critique, ... Or le développement des techniques numériques et hypermédia rend possible la réalisation de documents numériques composés d'un grand nombre de données hétérogènes, encore faut-il disposer d'outils adaptés à cette tâche. Ce projet PHILECTRE a donc abouti entre autres à la réalisation d'un prototype de poste de lecture et d'édition de documents numérisés, sous forme d'hypermédias [Lecolinet 99]. Ce prototype se base sur une représentation duale

texte/image des manuscrits et fournit un certain nombre d'outils destinés à faciliter le couplage entre les deux modes de représentation [Robert 97]. Il est dédié à la visualisation des sources numérisées mais également à l'édition de transcription de ces sources. Différents outils de sélection ou de saisie sont proposés à l'utilisateur pour faciliter la production de ces transcriptions. Il s'agit de transcriptions diplomatiques encodées selon une DTD XML/TEI. Le point intéressant du projet est qu'un certain nombre de travaux ont été menés sur l'apport des techniques et outils du traitement d'image, de l'analyse d'images de documents, du domaine de l'interaction homme-machine et plus généralement des techniques informatiques, à la production de bibliothèques virtuelles et d'éditions hypertextuelles. Ainsi par exemple dans le prototype proposé, des modules d'extraction de lignes de texte permettent de localiser automatiquement les parties textuelles ce qui facilite la saisie des transcriptions textuelles. Un module d'extraction interactive de traits, tels que les traits de biffures ou les lignes de renvoi, est également intégré [Likforman-Sulem 98]. Ces modules de traitement sont basés sur des processus collaboratifs, c'est à dire qu'ils sont supervisés par l'utilisateur, et que celui-ci a la possibilité d'interagir et de corriger les erreurs d'analyse. Ce point nous semble très important car sur des documents aussi complexes que les manuscrits anciens on ne peut pas s'appuyer sur des résultats de modules complètement automatiques. Des travaux ont également été menés sur la structuration et le codage de la représentation textuelle des manuscrits (transcriptions) et sur la comparaison des différentes versions du texte. Un éditeur de documents structurés gérant les multiples versions, et appelé THOT a notamment été développé dans le cadre de ce projet pour le codage des textes. Ce codage se base sur le méta-langage XML (initialement sur SGML puis ensuite sur XML) et la DTD TEI. Dans le cadre de ce projet, un certain nombre de travaux [Lecolinet 98] [Lecolinet 99][Robert 00], ont aussi été menés sur la problématique de la navigation et du repérage dans les documents hypermédias. En effet la représentation hypertextuelle bouleverse considérablement les habitudes de lecture. Des solutions inspirées des techniques de visualisation de l'information ont donc été proposées. Au travers de ce projet on voit donc que la numérisation des manuscrits pose un certain nombre de questions et de problèmes, à la fois sur le choix des modes de représentation, sur le passage entre ces modes de représentations, sur

l'analyse et l'indexation des sources numérisées, sur la constitution de documents structurés hypermédias à partir de ces sources hétérogènes, et sur l'édition et la navigation dans ces hypermédias. On pourra trouver dans [Likforman-Sulem 03] un pananorama de ces problématiques et un éclairage au travers des expériences menées au cours de ce projet PHILECTRE, qui est selon nous le projet le plus abouti dans le domaine de la génétique et de la philologie électronique.

- **Le projet DEBORA**

Le projet DEBORA (Digital accEss to BOoks of the RenAissance) [Belisle 99], est un projet européen qui s'est déroulé de 1999 à 2001, dont l'objectif était de développer un système facilitant l'accès à des collections de manuscrits numérisés du XVIe siècle, et permettant un travail collaboratif d'annotation sur ces sources numérisées. Le projet DEBORA s'est principalement concentré sur l'étude des formats d'image, dans le cadre de la constitution de bibliothèques numériques consacrées aux documents de la Renaissance, et un format de compression d'images, dédié aux ouvrages numérisés a été spécifiquement développé lors de ce projet. Ce format de compression est assez proche du format DjVu. Cependant la particularité de ce format de compression est qu'il n'est pas destiné aux pages de documents, mais aux ouvrages entiers. En effet l'algorithme de compression proposé exploite la redondance des informations sur toutes les pages d'un même ouvrage. Ce format de compression permet non seulement d'obtenir des taux de compression importants de l'ordre de 60 :1 mais à l'instar du format DjVu qui permet un affichage des différentes couches du document, il offre en plus aux usagers des fonctionnalités intéressantes comme la transcription assistée, l'étude de la régularité de la construction de l'ouvrage, la recherche d'illustrations ou de mots dans plusieurs ouvrages à la fois. Hormis ce format de compression d'image, le projet DEBORA a abouti à d'autres réalisations, dont un format structuré de gestion de documents numérisés intitulé "AdHoc" (Auto-Documented & Hierarchically Organized Contents format), représentant les documents par leurs images, leurs contenus textuels et un certain nombre de méta-données telles que la description de l'ouvrage, sa structure ou des commentaires associés. Un environnement spécifique pour la manipulation et l'annotation de sources numérisées du

XVIème siècle a également été développé sous la forme d'un démonstrateur librement téléchargeable (http ://rfv6.insa-lyon.fr/debora/client.htm). Ce démonstrateur permet la création et l'édition de projets constitués de sources numérisées, la structuration et l'enrichissement de ces sources numérisés par la description de métadonnées ou encore la recherche d'informations (mots, éléments graphiques, annotations...) dans plusieurs ouvrages à la fois. Les travaux menés dans le cadre de ce projet sont très intéressants, cependant certains aspects comme l'analyse de la structure physique ou logique n'ont pas été abordés, et de plus ils se sont concentrés uniquement sur les documents de la Renaissance qui ont des particularités propres par rapport à d'autres types de documents anciens.

- **Le projet BAMBI**

Le projet BAMBI, Better Access to Manuscripts and their Images, est également un projet pluridisciplinaire européen, lancé à l'occasion du programme "Télématique" de la Communauté Européenne [Calabretto 98]. Comme le projet DEBORA ce projet concerne également l'accès, la visualisation, la transcription, l'annotation et l'indexation de documents numérisés du XVIème siècle. Cependant d'autres aspects ont été abordés lors de ce projet, notamment l'analyse de la structure des documents et l'aspect couplage texte/image. Ce projet a en effet abouti à la réalisation d'une station de travail permettant l'annotation des sources numérisées et la constitution de documents hypermédias (figure 1.2).

Le couplage texte/image est réalisé grâce à des modules de segmentation des images des documents afin de détecter automatiquement les lignes de texte et les mots. Ces modules de segmentation utilisent une méthode simple de détection basée sur l'analyse des histogrammes de projection des pixels de l'image du document. Le couplage texte/image est ensuite effectué ligne par ligne par appariement entre les mots segmentés et les mots de la transcription saisie par l'utilisateur. Une telle méthode fonctionne bien dans le cas des documents de la Renaissance car les lignes d'écriture sont droites et relativement bien espacées, de mêmes que les mots. Ce n'est pas le cas de nombreux documents anciens. Le codage de la représentation textuelle des documents ainsi que le codage des liens entre les mots de

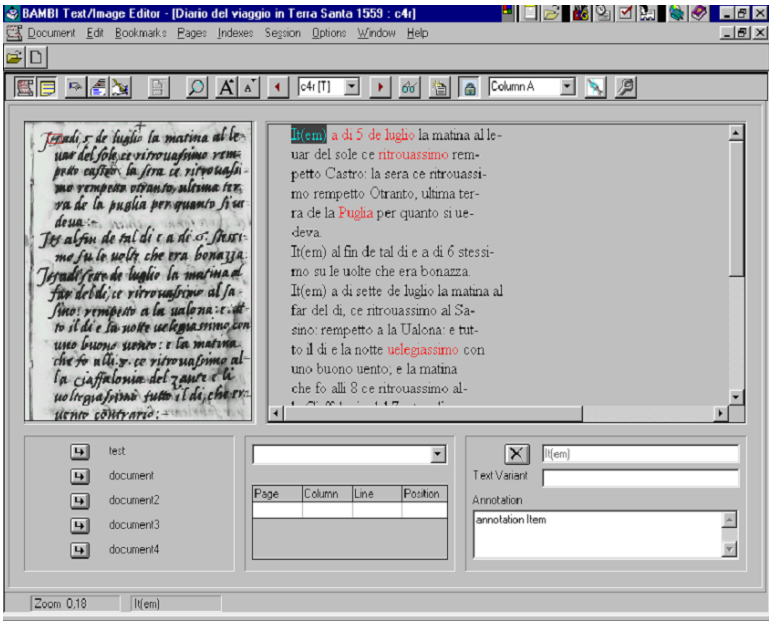


Fig. 1.2. Station de travail BAMBI

cette représentation et les portions d'image correspondantes est basé sur l'utilisation du méta-langage SGML et sur la norme de codage HyTime. Une DTD spécifique a également été développée dans le cadre de ce projet. Ce codage permet de faciliter la recherche par mots-clés, dans les transcriptions ou directement dans les images des documents numérisés.

- **Système d'annotation collaborative de manuscrits de Carrera [Carrera 05]**

Carrera propose un système permettant l'annotation collaborative de sources documentaires, qu'il s'agisse de documents anciens historiques, de documents imprimés ou manuscrits. Le système proposé comporte trois composants logiciels principaux :

– un module d'archivage et d'accès aux sources (archive assistant)
– un module d'aide à la transcription (transcription assistant)
– un module d'enregistrement des contributions et de gestion de crédits (contribution accountant)

Le module d'archivage permet de gérer l'accès aux sources numérisées et aux éventuelles métadonnées qui leur sont attachées. Ces sources sont stockées de manière centralisée dans une base de données hébergée sur une machine serveur. L'image et les métadonnées associées sont encapsulées dans un fichier au format XPG (eXtended jPeG). La deuxième application composant le système est une interface d'aide à la transcription (voir figure 1.3). Cette interface qui peut être utilisée en ligne en mode connecté ou en local de manière indépendante sur la machine de l'utilisateur, permet d'éditer de manière relativement simple ses propres transcriptions de sources numérisées ou des transcriptions existantes. Les transcriptions réalisées sont encodées dans un format basé sur le standard XML afin de permettre les échanges de données et la recherche d'information. Les auteurs ont défini leur propre format appelé Manuscript Markup Language (MML), pour l'encodage des transcriptions, mais ne fournissent pas plus de détails sur ce format d'encodage. L'interface permet la visualisation simultanée de l'image du document et de sa transcription, selon différentes vues, la vue du document XML ou une prévisualisation d'une version imprimable. La transcription des documents s'effectue au niveau mot. Après un paramétrage réalisé par l'utilisateur, le système est capable de localiser automatiquement les mots dans l'image du document, en utilisant une technique de lissage (smearing). Afin de visualiser les mots localisés ceux-ci sont encadrés par leur boîte englobante. L'utilisateur a ensuite la possibilité de rentrer la transcription de chaque mot dans une fenêtre de saisie apparaissant à côté de la boîte englobante du mot, par simple clic sur la boîte. Il a également la possibilité d'annoter simplement les zones transcrites à l'aide d'un système de menus contextuels. Cette annotation concerne la distinction entre éléments graphiques ou symboles, et texte. Pour les zones textuelles, plusieurs "tags" sont proposés pour distinguer notamment les abréviations, les corrections, ou les changements de scripteurs. Afin de distinguer les zones annotées des zones non encore annotées ou encore les zones textuelles des zones graphiques, des couleurs différentes sont utilisées pour afficher les boîtes englobantes. Dans la fenêtre de prévisualisation de la transcription les mots apparaissent à la même position relative que leur boîte englobante dans l'image. Dans le cas des zones graphiques l'imagette de la zone est extraite et insérée telle qu'elle dans la transcription, à la position correspondante.

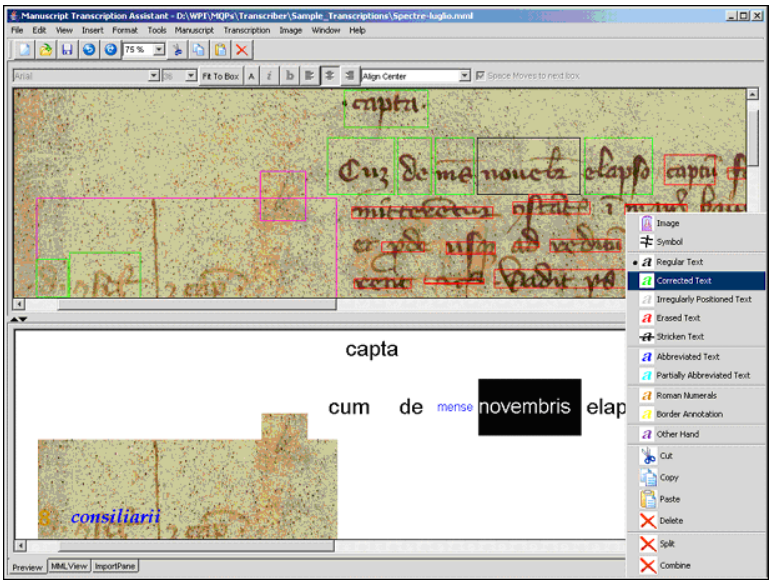


Fig. 1.3. système d'annotation collaborative de Carrera [Carrera 05]

La dernière composante logicielle du système permet l'attribution et la gestion de "crédits" pour les utilisateurs ayant participé à l'effort de transcription. Ces crédits offrent à l'utilisateur la possibilité d'utiliser des services complémentaires, notamment la possibilité de consulter les transcriptions réalisées par d'autres utilisateurs. Ce composant logiciel tourne également sur une machine serveur. L'ensemble de ce système a été développé en utilisant des langages et des standards ouverts tels que Java et XML, et est de ce fait parfaitement portable ce qui est important pour ce type d'application.

- **L'environnement d'indexation de documents anciens AGORA**

Un environnement dédié à l'analyse et à l'indexation d'images de documents anciens, appelé AGORA, a été proposé récemment par Ramel et al. [Ramel 05][Ramel 06]. Ce système a été développé dans le cadre du projet "Bibliothèques Virtuelles Humanistes", mené par le "Centre d'Etudes Supérieures de la Renaissance" (CESR) de la ville de Tours. Ce projet concerne plus spécifiquement la numérisation et l'indexation d'ouvrages datant de la Renaissance. Ces documents sont très variables dans leurs structures et les métadonnées à extraire peuvent être très différentes en

fonction des ouvrages et de l'objectif des tâches d'indexation considérées.
Parmi les nombreuses spécificités de la structure des documents de la Renaissance, les auteurs mettent en avant le fait qu'il n'existe généralement pas de style éditorial ou de structure logique facilement identifiable, que ces documents comportent de nombreux styles et fontes d'écriture, ainsi que de nombreuses ornementations, qu'il existe une forte variabilité à la fois dans la forme des blocs de texte qui ne sont pas toujours rectangulaires, dans la position des illustrations et de leurs légendes, mais également dans la taille des espaces séparant les caractères, les mots et les blocs de texte, et qu'il est fréquent de se trouver en présence de superpositions des différentes couches informatives (texte imprimé et notes manuscrites) ou non-informatives (bruit). Cette liste n'est évidemment pas exhaustive, et chaque document possède ses particularités. La caractéristique importante qui ressort donc, est la forte variabilité. Il est donc nécessaire de prendre en compte ce paramètre important dans la mise en oeuvre des traitements d'analyse, ce qui nécessite notamment une certaine souplesse et une certaine robustesse dans les solutions retenues.
C'est pourquoi Ramel et al. ont choisi une approche d'indexation interactive plutôt qu'une approche complètement automatique. Le système AGORA intègre donc des outils d'analyse d'images et manipule des scénarii d'indexation que l'utilisateur peut définir et modifier selon ses besoins. Ce système montre une certaine souplesse et une certaine efficacité puisqu'il est aujourd'hui utilisé de manière fonctionnelle par le CESR dans la réalisation de son projet de bibliothèque virtuelle. Cependant son inconvénient est qu'il nécessite malgré tout quelques connaissances en analyse d'images et n'est donc pas utilisable par n'importe qui sans un temps d'apprentissage et de prise en main du système. D'autre part, les outils d'analyse qu'il intègre sont selon nous bien adaptés aux structures présentées par les documents imprimés, mais moins à celles moins contraintes des documents manuscrits par exemple.

- Autres travaux sur le traitement des documents patrimoniaux

D'autres travaux ont été menés sur le traitement et l'analyse de documents anciens. Nous ne pouvons pas tous les détailler dans le cadre ce mémoire, mais nous pouvons citer notamment les travaux de Journet et al.[Journet 05][Journet 06b][Journet 06a], sur l'analyse de documents de la

Renaissance et la séparation texte-graphique ou encore les travaux de Coüasnon sur l'accès par le contenu aux documents manuscrits d'archives numérisés, à l'aide d'un système d'analyse baptisé DMOS (cf figure 1.4) reposant sur la description de la structure du document par un ensemble de règles de grammaire 2D [Coüasnon 03c][Coüasnon 03a][Coüasnon 04]. Ce système a également montré son efficacité sur des tâches d'indexation de grands corpus documentaires [Coüasnon 02], cependant la méthode d'analyse utilisée, ne peut s'appliquer qu'à des documents spécifiques et fortement structurés (partitions musicales, structures tabulaires, formulaires). De plus l'adaptation de ce système à différents types de documents, nécessite systématiquement la redéfinition de la grammaire, ce qui ne peut être fait que par des experts du système et représente une tâche assez fastidieuse.

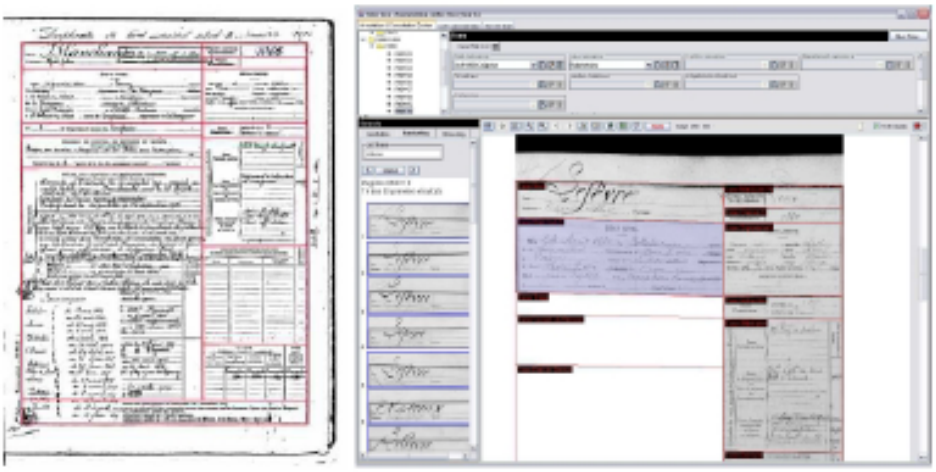


Fig. 1.4. Extraction de la structure physique et reconnaissance de champs spécifiques à l'aide du système DMOS [Coüasnon 03c]

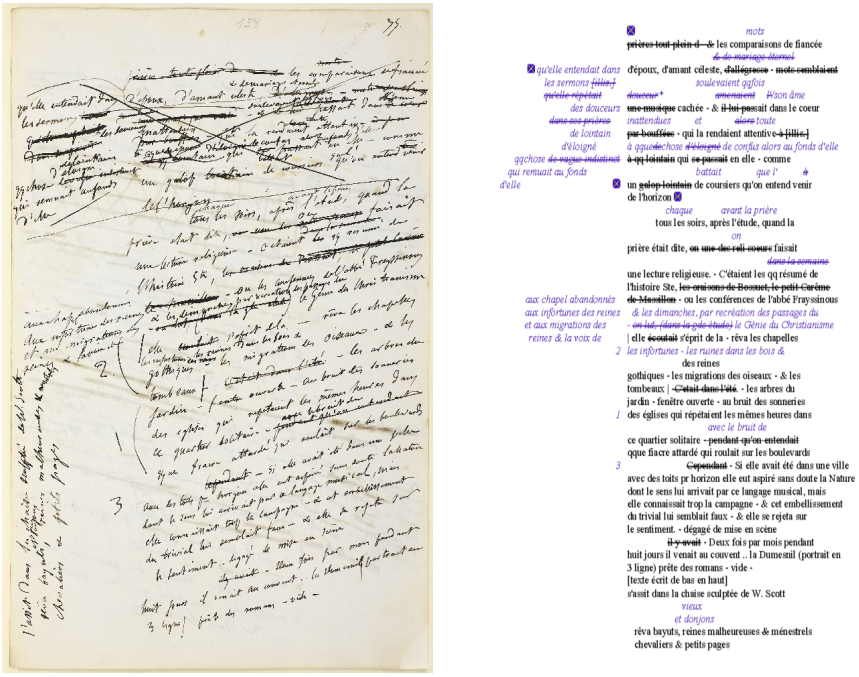
Tous ces projets dédiés à la numérisation des documents anciens, et à la production de bibliothèques numériques et de représentations numériques hypermédias, ont permis de poser un certain nombre de points importants, notamment sur ce que l'analyse d'images de documents peut apporter à la valorisation des sources numérisées. Ces projets ont souvent abouti à des prototypes fonctionnels ou à l'établissement de spécifications. Cependant ces résultats doivent être pris en compte comme des résultats préliminaires dans le cadre de projets prospectifs. Il n'existe pas encore à l'heure actuelle de solution satisfaisante pour traiter les documents anciens et tous les besoins

sont loin d'être satisfaits. Ainsi Thierry Delcourt, actuel directeur de la médiathèque de l'agglomération troyenne, exprimait en 2000 dans le Bulletin des Bibliothèques de France [Delcourt 00] les limitations de ces expériences en ces termes : "Pour le bibliothécaire, soucieux de simplicité et de standardisation, ces expériences trouvent toutefois leurs limites. Conduites par des équipes associant informaticiens et universitaires, elles risquent en effet de mener aux mêmes impasses que feu le PLAO (poste de lecture assistée par ordinateur) de la Bibliothèque Nationale de France. La complexité de ce dernier, conséquence du souci louable d'y intégrer le maximum d'attentes d'un maximum de chercheurs de tous horizons, l'a finalement rendu irréalisable et obsolète, car l'offre standard du commerce s'était suffisamment enrichie entre-temps pour couvrir la plupart des besoins exprimés à l'origine. Les bibliothécaires préféreront sans doute considérer ces projets comme des expériences pilotes, et chercher sur le marché les outils qui permettront de répondre, de manière standardisée, à des besoins moins spécifiques, mais plus généralisables." Cela montre bien qu'il reste beaucoup à faire pour montrer le réel apport de l'informatique à la numérisation du patrimoine, et la nécessité de structurer l'information et de ne pas se contenter de mettre en ligne simplement les images numérisées. Cela passe par le développement d'outils et de méthodes adaptés, robustes et simples d'utilisation.

1.3 Contexte de nos travaux : le projet Bovary

Les travaux de thèse que nous présentons dans ce mémoire s'inscrivent dans le contexte d'un projet de numérisation et de valorisation de documents anciens, et plus particulièrement de manuscrits d'auteurs [Nicolas 03]. Ce projet intitulé Projet Bovary est un projet de numérisation de manuscrits de Gustave Flaubert initié par la Bibliothèque Municipale de la ville de Rouen et mené en partenariat avec le Centre Flaubert du laboratoire CEREDI de l'Université de Rouen, et le laboratoire LITIS (ex PSI) au sein duquel nous menons nos travaux. L'ensemble des brouillons de "Madame Bovary" constituant la genèse de cette oeuvre emblématique de Gustave Flaubert, est concerné par ce projet. Au delà de la simple production de substituts numériques sous forme d'images haute définition, l'enjeu majeur de ce projet est la réalisation d'une véritable édition numérique sous une forme hypertextuelle, permettant de naviguer dans l'avant-texte du roman à la fois en mode

image, en visualisant les substituts numériques des brouillons, ou en mode texte, par l'intermédiaire de transcriptions textuelles, et ceci selon différents scénarii de navigation. Les brouillons de Flaubert étant particulièrement difficiles à déchiffrer, même pour des spécialistes, a été fait le choix de proposer des transcriptions textuelles dites "diplomatiques", c'est à dire des transcriptions qui sont les plus fidèles possibles à la mise en page du document original. Dans ces transcriptions, des conventions typographiques, de codage et de mise en page sont utilisées pour reproduire l'apparence des manuscrits (figure 1.5). La difficulté avec les transcriptions diplomatiques est de savoir jusqu'où la similitude entre l'original et sa transcription doit et peut aller. Cette limite est souvent fixée par les contraintes liées à la réalisation de ces transcriptions, et aux outils utilisés pour cela.



(a) (b)

Fig. 1.5. Un brouillon manuscrit (a) et sa transcription diplomatique (b)

La réalisation d'un projet aussi ambitieux présente de nombreuses difficultés. Tout d'abord se pose le problème délicat de la numérisation des brouillons. Quelle solution de numérisation retenir ? Quel format d'image choisir ? Ensuite puisqu'il s'agit de produire une édition génétique se pose

le problème de l'ordonnancement des sources manuscrites. Se pose également le délicat problème de la production des transcriptions textuelles. Et enfin, se pose le problème du choix d'une architecture et de solutions techniques pour la réalisation de l'édition hypertextuelle, qui soulève des questions concernant l'indexation des documents et la structuration du corpus. Chaque participant à ce projet s'est donc concentré sur un aspect spécifique en fonction de son champ d'expertise et de ses domaines de compétence. Les aspects relatifs à la numérisation des brouillons et à la réalisation de l'édition hypertextuelle finale ont été confiés à la Bibliothèque Municipale de Rouen qui est l'instigatrice du projet, les aspects liés à la transcription des brouillons, à la production des contenus éducatifs et pédagogiques, et au classement génétique du corpus ont été appréhendés par le Centre Flaubert, et le laboratoire LITIS a apporté son expertise en matière d'analyse d'images de documents et d'analyse de l'écriture manuscrite afin de proposer des solutions d'aide à l'indexation et à la transcription de grandes masses de documents.

La numérisation de la totalité du manuscrit, soit 4549 pages, a été réalisée de fin 2002 à septembre 2003 par la BMR, à l'aide d'un appareil photographique numérique haute résolution. En ce qui concerne la structuration de l'édition hypertextuelle a été fait le choix de se baser sur le classement génétique des brouillons établi dans le cadre de la thèse de Marie Durel réalisée sous la direction du Professeur Yvan Leclerc du Centre Flaubert [Durel 00]. Ce classement fournit un ordonnancement génétique des brouillons selon deux axes : la chronologie du récit (axe syntagmatique) et la chronologie des étapes successives de rédaction d'un même passage (axe paradigmatique) [Grésillon 94][Crasson 04]. Il peut être représenté par un graphe appelé graphe génétique qui définit les correspondances et relations génétiques entre les différents folios (figure 1.6). Afin d'étudier les solutions envisageables en termes de visualisation, de navigation et d'architecture de l'édition hypertextuelle finale qui sera développée, un prototype a été réalisé courant 2003 par le laboratoire LITIS pour les aspects techniques. La réalisation de ce prototype qui est consultable en ligne[5], a permis d'engager une réflexion sur les fonctionnalités de navigation et de visualisation des données qui seront proposées dans l'édition hypertextuelle finale.

Outre la numérisation et le classement des brouillons, une part im-

[5] www.univ-rouen.fr/psi/BOVARY

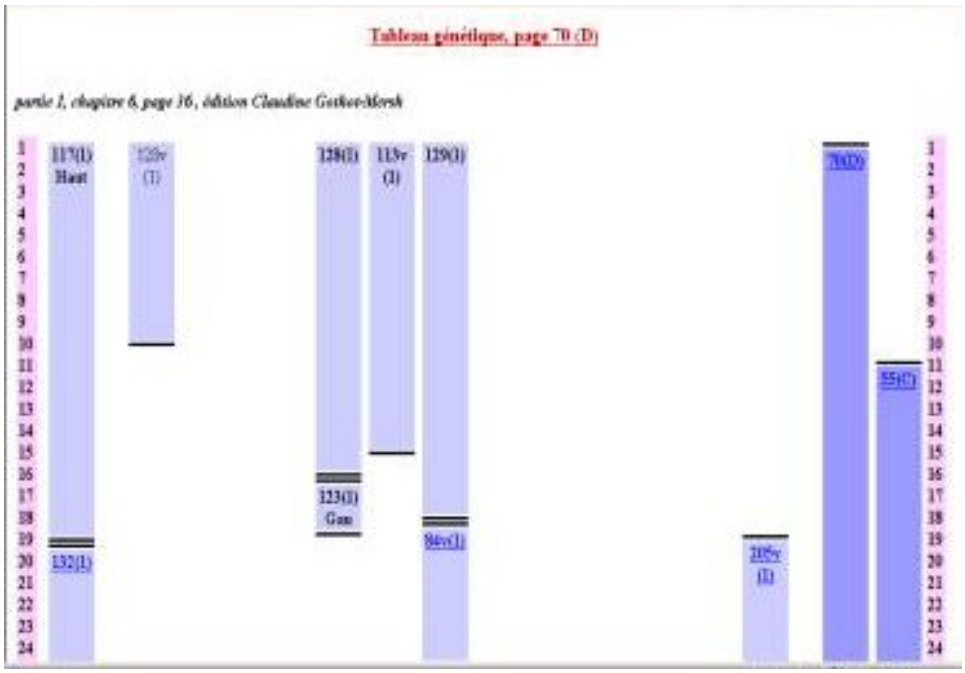


Fig. 1.6. Ordonnancement génétique des brouillons selon l'axe paradigmatique et l'axe syntagmatique (les brouillons sont désignés selon la foliotation établie par la BMR et les numéros sur les côtés sont les index des lignes de texte dans les brouillons)

portante du travail qu'implique ce projet, réside dans la transcription des brouillons. Transcrire presque 5000 folios manuscrits est un tâche titanesque, surtout lorsque l'on considère l'extrême complexité et l'aspect chaotique des brouillons de Gustave Flaubert (figure 1.7).

Afin de permettre la production des transcriptions diplomatiques de tous les folios manuscrits du corpus, un appel à transcripteurs volontaires a donc été lancé par le Centre Flaubert qui coordonne les travaux de transcription et de création des contenus de l'édition hypertextuelle dans le cadre du projet. L'intérêt de constituer un tel réseau de transcripteurs réside dans le fait de pouvoir répartir la charge de travail et ainsi transcrire à moindre coût l'ensemble du corpus. Cependant ce système demande un effort important de coordination, d'homogénéisation et de vérification. De plus, faute d'outils informatiques adaptés à ce travail fastidieux, les transcripteurs utilisent généralement les outils qu'ils maîtrisent le mieux, c'est à dire des éditeurs de texte ou éditeur HTML (Microsoft Word, Microsoft Frontpage, Dreamweaver,...). Les transcriptions diplomatiques sont donc produites dans un

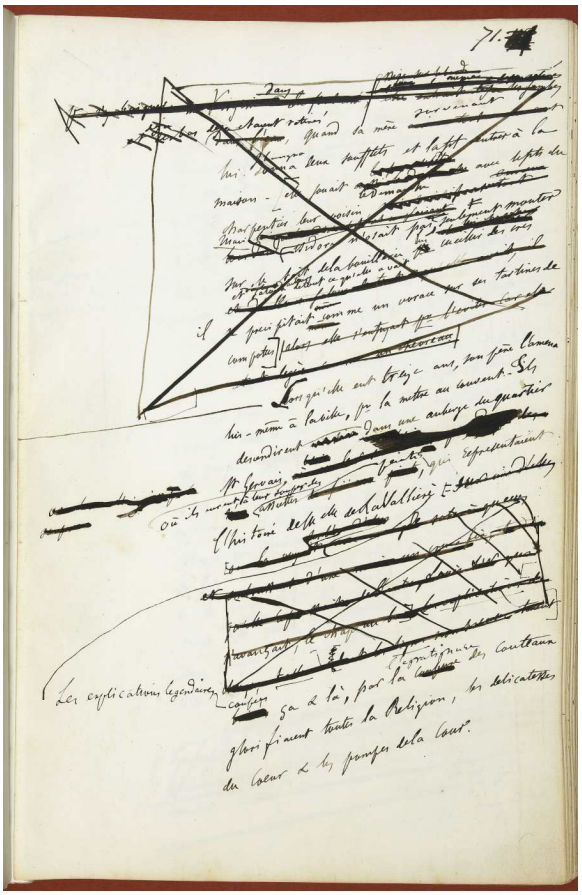


Fig. 1.7. Exemple de brouillon permettant d'apprécier la complexité des manuscrits autographes de Gustave Flaubert

format de document propriétaire peu ou pas structuré, ou dans le meilleur des cas en HTML, c'est à dire dans un format qui code le rendu à l'écran du document, et non pas sa structure logique (nous reviendrons dans le chapitre suivant sur ces notions de structure physique et logique). Cela ne facilite pas l'indexation de l'information, ni la réutilisabilité des documents produits.

La transcription des manuscrits est une phase nécessairement préalable à toute analyse du texte et à toute interprétation génétique. Bien que préalable et n'étant pas une finalité en soi, cette tâche de transcription représente une part très importante du travail du généticien des textes. De plus la transcription des images de documents permet la génération d'index nécessaires à toute recherche d'information dans les documents. Cette phase de transcription nécessaire est cependant très fastidieuse. Faute de pouvoir disposer d'un éventuel travail de transcription déjà effectué par d'autres avant lui,

lorsqu'un chercheur souhaite effectuer de nouvelles analyses, proposer de nouvelles interprétations et publier une édition critique, il est obligé de refaire lui-même ce travail de transcription. Cela représente une perte de temps considérable. D'autant plus que le travail de transcription peut-être relativement fastidieux en fonction des documents considérés. La transcription d'un folio manuscrit peut prendre de quelques heures à une journée entière, voire plusieurs jours si l'écriture est particulièrement difficile à déchiffrer. Pourtant ce travail de transcription pourrait être facilité par l'apport d'outils d'analyse automatique d'images de documents permettant d'extraire un certain nombre de métadonnées de manière automatique ou interactive, à défaut d'envisager une rétro-conversion qui serait illusoire sur ce type de documents fortement perturbés et parfois dégradés.

Dans le cadre du projet Bovary, afin de faciliter le travail de transcription et d'indexation des manuscrits du corpus, un environnement d'aide à la transcription et à l'indexation de documents a été développé. Cet environnement baptisé EMMA (Edition de Manuscrits Moderne Assistée), permet à l'utilisateur de définir des zones d'intérêt dans l'image d'un document, puis de transcrire et d'attacher des métadonnées d'indexation aux zones ainsi définies. L'ensemble des informations ainsi recueillies (coordonnées des zones, transcriptions, métadonnées d'indexation) sont stockées dans un format structuré défini par une DTD (Document Type Definition) XML. Nous avons pour cela établi un modèle de document XML, que nous avons baptisé GustaveML, spécifiquement adapté aux manuscrits modernes. Cependant ce schéma d'encodage peut être adapté et étendu à d'autres types de documents. Cette représentation numérique et structurée dans un format ouvert tel XML facilite l'échange des documents et l'interopérabilité, ainsi que la réutilisabilité des données. Ainsi à partir de cette représentation il est aisé de formater le document de différentes manières suivant les besoins et de produire par exemple des transcriptions diplomatiques au format HTML pour la visualisation sur Internet, ou des transcriptions au format PDF pour l'échange, l'archivage ou l'impression, simplement par application de techniques de transformation (XSLT) et de formatage (CSS, XSL-FO). Nous renvoyons à la lecture de l'annexe I pour une présentation plus technique et plus complète des fonctionnalités d'EMMA. Il est important de noter que cette interface n'est aucunement limitée au traitement des manuscrits de Flaubert. Dans sa version actuelle cependant l' indexation est réalisée

de manière totalement manuelle, et l'environnement proposé, s'il permet de structurer les informations attachées aux documents, ne facilite pas complètement le travail de l'utilisateur.

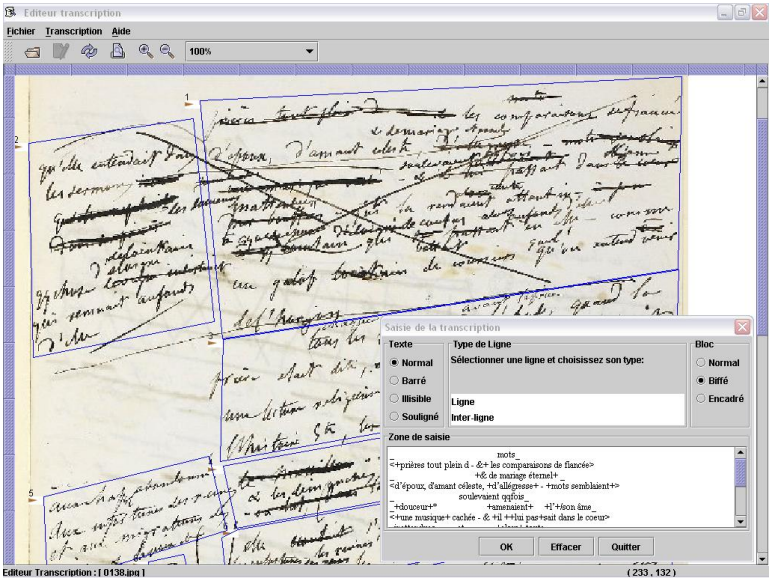


Fig. 1.8. Interface d'aide à l'indexation EMMA

Dans le cadre de nos travaux de thèse et de ce projet, nous nous sommes donc plus particulièrement intéressés à la problématique de l'aide à la production de transcriptions textuelles pour l'indexation des documents, grâce à l'apport d'outils informatiques permettant d'analyser de manière automatique ou semi-automatique les images des documents afin d'en extraire les structures, ce qui permet de faciliter le travail d'annotation et d'indexation, et de produire des contenus structurés permettant la création de véritables éditions et bibliothèques numériques. L'objectif de nos travaux est de développer des méthodes robustes de segmentation et de reconnaissance de la structure de pages de documents pouvant être intégrées dans un environnement d'aide à l'indexation comme l'éditeur EMMA que nous venons d'évoquer, ceci afin d'automatiser un certain nombre de tâches fastidieuses et répétitives, et permettre d'extraire de manière supervisée des métadonnées pouvant servir à l'indexation des documents. Si les techniques permettant de traiter des documents plus simples dans un contexte industriel, sont matures, encore peu de travaux ont été réalisés sur l'analyse de documents anciens. Certains aspects comme

la compression des images, la structuration des données ou le couplage entre texte et image ont été abordés ces dernières années dans le cadre de projets pluridisciplinaires, mais la problématique de l'extraction de structures dans des documents anciens n'a pas encore été résolue de manière satisfaisante, c'est pourquoi nous nous intéressons plus particulièrement à cet aspect.

L'intérêt des travaux que nous présentons ne se limite bien évidemment pas qu'au contexte de la transcription des manuscrits de Gustave Flaubert, puisque la problématique de l'analyse automatique d'images de documents pour l'indexation, est centrale dans tous les projets de numérisation des documents anciens, quels que soient les documents concernés. Il est donc important selon nous que les solutions mises en oeuvre soient suffisamment génériques ou tout au moins adaptables à différents contextes de numérisation de documents anciens voire même de documents imprimés modernes, et qu'ils s'intègrent dans un contexte de traitement de masses de données.

Nos travaux de recherche s'inscrivent donc également dans le cadre de l'ACI MADONNE (MAsse de DOnnées appliquées à la Numérisation du patrimoiNE) financée par le Ministère de la Recherche et de l'Enseignement. Ce projet a été mené de fin 2003 à fin 2006 par un consortium de laboratoires français ayant une expérience forte et complémentaire dans le domaine de l'analyse d'images de documents. Ce consortium comprenant le L3i (La Rochelle) qui pilote le projet, le LORIA (Nancy), le LITIS (Rouen), le LI (Tours), le LIRIS (Lyon), le CRIP5 (Paris) et l'IRISA (Rennes). Les problématiques abordées concernent le développement d'outils et de méthodologies facilitant l'indexation de masses de données et la navigation dans des bases d'images documentaires, en aval des projets de numérisation qui fournissent de larges bases d'images faiblement structurées [Ogier 06]. Ces problématiques englobent différents aspects tels que la modélisation des collections, la compression et le traitement des images, l'analyse de documents manuscrits, l'analyse de documents graphiques, l'analyse des structures. Notre contribution à ce projet concerne l'analyse de la structure de documents complexes, notamment les documents manuscrits.

1.4 Conclusion

Pour certains types de documents, comme les documents historiques, patrimoniaux ou culturels, l'image reste la meilleure représentation numérique, car c'est la seule qui permet de décrire fidèlement l'apparence physique du document. Néanmoins se pose le problème de l'indexation et de la gestion de ces documents dans des bibliothèques numériques, car la représentation sous forme d'image est pauvre en terme de structuration de l'information. Les techniques traditionnelles d'indexation reposent sur l'ajout de métadonnées ou sur l'analyse du contenu des documents, notamment le contenu textuel. Dans la mesure où la rétro-conversion de ces documents est difficilement envisageable et l'indexation manuelle est fastidieuse et très coûteuse, la solution pourrait résider dans une indexation automatique ou supervisée, reposant sur l'extraction et la reconnaissance partielle de certaines zones ou entités informatives. Une telle solution nécessite la mise en place d'outils et de méthodes d'analyse d'image intégrés éventuellement dans un environnement d'aide à l'indexation ou à l'annotation d'images de documents. C'est ce type d'approche que nous avons retenue en particulier dans le cadre du Projet Bovary pour la réalisation des transcriptions diplomatiques de manuscrits d'auteurs, mais également d'une manière plus générale pour l'indexation de documents anciens hétérogènes comme c'est le cas dans le contexte du projet MADONNE.

L'objectif est de proposer une méthodologie et des outils suffisamment souples, génériques et simples à utiliser, pour apporter une aide à la résolution de diverses tâches d'indexation de masses de documents. Il s'agit de faciliter le travail d'indexation en localisant de manière automatique ou supervisée les zones informatives , et en extrayant des métadonnées, à partir de l'analyse de la structure des documents. Ainsi par exemple pour des manuscrits d'auteurs il pourra s'agir d'étiqueter les zones telles que les blocs de texte, en marge, en en-tête ou en pied de page, les lignes de texte ou encore les zones de ratures et de surcharge, afin de permettre la production automatique de transcriptions diplomatiques. Pour des ouvrages de la Renaissance il pourra s'agir d'identifier les parties graphiques et certaines parties textuelles telles que les numéros de pages, pour par exemple permettre une indexation sur les illustrations et la construction de tables d'index. Les possibilités d'indexation sont multiples, mais dans tous les cas ceci nécessite des

méthodes d'analyse robustes et adaptables à différentes problématiques de localisation. Des méthodes et des outils d'aide à l'indexation ont certes été proposées ces dernières années comme nous avons pu le voir au cours de ce premier chapitre, mais les solutions mises en oeuvre sont systématiquement dédiées à un type précis de documents et présentent des limitations pour le traitement de documents moins spécifiques, notamment en ce qui concerne l'analyse de la structure des documents. Il est donc nécessaire de proposer des solutions plus génériques. Pour cela on se propose, dans le chapitre suivant, de passer en revue les principales méthodes utilisées en analyse de documents et montrer qu'elles sont souvent dédiées à des tâches et à des documents particuliers.

Chapitre 2

Analyse d'images de documents

2.1 Introduction

L'analyse d'images de documents est le processus informatique qui consiste à fournir une interprétation ou plusieurs interprétations de l'information véhiculée par le document, en s'appuyant éventuellement sur différents niveaux d'abstraction de l'information. Il s'agit donc de localiser, extraire puis reconnaître l'information contenue dans un document scanné afin d'en fournir une interprétation. L'objectif visé par ce processus peut-être soit la rétro-conversion complète du document dans un format structuré en permettant la manipulation sous forme électronique, soit une interprétation partielle de l'information permettant l'extraction de métadonnées nécessaire à l'indexation automatique de l'image du document. L'enjeu est donc la gestion et la manipulation de documents numérisés (et de l'information qu'ils contiennent), au sein d'applications de Gestion Electronique de Documents (GED), et de vastes bibliothèques numériques. L'interprétation d'une image de document nécessite la mise en place d'une chaîne complète de traitement, reposant sur l'utilisation et la coopération de différents niveaux d'analyse, permettant de construire différentes vues, ou niveaux hiérarchiques d'abstraction de l'information. Les résultats d'analyses intermédiaires obtenus sont non seulement nécessaires pour les traitements ultérieurs, mais constituent également des métadonnées intéressantes pour l'indexation automatique.

Dans la littérature, le processus d'analyse d'images de documents se décompose généralement en deux étapes plus ou moins dépendantes en fonction des approches utilisées, et de la nature des documents considérés : une étape de segmentation et une étape de reconnaissance. La première vise à isoler les différentes entités informatives et à déterminer la structure physique de mise en page du document, alors que la seconde vise à déterminer la structuration logique de l'information.
En effet on distingue généralement deux types de structuration de l'information dans un document. La structuration logique et la structuration physique. Ces deux types de structures correspondent à différents niveaux d'abstraction de l'information. Pour certains types de documents un niveau intermédiaire de structuration peut également être pris en compte.

Structure logique d'un document

La structure logique décrit l'organisation du discours de l'auteur, c'est-à-dire la façon dont il a articulé, structuré sa pensée pour communiquer l'information [Heroux 01]. Il s'agit d'une description hiérarchique et logique du contenu d'un document au moyen d'entités logiques telles que les sections, les chapitres, les paragraphes, les titres, ... Les entités logiques sont donc des concepts servant à structurer le message de l'auteur, et en retour elles servent de repères au lecteur pour mieux décrypter l'information transmise par l'auteur du document [Azokly 95]. L'intérêt de la structure logique est qu'il s'agit d'une représentation abstraite du document qui ne tient compte ni de son support ni de sa présentation. La structuration logique est donc une abstraction qui décrit l'organisation du contenu du document, la manière de décrypter l'information.

Structure physique d'un document

La structure physique décrit quant à elle la mise en page du document, c'est à dire qu'elle décrit au moyen d'entités physiques l'organisation visuelle de l'information véhiculée. Les entités physiques sont donc des concepts servant à structurer l'aspect graphique d'un document. Il s'agit des caractères, des mots, des lignes de texte, des blocs, des objets graphiques et typographiques, ... Ces entités décrivent la présentation graphique des entités lo-

giques. Il existe donc une relation entre la structure logique et la structure physique.

La difficulté de l'analyse d'images de documents est en effet liée à la variabilité dans les contenus, à la variabilité dans les structures, et à la grande hétérogénéité dans les types de documents. On peut en effet distinguer les documents selon leur nature (imprimé, manuscrit), selon leur contenu (données textuelles, graphiques ou hétérogènes), selon leur niveau de structuration (documents structurés ou documents non contraints) ou encore suivant leur niveau de dégradation (documents anciens ou bruités, et documents faiblement dégradés). En fonction de ces différents critères les structures physiques et logiques sont plus ou moins identifiables, et plus ou moins dépendantes. Les stratégies d'analyse à mettre en place en dépendent donc généralement de ces facteurs. Ainsi les problèmes rencontrés ne sont par exemple pas les mêmes pour les documents imprimés que pour les documents manuscrits.

2.2 Analyse des documents imprimés

Dans le cas des documents imprimés il existe une relation clairement identifiée entre la structuration logique du document, et sa structuration physique. En effet, la structure physique est issue du processus de formatage ou de mise en forme appliquée à la représentation logique du document lors de sa production. Le formatage du document est obtenu par application de règles de typographie et de mise en page, qui ont pour but de traduire graphiquement la structuration logique du contenu et d'en faciliter la compréhension par le lecteur.
Le processus d'analyse est alors dans ce cas le processus inverse du processus qui a abouti à la production du document. Il s'agit donc de retrouver l'information que l'auteur du document a voulu transmettre, ce qui nécessite de comprendre comment l'information est structurée. Pour cela on cherche généralement à déterminer la structuration physique à partir de l'image du document, puis à déduire de la structure physique les règles typographiques permettant de retrouver la structure logique du document à partir de sa structure physique. On distingue trois types de stratégies d'analyse : les stratégies ascendantes ou "guidées par les données" qui se basent sur l'analyse des données pour remonter jusqu'à la structure logique, les stratégies

descendantes ou "guidées par les modèles" qui exploitent la connaissance a priori de modèles de structuration pour aller rechercher l'information dans l'image du document, et les stratégies mixtes qui reposent à la fois sur l'analyse des données et sur l'exploitation de modèles. En pratique il n'existe généralement pas de méthode d'analyse purement ascendante ou purement descendante, et la plupart des approches proposées sont mixtes. La chaîne d'analyse des documents imprimés est illustrée sur la figure 2.1.

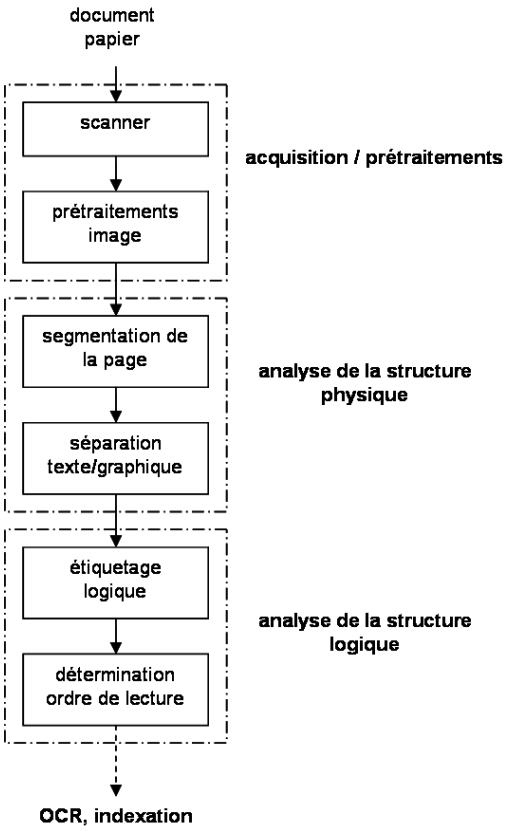


Fig. 2.1. Chaîne générale d'analyse des documents imprimés

2.2.1 Analyse de la structure physique

Dans le cas des documents imprimés structurés, beaucoup de méthodes d'analyse de la structure physique ont été proposées. On trouve principalement deux catégories de méthodes : celles qui se basent sur l'analyse des espaces (qui correspondent généralement à des stratégies

descendantes), et celles qui se basent sur l'analyse des formes de l'image (qui correspondent plutôt à des stratégies ascendantes). Nous présentons ici le principe des quelques méthodes parmi les plus connues et les plus utilisées.

Algorithme RLSA de Wong et al. [Wong 82]

L'algorithme RLSA (Run Length Smoothing Algorithm) [Wong 82], l'un des algorithmes de segmentation physique les plus populaires et les plus anciens, repose sur une stratégie ascendante. Il s'agit d'une méthode itérative basée sur des opérations morphologiques de traitement d'image, qui permet de segmenter des images de documents binaires en blocs, en fusionnant en une seule composante connexe, des composantes de l'image suffisamment proches, selon un seuil fixé. Le principe de cette méthode est de noircir toute séquence de pixels blancs comprise entre deux pixels noirs, de longueur inférieure à un seuil donné. Ceci permet de fusionner les composantes connexes proches. En pratique l'algorithme est appliqué horizontalement et verticalement sur l'image binaire originale avec des seuils éventuellement différents pour l'horizontale et la verticale, puis une opération ET logique est appliquée entre les deux images lissées obtenues. L'extraction des composantes connexes de l'image résultante permet d'obtenir les entités de la structure physique sur un niveau hiérarchique donné. On peut ainsi en répétant la procédure avec des seuils de lissage horizontal et vertical différents, extraire itérativement les blocs de l'image, puis les lignes de texte et les mots. Ces seuils de lissage sont les seuls paramètres de l'algorithme RLSA. Ils contrôlent la manière dont les composantes sont fusionnées sur un niveau de segmentation prédéterminé. Par exemple pour segmenter des lignes de texte horizontales, droites et bien espacées, on pourra utiliser un seuil de lissage vertical nul et un seuil de lissage horizontal suffisamment grand pour combler les espaces inter-lettres et inter-mots. Cet algorithme a l'avantage d'être très simple à mettre en oeuvre puisqu'il repose sur des opérations morphologiques élémentaires. De plus, il ne requiert qu'un nombre limité d'itérations pour atteindre la segmentation complète du document. La principale difficulté dans l'utilisation de cet algorithme est le réglage des seuils de lissage qui est délicat. Il est nécessaire de déterminer les seuils adéquats pour chaque niveau de segmentation. De tels seuils ne peuvent être déterminés qu'empiriquement. De plus pour une

itération donnée ces seuils sont constants, il faut donc que pour le niveau de segmentation considéré, les espaces entre les composantes de l'image à fusionner soient aussi constants.

Méthode Docstrum de O'Gorman [O'Gorman 93]

La méthode Docstrum par O'Gorman [O'Gorman 93] est également une méthode ascendante. Elle procède par regroupement itératif des composantes connexes de l'image, en se basant sur une analyse des distances et des orientations. Le regroupement est effectué par clustering de type "k plus proches voisins". Cette méthode repose donc sur une analyse locale, tout en intégrant une part d'information contextuelle. En effet le regroupement effectué se base sur l'analyse des histogrammes modélisant les distributions des distances et des orientations entre composantes voisines. Ces distributions font apparaître plusieurs modalités correspondant aux distances et aux orientations inter-caractères, inter-mots et inter-lignes. En effet en moyenne dans un document imprimé les espaces inter-caractères sont plus faibles que les espaces inter-mots, eux-mêmes plus faibles que les espaces interlignes. En ce qui concerne les orientations, les composantes appartenant à une même ligne de texte seront orientées selon une même direction, qui est en général de 0° si la page ne présente pas une légère inclinaison. Le principe de la méthode est de faire une analyse globale des distances et des orientations entre des couples de composantes voisines dans l'image, en traçant ce que l'auteur appelle le spectre du document (document spectrum ou docstrum). L'analyse de ce spectre permet d'estimer les distances intra et inter-lignes globales sur tout le document, ce qui permet ensuite de guider efficacement le processus d'analyse locale lors du regroupement. Cette méthode est assez robuste, et efficace, mais elle permet d'extraire la structure physique de documents textuels uniquement. Elle fait l'hypothèse que toutes les parties non-textuelles du document, c'est à dire les zones graphiques et les photographies, ont été filtrées préalablement.

Une approche ascendante assez similaire permettant de segmenter des images de documents binaires en blocs de texte, en lignes ou en mots, par regroupement de composantes connexes est présentée également dans [Kise 98]. Cependant dans cette méthode, le processus de regroupement est

basé sur l'utilisation d'un diagramme de Voronoï. Une méthode similaire est présentée dans [Lu 04] pour la segmentation en mots.

Algorithme X-Y Cut de Nagy et al. [Nagy 84]

Un autre algorithme de segmentation très connu et très utilisé est l'algorithme X-Y Cut initialement présenté par Nagy et al. [Nagy 84], puis amélioré par Ha et al. [Ha 95]. Cet algorithme s'applique également sur des images binaires. Cependant il s'agit cette fois d'un algorithme descendant qui permet de découper l'image récursivement en zones homogènes de plus en plus petites par analyse des profils de projection horizontaux et verticaux des pixels noirs de l'image. On peut ainsi aboutir avec cet algorithme à une segmentation en blocs (zones textuelles et graphiques), puis en lignes et éventuellement en mots pour les zones textuelles. Le principe est de calculer les profils de projection alternativement sur l'horizontale et la verticale, et de déterminer les minima sur l'histogramme des projections. Les minima correspondent en effet aux lignes (ou colonnes) comportant très peu de pixels noirs et donc pratiquement blanches, susceptibles de représenter des espaces entre entités. Si la valeur de ces minima est inférieure à un seuil donné, l'image est découpée en deux en ces points. Cet algorithme est également très simple, mais il est très sensible à l'inclinaison. En effet, vu que les projections se font sur l'horizontale et la verticale, si la page présente une certaine inclinaison, il n'est alors plus possible de déterminer de minima dans les profils de projection. Une alternative possible est alors d'effectuer les projections suivant différentes orientations, l'algorithme devient dans ce cas plus coûteux en temps de calcul. Un inconvénient majeur de cet algorithme est également le fait qu'il nécessite le réglage de plusieurs seuils qui sont toujours difficiles à déterminer et qui posent des problèmes d'adaptation à différents types de documents.

Nous avons présenté là les méthodes les plus connues, mais cette liste est évidemment loin d'être exhaustive, tant les travaux ont été nombreux dans ce domaine depuis plus de 20 ans. On pourra trouver des états de l'art relativement complets sur l'analyse d'images de documents et l'extraction de structures physiques et logiques dans [Haralick 94], [Nagy 00], [Mao 03].

Quelle que soit la méthode utilisée, le résultat de l'analyse de la structure physique est une segmentation en blocs homogènes contenant un certain type d'information (texte, graphique, photographie, ...). Dans le cas de documents hétérogènes, un processus de séparation texte/graphique peut ensuite être appliqué, afin de permettre l'application de traitements spécifiques aux différents types de données. Ainsi par exemple les éléments textuels peuvent être de nouveau segmentés en lignes de texte, en mots et éventuellement en caractères pour permettre l'application d'un module de lecture optique de caractères (OCR). Le problème est ensuite de déterminer la structuration de l'information véhiculée par le document pour en fournir une interprétation. C'est le problème de l'analyse de la structure logique.

2.2.2 Analyse de la structure logique

Alors que la phase d'analyse de la structure physique vise à segmenter (ou décomposer) en entités telles que des blocs, l'analyse de la structure logique vise à classifier ces blocs selon leur fonction logique dans le document, et à en déterminer l'ordre de lecture. Il s'agit donc d'attribuer des étiquettes logiques aux régions homogènes extraites, et de déterminer les relations logiques existantes entre ces entités. Cela s'effectue généralement en étant guidé par une certaine description a priori ou modèle du document [Cattoni 98].
Dans la mesure où il existe des relations fortes entre structure physique et logique, la structure logique d'un document est souvent déterminée à partir de sa structure physique extraite lors de la phase de segmentation préalable. En effet la structuration logique d'un document se traduit généralement par des règles de mise en page et des conventions typographiques bien identifiées et relativement stables. Ceci est d'autant plus vrai pour les documents imprimés et dactylographiés. Pour ces types de documents il est possible de déterminer des relations uniques entre les éléments de la structure physique et de la structure logique. Ainsi par exemple un titre de chapitre pourra apparaître en gras avec une certaine taille de police, alors qu'un titre de section utilisera une taille de police plus petite. Ces règles sont celles définies par la feuille de style utilisée lors de la production du document. Ces règles de mise en forme définies par la feuille de style étant généralement univoques, la détermination de la structure logique à partir de la structure physique ne pose pas trop de difficultés, à condition évidemment que le ré-

sultat de la segmentation soit correct, ce qui n'est pas toujours garanti. Dans ce cas là, une erreur de segmentation entraîne irrémédiablement une erreur de reconnaissance de la structure logique.

De nombreuses approches ont été proposées dans la littérature pour l'extraction de la structure logique. On distingue notamment les approches à base de transformation d'arbre [Tsujimoto 92][Dengel 92], les approches à base de langage de description [Schurmann 92][Higashino 86], les approches à base de tableau noir (blackboard) [Srihari 87][Wang 88], les approches syntaxiques à base de grammaires formelles [Nagy 92][Krishnamoorthy 93][Conway 93] ou encore les approches probabilistes à base de modèles de Markov cachés [Kopec 94][Kam 96].
On trouvera des états de l'art complets sur l'analyse de la structure logique dans [Haralick 94][Cattoni 98] [Nagy 00][Mao 03].

Pour les documents imprimés structurés, il est donc possible de déterminer la structure logique à partir de la structure physique. Cependant cela suppose d'être capable de segmenter correctement le document, ce qui ne pose généralement pas trop de difficultés dans le cas des documents imprimés non dégradés car les entités sont généralement naturellement bien espacées. Lorsque les documents sont bruités ou dégradés, la segmentation devient plus délicate, et il est donc souvent nécessaire de mettre en place des stratégies faisant coopérer segmentation et reconnaissance. C'est le cas notamment pour les documents manuscrits.

2.3 Analyse des documents manuscrits

Dans le cas des documents manuscrits, l'objectif visé par l'interprétation du document est généralement la reconnaissance des mots ou des lignes de texte. Les documents manuscrits sont en effet des documents qui sont majoritairement textuels, et dont la structure est en général relativement simple contrairement aux documents imprimés. Cependant cette structure est caractérisée par une très forte variabilité spatiale qui se traduit par exemple par des lignes de texte inclinées et fluctuantes, des chevauchement entre les lignes, des espaces irréguliers entre les mots, ...
Contrairement à ce qui est fait en analyse des documents imprimés, on ne cherche donc pas explicitement à déterminer la structure physique et

la structure logique du document, mais on considère plutôt une phase de segmentation et une phase de reconnaissance de l'écriture. Les notions de structure physique et de structure logique se rapportent plus en effet aux documents structurés et imprimés, qu'aux documents manuscrits. Les documents manuscrits sont des documents non contraints qui ne suivent pas nécessairement des règles de structuration explicitement définies.
La phase de segmentation, qui peut être assimilée à la phase d'analyse de la structure physique pour les documents imprimés, consiste à découper la page en entités textuelles élémentaires, telles que les lignes de texte et les mots, et la phase de reconnaissance a pour but la reconnaissance des entités textuelles segmentées.
Dans le domaine du traitement des documents manuscrits la segmentation est souvent considérée comme une étape de prétraitement visant à isoler les mots préalablement à une reconnaissance par un module de reconnaissance de mots isolés. Il s'agit rarement de déterminer la structure physique complète du document et de construire l'arborescence correspondante, mais plutôt d'isoler les entités textuelles afin de faciliter leur reconnaissance. Du fait de la variabilité de l'écriture manuscrite, des méthodes différentes sont souvent utilisées pour extraire les lignes et les mots. Les méthodes proposées sont donc toujours dédiées à une tâche de segmentation identifiée. Il ne s'agit pas d'appliquer une seule méthode de décomposition de la page permettant d'obtenir la hiérarchie des blocs, des lignes et des mots, mais plutôt d'appliquer successivement et séquentiellement une suite de traitements permettant d'isoler les mots afin de les reconnaître. En effet on est capable aujourd'hui de reconnaître des mots manuscrits isolés avec des taux de reconnaissance corrects, surtout avec un lexique réduit. L'étape de segmentation s'intègre donc plutôt dans une chaîne de traitements telle qu'illustrée sur la figure 2.2.

2.3.1 Processus de segmentation

Nous présentons dans cette section les méthodes permettant de segmenter un document manuscrit en lignes de texte puis les méthodes permettant de segmenter les lignes de texte en mots, en essayant d'être le plus exhaustif possible dans la mesure où il existe finalement peu de méthodes.

Segmentation en lignes de texte

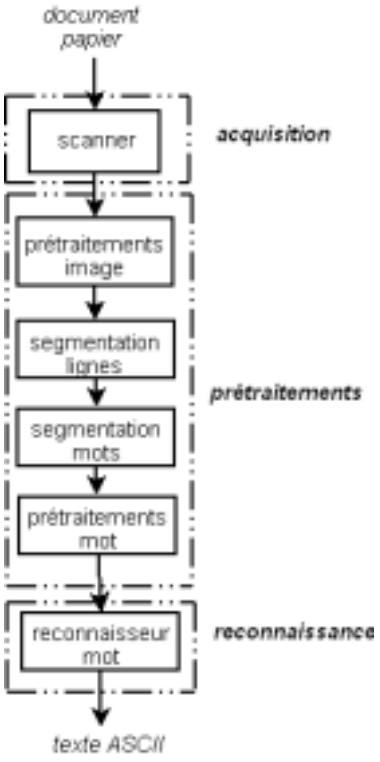


Fig. 2.2. chaine d'analyse de documents manuscrits

L'extraction de lignes de texte représente une étape dans le processus d'extraction de la structure physique. Dans le cas des documents manuscrits il s'agit souvent de la première étape, préalablement à l'extraction et à la reconnaissance des mots des lignes de texte. C'est une étape importante, car toute erreur produite à ce niveau a des répercussions importantes sur les traitements ultérieurs. Les trois types de stratégie vues précédemment, à savoir la stratégie ascendante, descendante et mixte peuvent être utilisées pour segmenter un document manuscrit en lignes de texte. Cependant en pratique la stratégie ascendante est la plus souvent utilisée car elle s'affranchit mieux de la variabilité liée à l'écriture manuscrite. Les méthodes proposées sont donc généralement basées sur une analyse locale de l'image du document.

Une méthode ascendante basée sur un regroupement itératif de composantes connexes est proposée dans [Likforman-Sulem 93] [Likforman-Sulem 94a] [Likforman-Sulem 94b] [Likforman-Sulem 95b] pour segmenter en lignes de texte des images binaires de textes manuscrits. Les primitives manipulées par la méthode sont les composantes connexes de

pixels noirs. Ces composantes sont tout d'abord extraites de l'image. Les composantes relativement allongées présentant une orientation privilégiée sont ensuite déterminées. Ces composantes appelées point d'ancrage par les auteurs permettent d'initialiser le processus de regroupement des composantes. C'est en effet à partir de ces points d'ancrage que les lignes sont formées, par regroupement des composantes voisines selon différents critères perceptifs tels que la proximité, la ressemblance, ou encore la continuité de la ligne. Dans la mesure où des conflits de regroupement peuvent apparaître si une même composante est assignée à des alignements différents, ou encore que des lignes puissent être fragmentées ou au contraire fusionnées par erreur, les auteurs proposent un certain nombre de règles de correction, appliquées en post-traitement pour affiner le résultat obtenu. Alors que le processus de regroupement est basé sur une analyse relativement locale, cette étape de résolution des conflits est plutôt globale. Il s'agit en effet d'évaluer la qualité des alignements de manière globale. Cette méthode repose donc sur une stratégie mixte, mais à tendance plutôt locale tout de même. Les auteurs l'ont testé sur quelques manuscrits de l'écrivain Gustave Flaubert et pour certains manuscrits pas trop bruités ils ont obtenus de bons résultats. Un exemple de résultat de segmentation obtenu avec cette méthode est présenté sur la figure 2.3.

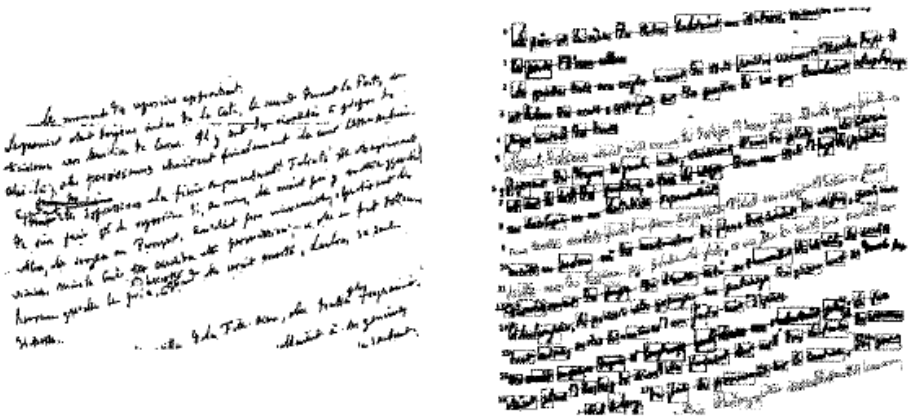


Fig. 2.3. Exemple de résultat de segmentation en lignes obtenu avec la méthode par regroupement de Likforman [Likforman-Sulem 93]

L'approche descendante basée sur l'analyse des profils de projection des pixels noirs de l'image, classiquement utilisée pour segmenter les documents

imprimés, peut également être utilisée pour segmenter les documents manuscrits en lignes de texte. Cette approche s'applique sur des images binaires. Comme nous l'avons déjà vu précédemment, l'idée est de considérer que les espaces interlignes permettent de séparer les lignes de texte. Ces espaces interlignes peuvent être repérés sur l'histogramme des projections des pixels noirs de l'image, par les vallées du profil. Il s'agit donc de réaliser une projection horizontale de l'ensemble des pixels noirs de l'image et de déterminer sur l'histogramme ainsi obtenu les minima locaux du profil correspondant aux espaces interlignes. La position de ces minima détermine les points de coupure pour segmenter l'image en lignes de texte. Une autre alternative possible est de détecter sur l'histogramme les maxima locaux plutôt que les minima. Ces extrema correspondent dans ce cas directement aux lignes de texte. Il est également possible d'utiliser l'histogramme des transitions horizontales noir/blanc de pixels pour déterminer les points de segmentation [Marti 99][Nosary 02][Kalcheva 03] . Les lignes ne présentant aucune ou peu de transitions noir/blanc sont en effet plus susceptibles de correspondre à des interlignes. Les méthodes à histogramme ont l'avantage d'être relativement peu coûteuse en temps de calcul, et d'être simples à mettre en oeuvre. Cependant elles supposent que les lignes sont droites, horizontales ou très peu inclinées et relativement espacées de manière à ce qu'il y ait le moins de recouvrements possibles entre les lignes du fait de chevauchement entre les jambages d'une ligne et les hampes de la ligne suivante. Ceci est rarement le cas dans les documents manuscrits non contraints dont la structure est complètement libre. En effet dans les documents manuscrits les lignes de texte présentent souvent une certaine inclinaison, sont souvent fluctuantes et ont généralement tendance à se chevaucher. En conséquence de quoi l'application d'une telle méthode qui calcule un histogramme global des projections sur toute la page du document est difficilement réalisable. Il est en effet très difficile voire impossible de détecter les maxima locaux sur l'histogramme des projections, car l'analyse est trop globale. C'est pourquoi certaines méthodes recherchant un compromis entre analyse locale et analyse globale ont été proposées [Bruzzone 99] Il s'agit de calculer l'histogramme des projections des pixels noirs sur des fragments de l'image plutôt que sur l'image compléte. Ce principe est également mis en oeuvre dans la méthode à ombrage (shading) proposée dans [Cohen 91]. L'image binaire du document est tout d'abord découpée en bandes verticales de largeur fixe. Dans chacune

des bandes ainsi obtenues, on détermine sur l'histogramme des projections horizontales la position des lignes de texte potentielles. Ces hypothèses sont ensuite éventuellement validées en vérifiant la continuité des lignes formées sur les bandes verticales consécutives. Ce type d'approche permet de mieux s'affranchir des fluctuations locales des lignes, mais la largeur des bandes verticales représente un paramètre critique.

Une méthode basée sur l'utilisation de la transformée de Hough est proposée dans [Likforman-Sulem 95a]. La transformée de Hough est une technique de reconnaissance de formes développée en 1962 par Paul Hough, qui permet de détecter des lignes ou des alignements. Le principe de cette technique est de faire correspondre à une droite dans l'espace de départ (l'image), un unique point dans l'espace d'arrivée (espace de Hough). Il s'agit en fait de changer d'espace de représentation, en passant d'un espace cartésien à un espace paramétrique en coordonnées polaires. Dans cet espace chaque ligne est alors représentée par un point. La méthode proposée dans [Likforman-Sulem 95a] utilise une stratégie de génération/validation d'hypothèses de détection de lignes de texte, basée sur cette transformée. Pour cela les composantes connexes de l'image sont d'abord extraites puis la transformée de Hough est appliquée. Pour chaque droite possible, le nombre d'intersections avec des composantes connexes de l'image est mémorisé dans l'espace de Hough. Seules les droites comptabilisant un maximum d'intersection sont retenues comme lignes potentielles (phase de génération d'hypothèses). Une validation est ensuite effectuée dans l'image en se basant sur des critères perceptifs, tels que la proximité, afin de rejeter certaines hypothèses qui ne correspondent pas véritablement à des lignes. Un exemple de résultat obtenu avec cette méthode est illustré sur la figure 2.4.

Une méthode similaire est utilisée dans [Shapiro 93]. La transformée de Hough est d'abord appliquée sur l'image originale binaire puis l'orientation des lignes de textes est estimée en analysant les paramètres dans le plan de Hough. La transformée de Hough inverse est ensuite utilisée pour déterminer comment découper l'image en bandes horizontales correspondant aux lignes détectées dans le plan de Hough. Certaines méthodes appliquent également la transformée de Hough sur les centres de gravité des composantes connexes. Les méthodes basées sur l'utilisation de la transformée de Hough ont l'avantage de permettre de détecter des lignes quelle que soit leur orientation. Cependant, bien qu'elles soient plus ou moins tolérantes à une

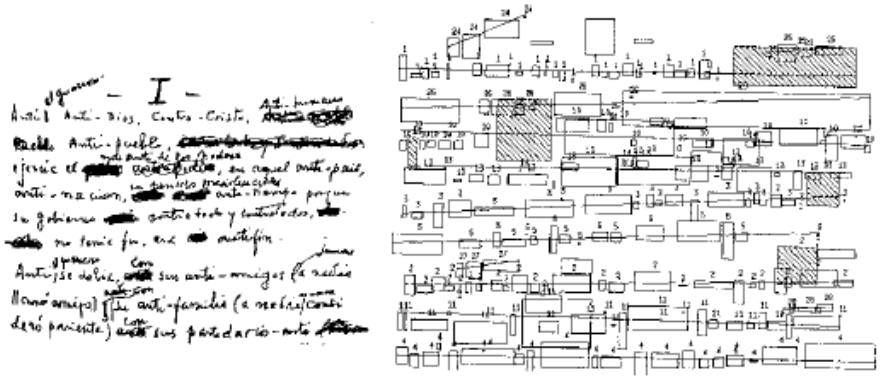


Fig. 2.4. Exemple de segmentation en lignes réalisée par la méthode basée sur la transformée de Hough présentée dans [Likforman-Sulem 95a]

certaine fluctuation, les résultats fournis par ces méthodes ne sont corrects que si les lignes sont globalement droites. De plus, la transformée de Hough est très coûteuse en temps de calcul.

A l'instar des travaux présentés dans [Likforman-Sulem 94a], Lemaître et al. [Lemaitre 06] proposent une approche d'extraction de lignes de texte qui se base sur des mécanismes de vision perceptive, en l'occurence sur l'observation qu'à une certaine distance, ou de manière équivalente à une certaine résolution, les lignes de texte d'un document peuvent être perçues comme de simples lignes droites. Le problème de l'extraction des lignes dans un document textuel, devient alors un problème de détection et d'extraction de segments de droite dans l'image du document. Les auteurs proposent donc d'appliquer un filtrage de Kalman à basse résolution sur l'image du document pour extraire les lignes de texte. Cela permet de réduire les détails lors de l'analyse et de pouvoir ainsi utiliser un extracteur de segments de droites basé sur un filtrage de Kalman. Les auteurs ont appliqué leur méthode sur différents types de documents d'archive : registres de naissance, mariages, décès, registres de décrets de naturalisation. Des résultats qualitatifs sont présentés dans [Lemaitre 06]. Une évaluation statistique a également été réalisée sur un corpus de 1124 pages de décrets de naturalisation. Malheureusement cette évaluation ne porte pas directement sur l'extraction des lignes de texte, mais plutôt sur la détection de certains types d'information dans les lignes de texte, en l'occurence ici les noms

des personnes naturalisées. Les avantages de cette méthode sont qu'elle permet d'extraire des lignes de texte inclinées, verticales ou présentant une certaine courbure, qu'elle peut s'appliquer aussi bien aux documents imprimés qu'aux documents manuscrits.

Dans [Feldbach 01b] et [Feldbach 01a] une méthode basée sur la recherche des minima locaux du contour externe des composantes connexes est proposée. Cette méthode consiste à relier itérativement les points potentiels extraits des contours afin de construire les lignes de base. Elle a été appliquée à la segmentation de registres paroissiaux datant du 17ème au 19ème, et a montré sa capacité à segmenter en lignes de texte des documents présentant des chevauchements importants de l'écriture. Une méthode basée sur la recherche d'arbre recouvrant minimal est proposée dans [Abuhaiba 96]. Cette méthode permet l'extraction de lignes de texte présentant une certaine inclinaison. Cependant elle repose sur l'hypothèse que les espaces entre les lettres et entre les mots, sont plus faibles que les espaces entre les lignes, ce qui en pratique n'est pas toujours vérifié. Plus récemment des méthodes basées sur l'utilisation de longueurs de plages floues [Shi 04] ou sur l'utilisation d'une carte de connectivité locale adaptative [Shi 05]) ont été proposées pour la segmentation en lignes de texte dans des documents anciens.

La détection et l'extraction de lignes de texte dans des documents manuscrits non contraints ou dans des documents historiques ou anciens est un problème difficile qui reste encore non résolu de manière satisfaisante, du fait de la très forte variablité de l'écriture manuscrite, de la grande complexité et de l'état de conservation de certains types de documents. Avec l'intérêt porté récemment à la problématique de la valorisation du patrimoine, de nouvelles méthodes sont proposées et d'autres seront assurément encore proposées dans les mois et années à venir [Li 06b] [Li 06a].

Segmentation des lignes de texte en mots

Une fois que le texte a été découpé en lignes, il faut segmenter chaque ligne en mots, afin de pouvoir appliquer un reconnaisseur mot pour remplacer la représentation des mots sous forme d'une image numérique par une représentation textuelle sous forme de caractères codés numériquement selon une norme de codage de texte telle que la norme ASCII ou Unicode par

exemple. La plupart des approches de segmentation des lignes de texte en mots se base sur l'analyse des espaces intra et inter-mots. L'idée est de considérer que les espaces entre les mots sont généralement plus importants que les espaces entre les lettres dans un mot. Les espaces inter-mots représentent donc les points de coupure recherchés pour segmenter une ligne de texte manuscrit en mots. Ces espaces sont mesurés par une distance entre composantes connexes dans une ligne. Différentes métriques peuvent être définies pour exprimer la distance entre les composantes [Seni 94], parmi lesquelles la distance horizontale minimale entre boîtes englobantes des composantes, la distance euclidienne minimale entre les composantes, la longueur moyenne des plages de pixels blancs entre composantes présentant un recouvrement vertical (figure 2.5), ou encore diverses mesures de distances qui combinent les méthodes précédentes avec l'utilisation de différentes heuristiques spécifiques (recouvrements horizontal et vertical supérieur à un certain seuil). Il est également possible de mesurer la distance entre les enveloppes convexes des composantes connexes. Il s'agit ensuite de modéliser la distribution des espaces intra et inter-mots afin de déterminer des seuils de segmentation. A partir de ces seuils les espaces entre composantes sont classés soit comme des espaces inter-mots, soit comme des espaces intra-mots. Cela permet ensuite de segmenter la ligne de texte en mots.

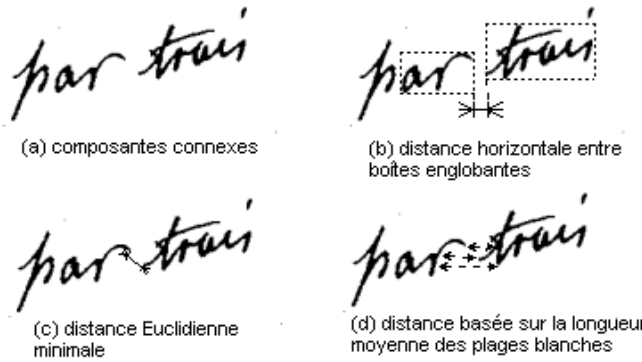


Fig. 2.5. exemples de métriques mesurant les espaces entre composantes connexes

Dans [Varga 05], plutôt que d'utiliser un seuil global pour séparer les espaces inter-mots des espaces intra-mots, les auteurs utilisent une structure d'arbre pour modéliser la ligne de texte et introduire du contexte dans la prise de décision. Le contexte est introduit en prenant également en consi-

dération la taille relative des espaces environnant dans la prise de décision. Cela permet en fait d'adapter la valeur du seuil en fonction du contexte. Il ne s'agit donc plus ici d'un seuil global, mais d'un seuil adaptatif, qui s'affranchit mieux de la variabilité inhérente à l'écriture manuscrite. La structure d'arbre utilisée permet de modéliser la ligne à segmenter. Les noeuds de cet arbre représentent les hypothèses de segmentation, c'est à dire que chaque noeud correspond à un mot candidat possible. L'arbre est parcouru de manière descendante afin de déterminer les noeuds qui correspondent aux mots de la ligne de texte.

Dans [Marti 01] les composantes connexes de la ligne de texte sont extraites, puis les distances sont mesurées entre les enveloppes convexes des composantes. Pour cela les centres de gravité des composantes sont calculés, puis les composantes connexes sont reliées deux à deux par leurs centres de gravité respectifs à l'aide d'un segment s. L'espace entre les composantes est mesuré par la distance d entre les deux points où le segment de droite s intersecte les enveloppes convexes des composantes. Cette distance d est affectée au segment s reliant les deux composantes c_1 et c_2. Par cette procédure on obtient un graphe G totalement interconnecté et pondéré, dont les noeuds sont les composantes connexes c de la ligne de texte, dont les arcs sont les segments de droite s reliant deux à deux les composantes connexes par leurs centres de gravité respectifs, et dont les pondérations associées aux arcs sont les distances entre enveloppes convexes des composantes. A partir de ce graphe, l'ordre des composantes dans la ligne est obtenu en calculant l'arbre recouvrant minimal du graphe G, c'est à dire en déterminant le chemin le plus court passant exactement une fois par toutes les composantes de la ligne de texte. La distance entre deux composantes successives est ainsi garantie d'être la plus petite possible. Ensuite comme dans la plupart des approches de segmentation en mots, il s'agit de distinguer dans cette suite ordonnée de composantes connexes, les paires de composantes qui appartiennent à un même mot et les paires appartenant à des mots différents, à l'aide d'un seuil. Ici le seuil est calculé à l'aide de trois mesures effectuées dans l'image de la ligne de texte, à savoir, la largeur de la ligne l_l, la largeur médiane des tracés d'écriture l_t, et la distance moyenne entre deux tracés verticaux d_t. Ce seuil de segmentation s_{seg} est calculé de la manière suivante : $s_{seg} = \alpha \frac{l_l - l_t}{d_t}$, où α est un facteur d'échelle qui doit être déterminé empiriquement. Toutefois les auteurs ne

précisent pas comment déterminer ce paramètre. Ce seuil de segmentation s_{seg} est recalculé systématiquement pour chaque ligne à segmenter. Une fois ce seuil déterminé, la segmentation de la ligne est obtenue en supprimant sur l'arbre recouvrant de la ligne, les arcs dont la pondération est supérieure au seuil s_{seg} calculé. Les composantes restant connectées correspondent donc aux mots de la ligne. La figure 2.6 illustre l'arbre recouvrant minimal (b) et le résultat de segmentation (c) obtenus pour une ligne de texte (a).

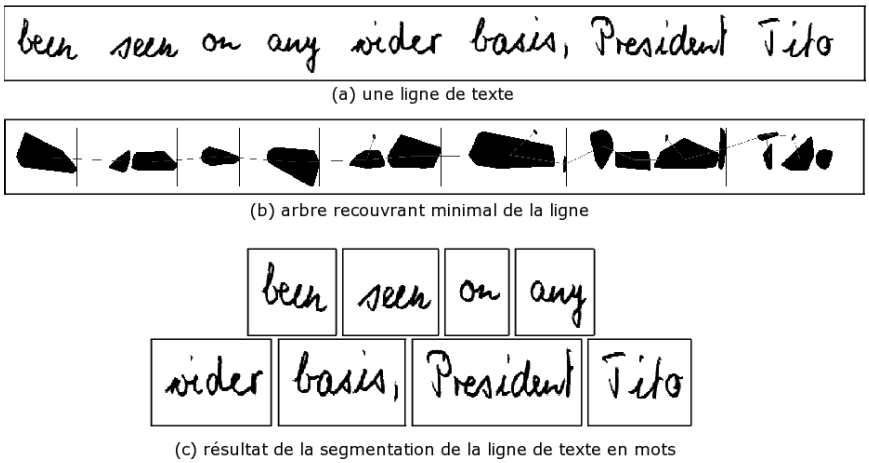


Fig. 2.6. Méthode de segmentation en mots de Marti et al.

Dans [Mahadevan 95] la distance entre entités connexes est également estimée en se basant sur l'espace entre les enveloppes convexes des entités. Les auteurs montrent que cette approche donne de meilleurs résultats que les approches traditionnellement utilisées pour estimer la distance entre composantes.

Feldbach et Tönnies [Feldbach 03] proposent une méthode basée sur l'utilisation combinée de connaissances a priori et de caractéristiques locales, pour segmenter en mots des dates manuscrites dans des documents historiques de type registres d'état civil. La segmentation de ce type de documents est difficile car non seulement les espaces entre les mots sont très faibles et non constants, mais il arrive également très souvent que plusieurs mots soient connectés. La connaissance a priori qu'ils utilisent est la connaissance de la syntaxe des champs manuscrits qu'ils cherchent à segmenter, en l'occurrence ici des champs de type dates. Cette connaissance

permet de générer des hypothèses sur la frontière des mots qui sont ensuite combinées à des mesures locales pour classifier les espaces entre les entités connexes. Elle est modélisée par les distributions de probabilité des combinaisons possibles entre les différentes parties d'un champ date.

Tout comme la segmentation en lignes, la segmentation en mots est difficile, du fait de la variabilité de l'écriture manuscrite. Il est en effet parfois ardu de différencier les espaces inter-mots des espaces intra-mots, car ces espaces ne sont pas uniformes. On est en fait confronté au fameux paradoxe énoncé par Sayre [Sayre 73] : "pour reconnaître il faut segmenter, mais il est difficile de segmenter sans reconnaître". C'est pour cette raison qu'ont également été proposées des approches "segmentation-reconnaissance", faisant coopérer les étapes de segmentation et de reconnaissance afin de corriger les éventuelles erreurs de segmentation, ainsi que des approches "sans segmentation" qui ne cherche pas à segmenter explicitement la ligne de texte en mots pour la reconnaître [Koch 06].

2.3.2 Processus de reconnaissance

Une fois l'information localisée et isolée par la phase de segmentation, il s'agit d'en fournir une interprétation. C'est le but de la phase de reconnaissance. Les premiers travaux réalisés dans le domaine de la reconnaissance de l'écriture manuscrite, se sont portés sur la reconnaissance de mots isolés.

Reconnaissance de mots isolés

La reconnaissance de mots isolés, est le problème qui consiste à convertir l'imagette d'un mot en une suite de caractères électroniques (selon une certaine norme de codage des caractères, telle que la norme ASCII). Il s'agit de formuler des hypothèses sur la nature du mot, à partir de sa représentation graphique (image). On suppose pour cela que le mot a préalablement été correctement isolé et extrait, lors de la phase de segmentation.
On distingue deux grandes familles d'approches pour la reconnaissance de mots manuscrits isolés : les approches globales (ou holistiques) et les approches analytiques.
Les premières considèrent le mot comme une entité indivisible et cherchent à reconnaître directement le mot dans sa globalité. Ces approches se

basent sur l'utilisation d'un lexique de mots à reconnaître. Elles mettent en concurrence les mots du lexique, et utilisent des caractéristiques globales sur la forme du mot à reconnaître, pour sélectionner la meilleure hypothèse. Il s'agit donc d'approches dirigées par un modèle (lexique). Ces approches ne sont toutefois efficaces que dans le cadre de problèmes de reconnaissance à lexique réduit, tels que la reconnaissance des montants littéraux dans les applications de traitement des chèques par exemple[Knerr 97][Guillevic 98].
La seconde famille regroupe les approches analytiques qui cherchent à reconnaître les mots en s'appuyant sur la reconnaissance des lettres les constituant. Cela suppose de pouvoir segmenter correctement le mot en lettres pour pouvoir les reconnaître, ce qui n'est pas évident, puisque l'on est une nouvelle fois confronté au paradoxe évoqué par Sayre. Il est donc souvent nécessaire de mettre en place des stratégies faisant coopérer, et alternant, le processus de segmentation et le processus de reconnaissance. Les approches utilisant une telle stratégie mixte "segmentation/reconnaissance", consistent à sur-segmenter l'image du mot en entités élémentaires appelées "graphèmes", puis à regrouper ces entités selon différentes hypothèses pour reconnaître les lettres et ensuite le mot. La reconnaissance s'effectue généralement à l'aide de modèles de lettres et de mots. Des modèles de types Modèles de Markov Cachés [El-Yacoubi 99] et modèles mixtes Neuro-Markoviens [Augustin 00][Tay 01] sont utilisés pour cela. Parmi les approches analytiques on distingue encore les approches basées sur une segmentation explicite [Kimura 94][El-Yacoubi 99] et les méthodes basées sur une segmentation implicite [Cho 95][Vinciarelli 00][Tay 01].
Les premières sur-segmentent de manière régulière l'image du mot en utilisant un pas fixé, ce qui permet de limiter les cas de sous-segmentation (et donc les erreurs de reconnaissance) mais démultiplie également les possibilités de combinaison. Les approches à segmentation explicite cherchent au contraire à réduire la sur-segmentation, en cherchant des points de segmentation potentiels qui sont généralement des points particuliers répondant à certains critères, tels que les extrema locaux du contour du mot. Les approches basées sur une segmentation explicite présentent cependant le risque de rater certains points de segmentation importants.
Les approches analytiques sont principalement utilisées dans les applications de reconnaissance à large lexique, à lexique dynamique, ou lorsque le lexique est inconnu. Basant leurs décisions sur la reconnaissance des lettres ces

approches sont en effet par définition plus souples vis-à-vis du lexique.

Notons enfin qu'outre les méthodes holistiques et analytiques, il existe une dernière famille d'approches qui se basent sur des modèles cognitifs inspirés du modèle cognitif humain [Côté 98][Pasquer 03]. Ces méthodes ont toutefois été très peu développées.

Quelle que soit l'approche utilisée, la reconnaissance de mots suppose évidemment que les mots soient au préalable correctement segmentés. La segmentation en mots étant un problème complexe, des approches visant à modéliser et à reconnaître directement les lignes de texte sans segmentation en mots ont donc également été proposées.

Reconnaissance de lignes texte

Il s'agit dans ce cas de reconnaître des lignes complètes d'écriture. Certains systèmes de lecture séparent la segmentation et la reconnaissance [Nosary 02], à la manière de ce qui est fait en analyse des documents imprimés, alors que d'autres cherchent au contraire à reconnaître les mots directement dans la ligne de texte, sans tenter de les isoler, de manière à pouvoir profiter du contexte pour lever des ambiguïtés [Marti 01][Vinciarelli 04]. Il s'agit donc dans ce cas de faire coopérer les processus de segmentation et de reconnaissance des mots. Des Modèles de Markov Cachés sont souvent utilisés pour modéliser les lettres, et ces modèles de lettres sont concaténés en modèles de mots, eux-mêmes concaténés pour former des modèles de ligne. La segmentation et la reconnaissance sont alors réalisés simultanément par un algorithme de décodage tel que l'algorithme de Viterbi [Rabiner 89].
En fonction de la nature des documents considérés les approches peuvent toutefois être différentes. Dans le cadre d'applications spécifiques différentes approches de reconnaissance des mots dans les lignes ont donc été proposées, notamment pour la lecture de montants littéraux dans des chèques bancaires [Knerr 97], la lecture d'addresses postales [El-Yacoubi 99], ou la lecture de documents non contraints [Marti 01][Vinciarelli 04], tels par exemple que des courriers entrants [Koch 06]

Comme pour la reconnaissance de mots isolés, on peut donc considérer qu'il existe en reconnaissance de lignes de texte deux stratégies d'analyse : celle qui consiste à séparer et à enchaîner séquentiellement les traitements, et

celle qui consiste à faire une analyse globale de la ligne en réalisant simultanément segmentation, reconnaissance et analyse syntaxique. Ainsi des cas de segmentation ambiguë peuvent être résolus par la reconnaissance ou l'analyse syntaxique, alors qu'avec la première stratégie les erreurs ont tendance à être propagées sans pouvoir être corrigées. Du fait de la forte variabilité de l'écriture manuscrite et de la difficulté de séparer les traitements, de telles approches globales sont donc de plus en plus privilégiées.

2.4 Evaluation des performances

Nous avons donc vu précédemment qu'il existe de nombreuses méthodes d'analyse d'images de documents, que ce soit pour les documents imprimés ou les documents manuscrits, et qu'en fonction de la nature des documents (imprimés ou manuscrits, structurés ou non-contraints, dégradés ou non, ...) les problèmes considérés et les solutions proposées peuvent être différents. Il n'existe certainement pas de méthode d'analyse d'images de documents universelle permettant d'obtenir des résultats satisfaisants sur n'importe quel type de documents. Aucune des méthodes présentées dans la littérature n'est optimale. Certaines méthodes sont cependant plus adaptées et plus robustes pour certains types de documents, ou dans certaines conditions, que d'autres, il faut donc être capable de choisir une méthode adaptée à ses besoins. C'est pourquoi il est très important de pouvoir évaluer le comportement des méthodes d'analyse sur différents types de documents. C'est le problème de l'évaluation des performances.

La manière la plus immédiate pour évaluer les résultats d'un module d'analyse d'images de documents, est de vérifier visuellement le résultat produit. Par exemple, si l'on souhaite évaluer les performances d'un module de segmentation en lignes de texte, on pourra visuellement vérifier le nombre de lignes correctement segmentées par le module. Ceci est possible si on évalue la méthode sur une base d'images de documents, d'une taille raisonnable (quelques dizaines de pages tout au plus), et si on s'intéresse à des tâches de segmentation relativement simples ou à une analyse à un niveau relativement grossier (segmentation en blocs). Cela peut cependant devenir très vite fastidieux. De plus, une telle méthode d'évaluation est complétement subjective. En effet il n'existe généralement pas une segmen-

tation unique du document, mais souvent plusieurs découpages possibles, et il peut parfois être difficile de délimiter avec exactitude certaines zones du document. De plus l'évaluation d'une méthode n'est pertinente que si elle a été effectuée sur un grand nombre de documents de différents types. Or bien souvent les auteurs de nouvelles méthodes ne proposent qu'une évaluation assez limitée sur des corpus de tailles réduites et sur certains types de documents uniquement, en définissant leurs propres critères ou en faisant une vérification visuelle quantitative ou qualitative. Dans ces conditions, l'évaluation et la comparaison de méthodes deviennent des tâches très difficiles. C'est pourquoi il est très important de pouvoir disposer d'un ensemble de critères pertinents et de procédures d'évaluation automatiques dont on est sûr qu'elles se comporteront toujours de la même manière sur des cas similaires. La problématique de l'évaluation des performances des modules d'analyse d'images de documents a donc connu un regain d'intérêt cette dernière décennie, et un certain nombre de systèmes d'évaluation ont été proposés, cependant il s'agit encore d'un domaine de recherche ouvert. Ainsi par exemple dans le cadre du programme Technovision[1] initié par le Ministère de la Recherche et le Ministère de la Défense, des projets centrés sur ces thématiques d'évaluation, tels que le projet RiMES[2] (pour des données manuscrites) ou le projet EPEIRES[3] (pour la reconnaissance de symboles), ont vu le jour récemment.

En ce qui concerne la segmentation, il existe principalement deux manières d'évaluer les performances. La première consiste à analyser le résultat obtenu en sortie de la chaîne d'analyse complète, c'est à dire après la reconnaissance, et sachant les performances du module de reconnaissance d'en déduire les performances du module de segmentation [Kanai 95]. Cette technique est bien adaptée aux documents textuels pour lesquels on peut disposer de la version ASCII correcte, cependant même si le comportement du module de reconnaissance est connu, il est très difficile de déterminer à quel module sont dues les erreurs produites en sortie de la chaîne d'analyse et on ne peut pas savoir non plus où se produisent les erreurs de segmentation. Cette approche d'évaluation des performances est

[1] http ://www.recherche.gouv.fr/technologie/infotel/technovision.htm
[2] http ://www.int-evry.fr/rimes/
[3] www.epeires.org

celle qui historiquement a été utilisée en premier, car on sait depuis de nombreuses années évaluer de manière relativement fiable les performances des modules de reconnaissance de texte. La deuxième approche consiste à évaluer directement et de manière isolée le module de segmentation en comparant les structures physiques et/ou logiques obtenues en sortie, avec une vérité-terrain représentée dans un formalisme identique, et à calculer des critères de performance. Ce type d'approche permet de déterminer exactement le type d'erreurs de segmentation. Ceci est très important car toutes les erreurs de segmentation n'ont pas les mêmes conséquences sur les traitements de reconnaissance ultérieurs. Les différents types d'erreurs de segmentation sont donc pondérés et ces pondérations sont prises en compte dans la mesure globale d'évaluation de performance. En effet, on pourra dans certains cas privilégier un module de segmentation qui effectue beaucoup d'erreurs mais sans grandes incidences sur les traitements ultérieurs, plutôt qu'un module faisant moins d'erreurs mais qui sont plus pénalisantes.

Approches basées sur l'analyse des résultats de reconnaissance

Les premiers travaux sur l'évaluation des performances des modules de segmentation d'images de documents ont consisté à comparer le résultat obtenu en sortie de la chaîne complète d'analyse, c'est à dire après la reconnaissance, avec la version ASCII correcte du contenu textuel du document. La mesure de performance est déterminée en calculant une distance d'édition entre le texte reconnu et le texte de référence. Les opérations d'éditions considérées sont classiquement, l'insertion, la substitution, ou la suppression. En supposant connues les performances du module de reconnaissance seul, il est possible d'en déduire la mesure de performance du module de segmentation. Bien évidemment ce type d'approche ne peut s'appliquer qu'aux documents textuels pour lesquels on peut disposer de la version plein texte ASCII et de modules de reconnaissance fiables, adaptés au vocabulaire et à la langue considérés. Le problème de ce type d'approche est que l'on n'évalue pas directement les performances du module de segmentation, mais qu'on les déduit à partir des performances globales du système complet. Avec ce type d'approche il est très difficile de savoir le type d'erreurs commises lors de la segmentation.

Approches basées sur une comparaison avec une vérité-terrain

Ce type d'approche d'évaluation des performances consiste à comparer directement le résultat obtenu en sortie du module de segmentation avec une vérité-terrain définie manuellement, de détecter les erreurs de segmentation et d'en déduire une mesure de performance. La difficulté de ce type d'approche réside dans le fait qu'il faut disposer d'une vérité-terrain au niveau segmentation (structure physique et/ou logique), c'est à dire à un niveau relativement bas (certains travaux vont même jusqu'au niveau du pixel [Randriamasy 94] [Shafait 06]), et qu'il faut déterminer une mesure de performance apte à comparer les différentes méthodes. L'avantage est qu'avec ce type d'approche on est capable d'évaluer directement les performances du module de segmentation, indépendamment du reste de la chaîne de traitement, et que de plus on peut déterminer de manière précise les erreurs de segmentation, ce qui permet éventuellement de les corriger. Le fait de disposer de la vérité-terrain permet également d'effectuer un apprentissage automatique des paramètres du module de segmentation.

Mao et al. proposent une méthodologie d'évaluation et de comparaison d'algorithmes de zonage et appliquent cette méthodologie d'évaluation sur trois méthodes dédiées à l'imprimé très connues, à savoir la méthode X-Y cut [Nagy 84] [Nagy 92] [Ha 95], la méthode Docstrum de O'Gorman [O'Gorman 93], et la méthode de Kise basée sur l'utilisation d'un diagramme de Voronoï [Kise 98]. La méthodologie utilise une mesure de performance basée sur la précision de la segmentation au niveau ligne. Les résultats obtenus en sortie du module de zonage sont comparés avec la vérité-terrain. Cette mesure de précision au niveau ligne est déterminée en calculant le rapport du nombre de lignes de la vérité-terrain correctement segmentées sur le nombre total de lignes de texte définies dans la vérité-terrain. Les lignes qui sont considérées par les auteurs comme étant mal segmentées sont celles qui sont soit fusionnées horizontalement en une seule ligne, soit fragmentées horizontalement en plusieurs lignes, ou qui n'ont tout simplement pas été détectées.

Dans le cadre des conférences ICDAR'2003 et ICDAR'2005, des com-

pétitions dédiées à la segmentation d'images de documents ont été proposées [Antonacopoulos 03] [Antonacopoulos 05], afin de comparer un certain nombre de méthodes selon un protocole bien établi et défini à l'avance, sur des données réelles, à savoir sur des documents imprimés scannés présentant des structures couramment rencontrées dans la vie courante, tels que des pages de journaux, des articles techniques, ou des pages de magazines. Il est ressorti de ces compétitions que bien que les techniques dédiées aux document imprimés sont arrivées à une certaine maturité, et que les résultats obtenus avec la plupart des méthodes sont bons, il existe toujours néanmoins un réel besoin de méthodes robustes capables de traiter n'importe quel type de document que nous manipulons tous les jours dans notre vie quotidienne [Antonacopoulos 05] . Évidemment cela est d'autant plus vrai pour les documents manuscrits, ou mixtes imprimé/manuscrit.

2.5 Conclusion

Comme nous avons pu le voir, il existe de très nombreuses méthodes pour l'analyse de la mise en page et la reconnaissance de documents notamment en ce qui concerne les documents imprimés. Cependant les méthodes proposées sont souvent très spécifiques à certains types de documents, et nécessitent le réglage de nombreux paramètres difficiles à déterminer. De plus elles reposent sur une séparation des différents traitements, notamment de la segmentation et de la reconnaissance, ce qui peut poser de nombreux problèmes dans le cas de documents fortement bruités. Ces méthodes présentent donc des problèmes d'adaptabilité à différents types de documents. En particulier elles ne sont pas directement applicables aux documents manuscrits du fait de la forte variabilité spatiale de ceux-ci.
Dans les documents imprimés la structure recherchée est généralement une structure hiérarchique en blocs. On cherche à extraire ces blocs, à les étiqueter en fonction de leur nature (texte, image, graphique), et à déterminer les relations spatiales entre ces différentes régions. Les blocs textuels sont récursivement segmentés en lignes, en mots puis en lettres (pour permettre éventuellement l'application d'un OCR). Dans le domaine des documents manuscrits on se restreint généralement à une segmentation en lignes puis en mots. Des méthodes différentes sont utilisées pour les deux types de segmentation. En ce qui concerne la segmentation en lignes, les approches sont

souvent ascendantes du fait de la variabilité de l'écriture. Ces approches reposent sur des hypothèses a priori très fortes (par exemple les lignes de texte sont droites et horizontales) et sur le réglage de nombreux seuils ou paramètres difficiles à déterminer. Les principaux défauts de ces méthodes résident dans le fait qu'elles n'exploitent pas ou pas suffisamment le contexte et que par conséquent l'analyse est souvent très locale. De plus, la détermination empirique des paramètres et l'utilisation implicite d'hypothèses ou de connaissances a priori, représentent une limitation pour l'adaptation de ces méthodes à différents types de documents et à différentes tâches de segmentation.

Nous proposons donc des approches permettant à la fois de prendre en compte la forte variabilité spatiale des documents manuscrits, mais également d'intégrer des connaissances sur la structure, pour procéder simultanément à la segmentation et la reconnaissance, tout en permettant un paramétrage simple par des procédures d'apprentissage supervisé, autorisant une adaptation souple de la méthode à différents cas. La théorie des modèles graphiques probabilistes, notamment celle des champs de Markov cachés et des champs aléatoires conditionnels, répond à l'ensemble des critères énoncés précédemment, et ceci dans un cadre mathématique bien établi. Nous présentons donc dans les chapitres suivants la théorie des champs de Markov puis celle des champs aléatoires conditionnels, dans un cadre bayésien d'analyse d'image, et nous proposons deux approches d'analyse d'images de documents, basées sur ces modèles et sur des techniques d'apprentissage automatique.

Chapitre 3

Analyse d'images de documents par champs de Markov

3.1 Introduction

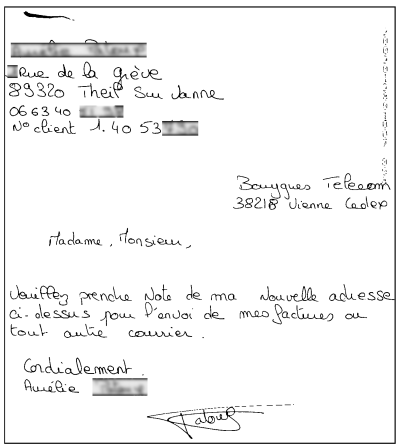
Les champs markoviens sont utilisés avec un certain succès depuis plus de 20 ans dans le domaine de l'analyse d'image pour résoudre, comme le rappelle Chevalier dans [Chevalier 04], des tâches aussi diverses et variées que l'analyse de textures [Derras 93][Lorette 99], de restauration et de débruitage [Geman 84], de segmentation [Bouman 94][Held 97] ou encore de reconnaissance de formes [Cai 01] [Geoffrois 04]. Pour résoudre ces différentes tâches différents modèles de Markov 2D ou pseudo-2D, y compris des modèles hiérarchiques et multirésolution ont été proposés [Bouman 94][Graffigne 95]. Pourtant à ce jour ils n'ont pratiquement pas été utilisés dans le cadre de l'analyse d'images de documents. A notre connaissance les seuls travaux concernant l'application de champs de Markov 2D pour des tâches d'analyse d'images de documents sont ceux de Cheng et al. [Cheng 97][Cheng 98][Cheng 01] sur l'étiquetage d'images de documents imprimés à l'aide de modèles de champs de Markov multi-échelles, appliqués à la séparation texte/graphique, ceux de Wolf et al. [Wolf 02] sur la binarisation de documents textuels dégradés, et ceux de Zheng et al. sur l'identification de texte manuscrit dans des documents dégradés [Zheng 03]. Toutefois quelques modèles markoviens pseudo-2D ont également été propo-

sés pour l'analyse de structures dans des documents, ou dans le domaine de la reconnaissance de l'écriture manuscrite. Ainsi par exemple Kopec et al. [Kopec 94] utilisent des modèles de Markov pour le décodage d'images de documents et l'appliquent à la reconnaissance de la structure de documents de type annuaires téléphoniques. Divers modèles markoviens 1D et 2D, ou pseudo-2D ont également été proposés et appliqués à la reconnaissance de l'écriture manuscrite pour la reconnaissance de caractères, de chiffres ou de mots sans avoir recours à une segmentation explicite [Gilloux 94] [Saon 97][Park 98][Choisy 00][Chevalier 03][Chevalier 04][Chevalier 05], et il semblerait que l'utilisation de champs de Markov connaisse un intérêt toujours croissant dans le domaine de la reconnaissance de l'écriture, notamment pour la segmentation ou la reconnaissance de caractères chinois [Wang 00][Zeng 05], car les modèles purement 2D ont montré leur supériorité sur les modèles 1D ou pseudo 2D pour l'analyse de signaux bidimensionnels.

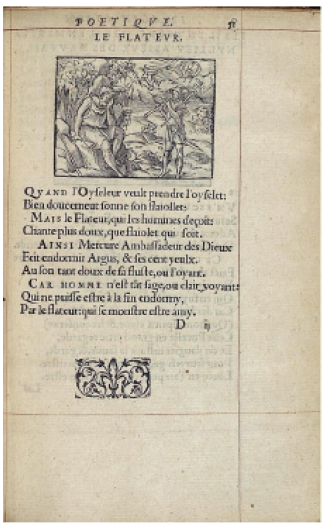
Le formalisme des champs de Markov se base à la fois sur la théorie des probabilités et sur la théorie des graphes. Il s'agit d'un modèle graphique probabiliste très puissant, qui selon nous peut également être utilisé pour l'analyse de documents complexes ou dégradés.

En effet tous les documents qu'il s'agisse de documents imprimés, de documents manuscrits ou de documents anciens présentent une certaine structure, même si cette structure est plus ou moins rigide et plus ou moins complexe (figure 3.1). Les documents sont régis par des règles de structuration. Ces règles de structuration sont implicites ou explicites, et peuvent être plus ou moins respectées par le producteur du document, mais elles existent. La spatialisation de l'information et de l'écriture sur la page traduit cette structuration de l'information que l'auteur cherche à transmettre ou à archiver. Il existe donc des relations fortes et des dépendances entre les différentes entités informatives de la page. C'est le cas également pour les documents manuscrits tels que les brouillons d'auteurs, et souvent l'apparence physique des documents est spécifique pour chaque auteur (figure 3.2).

Ainsi par exemple, si on considère un brouillon manuscrit de Gustave Flaubert, on peut remarquer une mise en page et un aspect visuel parti-

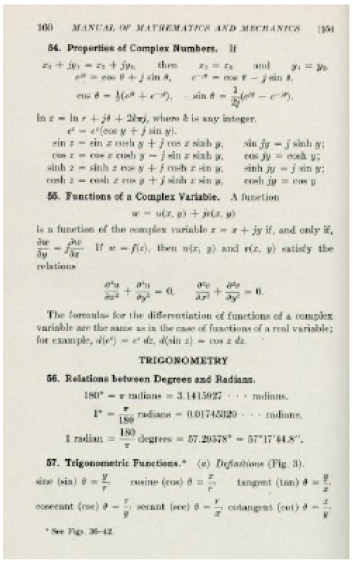


(a) lettre manuscrite



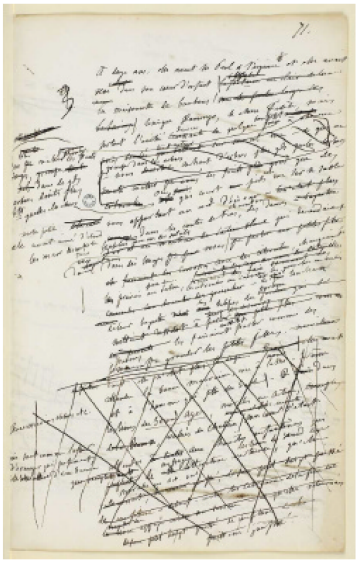
(b) document imprimé de la Renaissance

(c) article de journal

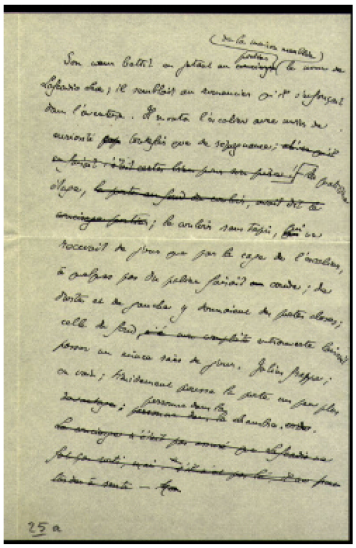


(d) document scientifique

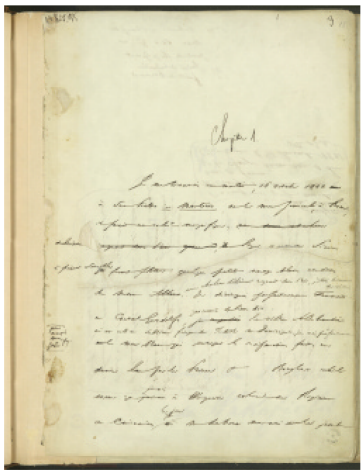
Fig. 3.1. Différentes structures associées à différents types de documents



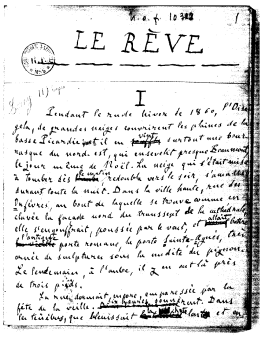
(a) manuscrit de G. Flaubert



(b) manuscrit de A. Gide



(c) manuscrit de Stendhal



(d) manuscrit de E. Zola

Fig. 3.2. Exemples de brouillons manuscrits de différents auteurs

culier et spécifique à l'écrivain, qui sont caractérisés notamment par un large corps de texte occupant environ 2/3 de la page, et par la présence de nombreuses ratures et annotations en marge (figure 3.3). Le modèle de structuration utilisé par l'auteur est implicite et réellement connu de lui seul, mais il existe. La difficulté réside dans le fait que pour certains types de documents, bien que la structure de la page sous-jacente ne soit pas forcément très complexe, la mise en page correspondante présente néanmoins une ambiguïté spatiale très importante dans la disposition des entités. C'est le cas notamment des documents manuscrits. Pour d'autres types de documents comme les documents anciens, ce sont les dégradations du document qui font qu'il n'est pas aisé de déterminer la structure de ces documents.

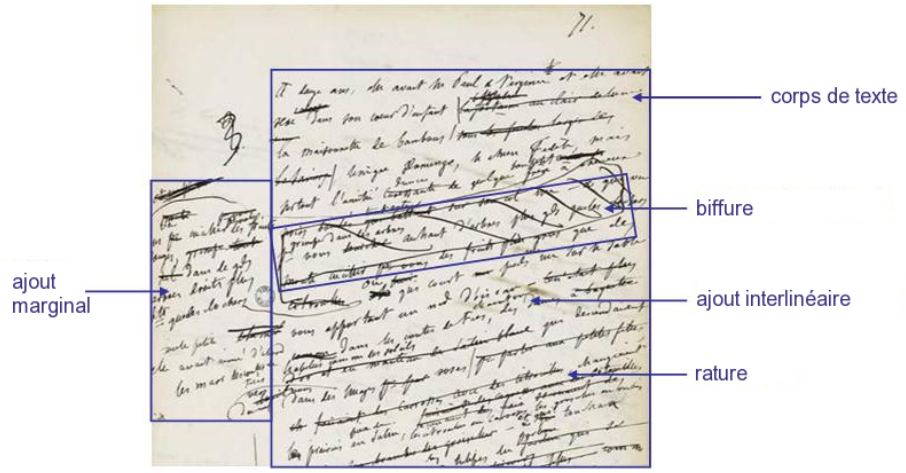


Fig. 3.3. Entités caractéristiques de la structure d'un manuscrit de Flaubert

Nous pensons que l'analyse de ces documents peut être guidée efficacement par une modélisation a priori des connaissances à la fois sur leur structure générique et sur le processus de spatialisation de l'information, de même qu'éventuellement une modélisation du processus de dégradation. Cependant cette modélisation doit être suffisamment souple pour prendre en compte les ambiguïtés. La modélisation par champs de Markov caché permet de répondre à ces exigences. Il s'agit non seulement d'une modélisation probabiliste, qui donc de ce fait est apte à appréhender la variabilité, mais

également d'une modélisation structurelle et contextuelle, qui permet de modéliser les dépendances spatiales entre les primitives de l'image. Nous allons maintenant présenter le cadre théorique des modèles de Markov cachés pour l'analyse d'image.

3.2 Cadre théorique

Nous nous plaçons dans un cadre bayésien classique d'analyse d'image, où l'image observée est supposée être générée par un processus inconnu que l'on cherche à déterminer. Ce processus est supposé être modélisé par une structure d'états ou configuration d'étiquettes. Dans ce contexte le problème de l'analyse d'image est donc de déterminer à partir d'observations et d'un modèle de connaissance a priori sur la structure de l'image, cette configuration d'étiquettes sous-jacente. Ce problème peut être modélisé et résolu de manière efficace à l'aide du formalisme des champs de Markov caché.

Un champ de Markov caché noté (X, Y), est un processus doublement stochastique formé de deux champs aléatoires X et Y dont les variables sont indexées par un ensemble S de sites ou positions s dans l'image noté $S = \{s\}$. Dans le domaine de l'analyse d'image, les sites sont généralement les éléments sur une grille ou un maillage bidimensionnel superposé à l'image. Classiquement pour des problèmes de débruitage d'image chaque site est associé à un pixel de l'image. Cependant pour d'autres problèmes comme la segmentation d'image ou des problèmes de plus hauts niveaux comme l'analyse de scènes, chaque site pourra être associé à un ensemble connexe de pixels ou à des primitives graphiques (composantes connexes ou segments de trait par exemple).

Le champ Y est le champ des observations et se note $Y = (Y_s)_{s \in S}$. Une réalisation de ce champ est notée $Y = y$ et la valeur du champ au site s pour une réalisation donnée y se note $Y_s = y_s$. Les variables aléatoires Y_s de ce champ correspondent à des observations ou mesures effectuées sur l'image. Ces variables aléatoires prennent une valeur scalaire ou vectorielle, continue ou discrète. Il peut par exemple s'agir de la valeur de l'intensité lumineuse au site s sur une image en 256 niveaux de gris auquel cas chaque

variable Y_s prendra une valeur dans un ensemble fini et discret de 256 valeurs, mais on considérera d'une manière plus générale qu'il s'agira d'un vecteur de caractéristiques (feature vector) mesurées dans l'image au site s. L'image donne donc accès à un ensemble de vecteurs de caractéristiques. Le champ X est le champ caché des étiquettes et se note $X = (X_s)_{s \in S}$. Une réalisation de ce champ est notée $X = x$ et la valeur du champ au site s pour une réalisation donnée x se note $X_s = x_s$. Les variables aléatoires X_s de ce champ prennent leur valeur dans un ensemble fini $L = \{l_1, l_2, ..., l_q\}$ de $q = |L|$ étiquettes discrètes. Ces étiquettes désignent les entités de la structure sous-jacente cachée. L'ensemble des réalisations possibles du champ X se note $\Omega = L^{|S|}$ où $|S|$ désigne le cardinal de l'ensemble S, c'est à dire le nombre total de sites de l'image.

La représentation graphique d'un champ de Markov caché est illustrée sur la figure 3.4. Il s'agit d'un modèle graphique non-orienté qui représente les dépendances mutuelles entre les variables du champ d'étiquettes X et génératif car il explicite comment les observations sont obtenues à partir d'une configuration d'étiquettes sous-jacente (par l'intermédiaire des probabilités d'émission des observations sachant les étiquettes $P(y_s|x_s)$). Un modèle de champs de Markov caché permet donc d'accéder à la probabilité conjointe du champ X et du champ Y :

$$P(X, Y) = P(Y/X)P(X)$$

Ce qui permet d'accéder, de manière indirecte grâce au théorème de Bayes, à la probabilité conditionnelle d'une configuration d'étiquettes sachant les observations :

$$\begin{aligned} P(X|Y) &= \frac{P(X,Y)}{P(Y)} \\ &= \frac{P(Y|X)P(X)}{P(Y)} \end{aligned}$$

La théorie de la modélisation par champs de Markov cachés pose deux hypothèses fondamentales. La première est l'hypothèse d'indépendance des observations par rapport aux étiquettes. Cela signifie que les variables Y_s du champ observé Y sont conditionnellement indépendantes sachant une configuration x du champ caché X. En conséquence, la probabilité conditionnelle

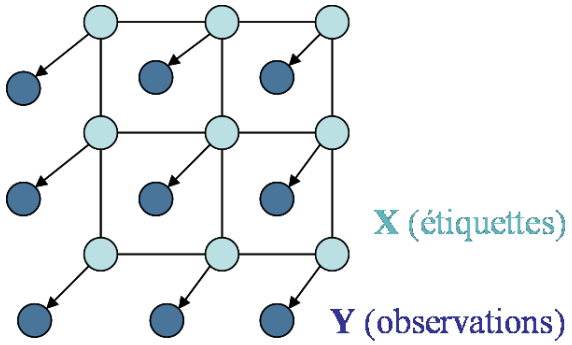


Fig. 3.4. Représentation graphique d'un champ de Markov caché

globale $P(Y = y|X = x)$ d'une réalisation y du champ Y étant donnée une réalisation x du champ X peut se factoriser en un produit de probabilités conditionnelles locales en chaque site de la manière suivante [Maître 03] :

$$P(Y = y|X = x) = \prod_{s \in S} P(y_s|x_s)$$

La seconde hypothèse forte posée par la théorie des champs de Markov cachés est que le champ caché des étiquettes X est un champ markovien selon le système de voisinage V défini sur l'ensemble des sites S de la manière suivante :

$$\forall s, t \in S \; V_s = t \text{tel que} \begin{cases} s \notin V_s \\ t \in V_s \Rightarrow s \in V_t \end{cases}$$

De la même manière que dans le domaine monodimensionnel la théorie des modèles de Markov cachés (Hidden Markov Models ou HMM dans la littérature anglophone) établit que la séquence d'états cachée est une chaîne de Markov, la théorie des champs de Markov cachés établit en 2D que le champ caché d'étiquettes est un champ de Markov. La définition d'un champ alétoire de Markov (Markov Random Field ou MRF) s'énonce de la manière suivante :

Un champ X de variables aléatoires X_s indexées par un ensemble de sites $S = s$ dont les relations sont définies selon un système de voisinage V, est un champ markovien si et seulement si il vérifie deux propriétés : la propriété de positivité et la propriété de dépendance Markovienne entre les variables.

– propriété de positivité :

$\forall x \in q^{|S|}\ P(X = x) \geq 0$

Cela signifie que quelle que soit la réalisation du champ X, sa probabilité est positive.

– propriété de Markov :

$\forall s \in S\ et\ \forall x \in q^{|S|}$
$P(X_s = x_s | X_r = x_r, r \in S - s) = P(X_s = x_s | X_r = x_r, r \in V_s$

Cette propriété implique que l'état du champ en tout site ne dépend que de l'état du champ sur les sites voisins.

Un modèle de champ de Markov caché introduit donc des dépendances contextuelles locales entre les variables cachées. La probabilité jointe $P(X, Y)$ sur le champ des étiquettes et sur le champ des observations s'exprime alors de la manière suivante en intégrant les deux hypothèses explicitées précédemment :

$$\begin{aligned} P(X,Y) &= P(Y|X)P(X) \\ &= \prod_{s \in S} P(Y_s|X_s) \prod_{s,t \in S} P(X_s|X_t, t \in V_s) \end{aligned}$$

Dans cette expression le premier terme est appelé terme d'attache aux données car il modélise le processus de formation de l'image (il s'agit également parfois d'un modèle de dégradation ou de bruit) en fonction d'une configuration d'états sous-jacente, c'est à dire le lien entre les observations et les étiquettes. Ce terme d'attache aux données est facilement calculable car du fait de l'hypothèse d'indépendance des observations conditionnellement aux étiquettes, il se factorise en un produit de probabilités conditionnelles sur l'ensemble des sites de l'image. Dans le domaine de l'analyse d'image, les observations étant généralement continues, ces probabilités conditionnelles sont souvent modélisées par des modèles de mélange de distributions gaussiennes [Chevalier 04] :

$$P(y|x) = \sum_{k=1}^{M} c_k \mathcal{N}(y, \mu_{k,x}, \Sigma_{k,x})$$

où $\mathcal{N}(y, \mu_{k,x}, \Sigma_{k,x})$ désigne la valeur en y de la k^{me} fonction gaussienne du mélange, de moyenne $\mu_{k,x}$ et de matrice de variance/covariance $\Sigma_{k,x}$. En pratique cette matrice est souvent choisie diagonale. M désigne le nombre de composantes gaussiennes du modèle de mélange, et les c_k sont les pondérations associées à chaque composante du mélange, telles que :

$$\sum_{k=1} c_k = 1$$

Dans l'expression de la probabilité jointe $P(X, Y)$ le second terme représente la probabilité a priori de la configuration d'étiquettes et est appelé terme de régularisation car il permet une certaine homogénéisation du champ d'étiquettes en prenant en compte le contexte sur le voisinage. C'est par l'intermédiaire de ce terme que l'on peut introduire de la connaissance contextuelle dans le modèle, en favorisant certaines configurations d'étiquettes.

Contrairement au terme d'attache aux données ce terme contextuel est très difficile à déterminer sous cette forme là, car un modèle de champ de Markov est a priori un modèle non causal, c'est à dire qu'il y a des interdépendance entre les sites voisins. En effet si d'après la propriété markovienne des champs de Markov cachés, l'état du champ caché en un site donné dépend de l'état du champ caché sur les sites voisins, réciproquement, l'état du champ sur les sites voisins dépend également de l'état du champ en ce site. Le problème est donc difficile à résoudre. Heureusement un théorème fondamental de la théorie des champs markoviens établit que cette probabilité peut se mettre sous une autre forme qui s'exprime en fonction d'une quantité appelée énergie globale. Cette énergie globale peut se décomposer en une somme d'énergies locales sur des sous-parties de l'ensemble des sites S, appelées cliques. Une clique est définie comme un ensemble de sites mutuellement dépendants selon le système de voisinage V. L'ensemble des cliques se note C et une clique de cet ensemble se note c.

Ce théorème fondamental de la théorie des champs de Markov est le

théorème d'Hammersley-Clifford [Hammersley 71]. Il établit qu'un champ markovien suit une distribution particulière appelée distribution de Gibbs. Toute la puissance des champs de Markov tient à ce théorème car il établit le lien entre la propriété de localité des champs de Markov (propriété markovienne) et la propriété de globalité des champs de Gibbs (distribution de Gibbs). En effet ce théorème exprime le fait qu'un champ aléatoire X est un champ de Markov si et seulement si sa distribution $P(X)$ est une distribution de Gibbs, c'est à dire si elle suit la loi suivante :

$$P(X = x) = \frac{\exp(-U(x))}{Z}$$

Dans cette expression $U(x)$ désigne une fonction positive appelée fonction d'énergie (ou simplement énergie globale) et $Z = \sum_{z \in \Omega} U(z)$ est une constante de normalisation appelée fonction de partition. Cette constante permet simplement de s'assurer que la quantité calculée est bien équivalente à une probabilité. La probabilité d'une configuration x sur le champ de Markov X s'exprime donc de la manière suivante :

$$\forall x \in \Omega \; P(X = x) = \frac{\exp(-U(x))}{\sum_{z \in \Omega} \exp(-U(z))}$$

L'énergie globale $U(x)$ se définit comme une somme d'énergies locales, également appelées fonctions de potentiel ou simplement potentiels, calculées sur les cliques c et notées V_c :

$$U(x) = \sum_{c \in C} V_c(x)$$

Les fonctions de potentiel permettent d'intégrer des connaissances a priori sur le processus que l'on cherche à modéliser en favorisant certaines configurations d'étiquettes. Plusieurs formes de fonctions de potentiel ont été proposées dans la littérature, mais les principaux modèles utilisés sont le modèle de potentiel d'Ising et le modèle de potentiel de Potts. Tous deux sont issus du domaine de la physique statistique [Maître 03].

- Modèle de potentiel d'Ising

Le modèle d'Ising est un modèle de potentiel binaire inspiré du domaine

de la physique statistique. Il s'agit du modèle de potentiel le plus ancien, développé initialement pour l'étude du ferro-magnétisme. Il n'est utilisable que dans le cadre de problèmes ne pouvant prendre que deux états ou étiquettes traditionnellement notées -1 et 1 (par analogie avec l'état des spins en ferro-magnétisme), soit $L = \{-1, 1\}$. Ce modèle s'emploie avec des systèmes de voisinage de type 4 ou 8-connexe. Dans le modèle d'Ising on ne prend en compte que des cliques d'ordre 1 et d'ordre 2. Les fonctions de potentiels, de type "tout ou rien", sont définies de la manière suivante pour les cliques d'ordre 2 :

$$V_{c=(s,t)}(x_s, x_t) = \begin{cases} -\beta \ si \ x_s = x_t \\ +\beta \ si \ x_s \neq x_t \end{cases} \quad avec \ x_s, x_t \in -1, 1$$

Cette fonction de potentiel peut également se mettre sous la forme suivante :

$$V_{c=(s,t)}(x_s, x_t) = -\beta x_s x_t$$

Pour les cliques d'ordre 1 ces fonctions de potentiel sont de la forme :

$$V_{c=s}(x_s) = -Bx_s$$

Au final l'énergie totale s'écrit donc de la manière suivante :

$$U(x) = \sum_{c \in C} V_c = - \sum_{c=(s,t) \in C} \beta x_s x_t - \sum_{c=(s) \in C} B x_s$$

En physique statistique β est appelée constante de couplage entre sites voisins et B représente un champ magnétique externe. Lorsque β est positif il agit comme un terme de régularisation, c'est à dire que les configurations les plus probables (ou de manière équivalente d'énergies plus faibles) sont celles pour lesquelles les spins sont de même signe (étiquettes identiques). Une valeur positive de β favorisera donc la formation de régions homogènes, alors qu'une valeur négative favorisera au contraire l'alternance de spins de signes opposés (étiquettes différentes). La valeur de la constante β conditionne donc la régularité du modèle. Le paramètre B (champ magnétique externe) quant à lui favorise a priori par son signe une étiquette (spin) ou une autre. La figure 3.5 représente différentes réalisations d'un champ markovien défini par un modèle d'Ising 4-connexe, avec différentes valeurs de la constante de couplage β. Ces réalisations sont le résultat de

tirages aléatoires par l'algorithme d'échantillonnage de Gibbs-Metropolis. On remarque que plus la valeur de β est élevée, plus le champ est homogène.

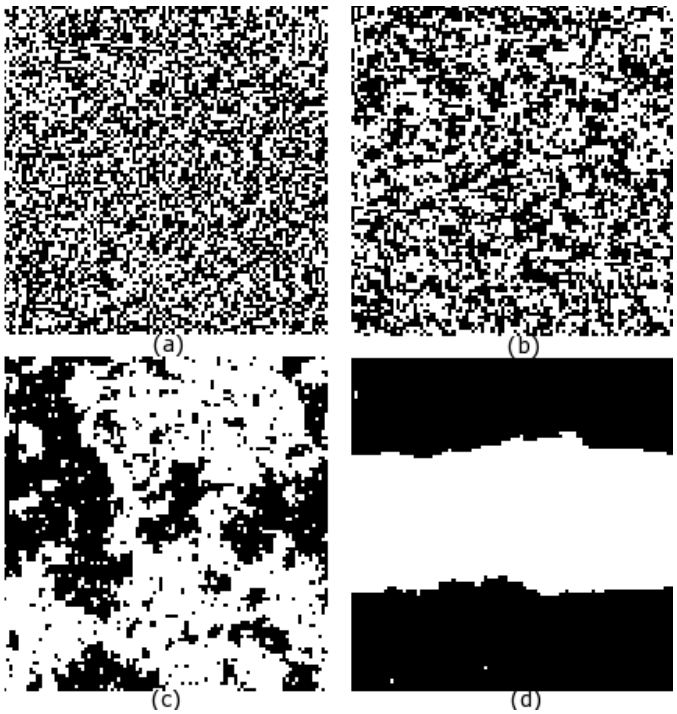


Fig. 3.5. Images simulées avec modèle d'Ising en 4 connexité avec différentes valeurs du paramètre β : (a) $\beta = 0$ (b) $\beta = 0,2$ (c) $\beta = 0,44$ (d) $\beta = 1$

- Modèle de potentiel de Potts

Le modèle de Potts correspond à une généralisation du modèle d'Ising à un espace de d'étiquettes quelconque à m dimensions. Le système de voisinage utilisé avec ce modèle est 4 ou 8-connexe et les fonctions de potentiel sont comme pour le modèle d'Ising de type "tout ou rien", mais ces fonctions ne sont définies que sur les cliques d'ordre 2 de la manière suivante :

$$V_{c=(s,t)}(x_s, x_t) = \begin{cases} -\beta \ si \ x_s = x_t \\ +\beta \ si \ x_s \neq x_t \end{cases}$$

Une valeur positive de β favorise la constitution de larges zones homogènes car en effet les configurations les plus probables (donc de plus faibles énergies) sont celles pour lesquelles les sites voisins ont la même étiquette. La valeur absolue de ce paramètre β contrôle la taille de ces zones homogènes.

Plus la valeur de β sera grande, plus les zones homogènes seront importantes.

Dans l'expression du modèle de Potts donnée ci-dessus, le paramètre β est le même quel que soit le couple d'étiquettes ou de descripteurs considéré, ou quelle que soit la direction de la clique. Il est également possible de définir des modèles de Potts avec des valeurs de β différentes en fonction de la direction des cliques, ce qui permet de privilégier a priori une orientation, ou certains couples d'étiquettes si ce paramètre β dépend des valeurs de ces couples.

En fonction du problème à modéliser, d'autres formes de fonctions de potentiel plus spécifiques sont possibles. Nous ne les détaillerons pas ici et nous renvoyons le lecteur à [Li 03] pour plus d'informations sur ces modèles qui concernent plus spécifiquement les problèmes de restauration et de débruitage.

En pratique la constante de normalisation est difficile à déterminer, car elle nécessite d'effectuer une sommation sur toutes les réalisations possibles du champ X. Ce problème est souvent difficile à résoudre de manière exacte du fait de la structure du graphe qui peut présenter des cycles, mais il peut être résolu de manière approchée à l'aide de différentes techniques telles que l'algorithme Loopy Belief Propagation [Murphy 99] ou des méthodes de simulation Monte-Carlo telles que l'échantillonnage de Gibbs ou de Metropolis [Geman 84][Chellappa 93]. Nous verrons cependant que pour inférer la configuration d'étiquettes optimale il n'est pas nécessaire d'évaluer cette constante.

Les champs de Markov permettent donc de modéliser des propriétés globales (distribution jointe) en utilisant des contraintes locales exprimées par l'intermédiaire des fonctions de potentiel définies sur les cliques du graphe. La probabilité jointe sur le champ des observations et sur le champ des étiquettes peut donc s'exprimer maintenant de la manière suivante :

$$\begin{aligned} P(X = x, Y = y) &= \frac{1}{Z} \prod_s P(y_s|x_s) \exp(-U(x)) \\ &= \frac{1}{Z} \prod_s P(y_s|x_s) \exp(-\sum_{c \in C} V_c(x)) \end{aligned}$$

Hormis la constante de normalisation qui reste difficile à évaluer, cette forme permet de ramener le calcul à la détermination de potentiels locaux.

Pour résumer, un modèle de champ de Markov caché est donc entièrement spécifié par une structure implicite définie par un ensemble de sites et un système de voisinage formant un graphe, un modèle de génération des observations et des potentiels sur les cliques du graphe. Modéliser un problème donné revient donc à expliciter les choix effectués pour chacun de ces points. Ayant exprimé ce modèle, se pose un certain nombre de problèmes à résoudre que nous allons maintenant brièvement présenter et détailler pour certains d'entre eux. Les différents problèmes que l'on peut être amené à résoudre avec un modèle de champ markovien sont les suivants :

- apprentissage des paramètres du modèle
- échantillonnage de réalisations suivant le modèle
- détermination de la probabilité d'une réalisation y du champ Y d'observation
- détermination de la configuration optimale $\hat{x}$ du champ d'étiquettes X sachant une réalisation y du champ Y d'observation

Un modèle de champ de Markov est comme nous l'avons vu précédemment un modèle génératif qui suit une certaine loi paramétrique qui est une loi de Gibbs. Il est donc possible de générer des réalisations ou échantillonner des images suivant les paramètres définis par le modèle spécifié. C'est le problème de l'échantillonnage. Ce problème peut être résolu par des algorithmes d'échantillonnage stochastique tels que l'échantillonneur de Gibbs ou de Metropolis [Maître 03]. Ce problème est rencontré dans le domaine de la synthèse d'image pour modéliser des textures notamment.

La détermination de la probabilité d'une réalisation y du champ observable Y est le problème qui consiste à déterminer la probabilité qu'un modèle donné ait généré les données observées. Les observations sur le champ Y pouvant avoir été générées par différentes configurations x du champ d'étiquettes X sous-jacent, il s'agit de déterminer la probabilité des observations sur toutes les configurations d'étiquettes possibles :

$$P(Y = y) = \sum_{x \in \Omega} P(X = x, Y = y)$$

C'est un problème rencontré typiquement en reconnaissance de forme (reconnaissance de chiffres, de caractères, d'idéogrammes chinois, ...), où une classe est représentée par un modèle, et il s'agit de déterminer la classe d'appartenance d'une forme inconnue représentée par un champ d'observations. On calcule donc pour chacun des modèles la probabilité que le modèle ait généré l'observation, et on attribue à la forme inconnue la classe correspondant au modèle qui maximise cette probabilité. En 1D dans le cadre d'une modélisation par chaînes de Markov cachés (HMM) ce problème peut être résolu efficacement à l'aide de l'algorithme Forward-Backward [Rabiner 89]. Cependant en 2D dans le cadre d'une modélisation par champs de Markov cachés, le calcul de cette probabilité est plus difficile car il nécessite le calcul de la constante de normalisation, ce qui implique le calcul de la probabilité de l'observation sur toutes les configurations d'états possibles. Du fait de la non-causalité des champs de Markov, ce problème ne peut généralement être résolu que de manière approchée en utilisant des techniques telles que l'algorithme Loopy Belief Propagation [Murphy 99] ou des techniques de simulations de type Monte-Carlo. Notons que l'algorithme Loopy Belief Propagation est une version sous-optimale de l'algorithme Belief Propagation de Pearl [Pearl 88] dont l'algorithme Forward-Backward est un cas particulier pour les modèles de séquences.

Dans le cadre de l'analyse d'images de documents à l'aide de champs markoviens nous nous intéressons plus spécifiquement au problème de l'apprentissage et de l'inférence de la configuration optimale du champ d'étiquettes pour une réalisation donnée du champ d'observation.

3.2.1 Problème de l'inférence

Étant donné un modèle de champ de Markov caché défini par les lois de probabilités du couple de champs aléatoires (X, Y), le problème de l'inférence consiste à déterminer la réalisation ou configuration optimale $\hat{x}$ du champ d'étiquettes X fonction d'une réalisation y du champ d'observation, au sens d'un certain critère ou d'une certaine fonction de coût :

$$\hat{x} = \phi(y) \; avec \; \phi : \mathrm{R}^p \rightarrow \Omega$$

où ϕ est une fonction déterministe de l'observation y et p la dimension du vecteur de caractéristiques intrinsèques sur l'observation.

Il s'agit typiquement d'un problème d'optimisation. Pour cela on considère une certaine fonction de coût, notée $\mathcal{C}$ et définie de $\Omega \times \Omega$ dans R^+. Cette fonction de coût définit comment est pénalisée une configuration x différente de la configuration retournée par $\phi(y)$. Cette fonction de coût est toujours positive et nulle si x et $\phi(y)$ sont identiques.

L'estimateur optimal ϕ^{opt} pour cette fonction de coût est alors la fonction ϕ qui minimise l'espérance du coût :

$$E[\mathcal{C}(X, \phi(y))|Y = y] = \sum_{x \in \Omega} \mathcal{C}(x, \phi(y)) P(x|y)$$

La fonction ϕ^{opt} minimise donc l'erreur moyenne conditionnellement à y, et la configuration optimale du champ X est alors :

$$\hat{(}x) = \phi^{opt}(y)$$

Différents fonctions de coût peuvent être définies. Suivant les fonctions de coût considérées, il y a alors différents estimateurs et différentes méthodes de résolution associées. Dans la littérature on considère principalement deux estimateurs : l'estimateur du Maximum A Posteriori et l'estimation du Maximum a Posteriori de la Marginale.

Estimateur du Maximum A Posteriori (MAP)

Dans le cas d'un estimateur MAP, on considère une fonction de coût $\mathcal{C}$ ayant la forme suivante :

$$\begin{aligned} \mathcal{C}(x, x') &= 1 \quad si \quad x \neq x' \\ \mathcal{C}(x, x') &= 0 \quad sinon \end{aligned}$$

Il s'agit donc, avec cette fonction de coût, de pénaliser toute différence entre deux configurations, quel que soit le nombre de sites qui diffèrent. L'espérance de cette fonction de coût est la suivante :

$$\begin{aligned} E[\mathcal{C}(X, \phi(y))|y] &= \textstyle\sum_{x \in \Omega} \mathcal{C}(x, \phi(y)) P(x|y) \\ &= 1 - P(X = \phi(y)|y) \end{aligned}$$

La fonction optimale $\phi^{opt}(y)$ qui minimise l'espérance de cette fonction de coût est donc celle qui maximise la probabilité a posteriori de la configuration d'étiquettes x sachant les observations y :

$$\hat{x} = \phi^{opt}(y) = \arg\max_{\phi}[P(X = \phi(y)|y] = \arg\max_{x}[P(X = x|y)]$$

Pour satisfaire cette fonction de coût, il s'agit donc de déterminer la réalisation $\hat{x}$ du champ X, qui maximise la probabilité a posteriori $P(X|y)$. On parle dans ce cas d'estimateur du Maximum A Posteriori (MAP).

Estimateur du Maximum a Posteriori de la Marginale (MPM)

Un autre estimateur souvent définit dans la littérature, est l'estimateur MPM, associé à la fonction de coût C suivante :

$$\mathcal{C}(x, x') = \sum_{s \in S} L(x_s, x'_s) = \sum_{s \in S} I_{x_s \neq x'_s}$$

Cette fonction de coût pénalise cette fois une configuration proportionnellement au nombre de différences entre deux configurations. Elle considère donc le nombre de sites étiquetés différemment, et se définit donc comme une somme de coûts en chaque site.

L'espérance de cette fonction de coût s'exprime alors de la manière suivante :

$$\begin{aligned} E[L(X, \phi(y))|Y = y] &= \sum_{x \in \Omega} \mathcal{C}(x, \phi(y)) P(x|y) \\ &= \sum_{s \in S} \sum_{x_s} \mathcal{C}(x_s, \phi(y)_s) \sum_{x^s, x_s | y} \end{aligned}$$

Comme $\sum_{x^s} P(x^s, x_s|y) = P(X_s = x_s|y)$, on peut faire apparaître les probabilités conditionnelles et les espérances en chaque site s, soit :

$$\begin{aligned} E[\mathcal{X}(X, \phi(y))|Y = y] &= \sum_{s \in S} \sum_{x_s} \mathcal{X}(x_s, \phi(y)_s) P(X_s = x_s|y) \\ &= \sum_{s \in S} E[\mathcal{C}(X_s, \phi(y)_s)|y] \end{aligned}$$

Cette expression permet donc de passer de la probabilité conditionnelle globale d'une configuration à la probabilité conditionnelle en un site s. C'est une somme de termes positifs, et la fonction ϕ optimale minimise donc par conséquent l'espérance conditionnelle du coût local $E[\mathcal{C}(X_s, \phi(y)_s)|y]$. Ce

résultat est valable pour toutes les fonctions de coût définies par une somme de coûts en chaque site.

On a alors dans ce cas :

$$E[L(X_s, \phi(y)_s)|y] = 1 - P(X_s = \phi(y)_s|y)$$

Et la valeur optimale $\hat{x}_s$ du champ d'étiquettes en chaque site est telle que :

$$\hat{x}_s = \phi^{opt}(y)_s = \arg\max_{\phi}[P(X_s = \phi(y)_s|y)] = \arg\max_{x}[P(X_s = x_s|y)]$$

On maximise donc en chaque site la marginale a posteriori $P(X_s|y)$, et on obtient des estimateurs locaux du maximum a posteriori, calculée en chaque site contrairement à la recherche du MAP qui est globale. C'est pourquoi cet estimateur s'appelle l'estimateur du Maximum Posteriori de la Marginale.

Du fait de la taille de l'espace des configurations Ω, il est en pratique impossible de déterminer par un calcul direct les marginales $P(x_s|)y$. La solution consiste alors à réaliser des approximations empiriques de la loi conditionnelle $P(X|y)$ par des simulation de type Monte-Carlo. Il s'agit de tirer des réalisations du champ X selon sa loi conditionnelle à y, et de calculer à partir de ces réalisations une approximation de l'estimateur MPM. La loi conditionnelle $P(X|y)$ étant une distribution de Gibbs, les tirages des réalisations peuvent s'effectuer avec l'échantillonneur de Gibbs et l'algorithme de Metropolis. Ces solutions algorithmiques sont toutefois très lourdes, très complexes et très coûteuse. En effet, une bonne estimation des marginales nécessite un nombre important de tirages aléatoires.

Nous allons voir qu'au contraire pour l'estimateur MAP, il existe plusieurs solutions algorithmiques très efficaces qui permettent une approximation correcte du MAP sous certaines conditions. Ces solutions sont donc en pratique plus facile à mettre en place et certaines d'entre elles sont très utilisées en analyse d'image.

Méthodes de résolution associées à l'estimateur MAP

Dans un champ de Markov caché, la probabilité a posteriori $P(X|Y)$ n'est pas directement accessible, mais peut être appréhendée au travers de la probabilité jointe $P(X,Y)$ modélisée par le champ. En effet, le théorème de Bayes établit la relation entre probabilité a posteriori, probabilité jointe et probabilité a priori :

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \propto P(Y|X)P(X) = P(X,Y)$$

Dans la mesure où la probabilité a priori des observations $P(Y)$ est indépendante de la configuration d'étiquettes, maximiser la probabilité a posteriori revient donc à maximiser la probabilité jointe $P(X,Y)$. Or comme nous l'avons vu, cette probabilité jointe peut se calculer efficacement à partir des énergies locales. En ne tenant pas compte de la constante de normalisation qui n'intervient pas dans la maximisation on obtient :

$$\begin{aligned}
\hat{x} &= \arg\max_x P(X = x, Y = y) \\
&= \arg\max_x P(Y = y|X = x)P(X = x) \\
&= \arg\max_x (\textstyle\prod_s P(y_s|x_s) \exp(-U(x)) \\
&= \arg\max_x (\textstyle\prod_s P(y_s|x_s) \exp(-\sum_{c\in C} V_c(x))
\end{aligned}$$

En passant en logarithme négatif on obtient l'énergie jointe globale $U(x,y)$, et le problème devient un problème de minimisation de cette énergie globale qui ne fait intervenir que des sommes sur les sites et sur les cliques :

$$\begin{aligned}
\hat{x} &= \arg\min_x U(x,y) \\
&= \arg\min_x (\textstyle\sum_s - \log(P(y_s|x_s)) + \sum_{c\in C} V_c(x)
\end{aligned}$$

Il existe un certain nombre de méthodes d'optimisation permettant de résoudre ce problème et de déterminer la configuration optimale du champ d'étiquettes au sens du critère du MAP. On peut effectuer une classification de ces méthodes d'optimisation en fonction de certains critères. Le premier critère que l'on peut considérer est l'optimalité de la solution trouvée. Il

existe d'une part des méthodes de résolution optimales qui permettent de résoudre le problème de manière exacte mais au prix d'une importante complexité combinatoire, et d'autre part des méthodes dites sous-optimales, qui fournissent une solution approchée, mais en un temps raisonnable. On distingue ensuite d'une part les méthodes basées sur le principe de la relaxation probabiliste et d'autre part les méthodes basées sur le principe de la programmation dynamique et le passage de messages. Enfin pour les méthodes de relaxation, on distingue les méthodes déterministes et les méthodes stochastiques. Les premières utilisent une stratégie d'analyse de l'espace des solutions dont le comportement est prédéterminé et connu alors que les secondes explorent l'espace de manière plus ou moins aléatoire afin d'éviter la convergence vers un optimum local. Le tableau suivant présente la classification d'un certain nombre de méthodes selon ces critères.

	relaxation déterministe	relaxation stochastique	programmation dynamique
optimale		SA [Geman 84] GA [Kim 00] ACS [Ouadfel 03]	
sous-optimale	ICM [Besag 86] HCF [Chou 90]		fusion de régions [Geoffrois 04] Loopy Belief Propagation [Murphy 99]

Nous allons maintenant détailler quelques unes des principales méthodes présentées dans la littérature et très utilisées en pratique pour certaines d'entre elles (ICM).

- L'algorithme de recuit simulé

Cet algorithme (cf algorithme 1) a été introduit en 1983 par Kirkpatrick [Kirkpatrick 83] et adapté par Geman [Geman 84] pour la restauration d'images à l'aide de champs de Markov. Il s'agit d'un algorithme de relaxation stochastique qui permet, en théorie, de trouver l'optimum global de la fonction d'énergie $U(X, Y)$, selon le critère du MAP. Cet algorithme introduit une notion de température qui autorise des changements

aléatoires d'étiquettes, même si ces changements ne permettent pas de réduire l'énergie locale $U_s(X, Y)$, ce qui permet d'explorer aléatoirement l'espace de recherche et donc d'éviter de converger vers un minimum local de la fonction d'énergie. Ces changements d'étiquettes sont contrôlés par le paramètre de température T. Plus cette température est élevée plus les changements d'étiquettes peuvent se produire, et inversement si elle est faible les seules remises à jour de sites autorisées sont celles qui permettent de diminuer l'énergie locale $U_s(X, Y)$. Pour atteindre l'optimum global de la fonction il faut d'une part fixer une température initiale T_0 suffisamment élevée, et d'autre part que cette température ne diminue pas trop rapidement afin d'éviter de converger vers un minimum local de l'énergie. Il est donc conseillé d'utiliser une fonction de décroissance de la température qui soit logarithmique. Pour chaque valeur de la température T, les sites sont parcourus de manière aléatoire en faisant en sorte que tous les pixels soient parcourus plusieurs fois. Pour cela il est nécessaire de fixer un nombre d'itérations à chaque température qui soit suffisamment élevé. Évidemment ce nombre d'itérations N_{iter} dépend du nombre de sites que comportent l'image. En définitive pour l'algorithme de recuit simulé deux paramètres doivent être fixés : la température initiale T_0 et le nombre d'itérations N_{iter} (ou nombre de sites visités pour chaque valeur de la température). Du choix de ces paramètres dépend l'optimalité de la solution. Le réglage de ces paramètres est cependant délicat car il n'existe pas de règles fiables permettant de les fixer. En général ces paramètres sont fixés empiriquement. Cela constitue une limitation de cette méthode. Un autre inconvénient est le temps de calcul nécessaire pour trouver la solution optimale. L'algorithme de recuit simulé est très coûteux en temps de calcul. En effet avec cet algorithme l'espace de recherche est exploré de manière aveugle et de nombreuses mises à jour de la configuration d'étiquettes sont nécessaires avant d'atteindre la convergence. Par contre, l'avantage de cet algorithme, outre le fait qu'il permette de trouver l'optimum global, est qu'il n'est pas dépendant des conditions d'initialisation de la configuration du champ d'étiquettes.

- L'algorithme de recuit simulé multirésolution

Partant du constat que l'algorithme du recuit simulé serait plus rapide

Algorithme 1: algorithme du recuit simulé

Debut Choisir une température initiale $T = T_0$;
Partir d'une configuration initiale x(0) du champ d'étiquettes;
répéter
  i=0;
  répéter
    Choisir un site s (selon stratégie de parcours) et changer aléatoirement son état x en z ;
    Calculer $\Delta U_s = U(X_s = x) - U(X_s = z)$; **si** $\Delta U > 0$ **alors**
      remplacer x par z ;
    fin
    sinon
      si $p < exp(\Delta U/T)$ $p \in [0, 1]$ **alors**
        remplacer x par z ;
      fin
    fin
    i++ ;
  jusqu'à $i = N_{iter}$;
  $T = f(T) = \alpha T$ avec $0 < \alpha < 1$
jusqu'à $T < \epsilon(gel)$;
Fin

si le nombre de sites à considérer était moins important, d'une part, et d'autre part qu'une analyse fine n'est pas nécessaire pour certaines zones de l'image où il y a peu d'ambiguïtés, comme les zones correspondant au fond notamment ou les larges zones homogènes, un algorithme de recuit simulé multirésolution (cf algorithme 2) a été proposé dans [Loncaric 99] pour la segmentation à l'aide de champs de Markov et a été appliqué dans le cadre de la segmentation d'images tomographiques. L'idée générale est de réduire le nombre de variables à optimiser et ainsi de permettre une réduction significative de la combinatoire. Il s'agit bien d'une analyse multirésolution mais en aucun cas d'une analyse hiérarchique, c'est à dire qu'à chacun des niveaux de l'analyse les étiquettes ou états pris en compte sont les mêmes. Le principe de cet algorithme est de procéder à une analyse descendante de l'image en partant d'une résolution faible jusqu'à atteindre la pleine résolution. Pour chacun des niveaux on procède à la segmentation de l'image à l'aide de l'algorithme classique de recuit simulé mais à chaque fois sur un nombre limité de sites ce qui permet d'accélérer le processus. En effet, initialement à la plus faible résolution le nombre de sites est largement moindre qu'à la plus forte résolution. Le résultat de la segmentation à un niveau

donné constitue la condition d'initialisation du champ d'états au niveau suivant. La propagation des hypothèses d'étiquetage d'un niveau au suivant se fait en utilisant un modèle de type arbre quaternaire (quadtree), selon un principe d'expansion qui est le suivant : sur un niveau donné n chaque site de coordonnées quelconques (i, j) est le père de 4 sites respectivement de coordonnées $(2i, 2j)$, $(2i+1, 2j)$, $(2i, 2j+1)$, $(2i+1, 2j+1)$ à la résolution suivante $n+1$, auxquels il propage son étiquette (figure 3.6). Dans la mesure où l'hypothèse d'étiquetage ainsi héritée n'est pas nécessairement la meilleure, il faut ensuite procéder à une phase de validation. Il s'agit en fait de vérifier la corrélation entre le modèle contextuel défini par l'arbre quaternaire et le modèle d'attache aux données, défini par les fonctions de vraisemblance $P^{(n)}(y/x)$ sur chaque niveau de résolution n. Les sites pour lesquels il y a une ambiguïté sur l'état propagé par leur père au niveau précédent, par rapport à l'image observée à la résolution correspondante (terme d'attache aux données), sont stockés dans une liste et seuls ces sites sont ensuite considérés lors de la phase d'optimisation par recuit simulé. A chaque niveau le nombre de sites "instables" à optimiser est donc réduit. Ce processus est appliqué récursivement jusqu'à la pleine résolution. Les paramètres des modèles markoviens sont différents pour chaque niveau de résolution et doivent être déterminés par apprentissage sur une base d'images annotées. La procédure d'apprentissage est inverse à celle correspondant à la segmentation. On part de l'image annotée à pleine résolution, on effectue l'apprentissage des paramètres, puis récursivement on sous-échantillonne l'image annotée et on réestime les paramètres, ceci jusqu'à la résolution la plus faible.

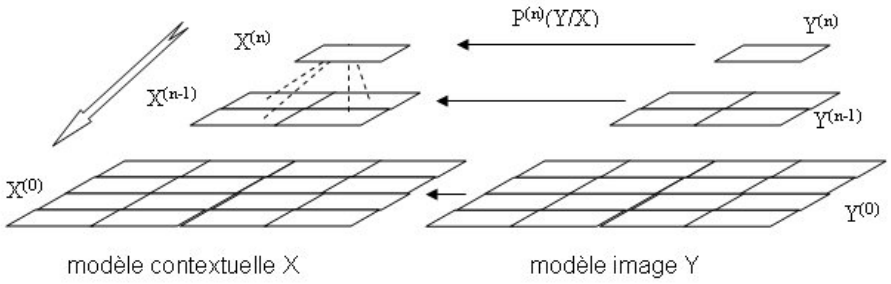


Fig. 3.6. modèle multirésolution quadtree

Algorithme 2: recuit simulé multirésolution

Debut Construire la pyramide multirésolution $S^0, S^1, ..., S^{k-1}$ de l'image par sous-échantillonage, avec S^0 l'image originale à pleine résolution et k le nombre de niveaux;
Appliquer l'algorithme de segmentation par recuit simulé sur l'image S^{k-1} de plus faible résolution;
pour $n = k-2, ..., 0$ **faire**
  Estimer le champ d'étiquette $\hat{x}^n$ au niveau n à partir du champ d'étiquettes $\hat{x}^{n-1}$ au niveau $n-1$ par interpolation selon le modèle quadtree;
  Initialiser la liste L;
  pour *chaque site s de l'image* S^n **faire**
    si $\hat{x}_s^n \neq \arg\max_l p(Y_s^n = y_s / X_s^n = l)$ $l = l_1, l_2, ..., l_q$ **alors**
      Ajouter à la liste L les 4 sites (site s compris) ayant hérités de l'étiquette du même site parent au niveau précédent $n-1$
    fin
  fin
  Appliquer l'algorithme de recuit simulé uniquement sur les sites de la liste L, pour déterminer la segmentation de l'image S^n au niveau n.
fin
Fin

- L'algorithme ICM

L'algorithme des Modes Conditionnels Itérés (Iterated Conditional Modes ou ICM) a été proposé en 1986 par Besag [Besag 86]. C'est un algorithme de relaxation déterministe et itératif sous-optimal qui converge rapidement vers un minimum local de la fonction d'énergie. Il s'agit en fait d'un algorithme de recuit simulé avec une température T nulle, et un balayage déterministe des sites (cf algorithme 3). Contrairement à l'algorithme du recuit simulé, ICM n'autorise donc pas les remontés d'énergie et converge donc vers la solution optimale la plus proche du point initiale, qui peut n'être qu'un minimum local de la fonction d'énergie. L'algorithme ICM peut donc être assimilé à un algorithme de descente de gradient.
Partant d'une configuration initiale d'étiquettes, le principe de cet algorithme consiste à balayer tous les sites de l'image selon un ordre de parcours déterminé, et de remettre à jour les étiquettes associées de manière à faire baisser l'énergie locale en chaque site parcouru. Ainsi l'énergie globale du

champs diminue à chaque itération. Évidemment la modification de l'étiquette d'un site peut avoir des répercussions sur le calcul de l'énergie locale de ses voisins, il est donc nécessaire de réitérer ce balayage de l'image jusqu'à stabilité de l'énergie ou lorsqu'un critère d'arrêt défini préalablement est atteint. Ce critère d'arrêt peut être par exemple le nombre de sites modifiés ou le nombre d'itérations effectuées. De la même manière, l'ordre de parcours des sites a une influence sur la convergence. Avec cette méthode, la qualité finale de la segmentation dépend beaucoup de la configuration d'étiquettes de départ, puisque seul un optimum local peut être atteint. Il est donc recommandé d'utiliser cet algorithme lorsque l'on dispose d'une bonne stratégie d'initialisation de la configuration.

Algorithme 3: algorithme ICM

Debut Initialisation de la configuration du champ d'étiquettes $x(0)$;
Calcul de la nouvelle configuration $x(n+1)$ à partir de la configuration précédente $x(n)$;

1. On balaie l'ensemble des sites s (selon une stratégie de visite des sites) et on change leurs étiquettes de la manière suivante :

$$x_s(n+1) = \arg\min_{l \in L} U_s(X_s(n) = l, Y_s = y_s) \; L = l_1, l_2, ..., l_q$$

$n = n + 1$

2. Retour en 1. Jusqu'à réalisation d'un critère d'arrêt.

Fin

D'après Frey et Jojic [Frey 04], un problème majeur d'ICM est qu'à chaque étape, après que l'étiquette donnant la plus forte énergie locale ait été attribuée au site courant, les informations sur les autres valeurs possibles en ce site ne sont pas mémorisées. Si par exemple deux étiquettes différentes donnent deux énergies locales relativement proches, ICM ne retiendra que l'étiquette de plus faible énergie, sans garder en mémoire le fait que les deux solutions sont très proches, et qu'a priori la deuxième solution est peut-être tout aussi bonne que la solution retenue. En clair le problème de l'algorithme ICM est qu'il ne tient pas compte des incertitudes sur les valeurs des autres variables cachées (états cachés) lorsque la valeur d'une variable est recalculée. Typiquement le calcul de l'énergie locale en un site du champ s'effectue en prenant en compte les valeurs courantes estimées sur les sites voisins. Comme l'expliquent Frey et Jojic un voisin

peu fiable devrait avoir moins de poids dans le calcul de la remise à jour de la valeur du site courant. Malgré ce problème et le fait que l'algorithme ICM soit sous-optimal, il reste très utilisé en pratique car il est d'une part très simple à mettre oeuvre, et d'autre part il converge rapidement vers une solution qui, si elle n'est pas la solution optimale, reste néanmoins très acceptable. En fait la qualité de la solution dépend beaucoup de l'initialisation.

- L'algorithme HCF de Chou et al. [Chou 90]

L'algorithme HCF (Highest Confidence First) est une optimisation de l'algorithme ICM. Il a été proposé par Chou et Brown [Chou 90]. Il s'agit également d'un algorithme de relaxation déterministe et itératif. Partant du principe que l'ordre de parcours des sites a une influence sur la convergence, plutôt que de parcourir les sites successivement sans a priori, l'idée de cet algorithme est d'utiliser une stratégie de parcours optimisée (cf algorithme 4). Il s'agit de remettre à jour les sites selon un critère de stabilité, en partant des sites les moins stables, la stabilité étant calculée à partir de la différence d'énergie locale, entre l'étiquette courante du site et l'étiquette optimale. Le site le moins stable est toujours traité en priorité. Pour gérer les sites à parcourir une pile est utilisée. Les sites sont rangés dans cette pile en fonction de leur stabilité. L'algorithme se termine lorsque tous les sites sont stables, c'est à dire lorsque la pile est vide.

- L'algorithme Loopy Belief Propagation

L'algorithme Loopy Belief Propagation (LBP), est une généralisation aux graphes quelconques présentant éventuellement des cycles, de l'algorithme Belief Propagation (BP) proposé par Pearl pour résoudre de manière exacte le problème de l'inférence sur des arbres ou des chaînes. A quelques modifications près, l'algorithme Loopy Belief Propagation est quasiment identique à l'algorithme Belief Propagation. Il s'agit toutefois d'un algorithme d'inférence inexact qui ne permet de trouver qu'une solution approchée. En effet l'inférence exacte est en général trop coûteuse voire même dans certains cas impossible. Le terme "loopy" vient du fait qu'il s'applique aux graphes cycliques. Les algorithmes Belief Propagation et Loopy Belief Propagation sont des algorithmes itératifs qui procèdent par passage de messages. Ils permettent de calculer soit les marginales

Algorithme 4: algorithme HCF

Debut Marquer tous les sites comme "non visités";
Calculer la stabilité G de tous les sites et ranger les sites dans une pile P par ordre croissant de stabilité;
Retirer le premier site s de la pile P;
si $G_s >= 0$ **alors**
    fin
fin
sinon
    si *s est "non visité"* **alors**
        marquer comme "visité" et attribuer étiquette donnant la plus faible énergie locale
    fin
    sinon
        changer son étiquette de manière à diminuer son énergie
    fin
fin
Mettre à jour stabilité du site s et de ses voisins;
Remettre le site s dans la pile P;
Réordonner la pile P;
Fin

(version sum-product), soit la configuration optimale (version max-product).

Leur principe de fonctionnement est le suivant : en chaque site on calcule le produit des messages entrants transmis par les sites voisins, et de l'évidence locale (attache aux données), que l'on marginalise (sum-product) ou maximise (max-product) ensuite pour former les messages sortants qui sont propagés aux voisins. Dans le cas de l'algorithme Loopy Belief Propagation cette procédure est répétée itérativement. Cependant l'algorithme Loopy Belief Propagation ne présente aucune garantie de convergence.

Les algorithmes Belief Propagation et Loopy Belief Propagation permettent dans leur version "sum-product" de calculer les fonctions marginales, mais en remplaçant les sommes par une maximisation, ils permettent également de trouver le maximum a posteriori. On parle alors d'algorithme "max-product". On notera d'ailleurs que l'algorithme de Viterbi utilisé pour les modèles de Markov cachés est un exemple très connu d'algorithme "max-product", qui est en fait un cas particulier d'application de l'algorithme Belief Propagation aux séquences.

L'inconvénient de ces algorithmes par passage de messages est qu'ils sont très coûteux puisqu'ils nécessitent une marginalisation sur un très grand nombre de variables. Il ne sont donc efficaces que pour un nombre restreint de sites et d'étiquettes dans le modèle. De plus la convergence n'est pas garantie sur des problèmes 2D.

- L'algorithme de programmation dynamique 2D de Geoffrois et al. [Geoffrois 03]

En 2003, Geoffrois a proposé d'étendre le principe de la programmation dynamique, fondamentalement monodimensionnel, aux espaces à 2 dimensions ou plus [Geoffrois 03]. Ce principe a été mis en oeuvre dans une méthode de décodage de champs Markoviens présentée dans [Geoffrois 04], et appliquée à la reconnaissance de chiffres manuscrits. L'idée de cette méthode de décodage est de diviser l'image en sous-régions et de déterminer indépendamment pour chacune de ces sous-régions l'ensemble des configurations possibles et leur énergie associée, puis de déterminer les configurations possibles de la région résultant de la fusion de ces deux sous-régions à partir des configurations des deux sous-régions. Ce processus de fusion est répété itérativement en partant de régions unitaires correspondant à un site, et pour lesquelles l'initialisation est triviale (calcul de Maximum de Vraisemblance sur le terme d'attache aux données de la fonction d'énergie), jusqu'à obtenir une région recouvrant toute l'image. Cette méthode utilise donc bien le principe du "diviser pour régner" de la programmation dynamique. En tenant compte de la nature des champs de Markov, de dépendance locale contextuelle entre les sites de l'image, il apparaît que lors de la fusion il ne peut y avoir que des interactions entre les sites appartenant aux frontières des deux sous-régions fusionnées. De ce fait il n'est pas nécessaire de déterminer toutes les configurations possibles des sites des régions entières, mais uniquement toutes les configurations des frontières, et de ne retenir pour chacune d'elles que la meilleure configuration des sites de l'intérieur des régions. Cela permet de réduire le nombre de configurations à mémoriser. Cependant en pratique ce nombre de configurations reste très élevé, surtout si l'image à décoder est très grande, et il y a alors un risque d'explosion combinatoire. C'est pourquoi un élagage est appliqué pour ne conserver que les meilleures configurations (celles de plus faible énergie) des frontières des

régions. Ce paramètre est appelé seuil d'élagage. Il s'agit du seul paramètre à régler de cette méthode. Cependant ce paramètre est critique et est très difficile à déterminer. En effet si ce seuil est trop faible la solution risque d'être largement sous-optimale, puisque de nombreux chemins de l'espace de recherche localement moins bons, mais menant en réalité à la solution optimale risquent d'être élagués lors du processus. Par contre si il est trop élevé le nombre de configurations à mémoriser et le nombre de combinaisons à tester lors de la fusion devient prohibitif. Cela constitue la limitation la plus contraignante de cette méthode, surtout s'il s'agit de traiter des images haute résolution de grandes dimensions. En théorie sans élagage l'ordre de fusion des régions n'a pas d'influence sur le résultat de la segmentation. En pratique, comme on procède à un élagage, le sens de parcours des sites a une importance. Il est donc possible de considérer différents sens de parcours, comme par exemple un parcours horizontal en ligne (figure 3.7, vertical en colonne, ou alternativement horizontal puis vertical. Beaucoup d'autres stratégies de fusion sont évidemment envisageables.

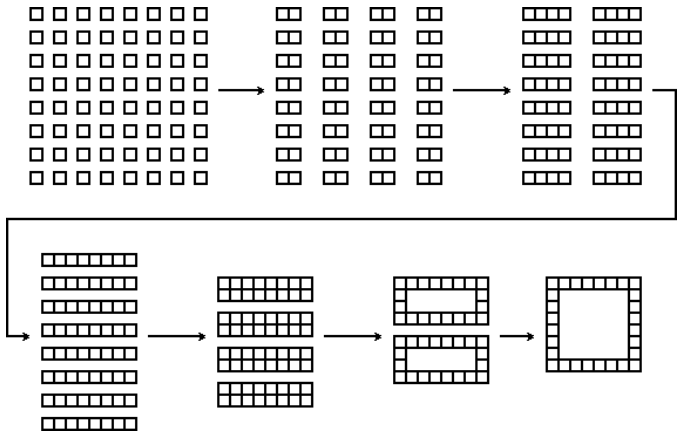


Fig. 3.7. stratégie de fusion horizontale

Nous n'avons fait ici que présenter les principales méthodes d'inférence pour les modèles de champs de Markov cachés présentées dans la littérature. Cette liste n'est bien évidemment pas exhaustive, et d'autres méthodes d'optimisation ont été proposées pour résoudre ce problème, notamment des méthodes basées sur les algorithmes génétiques [Kim 00] ou des méthodes basées sur des systèmes de colonies de fourmis [Ouadfel 03]. Comme l'algorithme du recuit simulé ces méthodes permettent en théorie de trouver la

solution optimale, cependant elles sont réputées pour converger lentement. Il n'existe vraisemblablement pas de méthode idéale, chacune ayant ses avantages et ses inconvénients.

3.2.2 Apprentissage d'un modèle de champ de Markov caché

Le problème de l'apprentissage consiste à déterminer les paramètres d'un modèle de champ de Markov. Comme nous l'avons vu précédemment, un modèle de champ de Markov caché est défini par plusieurs paramètres. D'une part les paramètres du modèle d'attache aux données et d'autre part , les paramètres du modèle de régularisation défini par un ensemble de fonctions de potentiel sur les cliques. Dans la résolution de problèmes par champs markoviens on considère généralement que la structure du graphe sous-jacent est connue et définie. L'apprentissage concerne donc la détermination des paramètres du modèle d'attache aux données, et la détermination des paramètres du modèle de régularisation. Cet apprentissage peut être réalisé de manière supervisée si on dispose d'exemples étiquetés, ou sinon de manière non-supervisée. Dans le cas où les données d'apprentissage sont complètes, c'est à dire que l'on dispose pour un ensemble d'observations de la configuration d'états correspondante, ces deux modèles peuvent en général être appris indépendamment car la modélisation par champs de Markov cachés sépare bien modèle d'attache aux données et modèle contextuel.

- Apprentissage du modèle d'attache aux données

Le modèle d'attache aux données consiste classiquement en un modèle de mélange. Chaque observation est supposée être générée par un mélange de noyaux gaussiens :

$$P(y|x) = \sum_{k=1}^{M} c_k \mathcal{N}(y, \mu_{k,x}, \Sigma_{k,x})$$

L'apprentissage consiste donc à déterminer les paramètres des mélanges modélisant chaque classe. Ces paramètres sont le nombre M de noyaux gaussiens, les pondérations C_k associées et les paramètres de chaque noyau, à savoir le vecteur moyen $\mu_{k,x}$ et la matrice de variance/co-variance $\Sigma_{k,x}$. Les

valeurs optimales $\hat{\theta}$ des paramètres $\theta = \{c_k, \mu_{k,x}\}, \Sigma_{k,x}$ des mélange sont classiquement appris en utilisant l'algorithme EM (Expectation-Maximization) de manière à maximiser la vraisemblance des données d'apprentissage $y = \{y_1, ..., y_n\}$:

$$\hat{\theta} = \arg\max_{\theta} \log[P(y|\theta)]$$

Une observation est supposée émise par l'un des M noyaux gaussiens :

$$p(y) = \sum_{k=1}^{M} P(k)p(y|k)$$

A l'étape d'estimation E, on calcule la probabilité que le noyau ait émis la i^{me} observation y_i :

$$p_k^i = \frac{c_k \mathcal{N}(y_i, \mu_k, \Sigma_k)}{\sum_{l=1}^{M} c_l \mathcal{N}(y_i, \mu_l, \Sigma_l)}$$

A l'étape de maximisation on obtient les nouveaux paramètres des noyaux :

$$\begin{aligned} c'_k &= \frac{1}{N} \sum_{i=1}^{N} p_k^i \\ \mu'_k &= \frac{\sum_{i=1}^{N} p_k^i y_i}{\sum_{i=1}^{N} p_k^i} \\ \Sigma'_k &= \frac{\sum_{i=1}^{N} p_k^i (y_i - \mu'_i)(y_i - \mu'_i)^t}{\sum_{i=1}^{N} p_k^i} \end{aligned}$$

Les étapes d'estimation E et de maximisation M sont ensuite réitérées jusqu'à la convergence.

- Apprentissage du modèle de régularisation

Le modèle de régularisation consiste généralement en un modèle d'Ising pour un problème binaire ou en un modèle de Potts pour un modèle n-aire. Pour le modèle d'Ising les paramètres à déterminer sont le coefficient de couplage β et le champ magnétique externe B. Pour le modèle de Potts seule la constante de couplage β doit être déterminée. Dans les deux cas à nouveau ce problème peut être résolu en utilisant un algorithme itératif de type EM [Maître 03].

3.3 Application des champs de Markov à l'analyse d'images de documents

On se propose d'utiliser des modèles de champs markoviens pour procéder à l'analyse d'images de documents, et d'illustrer l'intérêt de la méthode proposé sur des documents anciens et dégradés qui sont difficiles à traiter par les méthodes traditionnellement proposées dans la littérature, comme nous l'avons vu dans le chapitre précédent.

L'utilisation d'un modèle de champ de Markov caché pour résoudre un problème donné, nécessite d'expliciter un certain nombre de choix en ce qui concerne la structure du modèle, la modélisation de l'attache aux données et les fonctions de potentiel utilisées.

3.3.1 Modèle proposé

On considère que le modèle a une structure de grille régulière 2D. Pour cela on applique sur l'image un maillage régulier noté G de dimensions $n \times m$, nous donnant accès à un ensemble de sites ou régions dans l'image. Si la maille a une taille unitaire, chaque site correspond exactement à un pixel de l'image, sinon chaque site correspond à une région de l'image constituée d'un ensemble de pixels et délimitée par la maille. Sur la grille chaque site est repéré par ses coordonnées (i, j) et est noté $g(i, j)$.

On définit sur cette grille un système de voisinage 4 ou 8-connexe. Ceci définit de manière équivalente une structure de graphe, telle que les noeuds du graphe correspondent aux sites de l'image et les arêtes déterminent les relations de dépendance entre les sites selon le système de voisinage considéré. Cette structure de graphe nous donne accès à un ensemble de cliques. La forme des cliques obtenues avec une structure de grille régulière, en fonction du système de voisinage choisi, est illustrée sur la figure 3.8.

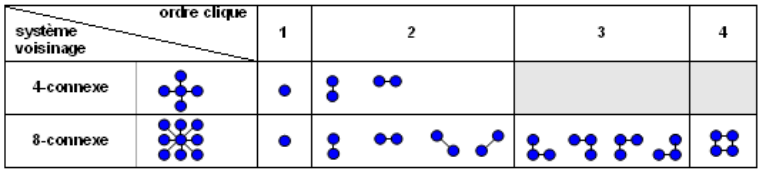


Fig. 3.8. cliques associées à un système de voisinage 4 et 8-connexe

Sur cette structure de graphe nous définissons donc un modèle de champ de Markov caché noté (X, Y), où X désigne le champ des étiquettes et Y est le champ des observations. Les variables aléatoires $X_{g(i,j)}$ du champ X, indexées par les sites $g(i, j)$ de l'image prennent leurs valeurs dans un ensemble fini et discret d'étiquettes. Nous considérons un problème d'analyse d'images de documents, les étiquettes traduisent donc l'appartenance à une entité de la structure du document. Ces étiquettes sont donc définies spécifiquement en fonction du problème d'analyse considéré. De la même manière chaque variable aléatoire $Y_{g(i,j)}$ du champ des observations, indexée par un site g de coordonnées (i, j) sur la grille G, prend pour valeur un vecteur de caractéristiques extraites de l'image. Les caractéristiques considérées peuvent être continues ou discrètes. Le modèle d'attache aux données faisant le lien entre le champ des observations et le champ des étiquettes prend la forme d'un modèle de mélange.

- Extraction des caractéristiques de l'image

Nous extrayons en chaque site de l'image un ensemble de caractéristiques. Dans la mesure où les vraisemblances conditionnelles des observations sachant les états $P(Y/X)$ sont modélisées par des mélanges de distributions gaussiennes, et qu'il est connu que ce type de modélisation ne se comporte correctement qu'avec des espaces de faibles dimensions, nous avons choisi d'extraire un nombre limité de caractéristiques. Nous considérons deux types de caractéristiques : des caractéristiques de densités de pixels et des caractéristiques de position.

Densités de pixels noirs

Nous déterminons la densité de pixels noirs pour le site courant et pour ses 8 voisins selon un système de voisinage en 8-connexité, sur plusieurs niveaux de résolution. Pour cela une pyramide multirésolution est construite pour chaque site de la manière suivante : un site et ses huit voisins sur un niveau de résolution donné correspond à un exactement un site parent sur le niveau de résolution inférieure suivant (figure 3.9).

Pour une pyramide à 2 niveaux de résolution on obtient 18 caractéristiques de densité, soit 9 pour le site courant et ses 8 voisins à la résolution

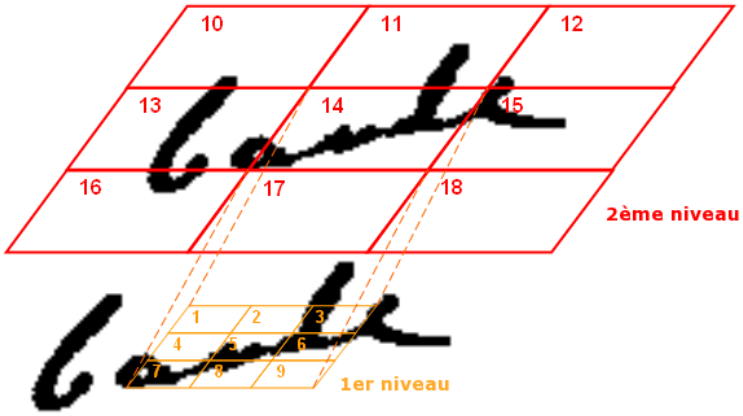


Fig. 3.9. Pyramide multirésolution pour l'extraction de caractéristiques de densité de pixels

courante, et 9 autres pour le niveau suivant. Les caractéristiques de densité de pixels noirs permettent de discriminer les zones de texte comportant une densité moyenne, les zones de ratures comportant une forte densité et les zones d'espaces et de fond comportant une densité faible.

Position relative du site

Une caractéristique très simple à extraire mais qui apporte pourtant beaucoup d'information, est la position du site dans l'image. En effet, certaines entités de la structure du document peuvent être identifiées simplement en tenant compte de leurs positions dans la page. Par exemple si on considère un titre, pour une catégorie de documents donnée, ce titre sera toujours globalement positionné de la même manière dans la page, c'est à dire souvent en haut de la page au-dessus du corps de texte. Il en est de même par exemple pour un numéro de page, ou encore pour une zone d'en-tête qui est toujours située comme son nom l'identique dans la partie supérieure de la page. Nous déterminons donc les coordonnées du site dans l'image, à savoir son abscisse x et son ordonnée y, que nous normalisons respectivement par la largeur et la hauteur de l'image de manière à être indépendant des dimensions de la page. Ces deux coordonnées nous fournissent un vecteur à deux dimensions qui code la position relative du site courant dans l'image. Ce vecteur sera combiné aux autres caractéristiques de manière à mieux discriminer les différentes objets de la structure de la page.

- Potentiels de cliques

En ce qui concerne les potentiels des cliques nous avons choisi d'utiliser les mêmes formes de potentiels que celles proposées par Chevalier dans [Chevalier 04], car elles permettent de tenir compte de l'orientation et des combinaison d'étiquettes. Ces fonctions de potentiel peuvent être vues comme une généralisation des fonctions de potentiel du modèle de Potts. Elles sont obtenues en calculant le logarithme négatif de fonctions d'interactions définies sur les cliques horizontales et verticales d'un système de voisinage 4-connexe, de la manière suivante :

$$I_H = \frac{P(x_k|x_l)}{P(x_k)P(x_l)} \quad I_V = \frac{P(\frac{x_k}{x_l})}{P(x_k)P(x_l)}$$

avec $P(x_k|x_l) = P(x(i,j) = x_k, x_{(i,j+1)} = x_l)$
et $P(\frac{x_k}{x_l}) = P(x_{(i,j)} = x_k, x_{(i,j+1)} = x_l)$

Il s'agit des probabilités jointes de configurations d'étiquettes sur les cliques du graphe. Les potentiels s'expriment donc de la manière suivante à partir de ces fonctions d'interaction :

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

où les termes C_1, C_2 et C_3 désignent les formes des cliques au maximum d'ordre 2 sur un voisinage 4-connexe :

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

avec

$$\begin{array}{rcll} C_1 & = & \{(i,j), & 1 \leq i \leq n, \quad 1 \leq j \leq m\} \\ C_2 & = & \{((i,j),(i+1,j)), & 1 \leq i \leq n, \quad 1 \leq j \leq m\} \\ C_3 & = & \{((i,j),(i,j+1)), & 1 \leq i \leq n, \quad 1 \leq j \leq m\} \end{array}$$

3.3.2 Apprentissage du modèle

L'apprentissage du modèle que nous venons d'expliciter est réalisé de manière totalement supervisée à partir d'une base d'images étiquetées. Pour chacune des images de la base nous disposons des étiquettes associées aux sites de l'image. Les paramètres du modèle à apprendre sont d'une part les paramètres du modèle d'attache aux données, c'est à dire les paramètres des modèles de mélanges, et d'autre part les paramètres du modèle de régularisation, c'est à dire les paramètres des fonctions de potentiel. Les paramètres de ces deux modèles, attache au données et régularisation, sont déterminés de manière indépendante.

- Apprentissage du modèle d'attache aux données

Pour chaque classe la vraisemblance conditionnelle $P(x/y)$ est modélisée par un mélange de distributions gaussiennes. Les paramètres à déterminer pour le modèle d'attache aux données sont donc les paramètres des modèles de mélange associés à chacune des classes. Ces paramètres sont le nombre de composantes gaussiennes du mélange, les pondérations associées à chaque composante du mélange et les paramètres de chaque composante gaussienne, à savoir le vecteur moyen et la matrice de variance/co-variance. L'apprentissage de l'ensemble de ces paramètres peut s'effectuer en utilisant une approche EM (Expectation-Maximization). La modélisation par mélanges gaussiens est classiquement utilisée dans le cadre des champs markovien et de l'analyse d'image, et il existe donc en conséquence un certain nombre d'outils logiciels pour l'apprentissage et l'utilisation de ces modèles. Nous avons utilisé la bibliothèque CLUSTER[1] proposée par Bouman pour l'apprentissage de ces modèles [Bouman 97]. Cette bibliothèque, développée en langage C, implémente l'algorithme EM avec le critère de Rissanen pour la détermination automatique du nombre de composantes gaussiennes du mélange à partir des données d'apprentissage.

- Apprentissage du modèle de régularisation

[1] http ://cobweb.ecn.purdue.edu/ bouman/software/cluster/

Le modèle de régularisation est également appris de manière supervisée à partir des images étiquetées de la base d'apprentissage, en estimant de manière statistique les fréquences d'apparition des différentes configurations d'étiquettes sur les cliques du graphe. Les configurations non rencontrées dans la base d'apprentissage se voient affecter une valeur très faible, mais non nulle de manière à autoriser tout de même ces configurations, mais à les rendre très peu probables.

3.3.3 Inférence à l'aide du modèle

Pour l'inférence du champ optimal d'étiquettes à partir d'observations données et des paramètres du modèle déterminés par apprentissage, nous avons implémenté trois méthodes : la méthode ICM, la méthode HCF et la méthode par fusion de régions de Geoffrois et al. [Geoffrois 03]. Les deux premières sont comme nous l'avons vu des méthodes de relaxation probabiliste sous-optimales et la troisième est une méthode basée sur le principe de la programmation dynamique qui sans élagage permet de résoudre le problème de l'inférence de manière exacte, mais qui en pratique nécessite un élagage et fournit donc une solution approchée elle aussi. Les trois méthodes permettent de trouver la meilleure solution au sens du critère du Maximum A Posteriori. La méthode ICM nécessite une initialisation du champ d'étiquettes. Pour réaliser cette initialisation nous nous basons uniquement sur le terme d'attache aux données en neutralisant le terme de régularisation. Lors de l'initialisation l'étiquette $\hat{y_s}$ affectée à chaque site s est celle qui maximise la vraisemblance conditionnelle de l'observation en ce site :

$$\hat{y_s} = max_y p(x_s|y_s = y)$$

L'initialisation est donc réalisée en appliquant le critère de maximum de vraisemblance en chaque site. L'algorithme ICM ne nécessite pas le réglage de paramètres particuliers si ce n'est le choix du sens de parcours des sites de l'image. Afin de vérifier si ce sens de parcours a une influence importante sur le résultat de l'étiquetage, et le cas échéant de déterminer la meilleure stratégie de parcours pour les différents problèmes d'analyse que nous allons considérer, nous avons implémenté plusieurs stratégie de parcours. Nous comparons les résultats obtenus avec ces différentes stratégies de parcours

dans la section suivante.

L'algorithme HCF ne nécessite ni d'initialisation préalable du champ d'étiquettes, ni le réglage de paramètres. Pour ces raisons l'algorithme HCF est très pratique à utiliser.

3.4 Expérimentations et analyse des résultats

3.4.1 Descriptions des bases d'images

Nous allons décrire les bases d'images de documents que nous avons utilisées pour mener nos expérimentations. Nous avons principalement utilisé deux bases d'images de documents : une base de brouillons manuscrits et une base de documents de la Renaissance. Contrairement aux documents imprimés pour lesquels il existe un certain nombre de bases de référence avec vérité-terrain (bases de l'Université de Washington, base MediaTeam,...) pour l'apprentissage et l'évaluation des performances des méthodes d'analyse d'images de documents, il n'existe pas à l'heure actuelle, à notre connaissance, de bases similaires de documents anciens et historiques. Nous avons donc été obligés de constituer nous mêmes nos propres bases et de produire la vérité-terrain pour différentes tâches de segmentation que nous avons définies. La vérité-terrain a été produite sous forme d'images bitmap couleur pour lesquelles chaque couleur correspond à une étiquette précise de la vérité-terrain, à l'instar de ce qui est proposé dans [Breuel 02] et dans [Shafait 06]. La vérité-terrain est donc définie sous forme d'un étiquetage à un niveau "pixel". Cet étiquetage a été produit très simplement à l'aide d'un outil d'édition d'images, en utilisant notamment des fonctionnalités telles que la sélection polygonale. L'avantage de ce mode de représentation de la vérité-terrain, outre le fait que l'on peut effectivement utiliser n'importe quel logiciel d'édition d'image pour la produire et n'importe quel visualisateur d'image pour la visualiser comme le souligne Breuel dans [Breuel 02], est qu'elle est directement dans le même format que le résultat produit en sortie par la méthode d'analyse d'images de documents que nous proposons. Dans le cas d'une évaluation au niveau pixel il est donc très facile de comparer le résultat obtenu avec la vérité terrain.

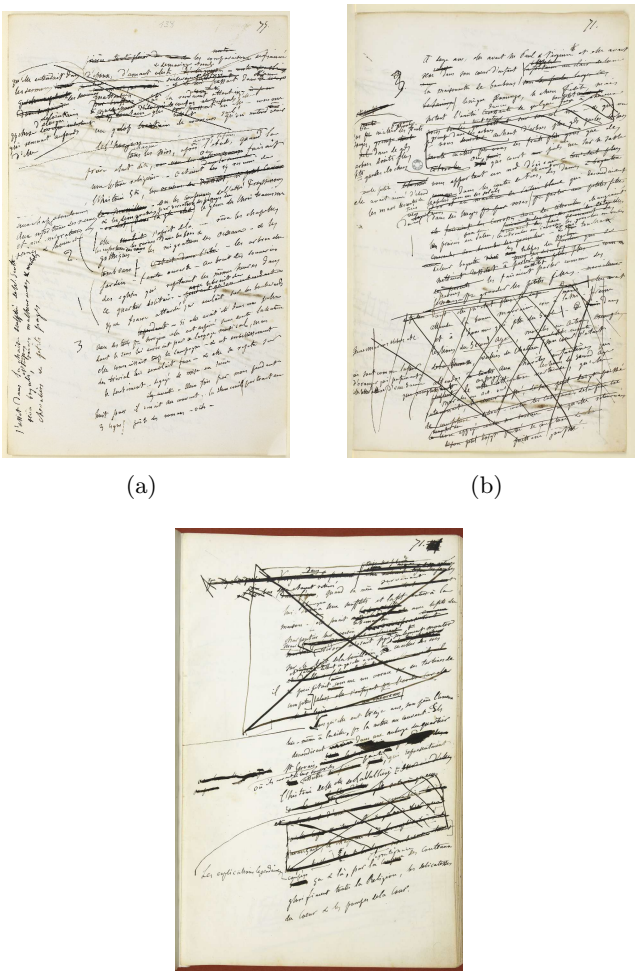
3.4.1.1 Base "Bovary"

Cette base est constituée de 69 images de documents manuscrits de l'écrivain Gustave Flaubert. Ces manuscrits sont issus du dossier génétique du roman "Madame Bovary". Il s'agit de brouillons manuscrits de l'auteur, c'est à dire de documents particulièrement complexes à analyser comme nous pouvons le voir sur la figure 3.10. Les images des documents qui nous ont été fournies par la Bibliothèque Municipale de Rouen, sont des images en 256 niveaux de gris numérisées à une résolution de 300 dpi. Elles sont en moyenne de dimensions 2450×3682, c'est à dire qu'il s'agit d'images de relativement grande taille. Nous avons tout d'abord appliqué quelques prétraitements sur ces images. Etant donné que les caractéristiques image que nous considérons sont essentiellement des caractéristiques extraites sur des images binaires (densités de pixels noirs), nous avons tout d'abord binarisé les images avec un seuillage manuel. Les images dont nous disposions étant des images brutes de scan, elles comportent un bord noir assez important dû à l'acquisition. Nous avons choisi d'effectuer un recadrage automatique de l'image pour supprimer ce bord noir. Il aurait été possible grâce à notre méthode d'ajouter un état supplémentaire pour modéliser ce bord, mais cela aurait demandé un effort supplémentaire d'étiquetage manuel. Nous avons donc choisi de le supprimer dès le départ lors des prétraitements. La base a été divisée en 3 sous-bases, une pour l'apprentissage des paramètres du modèle, une autre pour la validation des paramètres appris et la dernière pour l'évaluation, soit 23 images dans chaque sous-base.

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 considérons deux tâches de segmentation à deux niveaux différents afin de pouvoir évaluer les capacités d'adaptation de la méthode à différentes tâches d'indexation : une analyse au niveau bloc et une analyse au niveau ligne.

Etiquetage au niveau bloc

De manière à pouvoir évaluer précisément les performances de notre approche, et pouvoir comparer les méthodes de décodage entre elles, selon des critères que nous allons définir, nous considérons 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



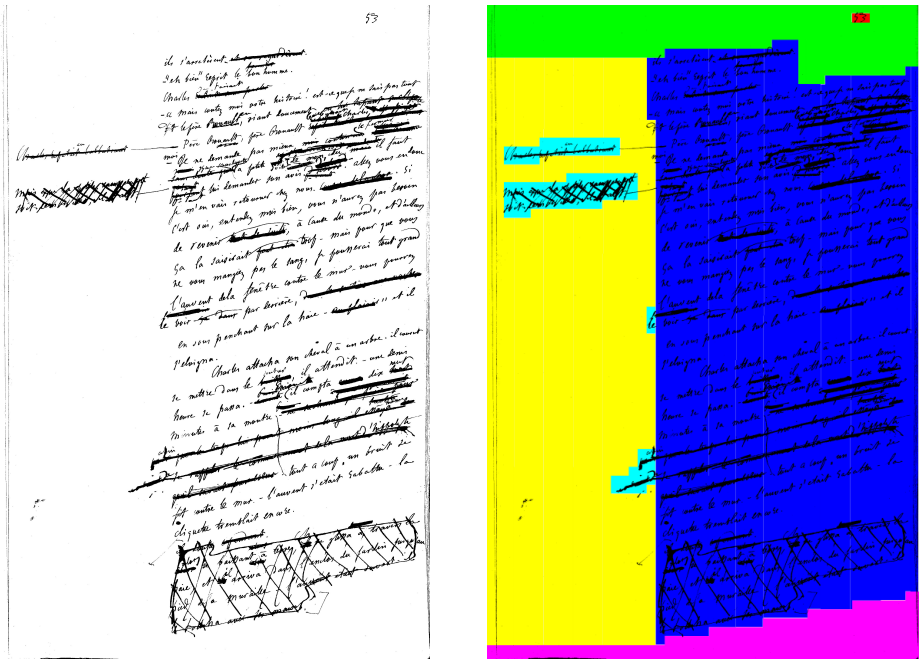
(a) (b)

(c)

Fig. 3.10. Manuscrits de la base "Bovary"

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. La première consiste donc 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 (grande maille d'analyse). Le modèle défini pour cette tâche d'étiquetage comporte les 6 états suivants : "corps de texte", "bloc de texte", "numéro de page", "marge", "haut de page", et "bas de page". Pour permettre l'évaluation sur cette tâche, les 69 images de la base ont été annotées manuellement à ce niveau en utilisant le modèle défini

précédemment, afin de définir la vérité-terrain associée (figure 3.11).



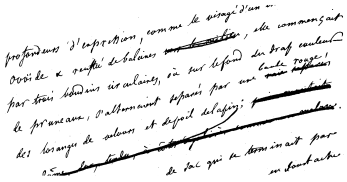
(a) (b)

Fig. 3.11. Etiquetage de zones d'intérêt au niveau bloc : (a) image bitonale d'un manuscrit de la base "Bovary" (b) image vérité terrain associée, manuellement étiquetée avec les conventions de couleurs suivantes : bleu = corps de texte, jaune = marge, cyan = bloc annotation, vert = en-tête, rose = pied de page, rouge = numéro de page

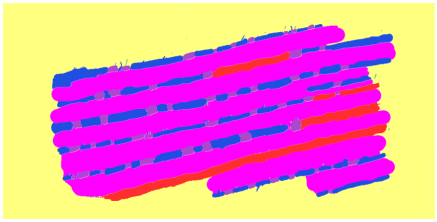
Etiquetage au niveau ligne

La seconde tâche que nous considérons est un étiquetage au niveau ligne. Avec cette seconde tâche nous souhaitons évaluer les capacités de notre approche à analyser l'image à un niveau plus fin, afin d'extraire les lignes de texte. Il s'agit d'étiqueter les entités constituant ces lignes, c'est à dire les composantes connexes correspondant à des mots ou des fragments de mots, des ratures, des symboles diacritiques et de ponctuation. Les espaces inter-mots et inter-lignes sont également considérés afin de mieux modéliser les lignes. Nous avons donc défini pour cette tâche un modèle avec les 6 états suivants : "pseudo-mot", "rature", "diacritique", "espace inter-mots", "espace inter-lignes" et "fond" (cf figure 3.12) La difficulté pour cette tâche d'étiquetage est de produire manuellement une vérité-terrain à un niveau aussi fin sur des documents très complexes comme les manuscrits de Flaubert où beaucoup d'entités sont connectées du fait de la forte densité

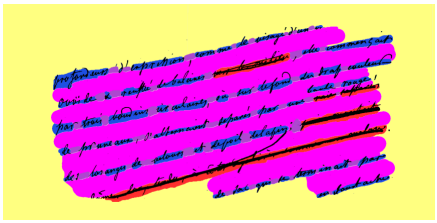
d'information et de la superposition des couches d'écriture. La séparation des entités connectées (ratures, mots) et l'étiquetage des espaces sont particulièrement difficiles à réaliser manuellement. Nous avons donc du nous limiter à l'étiquetage de quelques fragments d'images pour l'apprentissage des modèles uniquement. L'évaluation des résultats pour cette tâche se fera donc de manière qualitative par inspection visuelle.



(a)



(b)



(c)

Fig. 3.12. Exemple d'annotation manuelle de la vérité-terrain au niveau ligne : (a) image bitonale d'un fragment de manuscrit de la base "Bovary", (b) image vérité terrain associée manuellement étiquetée, (c) superposition de l'image et de la vérité terrain, avec les conventions de couleurs suivantes : bleu = pseudo-mot, rouge = rature, vert = diacritique ou ponctuation, violet = espace inter-mots, rose = espace inter-lignes, jaune = fond

3.4.1.2 Base "Imagination Poétique"

La base "Imagination Poétique" est une base que nous avons constituée à partir d'images de manuscrits de la Renaissance mises en ligne par le Centre

d'Etudes Supérieures de la Renaissance[2] (CESR) dans le cadre de son projet de bibliothèque numérique "Bibliothèques Virtuelles Humanistes"[3]. La base que nous avons constituée contient 46 images. Il s'agit à l'origine d'images couleur JPEG ayant une résolution de 96 dpi et de dimensions moyennes de 474×798. Ces images ont été converties en images TIFF niveau de gris, puis binarisées puisque nous ne considérons que des caractéristiques binaires. Cependant une fois de plus nous attirons l'attention du lecteur sur le fait qu'il serait possible d'utiliser la méthode d'analyse proposée, en utilisant des caractéristiques extraites sur les images couleur ou en niveaux de gris, voire même de combiner les caractéristiques. Nous ne l'avons pas fait pour des raisons de manque de temps.
Ces images ont ensuite été annotées manuellement pour produire la vérité terrain. Nous considérons pour cette base uniquement une tâche d'étiquetage des zones d'intérêt à un niveau bloc. En effet, il est intéressant sur de tels documents de pouvoir discriminer les différentes parties graphiques (lettrines, bandeaux, fleurons) des parties textuelles, et de localiser certaines parties textuelles, très importantes pour la création automatique d'index et de tables des illustrations par exemple, telles que les numéros de pages et les titres de chapitres. Nous avons donc défini l'alphabet de 7 étiquettes suivant : "fond", "corps de texte", "illustration", "titre", "sous-titre", "numéro de page", "fleuron" (cf figure 3.13). La détection de ces différentes entités informatives est importante pour l'indexation de grands corpus comme nous pouvons nous en rendre compte au travers de la bibliothèque virtuelle du CESR.

Comme cela a été effectué pour la base "Bovary", nous l'avons divisée en 3 sous-bases, une pour l'apprentissage (16 images), une pour la validation (15 images) et une pour les tests (15 images). Il est important de noter que l'ensemble de l'oeuvre a une structure relativement stable, et que les règles de mise en page sont relativement strictes. Cette mise en page est caractérisée par les règles suivantes : le titre est toujours en haut de la page et centré, suivi juste en dessous par le sous-titre. Il a toujours en-dessous du sous-titre une illustration suivie du texte lui-même. La fin du texte est délimitée par une ornementation appelée fleuron. Nous pouvons voir sur la figure 3.14 des exemples de l'annotation manuelle réalisée.

[2]http ://www.cesr.univ-tours.fr/
[3]http ://www.bvh.univ-tours.fr/

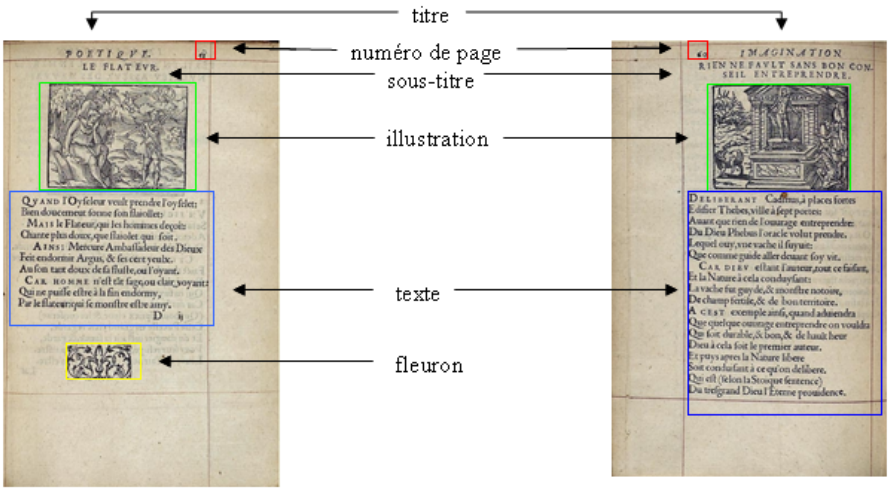


Fig. 3.13. Différentes zones d'intérêt dans des manuscrits de la Renaissance issus de l'oeuvre "Imagination Poétique"

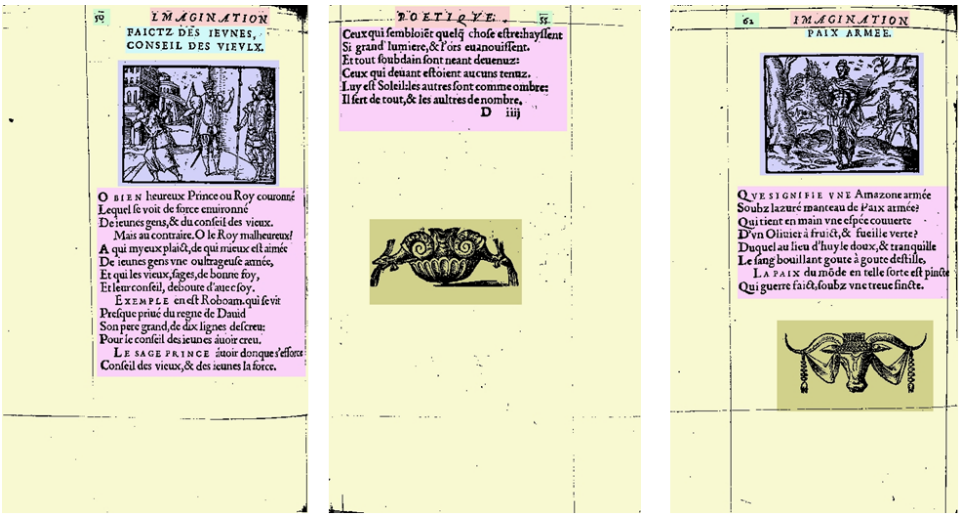


Fig. 3.14. Exemples de la vérité-terrain produite par étiquetage manuel, avec les conventions de couleur suivantes : jaune = fond, violet = illustration, beige = fleuron, rose = corps de texte, saumon = titre, cyan = sous-titre, vert = numéro de page

3.4.2 Critères d'évaluation

Dans l'approche que nous proposons nous cherchons à la fois à segmenter et à reconnaître les entités simultanément, en proposant un étiquetage de

l'image au niveau pixel. Cet étiquetage de l'image forme implicitement des régions connexes de pixels de même étiquette, que l'on souhaite les plus homogènes possibles et cohérentes vis à vis des observations effectuées sur l'image, qui correspondent aux entités de la page à un certain niveau d'analyse. Nous devons donc à la fois non seulement évaluer la qualité de l'étiquetage, mais également la qualité des régions formées par cet étiquetage, par rapport à une certaine vérité-terrain définie. Nous proposons donc pour évaluer notre méthode d'utiliser les deux catégories d'indicateurs, à savoir des indicateurs de reconnaissance au niveau pixel et des indicateurs de qualité au niveau des régions. Les critères d'évaluation que nous utilisons sont donc les suivants :

- **Critères d'évaluation au niveau pixel**

Dans la mesure où la méthode que nous proposons appréhende l'analyse d'images de documents comme un problème d'étiquetage d'image ou de classification de pixels, il apparaît assez naturel et évident d'évaluer les résultats en calculant des taux au niveau pixel. Dans [Baird 06b] où une approche d'analyse d'images de documents et d'extraction de contenu basée également sur un étiquetage de l'image au niveau pixel est présentée, l'évaluation s'effectue également au niveau pixel en considérant le taux d'étiquetage correct. Nous calculons deux taux. Le taux d'étiquetage correct global et le taux d'étiquetage correct normalisé de la manière suivante :

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

$$TEN = \frac{\sum_{i=1}^{q} \left(\frac{\text{nombre de pixels correctement étiquetés } l_i}{\text{nombre total de pixels de la classe } l_i} \right)}{q}$$

où q désigne le nombre de classes (ou étiquettes)

Le TEN est donc le taux moyen d'étiquetage correct par classe. L'utilisation de ce taux se justifie par le fait que les classes ne sont bien souvent pas réparties de manière homogène, si bien que les erreurs commises sur des classes peu représentées ne se détectent quasiment pas dans le taux global. Un autre outil permettant d'analyser efficacement les erreurs et les confusions entre les classes est la matrice de confusion. Nous utilisons donc les

matrices de confusion pour analyser plus en détails les erreurs commises par la méthode.
Il est à noter que pour le calcul de tous ces taux nous tenons compte de tous les pixels, qu'il s'agisse de pixels de fond ou de forme. Cependant les erreurs d'étiquetage commises sur les pixels de fond ont a priori moins d'importance que les erreurs commises sur les pixels de forme, car on cherche à segmenter et à reconnaître les formes. Il est donc également possible de ne tenir compte que des pixels forme dans le calcul de ces taux, sans tenir compte des pixels de fond. Dans ce cas les taux s'expriment de la manière suivante :

$$TEG_{forme} = \frac{\text{nombre de pixels forme correctement étiquetés}}{\text{nombre total de pixels forme}}$$

$$TEN_{forme} = \frac{\sum_{i=1}^{q} \left(\frac{\text{nombre de pixels forme correctement étiquetés } l_i}{\text{nombre total de pixels forme de la classe } l_i} \right)}{q}$$

- **Autres critères**

Un dernier critère numérique qu'il est selon nous important de prendre en compte, dans la mesure où nous nous plaçons dans un cadre de traitement en masse de documents, est le temps d'exécution de la méthode. Nous ne fournissons que le temps nécessaire au décodage de l'image. Le temps nécessaire à l'apprentissage du système est lui relativement plus long (de l'ordre de plusieurs minutes à quelques heures), mais cet apprentissage peut être réalisé hors-ligne, ce n'est donc pas un handicap trop important. Les temps sont exprimés en secondes.

Enfin l'évaluation des résultats peut s'effectuer également de manière complémentaire sur des critères qualitatifs par inspection visuelle de l'étiquetage obtenu, car les critères numériques ne rendent pas nécessairement compte de la qualité d'un résultat.

3.4.3 Evaluation et résultats

Nous nous proposons d'évaluer notre approche sur les deux bases que nous avons décrites précédemment en considérant les différentes tâches de segmentation et les critères d'évaluation que nous venons de définir. Nous

allons notamment étudier l'influence des quelques paramètres du modèle et de la méthode d'inférence utilisée.

Evaluation sur la Base "Bovary"

Nous évaluons d'abord la méthode sur la Base "Bovary" en considérant les deux tâches d'étiquetage définies : la tâche d'étiquetage au niveau bloc et la tâche d'étiquetage au niveau ligne.

Tâche d'étiquetage au niveau bloc

- Influence des paramètres du modèle d'attache aux données

Nous nous intéressons dans un premier temps à l'influence des paramètres du modèle d'attache aux données uniquement, sans prendre en compte le modèle contextuel de régularisation sur les étiquettes. Pour cette évaluation la taille du maillage a été fixée empiriquement. L'influence de la taille de la maille sera étudiée ensuite.

Les paramètres que nous considérons sont : le vecteur de caractéristiques intrinsèques utilisé, le nombre de composantes gaussiennes dans le modèle d'attache aux données, et la taille des bases d'apprentissage.

En ce qui concerne les caractéristiques intrinsèques, nous considérons 3 jeux différents :

– jeu 1 : un vecteur de 2 caractéristiques (caractéristiques de position)
– jeu 2 : un vecteur de 18 caractéristiques (caractéristiques de densité)
– jeu 3 : un vecteur de 20 caractéristiques (positions + densités)

L'apprentissage des mélanges de gaussiennes est réalisé en utilisant le critère de Rissanen qui permet de déterminer le nombre optimal de composantes gaussiennes. Cependant nous avons aussi réalisé des apprentissages en fixant le nombre de composantes afin de comparer les performances obtenues.

En ce qui concerne la base d'apprentissage comme nous pouvons le voir sur la figure 3.15, la répartition des classes n'est pas homogène. Nous avons donc réalisé différents apprentissages de mélanges, en considérant différentes réductions de la base d'apprentissage à un nombre maximal d'exemples par classe, de manière à homogénéiser un peu plus la répartition des classes. Nous considérons donc les trois bases d'apprentissage réduites suivantes, où les effectifs sont exprimés en nombre de sites :

- base 1 : 150 exemples maximum par classe et 900 exemples au total (cf figure 3.16(a))
- base 2 : 1000 exemples maximum par classe et 5186 exemples au total (cf figure 3.16(b))
- base 3 : 10000 exemples maximum par classe et 44085 exemples au total (cf figure 3.16(c))

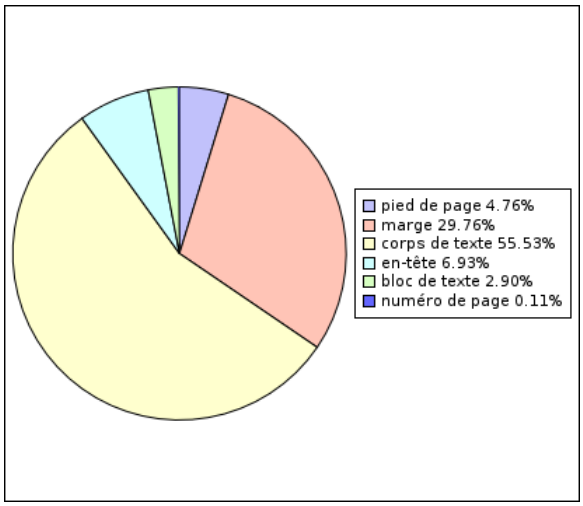


Fig. 3.15. Répartition des données dans la base d'apprentissage complète

Les graphes 3.17 et 3.18 illustrent les variations du taux global (TEG) et du taux moyen (TEN) d'étiquetage correct, en fonction de ces différents paramètres, alors que les graphes 3.19 et 3.20 illustrent la variation des taux d'étiquetage sur les pixels forme uniquement, en fonction des même paramètres.

Si on considère uniquement les taux d'étiquetage sur l'ensemble des pixels (TEG et TEN), les meilleurs résultats sont obtenus avec le jeu de caractéristiques n°3 qui intègre le plus de caractéristiques. Cependant l'analyse des

bas de page	150
marge	150
corps de texte	150
haut de page	150
bloc de texte	150
n° de page	150
	900

(a)

bas de page	1000
marge	1000
corps de texte	1000
haut de page	1000
bloc de texte	1000
n° de page	186
	5186

(b)

bas de page	7878
marge	10000
corps de texte	10000
haut de page	10000
bloc de texte	6021
n° de page	186
	44085

(c)

Fig. 3.16. Effectifs des bases d'apprentissage réduites : (a) base n°1, (b) base n°2, (c) base n°3

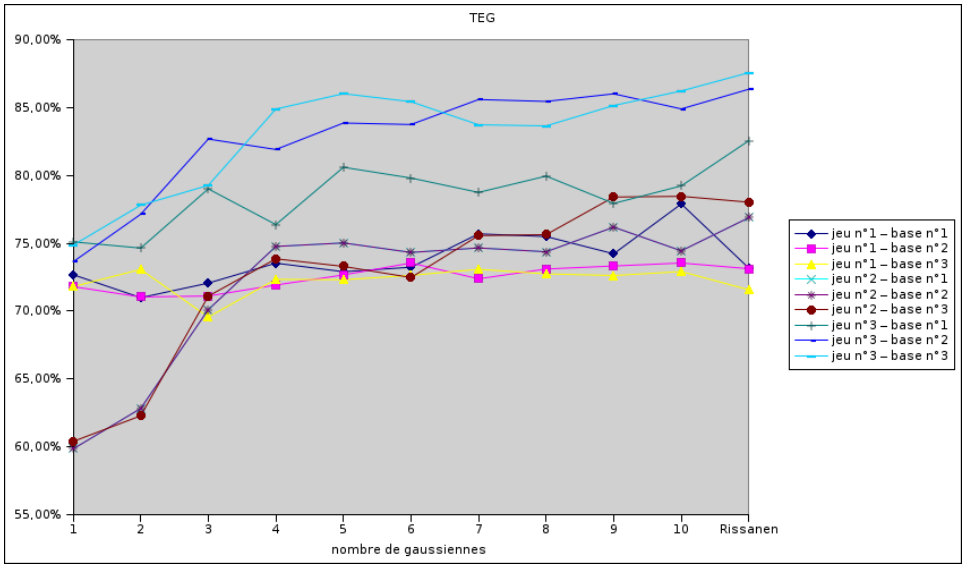


Fig. 3.17. Taux d'étiquetage correct (TEG) obtenus avec le modèle d'attache seul, sans décodage, en utilisant différents jeux de caractéristiques et des bases d'apprentissage de différentes tailles

courbes montre que sur cette base, et pour la tâche d'étiquetage considérée, les caractéristiques de position apporte beaucoup d'information, même quand sont utilisées seules (jeu n°1). Cela se vérifie d'ailleurs sur les courbes des taux d'étiquetage des pixels forme (TEG_{forme} et TEN_{forme}), puisque les meilleurs résultats sont obtenus avec le jeu n°1. Il apparaît donc que le jeu n°2 basé sur des caractéristiques de densité n'est pas très discriminant pour cette tâche d'annotation. Toutefois la combinaison des deux types de caractéristiques dans le jeu n°3 permet d'améliorer les taux d'étiquetage sur l'ensemble des pixels (fond et forme). Pour la suite des expérimentations nous utiliserons donc ce vecteur de caractéristiques.

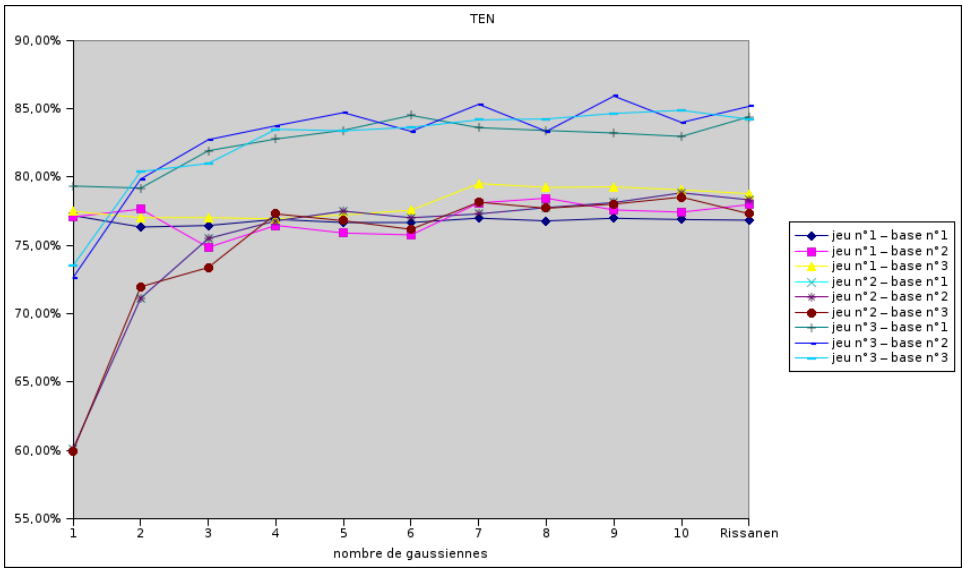


Fig. 3.18. Taux moyens d'étiquetage correct (TEN) obtenus avec le modèle d'attache seul, sans décodage, en utilisant différents jeux de caractéristiques et des bases d'apprentissage de différentes tailles

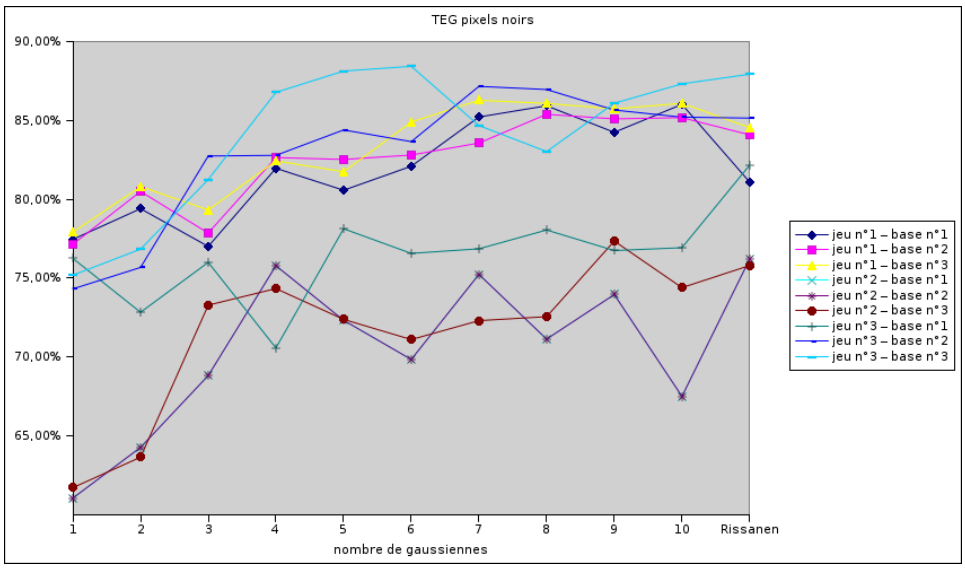


Fig. 3.19. Taux d'étiquetage correct des pixels forme (TEG_{forme}), obtenus avec le modèle d'attache seul, sans décodage, en utilisant différents jeux de caractéristiques et des bases d'apprentissage de différentes tailles

En ce qui concerne la taille des bases d'apprentissage, il est difficile d'après les résultats que nous avons de tirer des conclusions. On peut simple-

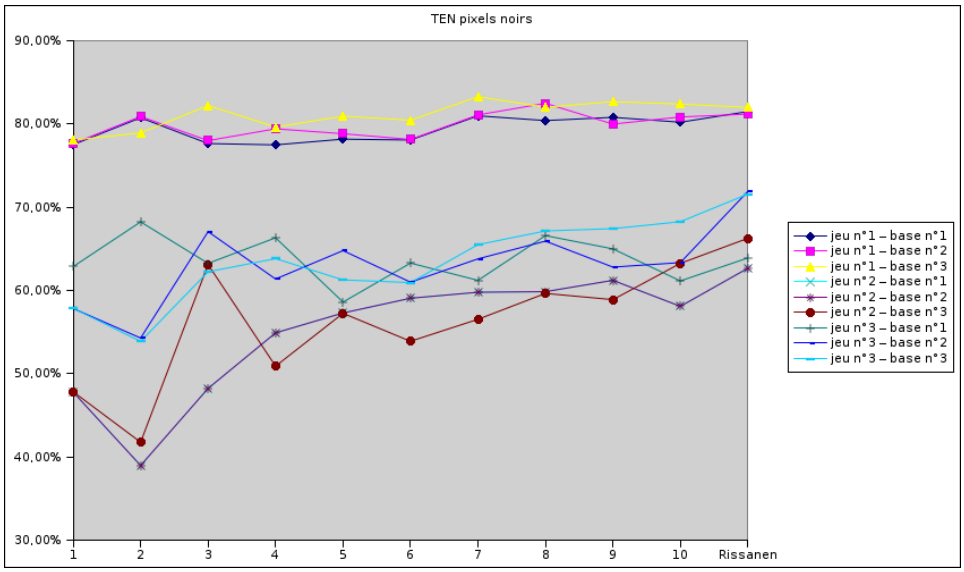


Fig. 3.20. Taux moyens d'étiquetage correct des pixels forme (TEN_{forme}), obtenus avec le modèle d'attache seul, sans décodage, en utilisant différents jeux de caractéristiques et des bases d'apprentissage de différentes tailles

ment remarquer que l'homogénéisation et la réduction du nombre d'exemples dans la base d'apprentissage à une influence seulement sur les taux d'étiquetage globaux, mais n'influence pas les taux moyens. Par contre le fait de réduire le nombre d'exemples dans la base d'apprentissage permet de réduire le temps nécessaire à l'apprentissage des paramètres des mélanges. Pour cette raison pour la suite des expérimentations les paramètres des mélanges seront déterminés sur la base d'apprentissage n°2 qui permet d'obtenir un bon compromis entre rapidité de l'apprentissage et qualité du modèle.

Enfin pour ce qui concerne le nombre de gaussiennes dans les mélanges, là non plus les résultats ne permettent pas vraiment de tirer des conclusions, puisque l'on peut voir que les taux varient beaucoup en fonction du nombre de gaussiennes. Sur les graphes, la valeur indiquée "Rissanen", est le nombre optimal de gaussiennes dans chacun des mélanges modélisant les classes, déterminé automatiquement par apprentissage en utilisant le critère de Rissanen. Avec ce critère le nombre de composantes gaussiennes est donc optimisé pour chacun des mélanges. Ce nombre de gaussiennes peut donc être différent d'un mélange à l'autre. Dans la suite des expérimentations nous utiliserons donc ce critère pour déterminer le nombre optimal de gaussiennes dans les mélanges.

- Influence des paramètres du modèle de régularisation

En ce qui concerne le modèle de régularisation, les potentiels sur les cliques sont déterminés par un apprentissage supervisé. Cependant pour définir un système de cliques il est nécessaire de définir un système de voisinage. Nous considérons au maximum des cliques d'ordre 2 sur un système de voisinage 4 ou 8-connexe. Afin de déterminer l'influence du système de voisinage sur le résultat, nous comparons les taux d'étiquetage obtenus avec les deux systèmes. Ces taux sont données dans le tableau 3.5 suivant :

	4-connexe	8-connexe
TEG (%)	85,8	84,1
TEN (%)	86,3	86,9
TEG_{forme} (%)	84,4	81,6
TEN_{forme} (%)	75,2	77,5

Tab. 3.1. Taux d'étiquetage obtenus en utilisant un système de voisinage 4-connexe et en utilisant un système 8-connexe

Les taux obtenus avec les deux systèmes de voisinage sont sensiblement identiques. Les taux moyens sont toutefois très légèrement supérieurs avec un voisinage 8-connexe. En effet, un tel système de voisinage permet a priori une meilleure régularisation en prenant en compte plus de sites voisins puisqu'il définit, en plus des cliques horizontales et des cliques verticales, des cliques diagonales. Pour la suite des expérimentations nous utiliserons donc un système de voisinage 8-connexe.

- Influence de la méthode de décodage

Nous comparons ensuite les taux moyens d'étiquetage obtenus avec les différentes méthodes de décodage suivantes : ICM, HCF et l'algorithme de programmation dynamique 2D de [Geoffrois 03]. Les taux moyens d'étiquetage obtenus avec les différents algorithmes de décodage sont donnés dans le tableau 3.2, et sont comparés avec le taux obtenu pour un étiquetage basé sur le modèle d'attache aux données (mélanges gaussiens) sans information contextuelle.

Nous constatons que sur cette base et pour cette tâche d'étiquetage,

	mélanges gaussiens	ICM	HCF	2D DP
TEN (%)	83,7	87,5	88,2	87,4

Tab. 3.2. Taux moyens d'étiquetage obtenus avec différents algorithmes de décodage comparativement au taux obtenu avec les mélanges gaussiens uniquement

les taux obtenus avec les trois algorithmes de décodage sont sensiblement identiques et nettement supérieurs aux taux obtenus avec un étiquetage basé uniquement sur l'information d'attache aux données. Ces résultats tendent donc à montrer que l'information contextuelle prise en compte dans le modèle de régularisation permet d'améliorer l'étiquetage. En ce qui concerne la méthode de décodage, ces résultats seuls ne permettent pas de tirer des conclusions sur la meilleure méthode à utiliser, même si la méthode HCF donne un taux moyen d'étiquetage très légèrement supérieur aux deux autres méthodes.

Si on considère les temps nécessaires au décodage, obtenus pour ces trois méthodes (tableau 3.3), on constate d'une part que pour les trois méthodes, ces temps sont relativement faibles (inférieurs à la seconde), et d'autre part que l'algorithme ICM semble être dans le cas présent le plus rapide. Cependant la différence entre ICM et HCF est une nouvelle fois minime. Bien évidemment ces temps de décodage dépendent du nombre de sites à étiqueter et donc de la taille de l'image et de la taille du maillage appliqué sur l'image. En adaptant correctement la taille de la maille à la taille de l'image et à la tâche d'étiquetage considérée, on voit donc que l'on peut obtenir des taux d'étiquetage corrects (presque 90% de pixels bien étiquetés par classe) en un temps relativement court, pour des images malgré tout de grande taille ($2500 \times 3600 pixels$).

	ICM	HCF	2D DP
temps (s)	0,21	0,29	0,61

Tab. 3.3. Temps nécessaire au décodage, par image, pour différents algorithmes

- Influence de la taille de la maille

La taille du maillage G appliqué sur l'image représente également l'un des paramètres réglables de notre approche. Nous avons donc cherché à étudier l'influence de ce paramètre sur le résultat de l'étiquetage. Nous avons donc déterminé les taux d'étiquetage correct obtenus pour différentes tailles de maille. Nous avons considéré des mailles de taille 20×20, 30×30, 40×40, 50×50. Les résultats obtenus sont illustrés sur le graphe (insérer numéro graphe).

	20×20	30×30	40×40	50×50
TEG	80,2	85,4	85,2	84,1
TEN	86,7	87,6	87,5	86,9
TEG_{forme}	82,3	85,3	82,6	81,7
TEN_{forme}	63,0	69,6	71,4	77,5

Tab. 3.4. Taux d'étiquetage obtenus sur la base "Bovary" avec différentes tailles du maillage

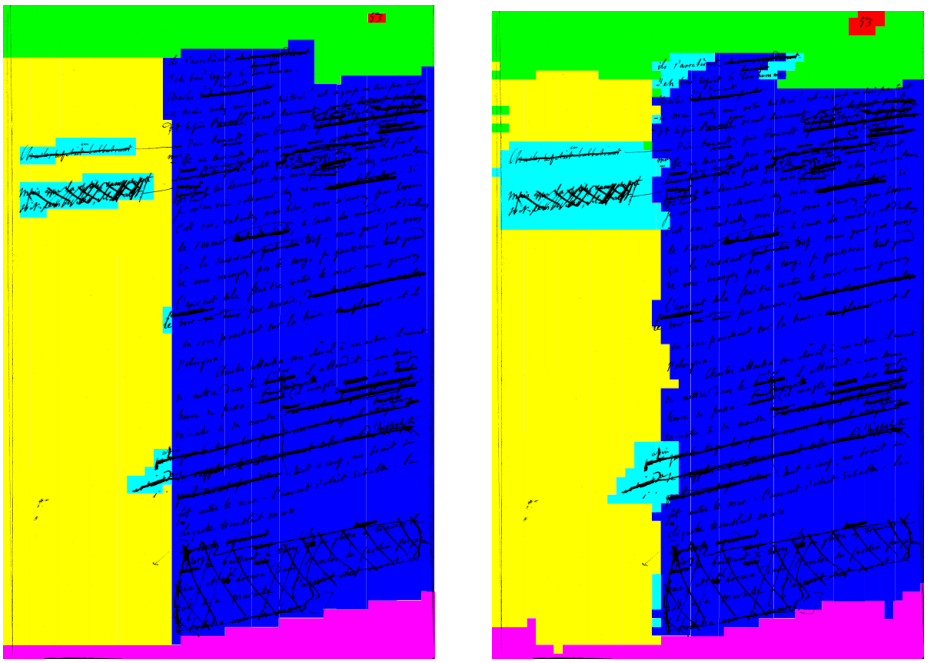
La taille du maillage G appliqué sur l'image peut représenter un paramètre critique de notre méthode. En effet, la taille de la maille doit être a priori adaptée à la tâche d'étiquetage considérée. Si la maille est trop grande l'étiquetage risque d'être grossier. Si la maille est au contraire trop petite le contexte pris en compte par le modèle de régularisation risque de ne pas être assez important pour permettre de bien étiqueter l'image.
Sur ces résultats on peut voir que la taille de la taille influe notamment sur le taux moyen d'étiquetage correct des pixels forme.

- Résultats qualitatifs

La figure 3.21, illustre un exemple de résultat d'étiquetage obtenu. On peut voir sur ce résultat que globalement la qualité est correcte, cependant, on peut voir qu'il existe un certain nombre d'erreurs d'étiquetage et de confusion notamment en ce qui concerne les entités de type "bloc de texte".

Tâche d'étiquetage au niveau ligne

Les manuscrits de Flaubert contiennent de nombreuses ratures et passages biffés, 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



(a) (b)

Fig. 3.21. Exemple de résultats de l'étiquetage de zones d'intérêt au niveau bloc : (a) vérité terrain (b) résultats de l'étiquetage 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

des mots, et d'extraire les lignes de texte, en utilisant une modélisation 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 unitaire 1×1. 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 3.22, on peut voir les résultats obtenus avec ce modèle sur un fragment de manuscrit. On peut constater que les mots, et les ratures sont correctement étiquetés. En ce qui concerne les marques de ponctuation et les symboles diacritiques, il y a quelques erreurs d'étiquetage, mais dans l'ensemble ils sont également correctement étiquetés. Ces erreurs sont dues au manque de données d'apprentissage, car

dans l'image d'un document, le nombre de sites appartenant à une entité "symbole diacritique" est nettement moins important que le nombre de sites appartenant à une entité "pseudo-mot" par exemple. On peut remarquer également que l'étiquetage produit en résultat par la méthode est beaucoup plus fin que l'annotation manuelle dans la vérité terrain. L'étiquetage résultat suit beaucoup plus finement les contours des mots et des ratures. Un autre point très positif de la méthode est sa capacité à séparer les différentes entités connectées du fait de chevauchement ou de superposition de l'écriture.

La figure 3.23(a) 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 peuvent 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 3.23(b), montre un résultat similaire obtenu sur un fragment contenant un mot et une rature connectés. 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 étiquetés (voir figure 3.22(b)).

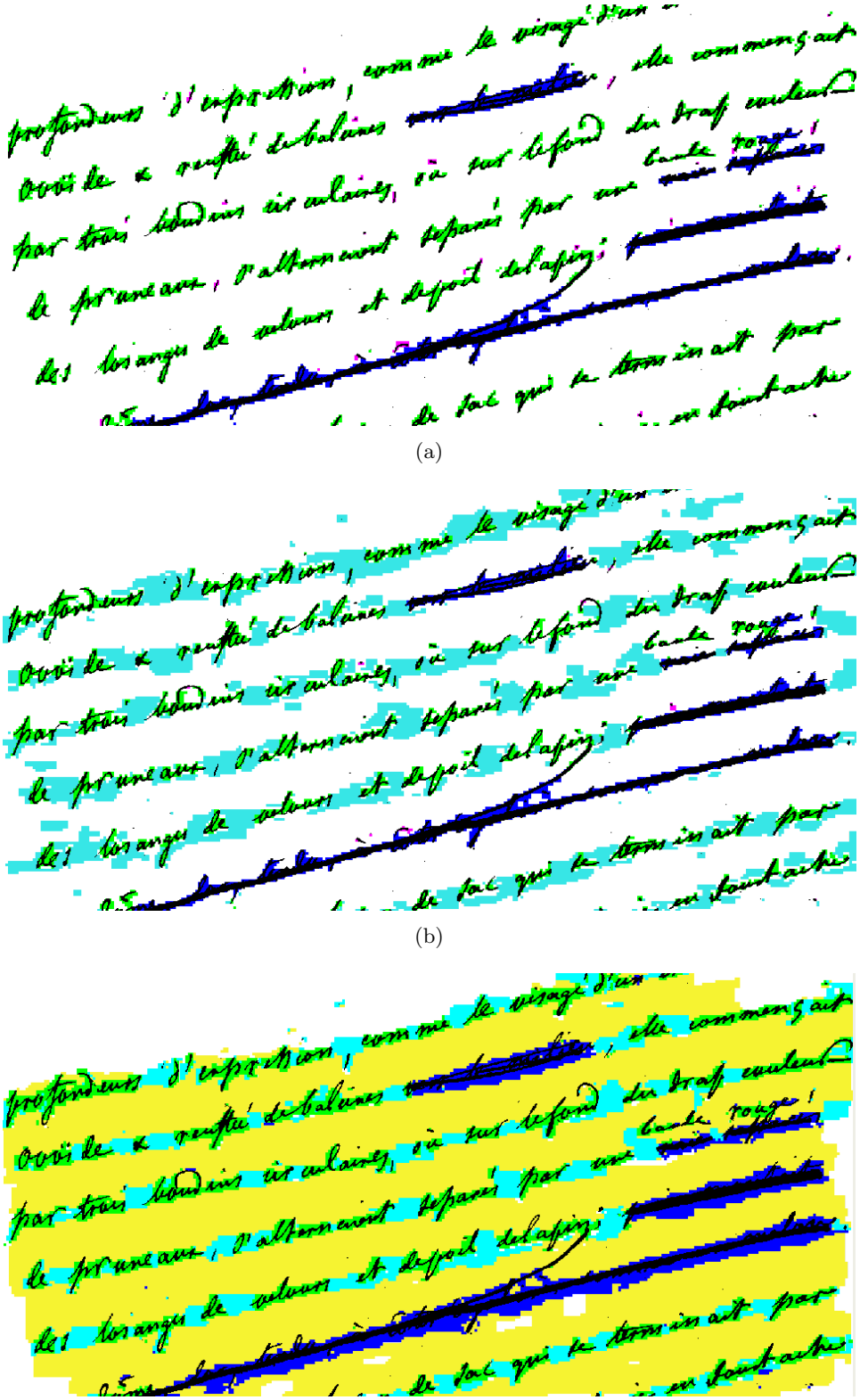
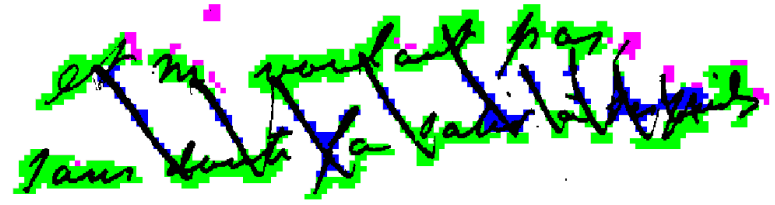


Fig. 3.22. résultats d'étiquetage au niveau "ligne" obtenus avec un modèle à 4 états (a), en ajoutant un état "espace inter-mots" supplémentaire (modèle 5-états) (b) et en ajoutant encore un état "interligne" (modèle 6-états) (c), avec les conventions de couleurs suivantes : blanc = fond, vert = pseudo-mot, bleu = rature, cyan = espace inter-mots, jaune = interligne, rose = diacritique



(a)

(b)

Fig. 3.23. résultats d'étiquetage à un niveau fin sur différents cas de connexion des composantes : (a) sur une zone de biffure où les couches d'écriture se superposent, (b) sur un cas de connexion entre un mot et une rature, avec les conventions de couleurs suivantes : blanc = fond, vert = pseudo-mot, bleu = rature, cyan = espace inter-mots, jaune = interligne, rose = diacritique

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 3.22(c) le type de résultat obtenu sur le même fragment que précédemment, et sur la figure 3.24 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 [Baird 06a] et [Baird 06b]), 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 est possible de segmenter ces entités de plus hauts niveaux, en utilisant des règles de fusion d'étiquettes, des techniques d'extraction de composantes connexes et des méthodes de

construction de graphes de régions.

Ne disposant pas de suffisamment de données étiquetées à un niveau aussi fin pour pouvoir faire une évaluation quantitative, nous nous limitons à cette étude qualitative. Cependant cette étude permet de montrer que la même méthode peut s'adapter facilement et à moindre coût à différentes tâches de segmentation à partir de quelques images étiquetées, par un apprentissage automatique. L'utilisateur n'a donc pas à paramétrer lui même le système ce qui permet une utilisation par des non-spécialistes du traitement d'image et de la reconnaissance de formes.

Evaluation sur la Base "Imagination Poétique"

Nous avons effectués sur la base "Imagination Poétique" les mêmes expérimentations que sur la base "Bovary" sur l'influence des paramètres du modèle. Nous ne considérons toutefois sur la base poétique qu'une tâche d'étiquetage au niveau bloc.

- Influence des paramètres du modèle d'attache aux données

Les différents taux d'étiquetage correct obtenus en faisant varier les paramètres du modèle d'attache aux données, sont présentés sur les graphes 3.25, 3.26, 3.27 et 3.28.

Contrairement aux résultats obtenus sur la base "Bovary", en ce qui concerne les jeux de caractéristiques il apparaît que les caractéristiques de densité apportent ici plus d'information que les caractéristiques de position. Cependant on observe une nouvelle fois qu'en combinant les deux types de caractéristiques (jeu n°3), on obtient les meilleurs résultats tant sur le taux global que sur le taux moyen.

En ce qui concerne le nombre de gaussiennes dans les mélanges, nous pouvons tirer les mêmes conclusions que sur la base Bovary, à savoir qu'il est difficile au vu des résultats obtenus de tirer une conclusion sur ce point, car ce paramètre dépend des données considérées. Nous utiliserons donc le critère de Rissanen pour déterminer automatiquement le nombre de composantes dans les mélanges.

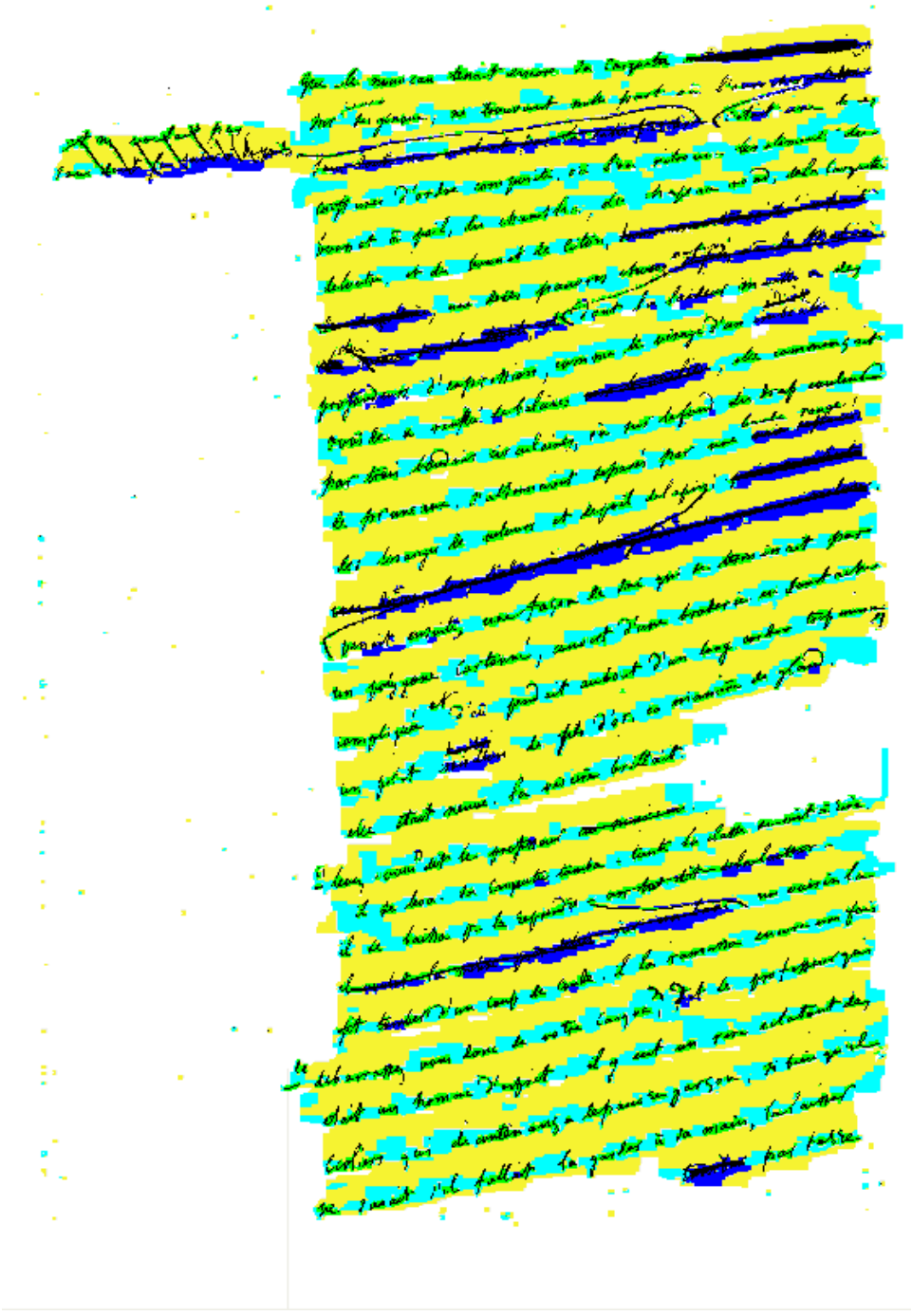


Fig. 3.24. Résultats d'étiquetage obtenus au niveau ligne avec le modèle à 6 états sur une page complète, avec les conventions de couleur suivante : blanc = fond, vert = pseudo-mot, bleu = rature, cyan = espace inter-mots, jaune = interligne, rose = diacritique

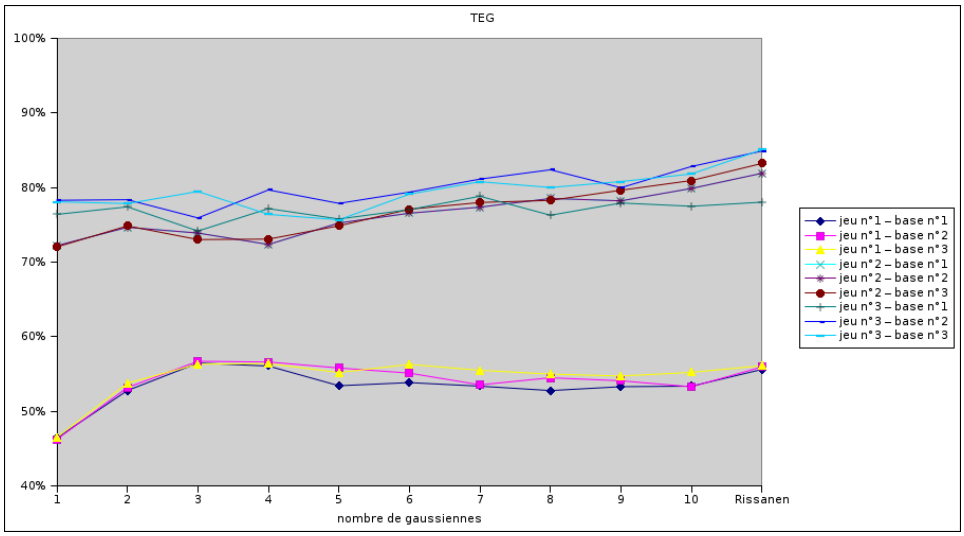


Fig. 3.25. TEG

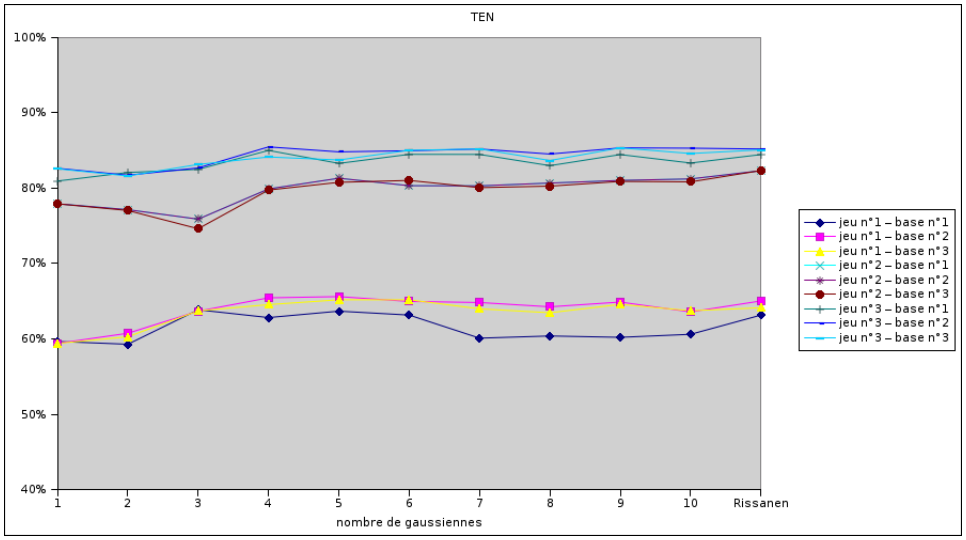


Fig. 3.26. TEN

Il est également difficile de tirer des conclusions sur l'influence des tailles des bases d'apprentissage. Il semble que ce paramètre n'a toutefois pas beaucoup d'influence sur les taux obtenus. Nous utiliserons pour la suite des expérimentations la base n°2.

Il apparaît donc que ce sont surtout les caractéristiques prisent en

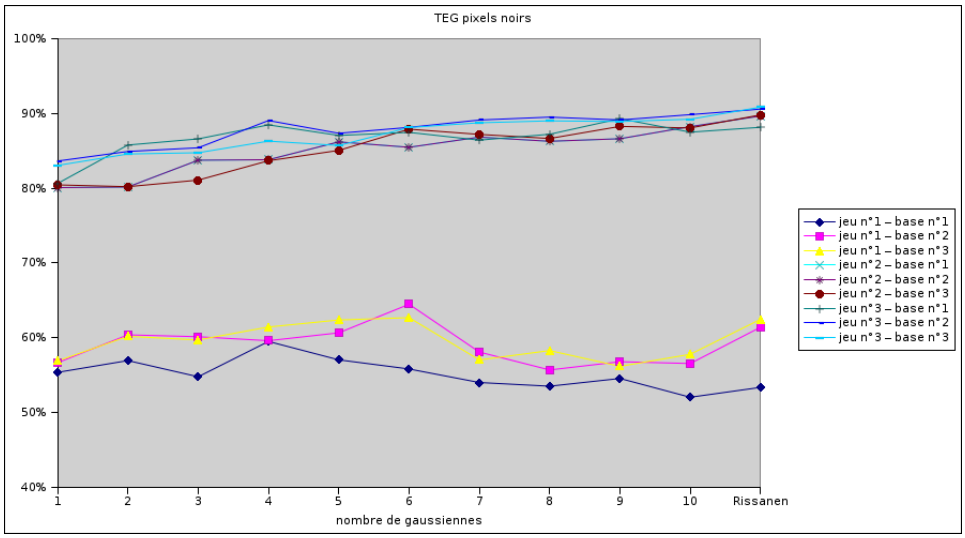


Fig. 3.27. TEG noir

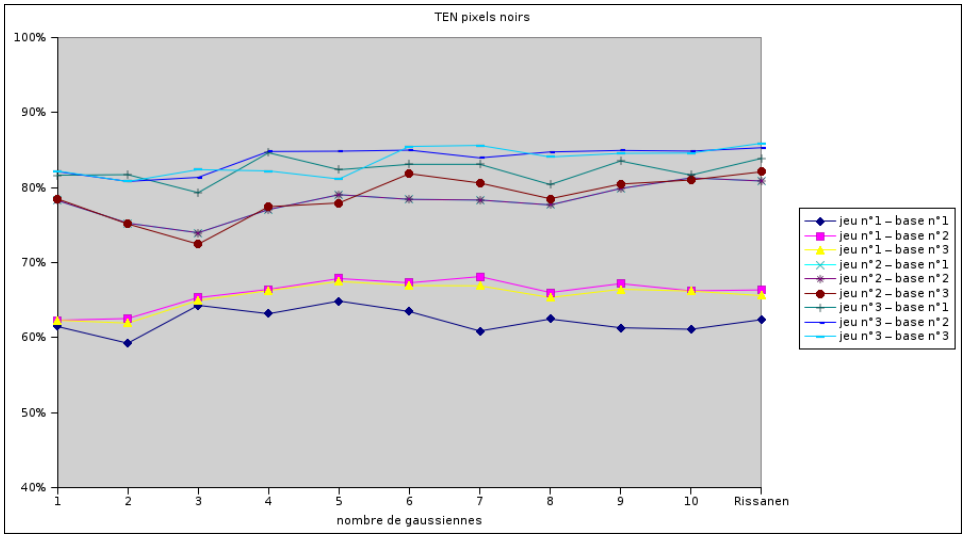


Fig. 3.28. TEN

compte qui influent sur le résultat produit par le modèle d'attache aux données.

- Influence des paramètres de régularisation

Comme sur la base Bovary, les résultats montrent que sur cette tâche

	4-connexe	8-connexe
TEG (%)	86,3	86,2
TEN (%)	75,9	75,7
TEG_{forme} (%)	91,6	91,8
TEN_{forme} (%)	75,0	74,5

Tab. 3.5. Taux d'étiquetage obtenus en utilisant un système de voisinage 4-connexe et en utilisant un système 8-connexe

d'étiquetage les deux systèmes de voisinages donnent des résultats sensiblement identiques.

- Influence de la taille de la maille

Nous avons déterminé les taux d'étiquetage obtenus avec les tailles de maille suivantes : 5×5, 10×10, 15×15, 20×20, 25×25, 30×30. Les mailles utilisées sont plus petites que sur la base "Bovary" car les images considérées sont également de dimensions plus faibles. Les taux obtenus sont donnés dans le tableau 3.6.

	5×5	10×10	15×15	20×20	25×25	30×30	35×35
TEG	84,6	83,3	84,8	86,2	86,1	84,1	85,8
TEN	89,3	75,4	76,7	75,7	73,3	75,0	75,7
TEG_{forme}	89,7	90,0	91,8	91,8	91,6	90,0	90,1
TEN_{forme}	85,8	73,3	74,7	74,5	73,5	75,4	76,7

Tab. 3.6. Taux d'étiquetage obtenus sur la base "Imagination Poétique" avec différentes tailles du maillage

Quelle que soit la taille de la maille les taux globaux obtenus sont sensiblement identiques. Le taux global d'étiquetage correct des pixels forme atteint les 90% pour les différentes tailles de maille. Cela signifie qu'une grande majorité des pixels formes sont correctement étiquetés. Par contre, si on considère les taux moyens (TEN et TEN_{forme}), les meilleurs résultats sont obtenus avec la maille 5×5. Il apparaît donc que la taille de la maille est un paramètre important de la méthode.En effet, elle conditionne la qualité de l'étiquetage produit.

Conclusion

Si on confronte les résultats des expérimentations sur les deux bases pour des tâches d'étiquetage assez similaire au niveau bloc, il apparaît que les paramètres qui influencent réellement les résultats obtenus, sont les caractéristiques utilisées et la taille de la maille. Nous n'avons pas réalisé dans les travaux que nous présentons d'apprentissage de la taille de la maille, mais c'est une perspective que nous envisageons, afin de permettre une meilleure adaptation du modèle à différentes tâches, sans avoir à définir manuellement ce paramètre, ce qui est toujours délicat. Le reste des paramètres étant déterminés automatiquement par apprentissage, la méthode est donc capable de s'adapter à différentes tâches d'analyse et à différents types de documents à partir simplement de quelques images étiquetées sans nécessiter le réglage de quelconques paramètres.

3.5 Conclusion

La théorie des champs de Markov fournit un cadre théorique robuste et efficace pour la segmentation d'images de documents. La modélisation par champ de Markov permet d'intégrer à la fois une certaine connaissance a priori sur le problème à traiter, et du contexte dans la prise de décision, cela dans un cadre probabiliste. De plus les caractéristiques locales et globales des champs de Markov, font que ce type de modélisation est bien adapté aux documents présentant une certaine variabilité, comme les documents manuscrits. Nous avons vu que dans le cadre des champs de Markov, le problème de la segmentation de l'image du document, est perçu comme un problème d'étiquetage de l'image. Contrairement à l'approche traditionnellement employée dans le domaine de l'analyse d'images de documents, qui consiste à dissocier l'étape d'extraction de la structure physique de celle d'extraction de la structure logique, l'approche que nous avons proposée, permet de combiner un étiquetage de l'image à un niveau physique et à un niveau logique. Cet étiquetage de l'image consiste en un problème de minimisation de la fonction d'énergie du champ Markovien associé à l'image. Nous avons proposé un étiquetage à un niveau image relativement fin afin de nous affranchir des problèmes de recouvrements et de connexités inhérents aux documents manuscrits et aux documents anciens, cependant la méthode peut s'adapter à une analyse à un niveau d'abstraction plus élevé. Les résultats que nous

avons obtenus avec cette méthode sont intéressants. Toutefois l'utilisation de champs de Markov pour la segmentation d'images de documents présente certains problèmes qui font que les résultats ne sont pas aussi bons que ce que l'on pourrait être en droit d'attendre. Tout d'abord les champs de Markov sont des modèles génératifs, c'est à dire qu'ils modélisent la manière dont les données sont générées, or le problème de la segmentation est un problème discriminant, il s'agit d'étiqueter les zones de l'image. Le modèle utilisé impose des hypothèses fortes d'indépendance des observations. Or dans le domaine de l'image ces hypothèses ne sont pas vérifiées, puisqu'il existe des dépendances statistiques entre les pixels de l'image, surtout entre pixels voisins. De plus cette hypothèse ne permet pas de prendre en compte des dépendances à plus ou moins long terme dans l'image. Pourtant il semble évident que de telles caractéristiques moins locales permettraient de segmenter plus efficacement l'image. En effet, on sait par exemple que les grands alignements de pixels blancs permettent de séparer les entités. La mesure de tels alignements ne peut pas s'effectuer à un niveau local. Pour toutes ces raisons nous nous sommes intéressés à l'utilisation de modèles markoviens discriminants pour la segmentation d'images de documents. Ces modèles sont appelés Champs Aléatoires Discrimants (ou Discriminative Random Fields en anglais), ou encore Champs Aléatoires Conditionnels (Conditional Random Fields). En effet ces modèles ne cherchent pas à représenter la probabilité jointe des observations et des étiquettes, mais modélisent directement la probabilité conditionnelle des étiquettes sachant les observations. Nous nous intéressons dans le chapitre suivant à l'étude de ces méthodes et à leur application à la problématique de la segmentation d'images de documents.

Chapitre 4

Vers une approche discriminante de la segmentation d'images de documents

4.1 Introduction

Les modèles génératifs vus dans le chapitre précédent présentent un certain nombre de limitations lorsqu'ils sont considérés dans le cadre d'une tâche d'étiquetage, qui est une tâche fondamentalement discriminante. La plus forte de ces limitations est l'hypothèse d'indépendance des données que ces modèles posent pour des raisons pratiques de réduction de la complexité combinatoire lors de l'apprentissage et de l'inférence. En effet, en pratique cette hypothèse d'indépendance est rarement vérifiée dans les données réelles considérées, notamment pour les images. Une autre limitation de ces modèles réside dans le fait qu'ils sont appris de manière non discriminante. Cela vient du fait qu'il s'agit de modèles génératifs qui définissent une loi de probabilité conjointe $P(X, Y)$ sur le processus observable Y et le processus label associé X [Do 06a], et qui modélisent ainsi le processus de génération des observations. Le problème de la segmentation et de l'étiquetage de données que nous considérons est clairement un problème discriminant, puisqu'il s'agit de discriminer des configurations d'étiquettes afin de déterminer la plus probable au regard du modèle en présence et des données

observées dont on dispose. Des approches et des modèles discriminants ont donc été proposés pour palier aux défauts des modèles de Markov cachés (chaînes en 1D et champs en 2D), traditionnellement et intensivement utilisés, dans de nombreux domaines, pour des tâches diverses d'extraction d'information, d'étiquetage, de segmentation et de reconnaissance. Ainsi, comme le rapportent Do et Artières dans [Do 06a], des méthodes ont été proposées pour introduire de l'information discriminante dans des systèmes basés sur des modèles génératifs. Ce sont principalement des approches reposant sur des méthodes à noyau et des machines à vecteur support [Do 05]. Cependant il ne s'agit pas de définir des modèles réellement discriminants, mais plutôt de faire de la sélection de modèles génératifs à l'aide d'approches discriminantes, et d'apprendre de manière discriminante les paramètres de ces modèles génératifs. On peut donc considérer que ces méthodes sont plutôt mixtes, génératives/discriminantes que complètement discriminantes. Des modèles réellement discriminants visant à palier l'ensemble des défauts des modèles génératifs ont donc été proposés ces dernières années. Ces sont des modèles conditionnels qui cherchent à modéliser la loi de probabilité $P(X|Y)$ des étiquettes conditionnellement aux observations, et qui peuvent être appris de manière discriminante. Ils modélisent donc ce qui différentie deux configurations d'étiquettes, étant donné un champ d'observations, plutôt que la manière dont les observations sont générées par une configuration d'étiquettes sous-jacente donnée. Ils sont donc a priori plus efficaces et plus adaptés à résoudre des tâches de segmentation/reconnaissance qui sont des tâches fondamentalement discriminantes.

4.2 Etat de l'art

Les modèles discriminants qui ont d'abord été proposés, sont les modèles de Markov à entropie maximale (Maximum Entropie Markov Models ou MEMM) [MacCallum 00] puis les champs aléatoires conditionnels [Lafferty 01]. Les modèles de Markov à entropie maximale (MEMM) sont des modèles conditionnels (ou discriminants), dont le modèle graphique associé est défini sur une structure de graphe orienté (figure 4.1(b)). Contrairement aux modèles de Markov cachés (figure 4.1(a)), les MEMM ne définissent pas de probabilités de transition et des probabilités d'émission

des observations mais des probabilités conditionnelles sachant toutes les observations et l'état à l'instant précédent $P(x_t|x_{t-1}, Y)$. L'avantage des MEMM est qu'ils ne nécessitent pas de poser d'hypothèses particulières sur les observations Y. Toutefois les MEMM présentent un problème soulevé par Laferty dans [Lafferty 01], et appelé "label bias". Ce problème est que dans un MEMM, les probabilités de transition sont normalisées ce qui fait que si la structure du modèle est telle qu'un état n'a qu'un successeur, l'observation Y n'a aucune influence sur le décodage. Pour régler ce problème des modèles de Markov à entropie maximale, tout en conservant leurs avantages sur les modèles génératifs, un autre type de modèle discriminant a été proposé. Il s'agit des champs aléatoires conditionnels (Conditional Random Fields ou CRF dans la littérature anglophone). Comme les HMM, les MEMM et les MRF, il s'agit de modèles graphiques probabilistes et markoviens, mais qui reposent d'une part sur une structure de graphe non orienté (figure 4.1(c)), comme les MRF, ce qui permet de s'affranchir du problème du "label bias", et d'autre part ils ne sont plus génératifs mais discriminants (comme les MEMM), ce qui permet de mieux modéliser des processus fondamentalement discriminants tels que l'annotation, et autorise un apprentissage dans un cadre réellement discriminant.

La figure 4.1 illustre les représentations graphiques des modèles de Markov cachés (HMM), des modèles de Markov à entropie maximale (MEMM), et des champs aléatoires conditionnels (CRF), d'après [Do 06a]. Les noeuds grisés correspondent aux variables observées du champ Y. On peut remarquer sur ces modèles graphiques que d'une part pour les MEMM et CRF chaque état X_t est potentiellement conditionné par l'ensemble des observations Y, car ces modèles ne requièrent pas d'hypothèses particulières sur Y (ce qui se traduit par les noeuds Y non décomposés dans les figures 4.1(b) et (c)). D'autre part on remarque que les HMM et les MEMM sont des modèles orientés dont les arcs expriment des dépendances causales entre les états (probabilités conditionnelle de transition), alors que les CRF sont des modèles non orientiés dont les arêtes traduisent des dépendances mutuelles entre les états.

Les champs aléatoires conditionnels ont d'abord été introduits par Lafferty et McCallum qui ont proposé des modèles CRF 1D pour la segmentation et l'annotation de séquences [Lafferty 01]. Ces modèles ont été

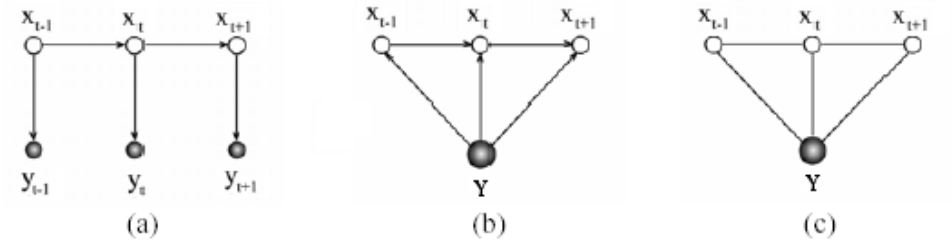


Fig. 4.1. Représentations graphiques d'un (a) modèle de Markov caché (HMM), (b) modèle de Markov à entropie maximale (MEMM) et (c) champ aléatoire conditionnel 1D (CRF 1D)

appliqués à différentes tâches dans le domaine du traitement des langages naturels (TALN), notamment à l'analyse de surface (shallow parsing) [Sha 03] et à l'étiquetage morpho-syntaxique de phrases (Part-of-Speech Tagging) [Lafferty 01]. Les CRF ont également été utilisés pour des tâches d'extraction d'information. Ainsi ils ont notamment été employés dans le domaine de la bio-informatique pour la recherche de séquences génomiques [McDonald] ou dans des tâches d'extraction d'information pour identifier des entités nommées relatives à des gènes , protéines ou molécules [Culotta 05]. Dans ces différents domaines ils ont montré leur supériorité par rapport aux modèles génératifs tels que les Modèles de Markov Cachés (MMC) classiquement utilisés auparavant pour ce genre de problème. Le modèle de champ conditionnel 1D proposé par Lafferty a ensuite été étendu à l'analyse des données bidimensionnelles avec la définition des champs aléatoires discriminants (Discriminative Random Fields) par Kumar et Hebert [Kumar 03b][Kumar 03a][Kumar 06]. Les auteurs ont utilisé ces modèles pour la détection de structures régulières, telles que des constructions ou des bâtiments, dans des images de scènes réelles. Dans [He 04] un modèle de CRF 2D basé sur la combinaison de réseaux neuronaux de type Perceptron Multicouches, est utilisé pour segmenter des images couleurs de scènes réelles, notamment de scènes animalières. Un modèle de CRF 2D est également proposé dans [Szummer 04], pour l'analyse et la reconnaissance de diagrammes manuscrits. Enfin, notons que les CRF 2D n'ont pas été utilisés uniquement dans le cadre de l'analyse d'image puisque d'une manière plus générale ils peuvent s'appliquer à l'analyse de données bidimensionnelles quelconques. Ainsi par exemple un modèle CRF causal pseudo-2D a également été proposés par Zhu et al. pour l'analyse et

l'extraction d'information dans les documents électroniques [J. Zhu 05].

Là encore, dans ces différents domaines, les modèles CRF 2D ont montré leur supériorité sur les modèles génératifs 2D tels que les champs de Markov. L'analyse de données et de signaux par champs aléatoires conditionnels connaît un intérêt croissant dans la communauté scientifique si on en juge par le nombre de travaux proposés sur cette thématique ces dernières années. Nous pouvons pourtant établir le même constat qu'avec les champs de Markov : malgré le succès des CRF dans de nombreux domaines, ils n'ont pratiquement pas encore été appliqués à l'analyse d'images de documents. A notre connaissance, les seules applications des CRF à l'analyse d'images de documents concernent les travaux sur la reconnaissance de documents graphiques manuscrits présentés dans [Szummer 04][Szummer 05][Qi 05] et les travaux de Feng et al. sur l'intérêt des modèles markoviens pour l'analyse de documents anciens [Feng 06]. Récemment des modèles CRF ont également été proposés et appliqués à la reconnaissance de l'écriture en-ligne [Do 06b].

Nous avons vu dans le chapitre précédent que la modélisation par champs de Markov peut apporter une solution souple et efficace à la problématique de l'analyse des documents anciens en vue de l'indexation de masses de documents complexes et hétérogènes. Par ailleurs les modèles CRF, en étant discriminants, ont montré leur supériorité sur les champs de Markov dans de nombreux domaines. Nous nous proposons donc naturellement dans ce chapitre de faire évoluer le modèle que nous avons présenté dans le chapitre précédent, vers un modèle discriminant s'inspirant de la théorie des champs aléatoires conditionnels et des modèles 2D proposés dans la littérature, en vue d'une application à l'analyse d'images de documents et à l'extraction de structures pour l'indexation.

4.2.1 Cadre théorique

Nous commençons dans cette section par poser le cadre théorique des champs aléatoires conditionnels en nous basant sur la description qui en est faite dans [Lafferty 01] et [Sutton 06].

Comme pour la modélisation par champs de Markov cachés on considère

que l'image donne accès à un ensemble de sites $S = \{s\}$ formant un réseau. On munit cet ensemble d'un système de voisinage noté V définissant les relations entre les sites. Selon ce système V, le voisinage V_s d'un site s est défini de la manière suivante :

$$V_s = \{t\} \text{ tels que } \begin{cases} s \notin V_s \\ t \in V_s \Rightarrow s \in V_t \end{cases} \quad \forall s, t \in S$$

Si deux sites s et t sont voisins selon ce système de voisinage V, on notera $s \sim t$

Selon la théorie des modèles graphiques, l'ensemble des sites S et le système de voisinage V associé, définissent de manière équivalente une structure de graphe $G = (N, E)$ telle que les noeuds N sont associés aux sites de l'image, et les arêtes E relient deux noeuds correspondant à deux sites voisins selon le système V. Ce graphe est appelé graphe d'indépendance.
On définit ensuite un champ aléatoire $X = (X_n)_{n \in N}$ dont les variables sont indexées par les noeuds du graphe G et un champ aléatoire noté Y. Le champ X est le champ caché des étiquettes, et Y le champ des observations.
Le couple (X, Y) est appelé champ aléatoire conditionnel, si les variables aléatoires X_n du champ X conditionnées par le champ Y obéissent aux propriétés de Markov vis à vis du graphe d'indépendance G, c'est à dire que :

$$P(X_n|Y, X^n) = P(X_n|Y, X_t, t \sim n)$$

Cela signifie qu'une variable aléatoire X_n du champ d'étiquettes X, dépend potentiellement de l'ensemble du champ observable Y et des variables X_t voisines selon le graphe d'indépendance G. D'un point de vue formel un champ conditionnel peut être vu comme un modèle graphique non orienté ou comme un champ de Markov globalement conditionné sur Y, le champ des observations.

Conditionnellement au champ Y des observations le champ X forme un champ de Markov, il est donc possible, d'après la théorie des champs markoviens, d'appliquer le théorème d'Hammersley-Clifford. La probabilité conditionnelle globale peut donc s'exprimer en fonction d'une fonction d'énergie U de la manière suivante :

$$P(X|Y) = \frac{1}{Z(Y)} \exp(-U(X|Y))$$

où Z est une constante de normalisation, appelée fonction de partition, qui ne dépend que de l'observation Y, et qui permet d'obtenir une probabilité. $U(X|Y)$ est l'énergie du champ X conditionné par le champ Y. Cette énergie est une fonction réelle à valeurs positives. Elle se définit comme une somme de fonctions de potentiel (ou potentiels) V_c calculées sur les cliques maximales c du graphe d'indépendance G :

$$P(X|Y) = \frac{1}{Z(Y)} \exp \sum_{c \in C} V_c(X|Y)$$

Les fonctions de potentiel sont des fonctions à valeurs réelles positives, et sont généralement définies sous la forme d'une combinaison linéaire de plusieurs fonctions de caractéristiques f_k évaluées sur les cliques maximales c du champ X et sur le champ Y :

$$V_c = \sum_k \lambda_k f_k(x_c, y, c)$$

où k désigne le nombre de fonctions de caractéristiques prises en compte dans la combinaison linéaire et les paramètres λ_k sont des pondérations associées. x_c représente la restriction de la configuration d'étiquettes sur le champ X, à la clique c.

La probabilité globale sur le champ X conditionnellement à la configuration sur le champ Y s'exprime alors ainsi :

$$P(X = x|Y = y) = \frac{1}{Z(Y = y)} \exp \left(\sum_{c \in C} \sum_k \lambda_k f_k(x_c, y, c) \right)$$

La constante de normalisation se calcule en sommant sur toutes les configurations d'étiquettes possibles :

$$Z(Y = y) = \sum_x \exp(\sum_{c \in C} \sum_k \lambda_k f_k(x_c, y, c))$$

En résumé un modèle CRF est donc défini par :

- un graphe d'indépendance qui définit les relations de dépendance entre les variables aléatoires du processus étiquette

- un ensemble de fonctions de caractéristiques f_k qui permettent de prendre en compte des connaissances sur le domaine

- un ensemble θ de coefficients de pondération des fonctions de caractéristiques $\theta = \{\lambda_k\}$

Dans le cas 1D, on considère des séquences d'observations. La structure du graphe d'indépendance est donc une chaîne linéaire et le modèle CRF défini sur cette structure est un modèle CRF 1D.

Dans le cas d'un modèle du premier ordre ne peut définir que des cliques d'ordre 1 (singleton) et des cliques d'ordre 2 (paires). Les cliques maximales que l'on peut considérer sont donc les paires de la forme (X_{i-1}, X_i) (cf figure 4.2) et les probabilités conditionnelles locales s'expriment en fonction de ces paires :

$$P(X_i|Y, X^i) = P(X_i|Y, X_{i-1}, X_{i+1})$$

Dans ce cas les caractéristiques sont donc de la forme : $f(x_{i-1}, x_i, y, i-1, i)$.

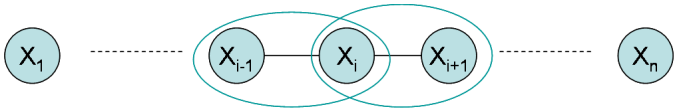


Fig. 4.2. cliques maximales d'une séquence pour un modèle 1D d'ordre 1

Dans le cas général les caractéristiques sont des fonctions à valeurs réelles positives qui s'appliquent sur les cliques maximales du graphe d'indépendance, et qui sont de la forme : $f(x_c, y, c)$.

La forme des caractéristiques utilisées dépend de l'application considérée. Ainsi par exemple en traitement automatique de la langue naturelle les caractéristiques considérés sont généralement des fonctions binaires qui testent la présence ou l'absence de certains mots et/ou de certaines structures syntaxiques. Dans le cas de l'analyse de signaux et de l'analyse d'image, il s'agit plus généralement de fonctions positives continues.

Comme pour les autres modèles probabilistes, qu'ils soient génératifs (modèles de Markov cachés, champs de Markov) ou discriminants (modèles de Markov d'entropie maximale), on considère trois tâches à résoudre avec

un champ aléatoire conditionnel :

- l'apprentissage des paramètres
- l'évaluation de la probabilité d'une réalisation du champ d'observations (reconnaissace)
- la recherche de la configuration d'étiquettes optimale étant donnée une réalisation du champ d'observations (segmentation/étiquetage)

Nous allons expliciter de quelles manières ces différents problèmes peuvent être résolu.

4.2.2 Apprentissage d'un modèle CRF

L'apprentissage d'un modèle CRF consiste à déterminer les paramètres du modèle, à savoir les caractéristiques à utiliser et leurs paramètres, ainsi que les pondérations associées à chacune de ces caractéristiques. L'apprentissage des caractéristiques dépend de la forme choisie pour ces caractéristiques. Dans le cas de caractéristiques binaires comme en TALN, l'apprentissage s'effectue généralement par sélection de caractéristiques. Dans d'autres travaux, comme nous allons le voir ensuite, les fonctions de caractéristiques sont continues et modélisées par des réseaux de neurones.
En ce qui concerne l'apprentissage des pondérations des caractéristiques, de nombreuses méthodes d'apprentissage existent. On suppose la structure du modèle fixée. Il peut s'agir d'une structure de chaîne, d'arbre, de grille régulière 2D ou d'une manière plus générale d'une structure de graphe quelconque. Cet apprentissage peut être réalisé de manière non supervisée comme cela est souvent le cas dans les problèmes de reconnaissance, mais pour des problèmes d'annotation il est généralement réalisé dans un cadre supervisé, à partir d'exemples étiquetés. Dans ce contexte il s'agit de déterminer les valeurs des paramètres qui maximisent la vraisemblance conditionnelle sur les données d'apprentissage. Ce problème ne peut pas être résolu de manière analytique, mais la vraisemblance étant une fonction convexe puisque le modèle est linéaire en fonction des paramètres, il peut être résolu de manière efficace par des méthodes d'optimisation basées soit sur le calcul de la dérivée de la fonction à optimiser (gradient), soit sur le calcul de sa dérivée seconde (hessien). Ces méthodes sont appelées respectivement, méthodes de gradient et méthode de Newton. Il s'agit dans les deux cas de méthodes itératives qui consistent à se déplacer le long de la courbe de la fonction

gradient en faisant varier les valeurs du vecteur de paramètres, de manière à trouver les valeurs qui permettent d'annuler le gradient, c'est à dire qui maximisent la vraisemblance. L'évolution le long du gradient s'effectue avec un pas fixe, il y a donc d'une part un risque de dépasser le point optimal, et d'autre part la convergence est relativement lente. L'utilisation de la dérivée seconde dans les méthodes de Newton permet de connaître le sens de variation de la dérivée, et d'adapter en conséquence la valeur du pas, de manière à converger très rapidement vers la solution en fixant un pas relativement grand au départ, puis à affiner de manière plus précise l'analyse, à mesure que l'on approche de la solution. Cependant le calcul de la dérivée seconde est relativement coûteux et complexe, et peut ne pas être possible. C'est pourquoi des méthodes basées sur l'approximation du hessien ont été proposées. Il s'agit des méthodes dites "quasi-Newton". Les méthodes de ce type sont très utilisées pour l'apprentissage des modèles CRF, en particulier la méthode L-BFGS qui est traditionnellement employée pour l'apprentissage des CRF 1D [Sha 03].

4.2.3 Etiquetage avec un modèle CRF

Le problème est sensiblement identique à celui de l'étiquetage par des champs de Markov cachés. Il s'agit de trouver l'étiquetage optimal au sens d'une certaine fonction de coût choisie. La différence est qu'un modèle CRF donne directement à la probabilité conditionnelle $P(Y|X)$ alors qu'avec un modèle MRF il faut passer par la règle de Bayes.
Comme nous l'avons dans le chapitre 3, il y a plusieurs estimateurs possibles, notamment le critère du Maximum A Posteriori (MAP) ou critère de Marginales Maximales (MPM). Les méthodes d'optimisation associées à ces estimateurs sont les mêmes que celles utilisées pour l'inférence dans des champs de Markov cachés. Il s'agit des méthodes de simulation basées sur la relaxation (recuit simulé, ICM) et des méthodes par passage de messages (algorithme de Viterbi en 1D, algorithmes Belief Propagation pour les arbres et Loopy Belief Propagation pour les graphes) également utilisées pour résoudre le problème de la marginalisation. Ces méthodes ont déjà été présentées dans le chapitre 3, nous ne reviendrons donc pas dessus.

4.2.4 Travaux existants sur l'application de modèles CRF 2D

Dans cette section nous passons en revue quelques exemples d'utilisation des CRF dans des cas bidimensionnels, notamment dans le domaine de l'analyse d'image. Nous nous attardons en particulier sur la forme générale des modèles proposés et sur les fonctions de caractéristiques utilisées, ainsi que sur la manière dont l'apprentissage et l'inférence sont réalisés avec ces modèles.

- Modèle de Kumar et Hebert [Kumar 03b]

Kumar et Hebert sont les premiers à avoir proposé une adaptation des modèles CRF introduits par Lafferty et al., à la segmentation et l'étiquetage de données bidimensionnelles, en l'occurrence, à la segmentation d'images. Le modèle qu'ils proposent se base directement sur le concept de champ aléatoire conditionnel de Lafferty et al., mais pour le différencier de celui-ci les auteurs l'ont baptisé de manière très générale, "Discriminative Random Field" (DRF), c'est à dire Champ Aléatoire Discriminant. En effet, il s'agit d'un modèle utilisé dans un cadre discriminant, comme les CRF, et non dans un cadre génératif comme les champs Markoviens classiques. Toutefois le modèle DRF proposé par Kumar et Hebert étend le modèle CRF de Lafferty et al. en proposant d'une part l'utilisation de modèles discriminants locaux pour modéliser l'attache aux données (association aux classes) en chaque site de l'image, et d'autre part en prenant en compte les interactions entre sites voisins sur une structure de grille régulière bidimensionnelle. Ainsi le modèle proposé permet de prendre en compte à la fois les interactions sur les données observées et sur les étiquettes cachées.

Forme du modèle

Le modèle CRF qu'ils proposent s'exprime donc sous la forme suivante :

$$P(x|y) = \frac{1}{Z}\left(\sum_{s\in S} A_s(x_s, y) + \sum_{s\in S}\sum_{t\in V_s} I_{st}(x_s, x_t, y)\right)$$

où A et I sont deux fonctions de caractéristiques. La fonction $A(x_s, y)$ est appelée potentiel d'association, car elle traduit l'attache aux données. Cette fonction est modélisée par un classifieur discriminant très simple, correspondant à une fonction logistique :

$$P(x_s = l \in \{-1, 1\}|y) = \frac{1}{1 + e^{-(\omega_0 + \omega_l^T f_s(y))}} = \sigma(\omega_0 + \omega_l^T f_i(y))$$

$I_{st}(x_s, x_t, y)$ désigne le potentiel d'interaction entre sites voisins. Ce potentiel prend la forme du modèle d'Ising auquel les auteurs ajoutent un second terme qui dépend de l'observation y. Ce second terme intervient également sous la forme d'une fonction discriminante logistique :

$$I_{st}(x_s, x_t, y) = \beta \left\{ K x_s x_t + (1 - K)(2\sigma(t_{st} \nu^T \mu_{st}(y)) - 1) \right\}$$

où K est la constante de couplage du modèle d'Ising, $\mu_{st}(y)$ est un vecteur de caractéristiques associé à la clique $x_s x_t$ et t_{st} désigne une caractéristique d'appariement des étiquettes voisines telle que :

$$t_{st} = 1 \ si \ x_s = xt$$

Apprentissage du modèle

L'apprentissage du modèle consiste à déterminer l'ensemble des paramètres $\theta = \omega, \nu, \beta, K$ qui comprend les paramètres des classifieurs logistiques, et les paramètres du modèle d'Ising. Cet apprentissage est réalisé de manière supervisée sur des données étiquetées. Il est réalisé par maximisation de la pseudo-vraisemblance conditionnelle, en utilisant une méthode de type "line search". Cette maximisation nécessite une bonne initialisation des paramètres, car la fonction de pseudo-vraisemblance n'est pas convexe, et il y a alors un risque de convergence vers un maximum local. Pour réaliser cette initialisation les paramètres ω et ν sont d'abord déterminés indépendamment en maximisant la log-vraisemblance par une méthode de Newton, c'est à dire une méthode basée sur le calcul de la dérivée seconde (hessien) de la log-vraisemblance.

Inférence

En ce qui concerne l'inférence les auteurs ont choisi d'utiliser l'algorithme des modes conditionnels itérés (ICM) pour déterminer une solution approchée de l'estimateur du Maximum a Posteriori de la Marginale

(MPM). Cet estimateur permet de miniser le nombre de sites mal classés, il favorise donc les bonnes décisions locales. L'algorithme ICM a été chosi par les auteurs pour sa simplicité et sa rapidité.

Evaluation du modèle

Les auteurs ont testé le modèle proposé sur un problème de détection de structures régulières correspondant à des bâtiments, dans des images naturelles. Il s'agit d'un problème d'étiquetage à deux classes. Les images considérées sont des images en niveaux de gris, issues de la base Corel[1] et de dimensions 256×384 pixels, ont été divisées en deux ensembles, l'un de 108 images pour l'apprentissage des paramètres du modèle, et l'autre de 129 images pour la réalisation des tests. Afin d'accélérer les traitements, un maillage régulier est appliqué sur les images et ainsi chaque image est divisée en blocs de 16×16 pixels. Chaque bloc correspond alors à un site de l'image. Pour chaque image de la base, la vérité-terrain associée a été produite en annotant manuellement les sites de l'image selon les deux classes suivantes : "structuré" ou "non structuré". Le modèle DRF a été évalué sur cette base et les résultats ont été comparés avec ceux obtenus avec un modèle de champ de Markov (MRF) et ceux obtenus avec le classifieur logistique de la fonction d'association appliqué seul en chaque site. Les critères d'évaluation choisis sont le taux de détection (nombre de sites correctement étiquetés "structuré") et le nombre de fausses détections (nombre de sites incorrectement étiquetés "structuré"). Les résultats montrent que pour un nombre de fausses alarmes identique à ce que fournit un classifieur linéaire logistique, le modèle DRF donne un meilleur taux de détection. La fausse alarme est également moindre qu'avec un modèle MRF, et le taux de détection est largement supérieur comme on peut le voir dans le tableau 4.3.

Méthode	Fausse alarme (par image)	Taux détection (%)
Classifieur linéaire	1.37	55.4
MRF	2.36	57.2
DRF	1.37	70.5

Fig. 4.3. Résultats de la détection de structures avec différentes techniques selon [Kumar 03b]

[1] http ://www.cs.cmu.edu/ skumar/manMadeData.tar

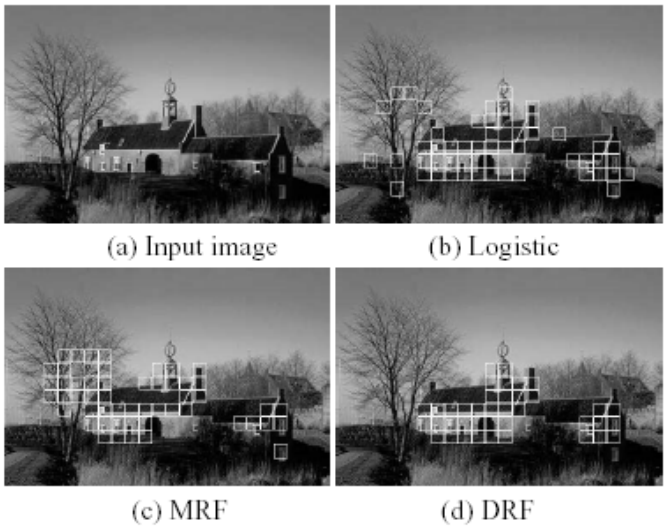


Fig. 4.4. Résultats de la détection de structures avec différentes techniques selon [Kumar 03b]

Ces résultats montrent, comme on peut également le vérifier visuellement sur la figure 4.4, que l'approche par champs de Markov (MRF) a tendance à homogénéiser à outrance sans toujours prendre en compte le modèle d'attache aux données, ce qui explique le nombre de fausses alarmes plus élevé et le taux de détection plus faible que pour les autres approches. A l'inverse le classifieur linéaire ne permet pas une homogénéisation car il repose sur une décision locale et ne prend pas en compte le contexte sur les sites voisins. Le modèle DRF quant à lui représente un bon compromis entre les deux, et permet ainsi une certaine homogénéisation sans toutefois négliger la composante d'attache aux données. En effet la modélisation par champs aléatoires conditionnels permet de combiner des classifieurs locaux discriminants, ce qui autorise l'extraction de caractéristiques quelconques sur l'image, même si celles-ci ne sont pas indépendantes, tout en permettant néanmoins une régularisation du champ d'étiquettes.

- Modèle de He et al. [He 04]

En 2004, He et al. ont également proposé une extension du modèle de champ aléatoire conditionnel de Lafferty, au cas bidimensionnel, pour la segmentation et l'étiquetage d'images. Comme pour les travaux de Kumar et al. il s'agit là également d'images naturelles, toutefois ce sont cette fois des images couleurs. De plus, le problème considéré n'est plus simplement

un problème binaire, mais un problème à plusieurs classes. En l'occurrence il s'agit de l'étiquetage d'images de scènes animalières ou de scènes routières. La principale différence avec le modèle proposé par Kumar et al. réside dans le fait que les classifieurs linéaires sont remplacés par des Perceptrons Multicouches (PMC) et que le modèle intègre différents niveaux de contexte sur le champ des étiquettes, ce qui permet une meilleure régularisation. Le modèle proposé est donc un modèle CRF multi-échelle qui définit une distribution conditionnelle sur le champ d'étiquettes X étant donnée une obervation Y, par un produit de distributions conditionnelles locales et contextuelles, qui capturent l'information à différentes échelles notées n. La distribution conditionnelle globale s'exprime donc de la manière suivante :

$$P(X|Y) = \frac{1}{Z} \prod_n P_n(X|Y)$$

où n désigne donc les différentes échelles d'analyse, et $P_n(X|Y)$ la distribution conditionnelle du champ X sachant le champ Y sur le niveau n.

Trois échelles différentes sont utilisées par les auteurs. Dans le contexte discriminant des CRF cela revient à combiner trois types de classifieurs : un classifieur local, un classifieur régional et un classifieur global.

Le classifieur local est un perceptron multi-couches à partir duquel la probabilité conditionnelle d'une configuration du champ X peut être calculée comme un simple produit des probabilités a posteriori locales en chaque site s :

$$P_{(L)}(X|Y, \lambda) = \prod_{s \in S} P_{(L)}(x_s|y_s, \lambda)$$

où λ désigne les paramètres du classifieur. Il s'agit d'un perceptron à une seule cachée, avec une fonction d'activation sigmoïde pour les neurones des couches cachées, et une fonction d'activation softmax pour les neurones de la couche de sortie. La couche de sortie comporte autant de neurones qu'il y a d'étiquettes définies dans l'alphabet. Des caractéristiques statistiques sont extraites sur une fenêtre d'analyse 3×3 centrée en chaque site. Ce sont ces caractéristiques qui sont appliquées en entrée du classifieur local. Ce classifieur détermine donc en chaque site les scores d'association du site considéré aux différentes étiquettes, en se basant uniquement sur des caractéristiques

intrinsèques locales extraites de l'image.

Le modèle régional permet de calculer en chaque région de l'image, la probabilité de l'étiquette sachant un ensemble de caractéristiques régionales $f_{r,a}$ détectés dans la configuration d'étiquettes de cette région. Ces caractéristiques régionales sont des détecteurs de configurations particulières comme par exemple une jonction, un bord, mais ces caractéristiques peuvent également concerner d'autres règles d'associations concernant les étiquettes des objets présents dans l'image : par exemple le ciel en-dessous de la mer est une configuration impossible tandis que l'inverse est une configuration toujours possible. Finalement la probabilité d'une configuration d'étiquettes x sur l'ensemble de l'image avec le modèle régionale est donnée par la relation ci-dessous où x_r désigne le vecteur des différentes configurations d'étiquettes locale dans la région r :

$$P_{\mathcal{R}}(X) \propto \prod_{r,a}[1 + \exp(\omega_a^T l_r)]$$

Le modèle global prend la même forme que le modèle régional mais avec un ensemble différent de caractéristiques extraites sur l'ensemble du champ d'étiquettes X :

$$P_{\mathcal{G}}(X) \propto \prod_{b}[1 + \exp(\nu_b^T X)]$$

Finalement, la combinaison des trois modèles fournit un modèle multi-résolution en faisant le produit des trois niveaux de modélisation :

$$P(X|Y,\theta) = \frac{1}{Z}\prod_{s} P_{\mathcal{L}}(x_s|y_s,\lambda) \times \prod_{r,a}[1 + \exp(\omega_a^T l_r)] \times \prod_{b}[1 + \exp(\nu_b^T X)]$$

Inférence Les auteurs ont choisi d'utiliser l'estimateur du Maximum a Posteriori de la Marginale (MPM) qui consiste à déterminer l'étiquette optimale en chaque site, c'est à dire à maximiser les marginales :

$$x_s^* = \arg\max_{x_s} P(x_s|Y),\ \forall s \in S$$

La maximisation des marginales est obtenue par tirage successifs de réalisations du champ, selon la loi conditionnelle, à l'aide de l'échantillonneur de Gibbs.

Apprentissage du modèle

L'apprentissage est réalisé de manière supervisée à partir d'images étiquetées, en utilisant un algorithme approché, appelé Contrastive Divergence (CD) [Carreira-Perpignan 05], qui permet d'éviter le calcul exacte des estimations en fonction des paramètres du modèle. Cet apprentissage est réalisé séquentiellement, c'est à dire que le classifieur local est d'abord entraîné séparément, puis ensuite les classifieurs contextuelles, une fois les paramètres du classifieur local fixés.

Evaluation du modèle

Les auteurs ont appliqué le modèle proposé à la segmentation et l'étiquetage d'images de scènes réelles et l'ont testé sur deux bases d'images. La première est constituée de 100 images de scènes naturelles de dimensions 180×120 pixels, issues de la base Corel et réparties aléatoirement en une base d'apprentissage de 60 images et une base de test de 40 images. La seconde base utilisée est constituée de 104 images de scènes extérieures de dimensions 96×64, issues de la base Sowerby [Collins 99] et réparties aléatoirement en une base d'apprentissage de 60 images et une base de test de 44 images. Les résultats obtenus (cf figure 4.5) sont comparés à ceux obtenus avec un simple classifieur non contextuel et un modèle génératif de champ Markovien. Ces résultats montrent que l'approche conditionnelle fournit une meilleure segmentation des images, puisqu'ils obtiennent sur la base Corel un taux d'étiquetage correct de 80.0% avec leur modèle, contre $66,9\%$ avec le classifieur non contextuel et 66.2% avec un modèle de champ de Markov. Sur la base Sowerby un taux de 89.5% est obtenu, contre 82.4% et 81.8% respectivement avec le classifieur et le modèle de champ markovien. Ces résultats montrent donc que d'une part l'approche discriminante est supérieure à l'approche générative (classifieur et CRF versus MRF), et que d'autre part l'approche contextuelle est supérieure à l'approche locale (CRF versus classifieur), pour des tâches de segmentation/reconnaissace. La difficulté réside toutefois dans la complexité des modèles conditionnels. Cette complexité se présente surtout lors de la phase d'apprentissage du modèle, car il n'existe pas de méthode de résolution exacte.

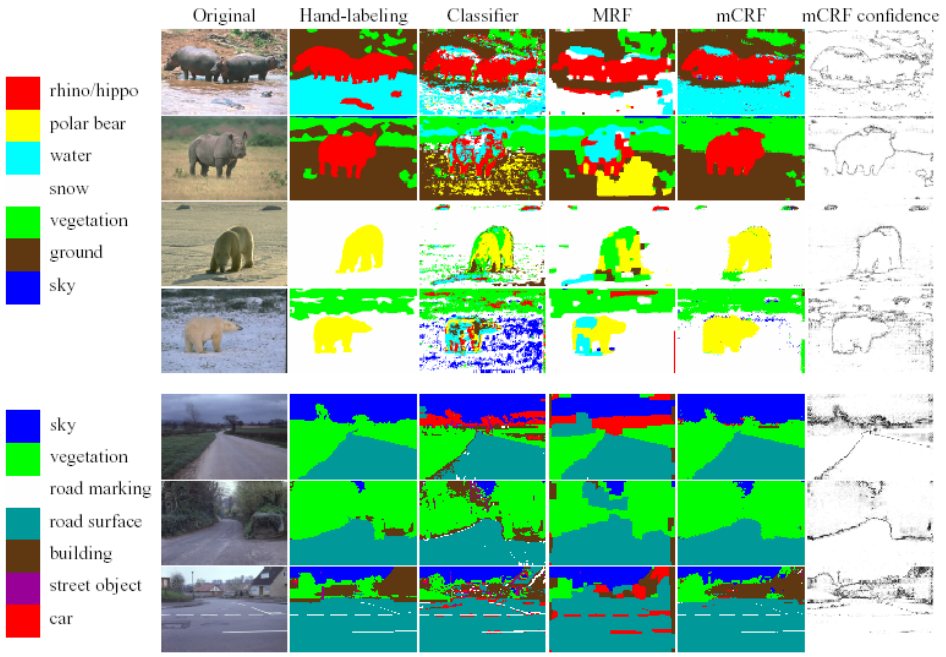


Fig. 4.5. Résultats de segmentation d'images de scènes avec le modèle CRF 2D de He et al. [He 04]

L'approche proposée est intéressante, car elle concerne une tâche d'étiquetage d'image, de plus le modèle proposé est un modèle conditionnel réellement 2D et multiéchelle. Le modèle proposé intègre des caractéristiques contextuelles dans un cadre probabiliste, qui combine les décisions de plusieurs composants. Chaque composant diffère dans l'information qu'il intègre. Certains se focalisent sur l'association entre l'image et le champ d'étiquettes, alors que d'autres se concentrent sur la détection de certains motifs dans le champ d'étiquettes. Ces composants diffèrent aussi sur le niveau d'analyse mis en oeuvre, puisque certains effectuent leurs analyses à un niveau d'analyse très fin alors que d'autres analysent la structure à un niveau plus global. Cette aspect multirésolution de l'analyse nous semble très important pour l'analyse d'image.

- Autres travaux sur les CRF 2D

D'autres travaux ont été réalisés sur la modélisation par CRF 2D, mais pour des problèmes autres que la segmentation d'image, telle que la reconnaissance ou l'extraction d'information dans des données bidimensionnelles.

Ainsi dans [Szummer 04], un modèle CRF 2D est proposé pour la reconnaissance de diagrammes manuscrits, dans [Do 06b] des modèles CRF sont utilisés pour la reconnaissance de l'écriture en-ligne, ou encore dans [J. Zhu 05] pour des tâches d'extraction d'information dans des pages Web.

4.2.5 Outils logiciels existants pour l'étiquetage à l'aide de CRF

Pour conclure cette étude théorique et bibliographique des modèles conditionnels, notons qu'il n'existe pas à l'heure actuelle, et à notre connaissance, d'outils logiciels, commerciaux ou libres, permettant l'analyse et l'étiquetage de données bidimensionnelles, à l'aide de modèles conditionnels. Il existe un certain nombre d'outils et codes sources pour l'étiquetage de séquences à l'aide de modèles conditionnels 1D, notamment pour des tâches de traitement automatique de la langue (TAL) telles que l'étiquetage morpho-syntaxique (part-of-speech tagging ou POS), mais rien pour le traitemenent et l'analyse d'image. Nous pouvons notamment citer le package CRF de Sunita Saragawi [2] qui est une implémentation en langage JAVA de champs aléatoires conditionnels pour l'étiquetage de données séquentielles, ou encore le package FlexCRF (Flexible Conditional Random Fields) [3] qui est une implémentation en langages C et C++, également pour l'étiquetage de données séquentielles et qui implémente la méthode L-BFGS pour l'apprentissage des modèles. Cet outil permet de traiter des centaines de milliers de séquences et peut prendre en compte des millions de caractéristiques. Toutefois il est dédié au traitement de données textuelles séquentielles, et les caractéristiques intégrées sont donc de ce fait des caractéristiques binaires. Le package MALLET [4] qui est une collection de code JAVA pour le traitement de données textuelles, intègre également une implémentation de modèles de champs conditionnels linéaires, et l'algorithme d'apprentissage L-BFGS. Là encore les outils proposés sont dédiés à l'analyse de données séquentielles et ne peuvent être appliqués à l'analyse de données bidimensionnelles, et en particulier à l'analyse d'image. Il en est de même du code CRF++ fourni par Taku Kudo [5]. Le seul code fournissant une implémentation de CRF 2D,

[2] http ://crf.sourceforge.net/
[3] http ://www.jaist.ac.jp/ hieuxuan/flexcrfs/flexcrfs.html
[4] *http : //mallet.cs.umass.edu/index.php/Main*$_{PAGE}$
[5] http ://www.chasen.org/ taku/software/CRF++/

est la toolbox Matlab CRF fournie par Kevin Murphy. Cette toolbox permet l'apprentissage et l'inférence de modèles CRF sur des grilles régulières 2D. Elle intègre une implémentation en langage C de l'algorithme Loopy Belief Propagation, ainsi que des modules d'extraction de caractéristiques et l'implémentation de la méthode d'optimisation Stochastic MetaDescent pour l'apprentissage. Les résultats publiés dans [Vishwanathan 06], ont été obtenus avec cette implémentation. Cependant cette toolbox, dans sa version actuelle, ne permet de résoudre que des problèmes d'étiquetage binaire et ne permet de prendre en compte que des cliques correspondant à des paires. Quoique très intéressant car c'est le seul code CRF 2D disponible, ce package ne permet pas, en l'état, de traiter n'importe quel problème. Un grand nombre de travaux récents et d'outils dans le domaine des champs conditionnels sont listés sur la page personnelle d'Hanna Wallach [6]. Nous voyons donc qu'il n'existe pas de code source pour résoudre des problèmes d'étiquetage d'images à plusieurs classes, et encore moins en ce qui concerne le traitement d'images de documents. En prenant comme point de départ le modèle de champ de Markov que nous avons proposé et présenté dans le chapitre précédent, pour l'étiquetage de documents manuscrits dégradés ou anciens, nous nous proposons de développer un modèle CRF 2D permettant d'effectuer la même tâche dans une approche discriminante, afin de pouvoir comparer les deux approches et déterminer celle qui est la plus adaptée au problème que nous considérons.

4.3 Application à l'analyse d'images de documents : approche proposée

4.3.1 Modèle général

Comme nous l'avons vu dans le chapitre précédent, l'étiquetage d'images en utilisant une approche générative par champs de Markov, fournit un cadre intéressant pour la segmentation et la reconnaissance d'images de documents anciens et dégradés. Cependant ces dernières années sont apparus des modèles discriminants plus performants pour de telles tâches d'étiquetage. Toutefois ces modèles sont récents, et il y a eu encore finalement peu de travaux les concernant. Cela est d'autant plus vrai pour des données

[6] http ://www.inference.phy.cam.ac.uk/hmw26/crf/

bidimensionnelles et continues comme nous avons pu le voir dans l'étude bibliographique précédente. Pourtant du point de vue purement modélisation, le passage au 2D ne pose pas de difficultés particulières, puisque définir un modèle CRF revient à définir un ensemble de fonctions de potentiel sur les cliques maximales du graphe représentant les dépendances entre les données, que l'on soit en 1D ou en 2D. La difficulté dans l'utilisation de ces modèles réside dans l'inférence et l'apprentissage des paramètres du modèle, car les dépendances entre les variables sont plus complexes, et le graphe des dépendances comportant des cycles, il n'existe pas alors de technique de résolution exacte. Cependant les méthodes de résolutions approchées, Loopy Belief Propagation ou méthodes Monte-Carlo, permettent souvent d'obtenir des résultats satisfaisants, en tous cas meilleurs que ceux obtenus avec les approches génératives.

La modélisation et la résolution d'un problème à l'aide de champs aléatoires conditionnels nécessitent de définir les points suivants et d'effectuer un certain nombre de choix :

- définir les variables cachées et observations du problème
- définir les dépendances entre les variables cachées, c'est à dire définir la structure du graphe d'indépendance
- définir la forme des fonctions de caractéristiques
- choisir une méthode d'apprentissage des paramètres du modèle
- choisir une méthode d'inférence

Nous allons maintenant expliciter le modèle que nous proposons et les choix que nous avons effectués pour ces différents points, notamment en ce qui concerne la définition des fonctions de caractéristiques, et la méthode d'inférence utilisée.

Le modèle que nous proposons est un modèle CRF classique qui définit la distribution conditionnelle globale du champ d'étiquettes X, sachant l'observation Y, comme un produit de fonctions de potentiel locales sur les sites s, qui s'exprime sous la forme d'une combinaison de k fonctions $f_k(x, y, s)$ de caractéristiques intrinsèques et contextuelles, extraites sur l'image et sur le champ des configurations d'étiquettes à différents niveaux d'analyse. La forme générale du modèle est donc donnée par la relation

suivante :

$$P(X = x|Y = y) = \frac{1}{Z} \prod_{s \in S} [\exp(\sum_{k} \lambda_k f_k(x, y, s)]$$

Fonctions de caractéristiques

A l'instar des travaux effectués dans le domaine de l'image, par Kumar et al. [Kumar 03b], puis plus récemment par He et al. [He 04], nous avons choisi de modéliser les fonctions de caractéristiques $f(x, y, s)$ par des classifieurs discriminants. Nous avons choisi pour cela d'utiliser des Perceptrons Multi-Couches (PMC) pour la rapidité en décision. Cependant on pourrait envisager d'utiliser d'autres types de classifieurs, tels que des SVM ou des classifieurs linéaires.
Nous considérons le modèle CRF que nous proposons comme un réseau de classifieurs interconnectés prenant leurs décisions non seulement en fonction de mesures effectuées sur l'image (observations), mais également d'après les décisions prises par les classifieurs voisins (étiquettes). Nous procédons ainsi car nous travaillons sur l'image, et nous devons considérer à la fois des caractéristiques continues (caractéristiques sur l'image) et des caractéristiques discrètes (caractéristiques sur les configurations d'étiquettes).
Contrairement à ce qui est fait pour l'étiquetage de données textuelles avec des modèles CRF 1D, nous ne pouvons pas définir manuellement un ensemble de fonctions de caractéristiques binaires car nous considérons des données continues. Nous choisissons donc de modéliser ces fonctions de caractéristiques par des scores retournés par des PMC en fonction d'informations extraites d'une part sur l'image et d'autre part sur les étiquettes. La fonction d'activation utilisée étant une fonction sigmoïde, et le critère minimisé lors de l'apprentissage étant celui des moindres carrés, les PMC fournissent donc en sortie une estimation des probabilités a posteriori des étiquettes $P(x_s = l, \forall l \in L|\mathcal{F}(x, y, s))$, en fonction des caractéristiques $\mathcal{F}(x, y)$ extraites sur le niveau d'analyse k considéré. La probabilité conditionnelle globale s'exprime donc de la manière suivante :

$$P(X|Y) = \frac{1}{Z} \prod_{s \in S} [\exp(\sum_{k} \lambda_k P(x_s|\mathcal{F}(x, y, s)))]$$

où k désigne le nombre de niveaux d'analyse ou composantes discrimantes prisent en compte dans le modèle, et les λ_k sont des coefficients associés à

ces composantes.

Nous considérons dans un premier temps deux niveaux d'analyse et définissons les deux types de fonctions de caractéristiques associées : une fonction de caractéristiques locales notée f_L et une fonction de caractéristiques contextuelles f_C. La fonction f_L ne prend en compte que des caractéristiques extraites des observations y sur une fenêtre d'analyse. Elle modélise l'attache aux données, et la valeur de cette fonction est assimilable à la probabilité conditionnelle locale $P(x_s|y_s)$. La fonction de caractéristiques f_C modélise l'information contextuelle sur les étiquettes dans un voisinage plus ou moins large défini par une fenêtre d'analyse. La probabilité conditionnelle locale $P(X_s|X^s, Y)$ en chaque site s s'exprime donc sous sous la forme d'une combinaison linéaire des deux fonctions de caractérisitiques f_L et f_C, soit :

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

Cette formulation combine donc un modèle discriminant local et un modèle contextuel ce qui permet de capturer les informations issues du champ d'observation Y et celles issues du champ d'étiquettes X dans un voisinage proche. Cette formulation permet de capturer un contexte plus riche, et donc permet une meilleure régularisation et homogénéisation du champ d'étiquettes X, tout en ne négligeant pas l'information d'attache aux observations. De plus, étant dans un contexte discriminant, il n'est pas nécessaire de poser d'hypothèses d'indépendance des observations conditionnellement aux étiquettes. Cela permet de prendre en compte des caractéristiques éventuellement corrélées sur un voisinage plus large.

Le schéma 4.6 illustre le principe de la combinaison d'information dans le modèle que nous proposons.

4.3.2 Caractéristiques locales

La fonction de caractéristiques locales, ne prend en compte que des caractéristiques liées à l'image observée, en un site donné. Elle modélise donc l'attache aux données, c'est à dire l'adéquation entre l'étiquette du site et l'observation locale au site s. Nous prenons en compte les mêmes jeux de caractéristiques que ceux que nous avons utilisés avec le modèle de champ markovien que nous avons présenté dans le chapitre précédent, à savoir des

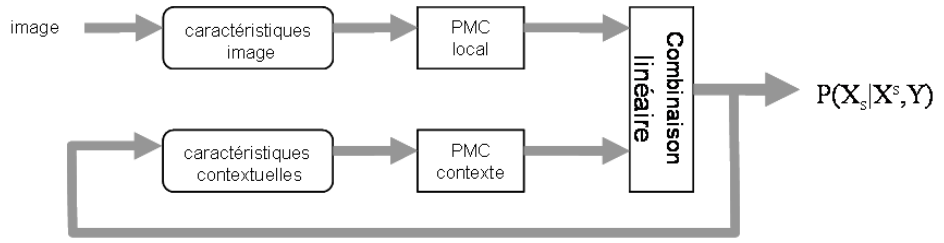


Fig. 4.6. schéma du modèle de combinaison de l'information locale et de l'information contextuelle

caractéristiques de densité de pixels noirs sur plusieurs niveaux de résolution ainsi que la position relative du site dans l'image. Ces caractéristiques sont extraites pour chaque site de l'image et le vecteur de caractéristiques ainsi formé est appliqué en entrée du PMC modélisant cette fonction de caractéristiques locales. Les scores retournés par le PMC modélisent donc la valeur de la fonction de caractéristiques pour les différentes étiquettes l_i possibles que peut prendre le site s, avec $l_i \in L = \{l_1, ..., l_q\}$.

4.3.3 Caractéristiques contextuelles

La fonction de caractéristiques contextuelles ne tient compte que des densités de probabilités conditionnelles locales $P(X_s = l_i, i = 1, ..., q|X^s, Y)$ sur le champ des étiquettes X dans un certain voisinage plus ou moins grand autour du site courant. Ce voisinage est déterminé par la définition d'une fenêtre d'analyse dont la taille est déterminée par la quantité de contexte que l'on souhaite intégrer. Par exemple, pour une fenêtre d'analyse de taille 3×3 et un alphabet d'étiquettes de taille $q = 3$ on obtient 27 densités de probabilités conditionnelles sur la fenêtre d'analyse, soit un vecteur de 27 caractéristiques.

4.3.4 Apprentissage du modèle proposé

L'apprentissage du modèle consiste d'une part à entraîner les deux PMC, local et contextuel, et d'autre part à déterminer les coefficients λ_L et λ_C de la combinaison linéaire (cf section 4.3.1). Nous considérons pour cela que nous disposons de données complètement étiquetées, c'est à dire que nous

disposons pour chaque image de la base d'apprentissage, de l'étiquetage correct associé.
Les deux PMC sont appris en utilisant l'algorithme de rétropropagation du gradient. Le PMC local est entraîné en premier à partir des caractéristiques extraites de l'image. Une fois appris, le PMC local est utilisé pour estimer les probabilités conditionnelles $P(X_s|Y_s)$ en chaque site de l'image qui sont utilisés comme caractéristiques pour l'apprentissage du PMC contextuel.
Les coefficients de la combinaison sont appris en utilisant une méthode de descente de gradient de manière à minimiser l'erreur d'étiquetage en chaque site.
Il est important de noter que dans la version actuelle du modèle que nous proposons et que nous avons implanté, les dimensions du maillage sont fixées manuellement et empiriquement. Il est clair que la taille de ce maillage influe directement sur le résultat de l'étiquetage et qu'elle détermine la finesse de l'étiquetage. C'est pourquoi le choix de ce maillage est primordial, et une procédure d'apprentissage automatique devrait permettre de le fixer plus efficacement. Cependant une telle procédure est complexe et lourde à mettre en place. En effet, les classifieurs PMC de notre modèle sont appris en fonction de la taille de maillage donnée, et procéder à l'apprentissage de la taille optimale du maillage avec une approche classique par maximisation de la vraisemblance des exemples d'apprentissage nécessiterait un certain nombre d'itérations avec des tailles de mailles différentes, et donc impliquerait systématiquement un ré-apprentissage des PMC. Sachant que l'apprentissage est quelque chose de relativement long, ceci n'a pas été envisagé en l'état.

4.3.5 Inférence avec le modèle proposé

Comme nous avons pu le voir dans notre étude bibliographique, les techniques utilisées pour l'inférence dans un champ aléatoire conditionnel, sont les mêmes que celles utilisées pour l'inférence dans un champ de markov. Dans le cas bidimensionnel le graphe d'indépendance est quelconque, il n'existe donc pas de méthode d'inférence exacte. Il est donc nécessaire d'avoir recours à des méthodes approchées. Parmi les méthodes les plus utilisées, on trouve l'algorithme Belief Propagation ou encore les techniques d'échantillonnage, telles que l'échantillonneur de Gibbs ou de Metropolis avec recherche des marginales maximales (critère des marginales maximales a posteriori). Les techniques de relaxation probabiliste telles que l'algorithme

du recuit simulé, l'algorithme ICM ou encore l'algorithme HCF peuvent également être utilisées. Ces algorithmes permettent de rechercher une solution approchée de la configuration d'étiquettes optimale au sens du critère du Maximum A Posteriori, c'est à dire la configuration $\hat{x}$ telle que :

$$\hat{x} = \arg\max_{x} P(X = x|Y = y)$$

Étant donné que nous avons implanté les algorithmes ICM et HCF dans le cadre de notre modèle de champ markovien, nous avons choisi de conserver ces mêmes algorithmes pour l'inférence avec le modèle CRF que nous proposons, car ce sont des algorithmes rapides. L'inférence se déroule sensiblement de la même manière que dans le cas du modèle de champ markovien présenté dans le chapitre précédent. Cependant il faut noter que notre modèle CRF utilise comme caractéristiques contextuelles, l'incertitude sur les étiquettes des sites voisins. De ce fait, les algorithmes ICM et HCF ne prennent pas seulement en compte la meilleure étiquettes des sites voisins comme dans leur version standard. Toute l'incertitide sur les étiquettes des voisins étant prise en compte, on se rapproche en quelque sorte du fonctionnement des algorithmes par passage de messages tels que Belief Propagation et Loopy Belief Propagation.

L'inférence se déroule donc de la manière suivante :
Un premier étiquetage de l'image est effectué en se basant uniquement sur le classifieur local. L'information contextuelle sur les labels des sites voisins n'est donc pas prise en compte lors de cette première phase. Cette première passe sur l'image permet d'initialiser le champ des étiquettes et de calculer en chaque site les valeurs de la fonction de caractéristiques locales, c'est à dire les probabilités a posteriori de chaque étiquette $l \in L$ sachant les caractéristiques extraites de l'image, soit $P(X_s = l, \forall l \in L|Y_s = y_s)$. Elles sont utilisées par la suite comme caractéristiques appliquées en entrée du classifieur contextuel. Pour les itérations suivantes la fonction de caractéristiques contextuelles est également prise en compte pour évaluer la fonction de potentiel sur chacun des sites de l'image. L'inférence consiste alors à parcourir tous les sites de l'image et à évaluer en chaque site, le score de la fonction de potentiel pour chaque étiquette de l'alphabet L, en combinant les sorties du classifieur local et du classifieur contextuel. Ce

score peut être assimilé à la probabilité d'affecter l'étiquette l au site s sachant les observations locales Y_s, et les probabilités des étiquettes sur un voisinage plus ou moins large. L'étiquette fournissant le score le plus élevé est affecté au site courant, et les probabilités conditionnelles sont mémorisés et constamment remises à jour au fur-et-à-mesure de l'inférence, comme le champ des étiquettes. Ce processus de remise à jour est répété jusqu'à convergence de la configuration d'étiquettes sur le champ X.

4.3.6 Expérimentations et résultats

Les bases utilisées pour les expérimentations et les tâches d'étiquetage considérées, sont les mêmes que celles que nous avons présentées dans le chapitre précédent.

Influence de la méthode d'inférence

Nous étudions dans un premier temps l'influence de la méthode de décodage. Nous comparons pour cela les résultats d'étiquetage obtenus avec le modèle CRF que nous proposons en utilisant les méthodes d'inférence ICM et HCF, avec les résultats obtenus en utilisant simplement en chaque site un classifieur local (MLP) prenant en compte les mêmes caractéristiques image locales que notre modèle CRF.
Ces expérimentations ont été menées sur la base "Imagination Poétique" en considérant la tâche d'étiquetage au niveau bloc définie dans le chapitre 3. Nous considérons comme critère d'évaluation pour ces expérimentations, le taux moyen d'étiquetage correct par classe (TEN).

Les résultats obtenus sont illustrés sur le graphe 4.7.

On peut remarquer que selon cet indicateur de performance, quelle que soit la méthode d'inférence utilisée, notre modèle CRF permet d'améliorer significativement le taux d'étiquetage correct de presque 5 points, par rapport au résultat fournit par un classifieur discriminant local. L'intégration du contexte semble donc permettre un meilleur étiquetage de l'image.
En ce qui concerne la méthode d'inférence, on ne constate pas de différence significative entre le taux d'étiquetage obtenu avec la méthode ICM et celui obtenu avec HCF. Pour la suite des expérimentations nous avons donc

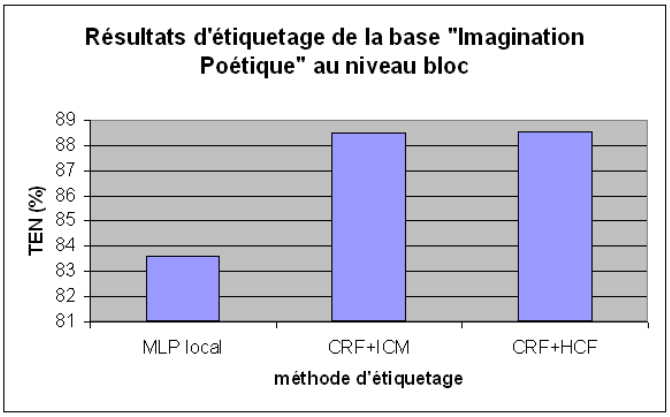


Fig. 4.7. Taux moyens d'étiquetage correct des pixels obtenus sur la base "Imagination Poétique" pour la tâche d'étiquetage au niveau bloc, avec : un MLP local, le modèle CRF proposé avec méthode d'inférence ICM, et le modèle CRF proposé avec méthode d'inférence HCF

choisi de ne conserver que la méthode ICM pour sa simplicité, sa rapidité, son efficacité et sa faible complexité combinatoire.

Si on analyse les matrices de confusion d'étiquetage des pixels (cf figure 4.8), on constate qu'il y a des confusions importantes entre la classe "numéro de page" et la classe "sous-titre" ou encore entre la classe "titre" et la classe "sous-titre". Ceci est dû au fait de la proximité géographique de ces entités dans la page, et de caractéristiques descriptives relativement proches pour certaines de ces classes (les classes "titre", "sous-titre" et "numéro" correspondent par exemple toutes à des zones de texte. Ce qui diffère, c'est leur position dans la page et relativement aux autres entités). On observe d'une manière générale une relative confusion entre les classes correspondant à des entités spatialement proches dans la page du document. Toutefois, si on compare la matrice de confusion obtenue avec un MLP local et les matrices de confusions obtenues avec le modèle CRF proposé, on peut remarquer que l'intégration d'une information contextuelle dans le modèle CRF, permet de réduire ces confusions entre entités spatialement proches, notamment des entités correspondant à des objets de type "forme" au détriment des zones de type "fond". Ceci peut s'expliquer par une répartition non homogène des classes dans les bases comme on peut le voir sur la figure 4.9. La classe "fond" est la classe la plus largement représentée dans la base. Le modèle contextuel ayant tendance à homogénéiser le champ d'étiquettes, c'est la classe la plus représentée qui a tendance à être pénalisée.

	MLP local						
	fond	corps de texte	illustration	sous-titre	titre	n° de page	fleuron
fond	0,89	0,04	0,01	0,02	0,02	0,02	0,01
corps de texte	0,05	0,89	0,02	0,01	0,01	0,00	0,02
illustration	0,00	0,03	0,94	0,03	0,00	0,00	0,00
sous-titre	0,04	0,00	0,04	0,77	0,12	0,03	0,00
titre	0,00	0,00	0,00	0,09	0,87	0,03	0,02
n° de page	0,12	0,00	0,00	0,12	0,08	0,65	0,04
fleuron	0,10	0,05	0,01	0,00	0,00	0,00	0,84
						TEN = 83,62 %	

(a)

	Modèle CRF + méthode inférence ICM						
	fond	corps de texte	illustration	sous-titre	titre	n° de page	fleuron
fond	0,87	0,06	0,01	0,01	0,01	0,03	0,01
corps de texte	0,05	0,94	0,01	0,00	0,00	0,00	0,01
illustration	0,00	0,05	0,93	0,03	0,00	0,00	0,00
sous-titre	0,04	0,00	0,04	0,80	0,08	0,04	0,00
titre	0,00	0,01	0,00	0,08	0,88	0,03	0,01
n° de page	0,00	0,04	0,00	0,08	0,00	0,88	0,00
fleuron	0,08	0,03	0,00	0,00	0,00	0,00	0,90
						TEN = 88,51 %	

(b)

	Modèle CRF + méthode inférence HCF						
	fond	corps de texte	illustration	sous-titre	titre	n° de page	fleuron
fond	0,87	0,06	0,01	0,01	0,01	0,03	0,01
corps de texte	0,05	0,94	0,01	0,00	0,00	0,00	0,01
illustration	0,00	0,05	0,93	0,01	0,00	0,00	0,00
sous-titre	0,06	0,00	0,04	0,82	0,06	0,02	0,00
titre	0,00	0,01	0,00	0,08	0,88	0,03	0,01
n° de page	0,00	0,04	0,00	0,08	0,00	0,88	0,00
fleuron	0,08	0,03	0,00	0,00	0,00	0,00	0,90
						TEN = 88,54 %	

(c)

Fig. 4.8. Matrices de confusion obtenues sur la base "Imagination Poétique" pour la tâche d'étiquetage au niveau bloc, avec : (a) un MLP local, (b) le modèle CRF proposé avec méthode d'inférence ICM, et (c) le modèle CRF proposé avec méthode d'inférence HCF

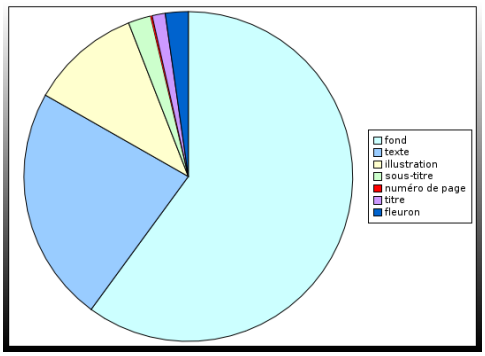


Fig. 4.9. Répartition des classes dans la base Imagination Poétique

Ces constatations se vérifient par une analyse qualitative des résultats, en inspectant visuellement les résultats d'étiquetage obtenus sur les images de la base (cf figure 4.10). Comme l'analyse des matrices de confusion a permis de le montrer, on vérifie en effet que les résultats présentent un certain nombre de cas de fausses détections et de confusions entre les étiquettes. Ainsi par exemple, certaines zones de fond sont étiquetées à tort comme des zones de numéro de page. Ceci est dû au bruit de l'image. Les frontières de certaines zones ne sont également pas toujours très bien délimitées, et on peut observer parfois un étalement assez important de certaines zones (zones étiquetées "corps de texte" débordant sur le fond). Ceci est dû en particulier au maillage. En effet si les mailles sont grandes, elles peuvent parfois être "à cheval" sur deux entités, et donc l'étiquetage sera grossier sur la frontière entre ces deux entités.
Cependant, malgré ces petits défauts, pratiquement toutes les entités informatives (numéro de page, illustration, titre, texte, ...) sont correctement détectées. Les erreurs d'étiquetage concernent principalement des pixels de fond (pixels blancs) et des zones non informatives en temps que telles (zones correspondant au fond ou à des espaces). Ces erreurs d'étiquetage, quoique non négligeables, sont donc peu préjudiciables dans la mesure où la majorité des pixels forme et des entités informatives sont correctement étiquetées et détectées. Ceci est important dans un contexte d'aide à l'annotation et à l'indexation, car une détection correcte des entités informatives permet déjà de réduire la complexité de la tâche de l'utilisateur. De plus, dans un tel contexte, il est toujours

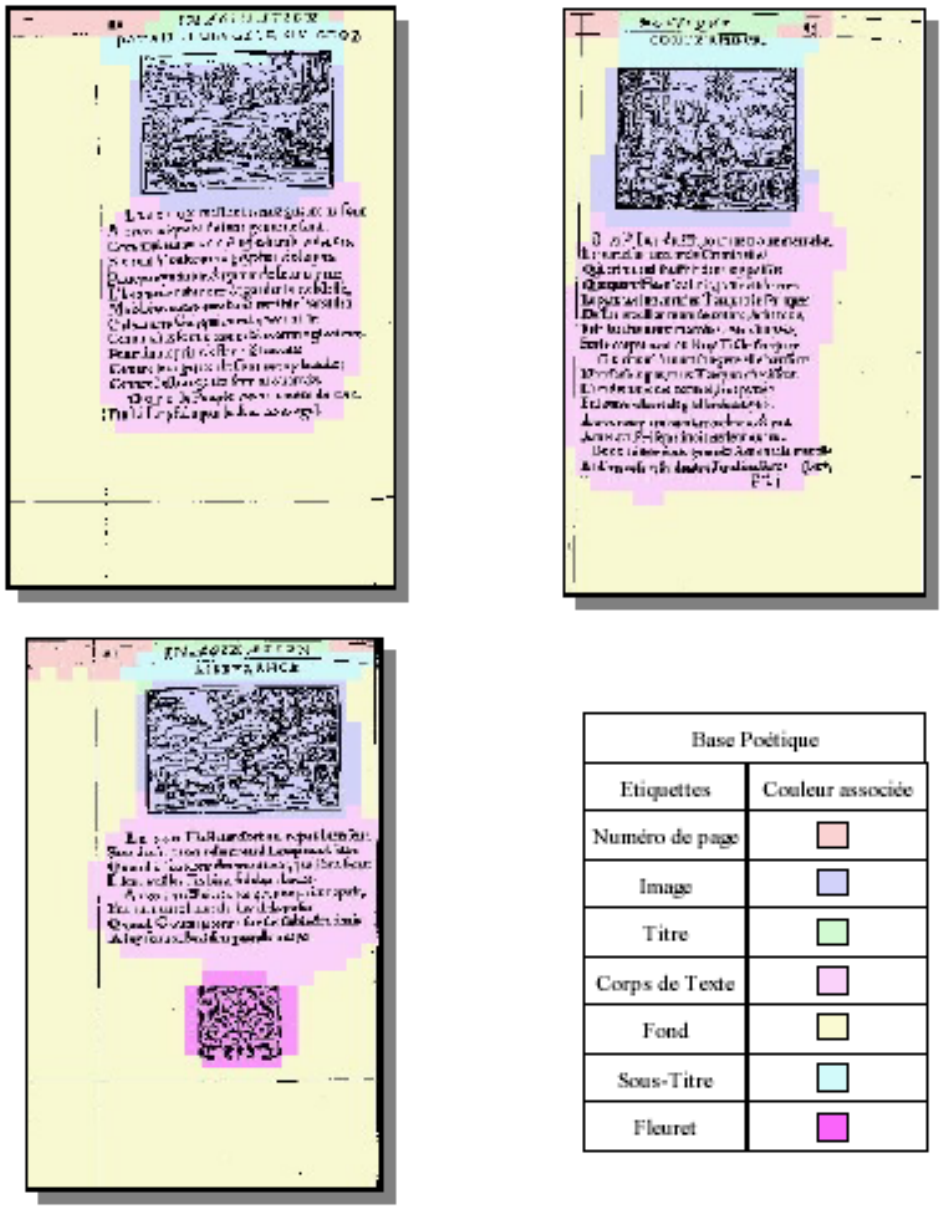
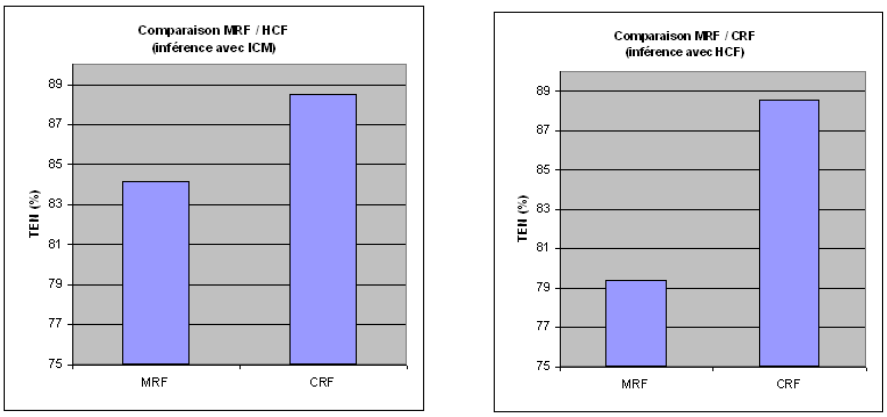


Fig. 4.10. Résultats de l'étiquetage sur des documents de la base Imagination Poétique

possible ensuite d'ajuster manuellement les limites des zones de manière interactive si celles retournées par la méthode ne sont pas tout à fait exactes.

Comparaison des résultats obtenus avec notre modèle MRF et notre modèle CRF

Comme nous l'avons vu précédemment, la prise en compte de l'information contextuelle dans le modèle CRF que nous proposons, permet d'améliorer significativement les résultats d'étiquetage. Nous comparons maintenant, sur la même base (base "Imagination Poétique"), en considérant la même tâche d'étiquetage au niveau bloc, les résultats obtenus avec le modèle de champ de Markov caché que nous avons présenté dans le chapitre 3 et le modèle de champ aléatoire conditionnel que nous présentons dans ce chapitre, en utilisant les mêmes algorithmes d'inférence, ICM et HCF. La comparaison que nous proposons, repose une nouvelle fois sur le taux moyen d'étiquetage correct des pixels par classe (TEN). Les résultats obtenus sont illustrés sur le graphe 4.11.



(a) (b)

Fig. 4.11. Comparaison des taux moyens d'étiquetage corrects des pixels obtenus sur la base "Imagination Poétique" pour la tâche d'étiquetage au niveau bloc, avec le modèle MRF et le modèle CRF, en utilisant pour l'inférence : (a) ICM et (b) HCF

Ces résultats montrent que pour la même base de test et la même tâche d'étiquetage des zones d'intérêt au niveau bloc, quelle que soit la méthode d'inférence utilisée (ICM ou HCF), le modèle CRF améliore significativement le taux moyen d'étiquetage correct des pixels par classe, par rapport au modèle MRF.

Conclusion

Nous avons pu voir au travers de ces premières expérimentations, que sur cette base et pour la tâche d'étiquetage au niveau bloc considérée, notre modèle CRF qui intègre à la fois des caractéristiques intrinsèques locales et des caractéristiques contextuelles sur les incertitudes du champ d'étiquette, par combinaison de classifieurs dans un cadre discriminant, semble donner des meilleurs résultats d'étiquetage que notre modèle MRF précédent, en termes de taux d'étiquetage. Nous avons également pu vérifier que pour notre modèle CRF, les algorithme d'inférence ICM et HCF donnent sensiblement les mêmes résultats. Pour la suite des expérimentations nous avons choisi de conserver uniquement ICM comme algorithme de décodage, pour sa simplicité et son efficacité.

4.4 Evolution du modèle CRF proposé : intégration de caractéristiques globales dans la fonction de potentiel

Comme les résultats précédents le montrent, le modèle CRF que nous proposons permet d'obtenir de meilleurs taux et de meilleurs résultats d'étiquetage que le modèle MRF que nous avons présenté dans le chapitre 3. Cette amélioration des résultats est due d'une part à une meilleure intégration de l'information contextuelle dans la fonction de potentiel, par l'intermédiaire de la fonction de caractéristiques contextuelles qui permet de prendre en compte plus facilement les dépendances entre les étiquettes, et d'autre part au fait que le modèle proposé est réellement discriminant. Afin d'améliorer encore les résultats obtenus avec ce modèle, on se propose d'intégrer une troisième fonction de caractéristiques permettant de prendre en compte un contexte plus large sur le champ X des étiquettes et d'évaluer la configuration d'étiquettes à un niveau plus global. Cette fonction est appelée fonction de caractéristiques globales et est noté f_G. Le but de l'intégration de cette fonction de caractéristiques globales, est de permettre une meilleure homogénéisation et une plus grande cohérence de la configuration d'étiquettes sur le champ X.

4.4.1 Fonction de caractéristiques globales

Une analyse du champ d'étiquettes X à un niveau plus global est maintenant également prise en compte par une troisième fonction de caractéristiques, dite globale. Cette analyse globale est réalisée à l'aide d'un troisième Perceptron Multicouche. Ce classifieur estime les probabilités a posteriori $P(X_s = l_i, \forall l_i \in L | \{_G(X))$ d'affecter l'étiquette l_i au site courant s, sachant un ensemble { de caractéristiques statistiques globales extraites sur la configuration des étiquettes dans un voisinage plus large que celui pris en compte par la fonction de caractéristiques contextuelles. Le classifieur global est donc également un classifieur contextuel qui tient compte de la configuration du champ d'étiquettes, cependant il s'agit ici de caractériser le champ d'étiquettes X à un niveau plus global, en mesurant des paramètres statistiques sur la configuration courante d'étiquettes x sur le champ X. Pour cela le champ est divisé en plusieurs zones par superposition d'un maillage H plus large que le maillage G initiale. Chaque zone sur le maillage H correspond donc à un ensemble de sites. Sur chacune de ces zones sont calculés des paramètres statistiques sur la configuration des étiquettes. En l'occurrence, on construit la matrice de co-occurrence des étiquettes dans chacune des zones, ceci pour différentes orientations. Plus précisément quatre matrices de co-occurrence sont déterminées pour les orientations 0, 45, 90 et 135 degrés. A partir de ces quatre matrices de co-occurrence, sont calculés cinq paramètres d'Haralick [Haralick 73][Haralick 79]. Les matrices de co-occurrence et les paramètres d'Haralick sont traditionnellement utilisés pour caractériser des textures dans des images en niveaux de gris. Les matrices de co-occurrence à elles seules suffisent à caractériser une texture, mais représentent une quantité d'information trop importante et trop lourde à manipuler. Les paramètres d'Haralick extraits de ces matrices permettent de réduire cette information à quelques descripteurs numériques pertinents et suffisants. Il existe en réalité 14 paramètres d'Haralick, mais 5 sont principalement utilisés, et reconnus comme discriminants. Ces paramètres sont : l'homogénéité, l'homogénéité locale, la corrélation, l'entropie et le contraste. De cette manière on peut réduire la caractérisation d'une texture à 5 paramètres seulement. L'originalité de l'approche que nous proposons repose sur le fait que nous déterminons ces caractéristiques non pas directement sur l'image, mais sur le champ des étiquettes. L'idée est de caractériser des configurations particulières d'étiquettes à l'aide d'indicateurs numériques. A

ces 5 paramètres d'Haralick, on ajoute la position de la zone H_s à laquelle appartient le site s considéré, ainsi que les probabilités des étiquettes en ce site. Ces caractéristiques constituent alors le vecteur appliqué en entrée du PMC modélisant la fonction de caractéristiques globales. Ce classifieur estime donc les probabilités a posteriori d'appartenance du site aux différentes classes, en fonction des caractéristiques du champ d'étiquettes dans la zone H_s dans laquelle se trouve le site courant s. L'apprentissage de ce troisième classifieur est réalisé séquentiellement après les deux autres.

4.4.2 Combinaison des sources d'information

Le modèle proposé intègre donc maintenant trois types d'information à des niveaux d'analyse différents. Chacune de ces sources d'information est modélisé par un PMC. Pour ces trois sources d'information en chaque site de l'image, un classifieur se prononce sur l'étiquette du site.

La probabilité conditionnelle locale $P(X_s|X^s, Y)$ en chaque site s, sachant les observations Y et le reste du champ X, s'exprime donc maintenant de la manière suivante :

$$P(X_s|X^s, Y) = h(f_L, f_C, f_G)$$

où h est une fonction de combinaison des 3 sources d'information : locale, contextuelle et globale.

En pratique il y a plusieurs manières d'implémenter cette fonction de combinaison h. Nous proposons deux solutions de combinaison :

1. une combinaison linéaire de la fonction de caractéristiques globales avec la sortie du modèle précédent (combinant lui-même linéairement les fonctions de caractéristiques locales et contextuelles)
2. une combinaison non linéaire des fonctions de caractérisitiques locales, contextuelles et globales à l'aide d'un PMC de combinaison.

Nous allons détailler ces deux solutions permettant d'intégrer une caractérisation globale des étiquettes au modèle précédent.

Combinaison linéaire des sources d'information

D'un point de vue pratique, la manière la plus simple d'intégrer la fonction de caractéristiques globales que nous venons de définir au modèle CRF que nous avons présenté précédemment, consiste à combiner linéairement cette fonction de caractéristiques globales à la sortie du modèle précédent, c'est à dire au résultat de la combinaison linéaire de la fonction de caractéristiques locales et de la fonction de caractéristiques contextuelles. Le modèle ainsi défini intègre donc deux niveaux de combinaison (cf figure 4.12). Le résultat est donc une combinaison non linéaire résultant de deux combinaisons linéaires en cascade.
L'apprentissage des coefficient de ces deux combinaisons est réalisé en deux temps. Tout d'abord les coefficients λ_L et λ_C de la première combinaison sont déterminés sur l'ensemble d'apprentissage de manière à maximiser le taux moyen d'étiquetage correct, en ne tenant pas compte de l'information globale. Une fois les coefficients de la première combinaison fixés, les coefficients λ_{L-C} et λ_G de la seconde combinaison sont déterminés de la même manière, en introduisant cette fois l'information globale.

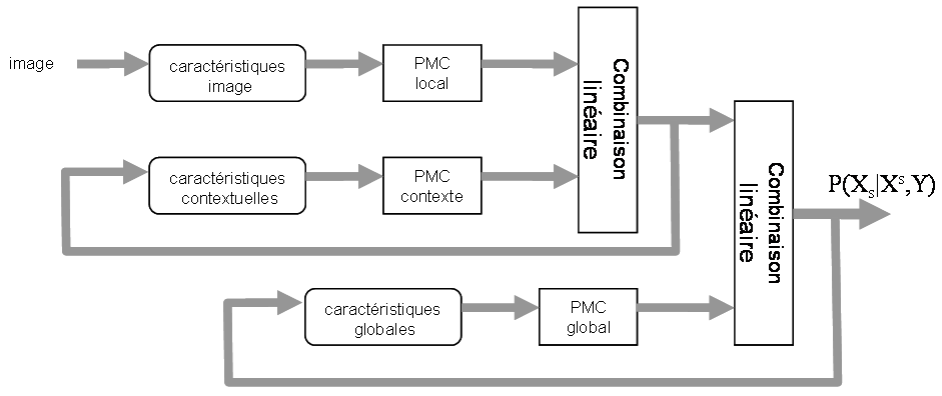


Fig. 4.12. Modélisation à deux niveaux de combinaison

Ce type de combinaison permet de contrôler facilement la part d'information apportée par la fonction de caractéristiques globales par rapport à l'information apportée par le modèle précédent. La première combinaison synthétise l'information locale alors que la seconde permet d'intégrer

l'information globale.

Combinaison des sources d'information par un PMC

L'alternative que nous proposons pour cette deuxième méthode de combinaison, consiste à remplacer les deux étages de combinaison linéaire successifs, par un Perceptron Multicouche, dont le but sera de fusionner les différentes sources d'information. En effet, la formulation d'un PMC étant assez proche de celle d'un champ aléatoire conditionnel, puisqu'il s'agit d'une combinaison non linéaire de caractéristiques. Les valeurs des trois fonctions de caractéristiques pour les différentes étiquettes possibles sont mises en entrée d'un seul PMC. Si l'alphabet L comporte q étiquettes, le vecteur de caractéristiques appliqué en entré du PMC de combinaison sera donc de dimension $3q$. L'implantation correspondant à cette solution de combinaison est illustrée sur la figure 4.13. Avec ce type de combinaison les 3 sources d'information sont considérées à niveau égal. La part d'information apportée par chaque composante dépend des poids synaptiques du PMC, mais ne peut être connue explicitement.
L'apprentissage de ce modèle consiste à entraîner sur des données d'apprentissage étiquetées, les 3 PMC modélisant les 3 fonctions de caractéristiques, puis à entraîner ensuite le PMC de combinaison. L'apprentissage de tous les PMC est réalisé en utilisant l'algorithme de rétropropagation du gradient.
L'avantage de cette alternative est qu'elle ne présent plus qu'un seul étage de combinaison. Cette méthode est donc très simple. Son inconvénient est par contre le temps nécessaire à l'apprentissage du PMC.

4.4.3 Expérimentations et résultats

Nous avons cherché à déterminer dans ces expérimentations, d'une part l'apport du contexte dans le résultat de l'étiquetage, et d'autre part l'apport de la fonction de caractéristiques globales ainsi que les performances obtenus avec les différentes solutions de combinaison des fonctions de caractéristiques. Nous comparons ensuite les résultats obtenus avec les différents modèles proposés, avec ceux obtenus avec le modèle MRF proposé dans le chapitre 3.

Pour ces expérimentations nous avons utilisé la base "Bovary", et

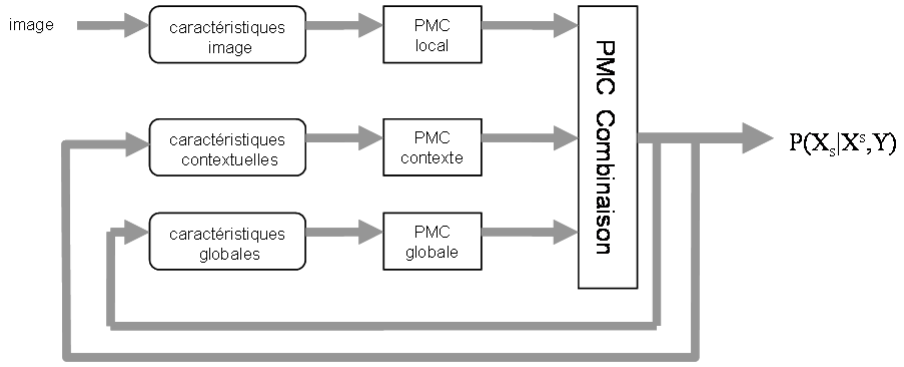


Fig. 4.13. Modèle de combinaison des 3 sources d'information à l'aide d'un PMC

nous avons considéré la tâche d'étiquetage au niveau bloc décrite dans le chapitre 3. Nous rappelons que pour cette tâche l'alphabet L comporte les 6 étiquettes suivantes : "corps de texte", "bloc de texte", "numéro de page", "marge", "haut de page", "bas de page". La taille de la maille a été fixée à une dimension de 50×50 pixels pour tous les tests que nous avons effectués. Cette taille de maille a été choisie empiriquement de manière à trouver un bon compromis entre finesse du résultat de l'étiquetage et la réduction de la complexité.

Apport de l'information contextuelle

Nous avons dans un premier temps cherché à évaluer l'apport de l'information contextuelle dans le résultat de l'étiquetage. Pour cette première évaluation nous considérons le premier modèle CRF que nous avons proposé, qui combine une fonction de caractéristiques locales et une fonction de caractéristiques contextuelles, en prenant en compte différentes tailles du contexte. Pour cela nous définissons les deux jeux de caractéristiques suivants :

- Jeu 1 : les caractéristiques utilisées pour le PMC local sont les 18 caractéristiques de densité de pixels noirs et les caractéristiques de position, soit au total 20 caractéristiques. Le PMC contextuel a été paramétré avec une fenêtre contextuelle de taille 3*3 sites, sur laquelle

sont extraites les probabilités conditionnelles $P(X_s = l, \forall l \in L | X^s, Y)$, soit 6 probabilités conditionnelles pour chacun des 9 sites sur la fenêtre contextuelle, donc au total 54 caractéristiques contextuelles.

– Jeu 2 : les caractéristiques locales sont les mêmes que pour le jeu 1 mais le PMC contextuel a été paramétré avec une fenêtre contextuelle de taille 5*5 sites, soit au total 150 caractéristiques contextuelles.

En ce qui concerne l'inférence, deux méthodes sont utilisées : la méthode ICM et la méthode HCF.

Les taux moyens d'étiquetage correct par classe que nous obtenons avec ce premier modèle en utilisant deux tailles différentes de fenêtre contextuelle sont synthétisés dans le tableau 4.1 et dans le tableau 4.2. Ces taux sont comparés avec ceux obtenus avec un MLP ne prenant en compte que les caractéristiques image locales.

	PMC local	CRF (local+contextuelle)
Jeu 1	90.56	92.55
Jeu 2	90.56	93.91

Tab. 4.1. Taux moyens d'étiquetage correct obtenus avec un classifieur local et notre modèle CRF avec l'algorithme d'inférence ICM

	PMC local	CRF (local+contextuelle)
Jeu 1	90.56	93.69
Jeu 2	90.56	93.83

Tab. 4.2. Taux moyens d'étiquetage correct obtenus avec un classifieur local et notre modèle CRF avec l'algorithme d'inférence HCF

Comme nous avions pu le constater avec les expérimentations effectuées précédemment sur la base "Imagination Poétique", on vérifie également sur la base "Bovary" que grâce à l'apport de l'information contextuelle, quelle que soit la méthode d'inférence utilisée, les résultats obtenus avec le modèle CRF sont meilleurs que ceux obtenus avec un classifieur local ne prenant en compte que des caractéristiques locales sur les observations. Nous pouvons constater sur ces résultats que plus le contexte pris en compte par le modèle CRF est important, meilleurs sont les résultats. Ainsi les résultats obtenus avec le jeu de caractéristiques n°2 qui intègre une fenêtre contextuelle 5×5

sont globalement meilleurs que ceux obtenus avec le jeux n°1 qui n'intègre qu'une fenêtre contextuelle de taille 3×3.
En ce qui concerne la méthode de décodage, on vérifie une nouvelle fois que les résultats obtenus avec les deux méthodes ICM et HCF sont sensiblement identiques. Le taux d'étiquetage obtenu avec ICM pour une fenêtre contextuelle plus large étant très légèrement supérieur, l'inférence sera réalisé avec ICM dans la suite des expérimentations.

Apport de l'information globale

Nous étudions ensuite l'apport de la fonction de caractéristiques globales dans le modèle CRF proposé, et l'influence de la solution de combinaison des 3 fonctions de combinaison, en comparant les taux d'étiquetage obtenus avec les trois implémentations du modèle CRF proposées :

1. combinaison linéaire de la fonction de caractéristiques locales et de la fonction de caractéristiques contextuelles
2. combinaison linéaire de la fonction de caractéristiques globales avec le résultat de la combinaison linéaire de la fonction de caractéristiques locales et de la fonction de caractéristiques contextuelles
3. combinaison non-linéaire des fonctions de caractéristiques locales, contextuelle et globale à l'aide d'un PMC

Les taux moyens d'étiquetage correct par classe obtenus avec ces trois implémentations en considérant les deux jeux de caractéristiques énoncés précédemment et en utilisant l'algorithme d'inférence ICM, sont présentés dans le tableau 4.3

	Implémentation 1	Implémentation 2	Implémentation 3
Jeu 1	92.55	93.90	94.04
Jeu 2	93.91	93.93	94.16

Tab. 4.3. Taux moyens d'étiquetage correct obtenus avec les différentes implémentations de notre modèle CRF avec l'algorithme d'inférence ICM

Les taux d'étiquetage obtenus en intégrant la fonction de caractéristiques globales (implémentations 2 et 3) sont très légèrement supérieurs à ceux obtenus avec le modèle CRF n'intégrant pas cette fonction de caractéristiques (implémentation 1). En ce qui concerne la solution de combinaison de

cette information globale, la combinaison par un PMC (implémentation 3) semble donner de meilleurs résultats. De plus on constate une nouvelle fois, que quelle que soit l'implémentation, un contexte plus important (jeu n°2) permet d'améliorer les taux d'étiquetage. Cette augmentation est toutefois plus sensible pour l'implémentation 1 qui n'intègre pas d'information globale.

Les tableaux suivants illustrent les poids accordés à chaque composante (locale, contextuelle, globale) dans les implémentations 1 et 2, en fonction de la taille de la fenêtre contextuelle prise en compte. Ces coefficients ont été déterminés par apprentissage.

	λ_L	λ_C
jeu n°1	0.45	0.55
jeu n°2	0.42	0.58

Tab. 4.4. Coefficients de la combinaison linéaire attribués à chaque composante dans l'implémentation 1

Nous pouvons voir dans le tableau 4.4 que dans le cas de l'implémentation n°1 (sans intégration d'information globale) l'information locale d'attache aux données et l'information contextuelle sur les étiquettes sont quasiment considérées à part égale. Toutefois, si on augmente la taille de la fenêtre contextuelle, l'information contextuelle a tendance à être d'avantage prise en compte. Cela se traduit par une homogénéisation plus importante du champ d'étiquettes et une amélioration du taux d'étiquetage comme nous l'avons vu précédemment.

Dans le cas de la deuxième implémentation proposée, s'ajoute l'information globale sur la configuration d'étiquettes. Le tableau suivant illustre l'importance accordée par le modèle à cette information globale par rapport à l'importance accordée à la combinaison de l'information locale et de l'information contextuelle.

	λ_{L-C}	λ_G
jeu n°1	0.83	0.17
jeu n°1	0.70	0.30

Tab. 4.5. Coefficients de la seconde combinaison linéaire dans l'implémentation 2

Nous pouvons voir dans ce tableau 4.5 que l'information globale

contribue moins à la décision que l'information locale d'attache aux données et l'information contextuelle dans un voisinage restreint. Néanmoins la part de cette information globale n'est pas négligeable, et il apparaît que les trois sources d'information ont un apport dans l'étiquetage final. L'information locale d'attache aux données garantit une certaine cohérence vis à vis des données observées, l'information contextuelle garantit une certaine homogénéité dans le champ d'étiquettes vis à vis du voisinage, et l'information globale garantit la cohérence de l'étiquetage au niveau de la page, en permettant une analyse des configurations d'étiquettes à une résolution plus faible.

Enfin, en ce qui concerne la troisième implémentation, nous ne pouvons pas déterminer précisément la part accordée à chaque composante dans la décision, dans la mesure où cela est déterminé par l'ensemble des poids synaptiques du PMC de combinaison.

Evaluation des temps de traitement

Le tableau 4.6 illustre le temps de traitement nécessaire pour étiqueter une image de la base Bovary de dimensions 3300×2500 pixels, avec les différentes implémentations proposées et le jeu de caractéristiques n°1 (20 caractéristiques locales et fenêtre contextuelle 3×3).

	ICM	HCF
implémentation 1	32.04 s	30.04 s
implémentation 2	40.02 s	38.22 s
implémentation 3	39.89 s	37.84 s

Tab. 4.6. Temps de traitement nécessaires à l'étiquetage d'une image de la base de test Bovary avec les différentes implémentations du modèle CRF et les méthodes d'inférence ICM et HCF

Nous constatons très naturellement que le temps nécessaire au décodage augmente avec le nombre de fonctions de caractéristiques intégrées dans le modèle CRF (implémentation 1 versus implémentations 2 et 3). Les temps obtenus avec les implémentations 2 et 3 sont quasiment identiques, l'implémentation 3 étant très légèrement plus rapide que la 2 du fait de la combinaison en cascade de celle-ci.
En ce qui concerne la méthode d'inférence, les temps de traitement obtenus

sont sensiblement les mêmes avec HCF qu'avec ICM. La méthode HCF est toutefois un peu plus rapide car elle nécessite moins d'itérations pour converger, du fait d'une meilleure stratégie de parcours des sites de l'image. Cependant, comme nous l'avons vu dans le chapitre 3, la méthode HCF est un peu plus coûteuse en ressources car elle nécessite la gestion d'une pile.
Ce temps nécessaire au décodage n'est pas négligeable, mais n'est pas non plus prohibitif dans le cadre d'une application d'aide à la transcription. Le temps nécessaire à l'apprentissage du modèle est par contre quant à lui nettement plus contraignant car il implique l'apprentissage de réseaux neuronaux qui est relativement long (plusieurs heures, voire plusieurs jours). Cela représente un frein pour ce type d'approche. Néanmoins cet apprentissage n'a besoin d'être réalisé qu'une seule fois sur quelques images annotées, pour permettre ensuite l'étiquetage de larges corpus sans avoir besoin de régler quelque paramètre que ce soit. Le gain est donc important.

Comparaison des résultats avec le modèle MRF

Pour terminer nous comparons les taux d'étiquetage obtenus avec les différentes implémentations de notre modèle CRF en utilisant le jeu de caractéristiques n°2 (fenêtre contextuelle 5×5), car c'est celui-ci qui semble être le plus performant, avec les taux d'étiquetages obtenus avec le modèle MRF présenté au chapitre 3. Pour l'inférence avec les modèles CRF et le modèle MRF, l'algorithme ICM a été utilisé. Les résultats obtenus sont synthétisés dans le tableau 4.7.

	mélange gaussien	MLP local	MRF	CRF impl.1	CRF impl.2	CRF impl.3
TEN (%)	83,70	87,50	90,56	93,91	93,93	94,16

Tab. 4.7. Comparaison des taux d'étiquetage correct obtenus avec différents modèles locaux et contextuels, génératifs et discriminants

Ces résultats montrent que pour le critère d'évaluation choisi (taux d'étiquetage correct des pixels), les modèles discriminants (MLP et CRF) permettent d'obtenir de meilleurs résultats d'étiquetage que les modèles génératifs (mélanges gaussiens et MRF). D'autre part les modèles intégrant une information contextuelle (MRF et CRF) sont plus performants que les modèles ne prenant en compte qu'une information locale (mélanges et MLP).

Les résultats que nous avons obtenus vont dans le même sens que les autres travaux effectués dans le domaine, puisque nous avons pu vérifier que les CRF, parce qu'ils sont discriminants, donnent de meilleurs résultats que les autres modèles. En effet, le caractère discriminant des CRF a l'avantage de permettre d'intégrer facilement dans les modèles proposés, différents niveaux d'analyse, sans aucune restriction.

4.5 Conclusion

En considérant les limitations des modèles génératifs pour des tâches clairement discriminantes de segmentation et d'étiquetage d'images, et en considérant les bons résultats apportés par des modèles conditionnels sur des tâches de segmentation de séquences et d'extraction d'information, nous avons proposé un modèle de champ aléatoire conditionnel pour l'étiquetage de données bidimensionnelles, notamment d'images. Ce modèle, qui s'inspire du modèle DRF de Kumar et al. [Kumar 03b], et du modèle CRF 2D de He et al. [He 04], intègre des fonctions de caractéristiques opérant à différents niveaux d'analyse à la fois sur le champ des observations et sur les champs d'étiquettes. Ces fonctions de caractéristiques sont définies sous la forme de classifieurs discriminants prenant leur décision en considérant une partie de l'information. Les sorties de ces classifieurs sont combinés pour fournir une estimation des probabilités marginales en chaque site. Le modèle que nous avons proposé peut donc être vu comme un ensemble de classifieurs coopérants. Nous nous sommes limités dans ces travaux à la combinaison de trois classifieurs, un classifieur local, un classifieur contextuel et un classifieur global, cependant la formulation du modèle dans un cadre discriminant comme celui des CRF, permet d'envisager la combinaison d'un nombre bien plus important de sources de décision, ce qui permettrait sûrement d'améliorer encore le pouvoir discriminant du modèle, et favorisait l'adaptation du modèle à différents problèmes d'analyses. Nous envisageons en perspectives de ces travaux de prendre en compte dans ce modèle plus de caractéristiques comme par exemple des caractéristiques de couleur afin de pouvoir bénéficier de toute l'information que des images couleur peuvent apporter en plus par rapport à des images en niveaux de gris ou binaires. L'utilisation de la couleur est selon nous très importante pour l'analyse de documents fortement bruités et dégradés comme les documents anciens.

L'ajout de caractéristiques contextuelles de plus haut niveau sur l'espace des étiquettes, telles que des caractéristiques de bords ou de jonctions, peut également être envisagé.

Nous avons choisi d'utiliser des Perceptron Multicouches, comme composants discriminants de notre modèle, pour leur rapidité en décision. Cependant l'inconvénient majeur des PMC est lié à la quantité importante de données nécessaires à leur apprentissage et au temps cet apprentissage. Une solution possible pour éviter cet inconvénient des PMC pourrait être de combiner dans le modèle des classifieurs plus simples pouvant être entraînés plus rapidement, tels que des classifieurs linéaires. Le modèle, tel qu'il est défini permet d'utiliser n'importe quel type de classifieur discriminant.

Un des avantages importants de ce modèle, est qu'il peut être entraîné à l'aide de procédures d'apprentissage automatique. Il ne nécessite donc aucun réglage de paramètres. Cela permet une adaptation facile à différents types de documents et à différentes tâches d'étiquetage. Les résultats que nous avons obtenus sur des manuscrits de Flaubert et sur des documents de la Renaissance, bien qu'étant des résultats préliminaires qui doivent encore être approfondis, montrent que le modèle proposé est supérieur à un modèle de champ de Markov cachés. Ces résultats vont dans le sens des résultats présentés dans d'autres travaux sur les champs conditionnels.

Conclusion générale

La numérisation en masse de notre patrimoine, conséquence du développement des technologies numériques ces dernières années, laisse entrevoir de nouvelles possibilités d'accès à la culture et au savoir, mais fait apparaître également de nombreux besoins en termes d'outils d'analyse d'images de documents. En effet, la réalisation d'éditions numériques et de bibliothèques virtuelles permettant l'accès aux sources documentaires numérisées, pose de véritables problèmes de structuration et d'indexation de l'information.

Nous avons abordé ces problèmes dans la première partie de ce mémoire, et nous avons vu que le processus de numérisation ne se résume pas uniquement au problème d'acquisition des images, mais vise également à extraire de ces documents des métadonnées permettant de les indexer et de structurer les vastes corpus de documents numérisés qu'ils constituent. Comme nous l'avons montré en étudiant les différents travaux réalisés dans ce domaine, l'analyse d'images de documents peut apporter des solutions viables à l'indexation des masses de documents, en permettant de déterminer automatiquement les structures et ainsi de localiser l'information nécessaire à leur indexation. Cependant la diversité des documents rencontrés et des tâches d'indexation considérées, fait apparaître de réels besoins en matière de méthodes robustes d'analyse d'images facilement adaptables à différentes tâches et permettant de s'affranchir de la variabilité rencontrée.

Dans la deuxième partie nous avons montré que les problèmes rencontrés en analyse d'images de documents et les solutions proposées dépendent beaucoup de la nature des documents concernés. Ainsi par exemple les problèmes rencontrés en analyse des documents imprimés ne sont pas tout à fait les mêmes que ceux posés par l'analyse des documents manuscrits. La

difficulté posée par l'analyse des documents imprimés se rencontre plutôt dans la reconnaissance de leur structure logique que dans l'extraction de leur structure physique. Cette difficulté est due à la grande variété des structures de ces documents. La segmentation par contre ne pose généralement pas de difficultés car les entités de la structure physique sont souvent naturellement bien séparées, et la variabilité spatiale est peu importante. L'approche adoptée pour l'analyse de ces documents est donc généralement une approche séquentielle qui consiste à déterminer dans un premier temps la structure physique à partir notamment de l'analyse des composantes connexes de l'image ou des espaces, puis dans un deuxième temps à reconnaître la structure logique du document en s'appuyant sur la structure physique extraite et sur la connaissance de règles de mise en page faisant le lien entre les deux types de structures. L'inconvénient de ce type d'approche est que les erreurs commises lors de la segmentation entraînent des erreurs lors de la reconnaissance.
Au contraire, les développements récents effectués dans le domaine de la reconnaissance de l'écrit, ont montré que des approches faisant coopérer les processus de segmentation et de reconnaissance permettent de mieux s'affranchir des problèmes de variabilité rencontrés dans les documents manuscrits. Ces approches sont souvent basées sur l'utilisation de modèles probabilistes et contextuels tels que des modèles markoviens ou des modèles neuro-markoviens faisant coopérer les approches génératives et les approches discriminantes d'analyse d'images. Ces modèles 1D ou pseudo 2D ont rencontré un certain succès en reconnaissance de mots manuscrits ou de lignes de texte.

Nous avons montré que ces modèles peuvent naturellement être étendus aux autres aspects de l'analyse, notamment à l'extraction des structures. En effet le cadre théorique des champs de Markov cachés permet de considérer conjointement la segmentation et la reconnaissance de l'image comme un problème d'étiquetage d'images. Les champs de Markov cachés permettent de résoudre ce problème efficacement en intégrant à la fois des caractéristiques intrinsèques sur l'image et des caractéristiques contextuelles sur les étiquettes affectées, dans un cadre probabiliste qui permet de mieux s'affranchir des problèmes de variabilité. De plus la modélisation par champs markoviens bénéficie de techniques d'inférence et d'apprentissage efficaces.

La contribution apportée par nos travaux réside donc dans l'application des champs de Markov cachés à l'analyse d'images de documents. Nous avons proposé un modèle qui peut s'adapter à différentes tâches d'étiquetage et à différents types de documents en utilisant des techniques d'apprentissage permettant d'apprendre de manière supervisée les paramètres du modèle sur quelques images étiquetées seulement, sans avoir à définir quel que paramètre que ce soit de manière manuelle. L'avantage apporté par une telle procédure d'apprentissage automatique est important pour le traitement de masses de documents, puisque l'étiquetage de quelques pages permet ensuite éventuellement de traiter automatiquement plusieurs milliers de documents. De plus cela permet au modèle de s'adapter à différentes tâches, comme les expérimentations effectuées l'ont montré. Cependant la modélisation par champs de Markov cachés présente un certain nombre de limitations. Ces limitations sont l'hypothèse d'indépendance des observations et la forme générative du modèle d'attache aux données, qui ne permettent pas d'intégrer un grand nombre de caractéristiques dans le modèle et de prendre en compte un contexte très large.

Ces problèmes peuvent être résolus en utilisant des modèles réellement discriminants. Nous avons donc proposé un modèle de champ aléatoire conditionnel 2D, qui permet de combiner plusieurs classifieurs dans un cadre discriminant, pour procéder à l'étiquetage de l'image, en considérant différents types de caractéristiques intervenant à différents niveaux. Cela permet d'améliorer les résultats de l'étiquetage en prenant en compte différents niveaux de contexte. Dans le modèle proposé nous avons considéré trois niveaux : un niveau local prenant en compte des caractéristiques intrinsèques extraites de l'image, un niveau contextuel prenant en compte les incertitudes sur l'étiquetage dans un certain voisinage, et un niveau global permettant d'appréhender la qualité de l'étiquetage à un plus haut niveau en considérant des caractéristiques statistiques extraites sur les étiquettes. L'avantage de ce type de modélisation discriminante est que l'on peut combiner facilement un grand nombre de caractéristiques tant intrinsèques que contextuelles afin d'améliorer le résultat de la décision finale. Le modèle que nous avons proposé peut être vu comme un réseau de classifieurs coopérants. Nous avons choisi pour ces travaux d'utiliser des classieurs de type Perceptron Multicouche, mais l'utilisation d'autres

classifieurs, tels que des classifieurs linéaires ou des SVM est envisageable et constitue une première perspective à ces travaux. Comme pour le modèle de champ markovien précédent, cette modélisation bénéficie de techniques d'inférence et d'apprentissage efficaces. L'adaptation du modèle à différentes tâches d'analyse est donc réalisée de manière automatique. L'apprentissage utilisé est un apprentissage supervisé sur des images étiquetées. Le problème rencontré est que l'apprentissage des PMC nécessite un nombre important de données. Afin de dépasser ces limitations, une autre perspective à cette étude, outre le fait d'utiliser des classifieurs plus simples à entraîner, réside dans l'utilisation de techniques d'apprentissage semi-supervisé combinant apprentissage supervisé et apprentissage non-supervisé. Il s'agirait alors d'effectuer un apprentissage supervisé sur quelques données étiquetées, puis de réaliser ensuite un apprentissage non-supervisé sur des données non étiquetées en alternant étiquetage des données par le modèle et réapprentissage des paramètres sur l'étiquetage obtenu, jusqu'à la convergence du modèle, en se fondant sur les principes de l'algorithme EM.

Les expérimentations que nous avons effectuées avec ce modèle de champ conditionnel, montrent que les résultats sont meilleurs qu'avec le modèle de champ de Markov. Ces résultats sont cohérents avec d'autres travaux effectués dans le domaine. Cependant les résultats que nous avons présentés ne constituent que des résultats préliminaires sur quelques tâches d'étiquetage uniquement. Afin de montrer que la méthode peut effectivement s'adapter à différents niveaux d'analyse et à différents types de documents, nous prévoyons d'effectuer une évaluation sur un plus grand nombre de documents en considérant d'autres tâches d'étiquetage, pas nécessairement à un niveau aussi fin que le pixel. De plus l'approche par champs aléatoires peut également être utilisée sur des données de plus haut niveau comme des graphes, tels que des graphes de primitives ou des graphes de régions par exemple.

Annexe A

Transcription et visualisation de larges corpus de sources littéraires manuscrites

1. Introduction

Nous avons choisi de présenter plus en détails dans cette annexe l'environnement de saisie des transcriptions réalisé dans le cadre du projet Bovary. Cet environnement baptisé EMMA (Editeur de Manuscrits Modernes Assisté) est une première version d'un environnement de travail que nous envisageons enrichir de fonctionnalités multiples pour à terme constituer un véritable poste de travail en philologie des grands corpus. L'expérience acquise lors de la réalisation de la première version de cet outil sera en effet mise à profit prochainement dans le cadre du projet OPTIMA (Outils Pour le Traitement et l'analyse de l'Information dans les Manuscrits modernes) labellisé dans le cadre du programme « corpus et outils de la recherche en sciences humaines et sociales » de l'ANR (Agence Nationale de la Recherche). Dans sa version actuelle, EMMA est un environnement de travail ne disposant d'aucune fonctionnalité de recherche d'information dans les images. Il est clair que les outils présentés dans cette thèse mais également d'autre travaux liés à l'analyse et la reconnaissance de l'écriture manuscrite réalisés au LITIS auraient vocation à pouvoir s'intégrer dans cette plateforme afin de proposer aux utilisateurs, chercheurs en SHS les moyens de rechercher, comparer, annoter, indexer, éditer… les grands corpus numérisés.

Les bibliothèques et musées du monde entier possèdent depuis longtemps des collections de grand intérêt et d'une grande richesse culturelle qui malheureusement ne peuvent être accessibles au grand public pour des raisons de conservation et préservation. Aujourd'hui avec l'essor des technologies numériques, il est enfin possible de valoriser cet extraordinaire patrimoine culturel en proposant des substituts numériques de grande qualité de ces œuvres originales, ce qui permet un partage de l'accès à la connaissance et au savoir, puisque ces documents numériques sont reproductibles à l'infini pour des coûts de réalisation minimes. Ceci permet de plus de conserver les originaux à l'abri de toutes dégradations du temps, ou du fait de manipulations répétées de ces ouvrages. Ainsi ces dernières années de nombreuses campagnes de numérisation ont vu le jour dans les bibliothèques et dans les institutions culturelles publiques ou privées. Cependant avec l'abondance de données numériques ainsi produites, le développement de bibliothèques numériques permettant un accès simplifié et distribué à ces données devient un enjeu majeur pour la valorisation du patrimoine culturel. Or l'utilisation des technologies numériques modifie considérablement nos habitudes documentaires et notre perception du document. Se pose notamment des problèmes d'encodage et d'indexation de l'information. Comment peut-on représenter numériquement des documents manuscrits? Comment produire des représentations numériques et comment les visualiser dans un environnement numérique ? C'est à cette question que nous nous intéressons dans cette annexe, dans le cadre de documents particuliers que sont les manuscrits d'auteurs, qui ont certaines spécificités. Nous abordons en particulier la problématique de la production et de la visualisation de transcriptions diplomatiques de ces documents. Dans un premier temps nous décrivons une interface d'annotation d'images de documents

manuscrits baptisée EMMA (Edition de Manuscrits Modernes Assistée). Cette interface permet la saisie et l'encodage des textes dans un format structuré basé sur la technologie XML. Ce schéma d'encodage permet de fournir une représentation pivot entre une description purement graphique des manuscrits sous forme de fac-similés numériques, et une description purement textuelle du contenu de ces manuscrits.

Dans la seconde partie de cette annexe nous discutons de la restitution et de la visualisation des larges corpus génétiques sous formes de documents hypermédia. La spécificité des documents manuscrits conduit à proposer des outils de transformation des transcriptions saisies qui prennent en compte les informations spatiales du document original afin de permettre la visualisation de la transcription électronique (html ou pdf) le plus conformément possible à son original manuscrit, c'est-à-dire en respectant la diplomatique de la page. Ceci est réalisé au moyen de différents outils et technologies tels que le formalisme SVG et les mécanismes de formatage XSL-FO.
La troisième et dernière partie de cette annexe illustre l'approche proposée dans le cadre de notre contribution au projet Bovary de la Bibliothèque Municipale de Rouen. Nous avons à cette occasion proposé différents scénarii de visualisation et de navigation dans le corpus Bovary en développant un prototype d'édition hypertextuelle de ce corpus, accessible via le Web[1]. Bien qu'appliquées à un corpus spécifique, les techniques mises en œuvre peuvent se généraliser facilement à d'autres œuvres littéraires, et présentent donc un intérêt certain pour la communauté des documentalistes et des chercheurs.

[1] http://www.univ-rouen.fr/psi/BOVARY

2. Un langage de description de manuscrits: le langage Gustave_ML

Le schéma d'encodage que nous proposons pour la description de pages manuscrites, comme les brouillons d'écrivains par exemple, est basé sur XML (eXtensible Markup Language). Ce schéma décrit la structure diplomatique d'un manuscrit et fait le lien entre la description purement textuelle et la description purement graphique par le biais d'un couplage texte/image.

Dans ce schéma d'encodage, l'entité principale est la page ou le folio. Nous cherchons effectivement à produire un couplage entre la représentation physique du manuscrit sous forme d'un fac-similé numérique, et son contenu textuel, par l'intermédiaire d'une représentation structurée intégrant des éléments de la topographique et de la diplomatique du manuscrit. Le schéma que nous proposons n'a pas pour but la structuration d'un dossier de manuscrits complet en terme d'analyse génétique, mais uniquement la structuration de pages manuscrites quelconques en termes de topographie et de spatialisation de l'écriture. En clair, l'objet de l'encodage est la page et non le dossier génétique. Dans notre schéma d'encodage, la page manuscrite est désignée par l'élément racine <transcription>. Cet élément ne possède qu'un attribut titre qui permet d'associer une description à la page, comme par exemple la cote du manuscrit et le numéro du folio. Nous considérons ensuite que l'entité de la mise en page la plus importante dans les manuscrits d'auteurs est le bloc, qu'il s'agisse de blocs de texte, ou d'éléments graphiques. Nous ne nous intéresserons ici qu'aux blocs de texte. La page est donc vue comme un agencement de blocs de texte. Ces blocs peuvent avoir des fonctions différentes dans la mise en page, en fonction de leur positionnement dans la page et des relations spatiales qui existent entre eux. Certains forment le corps du texte alors que d'autres correspondent à des ajouts ou des corrections de l'auteur dans les espaces libres de la page, comme les marges, l'en-tête ou le pied de page. Les blocs de texte sont désignés par l'élément <Bloc>. Chaque bloc est associé à un identifiant unique dans la page, spécifié par l'attribut Num. Les autres attributs de l'élément <Bloc> sont les attributs Type, Attribut et Rotation. L'attribut Type permet de spécifier la nature du bloc, à savoir s'il s'agit d'un bloc appartenant au corps du texte, à l'en-tête, au pied de page, à une marge. L'attribut Rotation permet de préciser si le contenu du bloc est orienté horizontalement (0°), verticalement (à 90° ou à 270°) ou est retourné (180°). L'attribut Attribut permet de spécifier si le bloc est encadré, biffé (recouvert par une grande croix de St André traduisant une suppression), ou les deux à la fois. Un bloc est associé à une zone dans l'image du manuscrit. Cette zone de forme polygonale quelconque est repérée par un ensemble de coordonnées désignées dans le fichier XML par l'élément <coordonnées> et constituées d'éléments <point>. A chaque zone sont également associées les coordonnées du rectangle minimal englobant la zone. Ces coordonnées sont désignées par les éléments <minx>, <miny>, <maxx>, <maxy>. Un bloc de texte est également constitué d'un ensemble de lignes de texte désignées par l'élément

<Ligne>. Cet élément comporte 3 attributs. Un attribut num, qui est un identifiant unique de la ligne dans le bloc. Un attribut Espace qui permet de spécifier la justification de la ligne par rapport au bord gauche du bloc en terme de nombre d'espaces. Enfin un attribut Type qui permet de spécifier la nature de la ligne, à savoir s'il s'agit d'une ligne ou d'un interligne. Chaque ligne est constituée elle-même de fragments textuels représentés par l'élément <texte>. Chaque élément textuel possède un identifiant Num unique dans la ligne et un attribut Type, qui définit si le texte est barré, souligné, illisible ou normal. Ce schéma d'encodage est représenté sur la Figure 1 et la DTD complète est donnée à titre indicatif en Figure 2.

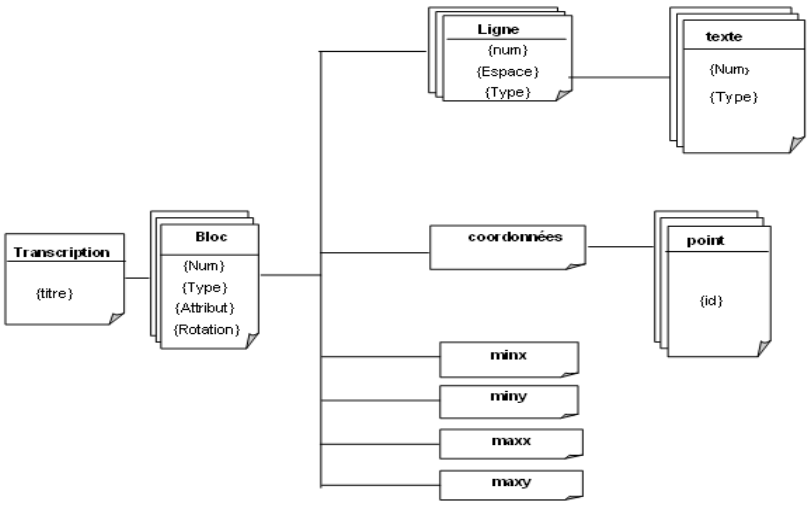


Figure 1. Structure XML d'une transcription

La génération "manuelle" du document XML à l'aide d'un éditeur de texte classique n'étant pas très pratique pour le transcripteur et pouvant potentiellement être source d'erreurs dans la structuration du document, nous avons développé un environnement d'aide à la transcription lui permettant de réaliser plus facilement ces transcriptions dans le langage de description GUSTAVE_ML que nous venons de définir. Cet environnement permet de plus de générer des versions diplomatiques aux formats HTML ou PDF pour permettre une diffusion Web ou classique sur un support papier. L'architecture de cette interface que nous avons baptisée EMMA, pour Edition de Manuscrits Modernes Assistée, est basée sur

l'utilisation de cette représentation pivot en XML du manuscrit, et utilise les outils JDOM et SVG.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!-- DTD GUSTAVE_ML pour la description de la mise en page de manuscripts d'auteurs -->

<!ELEMENT transcription (Bloc+)>
<!-- une transcription est constituée d'au moins un élément "Bloc" -->

<!ELEMENT Bloc (Ligne*,coordonnees,minx,miny,maxx,maxy)>
<!-- un élément "Bloc" est constitué éventuellement d'un ou plusieurs éléments "Ligne"
et de ses coordonnées -->

<!ATTLIST Bloc Num ID #REQUIRED Type (corps de texte | marge | bas de page | haut de
page| Fusion entre marge et corps de texte) #REQUIRED Attribut (Normal | Biffé | Encadré
| Biffé Encadré) #REQUIRED Rotation (0 | 90 | 180 | 270) #REQUIRED>
<!-- l'élément "Bloc" a pour attributs un identifiant unique "Num" et un "Type"
obligatoire défini parmi (corps_de_texte | marge | bas_de_page | haut_de_page) -->

<!ELEMENT Ligne (texte+)>
<!-- un élément "Ligne" est formé d'au moins un élément "texte" -->

<!ATTLIST Ligne Type (Ligne | Inter-ligne) #REQUIRED espace CDATA #REQUIRED num ID
#REQUIRED>
<!-- l'élément "Ligne" a un attribut obligatoire "Type" choisi parmi (Ligne | Inter-
ligne) -->

<!ELEMENT coordonnees (point,point*)>
<!-- les coordonnées sont définies par un ensemble de points -->

<!ELEMENT texte (#PCDATA)
<!-- comprend le texte de l'auteur en lui même -->

<!ATTLIST texte Num CDATA #REQUIRED Type (Normal | Barré | Souligné | Illisible)
#REQUIRED>
<!-- l'élément "texte" a pour attibuts un identifiant obligatoire "Num" et un "Type"
obligatoire choisi parmi (Normal | Barré | Souligné | Illisible) -->

<!ELEMENT point (#PCDATA) >
<!-- l'élément "point" comprend soit un absicse soit un coordonnee  -->
```

Figure 2. Définition de Type de Document Gustave_ML

3. EMMA une interface d'aide à la production de transcriptions diplomatiques

L'interface d'aide à la production de transcriptions diplomatiques EMMA, a été développée en langage JAVA, et utilise les technologies XML. Elle s'appuie notamment sur l'API JDOM. La fenêtre principale de l'interface EMMA, se compose, d'une barre de menus et d'une barre d'outils en haut de la fenêtre, ainsi que de deux sous-fenêtres qui séparent verticalement l'écran en deux parties (Figure 3). L'image du manuscrit s'affiche dans la fenêtre supérieure de l'interface, la fenêtre inférieure étant quant à elle réservée à la saisie de la transcription textuelle, et à l'ajout de métadonnées. La barre de menus et la barre d'outils permettent d'accéder aux fonctionnalités de l'éditeur. Ces fonctionnalités sont l'ouverture de l'image d'un manuscrit, la sélection et la transcription de zones de l'image, la

génération d'une vue destinée à l'impression au format PDF, la génération d'une vue destinée à l'affichage et à la diffusion Web, au format HTML, et l'affichage d'une aide sur l'utilisation de l'interface.

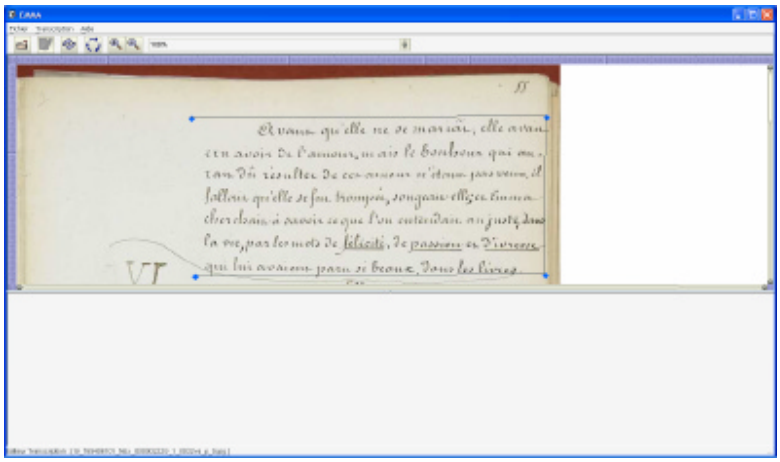


Figure 3. sélection polygonale d'une zone d'intérêt

La barre d'outils offre également des fonctionnalités pour manipuler l'image numérique du manuscrit et permettre ainsi une saisie plus confortable de la transcription. Il s'agit en l'occurrence de la possibilité d'agrandir ou de réduire l'image (zoom) et de retourner l'image de 90°, 180° ou 270° afin de permettre le déchiffrage et la transcription d'éléments retournés ou inscrits verticalement, comme c'est le cas par exemple sur la Figure 4 .

4. Un scénario type de transcription dans l'environnement EMMA

Dans l'éditeur EMMA, la transcription d'un manuscrit se fait par sélection puis annotation de zones d'intérêt dans l'image du manuscrit. Les zones d'intérêt que nous considérons dans la version actuelle de l'éditeur sont uniquement des éléments ou événements textuels, c'est-à-dire principalement des blocs de texte, qu'il s'agisse d'éléments appartenant au corps du texte ou des annotations effectuées en marge, en en-tête ou en pied de page. Un scénario de transcription commence d'abord par l'ouverture de l'image du manuscrit à transcrire. Dans sa version actuelle, l'éditeur accepte uniquement des images au format JPEG. Une fois l'image affichée dans la zone réservée à cet effet, l'utilisateur a donc ensuite la possibilité de sélectionner des zones d'intérêt dans cette image à l'aide de l'outil de sélection, puis de les

annoter. L'utilisateur ne peut sélectionner et annoter qu'une seule zone à la fois. Un type de sélection est proposé à l'heure actuelle dans l'éditeur, il s'agit d'une sélection polygonale. Ce type de sélection n'est certes pas forcément le plus intuitif au premier abord pour l'utilisateur mais il nous a semblé le plus adapté pour des documents

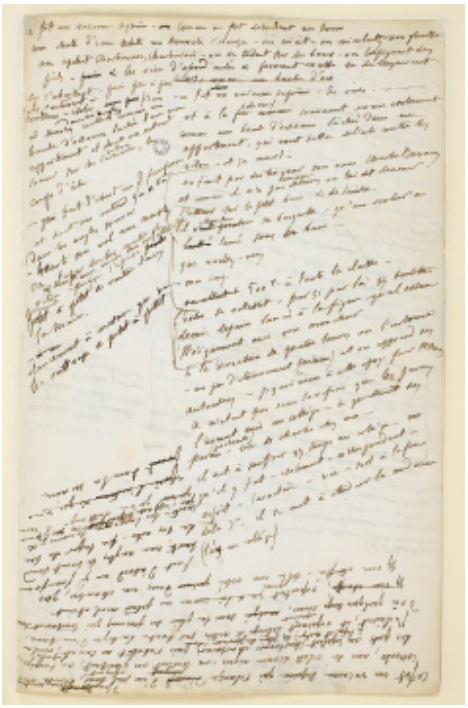


Figure 4. Exemple de manuscrit avec écritures inversées

ayant une mise en page complexe, non régulière et non linéaire, comme c'est le cas pour les documents manuscrits, en particulier les manuscrits d'auteurs, contrairement aux documents imprimés. En effet quoique plus simple d'utilisation, la sélection rectangulaire couramment utilisée dans les autres éditeurs proposés jusqu'à aujourd'hui ne permet pas de définir correctement les contours d'objets inclinés, ou non alignés. L'utilisateur clique sur le bouton "Transcrire" dans la barre d'outils, afin de pouvoir sélectionner une zone d'intérêt, et place à l'aide de la souris les points délimitant la zone qu'il souhaite sélectionner en prenant soin de fermer le contour de la zone en faisant coïncider le dernier point du contour avec le premier. S'ouvre alors une fenêtre permettant de spécifier la nature du bloc ainsi sélectionné. A savoir s'il s'agit d'un bloc appartenant au corps de texte, à l'en-tête,

au pied de page, ou à la marge (Figure 5). Ces métadonnées sur le type de bloc ainsi que les coordonnées de la sélection s'ajoutent automatiquement au fichier XML/GUSTAVE_ML de la transcription. Les coordonnées des blocs sont importantes car elles permettront par la suite la génération automatique de transcriptions diplomatiques en format HTML pour l'édition Web ou en format PDF pour une édition papier.

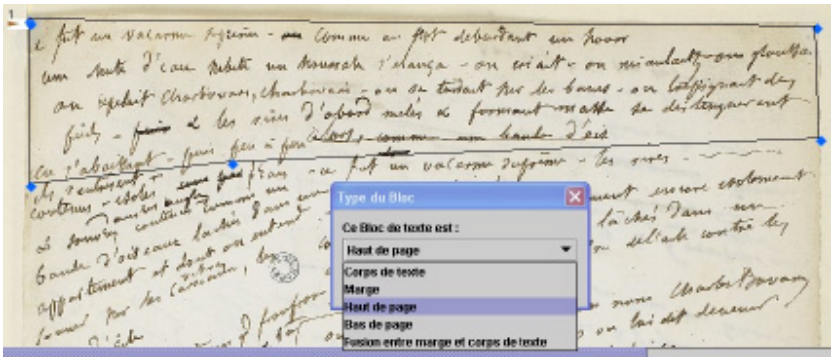


Figure 5. ajout de métadonnées sur le type de bloc

Une fois le type de bloc renseigné, l'utilisateur peut saisir à l'aide du clavier le contenu textuel de la zone sélectionnée dans la zone de saisie prévue à cet effet dans la partie inférieure de l'interface. L'utilisateur saisit également la mise en forme du texte dans le bloc en agençant les mots et les lignes de texte comme dans l'original. La zone de saisie comporte une boîte de boutons radio, permettant de spécifier des métadonnées de mise en forme sur le bloc, les lignes, et les mots transcrits. Ainsi l'utilisateur peut spécifier s'il y a des parties du texte qui sont barrées ou soulignées, s'il s'agit d'une ligne ou d'une interligne ou si le bloc est biffé, encadré ou les deux à la fois. Par exemple si un mot ou un groupe de mots est raturé dans l'original, l'utilisateur doit sélectionner le ou les mots dans la zone de saisie et cliquer sur le radio bouton "Barré" dans la boite "Texte" (Figure 6).

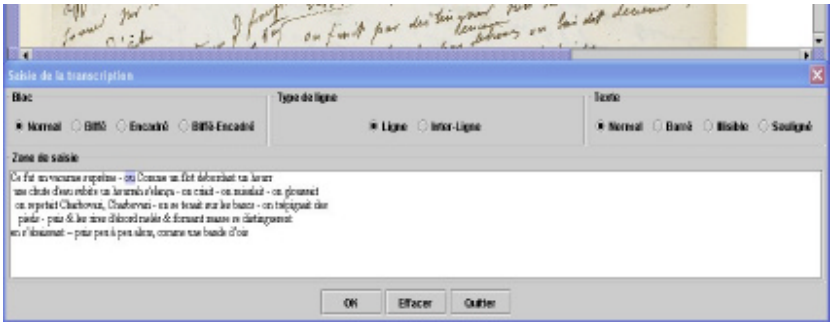


Figure 6. saisie de la transcription d'un bloc et des métadonnées associées

Lorsque l'utilisateur valide la transcription du bloc qu'il vient d'effectuer, le texte et les métadonnées saisies sont sauvegardés automatiquement dans le fichier XML associé au manuscrit, en respectant la DTD Gustave_ML. Dans l'image du manuscrit les blocs déjà transcrits restent encadrés en bleu pour les identifier plus facilement et une puce numérotée apparaît à côté de chaque bloc. L'utilisateur peut modifier la transcription d'un bloc déjà annoté en venant cliquer sur cette puce. Le texte de la transcription apparaît alors dans la zone de saisie pour en permettre la modification. L'utilisateur peut également supprimer un bloc déjà transcrit.
La transcription des blocs marginaux est un peu différente de celle des autres blocs, puisque les blocs marginaux correspondent généralement à des corrections ou à des insertions dans le texte et de ce fait sont donc en rapport avec un passage identifié dans le corps du texte. Du point de vue de la mise en page cela se traduit généralement par un alignement du début du bloc marginal sur la ligne du corps où l'insertion doit se faire. Pour cette raison, et pour permettre de respecter automatiquement cet alignement lors de la génération automatique de la transcription diplomatique, lors de la sélection d'un bloc marginal, l'utilisateur doit spécifier à quel bloc du corps de texte et à quelle ligne de ce bloc se rapporte cette insertion. Pour cela une fenêtre s'ouvre lui permettant de choisir dans une liste de choix le numéro du bloc, puis la ligne à associer (Figure 7).

Figure 7. association d'un bloc marginal à une ligne du corps

Dans le cas de blocs verticaux ou retournés l'utilisateur peut utiliser l'outil de rotation de l'image , situé dans la barre d'outil de l'interface, afin de déclencher la rotation de l'image de 90°. Lors de la validation de la transcription d'un bloc vertical ou retourné, l'information concernant l'angle de rotation sera automatiquement reportée dans le fichier XML comme attribut *Rotation* de l'élément <Bloc>.

5. Génération automatique de transcriptions diplomatiques

L'interface EMMA permet également, lorsque la transcription d'un manuscrit est terminée, ou même en cours de transcription, de générer une vue dite diplomatique de la transcription, c'est-à-dire respectant la topographie de l'original. Cette transcription diplomatique peut être générée dans deux formats. Soit en format HTML afin de permettre sa visualisation à l'aide d'un navigateur Web, ou de permettre l'élaboration d'une édition génétique sur le Web, soit en format PDF afin de permettre l'impression de cette transcription et sa diffusion sur support papier.
Ces versions diplomatiques sont générées dynamiquement à partir de la représentation structurée en XML du manuscrit, en utilisant les technologies XSL (eXtensible Styling Language) de transformation de documents XML, et de formatage XSL-FO (XSL Formating Objects) dans le cas de la génération de la sortie en PDF.
Le langage de feuilles de style XSL est un standard qui permet de transformer et de formater un document XML. Il se compose de deux éléments principaux: un langage de transformation qui permet de transformer un document XML en un autre document structuré, XML ou HTML : c'est le langage XSLT ; le second élément d'XSL est un vocabulaire XML permettant de spécifier des instructions de formatage de bas niveau (blocs de texte, marges, en-tête, pied de page, …), dans un document XML, c'est le langage XSL-FO. Ces deux langages utilisent des règles de production qui comportent une partie condition et une partie action. La première permet la sélection de nœuds dans la représentation arborescente du document XML, en utilisant le langage XPath (XML Path), la seconde permet de modifier cet arbre. Cette seconde partie liée à la modification diffère pour les 2 langages: il

s'agit de production de texte dans le cas de XSLT et de production d'objets de formatage avec XSL-FO [Gardarin 03].
Pour générer une sortie en HTML publiable en ligne sur le Web, nous avons donc défini une feuille de style XSL, permettant de transformer à l'aide d'un processeur XSLT la représentation pivot en XML/Gustave_ML du manuscrit, en une représentation finale en HTML, destinée à la visualisation dans un environnement web (Figure 8). La définition de cette feuille de style consiste à spécifier des règles de transformation des éléments de la structure du document XML/Gustave_ML en des éléments définis dans le langage HTML. Ces règles sont des règles de production qui sont appliquées si certains prédicats sont vérifiés dans l'arbre XML du document source. La transformation par le processeur XSLT consiste à parcourir les nœuds de l'arbre XML du document, et à rechercher les nœuds vérifiant les prédicats définis dans les règles XSL, en faisant de la sélection de nœud à l'aide du langage XPath, puis à appliquer ces règles de production sur les nœuds sélectionnés afin de produire une arborescence HTML.
Dans le cas de la production de transcriptions en format PDF, le processus se déroule en 2 étapes. Une feuille de style XSL est d'abord appliquée au document XML à l'aide du processeur XSLT afin d'y insérer des objets de formatage (Formating Objects) et de produire ainsi un document XSL-FO. Les informations retenues sont en fait extraites du document source XML et marquées avec des balises XSL-FO afin d'indiquer la mise en forme souhaitée. Dans le cas d'un document destiné à une impression papier, cette mise en forme prend en compte les dimensions de la page, les dimensions des marges, la police de caractère utilisée, la taille des fontes, les alignements entre paragraphes, le contenu de l'en-tête et du pied de page, etc. . Cette étape correspond au processus de transformation du document XML.
La 2ème étape est l'étape de mise en forme ou formatage du document XSL-FO obtenu précédemment, à l'aide d'un moteur de formatage, afin d'obtenir un document final adapté au support de publication, dans un format standard comme PDF ou Postscript.

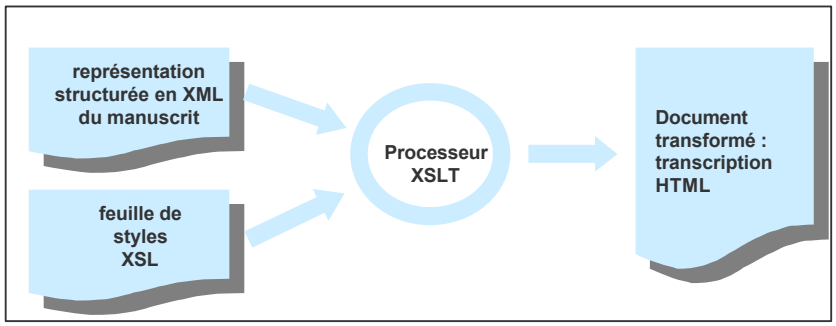


Figure 8. Transformation de la représentation XML d'un manuscrit en une transcription diplomatique HTML

Le moteur de formatage utilisé est le processeur FOP (Formating Object Processor). Le processeur FOP a été développé dans le cadre du projet Apache XML Graphics[2], et il permet à partir d'un document XSL-FO de générer un document au format PDF (format propriétaire Acrobat de la société Adobe). L'ensemble du processus de formatage est représentée sur la Figure 9.

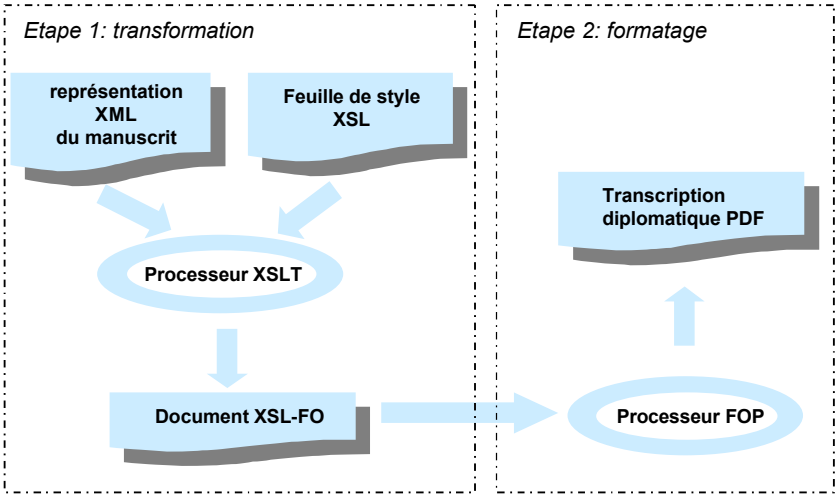


Figure 9. Processus de génération de transcriptions diplomatiques PDF

Pour la génération de transcription diplomatique comportant des blocs verticaux ou retournés, le standard SVG (Scalable Vector Graphic) est utilisé lors de la transformation du document XML par le processeur XSLT, pour transformer tout élément bloc donc l'attribut rotation n'est pas 0°, en une image vectorielle qui est alors insérée dans le document HTML selon le bon angle de rotation. Selon le même principe le standard SVG pourrait être utilisé pour la restitution de symboles graphiques, comme des traits ou des flèches de renvoi. Cependant nous n'avons pas encore envisagé le codage des symboles graphiques dans le langage de description Gustave_ML.

Comme nous l'avons dit précédemment deux feuilles de style XSL ont été définies. La première pour la transformation du document XML en un document HTML utilise les conventions de mise en forme suivantes: les blocs biffés apparaissent sur un fond bleu clair, les éléments raturés sont en bleu également et apparaissent raturés, les éléments interlinéaires sont en italique et en bleu, les éléments illisibles en rouge, et les éléments encadrés apparaissent encadrés. La mise en page est quant à elle réalisée suivant les modèles de la Figure 10. On remarquera en particulier

[2] http://xmlgraphics.apache.org/

que l'on a défini une marge de taille fixe, et que le positionnement des blocs en marge s'effectue relativement par rapport aux lignes d'un bloc du corps de texte comme cela est défini dans la DTD Gustave_ML (Figure 10).
Cette mise en page a été définie de manière spécifique d'après les caractéristiques des manuscrits de Flaubert, néanmoins elle peut être utilisée pour d'autres types de manuscrits, et il est de plus possible de définir une autre forme de mise en page de la feuille de style XSL.

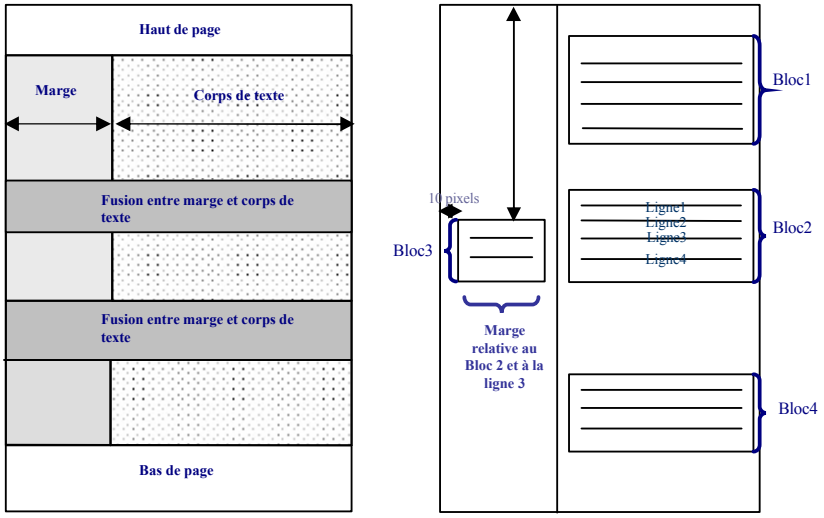


Figure 10. Mises en forme définies par les feuilles de style

La deuxième feuille de style XSL est utilisée pour la transformation du document XML en un document XSL-FO contenant des objets de formatage. Les mêmes conventions de mise en forme du texte et de même mise en page que précédemment sont utilisées.
Nous pouvons voir à titre d'exemple sur la Figure 11, une transcription diplomatique HTML générée avec ces conventions.

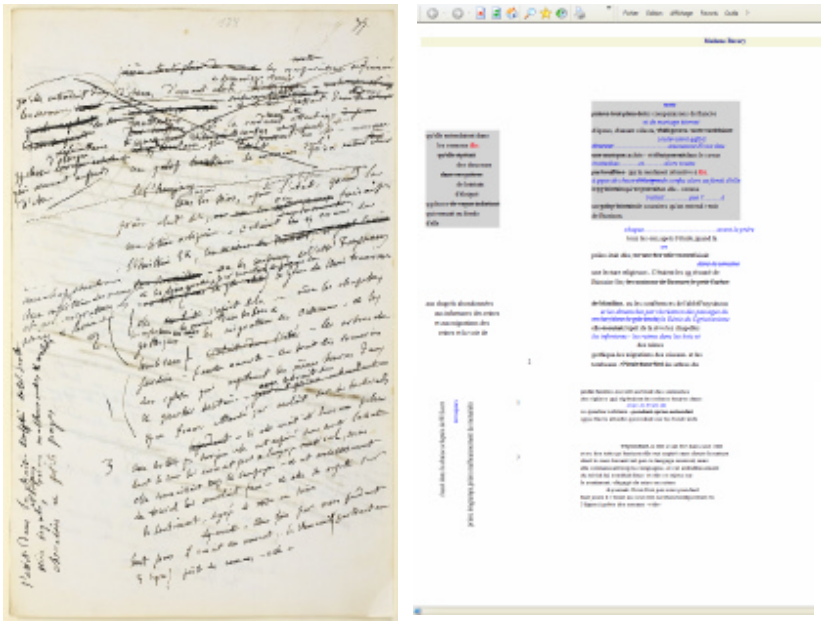


Figure 11. Fac-similé d'un manuscrit de Flaubert et sa transcription diplomatique associée, générée au format HTML par l'éditeur EMMA.

6. Visualisation de sources numérisées sous forme de documents hypertextes

L'objectif du projet Bovary étant la mise en ligne de l'ensemble des brouillons manuscrits du roman "Madame Bovary" de l'écrivain Gustave Flaubert, sous la forme d'une édition Web hypertextuelle donnant accès à l'ensemble des fac-similés numériques ainsi qu'à des transcriptions diplomatiques associées, nous nous sommes intéressés à la problématique de la visualisation de sources numérisées et de la navigation dans un hypertexte génétique. Ces études ont abouti à la réalisation d'un prototype Web accessible à l'adresse suivante: http://www.univ-rouen.fr/psi/BOVARY, au travers duquel sont proposés différents modes de visualisation et de navigation. Ce prototype permet de naviguer dans une séquence génétique de l'avant-texte du roman selon deux axes de temporalité. Soit selon la progression du récit ou axe syntagmatique, à travers l'édition finale du texte, et les états stabilisés des brouillons (brouillons définitifs, version du copiste), ou alors selon la chronologie des opérations d'écriture et de réécriture d'une séquence ou axe paradigmatique, au travers des différents états des brouillons. Pour chaque folio du manuscrit, l'utilisateur a la possibilité de visualiser au choix soit le fac-similé du

folio, soit une transcription textuelle diplomatique associée, ou les deux à la fois (Figure 12). Dans le cas d'une visualisation simultanée du fac-similé et de sa transcription, nous proposons deux modes d'affichages. Un affichage vertical où le fac-similé et sa transcription sont disposés cote à cote en regard, ou un affichage horizontal, où le fac-similé et sa transcription sont disposés l'un en dessous de l'autre (Figure 12). Une fonctionnalité de défilement synchronisé du fac-similé et de sa transcription, ainsi que de re-synchronisation des deux est également proposée. En ce qui concerne la visualisation des fac-similés numériques, des fonctionnalité de grossissement (zoom) sont également proposées afin de permettre par exemple une lecture plus facile de certaines graphies un peu complexes à déchiffrer.

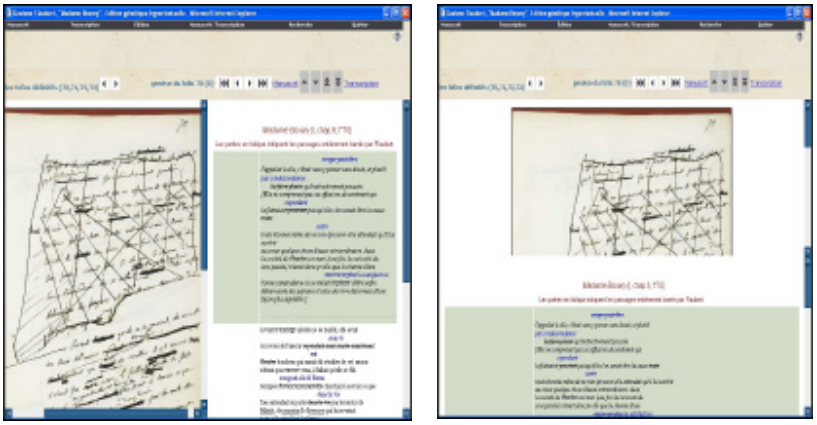


Figure 12. Visualisation de fac-similés et de transcriptions associées

Le format couramment utilisé pour la diffusion d'images sur Internet est le format JPEG, car il permet d'avoir des tailles de fichiers raisonnables, pour une qualité d'image correcte dès lors qu'il s'agit d'images photographiques de dimensions réduites. Ce format a surtout l'avantage de s'être imposé comme standard depuis de nombreuses années et donc d'être très largement reconnue par de nombreuses applications informatiques et navigateurs Internet. Pour cette raison, nous avons également choisi ce format d'image pour les fac-similés numériques des transcriptions. Néanmoins il nous est apparu que ce format s'il est adapté aux images photographiques, l'est beaucoup moins pour des images de documents. De nouveaux formats de compression bien plus efficaces sont apparus sont apparus ces dernières années. Il s'agit des formats JPEG2000 [Santa-Cruz 00] et surtout DjVu [Bottou 00]. Avec JPEG, la qualité de visualisation se dégrade très largement lorsque l'on effectue des agrandissements sur l'image. Or lorsque l'on visualise des documents aussi complexes et difficiles à déchiffrer que les manuscrits d'auteurs, il est primordial de pouvoir obtenir une bonne qualité d'image même lors de forts

agrandissements. Le format DjVu permet d'obtenir cela sur des images de documents, et qui plus est avec de bons taux de compression qui permettent d'aboutir à des tailles de fichiers très raisonnables. De plus le format DjVu offre de nombreux avantages, comme notamment la possibilité d'afficher séparément la couche d'écriture, et le fond de l'image correspondant au support. Pour ces raisons nous pensons que ce format pourrait bien s'imposer ces prochaines années comme standard de diffusion pour les images de documents, et nous avons donc choisi de proposer également dans notre prototype une visualisation des fac-similés numériques des manuscrits dans ce format, même si ce choix nécessite pour l'utilisateur de télécharger un module spécifique DjVu pour son navigateur Web.
En ce qui concerne la visualisation des transcriptions diplomatiques, celle-ci ne pose pas de problème puisque ces transcriptions sont réalisées en HTML, et par conséquent sont directement affichables par n'importe quel navigateur Internet. Ces transcriptions peuvent être générées de manière très simple grâce à l'environnement d'aide à la transcription que nous avons présenté dans la section précédente.

7. Perspectives

Le travail de transcription de manuscrits reste toujours une tâche fastidieuse, surtout lorsqu'il s'agit de produire des transcriptions dites "diplomatiques" qui respectent au maximum l'apparence des documents originaux. L'éditeur que nous proposons a l'avantage de produire des transcriptions dans un format numérique unifié simplifiant l'échange de documents et prenant en compte les spécificités des manuscrits d'auteurs. Cependant dans sa version actuelle il ne permet pas réellement de simplifier le travail du transcripteur même si l'environnement de travail que nous proposons est plus adapté qu'un simple éditeur de texte. Nous prévoyons d'intégrer prochainement à cet environnement de travail sur les manuscrits, des modules d'analyse automatique d'images de documents permettant de détecter automatiquement les zones d'intérêts. C'est pourquoi, en marge du Projet Bovary, nous menons au laboratoire PSI des travaux sur l'analyse automatique de la structure de documents manuscrits complexes, afin de développer des méthodes et des outils permettant une indexation, automatique ou semi supervisée, de ces documents par leurs structures physiques et logiques [Nicolas 03]. A terme nous espérons pouvoir intégrer ces outils dans l'éditeur EMMA, afin de faciliter encore le travail du transcripteur, en proposant une détection interactive des zones d'intérêts, comme les blocs de texte, les zones de ratures ou les zones de surcharge. Nous travaillons notamment sur une méthode de segmentation d'images de documents basée sur une modélisation probabiliste de la connaissance a priori sur la structure du document, par des champs Markoviens [Nicolas 05]. Le laboratoire PSI ayant depuis de nombreuses années une expertise dans le domaine de la reconnaissance de l'écriture, nous pouvons également envisager à plus long terme d'intégrer des modules de reconnaissance de l'écriture. Toutefois la complexité des manuscrits d'auteurs, et les performances des moteurs de reconnaissance ne permettent pas d'envisager une telle intégration à court terme.

Une autre limitation de notre éditeur est que, dans sa version actuelle, il est développé sous forme d'une application Java indépendante, si bien qu'il ne peut pas faire profiter l'utilisateur des avantages du concept de travail collaboratif que permettrait une version distribuée de cet environnement. Nous réfléchissons donc à une version distribuée client-serveur de cet environnement, qui permettrait à un utilisateur de partager son travail et de profiter du travail réalisé par d'autres utilisateurs, de générer à la volée des transcriptions diplomatiques ainsi que de charger ses propres images dans une base de données sur le serveur. En ce qui concerne le langage de description de manuscrits d'auteurs, Gustave_ML, nous envisageons une mise en conformité avec le standard TEI, afin de permettre une plus grande interopérabilité entre les projets existants et futures de bibliothèques numériques [Baird 03] L'édition Web de la genèse de "Madame Bovary" est quant à elle en cours de réalisation par une société extérieure, d'après les spécifications établies par les trois partenaires du projet, le laboratoire PSI, le Centre Flaubert et la Bibliothèque Municipale de Rouen qui a la direction de ce projet. Elle sera disponible en ligne sur le portail Web des bibliothèques de la ville de Rouen (http://bibliotheque.rouen.fr/) courant 2006.

8. Conclusion

Nous avons abordé dans cet article la problématique de l'encodage et de la représentation sous forme numérique des manuscrits d'auteurs, ainsi que de la visualisation de ces transcriptions numériques. Nous avons proposé un langage spécifique basé sur le standard XML, pour l'encodage des manuscrits d'auteurs, le langage Gustave_ML, ainsi qu'un environnement d'aide à la transcription permettant d'annoter et d'encoder facilement les manuscrits en se basant sur un principe de couplage entre l'image du manuscrit et sa transcription textuelle saisie par l'utilisateur. Cet environnement, baptisé EMMA, permet également de produire automatiquement des transcriptions diplomatiques respectant la mise en forme et la topographie du manuscrit original, aux formats HTML et PDF, à partir de la représentation pivot XML saisie par l'utilisateur, en utilisant les standards XSLT, XSL-FO et SVG. Ces transcriptions diplomatiques peuvent ensuite être utilisées pour produire des éditions génétiques Web ou papier. Nous avons également présenté le prototype d'une future édition Web des brouillons du roman "Madame Bovary" de Gustave Flaubert, dans lequel nous mettons en œuvre différents scénarii de visualisation et de parcours de l'avant-texte. Ce prototype, accessible en ligne à l'URL suivante: http://www.univ-rouen.fr/psi/BOVARY, illustre différentes possibilités d'utilisation de la représentation numérique des manuscrits, dans une édition génétique et pédagogique. L'environnement d'aide à la production de transcriptions diplomatiques EMMA est librement téléchargeable à partir de la rubrique "Outils" de ce site web. L'utilisation de ce logiciel est également libre pour quiconque souhaite transcrire des documents. Les outils présentés ne se limitent bien évidemment pas qu'à une utilisation dans le cadre de la transcription des manuscrits de Flaubert, ils sont suffisamment génériques pour utilisés avec différents types de manuscrits. De tels outils présentent un intérêt certain pour les

transcripteurs et les chercheurs qui analysent les textes littéraires, car il permettent de simplifier en partie, des tâches très lourdes et fastidieuses. Ce logiciel est un prototype, qui se veut évolutif. Notre objectif est maintenant de faire évoluer le langage de description Gustave_ML vers une mise en conformité avec TEI et les standards existants, et d'enrichir l'environnement EMMA en proposant une automatisation partielle des traitements de saisie, avec l'espoir qu'un jour une reconnaissance complète de l'écriture soit possible.

Bibliographie

[Abed 05] A. El Abed, V. Eglin, F. Lebourgeois & H. Emptoz. *Frequencies decomposition and partial similarities retrieval for patrimonial handwriting documents compression.* In proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR'05), pages 996–1000, 2005.

[Abuhaiba 96] I.S.I. Abuhaiba, M.J.J. Holt & S. Datta. *Processing of binary images of handwritten text documents.* Pattern Recognition, vol. 29, no. 7, pages 1161–1177, 1996.

[André 99] J. André & Marie-Anne Chabin. "les documents anciens", numéro spécial de la revue document numérique, volume 3. Juin 1999.

[Antonacopoulos 03] A. Antonacopoulos, B. Gatos & D. Karatzas. *ICDAR 2003 Page Segmentation Competition.* In proceedings of 7th International Conference on Document Analysis and Recognition (ICDAR'03), volume 2, pages 688–692, Edinburgh, Scotland, August 2003.

[Antonacopoulos 05] A. Antonacopoulos, B. Gatos & D. Bridson. *ICDAR 2005 Page Segmentation Competition.* In proceedings of 8th International Conference on Document Analysis and Recognition (ICDAR'05), volume 1, pages 75–79, Seoul, South Korea, August 2005.

[Augustin 00] E. Augustin, D. Price & O. Baret. *Reconnaissance de Mots Manuscrits par un Système hybride Mo-*

dèles de Markov Cachés et Réseaux de Neurones. RFIA, vol. 3, pages 365–374, 2000.

[Azokly 95] A.S. Azokly. *Une approche générique pour la reconnaissance de la structure physique de documents composites.* Thèse de Doctorat, Université de Fribourg, 1995.

[Baird 03] H.S. Baird. *Digital libraries and document image analysis (Invited paper).* In Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR'03), pages 2–14, Edinburgh, Scotland, August 4-6 2003. IEEE Computer Society.

[Baird 06a] H.S. Baird & M.R. Casey. *Towards Versatile Document Analysis Systems.* In proceedings of the 7th IAPR International Workshop on Document Analysis Systems (DAS'06), pages 280–290, Nelson, New Zealand, February 2006.

[Baird 06b] H.S. Baird, M.A. Moll, J. Nonnemaker, M.R. Casey & D.L. Delorenzo. *Versatile Document Content Extraction.* In proceedings of SPIE/IS&T Document Recognition & Retrieval XIII Conference, San Jose, CA, January 18-19 2006.

[Belisle 99] C. Belisle & C. Hembise. *Etat de l'art sur les pratiques et sur les usagers des bibliothèques virtuelles, rapport d'activité du projet DEBORA, release n°2.1*, 1999.

[Besag 86] J. E. Besag. *On the statistical analysis of dirty pictures.* Journal of the Royal Statistical Society B, vol. 48, no. 3, pages 259–302, 1986.

[Bottou 00] L. Bottou, P. Haffner, Y. Le Cun, P. Howard & P. Vincent. *DjVu : Un Système de Compression d'Images pour la Distribution Réticulaire de Documents Numérisés, (DjVu : An image compression system for distributing scanned document on the Internet).* In Actes de la Conférence Interna-

tionale Francophone sur l'Ecrit et le Document (CIFED'00), pages 453–462, Lyon, France, July 2000.

[Bouman 94] C. Bouman & M. Shapiro. *A multiscale random field model for Bayesian image segmentation.* IEEE Transactions on Image Processing, vol. 3, no. 2, pages 162–177, March 1994.

[Bouman 97] C.A. Bouman. *Cluster : An unsupervised algorithm for modeling Gaussian mixtures*, April 1997. Available from http ://www.ece.purdue.edu/˜bouman.

[Breuel 02] T.M. Breuel. *Representations and Metrics for Off-Line Handwriting Segmentation.* In proceedings of the 8th International Workshop on Frontiers in Handwriting Recognition (IWFHR'02), pages 428–433, Ontario, Canada, 2002.

[Bruzzone 99] E. Bruzzone & M.C. Coffetti. *An Algorithm for Extracting Cursive Text Lines.* In proceedings of the Fifth International Conference on Document Analysis and Recognition (ICDAR'99), pages 749–752, Bangalore, India, 1999.

[Bunke 03] H. Bunke. *Recognition of Cursive Roman Handwriting Past, Present and Future.* In proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR'03), pages 448–459, Edinburgh, Scotland, 2003.

[Cai 01] J. Cai & Z. Liu. *Pattern recognition using Markov random field models.* Pattern Recognition, vol. 35, no. 3, pages 725–733, 2001.

[Calabretto 98] S. Calabretto & A. Bozzi. *The Philological Workstation BAMBI (Better Access to Manuscripts and Browsing of Images).* International Journal of Digital Information (JoDI), vol. 1, no. 3, pages 1–17, 1998.

[Carreira-Perpignan 05] M.A. Carreira-Perpignan & G.E. Hinton. *On Contrastive Divergence Learning.* In Proceedings of

the Tenth International Workshop on Artificial Intelligence and Statistics (AISTATS'05), pages 33–40, The Savannah Hotel, Barbados, January 6-8 2005.

[Carrera 05] F. Carrera. *Making History : an Emergent System for the Systematic Accrual of Transcriptions of Historic Manuscripts*. In proceedings of the Eighth International Conference on Document Analysis and Recognition (ICDAR'05), pages 543–449, 2005.

[Cattoni 98] R. Cattoni, T. Coianiz, S. Messelodi & C.M. Modena. *Geometric Layout Analysis Techniques for Document Image Understanding : a Review*. Technical Report 9703-09, ITC-IRST, 1998.

[Chellappa 93] R. Chellappa & A. Jain. Markov random fields - theory and application. Boston : Academic Press, 1993.

[Cheng 97] H. Cheng, C.A. Bouman & J.P. Allebach. *Multiscale Document Segmentation*. In IS&T 50th Annual Conference, pages 417–425, Cambridge, MA, USA, May 18-23 1997.

[Cheng 98] H. Cheng & C.A. Bouman. *Trainable Context Model for Multiscale Segmentation*. In IEEE International Conference on Image Processing, volume 1, pages 610–614, Chicago, IL, USA, October 4-7 1998.

[Cheng 01] H. Cheng & C. A. Bouman. *Multiscale Bayesian Segmentation Using a Trainable Context Model*. IEEE Transactions on Image Processing, vol. 10, no. 4, pages 511–525, April 2001.

[Chevalier 03] S. Chevalier, E. Geoffrois & F. Preteux. *A 2D dynamic programming approach for Markov random field-based handwritten character recognition*. In proceedings of the IAPR International Conference on Image and Signal Processing (ICISP'2003), pages 617–630, Agadir, Morocco, 25-27 June 2003.

[Chevalier 04] S. Chevalier. *Reconnaissance d'écriture manuscrite par des techniques markoviennes : une approche bidimensionnelle et générique.* Thèse de Doctorat, Université Paris V - René Descartes, 2004.

[Chevalier 05] S. Chevalier, E. Geoffrois, F. Preteux & M. Lemaitre. *A generic 2D approach of handwriting recognition.* In proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR'05), volume 1, pages 489–493, Seoul, Korea, August 29 - September 1 2005.

[Cho 95] W. Cho, S.W. Lee & H. Kim. *Modeling and Recognition of Cursive Words with Hidden Markov Models.* Pattern Recognition, vol. 28, no. 12, pages 1941–1953, 1995.

[Choisy 00] C. Choisy & A. Belaid. *Analytic Word Recognition Without Segmentation Based on Markov Random Fields.* In proceedings of the Seventh International Workshop on Frontiers in Handwriting Recognition (IWFHR'00), pages 487–492, Amsterdam, The Nederlands, September 2000.

[Chou 90] P. Chou & C. Brown. *The theory and practice of Bayesian image labeling.* vol. 4, pages 185–210, 1990.

[Coüasnon 02] B. Coüasnon & J. Camillerapp. *DMOS, une méthode générique de reconnaissance de documents : évaluation sur 60 000 formulaires du XIXe siècle.* In Actes du Colloque International Francophone sur l'Écrit et le Document (CIFED'02), pages 225–234, Hammamet, Tunisie, Octobre 2002.

[Coüasnon 03a] B. Coüasnon & J. Camillerapp. *Accès par le contenu aux documents manuscrits d'archives numérisés.* Document Numérique, vol. 7, pages 61–84, 2003.

[Coüasnon 03b] B. Coüasnon, J.P. Dalbéra & H. Emptoz. "numérisation et patrimoine", numéro spécial de la revue document numérique, volume 7. 2003.

[Coüasnon 03c] B. Coüasnon & I. Leplumey. *A Generic System for Making Archives Documents Accessible to Public.* In proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR'03), pages 228–232, Edinburgh, UK, August 2003.

[Coüasnon 04] B. Coüasnon, J. Camillerapp & I. Leplumey. *Making Handwritten Archives Documents accessible to Public with a Generic System of Document Image Analysis.* In proceedings of the International Workshop on Document Image Analysis for Libraries (DIAL'04), pages 270–277, Palo Alto, USA, January 2004.

[Cohen 91] E. Cohen, J. J. Hull & S. N. Shrihari. *Understanding handwritten text in a structured environment : determining ZIP codes from addresses.* Character & handwriting recognition expanding frontiers, vol. 30, pages 221–264, 1991.

[Collins 99] D. Collins, W.A Wright & P. Greenway. *The Sowerby Image Database.* In proceedings of the Seventh International Conference on Image Processing And Its Applications, volume 1, pages 306–310, Manchester, UK, July 13-15 1999.

[Conway 93] A. Conway. *Page grammars and page parsing : A syntatic approach to document layout recognition.* In proceedings of International Conference on Document Analysis and Recognition, pages 761–764, Tsukuba Science City, Japan, October 1993.

[Crasson 04] A. Crasson & J.D. Fekete. *Structuration des manuscrits : Du corpus à la région.* In actes du huitième Colloque International Francophone sur

l'Ecrit et le Document (CIFED'04), pages 163–168, La Rochelle, France, juin 2004.

[Côté 98] M. Côté, E. Lecolinet, M. Cheriet & C.Y. Suen. *Automatic reading of cursive scripts using a reading model and perceptual concepts. The PERCEPTO system.* International Journal of Document Analysis and Recognition (IJDAR), vol. 1, no. 1, pages 3–17, 1998.

[Culotta 05] A. Culotta, D. Kulp & A. McCallum. *Gene Prediction with Conditional Random Fields.* Technical Report UM-CS-2005-028, University of Massachusetts, Amherst, 2005.

[Delcourt 00] T. Delcourt. *Document numérique, Les Documents anciens.* Bulletin des Bibliothèques de France, t.45, no. 4, 2000.

[Dengel 92] A. Dengel, R. Bleisinger, R. Hoch, F. Fein & F. Hoenes. *From Paper to Office Document Standard Representation*, July 1992.

[Derras 93] M. Derras. *Segmentation non supervisée d'images texturées par champs de Markov : Application à l'automatisation de l'entretien des espaces naturels.* Thèse de Doctorat, Université Blaise Pascal de Clermont-Ferrand, 1993.

[Do 05] T.M.T. Do, T. Artières & P. Gallinari. *Sélection de modèles par des Méthodes à Noyaux pour la classification de données séquentielles.* In Actes des cinquièmes journées Extraction et Gestion des Connaissances (EGC'2005), Paris, France, 2005.

[Do 06a] T.M.T. Do & T. Artières. *Champs de Markov conditionnels pour le traitement de séquences.* In Actes des sixièmes journées Extraction et Gestion des Connaissances (EGC'2006), pages 639–650, Lille, France, 2006.

[Do 06b] T.M.T. Do & T. Artières. *Conditional Random Fields for Online Handwriting Recognition.* In pro-

ceedings of the Tenth International Workshop on Frontiers in Handwriting Recognition (IWFHR'06), La Baule, France, octobre 2006.

[Durel 00] M. Durel. *Classement et analyse des brouillons de Madame Bovary.* Thèse de Doctorat, Université de Rouen, janvier 2000.

[El-Yacoubi 99] A. El-Yacoubi, M. Gilloux, R. Sabourin & C. Y. Suen. *An HMM Based Approach for Off–line Unconstrained Handwritten Word Modeling and Recognition.* IEEE Trans. on PAMI, vol. 21, no. 8, pages 752–760, 1999.

[Feldbach 01a] M. Feldbach & K. D. Tönnies. *Line Detection and Segmentation in Historical Church Registers.* In proceedings of the Sixth International Conference on Document Analysis and Recognition (ICDAR'01), pages 743–747, Seattle, USA, September 2001.

[Feldbach 01b] M. Feldbach & K. D. Tönnies. *Robust Line Detection in Historical Church Registers.* In proceedings of the 23rd DAGM-Symposium, pages 140–147, September 2001.

[Feldbach 03] M. Feldbach & K.D. Tönnies. *Word Segmentation of Handwritten Dates in Historical Documents by Combining Semantic A-Priori-Knowledge with Local Features.* In proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR'03), pages 333–337, Edinburgh, Scotland, August 2003.

[Feng 06] S. Feng, R. Manmatha & A. McCallum. *Exploring the use of conditional random field models and HMMs for historical handwritten document recognition.* In proceedings of the Second International Conference on Document Image Analysis for Libraries (DIAL'06), pages 30–37, 27-28 April 2006.

[Frey 04] B. Frey & N. Jojic. *Advances in algorithms for inference and learning in complex probability models.* IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), page à paraître, 2004.

[Geman 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, no. 6, pages 721–741, novembre 1984.

[Geoffrois 03] E. Geoffrois. *Multi-dimensional Dynamic Programming for statistical image segmentation and recognition.* In proceedings of the International Conference on Image and Signal Processing (ICISP'03), pages 397–403, 2003.

[Geoffrois 04] E. Geoffrois, S. Chevalier & F. Prêteux. *Programmation dynamique 2D pour la reconnaissance de caractères manuscrits par champs de Markov.* In actes Reconnaissance de Formes et Intelligence Artificielle (RFIA'04), pages 1143–1152, Toulouse, France, Janvier 2004.

[Gilloux 94] M. Gilloux. *Reconnaissance de chiffres manuscrits par modèle de Markov pseudo-2D.* In Actes du 3ème Colloque National sur l'Ecrit et le Document (CNED'94), pages 11–17, Rouen, France, 1994.

[Graffigne 95] C. Graffigne, F. Heitz, P. Pérez, F. Prêteux, M. Sigelle & J. Zerubia. *Hierarchical Markov random field models applied to image analysis : a review.* In proceedings of SPIE Neural Morphological and Stochastic Methods in Image and Signal Processing, volume 2568, pages 2–17, San Diego, CA, USA, 10-11 July 1995.

[Grésillon 94] A. Grésillon. Éléments de critique génétique, lire les manuscrits modernes. Paris, 1994.

[Guillevic 98] D. Guillevic & C.Y. Suen. *Recognition of Legal Amounts on Bank Cheques.* IEEE Trans. on PAMI, vol. 1, no. 1, pages 28–41, 1998.

[Ha 95] J. Ha, R. Haralick & I. T. Phillips. *Recursive XY Cut using Bounding Boxes of Connected Components.* In proceeding of the Third International Conference on Document Analysis and Recognition (ICDAR'95), pages 952–955, 1995.

[Hammersley 71] J. Hammersley & P. Clifford. *Markov fields on finite graphs and lattices.* In unpublished manuscript, 1971.

[Haralick 73] R. Haralick, K. Shanmugan & I. Dinstein. *Textural features for image classification.* IEEE Trans. on SMC, vol. 3, no. 6, pages 610–621, 1973.

[Haralick 79] R. Haralick. *Statistical and structural approaches to textures.* In Proceedings IEEE, volume 67, pages 786–804, 1979.

[Haralick 94] R.M. Haralick. *Document image understanding : geometric and logical layout.* In proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 385–390, Seattle, USA, 1994.

[He 04] X. He, R. S. Zemel & M. A. Carreira-Perpinan. *Multiscale Conditional Random Fields for Image Labeling.* In IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04), volume 2, pages 695–702, 2004.

[Held 97] K. Held, E. Kops, B. Krause, W. Wells, R. Kikinis & H. Muller-Gartner. *Markov random field segmentation of brain MR images.* vol. 16, no. 6, pages 878–886, 1997.

[Heroux 01] P. Heroux. *Contribution au problème de la rétroconversion des documents structurés.* Thèse de Doctorat, Université de Rouen, 2001.

[Higashino 86] J. Higashino, H. Fujisawa, Y. Nakano & M. Ejiri. *A knowledge based segmentation method for document understanding.* In proceedings of the 8th International Conference on Pattern Recognition, pages 745–748, Paris, France, 1986.

[J. Zhu 05] J.R. Wen B. Zhang W.Y. Ma J. Zhu Z. Nie. *2D Conditional Random Fields for Web Information Extraction.* In proceedings of the 22nd International Conference on Machine Learning (ICML 2005), pages 1044–1051, Bonn, Germany, 2005.

[Journet 05] N. Journet, V. Eglin, J.Y. Ramel & R. Mullot. *Text/graphic labelling of ancient printed documents.* In proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR'05), pages 1010–1014, Seoul Olympic Parktel, Seoul, South Korea, August 2005.

[Journet 06a] N. Journet, R. Mullot, V. Eglin & J.Y. Ramel. *Analyse d'images de documents anciens : Catégorisation de contenus par approche texture.* In actes du Neuvième Colloque International Francophone sur l'Ecrit et le Document (CIFED'06), pages 247–252, 2006.

[Journet 06b] N. Journet, R. Mullot, V. Eglin & J.Y. Ramel. *Dedicated texture based tools for characterisation of old books.* In proceedings of the Second International Conference on Document Image Analysis for Libraries (DIAL'06), pages 60–70, Lyon, France, April 2006.

[Kalcheva 03] E. Kalcheva & G. Gluchev. *Segmentation and analysis of handwritten scripts from patients with neurological diseases.* In proceedings of the 4th international conference conference on Computer systems and technologies : e-Learning, pages 272–277, Rousse, Bulgaria, 2003.

[Kam 96] A.C. Kam & G.E. Kopec. *Document Image Decoding by Heuristic Search.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 18, no. 9, pages 945–950, September 1996.

[Kanai 95] J. Kanai, S.V. Rice, T.A. Nartker & G. Naggy. *Automated evaluation of OCR zoning.* IEEE transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 17, pages 86–90, 1995.

[Kim 00] E.Y. Kim, S.H. Park & H.J. Kim. *A genetic algorithm-based segmentation of Markov Random Field modeled images.* IEEE Signal processing letters, vol. 11, no. 7, pages 301–303, 2000.

[Kimura 94] F. Kimura, S. Tsuruoka, Y. Miyake & M. Shridhar. *A Lexicon Directed Algorithm for Recognition of Unconstrained Handwritten Words.* IEICE Trans. Inf. & Syst., vol. E77-D, no. 7, 1994.

[Kirkpatrick 83] S. Kirkpatrick, C. D. Gellatt & M. P. Vecchi. *Optimization by Simulated Annealing.* Science, no. 220, pages 671–680, 1983.

[Kise 98] K. Kise, A. Sato & M. Iwata. *Segmentation of Page Image Using the Area Voronoi Diagram.* Computer Vision and Image Understanding, vol. 70, no. 3, pages 370–382, June 1998.

[Knerr 97] S. Knerr, V. Asimov, O. Baret, N. Gorsky, D. Price & J.C. Simon. *The A2iA intercheque system : Courtesy amount and legal amount recognition for french checks.* In Automatic Bankcheck Processing, pages 43–86. World Scientific, 1997.

[Koch 06] G. Koch. *Catégorisation automatique de documents manuscrits : application aux courriers entrants.* Thèse de Doctorat, Université de Rouen, 2006.

[Kopec 94] G.E. Kopec & P.A. Chou. *Document Image Decoding Using Markov Source Models.* IEEE Transac-

tions on Pattern Analysis and Machine Intelligence, vol. 16, no. 6, pages 602–617, June 1994.

[Krishnamoorthy 93] M. Krishnamoorthy, G. Nagy, S. Seth & M. Viswanathan. *Syntactic Segmentation and Labeling of digitized Pages from Technical Journals*. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 15, no. 7, pages 737–747, 1993.

[Kumar 03a] S. Kumar & M. Hebert. *Discriminative Fields for Modeling Spatial Dependencies in Natural Images*. In proceedings of advances in Neural Information Processing Systems (NIPS), December 2003.

[Kumar 03b] S. Kumar & M. Hebert. *Discriminative random fields : A discriminative framework for contextual interaction in classification*. In proceeding of the 9th IEEE International Conference on Computer Vision (ICCV'03), volume 2, pages 1150–1159, 2003.

[Kumar 06] S. Kumar & M. Hebert. *Discriminative Random Fields*. International Journal of Computer Vision, vol. 68, no. 2, pages 179–202, 2006.

[Lafferty 01] J. Lafferty, F. Pereira & A. McCallum. *Conditional random fields : Probabilistic models for segmenting and labeling sequence data*. In proceedings of the International Conference on Machine Learning (ICML'01), pages 282–289, 2001.

[Lebourgeois 03] F. Lebourgeois, H. Emptoz & E. Trinh. *Compression et accessibilité aux images de documents numérisés : application au projet debora*. Document Numérique, vol. 7, no. 3-4, pages 103–125, 2003.

[Lecolinet 98] E. Lecolinet, L. Likforman-Sulem, L. Robert, F. Role & J-L. Lebrave. *An Integrated Reading and Editing Environment for Scholarly Research on Literary Works and their Handwritten Sources*. In Third ACM Digital Libraries Conference (DL'98), pages 144–151, Pittsburgh, PA, USA, June 1998.

[Lecolinet 99] E. Lecolinet & L. Robert. *Conception d'un poste d'édition et de lecture d'hypermédias littéraires.* Numéro spécial "Numérisation et structuration des documents anciens" de la revue "Document Numérique", vol. 3, no. 1-1, pages 103–117, Juin 1999.

[Lemaitre 06] A. Lemaitre & J. Camillerapp. *Text Line Extraction in Handwritten Document with Kalman Filter Applied on Low Resolution Image.* In proceedings of the Second International Conference on Document Image Analysis for Libraries (DIAL'06), pages 38–45, Lyon, France, April 2006.

[Li 03] S.Z. Li. Modeling image analysis problems using markov random fields, volume 20, chapitre Stochastic Processes : Modeling and Simulation. Elsevier Science, 2003.

[Li 06a] Y. Li, Y. Zheng & D. Doermann. *Detecting Text Line in Handwritten Documents.* In proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), pages 1030–1033, Hong-Kong, Chine, 2006.

[Li 06b] Y. Li, Y. Zheng, D. Doermann & S. Jaeger. *A New Algorithm for Detecting Text Line in Handwritten Documents.* In 10th International Workshop on Frontiers in Handwriting Recognition (IWFHR'06), page (submitted), La Baule, France, 2006.

[Likforman-Sulem 93] L. Likforman-Sulem & C. Faure. *Extracting text lines in handwritten documents by perceptual grouping, une méthode de résolution de conflits d'alignements pour la segmentation des documents manuscrits.* In International Conference on Handwriting and Drawing, pages 192–194, Paris, France, juillet 1993.

[Likforman-Sulem 94a] L. Likforman-Sulem & C. Faure. *Extracting text lines in handwritten documents by perceptual grouping.* In C. Faure, P. Keuss, G. Lorette & A. Win-

ter, editeurs, Advances in handwriting and drawing : a multidisciplinary approach, pages 117–135, Europia, Paris, 1994.

[Likforman-Sulem 94b] L. Likforman-Sulem & C. Faure. *Une méthode de résolution de conflits d'alignements pour la segmentation des documents manuscrits.* In 3ème Colloque National sur l'Ecrit et le Document (CNED'94), pages 265–272, Rouen, France, juillet 1994.

[Likforman-Sulem 95a] L. Likforman-Sulem & C. Faure. *A Hough Based Algorithm for Extracting Text Lines in Handwritten Documents.* In proceedings of the Third International Conference On Document Analysis and Recognition (ICDAR'95), pages 774–777, Montréal, Canada, 1995.

[Likforman-Sulem 95b] L. Likforman-Sulem & C. Faure. *Une méthode de résolution des conflits d'alignements pour la segmentation des documents manuscrits.* Traitement du Signal, vol. 12, pages 541–549, 1995.

[Likforman-Sulem 97] L. Likforman-Sulem, L. Robert, E. Lecolinet & J-L. Lebrave. *Edition hypertextuelle et consultation de manuscrits : le projet Philectre.* Revue Hypertextes et Hypermédias (H2PTM'97), vol. 1, no. 2-3-4, pages 299–310, Septembre 1997.

[Likforman-Sulem 98] L. Likforman-Sulem. *Extraction d'éléments graphiques dans les images de manuscrits.* In actes du Colloque International Francophone sur l'Ecrit et le Document (CIFED'98), pages 223–232, Quebec, Canada, 1998.

[Likforman-Sulem 03] L. Likforman-Sulem. *Apport du traitement des images a la numerisation des documents anciens.* Document Numerique, vol. 7, no. 3-4, pages 13–26, 2003.

[Loncaric 99] S. Loncaric & Z. Majcenic. *Multiresolution Simulated Annealing for Brain Image Analysis.* In procee-

dings of SPIE Medical Imaging, volume 3661, pages 1139–1146, San Diego, USA, 1999.

[Lorette 99] A. Lorette. *Analyse de texture par méthodes markoviennes et par morphologie mathématique : application à l'analyse des zones urbaines sur des images satellitales.* Thèse de Doctorat, INRIA Sophia-Antipolis, 1999.

[Lu 04] Y. Lu, Z. Wang & C. L. Tan. *Word Grouping in Document Images Based on Voronoi Tessellation.* In Document Analysis Systems (DAS'04), pages 147–157, 2004.

[MacCallum 00] A. MacCallum, D. Freitag & F.C.N. Pereira. *Maximum entropy Markov models for information extraction and segmentation.* In proceedings of the 7th International Conference on Machine Learning, pages 591–598, 2000.

[MADONNE 06] Consortium MADONNE. *MADONNE : MAsse de DOnnées issues de la Numérisation du patrimoiNE.* In Atelier ANAGRAM'06, Fribourg, Suisse, Septembre 2006.

[Mahadevan 95] U. Mahadevan & R.C. Nagabushnam. *Gap metrics for word separation in handwritten lines.* In Third International Conference on Document Analysis and Recognition (ICDAR'95), volume 1, pages 124–127, Montréal, Canada, 1995.

[Mao 03] S. Mao, A. Rosenfeld & T. Kanungo. *Document structure analysis algorithms : a literature survey.* In T. Kanungo, E.H. Barney Smith, J. Hu & P.B. Kantor, editeurs, proceedings of SPIE, Document Recognition and Retrieval X, volume 5010, pages 197–207, Santa Clara, CA, January 2003.

[Marti 99] U.-V. Marti & H. Bunke. *A full English sentence database for off-line handwriting recognition.* In proceedings of the 5th International Conference on

Document Analysis and Recognition (ICDAR'99), pages 705–708, Bangalore, India, 1999.

[Marti 01] U.-V. Marti & H. Bunke. *Text Line Segmentation and Word Recognition in a System for General Writer Independent Handwriting Recognition.* proceedings of the Sixth International Conference on Document Analysis and Recognition (ICDAR'01), pages 159–163, September 2001.

[Maître 03] H. Maître. Le traitement des images, (dans la collection traité ic2 : Information-commande-communication, série traitement du signal et de l'image). 2003.

[McDonald] R. McDonald & F. Pereira. *Identifying Gene and Protein Mentions in Text Using Conditional Random Fields.* BMC Bioinformatics, vol. 6, no. (Suppl 1) :S6.

[Murphy 99] K.P. Murphy, Y. Weiss & M.I. Jordan. *Loopy Belief Propagation for Approximate Inference : An Empirical Study.* In proceedings of Uncertainty in AI, pages 467–475, 1999.

[Nagy 84] G. Nagy & S. Seth. *Hierarchical representation of optically scanned documents.* In proceedings of International Conference on Pattern Recognition (ICPR'84), pages 347–349, Montréal, Canada, July 1984.

[Nagy 92] G. Nagy, S. Seth & M. Viswanathan. *A prototype document image analysis system for technical journals.* Computer, vol. 25, pages 10–22, 1992.

[Nagy 00] G. Nagy. *Twenty years of document image analysis in PAMI.* IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 22, no. 1, pages 38–62, 2000.

[Niblack 92] W. Niblack, J. Sheinvald, B. Dom & D. Steele. *Unsupervised image segmentation using the minimum description length principle.* In Proceedings

of the 11th IAPR International Conference on Pattern Recognition (ICPR'92), pages 709–712, The Hague, Netherlands, 1992.

[Nicolas 03] S. Nicolas, T. Paquet & L. Heutte. *Digitizing cultural heritage manuscripts : the Bovary project.* In ACM Symposium on Document Engineering (DocEng'03), pages 55–57, Grenoble, France, 2003.

[Nosary 02] A. Nosary. *Reconnaissance Automatique de Textes Manuscrits pas adaptation du scripteur.* Thèse de Doctorat, Université de Rouen, 2002.

[Ogier 06] J.M. Ogier & K. Tombre. *Madonne : Document image analysis techniques for cultural heritage documents.* In Accepted for presentation at International Conference on Digital Cultural Heritage, Vienna, Austria, September 2006.

[O'Gorman 93] L. O'Gorman. *The document spectrum for page layout analysis.* IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 15, pages 1162–1173, 1993.

[Ouadfel 03] S. Ouadfel & M. Batouche. *MRF-based image segmentation using Ant Colony System.* Electronic Letters on Computer Vision and Image Analysis (LCVIA), vol. 2, no. 1, pages 12–24, August 2003.

[Park 98] Hee-Seon Park & Seong-Whan Lee. *A truly 2-D hidden Markov model for off-line handwritten character recognition.* Pattern Recognition, vol. 31, no. 12, pages 1849–1864, 1998.

[Pasquer 03] L. Pasquer & G. Lorette. *Système de perception et d'interprétation de formes structurées (SPI).* TSI, vol. 22, no. 7-8, pages 879–902, 2003.

[Pearl 88] J. Pearl. Probabilistic reasoning in intelligent systems : Networks of plausible inference. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

[Qi 05] Y. Qi, M. Szummer & T.P. Minka. *Diagram Structure Recognition by Bayesian Conditional Random Fields.* In proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), pages 191–196, San Diego, CA, USA, June 2005.

[Rabiner 89] L. R. Rabiner. *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition.* Proceedings of the IEEE, vol. 77, no. 2, pages 257–286, 1989.

[Ramel 05] J.Y. Ramel & S. Leriche. *Segmentation et analyse interactives documents anciens imprimés.* Traitement du Signal (TS), vol. 22, no. 3, pages 209–222, 2005.

[Ramel 06] J.Y. Ramel, S. Busson & M.L. Demonet. *AGORA : the Interactive Document Image Analysis Tool of the BVH Project.* In proceedings of the Second International Conference on Document Image Analysis for Libraries (DIAL'06), pages 145–155, Lyon, France, April 2006.

[Randriamasy 94] S. Randriamasy & L. Vincent. *Benchmarking Page Segmentation Algorithms.* In proceedings of IEEE Computer Vision and Pattern Recognition (CVPR'94), pages 411–416, Seattle, WA, June 1994.

[Robert 97] L. Robert, L. Likforman-Sulem & E. Lecolinet. *Image and Text Coupling for Creating Electronic Books from Manuscripts.* In Fourth IEEE International Conference on Document Analysis and Recognition (ICDAR'97), Ulm, Germany, August 1997.

[Robert 00] L. Robert & E. Lecolinet. *Techniques d'interaction et de visualisation pour l'accès à des documents numérisés.* In actes de la conférence Ergonomie et Interaction Homme-Machine (ERGO-

IHM'2000), pages 178–185, Biarritz, France, Octobre 2000.

[Saon 97] G. Saon. *Modèles markoviens uni- et bidimensionnels pour la reconnaissance de l'écriture manuscrite hors-ligne.* Thèse de Doctorat, Université Henri Poincaré - Nancy I, Vandoeuvre-lès-Nancy, 1997.

[Sayre 73] K. M. Sayre. *Machine recognition of Handwritten words : A project report.* Pattern Recognition, vol. 5, pages 213–228, 1973.

[Schurmann 92] J. Schurmann, N. Bartneck, T. Bayer, J. Franke, E. Mandler & M. Oberlander. *Document Analysis - From Pixels to Contents.* Proceedings of the IEEE, vol. 80, no. 7, pages 1101–1119, 1992.

[Seni 94] G. Seni & E. Cohen. *External Word Segmentation of Off-Line Handwritten Text Lines.* Pattern Recognition, vol. 27, no. 1, pages 41–52, 1994.

[Sha 03] F. Sha & F. Pereira. *Shallow parsing with conditional random fields.* Technical Report CIS TR MS-CIS-02-35, University of Pennsylvania, 2003.

[Shafait 06] F. Shafait, D. Keysers & T. Breuel. *Pixel-Accurate Representation and Evaluation of Page Segmentation in Document Images.* In proceedings of ICPR 2006, International Conference on Pattern Recognition, volume 1, pages 872–875, Hong Kong, China, August 2006.

[Shapiro 93] Vladimir Shapiro, Georgi Gluhchev & Vassil Stoyanov Sgurev. *Handwritten document image segmentation and analysis.* Pattern Recognition Letters, vol. 14, no. 1, pages 71–78, 1993.

[Shi 04] Z. Shi & V. Govindaraju. *Line Separation for Complex Document Images Using Fuzzy Runlength.* In proceedings of the International Workshop on Document Image Analysis for Libraries (DIAL'04), pages 306–312, Palo Alto, USA, January 23-24 2004.

[Shi 05] Z. Shi, S. Setlur & V. Govindaraju. *Text Extraction from Gray Scale Historical Document Images Using Adaptive Local Connectivity Map.* In proceeding of International Conference on Document Analysis and Recognition, volume 2, pages 794–798, Seoul, South Korea, August 2005.

[Srihari 87] S.N. Srihari, C. Wang, P. Palumbo & J. Hull. *Recognizing Address Blocks on Mail Pieces : Specialized Tools and Problem Solving Architecture.* AI Magazine, vol. 8, no. 4, pages 25–40, 1987.

[Sutton 06] C. Sutton & A. McCallum. Introduction to statistical relational learning, chapitre An Introduction to Conditional Random Fields for Relational Learning. MIT Press, 2006.

[Szummer 04] M. Szummer & 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.

[Szummer 05] M. Szummer. *Learning Diagram Parts with Hidden Random Fields.* In proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR'05), volume 2, pages 1188–1193, Seoul Olympic Parktel, Seoul, South Korea, August 2005.

[Tay 01] Y. H. Tay, P.M Lallican, M. Khalid, C. Viard-Gaudin & S. Knerr. *An Offline Cursive Handwritten Word Recognition System.* Proceedings of IEEE Region 10 Conference, 2001.

[Tsujimoto 92] S. Tsujimoto & H. Asada. *Major components of a Complete Text Reading System*, 1992.

[Varga 05] T. Varga & H. Bunke. *Tree Structure for Word Extraction from Handwritten Text Lines.* In proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR'05),

volume 1, pages 352–356, Seoul Olympic Parktel, Seoul, South Korea, August 2005.

[Vinciarelli 00] A. Vinciarelli & J. Luettin. *Offline cursive script recognition based on continuous density HMM.* IWFHR, pages 493–498, 2000.

[Vinciarelli 04] A. Vinciarelli, S. Bengio & H. Bunke. *Offline Recognition of Unconstrained Handwritten Texts Using HMMs and Statistical Language Models.* IEEE Trans. on PAMI, vol. 26, no. 6, pages 709–720, 2004.

[Virbel 93] J. Virbel. Reading and managing texts on the bibliothèque de france station, pages 31–52. MIT Press, Cambridge, MA, USA, 1993.

[Vishwanathan 06] S. V. N. Vishwanathan, Nicol N. Schraudolph, Mark W. Schmidt & Kevin P. Murphy. *Accelerated training of conditional random fields with stochastic gradient methods.* In ICML '06 : Proceedings of the 23rd international conference on Machine learning, pages 969–976, New York, NY, USA, 2006. ACM Press.

[Wang 88] C. Wang & S.N. Srihari. *A Framework for Object Recognition in a Visual Complex Environment and its Application to Locating Address Blocks on Mail Pieces.* International Journal of Computer Vision, vol. 2, pages 125–151, 1988.

[Wang 00] Q. Wang, R. Zhao, Z. Chi & D.D. Feng. *Hidden Markov Random Field Based Approach for Off-Line Handwritten Chinese Character Recognition.* In proceedings of the 15th International Conference on Pattern Recognition (ICPR'00), volume 2, pages 347–350, 2000.

[Wolf 02] C. Wolf & D. Doermann. *Binarization of Low Quality Text using a Markov Random Field Model.* In Proceedings of the International Conference on Pattern Recognition (ICPR'02), volume 2, pages

160–163, Quebec City, Canada, August 2002. IEEE Computer Society.

[Wong 82] K. Wong, R. Casey & F. Wahl. *Document Analysis system.* IBM J. R D, vol. 26, pages 647–656, 1982.

[Zeng 05] J. Zeng & Z.Q. Liu. *Markov Random Fields for Handwritten Chinese Character Recognition.* In proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR'05), volume 1, pages 101–105, Seoul Olympic Parktel, Seoul, South Korea, August 2005.

[Zheng 03] Y. Zheng, H. Li & D. Doermann. *Text Identification in Noisy Document Images Using Markov Random Field.* In proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR'03), volume 1, pages 599–603, 2003.