Combinaison de données hétérogènes pour la reconnaissance d'images de documents

Aurélie Lemaitre Legargeant

UNIVERSITÉ DE RENNES 1

Habilitation à Diriger des Recherches

Spécialité informatique

Combinaison de données hétérogènes pour la reconnaissance d'images de documents

Présentée devant l'Université de Rennes 1 par

Aurélie Lemaitre Legargeant

Rapporteurs :
Véronique Eglin
Rolf Ingold
Thierry Paquet
Examinateurs :
Bertrand Coüasnon
Guillaume Gravier
Jean-Yves Ramel

Soutenue le 3 novembre 2020

Résumé

Ce manuscrit est une synthèse de mes travaux de recherche depuis 2008, au sein de l'équipe Intuidoc de l'Irisa. Ces travaux portent sur l'analyse automatique d'images de documents numérisés, et plus particulièrement la reconnaissance de la structure de documents. S'il existe des OCR du commerce de bonne qualité, la reconnaissance d'images de documents reste un problème ouvert pour les documents anciens, abîmés, à structure complexe, ou avec une forte interaction entre écriture manuscrite et texte imprimé.

Mes travaux se sont focalisés autour de la méthode DMOS. Il s'agit d'une méthode à base de règles grammaticales, permettant une description physique, syntaxique et sémantique des documents à reconnaître. En particulier, nous avons mis au point le mécanisme de calque perceptif, qui permet une combinaison de données hétérogènes, guidée par une description grammaticale des contenus.

Les données hétérogènes peuvent être des primitives basiques extraites de l'image (segments, composantes connexes), des objets construits (lignes de texte, alignements), des résultats de classifieurs, des données issues d'interactions avec l'utilisateur, des résultats de systèmes de reconnaissance d'écriture, et plus récemment des résultats produits par des systèmes à base d'apprentissage profond. Grâce à l'utilisation des calques perceptifs, ces données sont combinées simplement en suivant des règles symboliques décrivant l'organisation physique et logique des documents.

Dans ce manuscrit, nous présentons l'intérêt de la combinaison de données hétérogènes sur de nombreuses problématiques concrètes. Nous abordons les tâches de segmentation physique des pages : la localisation de lignes de texte et la segmentation en paragraphes. Nous étudions l'analyse de documents à structure complexe comme la presse ancienne, le corpus hétérogène Maurdor, des diagrammes de type flowchart, ou des documents tabulaires d'histoire de la finance. Nous présentons également des résultats sur des formulaires pré-imprimés.

Enfin, nous abordons des perspectives de recherches autour de l'apprentissage avec peu de données, l'utilisation de réseaux de neurones profonds, ainsi que des interactions avec le domaine du traitement automatique de la langue.

Remerciements

Je tiens à remercier les membres de mon jury qui m'ont fait l'honneur d'assister à ma soutenance, ainsi que les trois rapporteurs qui ont évalué mon travail en détails. Les conditions sanitaires ont fait que nous n'avons même pas pu partager un pot, j'espère que nous aurons l'occasion de remédier à cela.

Je remercie l'équipe Intuidoc, pour la bienveillance et la sympathie qui y règnent. J'adresse mes remerciements au responsable de l'équipe, Eric Anquetil, qui, avec la direction de l'Irisa, a toujours facilité mon appartenance à l'équipe, malgré une situation administrative compliquée.

Je remercie particulièrement les doctorants, ingénieurs, stagiaires, post-doc que j'ai encadrés. Merci pour toutes les collaborations fructueuses, les échanges enrichissants.

Enfin, mes remerciements les plus grands vont à mes deux collègues, Bertrand Coüasnon et Jean Camillerapp.

Merci Bertrand de m'avoir embarquée dans tous ces projets, et fait confiance pour réaliser des co-encadrements. J'apprécie notre vision des choses complémentaires, et nos discussions à bâtons rompus sur la manière d'exploiter DMOS au mieux. Après avoir trouvé une application au UNTIL du FIND, nous trancherons peut-être sur l'utilité des opérateurs d'affectation :-). Merci de toujours me pousser en avant, en m'encourageant à prendre de nouveaux engagements.

Merci Jean pour votre avis éclairé. Le fait d'avoir partagé mon bureau avec vous m'a permis de faire progresser ma recherche au quotidien. Merci pour votre disponibilité quand, finalement, je veux bien la fonctionnalité que vous m'aviez proposée une semaine plus tôt mais que j'avais déclinée sur le moment :-) Vous êtes un guide précieux pour ma carrière scientifique.

Table des matières

Chapitre 1

Introduction

Ce manuscrit présente une synthèse de mes travaux de recherche, au sein de l'équipe Intuidoc de l'IRISA, depuis ma soutenance de thèse en 2008. Les travaux présentés ont été réalisés principalement en collaboration avec mes collègues Bertrand Couäsnon et Jean Camillerapp.

1.1 Analyse d'images de documents

Mes travaux portent sur l'analyse automatique d'images de documents numérisés. En partant de l'image d'un document qui a été scanné, l'objectif est de reconnaître l'organisation logique et le contenu textuel de ce document. Cette reconnaissance de tout ou partie des contenus permet alors, selon les besoins, la classification automatique des flux de documents, la localisation de champs d'intérêt dans des documents, ou l'interrogation plein texte de bases de données issues des images.

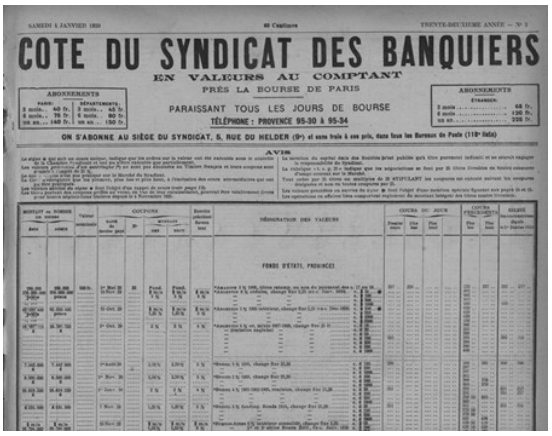
A l'heure actuelle, il existe dans le commerce des systèmes de reconnaissance optique, nommés OCR (Optical Character Recognition). On peut citer par exemple Abbyy Fine Reader, Omnipage, Tesseract. Ces systèmes sont très performants pour la reconnaissance d'images de documents imprimés, de bonne qualité. En revanche, ils présentent des limites lorsqu'il s'agit de traiter des documents anciens, à structure complexe, mal imprimés ou manuscrits (voir figure 1.1).

C'est donc sur ces documents difficiles que se concentrent mes thématiques de recherche. Les verrous scientifiques sont essentiellement les suivants :

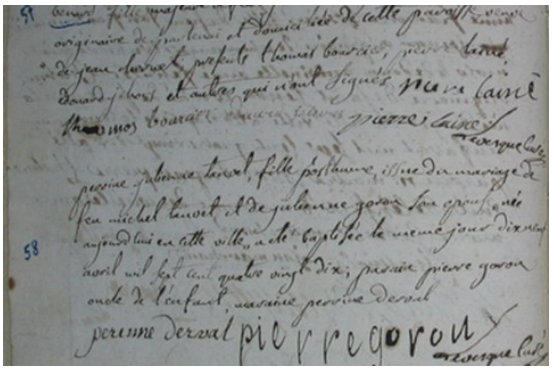
- les problèmes de qualité liés aux documents anciens : dégradations des documents, déchirures, pâleur de l'encre, visibilité du verso par transparence, mauvaise qualité d'impression ;
- la complexité de l'organisation sémantique de certains documents : présence de tableaux avec des niveaux imbriqués de données, ou organisation physique variable pour des contenus similaires ;
- la variabilité de l'écriture manuscrite, qu'un humain novice a parfois des difficultés à déchiffrer sur des documents anciens ;
- la difficulté et le coût d'obtention de bases étiquetées pour lancer des grandes campagnes d'apprentissage automatique.

La reconnaissance d'images de documents se décompose habituellement en deux étapes principales :

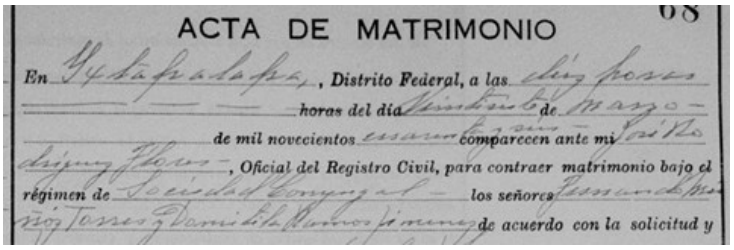
1. l'analyse de la *structure* des documents, pour détecter l'organisation physique et logique des pages, et typer les différents contenus à reconnaître : blocs de texte manuscrit ou imprimé, tableaux, illustrations ...
2. la reconnaissance du *contenu textuel*, manuscrit ou imprimé, pour chaque ligne de texte ayant été identifiée.



(a) Document imprimé à structure tabulaire complexe



(b) Document manuscrit ancien



(c) Formulaire avec superposition entre texte imprimé et écriture manuscrite

FIGURE 1.1 – Exemples de documents pour lesquels la reconnaissance automatique est encore une problématique ouverte

Les contenus textuels ainsi extraits et transcrits peuvent ensuite être utilisés par des systèmes d'analyse de corpus : extraction d'information, construction de bases de données, exploitation dans le domaine des humanités numériques...

Dans ce processus, mes travaux se focalisent principalement sur la première étape de *reconnaissance de la structure des documents.*

1.2 Contexte applicatif

Mes travaux sont directement liés à des besoins applicatifs concrets. En effet, quelle que soit la méthode de reconnaissance choisie, l'adaptation d'un système de reconnaissance à un corpus compliqué est toujours coûteuse. Il est donc nécessaire que le bénéfice d'un système automatique soit plus grand que ne serait le coût d'une transcription manuelle, pour s'engager sur des travaux de reconnaissance.

La figure 1.2 présente une vue chronologique des contextes applicatifs sur lesquels j'ai travaillé de 2009 à 2020. J'ai choisi de regrouper ces applications en quatre thématiques : la reconnaissance de l'organisation physique des documents (en orange sur la figure), l'analyse de documents à structure complexe (en bleu), l'analyse de documents de type formulaires (en jaune), et l'analyse interactive synchrone de documents (en vert). La figure 1.3 illustre la variété des documents analysés.

Historiquement, j'ai focalisé mes travaux de recherche sur la localisation de ligne de texte dans les images. J'ai ainsi travaillé sur les corpus proposés dans les compétitions internationales de reconnaissance : la compétition ICFHR'2010 [GSL10], ICDAR'2013 [Sta+13], et plus récemment le corpus cBAD pour ICDAR'2017 [Die+17] (figure 1.3a). J'ai aussi appliqué la segmentation en lignes de texte sur le corpus de la compétition Rimes [Gro+09] (figure 1.3b). Dans ce corpus,

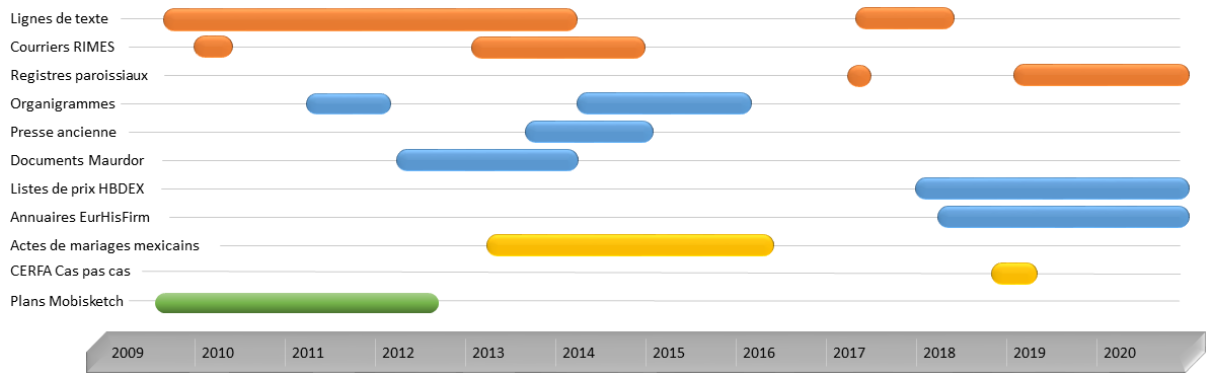


FIGURE 1.2 – Chronologie de l'étude des différents corpus présentés dans la figure 1.4 : travaux sur l'organisation physique des pages en orange, documents à structure complexe en bleu, documents de type formulaires en jaune, analyse interactive synchrone en vert.

l'objectif est de construire des blocs de texte sémantiques dans des courriers manuscrits. Plus récemment, j'ai également travaillé sur la construction de blocs de texte dans des registres paroissiaux (figure 1.3g), en partenariat avec la startup Doptim et les Archives Départementales d'Ille et Vilaine. Il s'agit ici de regrouper les lignes de texte en actes de baptêmes, mariages ou sépultures.

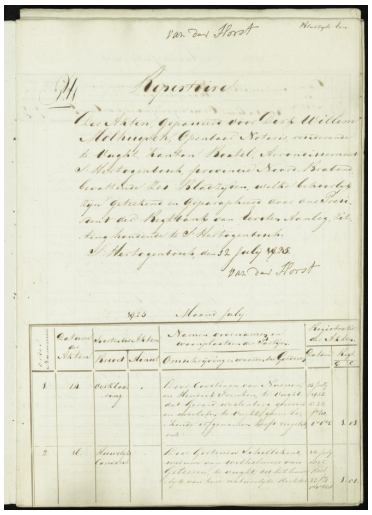
Je me suis également intéressée à la reconnaissance d'images de documents présentant une organisation logique plus complexe. J'ai d'abord travaillé sur la reconnaissance d'images d'organigrammes (figure 1.3j), en collaboration avec les collègues du LS2N de Nantes. Il s'agit, dans ces documents, de comprendre l'enchaînement logique des boîtes reliées par des flèches, respectant certaines contraintes sémantiques. Puis j'ai pris part au projet Maurdor, portant sur la reconnaissance d'images de documents hétérogènes tout venant (figure 1.3d), dans un contexte de renseignement militaire. Dans ces documents, la complexité vient de la très grande variabilité des documents à reconnaître. J'ai travaillé sur un projet de maturation portant sur la reconnaissance de pages de presse ancienne (figure 1.3c). Dans ces documents, il s'agit de comprendre l'organisation physique des journaux en colonnes, puis la structure logique en articles et les illustrations associées. Enfin, depuis 2018, je participe à deux projets portés par des chercheurs en économie, le projet ANR HBDEX, et le projet européen EurHisFirm. Dans ce contexte, l'objectif est d'analyser des documents d'histoire de la finance tels que des journaux de cotations boursières (figure 1.3h) ou des pages d'annuaires d'entreprises (figure 1.3f).

Concernant l'analyse de formulaires, j'ai travaillé sur des registres de mariages mexicains, proposés par FamilySearch, pour la compétition de HIP'2013 [LC13] (figure 1.3e). Dans ces documents, l'objectif est de localiser les champs de date et de lieu de l'acte de mariage. J'ai également mené des premières expérimentations pour l'administration territoriale, dans le cadre du projet IAT, sur la reconnaissance de formulaires de type CERFA (figure 1.3i).

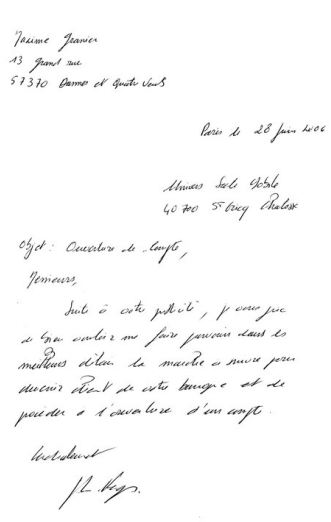
Enfin, dans le cadre du projet ANR Mobisketch, j'ai travaillé sur une analyse de documents interactive, en présence de l'utilisateur qui interagit de manière synchrone. Ce projet a porté sur la reconnaissance de plans d'architecture manuscrits, acquis par un signal en ligne sur tablette (figure 1.3k).

1.3 Combinaison de données hétérogènes

Pour la reconnaissance de structure de documents, différentes familles de systèmes ont été développées dans la littérature. Historiquement, des méthodes statistiques à base, par exemple, de HMM (Hidden Markov Models), concurrencent des approches structurelles à bases de règles grammaticales ou des modèles à base de graphe [DT14]. Plus récemment, les systèmes basés sur les techniques d'apprentissage profond obtiennent de très bon résultats dans l'analyse d'images



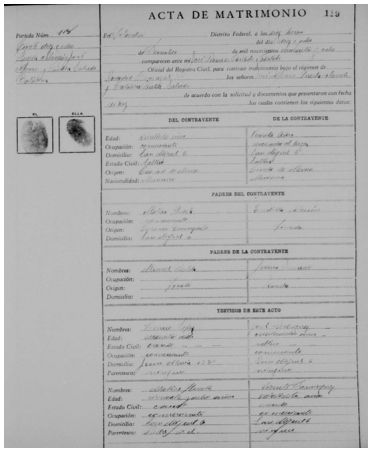
(a) Page de lignes de texte - cBad



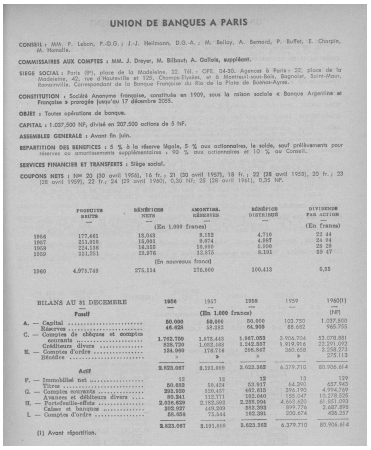
(b) Courriers - Rimes

(c) Presse ancienne

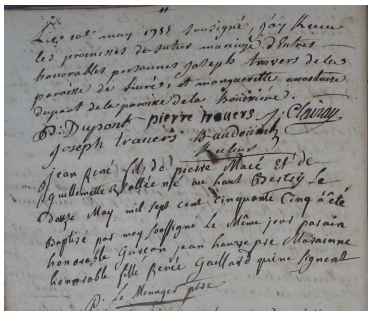
(d) Documents hétérogènes - Maurdor



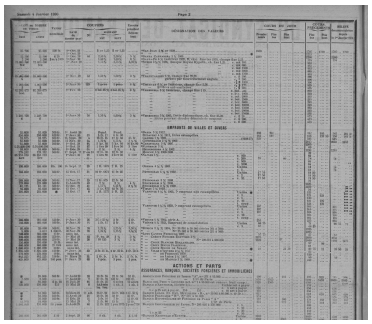
(e) Actes de mariages mexicains - HIP'13



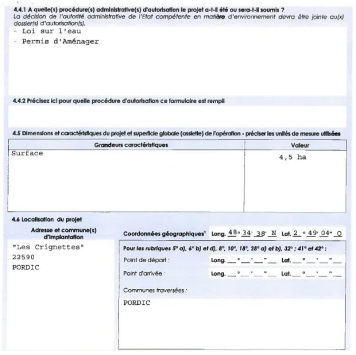
(f) Annuaire d'entreprise - EurHisFirm



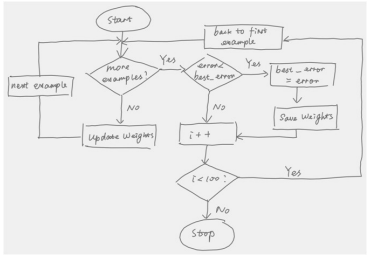
(g) Registre paroissial - Doptim



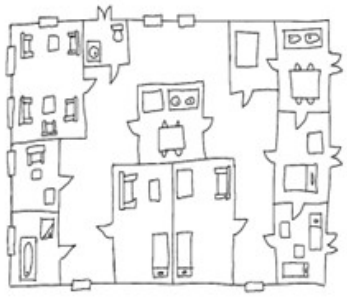
(h) Liste de prix - HBDEX



(i) CERFA - IAT



(j) Organigramme



(k) Plan d'architecture - Mobisketch

FIGURE 1.3 – Exemples de documents sur lesquels ont porté mes travaux de recherche

de documents [Moy+15] [Ren+18] [Alb+19].

Mes travaux se sont placés dans le cadre de l'utilisation de méthodes structurelles, à base de description grammaticale. En effet, je suis convaincue de l'intérêt de pouvoir exprimer de manière claire les connaissances que possède un humain pour pouvoir reconnaître une page de documents. Cela permet d'éviter un effet boite noire.

Lorsqu'un humain analyse une page de documents, il utilise également des connaissances apprises dans d'autres contextes sur l'agencement des éléments, la régularité des lignes, des caractères... La prise de décision est issue d'une combinaison entre une analyse très locale, presque au niveau pixel, et une vision plus globale de l'organisation du document. Il me semble donc intéressant de pouvoir utiliser les systèmes à base de règles pour pouvoir fusionner des données très hétérogènes. Ces données peuvent être issues de différents traitements de l'image : extraction de primitives, combinaison de plusieurs résolutions d'images, utilisation de techniques d'apprentissage, réseaux de neurones profonds... Les règles peuvent ensuite être utilisées pour exprimer la connaissance qui permet une combinaison intelligente des primitives disponibles selon les besoins.

Cette idée de combinaison de données hétérogènes, pour faciliter la reconnaissance de documents, a été le fil directeur de l'ensemble de mes travaux de recherche. La figure 1.4 synthétise les différents types de données hétérogènes étudiés, dans les différents contextes applicatifs. Nous donnons ici les clés de lecture de ce tableau.

La colonne de gauche sur la figure 1.4 présente le contexte applicatif (évoqué dans la partie 1.2) : analyse de courriers manuscrits, de registres anciens, de formulaires, de journaux... Ces diverses problématiques nous ont amenés à envisager des solutions de plus en plus variées, basées sur la combinaison de données hétérogènes.

De manière basique, dans un système à base de règles, on extrait dans l'image des primitives telles que des composantes connexes ou des segments... Ces primitives images sont alors combinées grâce à des règles décrivant l'organisation du document à reconnaître.

Selon ce principe, on combine deux types de données à priori hétérogènes :

— des primitives images, extraites à différentes résolution d'images (colonne rouge du tableau)
— des connaissances à base de règles sur l'organisation physique et logique du document (colonne bleu clair).

Les connaissances à base de règles, représentées dans la colonne bleu clair du tableau, peuvent être également utilisées pour combiner d'autres types de données. Selon les types de documents, nous avons ainsi combiné :

— des objets construits, tels que des lignes de texte ou des alignements (colonne orange),
— des résultats de système d'apprentissage, notamment basés sur des réseaux de neurones profonds (colonne bleu foncé),
— des systèmes de reconnaissance d'écriture manuscrits ou imprimés (colonne violette),
— des résultats d'interaction avec l'utilisateur ou avec des résultats provenant d'autres pages de la collection (colonne verte).

Dans la suite du manuscrit, nous détaillerons l'intérêt des différentes combinaisons entre éléments hétérogènes.

La plupart de mes travaux s'appuient sur la méthode DMOS (Description et MOdification de la Segmentation). Il s'agit d'une méthode grammaticale de reconnaissance de la structure de documents, qui a été développée de longue date dans l'équipe Intuidoc [Coü06]. Dans mes travaux de thèse, j'ai adapté la méthode DMOS existante, pour lui permettre de prendre en compte l'analyse conjointe de plusieurs résolutions d'images, et d'imiter ainsi les mécanismes de la vision perceptive humaine. DMOS est ainsi devenu DMOS-P (pour Perceptive). Depuis ma thèse, j'ai travaillé à étendre ces mécanismes de vision multirésolution, pour utiliser la méthode DMOS-P afin de combiner tous types de données hétérogènes.

FIGURE 1.4 – Synthèse des différents types de données hétérogènes combinées

1.4 Plan du manuscrit

Dans ce rapport, je présente tout d'abord comment il a été possible d'utiliser les mécanismes de vision perceptive présents dans DMOS pour combiner plus largement différents types de données hétérogènes. Le chapitre 2 présente le socle de la méthode DMOS, ainsi que les différentes contributions scientifiques que j'y ai apportées : analyse de données hétérogènes, analyse stochastique, inférence grammaticale.

Les chapitres suivants présentent ensuite différentes problématiques précises, pour lesquelles la combinaison de données hétérogènes a permis de lever des verrous dans le traitement de reconnaissance automatique. J'ai choisi de regrouper les différents documents étudiés selon la typologie de l'analyse de document (voir figure 1.2) : organisation physique des pages, analyse de documents à structure complexe et documents de types formulaires.

Le chapitre 3 montre l'intérêt de la combinaison de données pour la reconnaissance de l'*organisation physique des pages* : segmentation en lignes de texte et découpage en paragraphes. J'évoque notamment dans ce chapitre les compétitions internationales de segmentation de lignes de texte auxquelles nous avons participé, ainsi que nos travaux sur la segmentation en actes de registres paroissiaux.

Le chapitre 4 présente plusieurs corpus de documents à *structure complexe* pour lesquels la combinaison de données hétérogènes permet d'extraire les informations pertinentes. J'aborde notamment dans ce chapitre la reconnaissance d'organigrammes, le découpage de pages de presse ancienne, l'analyse de tableaux de cours de bourse, et la segmentation de documents fortement hétérogènes.

Le chapitre 5 traite de l'analyse de documents de types *formulaires* pré-imprimés. Je présente les travaux que nous avons menés sur l'extraction d'information dans des formulaires CERFA pour l'administration territoriale, ainsi que l'extraction d'information dans des actes de mariages mexicains pré-imprimés.

Dans le chapitre 6, nous aborderons des travaux réalisés dans un contexte légèrement différent : la reconnaissance interactive synchrone de documents, dans le cadre du projet ANR Mobisketch.

Enfin, dans le chapitre 7, je détaillerai mon projet et mes perspectives de recherche.

Chapitre 2

DMOS : de la vision perceptive à la combinaison de données hétérogènes

Dans ce chapitre, nous présentons d'abord la méthode DMOS [1] existante, puis montrons comment elle a été adaptée pour permettre la combinaison de données hétérogènes, avec les calques perceptifs. Nous évoquons ensuite quelques travaux d'extensions qui ont été réalisés sur la méthode DMOS, avant d'évoquer sa diffusion.

2.1 DMOS

La méthode DMOS (Description et MOdification de la Segmentation) a été développée dans l'équipe Intuidoc depuis de nombreuses années [Coü06]. Il s'agit d'un méthode générique qui permet de construire facilement un analyseur de documents, spécifique à une collection donnée. En effet, la méthode DMOS est basée sur un langage de description grammatical. Avec ce langage, il est possible d'exprimer les règles de composition physique et logique d'un corpus de documents à reconnaître. Puis, DMOS dispose d'un compilateur qui permet de générer automatiquement, à partir des règles grammaticales, un analyseur de documents dédié au corpus décrit. La figure 2.1 synthétise cette méthode.

Dans DMOS, la description du contenu des documents d'un corpus est réalisée avec le langage grammatical EPF (Enhanced Position Formalism). Il s'agit d'un langage bi-dimensionnel, qui à chaque étape de l'analyse va indiquer la position dans l'image des éléments à analyser. Ce langage s'appuie sur la présence d'indices visuels dans l'image qui sont considérés comme les terminaux de l'analyse grammaticale.

Par exemple, sur la figure 2.1, la règle EPF en violet exprime qu'une page est constituée d'un entête `E`, et indique que sous cet entête `E` se trouve un corps de page. Dans cet exemple, les indices visuels servant de base à la reconnaissance sont les lignes de texte extraites dans l'image. La description de ce que sont `entête` et `corpsDePage` sera également exprimée en EPF, en se basant sur la présence de lignes de texte.

On note ici que l'aspect bi-dimensionnel de la description est matérialisé par l'instruction `AT(sousEntete E)`. C'est l'opérateur `sousEntete` qui va permettre de définir où se place la suite de l'analyse dans la page. Sur cet exemple, l'opérateur `sousEntete` est placé relativement à l'entête `E` qui a été trouvé. On peut également placer des opérateurs de position de manière absolue, par exemple `AT(hautPage)`, qui se placerait en haut de la page. Une des limitations de la méthode DMOS est que ces opérateurs de position doivent être définis à la main par le concepteur de la grammaire. La difficulté est que la zone de recherche doit être assez grande pour ne pas risquer d'oublier un élément, mais elle doit être assez petite pour restreindre la combinatoire. Cette partie de la description grammaticale nécessite donc la définition intuitive

1. Pour alléger la lecture, on utilisera dans la suite de ce manuscrit uniquement le nom *DMOS* et pas DMOS-P, bien qu'il s'agisse de la version intégrant les mécanismes perceptifs.

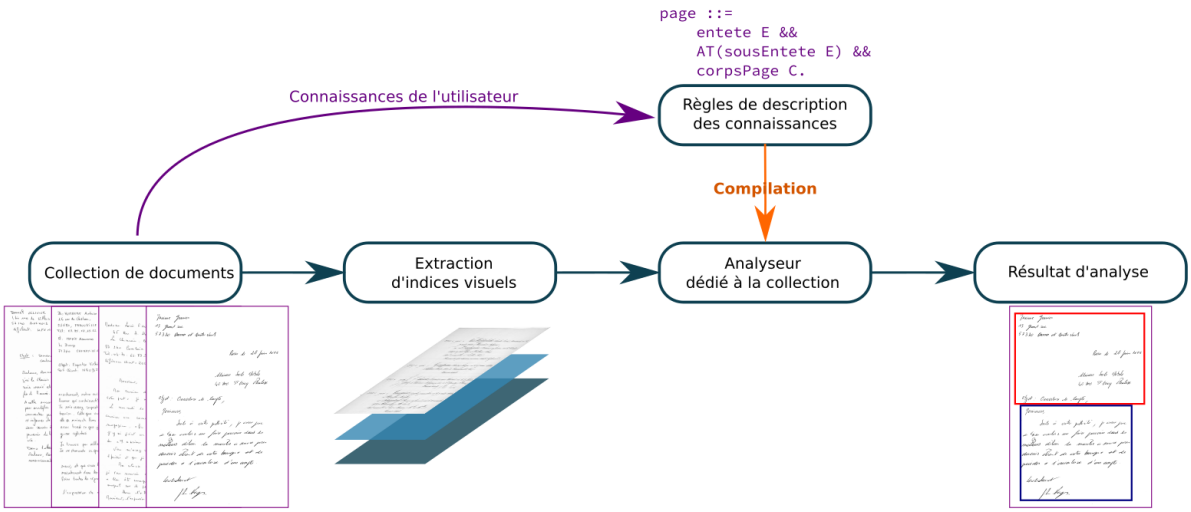


FIGURE 2.1 – Principe de la méthode DMOS : l'utilisateur décrit ses connaissance sur le corpus en langage EPF. Ce code EPF est compilé pour produire un analyseur capable d'interpréter les documents en se basant sur la présence d'indices visuels.

de seuils numériques par le concepteur. Nous verrons dans la partie 2.3.2 comment nous avons travaillé sur ce point.

Une fois la description grammaticale écrite par le concepteur, elle est compilée pour produire automatiquement un analyseur, dédié à la collection. L'implémentation de la méthode DMOS est basée sur des mécanismes de programmation logique (λProlog). Ainsi, les indices visuels sont interprétés par l'analyseur, en trouvant la première combinaison qui satisfait l'ensemble des contraintes décrites dans les règles de grammaire. Au fur et à mesure, les indices visuels utilisés sont consommés, comme des terminaux d'une analyse grammaticale. Si besoin, l'analyseur peut tester plusieurs combinaisons, avec une gestion automatique du retour arrière, avant de trouver une solution qui convient. Si les contraintes ne peuvent être résolues, l'analyse échoue.

Selon les besoins, il est possible de définir des règles grammaticales plus ou moins générales, ou adaptées spécifiquement au corpus. Par exemple, nous avons utilisé la méthode DMOS pour décrire ce qu'est une ligne de texte : cette description peut être utilisée sur des corpus de documents très divers. A l'opposé, nous avons décrit spécifiquement certains documents. Citons par exemple la localisation du champ de coordonnées GPS dans un formulaire CERFA. Cette description est typiquement très spécifique à un besoin précis. La méthode DMOS a ainsi été appliquée dans des contextes très variés, depuis plus de vingt ans : documents d'archives, structures tabulaires, partitions musicales, formules mathématiques, documents hétérogènes...

Pour réaliser des combinaisons de données hétérogènes, nous avons adapté la méthode DMOS pour que les indices visuels pris en entrée de l'analyse puissent être de nature variée.

2.2 Utilisation des calques perceptifs

Dans mes travaux de thèse, j'ai travaillé à la possibilité d'utiliser des indices visuels provenant de plusieurs résolutions d'images. J'ai pour cela proposé la notion de calque perceptif [LCC08]. Ici, nous allons étudier comment les calques perceptifs ont été utilisés pour stocker des données hétérogènes, avant de voir la manière dont ces calques sont exploités dans la description grammaticale, puis comme support à une mémoire visuelle.

2.2.1 Calque perceptif et données hétérogènes

Le calque perceptif est une structure de stockage des indices visuels, représentée en bleue sur la figure 2.1, qui permet de positionner des éléments extraits d'une image dans un référentiel unique. Initialement, chaque calque perceptif était associé aux données extraites dans une résolution de l'image.

Depuis ma thèse, j'ai généralisé l'utilisation des calques perceptifs pour leur permettre de contenir des données hétérogènes. La seule contrainte que doivent avoir ces données est :

— être localisées dans l'image avec des coordonnées d'un référentiel commun,
— être incluses dans un des deux conteneurs existants :
  — un rectangle englobant, représenté par les coordonnées de ses deux coins opposés,
  — un élément linéaire, de type segment, représenté par ses extrémités.

Ainsi, le concept de calque perceptif est à la base des travaux de fusion de données hétérogènes, puisqu'il est possible de stocker des données très variées dans ces calques. On peut notamment citer, parmi les primitives que nous avons utilisées dans différentes applications :

— les composantes connexes de l'image, extraites à plusieurs résolutions,
— les segments de droites, extraits avec un filtre de Kalman, à plusieurs résolutions,
— les indices de lignes de texte extraits par l'analyse d'une image floue,
— les strokes d'un signal en-ligne, et leur type issu d'un résultat de classification,
— des mots clés localisés par word-spotting,
— les sorties de l'OCR Abbyy Fine Reader,
— la carte de probabilité de la présence de lignes de texte, issue d'un apprentissage profond,
— des informations provenant d'autres pages de la collection,
— des données nativement numériques issues de documents pdf,
— des résultats d'interaction avec un utilisateur.

Selon les besoins de chaque application, il est possible de sélectionner l'ensemble des calques pertinents. Chaque calque est vu comme une couche supplémentaire de l'image, contenant des indices visuels à combiner lors de l'analyse.

L'intérêt majeur de la méthode DMOS est que la combinaison de toutes ces données hétérogènes se fait de manière simple, en étant guidé par les connaissances qui sont décrites dans le langage EPF. Le fait d'utiliser la structure de calque perceptif permet de garantir le référentiel commun et donc de simplifier la mise en correspondance des éléments selon leur positionnement dans l'image. Voyons maintenant comment l'utilisation des calques est exprimée dans le langage EPF.

2.2.2 Description grammaticale basée sur les calques perceptifs

Le langage EPF est un langage grammatical bi-dimensionnel. Avec l'introduction des calques perceptifs, ce langage est devenu tri-dimensionnel, puisqu'il permet de préciser, à chaque étape de l'analyse :

— la localisation dans l'image (2 dimensions), avec l'opérateur `AT`,
— le calque perceptif dans lequel l'analyse va se situer (une dimension), avec l'opérateur `USE_LAYER`.

Supposons qu'on analyse un titre, souligné. On dispose d'un calque contenant les lignes de texte, et d'un calque contenant les segments. On peut donc décrire de la manière suivante :

```
titreSouligné ::=
    AT(hautPage)&&
    USE_LAYER(lignesDeTexte) FOR(ligneTitre T) &&
    AT(sousLigneTitre T) &&
    USE_LAYER(segments) FOR(traitSouligné S).
```

On localise l'analyse en haut de la page avec l'opérateur `AT`. Puis, on va piocher dans le calque des lignes de texte, grâce à `USE_LAYER` pour trouver la ligne de titre `T`. Ensuite, on replace

l'analyse sous la ligne de titre, à nouveau avec `AT`, et on se place dans le calques des segments avec `USE_LAYER`, pour consommer un trait de souligné `T`.

Cet exemple illustre la manière dont la combinaison des opérateurs `AT` et `USE_LAYER` offre un grand pouvoir d'expressivité pour la combinaison des données. Ceci s'ajoute à toutes les autres fonctionnalités offertes par le langage EPF, qui sont décrites dans [Coü06].

Les chapitres suivants de ce rapport détailleront des exemples dans lesquels la combinaison très riche de données permet de résoudre des problèmes complexes de reconnaissance, en décrivant de manière simple la manière dont sont combinées les informations.

2.2.3 Calques perceptifs et mémoire visuelle

Les calques perceptifs peuvent servir à combiner des indices visuels variés, mais ils sont aussi le support d'interactions du système avec la collection ou avec un utilisateur. Cet aspect a été étudié en détails dans les travaux de thèse de Joseph Chazalon [Cha13]. Je n'ai pas encadré cette thèse, mais j'ai eu l'occasion de travailler avec Joseph sur l'utilisation des calques perceptifs comme support des mécanismes d'interaction.

Nous avons considéré que pour interagir avec un utilisateur ou avec une collection, il était nécessaire de pouvoir garder une « mémoire » des résultats des analyses des pages précédentes ou de l'analyse en cours. Nous avons mis ainsi au point un système de *mémoire visuelle* qui est en fait basé sur *l'import et l'export de calques perceptifs* lors de l'analyse de pages de documents. Dans le cadre de l'analyse d'une collection, ceci permet notamment de conserver une mémoire de ce qui a été vu dans les pages précédentes.

Par exemple, la figure 2.2 présente une stratégie qui a été implémentée. Il s'agit de document pour lesquels les numéros d'ordre présent dans les pages sont croissants, au travers des pages de la collection. En sortie d'une analyse, le système exporte donc le résultat sous la forme d'une mémoire visuelle. Cette mémoire est alors importée sous la forme d'un calque perceptif pour le traitement de la page suivante.

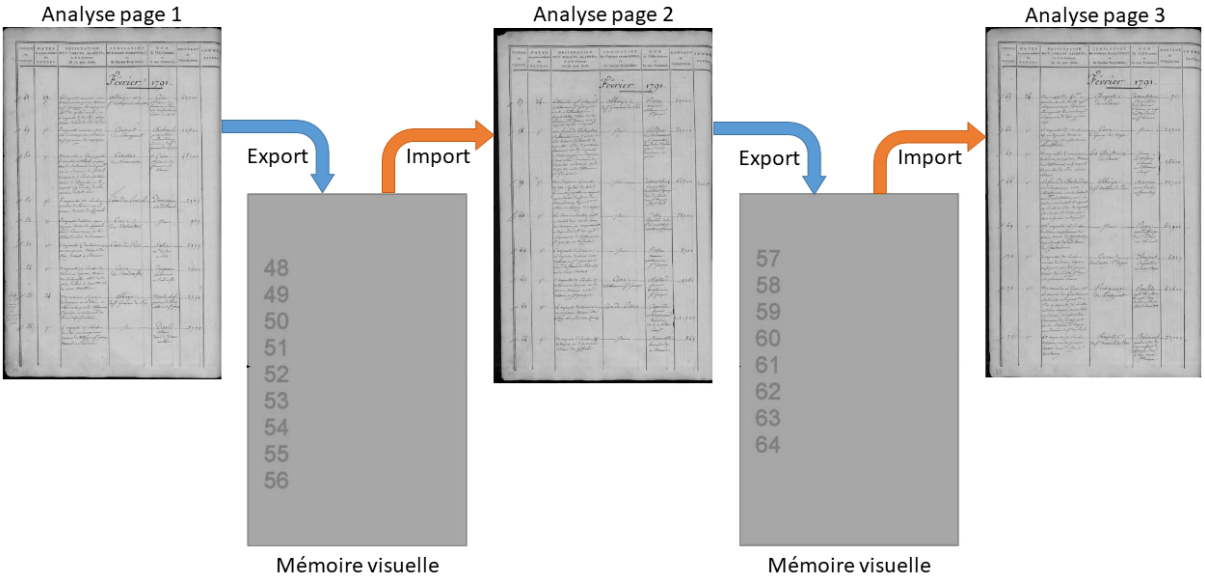


FIGURE 2.2 – Utilisation de la mémoire visuelle pour exploiter le principe de croissance des actes en transmettant les numéros trouvés. Mécanisme basé sur l'export et l'import successifs de calques perceptifs.

Une seconde stratégie testée (présentée sur la figure 2.3) consiste à générer séparément des résultats sur des pages de documents. Puis une étape d'analyse transversale vient fiabiliser un résultat attendu, en fonction des premiers résultats trouvés. Ce résultat est alors injecté, sous la

forme d'un calque perceptif, en entrée d'une seconde passe de traitement de toutes les images. Lors de ce second traitement, les images bénéficieront donc d'une connaissance supplémentaire, inférée par la collection.

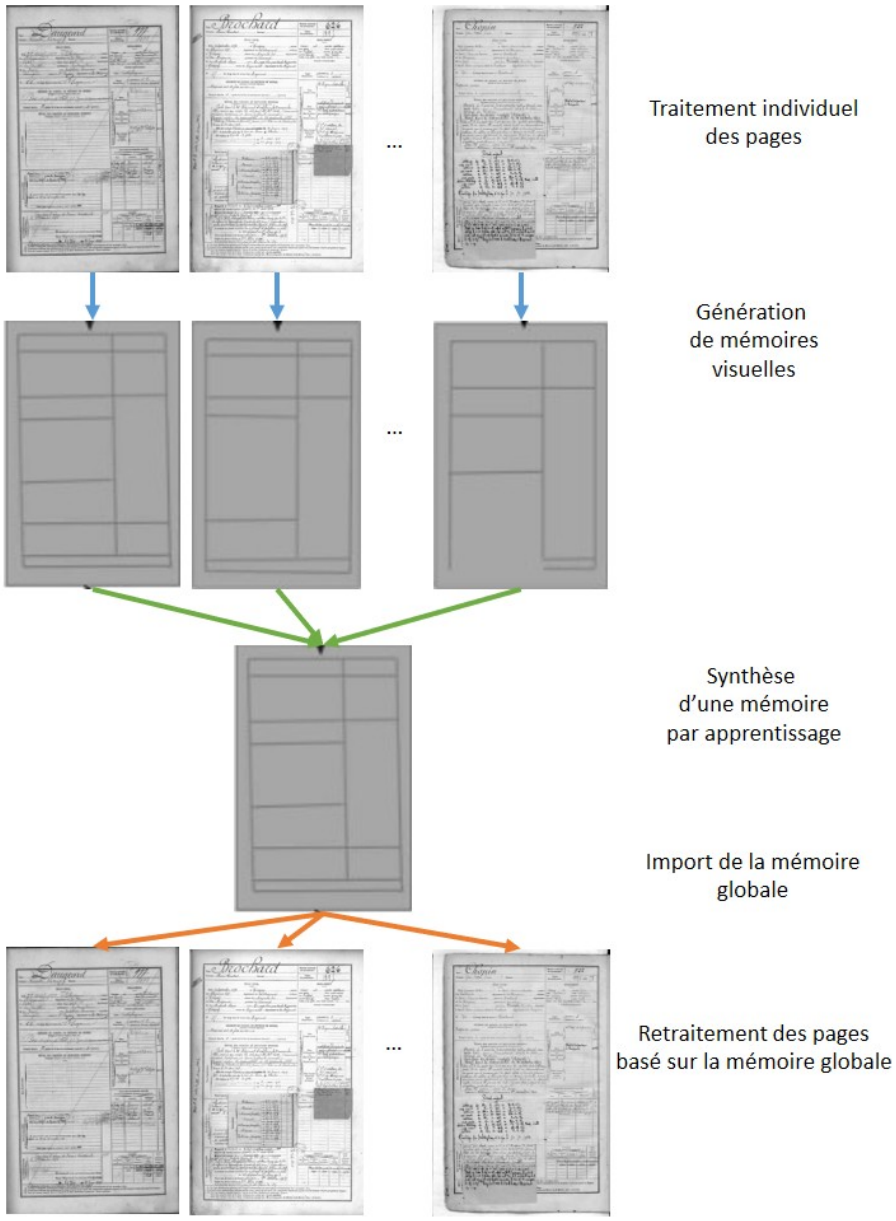


FIGURE 2.3 – Utilisation de la mémoire visuelle pour intégrer de l'apprentissage de dimensions dans la reconnaissance de pages de registres matricules

La mémoire visuelle se base donc sur un export et un import des calques perceptifs. Elle permet ainsi au système d'interagir avec les résultats obtenus sur les autres pages de la collection, ou avec un utilisateur.

Ces travaux ont été publiés, à différents stades d'avancements, dans [CCL12a ; CCL12b ; CCL10 ; CCL11].

Un autre cas d'utilisation de cette interaction sera développé dans la partie 4.4.2.

2.3 Travaux connexes à la méthode DMOS

Pour terminer ce chapitre sur la présentation de la méthode DMOS, nous abordons des sujets connexes. Il s'agit de travaux portant également sur le cœur de la méthode DMOS, qui ont été menés ces dernières années. Je présente en particulier l'intégration d'analyse stochastique dans la méthode DMOS, ainsi que la création d'un module d'inférence grammaticale.

2.3.1 Analyse stochastique

Dans le cadre du stage de master de Andre Oliveira Maroneze, nous avons travaillé à l'amélioration de la méthode DMOS, pour pouvoir prendre en compte des résultats de reconnaisseurs disposant d'un score de probabilité [OCL11].

En effet, par défaut, dans une analyse grammaticale, une règle de grammaire va réussir avec le premier élément trouvé, qui répond aux critères attendus. Cependant, il serait parfois utile de tester, au contraire, tous les éléments compatibles, pour réussir avec l'élément qui répond *le mieux* aux critères attendus. Mais, ceci nécessite une analyse en largeur, qui doit être appliquée avec parcimonie, sous peine de voir une explosion combinatoire.

Dans les travaux de master d'Andre, nous avons proposé un nouvel opérateur dans le langage EPF, `FIND_BEST_FIRST`, qui permet de modifier l'ordre de l'analyse pour une branche donnée, afin de réussir d'abord avec le meilleur élément. Ainsi, cet opérateur permet l'application locale d'une grammaire stochastique.

Ces travaux ont été mis en application pour la reconnaissance de courriers manuscrits, sur la base RIMES (voir partie 3.2.1), ainsi que pour l'analyse de flowcharts (voir partie 4.1).

2.3.2 Inférence de règles grammaticales

Nous avons vu que la méthode DMOS nécessite la définition manuelle de règles de grammaires permettant de décrire les connaissances sur l'organisation physique et logique du document. Parmi ces règles, certaines sont particulièrement fastidieuses à écrire : il s'agit des opérateurs de positions qui permettent d'indiquer, dans l'analyse, où se trouve le prochain terminal à reconnaître. Ces règles nécessitent l'expression arbitraire, en pixels, de dimensions de zones, absolues ou relatives à un élément précédemment trouvé.

C'est sur ce problème que s'est penchée la thèse de Cérès Carton [Car16], que j'ai co-encadrée. Le premier objectif était de proposer une inférence des opérateurs de position, par une analyse statistique d'un corpus étiqueté. Ainsi, elle a proposé le système EWO (Eyes Wide Open) qui permet de générer des opérateurs de position [CLC14a ; CLC16 ; CLC14b]. EWO se base sur le le clustering EAC (Evidence Acculumulation Clustering) [FJ02], un algorithme qui permet une analyse non supervisée d'un corpus étiqueté.

Les travaux de thèse de Cérès sur l'inférence se sont également penchés sur la détection automatique de variantes physiques [CLC15 ; CLC17]. Par exemple, dans un courrier manuscrit, les coordonnées de l'expéditeur sont souvent en haut à gauche. Une analyse exhaustive du contenu du corpus, avec la méthode statistique EWO, peut permettre de montrer l'existence de cas rares où ces coordonnées se situent à droite ou en bas du courrier. Dans une méthode à base de règles telle que DMOS, ces cas rares doivent être connus et décrits. Une analyse statistique de ces configurations, avec la méthode EWO, est donc un plus pour le concepteur de la grammaire.

Pour tous ces travaux d'inférence, la limite est qu'il était nécessaire de disposer d'une vérité terrain étiquetée pour générer des opérateurs de position. La méthode EWO a donc été complétée pour proposer la construction semi-automatique d'une vérité terrain. Le principe est de réaliser une analyse du corpus en se focalisant sur les redondances dans un grand volume de documents non annotés. La détection des redondances est faite automatiquement grâce à un algorithme de clustering. Les éléments détectés automatiquement sont ensuite fiabilisés par l'utilisateur afin d'obtenir les données étiquetées d'apprentissage. L'étiquetage ainsi produit est nommé « pseudo-

vérité terrain », puisqu'il n'est pas totalement vérifié, mais il est suffisamment fiable pour servir de support à l'inférence grammaticale.

La méthode EWO peut donc être utilisée d'une part pour générer une pseudo-vérité terrain, et d'autre part pour inférer des règles de description du corpus. La méthode EWO apporte une vision exhaustive et synthétique des données à analyser. Cela permet une meilleure exploitation du corpus que pour les méthodes syntaxiques décrites manuellement, avec une meilleure gestion des cas rares.

La méthode a été validée sur plusieurs corpus. Nous évoquerons des exemples d'application de la méthode EWO sur le corpus RIMES (partie 3.2.1), ainsi que sur le corpus des mariages mexicains (partie 5.1).

2.4 Diffusion de la méthode DMOS

Depuis 2014, nous avons travaillé à faciliter la diffusion de la méthode DMOS. En effet, il s'agit d'un système complet, et complexe, dans une machine virtuelle, comprenant un environnement de développement sous Eclipse. J'ai notamment rédigé un tutoriel, permettant de prendre en main l'environnement de développement de la méthode DMOS. Ce tutoriel a permis aux nouveaux arrivants dans l'équipe d'appréhender plus rapidement la méthode DMOS.

J'ai également rédigé un manuel utilisateur interne, le "DMOS-book", qui recense les fonctionnalités de la méthode DMOS. L'existence de ce manuel, ainsi que du tutoriel, a permis de commencer à envisager de transférer la méthode DMOS à des universitaires, dans des buts de recherche. Nous avons ainsi travaillé en coopération avec Harold Mouchère et Geoffrey Roman-Jimenez du LS2N pour qu'ils puissent utiliser la méthode DMOS dans leur projet CIRESFI [Rom+18].

Nous avons été invités pour animer un tutoriel sur la méthode DMOS lors de l'atelier SIFED à Nantes, en juin 2018 [LC17]. Nous avons également publié sur ce thème lors de l'atelier ICDAR-OST [CL17].

2.5 Conclusion

La méthode DMOS permet une génération simple d'un analyseur dédié à un corpus de documents. Nous avons mis au point la notion de calque perceptif qui permet de prendre en compte, en entrée de l'analyse, des indices visuels de nature hétérogène.

La description d'un document, réalisée en langage EPF, est maintenant tri-dimensionnelle : à chaque étape de la description, l'utilisateur précise l'emplacement d'analyse dans l'image, mais aussi le calque permettant de choisir les indices visuels utilisés. La combinaison des indices visuels est facilitée grâce à l'utilisation d'un référentiel unique.

Nous allons maintenant présenter des exemples concrets d'utilisation pour démontrer la faisabilité et l'intérêt de la combinaison des données hétérogènes. Ces cas concrets seront aussi l'occasion de valider les développements supplémentaires autour de DMOS : l'analyse stochastique avec `FIND_BEST_FIRST` ainsi que la méthode EWO.

Chapitre 3

Reconnaissance de l'organisation physique de pages

Dans ce chapitre, nous nous focalisons sur deux tâches « basiques » de l'analyse de structure de documents : la segmentation en lignes de texte et l'organisation de ces lignes de texte en paragraphes. Nous avons abordé ces travaux sous l'angle des systèmes à base de règles.

Nous allons montrer dans ce chapitre que nous avons utilisé la combinaison de données hétérogènes de plus en plus élaborées pour cette analyse. Ainsi, la construction des lignes de texte se base sur la présence de composantes connexes, puis sur l'analyse d'une image floue, et enfin sur le résultat d'un réseau profond. L'analyse des paragraphes se base sur une combinaison de lignes de texte, mais également d'indices issus de réseaux de neurones profonds tels que les signatures ou les bords de page.

3.1 Localisation de lignes de texte

La localisation de lignes de texte est une tâche essentielle dans les mécanismes de reconnaissance de documents. En effet, les systèmes de reconnaissance optique de caractères s'appliquent généralement sur des imagettes de lignes de texte. La segmentation de documents en lignes de texte peut sembler un problème simple, mais c'est un challenge encore ouvert dans le cas de documents manuscrits ou à structure complexe. Il existe notamment de nombreuses compétitions internationales qui portent sur la localisation de lignes de texte.

Dans cette partie, nous montrons comment la détection des lignes de texte peut être favorisée par la combinaison de différentes primitives images, utilisées à plusieurs résolutions, guidées par des règles symboliques, puis par l'utilisation de résultats d'apprentissage (figure 3.1).

	Primitives images			Règles symboliques	Résultats de systèmes d'apprentissage
	Segments	Composantes connexes	Indices de texte dans l'image floue	Organisation physique	Réseaux de neurones profonds
Lignes de texte	✓	✓	✓	✓	✓

FIGURE 3.1 – Données hétérogènes combinées pour la reconnaissance de lignes de texte

3.1.1 Vision perceptive des lignes de texte

Dans la continuité de ma thèse, les premiers travaux que j'ai réalisés pour la détection de lignes de texte sont inspirés des mécanismes de la vision perceptive humaine. On considère ainsi que, vue de loin, une ligne de texte peut être perçue comme un segment de droite, et qu'une vision de près permet de se focaliser sur les caractères.

Ce constat sur le fonctionnement humain peut être modélisé de manière simple grâce aux mécanismes de calques perceptifs de DMOS. La figure 3.2 présente les données utilisées pour cette modélisation de vision perceptive. Deux calques perceptifs sont utilisés : les segments horizontaux, extraits par un filtrage de Kalman à basse résolution, ainsi que les composantes connexes extraites dans l'image à haute résolution. Les connaissances exprimées dans le langage EPF permettent de guider la combinaison des données : une ligne de texte est vue de de loin comme une segment, et de près comme un ensemble de composantes connexes. Nos travaux permettent notamment de localiser le corps du mot, et de resegmenter des composantes connexes qui correspondent à des lettres qui se touchent entre deux lignes.

(a) Indices visuels : segments horizontaux

(b) Indices visuels : composantes connexes

(c) Résultat : segmentation en lignes de texte, attribution de chaque pixel à la bonne ligne

FIGURE 3.2 – Combinaison de données hétérogènes pour la vision perceptive de lignes de texte. Exemple sur un image du concours ICFHR 2010

Ces travaux nous ont permis de participer à la compétition ICFHR 2010 [GSL10]. Les résultats sont présentés sur la figure 3.3. Ces travaux ont été publiés dans [LCC11].

1. **CUBS** (*FM*=97,63%)
2. **NifiSoft-a** (*FM*=97,40%)
3. **NifiSoft-b** (*FM*=97,40%)
4. **IRISA** (*FM*=96,66%)
5. **ILSP-a** (*FM*=95,40%)
6. **ILSP-b** (*FM*=94,95%)
7. **TEI** (*FM*=94,86%)

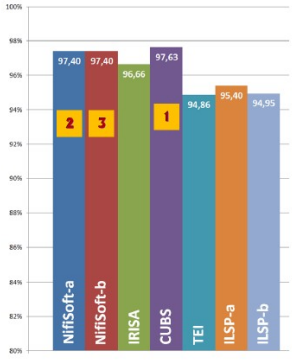


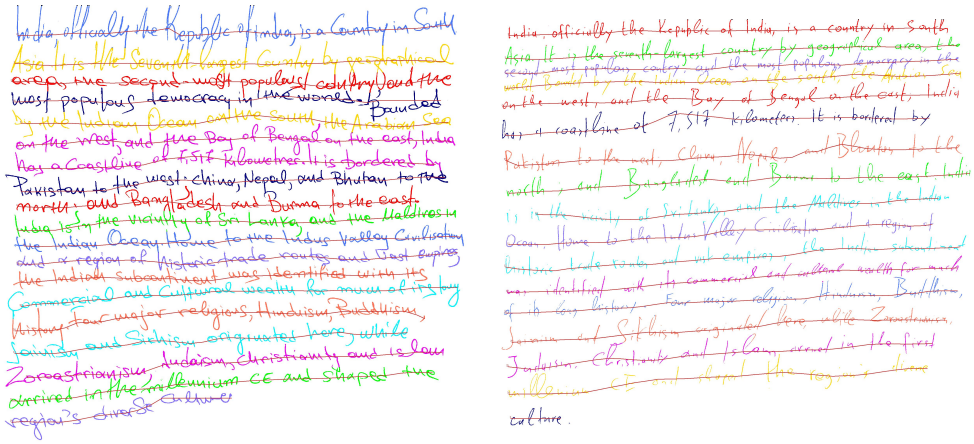
FIGURE 3.3 – Classement des participants pour le score de F-mesure de segmentation en mots [GSL10]. Notre système est "IRISA".

3.1.2 Utilisation d'images floues

Dans un second temps, des travaux en collaboration avec Jean Camillerapp ont permis d'expérimenter un nouveau système d'extraction des lignes de texte directement à partir de l'image. Il s'agit de construire une image floue, pour repérer les zones candidates dans lesquelles peuvent apparaître des lignes de texte. L'idée consiste toujours à considérer une vision « de loin » de l'image, mais en limitant les erreurs de quantification liées au changement de résolution de l'image. On obtient ainsi des indices sur la présence des lignes de texte (figure 3.6a).

Ces indices de lignes de texte sont alors combinés par une description grammaticale, en tenant compte du contexte, qui va permettre de construire les lignes de texte. Par exemple, dans un corpus qui ne contient que des pages de texte monocolonne, la description grammaticale peut prévoir d'agglomérer horizontalement les indices de lignes (figure 3.4). Dans des pages qui peuvent contenir du texte sur plusieurs colonnes, ou séparées par des tableaux, la description grammaticale devra être plus contraignante pour agglomérer les différents indices de lignes de texte, en tenant en compte la pente locale du texte. La détection des bords de page permet d'éviter de construire des lignes de texte dans les fragments de page voisine qui pourraient apparaître sur l'image.

Ces travaux ont été développés en deux phases. Une première version de l'image floue, construite à base de filtrage particulaire, a été soumise au concours ICDAR 2013 [Sta+13]. Les données proposées pour ce concours sont des textes manuscrits contemporains, en plusieurs alphabets, mais toujours monocolonne. La figure 3.4 montre que notre méthode basée sur une combinaison des données permet ici de résoudre les problèmes liés au recouvrement des lignes de texte, ainsi qu'à la pente variable. Les résultats ont été publiés dans [LCC14], et sont rappelés sur la figure 3.5. Notre système obtient la quatrième place avec un score FM de plus de 98%.



(a) (b)

FIGURE 3.4 – Exemple de résultats de segmentation en ligne, concours ICDAR'13, sur des images présentant des difficultés : recouvrement des hampes et des jambages, pente et courbure variées.

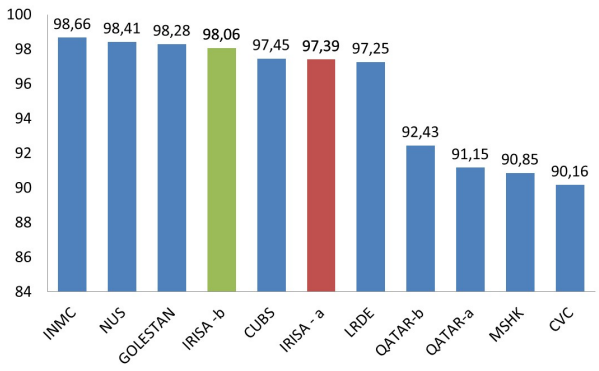
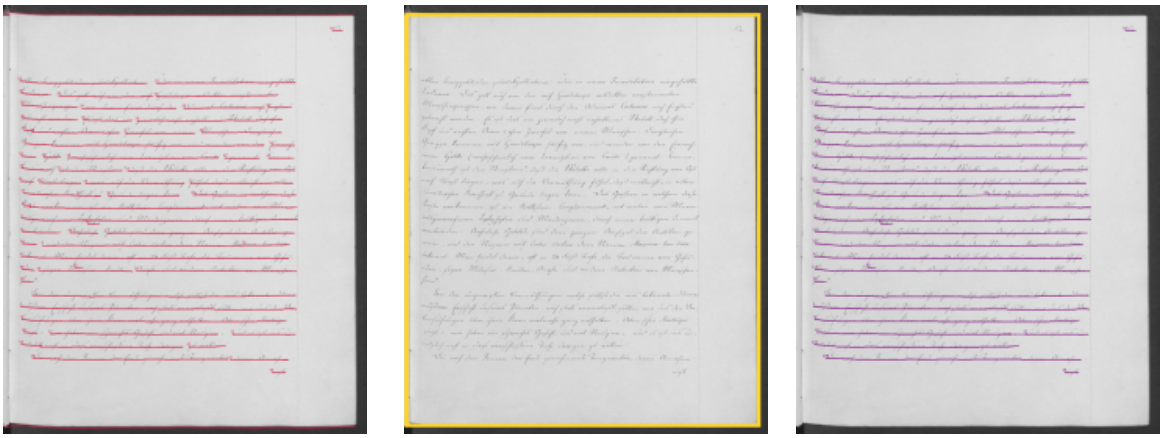


FIGURE 3.5 – Résultats de la compétition ICDAR'2013 [Sta+13] : score de F-measure. Nos systèmes sont IRISA-a et IRISA-b.

Une seconde version de ces travaux, construite à base d'un filtrage passe-bas, a été soumise pour le concours CBAD organisé pour la conférence ICDAR 2017 [Die+17]. Ce concours est

composé de deux bases de documents anciens : une base dite *simple* contenant des tableaux monocolonnes, et une base dite *complexe*, contenant des documents sur plusieurs colonnes ou présentés dans des tableaux. Nous avons dans un premier temps appliqué nos travaux, sur les documents *simples*, puisque cela nous a permis d'utiliser les mêmes connaissances que précédemment, à savoir une organisation des indices visuels en lignes de texte monocolonnes (figure 3.6). Nous avons également pris en compte des connaissances sur le bord de page.



(a) Indices de lignes de texte (b) Bords de page (c) Résultat

FIGURE 3.6 – Combinaison de données issues d'une image floue pour la construction de lignes de texte. Exemple sur une image du concours CBAD 2017.

La figure 3.7 présente les résultats de la compétition pour ICDAR 2017, sur le corpus *simple*. Sans adaptation particulière au corpus de documents, notre système obtient la quatrième place de la compétition. Les méthodes ayant obtenu de meilleurs résultats sont toutes basées sur un système d'apprentissage, ce qui nous a encouragé par la suite à nous tourner vers des méthodes à base de réseaux de neurones profonds.

Method	P-value	R-value	F-value	Rank
DMRZ	0.973	0.970	0.971	1
UPVLC	0.937	0.855	0.894	2
BYU	0.878	0.907	0.892	3
IRISA	0.883	0.877	0.880	4
LITIS	0.780	0.836	0.807	5

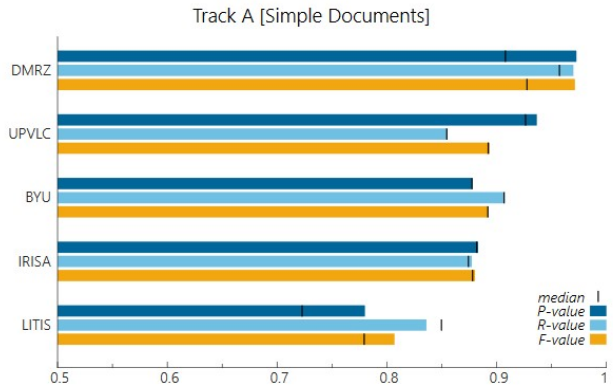


FIGURE 3.7 – Résultats de la compétition ICDAR'2017, sur la base CBAD *simple* [Die+17]. Notre système est IRSIA.

3.1.3 Utilisation de réseaux de neurones profonds

L'apparition de techniques utilisant des réseaux de neurones profonds dans l'état de l'art a permis d'envisager de nouvelles manières de combiner des informations pour la détection de lignes de texte.

Dans le cadre du travail de stage de master de Camille Guerry, nous avons proposé d'utiliser une description grammaticale des lignes de texte contenues dans des tableaux. Les indices visuels utilisés sont obtenus sur des cartes de probabilités résultant de l'application de réseaux de neurones. Nous avons utilisé le réseau profond dhSegment [ASK18], que nous avons entraîné pour pouvoir obtenir : les positions des lignes de textes, les points de début et de fin des lignes de texte, les positions des bords de page (figure 3.8).

Par rapport au fait d'utiliser le seul réseau dhSegment [ASK18], nous avons montré dans ces travaux que la combinaison des données permet d'améliorer la détection des lignes de texte, notamment dans le cas précis des tableaux. En effet, les lignes de texte produites par le réseau de neurones ne sont pas toujours correctement segmentées au niveau des frontières des colonnes des tableaux. La prise en compte de la présence des filets verticaux permet d'améliorer la segmentation au niveau des colonnes. Cette prise en compte des filets est réalisée de manière très simple grâce à l'expression de la connaissance dans le langage EPF.

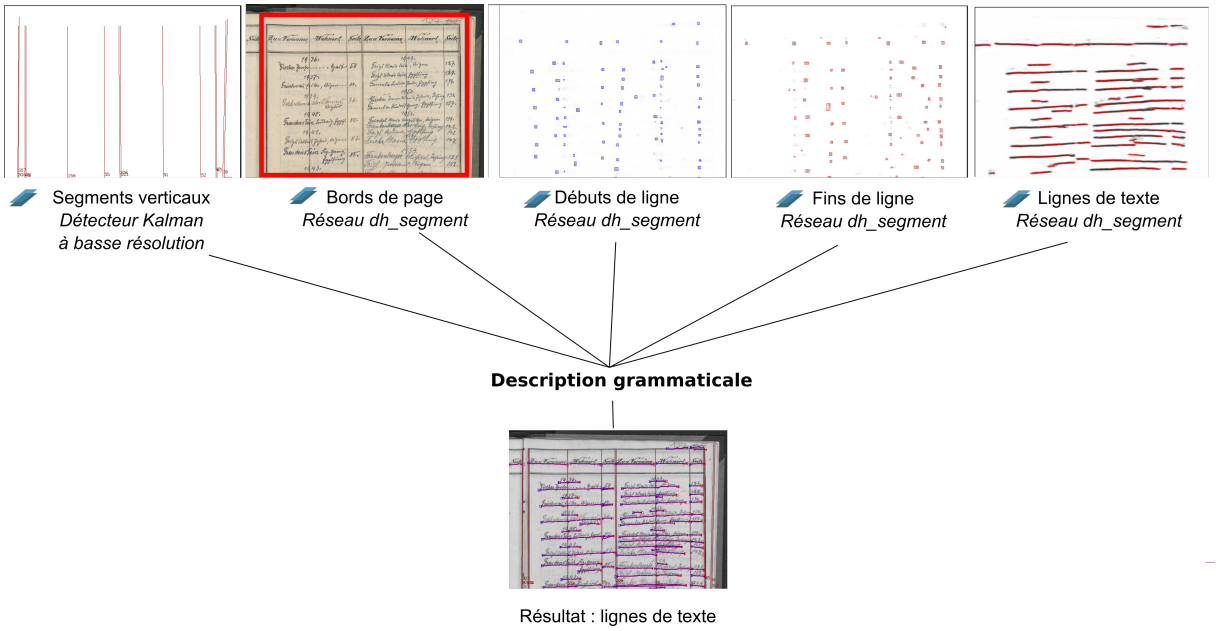


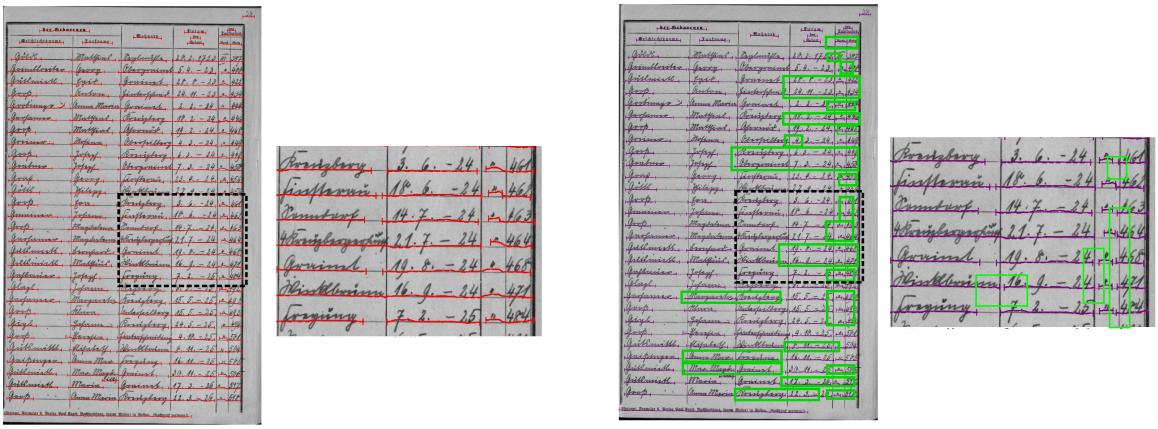
FIGURE 3.8 – Combinaison de données hétérogènes issues de réseaux de neurones pour la détection de lignes de texte dans des structures tabulaires.

La combinaison des tous ces indices permet de fiabiliser la localisation des lignes de texte. Le contexte de détection dans des tableaux est en effet particulièrement complexe puisque la prise de décision de mettre bout à bout deux indices de lignes de texte ne dépend pas seulement de leur proximité, mais aussi de la présence de séparateurs verticaux de tableaux. La figure 3.9 montre que la grammaire a permis de resegmenter les lignes de textes initialement trouvées par dhSgment.

Ces travaux ont été appliqués à la base de la compétition CBAD - ICDAR 2017, mais cette fois-ci nous avons pu aborder la base *complexe*, constituée d'images avec des tableaux. Ceci a permis la publication [GCL19]. Le tableau 3.1 montre que la méthode proposée permet d'obtenir de bons résultats sur cette base *complexe*, notamment en améliorant les résultats bruts de dhSegment.

3.2 Segmentation en paragraphes

En disposant des lignes de textes, il est fréquent de chercher à reconstruire les paragraphes d'un texte. Si dans des documents imprimés, la notion de paragraphe répond à des règles physiques très précises, leur segmentation n'est pas si aisée dans le cas de documents manuscrits, ou



(a) Résultats bruts de dhSegment

(b) Lignes de texte après combinaison des indices visuels

FIGURE 3.9 – Intérêt de la combinaison des données pour la détection des lignes : en vert, les points corrigés par l'utilisation de la description grammaticale

Système	Précision	Rappel	F-mesure
UPVLC [Die+17]	0.833	0.606	0.702
IRISA [Die+17]	0.692	0.772	0.730
BYU [Die+17]	0.773	0.820	0.796
DMRZ [Die+17]	0.854	0.863	0.859
dhSegment [ASK18]	0.826	0.924	0.872
Notre méthode	**0.858**	0.935	**0.895**

TABLE 3.1 – Résultats sur la compétition CBAD, base *complexe* de 1010 images

mixtes manuscrits/imprimés.

Un système de règle est alors bien adapté pour décrire de manière simple quelles sont les contraintes propres à chaque corpus. Les primitives restent toujours les lignes de texte qui ont été extraites dans une étape précédente. La grande souplesse de la méthode DMOS vient du fait que les primitives sont interchangeables de manière transparente. Une fois la description d'un document écrite, il est donc très aisé de changer les lignes de texte en entrée si nécessaire.

De nombreux corpus nécessitent le regroupement de lignes de texte en paragraphes. Nous évoquons dans cette section les travaux sur les courriers manuscrits du corpus Rimes ainsi que des actes de registres paroissiaux.

3.2.1 Corpus Rimes

Le corpus Rimes est un corpus de compétition internationale, contenant des courriers manuscrits. L'objectif de la tâche de segmentation sur ces documents est d'identifier les différents blocs de texte contenus dans ces documents : coordonnées destinataire, coordonnées expéditeur, date et lieu, corps de texte, signature... (figure 3.10).

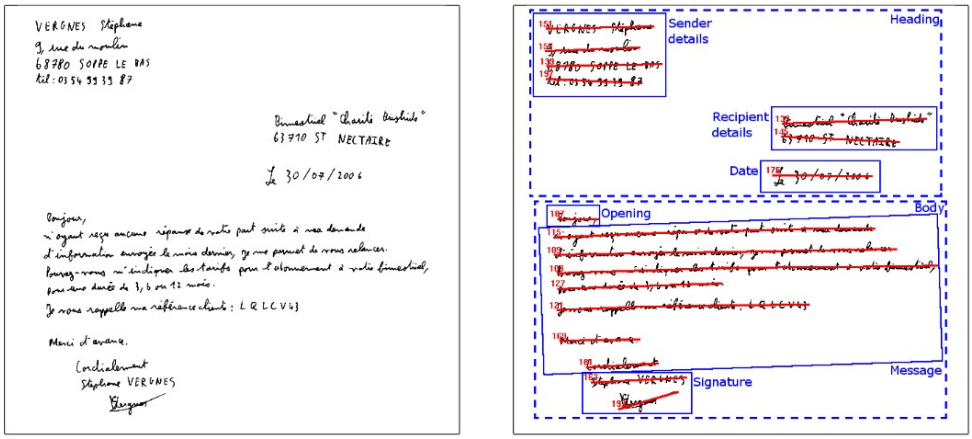


FIGURE 3.10 – Exemple de courrier manuscrit, et le résultat attendu : lignes de texte regroupées en paragraphes typés

Dans mes travaux de thèse, j'avais proposé une première version d'analyse de ces documents, et participé aux compétitions de reconnaissance. Nous avons continué à travailler sur ce corpus, d'une part pour valider les travaux d'analyse stochastique, d'autre part pour valider les mécanismes d'inférence d'opérateurs de position.

3.2.1.1 Analyse stochastique

Nous avons repris les documents RIMES pour valider l'intérêt des travaux d'analyse stochastique proposés lors du stage d'André Maroneze, et évoqués dans la partie 2.3.1. Nous synthétisons ici des travaux détaillés dans [OCL11].

Ces travaux se basent sur le constat suivant : des méthodes syntaxiques sont adaptées pour la reconnaissance de courriers manuscrits, puisque l'organisation de ces documents répond à des règles usuelles très précises. Ainsi, l'organisation physique d'un courrier peut être décrite aisément. En revanche, des méthodes statistiques obtiennent également de bons résultats. C'est pourquoi, nous avons proposé une fusion de connaissances structurelles et statistiques pour l'analyse de ces courriers.

Nous avons intégré pour cela une interaction avec un reconnaisseur d'écriture manuscrite. Ceci a été testé notamment pour le champ 'ouverture' qui contient des textes avec un vocabulaire réduit tel que « Bonjour », « Madame, Monsieur »... La description grammaticale doit donc

être capable de tester différentes lignes de texte, pour voir quelle est la ligne la plus probable de contenir le texte d'ouverture.

Le choix de la ligne « la plus probable » se fait grâce à l'utilisation de l'opérateur grammatical `FIND_BEST_FIRST`, présenté dans la partie 2.3.1. Cet opérateur permet d'analyser prioritairement la meilleure combinaison globale du document, en tenant compte des différents scores de probabilités attribués aux premières lignes candidates.

Ces travaux, publiés en 2011 nous ont permis d'améliorer, à l'époque, les résultats de reconnaissance de l'état de l'art (tableau 3.2).

	Campagne Rimes 2009				
	CEP	IRISA	LITIS	CEP/LITIS	Méthode [OCL11]
Taux d'erreur (%)	8.53	8.97	12.62	12.88	5.53

TABLE 3.2 – Résultats de la campagne RIMES 2009[Gro+09], à laquelle nous avions participé sous le nom IRISA, comparés avec nos travaux plus récents [OCL11]

L'article d'Andre Maroneze [OCL11] a obtenu le 'Best student paper award' à la conférence DRR 2011 pour ces travaux.

La figure 3.11 synthétise la combinaison de données hétérogènes mise en place dans le cadre des documents RIMES. La reconnaissance est guidées par des règles symboliques : les lignes de texte sont organisées de manière logique dans les pages, et combinées avec l'utilisation de reconnaisseurs de quelques mots manuscrits (figure 3.11).

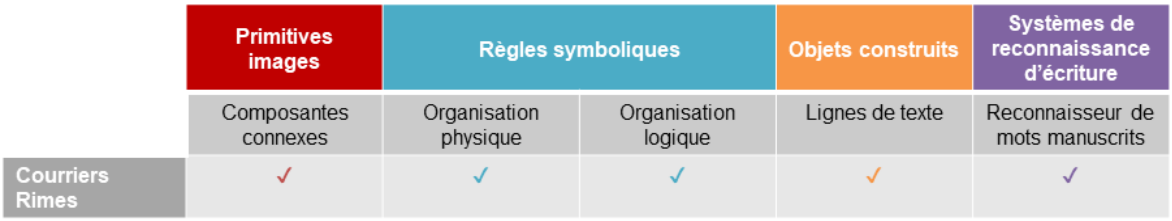


FIGURE 3.11 – Données hétérogènes combinées pour la reconnaissance de courriers manuscrits

3.2.1.2 Inférence de règles grammaticales

Les documents de RIMES ont également servi de données d'application dans le travail de thèse de Cérès Carton [Car16], en particulier sur l'inférence d'opérateurs de positions, présentée dans la partie 2.3.2.

Dans le cadre de la méthode DMOS, il est nécessaire de définir des opérateurs de position pour localiser la suite de la recherche. Par exemple, pour indiquer qu'un bloc de post scriptum se trouvera en bas de la page on écrira :

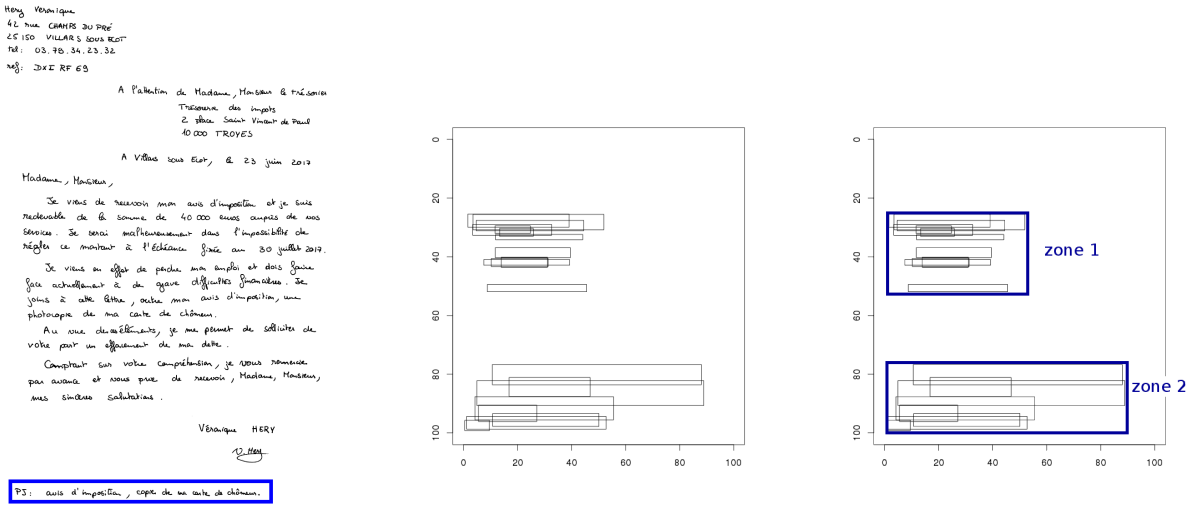
```
AT (basPage) &&
blocPS.
```

Toute la difficulté réside alors dans la définition, jusqu'à présent manuelle, de l'opérateur `basPage`. Qu'est-ce que le bas de la page, où délimiter la zone ? Il est nécessaire d'avoir une zone assez grande pour ne pas faire d'oubli, mais pas trop grande pour limiter les confusions.

Le travail de thèse de Cérès a consisté à remplacer, lorsque c'est possible, la définition manuelle des opérateurs de position par des zones inférées à partir de l'analyse d'un échantillon d'apprentissage. Elle a mis au point, notamment, la brique LearnPos, dont les détails sont donnés dans [CLC14a ; CLC14b].

LearnPos est basé sur des algorithmes de clustering, et une interaction utilisateur. Il permet ainsi de définir des opérateurs de position absolu ou relatifs, de signaler les cas rares. Les opérateurs de position obtenus peuvent être utilisés pour la définition de la grammaire, tout en réduisant l'exploration manuelle du corpus de documents.

La figure 3.12 présente un exemple d'inférence d'opérateur de position, pour la classe PS/PJ du corpus Rimes. Le clustering est appliqué sur la vérité terrain de la position des PS/PJ sur 300 pages. Cela permet ainsi la génération de deux zones, dont les coordonnées seront directement utilisées pour l'inférence des opérateurs de position dans la grammaire.



(a) Exemple de courrier contenant PS/PJ

(b) Positionnement normalisé des PS/PJ pour un corpus de 300 pages

(c) Calcul d'une zone par groupe

FIGURE 3.12 – Exemple d'inférence de l'opérateur de position pour les éléments "PS/PJ" par clustering

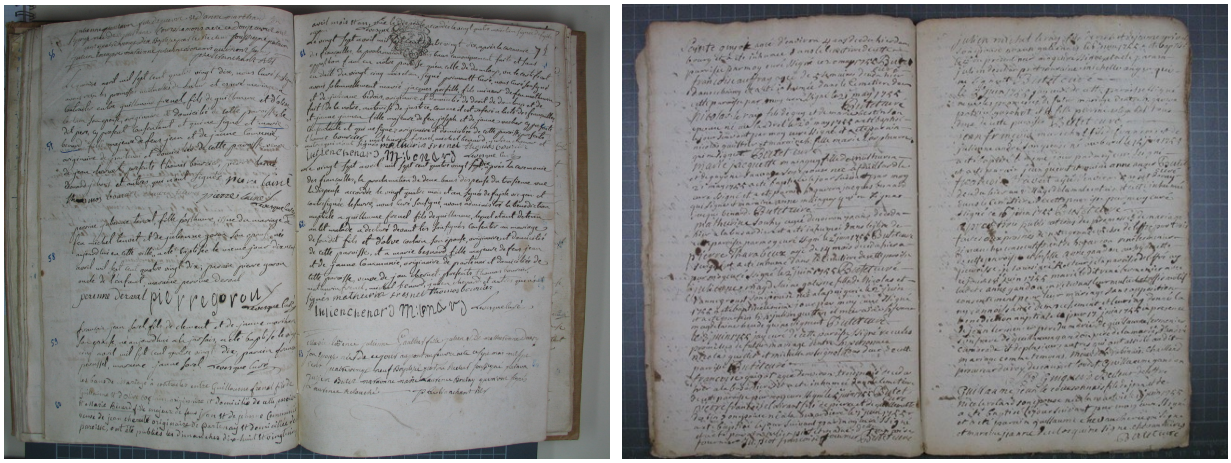
Ces travaux n'ont pas permis de faire progresser les scores de reconnaissance sur le corpus RIMES, mais ont permis de montrer que des opérateurs de positions inférés étaient au moins aussi efficaces que des opérateurs manuels, avec un coût de codage plus faible pour l'utilisateur. De manière chiffrée, la grammaire de description des courriers RIMES contient initialement 102 paramètres, définis manuellement. Avec l'inférence d'opérateurs de position, 48 de ces paramètres sont définis automatiquement.

3.2.2 Registres paroissiaux

Plus récemment, nous travaillons sur la segmentation en paragraphes dans des registres paroissiaux (figure 3.13). Ceci se place dans le cadre d'un partenariat avec la startup Doptim, et la thèse CIFRE de Solène Tarride, démarrée en 2019, que j'encadre avec Bertrand Coüasnon.

Les registres paroissiaux BMS (Baptèmes, Mariages et Sépultures) sont des documents indispensables pour les généalogistes. Ce sont des documents anciens (datés de 1539 à 1789). La problématique est de produire une transcription textuelle la plus complète possible. Le premier travail consiste à segmenter les pages de registres en actes.

Nous avons envisagé plusieurs manières de combiner des indices visuels pour la segmentation de pages en acte : l'utilisation des lignes de texte, des signatures, d'un réseau entraîné sur les actes, puis une combinaison de plusieurs indices visuels. Nous présentons ces différentes approches.



(a) (b)

FIGURE 3.13 – Exemples de documents de type registre BMS

3.2.2.1 Grammaire basée sur les lignes de texte

Dans une première stratégie, nous avons proposé d'utiliser les indices sur la position des lignes de texte, et de regrouper ces lignes de texte en paragraphes, selon la proximité. Nous avons ainsi utilisé une grammaire décrivant les actes comme des lignes de textes agencées en blocs.

Cette stratégie convient pour quelques images, mais dans la majorité des cas, les registres sont trop denses pour pouvoir être segmentés de cette manière par un oeil humain. Par exemple, sur la figure 3.13b, la seule disposition des lignes de texte ne permet pas de conclure sur la segmentation en actes.

Nous n'avons pas évalué cette approche de manière quantitative, car elle ne semblait pas assez pertinente.

3.2.2.2 Grammaire basée sur les signatures

Dans une seconde stratégie, nous avons envisagé de nous baser sur la présence de signatures pour délimiter les fins d'actes. En effet, les actes de registres sont souvent terminés par la signature du prêtre, et éventuellement des personnes concernées. La figure 3.14 montre les signatures présentes dans le document 3.13b : cette figure illustre bien le fait que les signatures sont de bons indices pour la segmentation en acte.

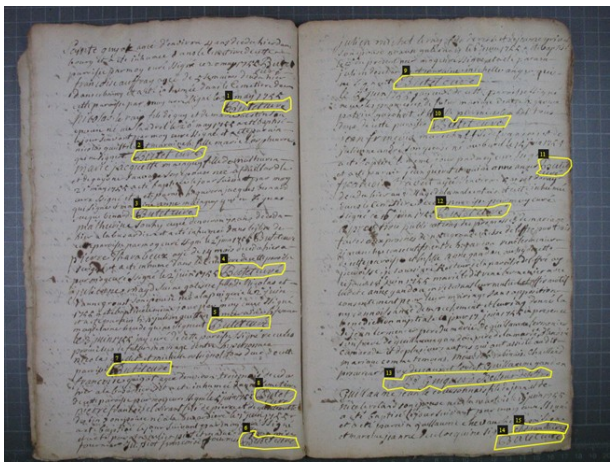


FIGURE 3.14 – Intérêt des signatures pour visualiser la présence d'actes

La détection des signatures pourrait sembler simple : on peut imaginer qu'il s'agit de "pâtés" d'écriture, s'apparentant à un dessin graphique. En réalité, sur les documents qui nous intéressent, il arrive fréquemment que les signatures soient de simples prénoms et noms, écrits de la même manière que le corps de texte. Pour la détection des signatures, nous avons proposé d'entraîner un réseau basé sur de l'apprentissage profond. Nous avons comparé dans [Tar+19], deux variantes de réseau U-net : dhSegment[ASK18] et ARU-Net[Grü+18].

Un fois que nous disposons des signatures et des lignes de texte, nous avons défini dans le langage EPF les règles suivantes pour combiner les données :

— un registre est constitué de deux pages,
— une page est organisée en actes,
— un acte est composé de lignes de texte et terminé par une ou plusieurs signatures.

Cette méthode a obtenu des résultats prometteurs, qui ont permis une publication dans [Tar+19]. Les résultats obtenus par cette approche sont présentés dans le tableau 3.4. Toutefois, cette approche présente quelques limites. En effet, il arrive que les signatures ne soient pas bien reconnues par le réseau de neurones, ou bien simplement que l'acte ne soit pas signé dans le registre.

3.2.2.3 Apprentissage des actes

Après avoir cherché à entraîner un réseau pour localiser les signatures dans des actes, il nous a semblé naturel de tester l'utilisation d'un réseau pour la localisation d'actes complets dans des pages. Nous avons pour cela utilisé des méthodes de détection d'objets. Solène Tarride a réalisé l'entraînement de trois types de réseaux : Mask R-CNN [He+17], RetinaNet [Lin+17] et YOLOv3 [RF18], à partir d'une base étiquetée manuellement contenant 200 images.

Un article est en cours de soumission avec ces résultats. Le tableau 3.3 présente une synthèse des trois meilleurs modèles. Des exemples qualitatifs de résultats sont présentés sur la figure 3.15. Les résultats ont montré que Mask R-CNN était le réseau le plus adapté. Nous avons poursuivi les expérimentations avec ce réseau, afin d'améliorer les résultats. Les scores finaux sont présentés dans le tableau 3.4.

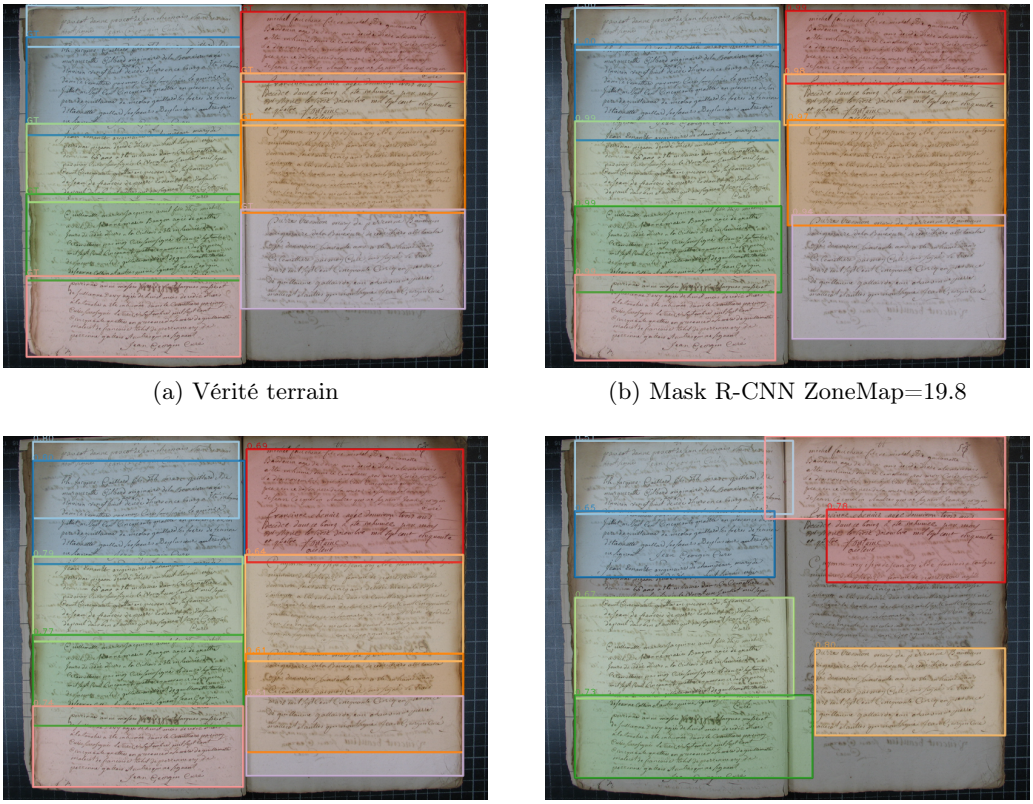
Ces résultats ont montré que les réseaux de neurones profonds permettent d'obtenir des résultats corrects en terme d'analyse de structure d'images de documents complexes. Cependant, il reste quelques erreurs de confusion sur des cas peu fréquents. Pour augmenter les performances, il serait probablement nécessaire d'annoter davantage d'images.

Model	ZoneMap	AP@.50	AP@.75
Mask R-CNN	**31.9**	**86.8**	**66.1**
RetinaNet	47.4	68.9	36.5
YOLOv3	76.3	24.3	0.8

TABLE 3.3 – Comparaison de trois réseaux pour la détection des actes. Pour la métrique ZoneMap[GKO14], le score est meilleur lorsqu'il est proche de 0. AP, Average Precision, doit se rapprocher de 100.

3.2.2.4 Grammaire basée sur la combinaison de connaissances

Au vu des résultats obtenus par les précédentes méthodes, nous avons mis en place une dernière approche, basée sur la combinaison de plusieurs indices visuels. La figure 3.16 synthétise cette approche. Les règles logiques sont utilisées pour combiner les lignes de texte, en tenant compte des bords de pages, des signatures des actes, ainsi que d'un nouvel indice : les premières lignes d'actes.



(a) Vérité terrain

(b) Mask R-CNN ZoneMap=19.8

(c) Retina-Net ZoneMap=32.2

(d) YOLOv3 ZoneMap=76.7

FIGURE 3.15 – Exemple de résultats comparatifs de détection des actes, avec score d'erreur ZoneMap

Nous avons entraîné un nouveau réseau, basé sur LARU-net [Grü+18], capable de localiser les lignes de texte qui sont les premières lignes d'un acte (figure 3.17b). En effet, les premières lignes d'actes présentent des motifs répétitifs : un léger espace avec la ligne précédente, et des mots clés souvent semblables.

Cet extracteur de premières lignes n'est pas parfait, et génère parfois des faux positifs. Par ailleurs, l'extraction des signatures provoque parfois des oublis. L'objectif de la nouvelle grammaire est donc de combiner les résultats de ces deux réseaux, pour fiabiliser la détection des actes (figure 3.17). Les premiers résultats dans le tableau 3.4 montrent que cette grammaire combinant des données de première ligne et de signature obtient de meilleurs résultats que la version optimisée de l'apprentissage avec Mask R-CNN.

Un article de revue synthétisant ces résultats a été soumis à IJDAR.

	Score ZoneMap	Nb match	Nb split	Nb merge
Grammaire signature (partie 3.2.2.2)	32.1	1228	166	82
Mask R-CNN (partie 3.2.2.3)	29.1	1293	40	107
Grammaire combinaison (partie 3.2.2.4)	27.1	1401	71	42

TABLE 3.4 – Comparaison des différentes approches de segmentation en actes, sur une base de 200 pages, 1565 actes. La grammaire avec combinaison des données obtient de meilleurs scores que le réseau Mask R-CNN.

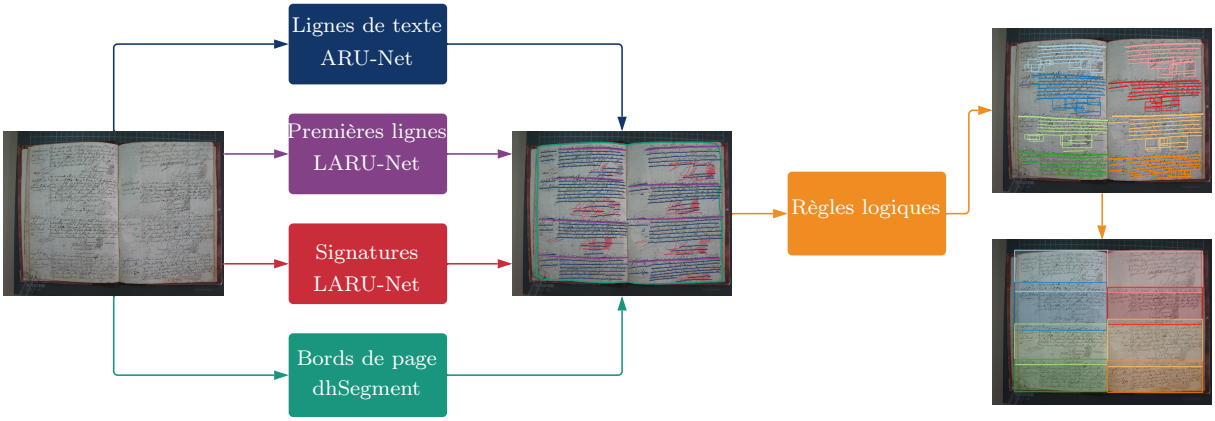
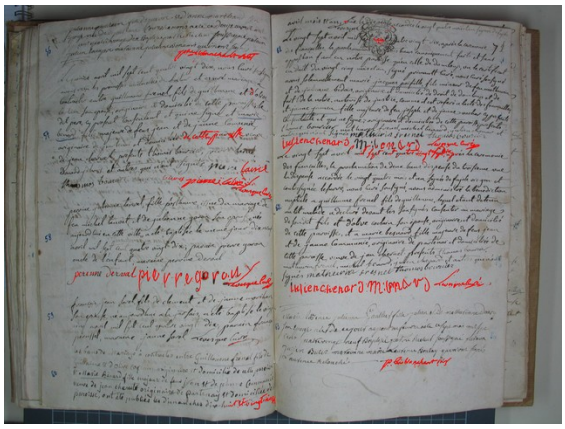
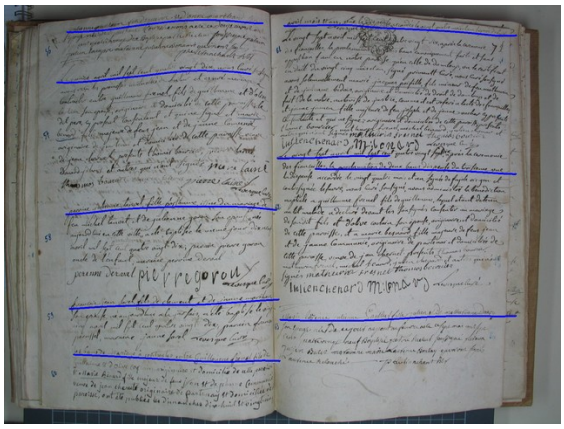


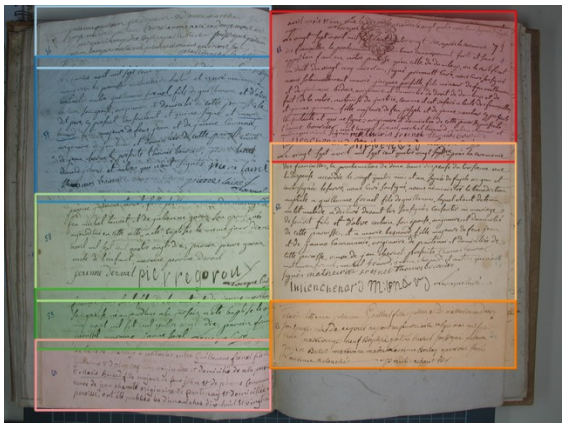
FIGURE 3.16 – Combinaison de données hétérogènes pour la segmentation de pages de registres BMS en actes



(a) Signatures détectées (faux positifs au milieu de la page de gauche)



(b) Premières lignes détectées (dédoublement sur page de droite)



(c) Actes construits grâce à la description grammaticale

FIGURE 3.17 – Combinaison de données signature et premières lignes pour la segmentation d'actes

3.2.3 Bilan sur la combinaison de données

La figure 3.18 synthétise les données combinées dans le cadre de l'analyse d'actes de registres paroissiaux. Il s'agit de lignes de texte, et des résultats de systèmes d'apprentissages (figure

3.18) : extraction des signatures, premières lignes, bords de pages. La combinaison est guidée par une connaissance de l'organisation physique et logique des pages. Les résultats obtenus dans le tableau 3.4 démontrent de manière chiffrée l'intérêt de cette combinaison dans ce contexte précis.

	Règles symboliques		Objets construits	Résultats de systèmes d'apprentissage		
	Organisation physique	Organisation logique	Lignes de texte	Réseaux de neurones Bords de pages	Réseaux de neurones Signatures	Réseaux de neurones Premières lignes
Registres paroissiaux	✓	✓	✓	✓	✓	✓

FIGURE 3.18 – Données hétérogènes combinées pour la reconnaissance de registres paroissiaux

Nous démarrons actuellement des expériences sur les capacités de généralisation de notre système, avec une plus grande base de données. Ces expériences, en cours de réalisation, tendent à montrer que la combinaison de données permet une meilleure généralisation que le système Mask R-CNN basé uniquement sur de l'apprentissage. Pour obtenir des performances équivalentes, Mask R-CNN nécessiterait des données étiquetées supplémentaires, ce qui est difficilement envisageable dans un contexte applicatif réel. La combinaison de données hétérogènes à base de règles permet donc de pallier ici l'absence de grandes quantités de données étiquetées.

3.3 Conclusion

Ce chapitre a présenté les travaux menés sur l'analyse de structure physique de pages : segmentation en lignes de texte et organisation en paragraphes.

Les travaux présentés peuvent être considérés selon un ordre chronologique : à chaque étape, la combinaison de données hétérogènes a permis de résoudre des problèmes de plus en plus complexes. Dans les travaux les plus anciens, pour la localisation de lignes de texte, nous avons combiné des primitives images, segments à basse résolution et composantes connexes à haute résolution, en étant guidé par des règles liées à la perception visuelle des éléments. Puis nous avons combiné des données construites, les lignes de texte, pour les organiser en paragraphes.

Plus récemment, les réseaux de neurones profonds sont devenus un concurrent incontournable pour l'analyse de structure de pages. Nous avons alors orientés nos travaux vers la combinaison de résultats de réseaux profonds, combinés par des règles de grammaires. Ainsi, il est possible de pallier les limites des différents réseaux par une prise de décision liée à des connaissances logiques sur les documents.

Chapitre 4

Analyse de documents à structure complexe

Dans ce chapitre, je regroupe les travaux sur plusieurs types de documents dont je qualifie la structure de « complexe ». En effet, leur organisation physique est directement liée à la compréhension logique et sémantique de leurs contenus.

Chaque corpus présenté ici a été l'occasion d'expérimenter et de valider de nouvelles combinaisons d'indices visuels pour la reconnaissance. Ainsi, l'analyse d'organigrammes se base sur l'analyse de strokes de signal en ligne, qui ont été étiquetées par un classifieur, ainsi que sur l'image construite. Dans des documents fortement hétérogènes comme ceux du corpus Maurdor, la grammaire guide l'analyse d'éléments prégnants pour faire émerger la structure des données. Pour l'analyse de pages de presse ancienne, on utilise les résultats d'un OCR du commerce en lien avec des règles d'agencements d'articles et d'illustrations sur plusieurs colonnes. Des règles d'agencement sont aussi appliquées pour l'analyse de documents d'histoire de la finance, pour lesquels on exploite également les informations venant de la collection.

4.1 Organigrammes

Les travaux sur les documents de type organigrammes (nommés flowcharts dans la suite) ont été réalisés avec les collègues du LS2N de Nantes, notamment en collaboration avec Harold Mouchère. L'objectif est de reconnaître des diagrammes tels que celui présenté sur la figure 4.2. Le corpus est composé de documents en ligne : les images sont des strokes ordonnées.

Dans cette partie, nous montrons comment la reconnaissance de flowcharts peut être favorisée par la combinaison de différentes primitive images, enrichies par les résultats d'un classifieur de strokes et guidées par des règles symboliques (figure 4.1). Les indices visuels sont issus de deux sources : d'une part les strokes issues du signal en ligne (correspondant à la colonne Inkml du tableau), et d'autre part des primitives usuelles (segments et composantes connexes) extraites dans l'image reconstruite à partir du signal en ligne.

	Primitives images			Règles symboliques		Résultats de systèmes d'apprentissage
	Segments	Composantes connexes	Inkml	Organisation physique	Organisation logique	Classifieur de strokes
Flowcharts	✓	✓	✓	✓	✓	✓

FIGURE 4.1 – Données hétérogènes combinées pour la reconnaissance de flowcharts

Les flowcharts sont des documents très fortement structurés, ce qui nous a orientés vers l'utilisation d'une description grammaticale des contenus. Puisque les images en entrée sont constituées

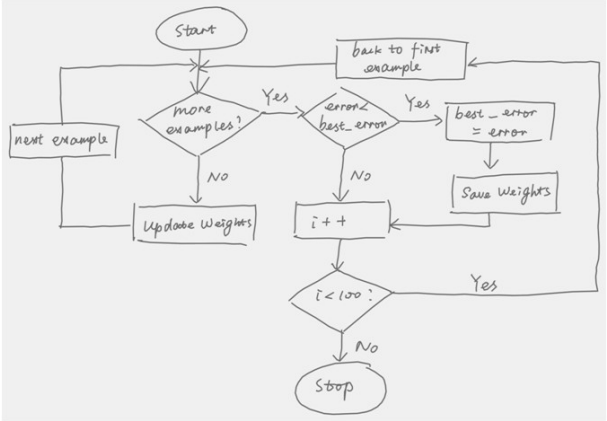


FIGURE 4.2 – Exemple d'image en-ligne de flowchart

de signal en ligne (inkml), nous avons proposé une nouvelle extension à la méthode DMOS pour qu'elle puisse prendre en compte les strokes, représentées par leurs boites englobantes. Nous avons également utilisé les traitements d'images usuels, tels que la détection de composantes connexes, ainsi que l'extraction des segments horizontaux et verticaux (figure 4.3).

La description grammaticale est réalisée en deux étapes. On cherche tout d'abord les *boites* qui sont les formes utilisées pour représenter les étapes du schéma : rectangles, losanges, cercle, ovales, parallélogrammes. Puis, dans un second temps, la grammaire décrit les enchaînements permis dans la symbolique des flowcharts. Par exemple, un losange est un point de décision, il est donc précédé d'une flèche entrante, et suivi de deux flèches sortantes, l'une pour le cas positif et l'autre pour le cas négatif .

Cette description grammaticale a permis de mettre au point un premier système de reconnaissance des flowcharts, qui a été publié dans [Lem+11] et [Lem+13].

Une des limites de cette description est qu'elle repose sur une confiance totale dans la qualité des boites qui sont détectées lors de la première phase d'analyse. C'est typiquement le genre de cas où il est nécessaire de valider la détection des *meilleures boites possibles* et pas simplement des premières trouvées. J'ai donc co-encadré le stage de master de Cérès Carton pour utiliser les outils de `FIND_BEST_FIRST` présentés dans la section 2.3.1, et les appliquer pour améliorer la détection des flowcharts. Ceci a mené à une publication [CLC13].

Nous avons ensuite poursuivi les travaux sur ce thème avec Harold Mouchère, qui a supervisé un stagiaire, Chengcheng Wang, dont l'objectif était d'entraîner un classifieur pour attribuer des labels aux strokes. J'ai ainsi pu enrichir la description grammaticale pour améliorer la description de la structure. Par exemple, une des difficultés rencontrées était la distinction entre les extrémités de flèche et le texte contenu à l'intérieur d'une boîte. L'interaction avec le classifieur a permis de corriger cela. Ces travaux ont été publiés dans [Wan+17].

Dans tous ces travaux réalisés avec les flowcharts, la combinaison de données hétérogènes a permis à chaque fois d'obtenir les meilleurs résultats. En effet, l'utilisation des règles logiques est particulièrement adaptée pour décrire des organigrammes construits également avec une organisation logique précise. Pour interpréter ces documents, un humain combine l'interprétation du signal a plusieurs niveaux : formes graphiques, enchaînement logique et contenu textuel. C'est ce que permet notre système basé sur la combinaison des données.

4.2 Documents hétérogènes : corpus Maurdor

Le projet Maurdor [Bru+14] est un PEA (Projet d'Étude Amont) financé par la DGA, entre 2012 et 2014, et piloté par Cassidian (Airbus Defence & Space). Son objectif est de produire un démonstrateur capable d'extraire le maximum d'informations à partir d'images de documents

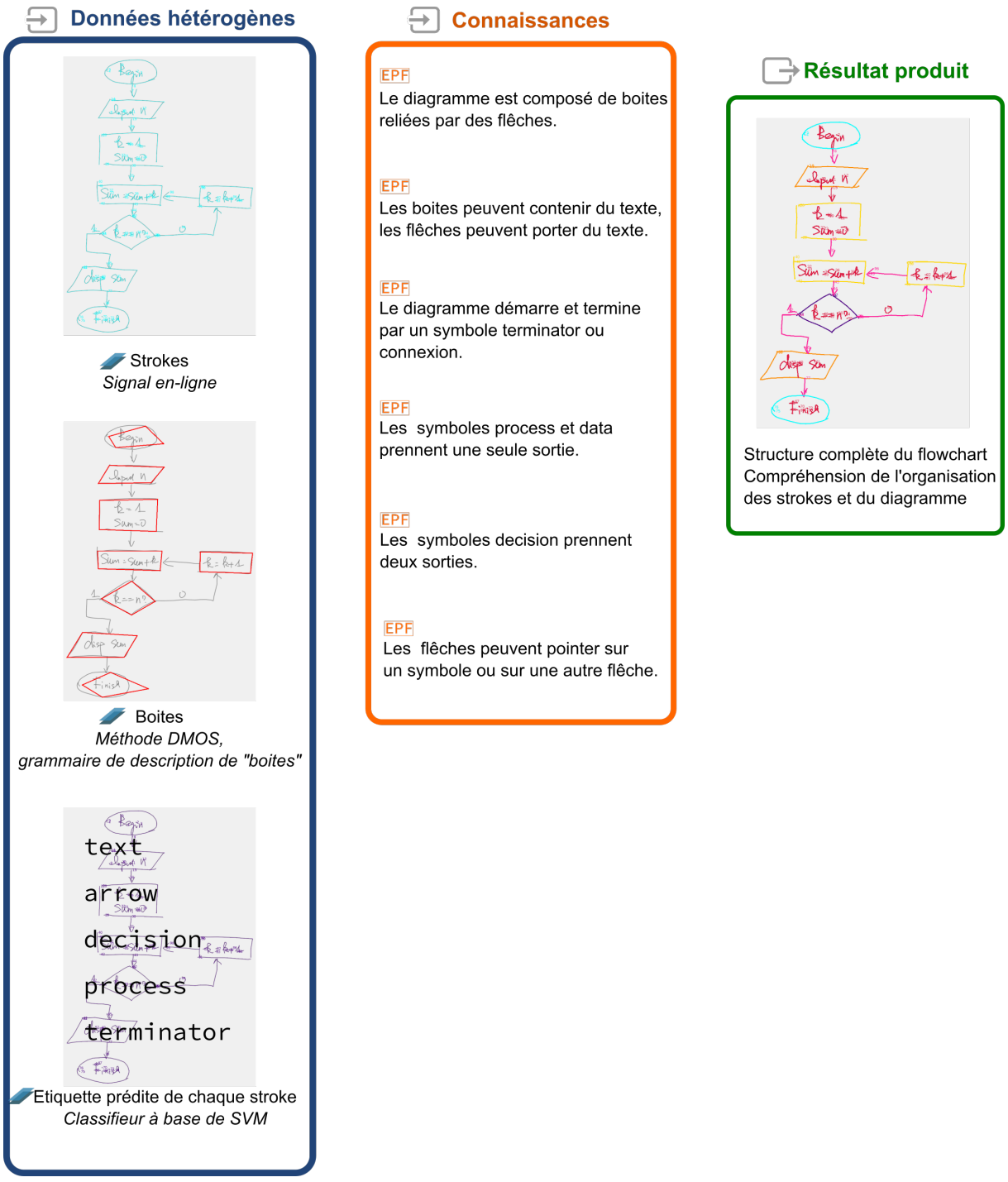
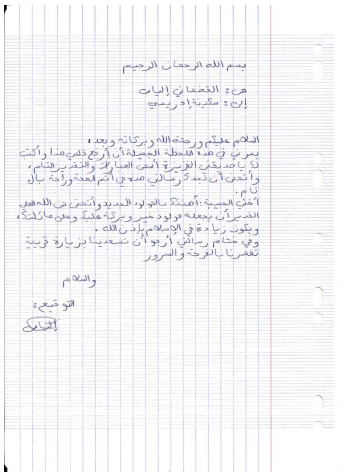
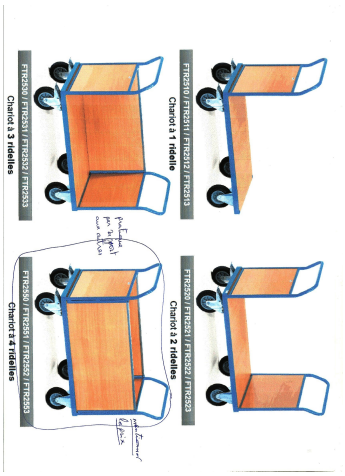


FIGURE 4.3 – Combinaison de données hétérogènes pour la segmentation de flowcharts en ligne

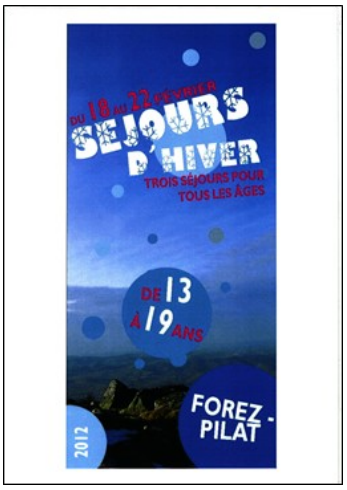
tout-venant : factures, publicités, courriers manuscrits, plans, scans, fax...(voir figure 4.4). Pour ce projet, j'ai co-encadré l'ingénieur Baptiste Poirriez d'avril 2012 à décembre 2014.



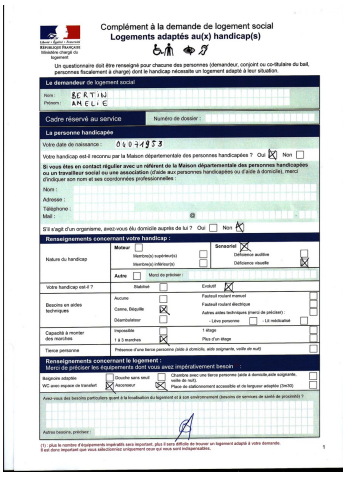
(a) Courrier personnel, en arabe



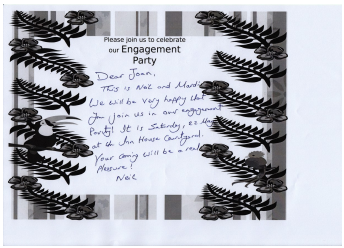
(b) Page de catalogue annotée manuellement



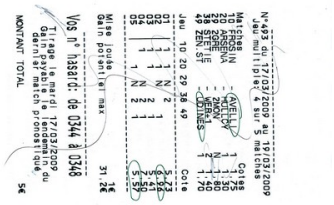
(c) Tract publicitaire



(d) Formulaire rempli en français



(e) Carte d'invitation avec illustrations graphiques



(f) Ticket de caisse, pivoté

FIGURE 4.4 – Exemples de l'hétérogénéité des documents du corpus Maurdor

Le projet Maurdor est découpé en 5 modules de traitements qui forment une chaîne complète :

1. segmentation et classification des régions,
2. identification du script : manuscrit ou imprimé,
3. identification de la langue : français, anglais ou arabe,
4. reconnaissance du contenu textuel,
5. définition de l'ordre de lecture et des relations logiques entre éléments.

Nous avons travaillé sur les modules 1,2 et 5, mais je me focalise ici sur la présentation de notre travail sur le module 1 : la segmentation du document en zones homogènes.

4.2.1 Processus de segmentation de documents

La tâche de segmentation de documents consiste à localiser les régions homogènes dans les documents : régions de texte imprimé, régions de texte manuscrit, tableaux, formulaires, gra-

phiques, dessins, tampons, logos.

Pour résoudre ce problème, nous avons choisi d'utiliser les mécanismes de la vision perceptive humaine, que nous avons déjà exploités pour la localisation des lignes de texte (partie 3.1.1). Nous considérons que, lorsqu'un humain regarde un document, il sait intuitivement faire le tri entre les structures tabulaires, les zones de texte, les zones de graphiques, avant même d'avoir identifié précisément le texte contenu dans le document. En particulier, deux mécanismes sont mis en jeux :

— certains contenus sont prégnants pour la vision humaine, ce sont souvent les éléments les plus structurants,

— la combinaison de plusieurs points de vue d'un même document permet de mettre en place un mécanisme de prédiction/vérification. Ainsi, la structure générale du document est prédite avec une vision globale, avant d'être vérifiée avec un point de vue plus précis.

Nous avons choisi de mettre en place un processus itératif, dans lequel nous appliquons successivement la méthode DMOS pour extraire des éléments dans la page. Le schéma de la chaîne de traitement est présenté sur la figure 4.5.

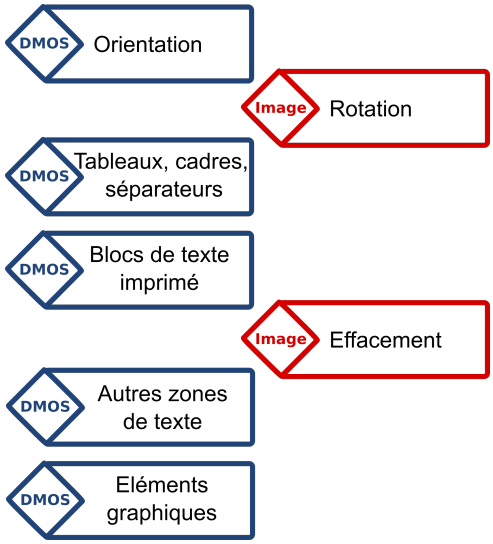


FIGURE 4.5 – Processus d'analyse des documents hétérogènes, alternant analyse avec DMOS (en bleu) et traitement de l'image (en rouge)

L'analyse s'effectue en détectant à chaque fois les éléments les plus prégnants. Ainsi, après avoir déterminé l'orientation de la page, nous extrayons dans cet ordre les constituants du document : les tableaux, cadres et séparateurs, les blocs de texte imprimés en latin, les blocs de texte manuscrit et en arabe, les régions graphiques.

L'intérêt d'un tel mécanisme est qu'il permet de modifier l'image en fonction du résultat d'un traitement précédent. Par exemple, une fois que les filets constituant les tableaux et les cadres sont interprétés dans un document, ceux-ci vont être effacés de l'image avant que l'étape d'extraction des primitives ne soit lancée pour la localisation des zones de texte manuscrit et des graphiques.

Pour chacune des étapes réalisées par DMOS sur la figure 4.5, nous avons mis au point une description grammaticale. Dans toutes ces tâches d'analyse, les calques perceptifs utilisés sont les mêmes : il s'agit des segments, des composantes connexes, ainsi que des mots imprimés qui ont été reconnus par un OCR, en l'occurrence l'OCR Abbyy Fine Reader.

Cette exploitation de l'OCR représente l'originalité de ces travaux, puisque ce corpus est le premier pour lequel nous avons introduit un OCR. Le concept de calque présent dans DMOS permet en effet de considérer comme primitive la boite englobante de chaque mot détecté par l'OCR, sa position, ainsi que l'hypothèse de transcription fournie par l'OCR.

Nous détaillons maintenant chacune des étapes de la figure 4.5.

4.2.1.1 Orientation

Un premier traitement se charge de détecter l'orientation du document. En effet, l'exemple 4.4f montre que certaines images de la base n'ont pas été numérisées dans le bon sens. L'oeil humain le comprend vite, car sa vision est guidée par l'orientation principale des lignes de texte. Nous avons donc utilisé ce principe : on applique un extracteur de segments, à assez basse résolution, sur l'image. Si le nombre de segments verticaux est significativement plus grand que le nombre de segments horizontaux, alors on considère qu'il faut appliquer une rotation à l'image (figure 4.6).

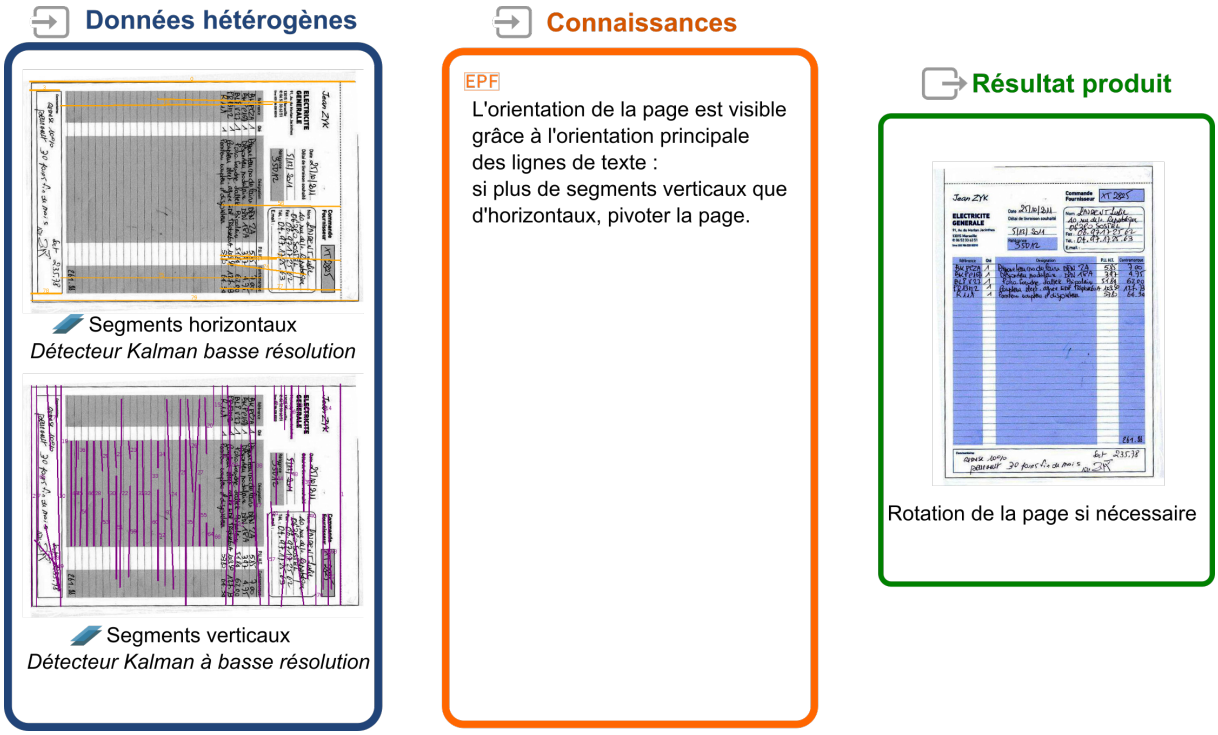


FIGURE 4.6 – Vision perceptive pour la détection de l'orientation dans des pages de documents

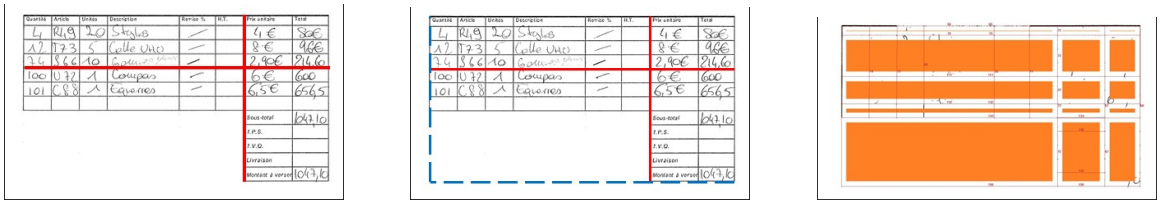
4.2.1.2 Tableaux, cadres et séparateurs

Pour la localisation des tableaux, la grammaire se base en entrée sur les filets horizontaux et verticaux. L'analyse considère tout d'abord la présence de tableaux. Un tableau est caractérisé par la présence d'au moins 2 filets qui se croisent (figure 4.7a). Puis, l'analyse recherche des filets horizontaux et verticaux parallèles. Notons que la description autorise la construction d'un tableau, même si les filets n'ont pas tous la même longueur. C'est le cas de la figure 4.7b qui permet la reconnaissance d'un tableau non rectangulaire, en construisant des "filets virtuels" en bleu. La description grammaticale précise qu'à l'intérieur de chaque cellule, il est possible de trouver récursivement un tableau ou tout autre constituant d'un document.

Lorsque les tableaux ont été détectés, la grammaire exprime qu'un cadre est constitué de quatre filets isolés, qui constituent un rectangle. Après cette passe, les filets restants, suffisamment longs, sont considérés comme des séparateurs.

4.2.1.3 Blocs de texte latin imprimé

Concernant l'analyse des blocs de texte latin, imprimé, nous utilisons l'OCR Fine Reader. Cet OCR n'est pas capable de produire une segmentation correcte pour les documents hétérogènes



(a) Localisation de deux filets qui se croisent (en rouge)

(b) Construction de filets virtuels (en bleu) pour permettre la reconnaissance de tableaux non rectangulaires

(c) Cellules finales

FIGURE 4.7 – Localisation de tableaux

complexes, tels que ceux du corpus Maurdor, mais il peut localiser les mots imprimés et en fournir une transcription. Notons que la difficulté est ici que l'OCR sait parfois reconnaître également des mots manuscrits en script bien écrit.

L'OCR est donc utilisé en entrée de notre description grammaticale. La grammaire considère qu'un bloc de texte se base sur la présence de mots pour lesquels l'OCR est suffisamment confiant : appartenance à un dictionnaire, score de reconnaissance élevé, taille du mot. Ces critères nous permettent de sélectionner les mots détectés comme imprimés par l'OCR. Les mots sont alors combinés en lignes de texte puis en blocs par l'application de règles grammaticales.

Les règles grammaticales expriment qu'un bloc de texte est un ensemble de mots organisés en lignes.

4.2.1.4 Effacement

A ce stade, les éléments les plus structurants du document ont été détectés : tableaux, cadres, blocs de texte imprimés. Cependant, ces éléments pourraient interférer avec la détection des éléments restants (texte manuscrit et graphiques). Par exemple, dans une page de formulaire, l'écriture peut croiser des filets ou du texte imprimé, ce qui causera des difficultés pour la détection des composantes connexes appartenant au texte manuscrit.

Nous proposons donc d'effacer dans l'image les éléments reconnus, pour permettre une nouvelle segmentation de l'image avant de poursuivre l'analyse. L'effacement consiste à remplacer les pixels noirs de l'image par un niveau de gris correspondant à l'arrière plan local dans le document. Il est alors possible de réaliser une nouvelle extraction des composantes connexes.

4.2.1.5 Autres zones de texte

L'étape suivante consiste à construire les zones de texte restantes. Il peut s'agir de texte latin manuscrit, ou bien de texte en arabe manuscrit et imprimé. La grammaire prend en entrée des indices sur les lignes de texte, et l'objectif est de regrouper les lignes de texte selon des proximités de paragraphes, tel que présenté dans la section 3.2. Nous avons également travaillé sur l'utilisation d'un classifieur de type "BonzaiBoost" pour la discrimination de texte imprimé/manuscrit [Ric+14].

4.2.1.6 Éléments graphiques

La dernière étape de l'analyse consiste à localiser les régions graphiques de l'image. Nous utilisons là encore le principe de la vision perceptive : les régions graphiques sont perçues comme des éléments prégnants, dans une vue globale du document. Nous avons donc choisi de travailler dans une image à basse résolution (dimensions de l'image divisées par 4 par sous-échantillonnage). La grammaire considère qu'un élément graphique est une assez grosse composante connexe dans l'image (figure 4.8).

(a) Document après la localisation du bloc de texte

(b) Détection des éléments graphiques prégnants

FIGURE 4.8 – Détection des éléments graphiques (dont fait partie la signature)

4.2.2 Résultats

Notre travail a été évalué dans le cadre de la compétition internationale Maurdor, en novembre 2013. Les participants disposaient d'un lot de 6127 pages pour l'apprentissage, et d'un lot de 1000 pages pour la validation. La compétition a été évaluée sur un troisième jeu de données de 1000 pages, manuellement étiquetées. Plusieurs expériences, ainsi que les métriques, sont présentées dans une publication [PLC14]. Nous présentons dans le tableau 4.1 les résultats de notre systèmes, évalués par le score de Jaccard, et présenté classe par classe.

Participants à la compétition	Zone de texte	Zone graphique	Tableaux
Participant 1	0.552	0.394	**0.363**
Notre méthode	***0.553***	***0.402***	*0.307*
Participant 2	0.307	0.176	0.174

TABLE 4.1 – Score de bonne segmentation des zones, classe par classe, avec la métrique Jaccard (meilleur score proche de 1)

Les résultats obtenus montrent que l'analyse de structure de pages dans le cadre de corpus très hétérogènes reste une tâche difficile. Notre système trouve principalement ses limites dans le cas des superpositions entre zones graphiques et zones de texte. Il est également difficile de construire des blocs de texte homogènes, c'est à dire contenant uniquement du manuscrit ou de l'imprimé.

Pour poursuivre ce travail, il serait nécessaire d'introduire l'utilisation de classifieurs au niveau des composantes connexes, ou des pixels, qui permettraient de donner des indices supplémentaires sur la nature des éléments à classer.

4.2.3 Bilan sur la combinaison de données

Dans le cadre de l'analyse du corpus Maurdor, nous avons combiné (figure 4.9) un calque contenant les segments, un calque de lignes de texte, un calque d'OCR, en étant guidé par la prégnance de certains éléments physiques.

Depuis 2013, les systèmes de reconnaissance et de segmentation ont énormément progressé. Ce corpus Maurdor, réputé « difficile », devient plus abordable grâce notamment aux développement de techniques à base d'apprentissage profond. Il me semblerait intéressant de reprendre ce corpus avec les techniques de l'état de l'art de 2020.

	Primitives images	Règles symboliques	Objets construits	Systèmes de reconnaissance d'écriture
	Segments	Organisation physique	Lignes de texte	OCR commercial
Documents hétérogènes Maurdor	✓	✓	✓	✓

FIGURE 4.9 – Données combinées pour la reconnaissance de documents hétérogènes du corpus Maurdor

4.3 Presse ancienne

Les techniques de description grammaticale de documents sont particulièrement appropriées pour la reconnaissance de documents structurés telles que des pages de presse ancienne. Dans le cadre de la SATT Ouest Valorisation, nous avons pu monter un projet de maturation pour lequel j'ai co-encadré l'ingénieur Eric Le Lay de septembre 2013 à mars 2015.

Nous avons travaillé sur la segmentation de pages de presse ancienne : il s'agit de découper les pages en articles, chaque article étant constitué d'un titre, de plusieurs paragraphes et éventuellement d'illustrations associées (figure 4.10).

La difficulté dans ces documents est que l'organisation des contenus est variable : les articles peuvent être disposés sur plusieurs colonnes, mais l'ordre de lecture est très guidé par la présence d'éléments physiques.

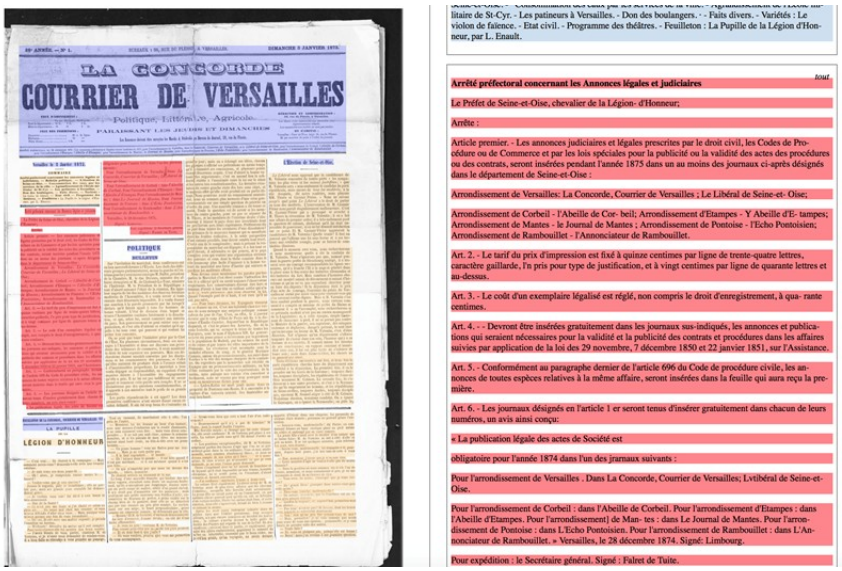


FIGURE 4.10 – Visualisation des résultats de l'analyse de pages de presse : résultat d'une segmentation d'un article à gauche et visualisation de sa transcription plein texte à droite

Dans ces travaux sur la presse ancienne, nous avons combiné d'une part un calque contenant des filets horizontaux et verticaux, et d'autre part des résultats d'OCR (figure 4.11). Afin de gérer les différentes tailles de caractères (très gros caractères des titres, et petits caractères dans le corps de page), nous avons combiné une double application de l'OCR, l'une à basse résolution, et l'autre à haute résolution. Ces éléments sont combinés en suivant des règles physiques et logiques de l'organisation d'une page de presse : articles, rubriques, tableaux, illustrations, pages de une.

Ces travaux de segmentation de pages de presse ont été évalués sur un corpus annoté de 250 pages du journal "L'avenir de Saint-Germain". Pour la segmentation des pages en articles et en

	Primitives images		Règles symboliques		Systèmes de reconnaissance d'écriture	
	Segments	Composantes connexes	Organisation physique	Organisation logique	OCR commercial Haute résolution	OCR commercial Basse résolution
Presse ancienne	✓	✓	✓	✓	✓	✓

FIGURE 4.11 – Données hétérogènes combinées pour la reconnaissance de pages de presse ancienne

tableaux, on obtient un score ZoneMap de 4.50 sur cette base (note : avec la métrique ZoneMap [GKO14], le meilleur score est proche de zéro).

Ces travaux n'ont finalement pas abouti sur un transfert technologique, mais nous ont permis de valider le pouvoir d'expressivité de la méthode DMOS pour combiner des informations provenant de sources variables, notamment la gestion de l'OCR à plusieurs résolutions.

Lorsque nous avons réalisé ces travaux, les techniques d'apprentissage profond n'étaient pas encore très développées. A l'heure actuelle, il me semblerait clairement bénéfique d'utiliser ces techniques pour aider à la segmentation et à la classification d'objets. En revanche, dans ces documents, certaines règles logiques sur l'organisation en colonnes et en articles sont très simples à écrire. Il me semblerait donc utile de continuer à utiliser un système de règles pour fiabiliser et améliorer à peu de coût certaines imprécisions dans les résultats de réseaux de neurones, éventuellement en combinant les résultats issus de plusieurs réseaux. La construction de l'ordre de lecture reste également, à mon avis, plus facile à décrire par des règles logiques.

4.4 Documents d'histoire de la finance

4.4.1 Contexte global

Depuis janvier 2018, nous participons au projet ANR HBDEX ainsi qu'au projet européen EurHisFirm. Il s'agit de deux projets avec plusieurs partenaires, dont le LITIS à Rouen et l'École d'Économie de Paris (PSE).

Ces deux projets, l'un à l'échelle nationale, l'autre à l'échelle européenne, sont portés par des économistes. Ils ont pour but de réaliser une analyse de données financières historiques afin d'aider à la prédiction de nouveaux modèles économiques. Dans ce contexte, je co-encadre la doctorante Camille Guerry et les ingénieurs Simon Bouvier et Iwan Le Floch. Notre objectif est de produire des systèmes d'analyse d'images permettant de reconnaître le contenu de pages de cotations boursières (listes de prix) et d'annuaires d'entreprises (figure 4.12).

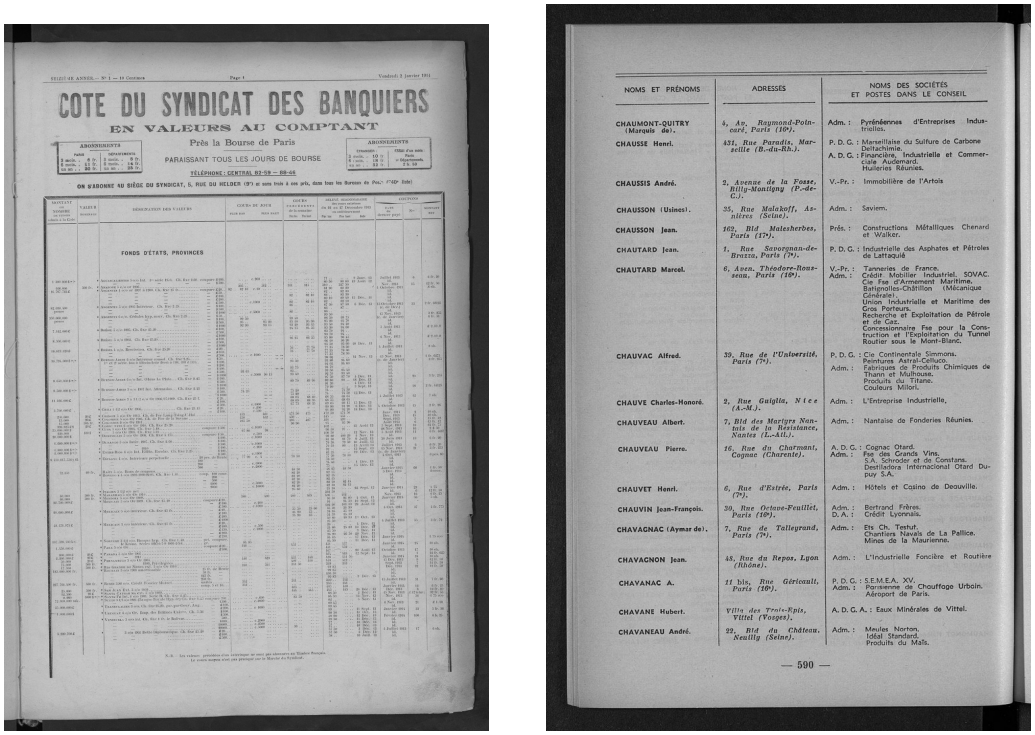
4.4.2 Listes de prix

4.4.2.1 Spécificités du corpus

Le projet HBDEX se focalise sur la reconnaissance de pages de listes de prix, du marché de la Coulisse, à Paris, entre 1899 et 1939. Un exemple de page est présenté sur la figure 4.12a. Il s'agit de documents anciens imprimés, dont la qualité est parfois mauvaise. Chaque jour, un journal de cotation contient entre 5 et 10 pages (selon les années). Nous disposons des documents numérisés pour tous les jours de cotation.

La grande particularité de ces documents est leur stabilité : globalement, d'un jour à l'autre, les seuls éléments variables dans les tableaux sont les valeurs et dates associées à un titre boursier. L'organisation de la page reste en général stable, ainsi que la liste des titres boursiers.

Dans ce corpus, il nous a donc semblé indispensable de prendre en compte le contexte de la collection pour pallier les difficultés dues à la mauvaise qualité des documents. Il s'agit de mettre en place une stratégie d'analyse de la collection, qui donne des indices sur la manière de mieux reconnaître un document. Les informations provenant de la stratégie globale doivent alors être



(a) Journal de cotation boursière (b) Page d'annuaire d'entreprise

FIGURE 4.12 – Exemples de documents financiers analysés

combinées avec les autres sources d'information, au sein de la description grammaticale. Ainsi, si une page est mal reconnue à cause des dégradations du document, sa reconnaissance peut être améliorée en prenant en compte les connaissances provenant des autres pages de la collection.

4.4.2.2 Analyse itérative transversale

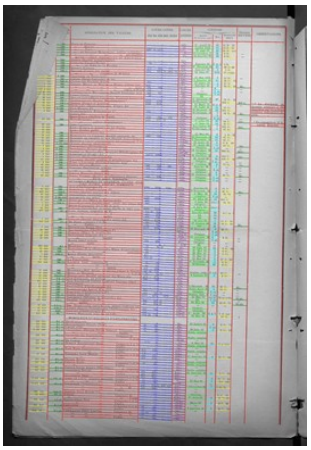
Les mécanismes d'interaction avec la collection, utilisés dans ce contexte, sont ceux proposés dans la thèse de Joseph Chazalon [CC12], présentés dans la partie 2.2.3. Ces mécanismes permettent une analyse itérative des documents au sein de la collection.

Le principe de l'analyse itérative avec stratégie est le suivant : le système de reconnaissance des pages est appliqué sur l'intégralité du corpus. Lors de cette première phase, les pages sont analysées de manière individuelle, sans disposer d'aucun indice sur la collection. L'analyse génère des informations sur les structure rencontrées. Par exemple, sur la figure 4.13a, un premier traitement permet de produire les colonnes présentes dans les documents. Notons que sur le document du 29/06/1899, le document est plié et les deux premières colonnes ne sont pas détectées.

A la fin de cette première itération, les résultats sont collectés par un module de stratégie, qui va répondre aux questions en sollicitant plusieurs acteurs. Il peut s'agir d'un appel à l'OCR, ou bien d'une analyse statistique transversale des résultats. Par exemple, si on sait que le nombre de colonnes des documents est stable dans la collection, le module d'analyse transversale va pouvoir prédire le nombre de colonnes attendu. Dans le cas précis des colonnes, ce traitement transversal sera détaillé dans la section suivante.

Les résultats ainsi produits vont alors être placés en entrée d'une seconde itération : le système va à nouveau traiter de manière individuelle chaque document. La différence est que, lors de cette seconde itération, les résultats de l'analyse transversales sont placés dans un calque perceptif qui contient maintenant les informations provenant de la collection. Par exemple, le calque perceptif peut contenir le nombre de colonnes probable à trouver dans le document. La description grammaticale pourra donc forcer la reconnaissance du nombre de colonnes attendu,

(a) Traitement initial : mauvaise détection des colonnes le 29/06/1899 à cause de la pliure du papier



(b) Résultat du 29/06/1899 après prise en compte du contexte

FIGURE 4.13 – Intérêt de l'utilisation de la collection. Une première phase permet d'estimer que sur cette période, les pages contiennent 10 colonnes. Lors de la seconde phase, on force la grammaire à trouver 10 colonnes, ce qui améliore la reconnaissance sur la page du 29/06/1899.

comme illustré sur la figure 4.13.

4.4.2.3 Analyse transversale des colonnes

Les travaux de thèse de Camille Guerry, dans le cadre du projet ANR HBDEX, portent essentiellement sur la mise au point de la stratégie à adopter : quelles sont les différentes itérations à proposer, et quelles sont les informations qui peuvent être calculées par le module d'analyse transversale des résultats.

Nous détaillons ici le module d'analyse transversale des colonnes, utilisé dans la figure 4.13. Les colonnes sur les documents sont a priori stables sur de longues périodes (plusieurs mois). Il arrive qu'il y ait une rupture, avec l'ajout ou la suppression d'une colonne. Une fois ce changement effectué, le nombre de colonnes reste stable à nouveau pour une grande période. L'objectif est donc de détecter les documents dans lesquels le nombre de colonnes est mal reconnu, mais également les ruptures dans la collection. Ces travaux se sont basés pour cela sur des algorithmes de détection de points de rupture dans des séries temporelles [TOV20].

Une publication est en cours de rédaction. Dans ce papier, nous présentons des résultats qui permettent de montrer :

— la capacité du système à corriger des erreurs de reconnaissance dans des documents bruités, en utilisant le contexte de la collection,
— la capacité du module de validation transversale à détecter et gérer les points de rupture (ajout ou suppression de colonnes dans le modèle).

Nous avons évalué des premiers résultats sur deux ensembles de la bourse La Coulisse de Paris. D'une part 194 images de 1899 (base 1), dans lesquelles le nombre de colonnes est stable, et d'autre part 137 images de 1926 (base 2), dans lesquelles on trouve une rupture de mise en page des colonnes.

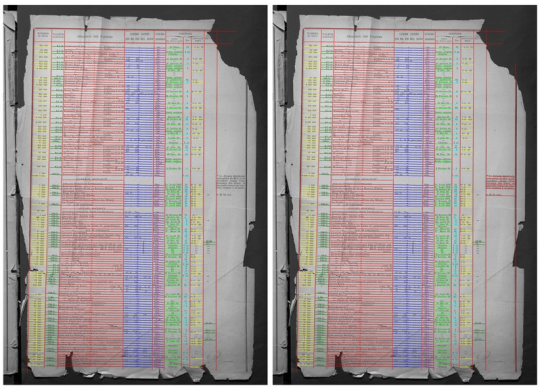
Le tableau 4.2 montre des résultats chiffrés : dans les deux corpus, la validation transversale permet de corriger toutes les erreurs de reconnaissance des colonnes. Par ailleurs, la rupture de mise en page dans la base 2 est bien prise en compte, et ne perturbe pas la correction des erreurs. La figure 4.14 présente des exemples d'images qui ont été corrigées en utilisant le contexte de collection.

	Base 1	Base 2
Nb de pages	194	117
Nb de points de rupture	0	1
Nb d'erreurs avant validation	8	6
Nb d'erreurs après validation transversale	0	0

TABLE 4.2 – Résultats quantitatifs de la correction d'erreurs de détection des colonnes

Les travaux de thèse de Camille Guerry vont maintenant se poursuivre pour exploiter la stabilité de la collection, pour la reconnaissance des différentes sections du tableau, ainsi que pour la reconnaissance de certains champs textuels stables, tels que les intitulés de titres boursiers.

Nous sommes également en train de mettre au point une interface d'interaction avec l'utilisateur : lorsque le module d'analyse transversale ne disposera pas d'assez d'information, il pourra solliciter ponctuellement l'utilisateur pour une prise de décision. Les résultats de l'interaction avec l'utilisateur seront alors inclus dans le processus d'analyse grammatical, comme une source de données supplémentaire.



(a) Création d'une colonne manquante à droite (en rouge)

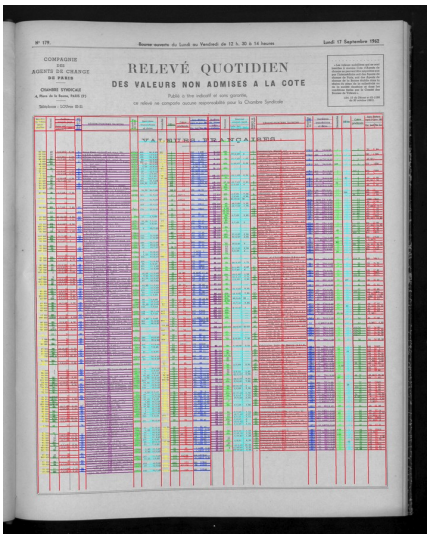
(b) suppression d'un faux positif dans la marge de droite (en bleu)

FIGURE 4.14 – Exemple de corrections réalisées par l'analyse transversale de la collection

4.4.2.4 Généralisation

Dans le projet EurHisFirm, il est prévu de pouvoir appliquer les travaux sur les listes de prix de La Coulisse à des corpus similaires provenant de plusieurs pays. Notre objectif est de s'assurer que les descriptions de documents sont assez génériques pour pouvoir être appliquées sur des documents d'autres pays, tout en identifiant les parties spécifiques à décrire.

Des premiers tests ont été lancés par exemple pour adapter la description des listes de prix à des listes de prix Belges (figure 4.15).



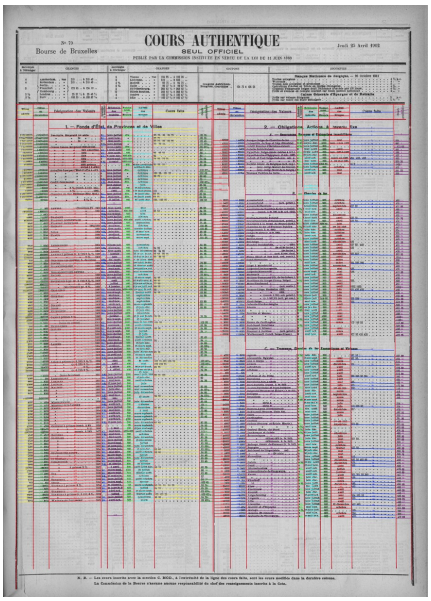


FIGURE 4.15 – Application des systèmes de reconnaissances sur les listes de prix belges, Bruxelles 1912

Un des principaux avantages de l'utilisation d'une description à base de règle est que le système ne va pas nécessiter de ré-apprentissage pour s'adapter à un autre corpus. Ainsi, un tableau sur une liste de prix belge respecte la même organisation physique qu'un tableau sur des listes de prix françaises. Il va simplement être nécessaire d'enrichir la description grammaticale pour pouvoir gérer plusieurs cas. Par exemple, il faudra prendre en compte la possibilité d'avoir

deux tableaux côte à côte dans les listes belges. Cela peut se faire de manière très simple. Préalablement, la règle décrivait qu'un page était composée d'un tableau, il suffit maintenant de dire qu'une page peut aussi être composée de deux tableaux côte à côte. La définition d'un tableau ne change pas.

4.4.2.5 Bilan sur la combinaison de données hétérogènes

Ces travaux sur les listes de prix font partie des travaux les plus récents de ce manuscrit. L'utilisation conjointe d'un grand nombre de sources de données est devenue une habitude dans l'élaboration des descriptions grammaticales avec DMOS. Dans ce corpus, la combinaison de données hétérogènes est finalement basée sur (figure 4.16) :

— des primitives images : les segments sont détectés pour construire les filets du tableau, un détecteur de pointillés est appliqué pour gérer les cellules sans texte.
— des résultats issus de l'analyse par des réseaux de neurones profonds : un réseau ARU-net[Grü+18] est appliqué pour la localisation de lignes de texte,
— l'utilisation d'OCR, d'une part un OCR commercial Abbyy Fine Reader, et d'autre par un OCR fourni par nos partenaires du LITIS, entraîné spécifiquement sur ce corpus.
— des données d'interaction, provenant de l'analyse transversale de la collection (telles que celles détaillées sur la fiabilisation des colonnes). Il est également prévu des interactions avec l'utilisateur pour lever des ambiguïtés.

Tous ces éléments sont stockés dans des calques perceptifs, et combinés par des règles symboliques de description physique de la structure tabulaire, ainsi que des règles de description logique portant sur la sémantique des lignes et des colonnes dans les tableaux de cours boursiers.

	Primitives images		Règles symboliques		Objets construits	Résultats de systèmes d'apprentissage	Systèmes de reconnaissance d'écriture	Interaction	
	Segments	Pointillés	Organisation physique	Organisation logique	Lignes de texte	Réseaux de neurones profonds	OCR commercial	Avec la collection	Avec l'utilisateur
Listes de prix HBDEX	✓	✓	✓	✓	✓	✓	✓	✓	✓

FIGURE 4.16 – Données hétérogènes combinées pour la reconnaissance de listes de prix

Les travaux sur ces documents sont probablement ceux sur lesquels nous avons réalisé la combinaison avec le plus grande variété de données hétérogènes. Ces travaux peuvent être vus comme une validation d'une utilisation complexe des mécanismes de calques perceptifs et de description grammaticale 3D.

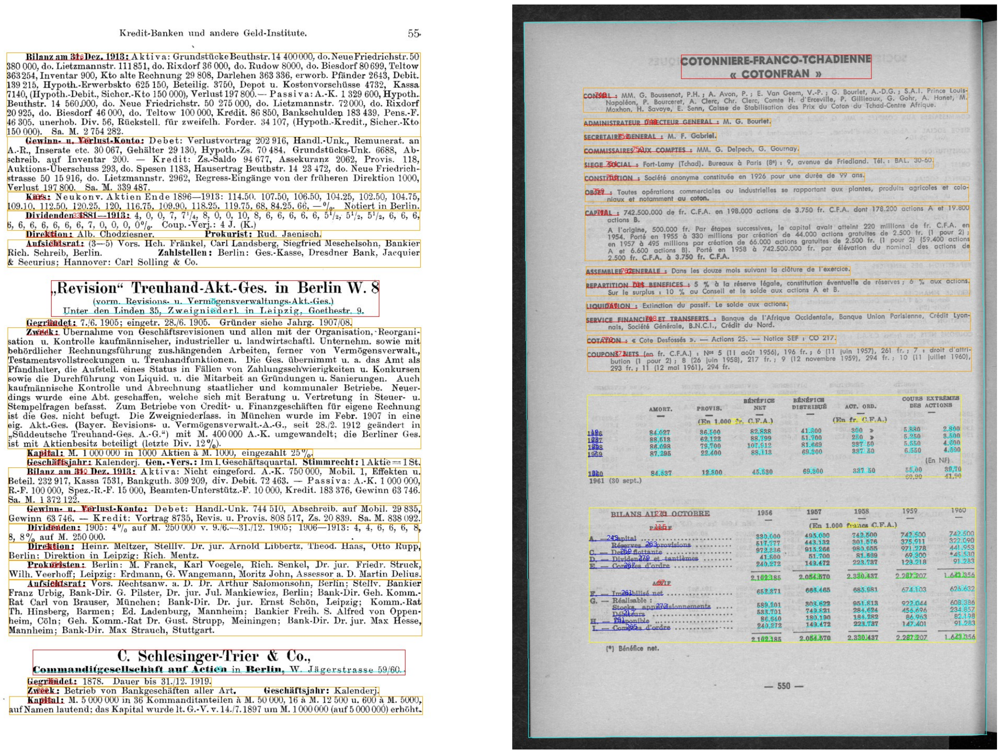
4.4.3 Annuaires d'entreprises

Nous nous focalisons maintenant sur l'analyse des annuaires d'entreprises, qui sont un des corpus étudiés dans le cadre du projet européen EurHisFirm.

4.4.3.1 Spécificités du corpus

Le corpus est constitué d'annuaires donnant des caractéristiques sur les entreprises cotées en bourse. Deux corpus sont étudiés actuellement : l'annuaire français Desfossé et le Handbuch allemand. Ces annuaires sont organisés en fiches associées à des entreprises (émetteurs). Deux exemples de pages sont présentés sur la figure 4.17.

Chaque fiche est constituée d'un ensemble de rubriques. Les rubriques correspondent globalement à des paragraphes, commençant par un titre. On trouve également des tableaux dans les rubriques des annuaires français. La particularité de ces tableaux est qu'il ne sont pas délimités par des filets (figure 4.20b).



(a) Handbuch (b) Annuaire Desfossé

FIGURE 4.17 – Exemples de pages d'annuaires étudiés : les rubriques sont encadrées en orange, le titre des rubriques est souligné en rouge. Les tableaux détectés sont encadrés en jaune.

4.4.3.2 Méthodes proposées

Dans ce corpus, il s'agit essentiellement de segmenter les pages en rubriques (en orange sur la figure 4.17). On localise les titres comme des éléments centrés dans la page, en caractères de grande taille. Puis les rubriques sont matérialisées par un titre de rubrique, suivi de lignes de texte. La difficulté est d'identifier les titres de rubrique. Dans les annuaires français, un titre de rubrique est caractérisé par un alinéa. Dans les annuaires allemands, les alignements ne permettent pas de prendre une décision.

Pour localiser les titres de rubriques dans les annuaires allemands, nous avons donc entraîné un réseau de type ARU-net [Grü+18], qui prédit la position des titres de rubriques. La figure 4.18 présente un exemple de prédictions. Ces prédictions sont placées dans un calque puis combinées avec la présence d'autres lignes de texte pour construire une segmentation de la page en rubrique. Un exemple de résultat est présenté sur la figure 4.20a.

Norderelbe-Dampfschifffahrts-Gesellschaft in Hamburg.
(In Liquidation.)
Gegründet: 1867. Die a.o. G.-V. 27./8. 1913 beschloss den Verkauf des D. Fortuna an eine Genossenschaft m. b. H. u. die Auflös. der Akt.-Ges.
Kapital: M. 48 000 in Nam.-Aktien Lit. A à M. 300, Lit. B à M. 60.
Geschäftsjahr: Kalenderj. Gen.-Vers.: Bis April. Stimmrecht: 1 Aktie = 1 St.
Bilanz am 31. Dez. 1912: Aktiva: Dampfschiff „Fortuna II" 36 000, Bankguth. 3485, Kassa 1494, Wertp. 70 476. — Passiva: A.-K. 48 000, Div. 3600, R.-F. 48 000, Disposit.-F. 11 857. Sa. M. 111 457.
Gewinn- u. Verlust-Konto: Debet: Unk. 31 427, Div. 3600, Wertp., Abschreib. 1042. — Kredit: Passage u. Frachten 23 065, div. Einnahmen 9946, Zs. 3057. Sa. M. 36 070.
Dividenden 1894—1912: 4, 4, 5, 5, 6, 7½, 8, 8, 9, 10, 12, 16, 20, 18, 19, 18, 18, 18, 7¼%.
Liquidatoren: P. E. Bielenberg, Dr. jur. Joh. Friedr. Voigt. Aufsichtsrat: Heinr. Schween, P. E. Bielenberg jr., Heinr. Lemke, Wilh. Annen, Heinr. Bartels.

FIGURE 4.18 – Exemple de prédictions de titres de rubriques dans un système entraîné avec ARU-net

La seconde grande difficulté dans ce corpus est l'absence de filets dans les tableaux. Cependant, l'humain est capable d'analyser le contenu en repérant de manière intuitive des éléments alignés. Pour résoudre cette difficulté, nous avons construit un détecteur d'alignements. Il s'agit de regrouper en blocs, les lignes de texte suffisamment proches, partageant un même alignement droite ou gauche. La figure 4.19 présente des exemples d'alignement trouvés. Notre détecteur d'alignements est une grammaire propre, basée sur la méthode DMOS. Elle prend en entrée les lignes de texte et construit les alignements. Il s'agit d'une description générique, qui exprime un alignement comme des lignes de textes ayant une indentation régulière, indépendamment du corpus de documents traité. Les alignements ainsi trouvés sont placés dans un calque et utilisés en entrée de la grammaire de reconnaissance des pages d'annuaires d'entreprise.

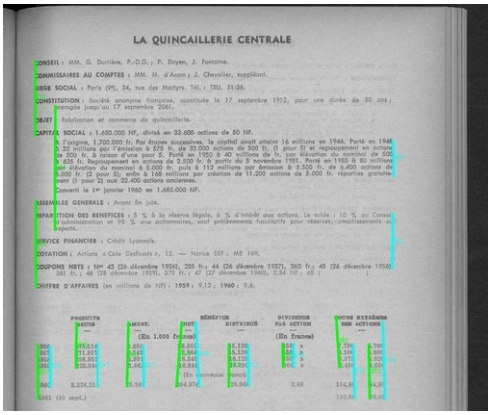
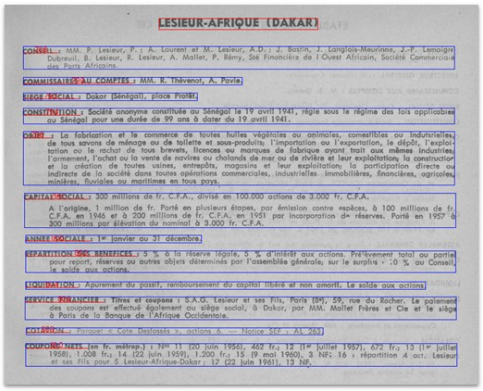
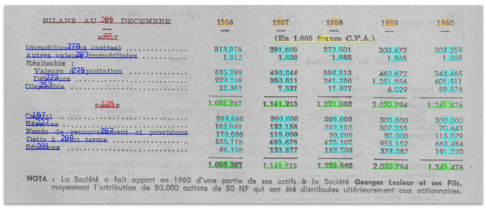


FIGURE 4.19 – Exemple de détection d'alignements dans une image : alignements à gauche en vert, alignements à droite en cyan

Muni de ce calque d'alignements, on peut décrire un tableau comme constitué d'une ligne d'entête, sous laquelle sont placées des données en alignements verticaux. Un exemple de tableau reconstruit est présenté dans la figure 4.20b.



(a) Localisation des rubriques



(b) Segmentation des tableaux sans filets

FIGURE 4.20 – Segmentation des pages d'émetteurs

4.4.3.3 Résultats

Nous avons évalué les résultats de segmentation des fiches d'émetteurs. On utilise la métrique ZoneMap [GKO14] (utilisée pour la compétition Maurdor) , qui est particulièrement bien adaptée pour évaluer les problèmes de sur-segmentation et sous-segmentation. Cette métrique fournit un score d'erreur de surface, qui doit être le plus près de 0 possible, mais peut dépasser 100 lorsque

le résultat contient plus d'erreurs que d'informations correctes. Les résultats sont présentés dans le tableau 4.3.

Dans les pages françaises de l'annuaire Desfossé, il s'agit d'isoler les titres, les rubriques, les tableaux de bilan et les autres tableaux. Les pages allemandes du Handbuch ne contiennent que des titres et des rubriques.

Base de données	Nombre de documents	Titre	Rubrique	Tableau de bilan	Autre tableau
Desfossé	61	14.94	3.06	9.08	35.13
Handbuch	172	12.12	3.32	-	-

TABLE 4.3 – Scores de segmentation (erreur ZoneMap) sur les annuaires Desfossé, et Handbuch

4.4.3.4 Bilan sur la combinaison de données hétérogènes

Tout comme les listes de prix HBDEX, les annuaires d'entreprise du projet EurHisFirm sont des travaux assez récents qui nous ont permis de mettre en place une collaboration entre des données de types variés (figure 4.21).

	Primitives images		Règles symboliques		Objets construits	Résultats de systèmes d'apprentissage		Systèmes de reconnaissance d'écriture
	Segments	Composantes connexes	Organisation physique	Organisation logique	Alignements	Réseau de neurones Lignes de texte	Réseau de neurones Titres de rubriques	OCR commercial
Annuaires EurHisFirm	✓	✓	✓	✓	✓	✓	✓	✓

FIGURE 4.21 – Données hétérogènes combinées pour la reconnaissance d'annuaires d'entreprises

Ainsi, dans ces documents, la combinaison de données hétérogènes est constituée de :
- des primitives images : les composantes connexes permettant de confirmer la présence de texte, ainsi que les segments, quand ils sont présents dans les structures tabulaires ;
- les alignements qui sont des objets construits par une autre description grammaticale, et qui permettent de gérer l'absence de filets dans les tableaux ;
- des données produites par l'application de réseaux de neurones : la localisation des lignes de textes, ainsi que la localisation des titres de rubriques dans les annuaires allemands ;
- le résultats de l'OCR commercial Abbyy Fine Reader.

La combinaison de ces données est guidée par des règles sur l'organisation physique et logique des documents.

Chapitre 5

Analyse de formulaires pré-imprimés

L'analyse d'images de formulaires pré-imprimés peut sembler une tâche simple, voire triviale. En effet, lorsque le pré-imprimé est bien connu, il suffit de comparer les images de documents complétés avec l'original vierge pour trouver quels champs ont été remplis.

Pourtant, dans de nombreux cas, la tâche n'est pas aussi aisée. Lorsqu'on traite des formulaires historiques, sur une grande période, deux problèmes se présentent : le formulaire vierge, n'est pas toujours disponible, et la mise en page du formulaire peut varier au cours du temps, au sein d'une collection de documents. C'est dans ce contexte que j'ai travaillé sur l'analyse d'actes de mariages mexicains.

L'extraction d'information peut également être rendue difficile, même dans des documents standardisés de type CERFA, par la manière de remplir les formulaires, la superposition entre des champs imprimés et le texte manuscrit, les mauvaises conditions de numérisation. J'aborderai cet exemple dans un second temps.

5.1 Actes de mariages mexicains

De 2013 à 2017, j'ai travaillé sur l'analyse d'actes de mariages mexicains. Ces travaux m'ont permis d'explorer de nouvelles pistes de combinaisons de données hétérogènes, notamment la recherche de mots clés avec des points d'intérêts, ainsi que l'inférence grammaticale de règles.

5.1.1 Corpus et compétition

Nous travaillons sur un corpus qui a été proposé par Family Search dans le cadre d'une compétition organisée pour HIP'2013, un workshop satellite d'ICDAR sur les images historiques [LC13].

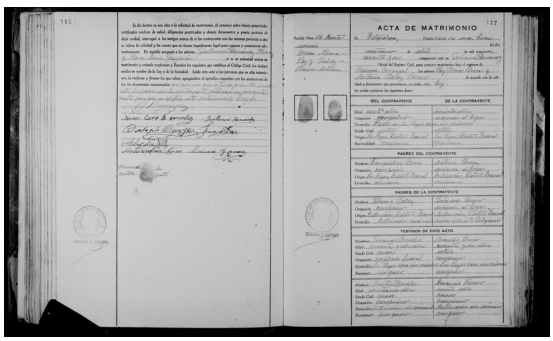


FIGURE 5.1 – Exemple de page d'acte de mariage mexicain

Le corpus est constitué de 30 000 pages d'actes de mariages mexicains de 1946 à 1950 (figure 5.1), qui sont des pré-imprimés remplis à la main. Pour la compétition proposée lors de HIP'2013,

il s'agissait de localiser quatre régions d'intérêts dans les actes : le mois, l'année de l'acte, et les villes d'origines pour les deux mariés. La métrique de la compétition évaluait ensuite la capacité des systèmes à regrouper des imagettes contenant le même texte. Par exemple, il fallait regrouper toutes les imagettes portant le mois de "juin". La transcription des imagettes n'était pas demandée.

Nous avons choisi de nous focaliser sur la première partie de ce problème, à savoir trouver la localisation de l'année et du mois de l'acte de mariage (figure 5.2).

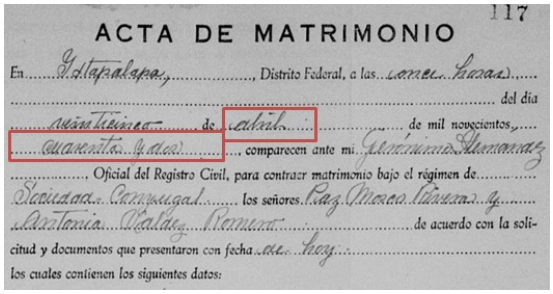


FIGURE 5.2 – Tâche étudiée : localisation des champs manuscrits du mois et de l'année. Le mois est précédé du mot *de*, l'année de *de mil novecientos*

.

Les difficultés de cette problématique sont les suivantes (voir figure 5.3) :

— les difficultés propres aux documents anciens, abîmés, avec une encre parfois pâle ;
— les zones de superposition entre le texte manuscrit et le texte imprimé, qui compliquent l'extraction de mots clés imprimés ;
— l'hétérogénéité de la base de données : le corpus est présenté dans un ordre aléatoire, les types de pré-imprimés varient, et on ne connaît pas le nombre de pré-imprimés différents au sein de la collection.

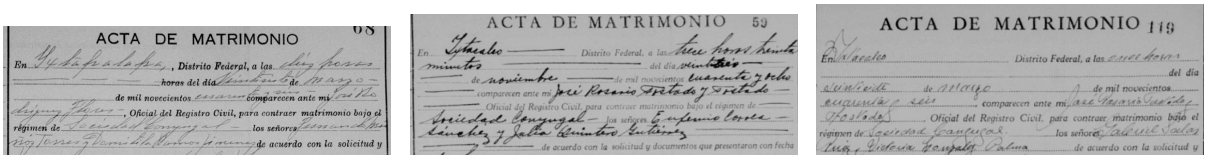


FIGURE 5.3 – Difficultés du corpus : documents anciens dont l'encre présente un contraste variable, interaction entre manuscrit et imprimé, variété dans les types de pré-imprimés (voir les différentes position du mot "del dia" sur la deuxième ligne).

5.1.2 Stratégie globale d'analyse

Pour extraire les champs mois et année des actes de mariage, nous avons mis en place une description grammaticale permettant d'appliquer la stratégie suivante :

1. dans la page, trouver la position de la colonne de droite, puis se focaliser sur la partie supérieure de la colonne ;
2. localiser les mots clés pré-imprimés qui délimitent les champs d'intérêt : "de" et "de mil novecientos" pour le mois ; "de mil novecientos" et "comparecen" pour l'année (figure 5.2) ;
3. construire les champs de texte manuscrits entre les mots clés.

Pour appliquer cette description grammaticale, il est nécessaire de combiner deux types de données : des segments verticaux permettant de localiser la structure physique des pages, des mots clés imprimés (figure 5.4).

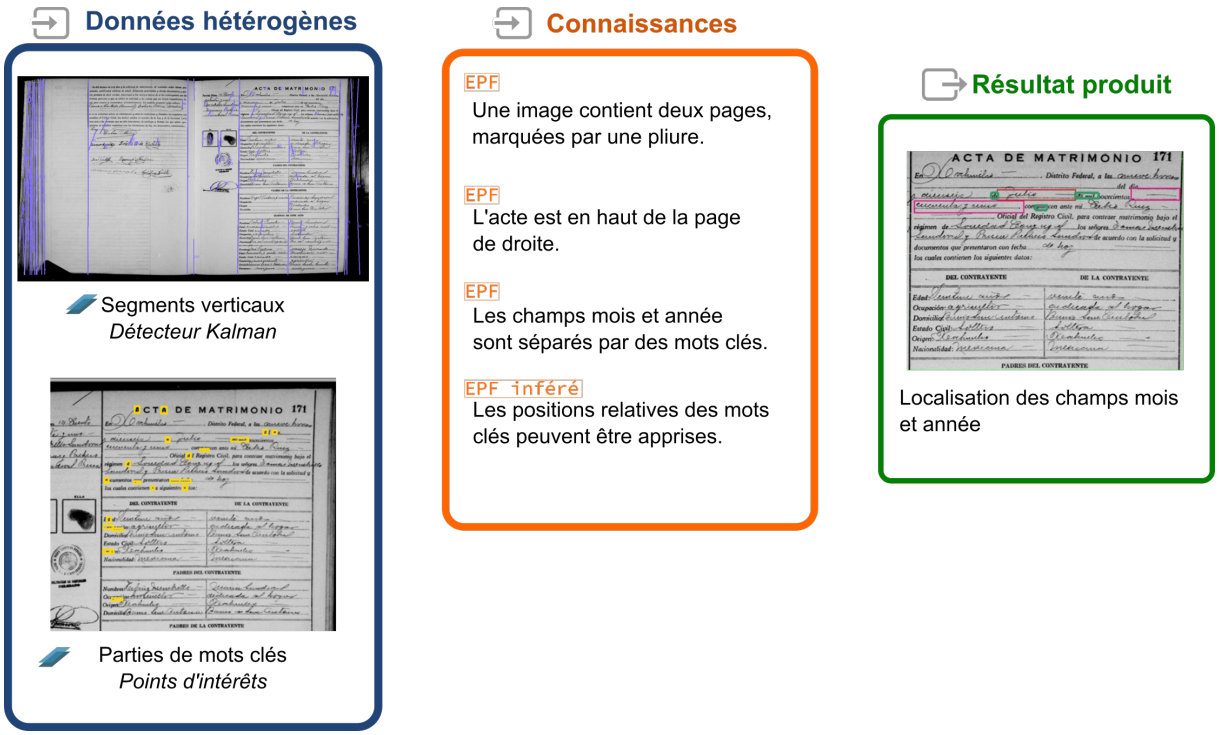
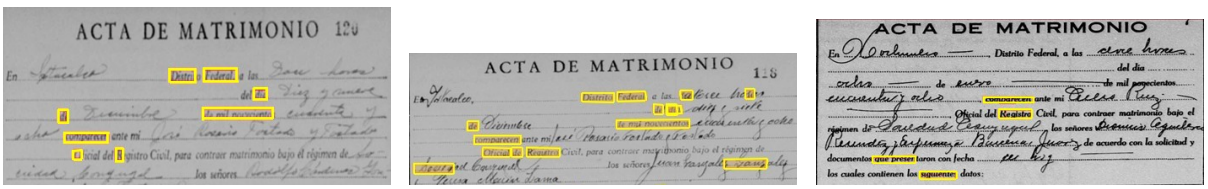


FIGURE 5.4 – Combinaison de données hétérogènes pour la localisation de champs manuscrits dans des actes de mariages mexicains.

Pour l'extraction des mots clés imprimés, nous avons testé deux méthodes. Nous avons tout d'abord voulu utiliser un OCR du commerce, Abbyy Fine Reader. Cependant, les résultats obtenus par cet OCR ne sont pas satisfaisants, principalement à cause de l'interaction forte entre le texte manuscrit et le texte imprimé : de nombreuses zones du document sont considérées comme zones graphique par l'OCR, qui n'y détecte pas de mots imprimés.

Nous avons donc choisi d'appliquer un mécanisme de word-spotting pour lequel les mots clés sont détectés en fonction de la présence de points d'intérêts au niveau pixel. La figure 5.5 montre un exemple des mots clés détectés par la méthode des points d'intérêt. Les résultats sont de meilleure qualité que ceux obtenus par l'OCR, mais la superposition avec le texte manuscrit provoque toutefois des erreurs de reconnaissance.



(a) Cas idéal : détection de 8 mots clés imprimés

(b) Faux positifs : détection de mots clés dans le texte manuscrit

(c) Oublis dûs à l'interaction entre imprimé et manuscrit : par exemple "de mil novecientos"

FIGURE 5.5 – Exemples de mots imprimés localisés par la méthode des points d'intérêt (en jaune)

Dans ce cas, la stratégie de combinaison des données hétérogènes doit prendre en compte l'imprécision des mots clés : il est nécessaire de prévoir les cas où des mots clés sont trouvés en

plus, et les cas où les mots clés sont manquants.

5.1.3 Inférence grammaticale à partir de données non étiquetées

Nous avons envisagé plusieurs stratégies pour pallier l'imprécision dans la détection des mots clés imprimés. Ces stratégies sont détaillées dans [Lem+18]. Nous nous focalisons ici sur la stratégie la plus originale : l'analyse de masses de données pour l'inférence grammaticale. Il s'agit d'une partie du travail de thèse de Cérès Carton [Car16], évoqué dans la section 2.3.2.

Le principe est le suivant : à partir d'une vérité terrain de la position des mots clés imprimés, il est possible, par clustering, d'inférer les différents modèles de pré-imprimés disponibles dans la base. La connaissance de ces modèles nous permet alors, en production, d'appliquer le modèle qui correspond le mieux à un document, même si tous les mots clés ne sont pas détectés. Mais ceci nécessite une base annotée des mots clés. Nous proposons de construire automatiquement une pseudo-vérité terrain du corpus, en étudiant les redondances de configurations dans une grande masse de données non annotées.

Nous avons ainsi utilisé la méthode EWO, présentée dans la section 2.3.2, pour réaliser deux phases de l'analyse :

— la construction d'une pseudo-vérité terrain, avec un clustering automatique validé par l'utilisateur,
— l'analyse non supervisée du corpus pour générer automatiquement les différents modèles de documents.

5.1.3.1 Construction d'une pseudo vérité terrain

Dans notre contexte, l'objectif est de détecter automatiquement le nombre de modèles de pré-imprimés présents dans la base d'apprentissage.

EWO prend en entrée les mots clés extraits par la méthode des points d'intérêts. L'analyse est basée sur 8 mots clés (figure 5.5a). Comme montré sur la figure 5.5, le résultat obtenu n'est pas parfait, on note la présence d'oublis et de faux positifs. Cependant, la grande quantité de données (7 000 images non annotées) permet d'étudier les redondances dans les mots clés trouvés. Ainsi, la méthode EWO génère automatiquement des clusters de mots clés, en fonction de leur position, leur label et leur dimension. Ce clustering est basé sur l'EAC (Evidence Accumulation Clustering) [FJ02].

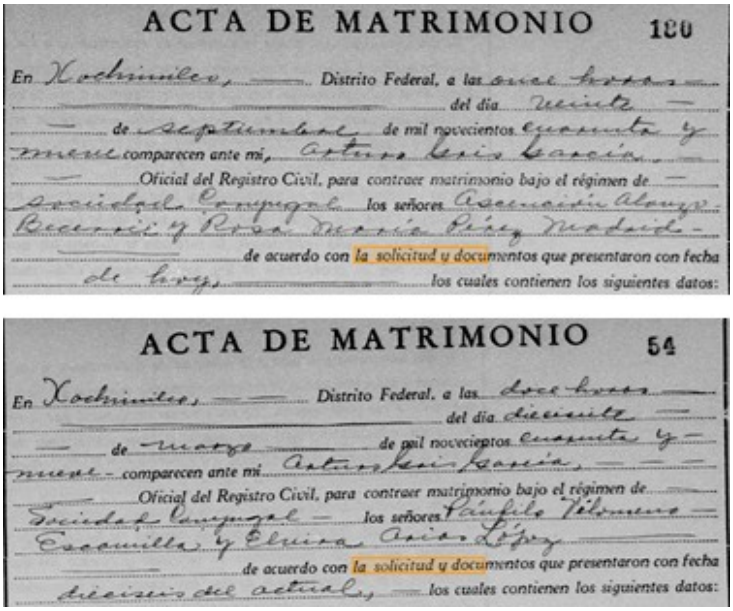
En sortie, pour chaque mot clé, EWO construit des clusters de mots clés similaires, à la même position. Une interaction utilisateur est alors requise : pour chaque cluster, quelques exemples sont présentés à l'utilisateur (figure 5.6). En voyant seulement quelques images de chaque cluster, l'utilisateur valide l'intégralité de la vérité terrain associée au reste du cluster.

Suite à ces interactions de validation de cluster, le système génère une pseudo-vérité terrain pour les mots clés. Cette vérité terrain est partielle, mais la grande masse de donnée initiale permet de disposer de suffisamment d'informations pour permettre une étape d'apprentissage.

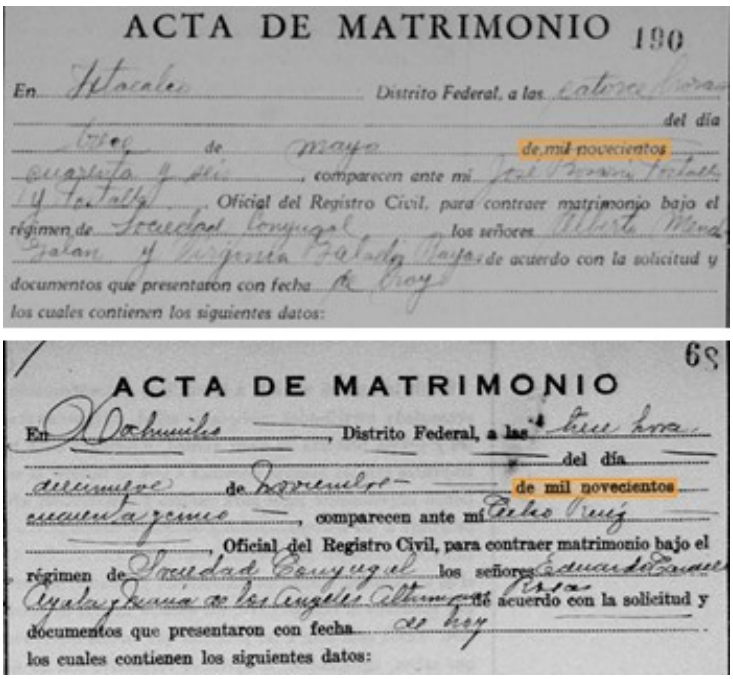
5.1.3.2 Inférence des modèles de formulaires

La présence d'une pseudo vérité-terrain permet d'inférer les différents modèles de formulaires. On considère qu'un modèle de pré-imprimé est caractérisé par le position des 8 mots clés étudiés. Le système se focalise sur les documents pour lesquels les 8 mots clés ont été validés dans la pseudo-vérité terrain, ce qui représente 5 406 documents sur les 7000 documents initiaux. L'analyse de la position relative des 8 mots clés du document permet de constituer une signature du document. La méthode EWO est alors utilisée pour construire des clusters de documents ayant des signatures similaires, toujours avec le clustering EAC.

La méthode EWO produit automatiquement les modèles de pré-imprimés, ici 11 modèles dans la vérité terrain. Des règles de grammaire sur la position des éléments sont alors inférées pour permettre la localisation automatique des champs.



(a) Exemple de cluster que l'utilisateur va rejeter



(b) Exemple de cluster que l'utilisateur va accepter

FIGURE 5.6 – Exemple d'interaction avec l'utilisateur pour la sélection des clusters : en voyant seulement deux images de chaque cluster (ici pour le mot clé "de mil novecientos") l'utilisateur valide l'intégralité de la vérité terrain associée au reste du cluster.

5.1.3.3 Résultats obtenus

Nous avons appliqué la méthode EWO sur 7000 pages de la base d'apprentissage, ce qui nous a permis de détecter 11 modèles de pages différents. Nous avons ainsi pu constater la non équi-répartition des modèles, le modèle le plus fréquent représentant 20% des formulaires, le modèle le moins fréquent correspondant seulement à 25 images sur 7000.

Nous avons annoté manuellement les positions des mois et des années sur 2000 pages, afin de créer une base de test pour évaluer notre système. Nous avons réalisé plusieurs expériences comparatives qui sont présentées dans [Lem+18].

Avec la méthode combinant les données présentées sur la figure 5.4, nous obtenons un taux de localisation des champs de 97.2%.

5.1.4 Bilan sur la combinaison de données hétérogènes

L'analyse du corpus de mariages mexicains nous a permis de valider une nouvelle combinaison de données (figure 5.7). Le contenu est décrit par des règles d'organisation logique, et de structure physique. Ces règles permettent de combiner des primitives images (segments), le résultat d'un extracteur de mots clés avec points d'intérêts, et des résultats d'analyse statistique portant sur la collection.

	Primitives images	Règles symboliques		Systèmes de reconnaissance d'écriture	Interaction
	Segments	Organisation physique	Organisation logique	Points d'Intérêts sur l'écriture (POI)	Avec la collection
Actes de mariages mexicains	✓	✓	✓	✓	✓

FIGURE 5.7 – Données hétérogènes combinées pour la reconnaissance de registres de mariages mexicains

5.2 Documents CERFA

Les documents administratifs français de type CERFA sont typiquement des formulaires imprimés, remplis de façon manuscrite ou imprimés, dont l'analyse peut être réalisée par une description grammaticale. Nous avons ébauché un travail sur ces documents dans le cadre du projet IAT.

5.2.1 Projet IAT

En 2019, nous avons pris part au projet IAT, porté par David Gross Amblard de l'IRISA. Ce projet d'Intelligence Artificielle pour l'Administration Territoriale, vise à produire un démonstrateur sur une manière d'utiliser l'IA dans le traitement des documents de la préfecture de la région Bretagne. Pour ce projet, j'ai co-encadré l'ingénieure Solène Tarride sur une durée de 2 mois.

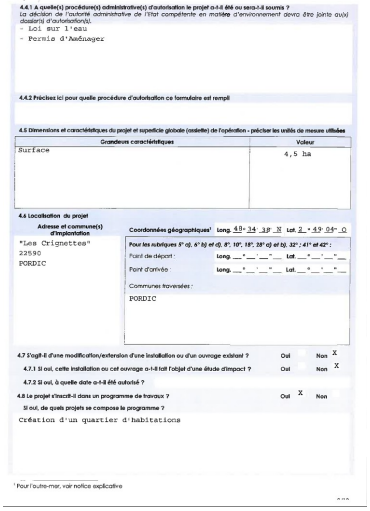
Le prototype a porté sur l'analyse de dossiers d'"Examen au cas par cas à la réalisation d'une étude d'impact". Ces formulaires sont remplis lors de l'installation d'infrastructures pouvant avoir un impact sur l'environnement. La DREAL (Direction Régionale de l'Environnement, de l'Aménagement et du Logement) est chargée d'instruire ces dossiers et d'évaluer si oui ou non une étude d'impact plus poussée est nécessaire.

La DREAL souhaiterait, lors de l'instruction d'un dossier, pouvoir accéder plus rapidement aux conclusions qui auraient été prises précédemment sur des dossiers similaires, de par leur localisation ou leur nature. Nous avons mené des travaux d'extraction d'information dans ces documents. Les informations extraites ont ensuite été traitées par les autres partenaires du projet.

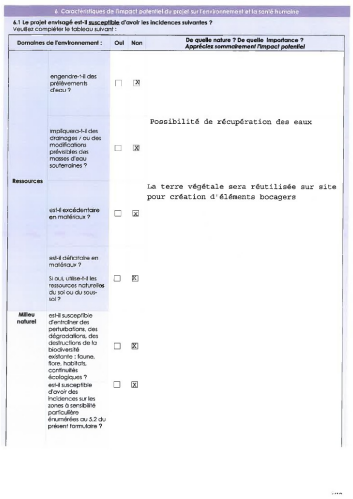
5.2.2 Extraction de champs spécifiques

Les dossiers traités sont constitués de formulaires CERFA, suivis par des pages d'annexes à formulation libre. Étant données les contraintes de temps du projet (2 mois), nous avons ciblé le travail sur l'extraction de champs spécifiques au sein des formulaires CERFA. La figure 5.8 présente des exemples de pages traitées. L'objectif est d'extraire :

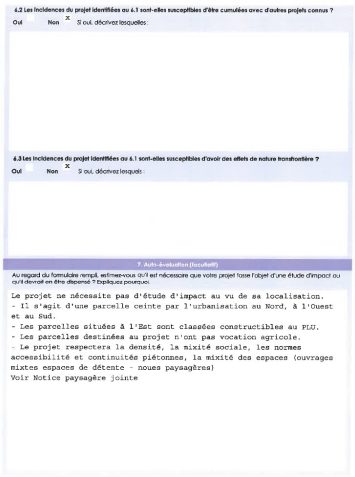
— les informations de localisation du projet : coordonnées GPS et adresse (figure 5.8a),
— les informations renseignées par des cases à cocher OUI/NON (figure 5.8b),
— les informations contenues dans le champs de texte libre d'évaluation du projet (figure 5.8c).



(a) Page contenant des informations de localisation du projet



(b) Page contenant des informations sous forme de cases à cocher



(c) Page contenant une évaluation sous forme de texte libre

FIGURE 5.8 – Exemple de documents CERFA analysés : extraits de dossiers d'"Examen au cas par cas à la réalisation d'une étude d'impact"

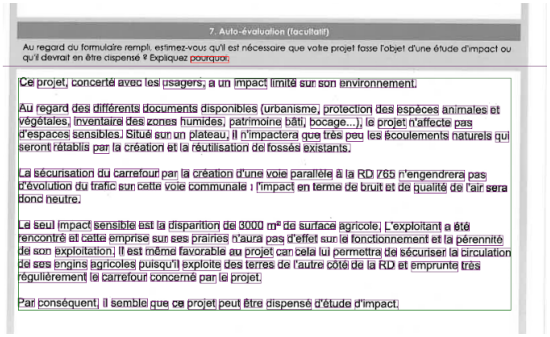
Pour extraire ces champs spécifiques, nous avons réalisé une description grammaticale simple. Nous avons considéré que les mots clés imprimés dans les formulaires pouvaient être suffisants pour localiser les champs d'intérêt. Nous avons donc appliqué un OCR commercial, Abbyy Fine Reader sur les pages de documents. La description grammaticale est basée sur la présence de mots clés prédéfinis. Concernant les cases à cocher, nous avons étudié les composantes connexes. Une comparaison du nombre de pixels noirs entre la case "oui" et la case "non" a permis d'identifier simplement la case cochée. L'expressivité de la méthode DMOS a permis de décrire ces règles de manière simple.

5.2.3 Résultats

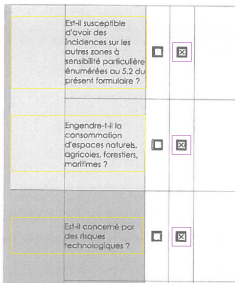
Ce premier prototype a été appliqué sur un lot de 1000 documents d'examens au cas par cas. La figure 5.9 présente des exemples de résultats.

Ces travaux préliminaires ont surtout permis de montrer :

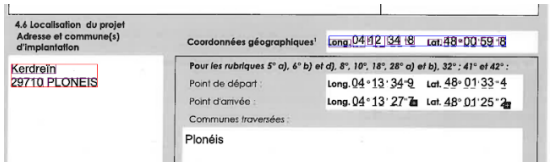
— la faisabilité d'un système d'extraction de données, basé sur une description grammaticale,
— la nécessité de disposer de bonnes conditions de numérisation des documents, puisque cela influe fortement sur les capacités de l'OCR,
— la possibilité d'inclure les résultats de reconnaissance de documents dans une chaine plus globale, permettant une exploitation statistique des données.



(a) Analyse des champs d'auto-évaluation



(b) Analyse des cases à cocher



(c) Analyse des champs de localisation

FIGURE 5.9 – Exemples de résultats

5.2.4 Bilan sur la combinaison de données

La figure 5.10 présente les données hétérogènes qui sont donc combinées pour ces documents. Il s'agit essentiellement du résultat d'un OCR, guidé par des règles d'organisation physique des contenus (figure 5.10). Les composantes connexes sont utilisées également pour typer les cases à cocher. A l'heure actuelle, il n'est pas exploité de connaissances logiques sur l'organisation des documents.

	Primitives images	Règles symboliques	Systèmes de reconnaissance d'écriture
	Composantes connexes	Organisation physique	OCR commercial
CERFA Cas par cas	✓	✓	✓

FIGURE 5.10 – Données hétérogènes combinées pour la reconnaissance de documents CERFA

Dans des prochains travaux, nous aimerions pouvoir généraliser l'analyse des formulaires de type CERFA. Une piste serait de voir comment faire pour générer les règles de description de l'organisation physique des contenus, de manière automatique en connaissant quelques exemplaires des formulaires initiaux.

Chapitre 6

Analyse interactive de documents

J'évoque ici la thèse d'Achraf Ghorbel [Gho12] que j'ai co-encadrée avec Eric Anquetil dans le cadre du projet ANR Mobisketch, entre 2009 et 2012. Ces travaux se placent dans un contexte légèrement différent de ceux présentés dans le reste du manuscrit. En effet, il s'agit ici de réaliser une analyse de documents *interactive*, en présence de l'utilisateur qui peut interagir de manière synchrone. Notons que dans les autres travaux présentés dans ce manuscrit, les mécanismes d'interactions utilisateurs sont asynchrones.

Dans ce chapitre, nous présentons le contexte du projet Mobisketch, avant de détailler la méthode Imisktech mise en place, et de donner quelques résultats chiffrés de ces travaux.

6.1 Projet Mobisketch

Le projet ANR Mobisketch vise à élaborer une solution logicielle générique orientée stylo pour la réalisation de documents techniques : schémas, plans... L'objectif est d'aboutir à un *continuum* entre un document technique sous sa forme papier et ce même document sous sa forme numérique interprétée. Ainsi, la figure 6.1 présente les différentes étapes souhaitées lors de l'analyse : l'utilisateur saisit un croquis sur papier, il prend en photo son schéma, le schéma est reconnu puis l'utilisateur peut poursuivre l'édition de son schéma de manière numérique.

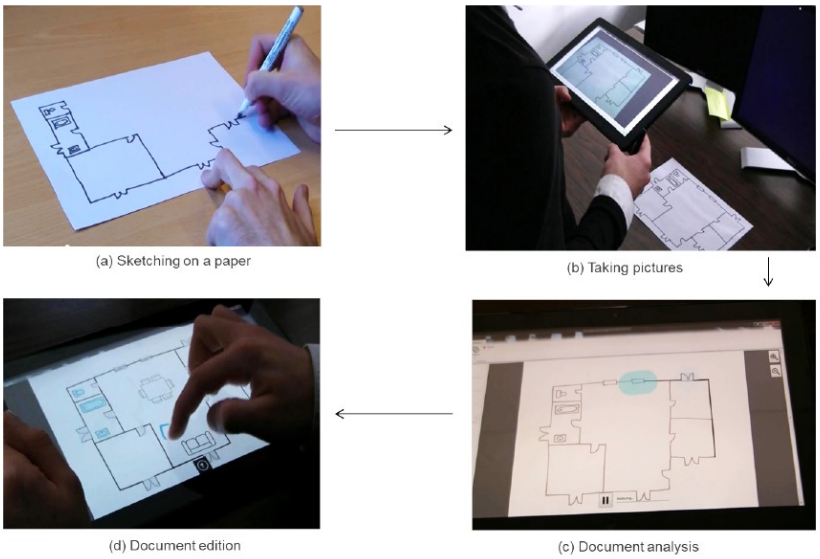


FIGURE 6.1 – Mobisketch : un continuum entre un document papier et le même document sous sa forme digitale interprétée.

Pour ce projet ANR, les travaux ont été menés en partenariat avec le CRPCC de l'Université

Rennes 2 (laboratoire de psychologie : cognition, comportement et communication), ainsi qu'avec la société Script&Go. La collaboration avec le CRPCC a notamment permis d'étudier l'impact du choix des interfaces pour les différentes étapes d'interaction avec l'utilisateur.

Ces travaux se placent à la croisée de deux thématiques étudiées dans l'équipe Intuidoc : l'analyse de documents hors-ligne, et l'édition de documents à la volée sur tablette. Ainsi, l'expertise de l'équipe se concentre autour de deux frameworks, la méthode DMOS [Coü06] pour l'analyse de documents hors-ligne, et le framework DALI [MA09] pour l'interprétation et l'édition de documents dessinés à la volée. Il était nécessaire ici de mettre en place un continuum entre ces deux étapes, et donc de disposer de deux analyseurs cohérents : un pour la phase de reconnaissance et un autre pour la composition/édition.

Le travail de thèse d'Achraf Ghorbel a consisté à réaliser un système de reconnaissance interactive de documents, qui puisse prendre en compte la présence de l'utilisateur pour valider ou corriger la reconnaissance. Nous avons donc mis en place un système, s'inspirant de la méthode DMOS pour la reconnaissance de documents, mais basé sur le formalisme DALI [MA09], de manière à garder une cohérence avec l'étape d'édition des documents, qui est basée sur DALI. C'est ce que nous appellerons le système Imisketch.

6.2 Méthode Imisketch

La méthode Imisketch est composée de quatre blocs majeurs, présentés sur la figure 6.2 :
- un module de pré-traitement, chargé de l'extraction des primitives dans l'image,
- un module de description grammaticale, permettant la modélisation de la connaissance associée aux documents à reconnaître,
- un module de construction d'arbres d'analyse, qui permet l'exploration de plusieurs interprétations,
- un module de prise de décision, qui permet de valider la reconnaissance, de manière implicite ou explicite, avec sollicitation de l'utilisateur.

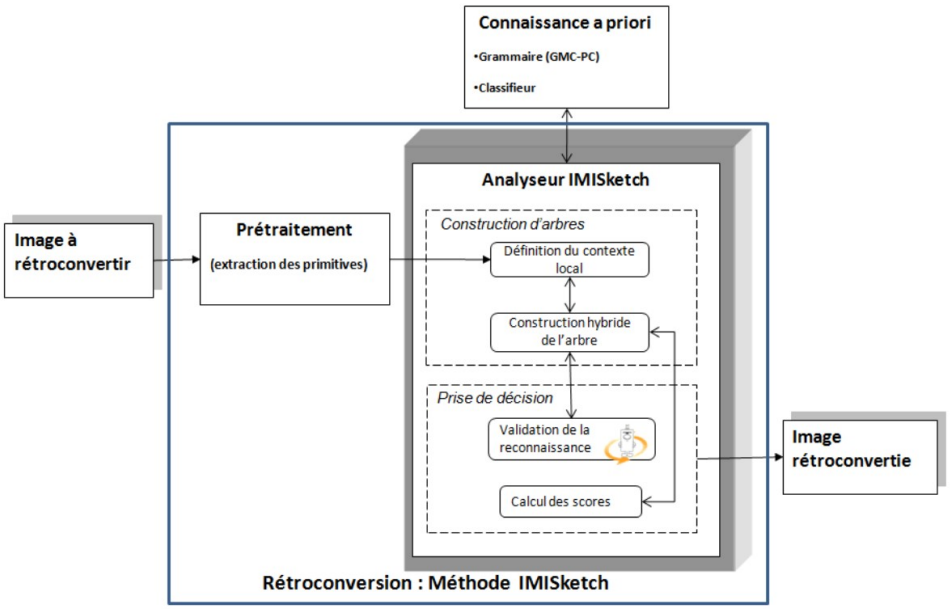


FIGURE 6.2 – Processus global d'analyse associé à IMISketch

Les connaissances structurelles a priori du document sont exprimées à travers un langage visuel grammatical, le formalisme DALI. Cette description grammaticale permet de piloter l'analyseur.

Contrairement à l'analyse avec DMOS, qui réussissait avec la première combinaison possible, ici l'analyseur attribue un score à toutes les branches de l'arbre de l'analyse. Afin de limiter la combinatoire, l'analyseur réduit la recherche à l'étude d'un contexte local. Nous avons également mis en place un processus d'exploration hybride original, guidé par la description grammaticale des schémas, qui permet d'accélérer localement l'analyse tout en limitant le risque de réaliser une fausse interprétation. Des détails sur la construction des arbres et l'analyse en largeur et en profondeur ont été publiés dans [GAL12 ; GLA12 ; Gho+11b].

Notre analyseur à base de règles est capable de mettre en concurrence des hypothèses d'interprétation, afin de solliciter l'utilisateur lorsque c'est nécessaire, durant la phase d'analyse. Deux cas d'ambiguïté peuvent se présenter : l'ambiguïté structurelle et l'ambiguïté de forme. L'ambiguïté structurelle est signalée par le système d'analyse quand il hésite entre deux segmentations différentes pour interpréter un symbole. Par exemple, dans un plan d'architecture, une ambiguïté structurelle peut être levée pour trouver la bonne segmentation des primitives entre un mur et un ouvrant (porte, fenêtre, etc.). L'ambiguïté de forme est levée s'il existe plusieurs hypothèses concurrentes pour étiqueter un symbole, par exemple une ambiguïté entre une porte et une fenêtre. L'intégration de l'utilisateur dans la boucle de reconnaissance évite une correction a posteriori fastidieuse des erreurs de reconnaissance tout en permettant d'avoir un système auto-évolutif au fur et à mesure de l'analyse [Gho+11a].

6.3 Résultats

Notre méthode interactive a été validée sur des plans d'architecture dessinés à main levée. Ces plans sont composés de murs, de trois types d'ouvrants et d'une dizaine de classes de mobilier. Des exemples de plans sont présentés sur la figure 6.3.

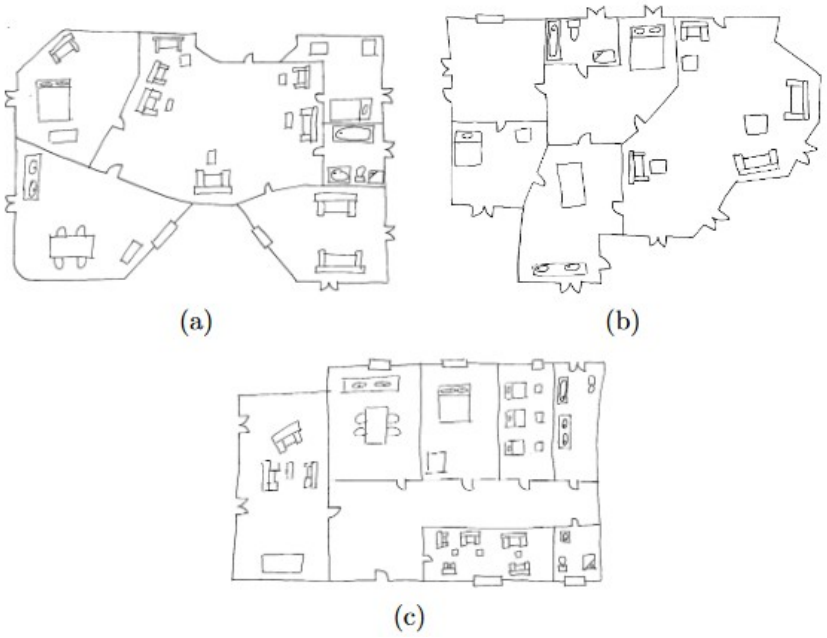


FIGURE 6.3 – Exemple de plans d'architectures analysés

Des expériences ont été menées pour évaluer la pertinence de la sollicitation utilisateur. Ces résultats sont détaillés dans [Gho+15 ; Gho+14].

Le tableau 6.1 présente quelques résultats. Sur une base composée de 15 plans d'architecture, nous avons mesuré un taux de reconnaissance des symboles de 93.4%. Un symbole est considéré comme bien reconnu lorsque sa boîte englobante est correctement construite, et que l'étiquette associée est correcte. Les erreurs restantes sont liées à des erreurs d'étiquettes générées par le

classifieur. L'utilisateur intervient en moyenne 5 fois par plan pour lever des ambiguïtés structurelles, et dans 25% des cas, l'utilisateur ne valide pas l'hypothèse proposée par le système. On peut donc considérer que son intervention était utile. Le système sollicite également l'utilisateur pour résoudre des ambiguïtés sur la classification des symboles. Sur la base de test, 49% des interactions sont utiles, c'est à dire que l'utilisateur ne sélectionne pas l'étiquette qui aurait été proposée à tort par le classifieur.

Nombre de plans d'architecture	15
Nombre de symboles	961
Taux de reconnaissance	93.4%
Nombre moyen de sollicitations structurelles par plan	5
Pourcentage de sollicitations structurelles utiles	25%
Nombre moyen de sollicitations pour classification par plan	5
Pourcentage de sollicitations pour classification utiles	49%

TABLE 6.1 – Evaluation de la reconnaissance des plans d'architecture

Ces travaux montrent que la sollicitation de l'utilisateur permet d'améliorer la qualité de reconnaissance des documents.

6.4 Expérimentations de psychologie cognitive

La mise en place de la méthode Imisketch a permis aux collègues du CRPCC de mener en parallèle l'encadrement de la thèse de Sylvain Fleury, sur le rôle de l'utilisateur dans les systèmes de traitements automatiques [Fle14]. L'objectif de cette thèse était d'améliorer la compréhension des interactions homme-machine afin de déterminer comment améliorer la performance du système. Plusieurs études ont ainsi été réalisées dans le cadre d'une démarche de conception centrée-utilisateur de notre système d'interprétation automatique de plans d'architecture.

Ces études se sont appuyées sur un cadre théorique varié, puisant à la fois dans la psychologie cognitive, les interactions homme-machine et l'ergonomie des documents multimédias. Elles ont mis en évidence des résultats débouchant sur des préconisations pour la conception de systèmes de reconnaissance. Ainsi, l'affichage de l'interprétation superposée au document interprété facilite la co-référenciation des informations à contrôler. De plus, permettre à l'utilisateur d'assister en temps réel à l'analyse, et d'intervenir au fur et à mesure améliore, les performances. Enfin, le signalement par la machine des risques d'erreurs est bénéfique, mais cela peut générer une réduction de la vigilance due à un phénomène de sur-confiance.

Ces expérimentations nous ont permis de publier plusieurs articles conjointement avec les collègues du CRPCC [Fle+15 ; Fle+13b ; Fle+13a].

Chapitre 7

Perspectives et projets de recherche

Dans le domaine de la reconnaissance d'images de documents, l'arrivée des systèmes à base de réseaux de neurones profonds a bouleversé les méthodes de reconnaissance, faisant dire à certains qu'il suffisait désormais d'un réseau bien entraîné pour pouvoir résoudre tous les problèmes.

Pourtant, je pense que l'utilisation de réseaux de neurones profonds ne résout pas tout, et fait apparaître de nouveaux problèmes. En effet, pour fonctionner, ces réseaux nécessitent de grandes bases de données annotées. Ces bases existent ou peuvent se générer facilement pour les documents nativement numériques (textes imprimés, pdf etc...). Mais il est beaucoup plus difficile d'obtenir de telles bases dans le cadre des documents anciens, qui n'ont jamais eu d'équivalent numérique, et pour lesquels il n'existe pas d'annotation.

Pour illustrer le propos, prenons l'exemple des registres paroissiaux du 17ème siècle : l'écriture parfois brouillonne du curé est souvent difficile à lire pour un humain. C'est typiquement le genre de document pour lequel des systèmes de reconnaissance automatique fourniraient une vraie plus-value, mais obtiennent actuellement des performances faibles, faute de données annotées. La réalisation de transcriptions intégrales est fastidieuse, coûteuse et l'annotation produite peut différer selon la manière dont l'humain interprétera un symbole.

Pour autant, même si l'utilisation des réseaux de neurones profonds trouve ses limites face à la difficulté d'annoter des images, il me semble évident qu'il faut poursuivre les recherches dans ce domaine.

Pour ce qui est de la reconnaissance de la structure de document, l'utilisation de réseaux profonds permet, par exemple, d'apprendre des séparations locales, en tenant compte du contexte de l'image. Mais l'utilisation de règles logiques pour réaliser un post-traitement des données permet de générer des résultats plus cohérents en fonction de ce qui est attendu dans un document.

De manière plus générale, je pense que pour améliorer la reconnaissance des documents, il faut que l'étape de reconnaissance de la structure puisse servir la reconnaissance de l'écriture, mais également en sens inverse, que le résultat de la reconnaissance d'écriture permette de remettre en cause la structure détectée pour en améliorer la détection. Et dans la même idée, les connaissances sémantiques qui sont extraites des documents par des méthodes de TAL (Traitement Automatique de la Langue) doivent pouvoir être utilisées pour valider la reconnaissance des caractères, mais également pour améliorer la détection de la structure.

Pour mes travaux à venir, je souhaite donc focaliser mes recherches autour des points suivants :

— la reconnaissance d'écriture manuscrite sur des images de documents anciens,
— la collaboration entre résultats de réseaux de neurones profonds et règles logiques,
— le développement de méthodes d'apprentissage avec peu de données,
— la mise en place de systèmes itératifs incluant des connaissances du TAL (Traitement Automatique de la Langue).

7.1 Reconnaissance d'écriture manuscrite dans les documents anciens

Jusqu'à présent, mes travaux de recherche se sont beaucoup focalisés sur la reconnaissance de structure de documents : construction de lignes de texte, de paragraphes, organisation logique des documents.

Pour ce qui est de la reconnaissance d'écriture, nous avons toujours préféré utiliser des systèmes du commerce (notamment Abbyy FineReader) ou collaborer avec des équipes de recherche plus avancées dans le domaine de la reconnaissance d'écriture (telles que le LITIS à Rouen). Mon expertise est donc moins grande en terme de système de reconnaissance d'écriture. A court terme, mes projets vont donc tourner autour de la reconnaissance d'écriture.

Si les systèmes de reconnaissance d'écriture actuels sont de plus en plus performants pour la reconnaissance d'écriture imprimée, et manuscrite récente, les systèmes restent améliorables sur la reconnaissance d'écriture dans les documents anciens. Ceci est dû à la difficulté d'accéder à des données étiquetées, mais également à la difficulté d'entraîner des architectures qui sont de plus en plus profondes et donc de plus en plus complexes.

Dans le cadre de la thèse de Solène Tarride, sur les registres paroissiaux anciens, un des prochains axes de recherche va être de mettre en place des techniques de word-spotting pour reconnaître certains mots clés. Nous souhaitons dans un premier temps pouvoir typer les actes (naissance, mariage, décès) avec la présence de mots clés tels récurrents, comme par exemple "ai baptisé" ou "les futurs époux". Puis, nous poursuivrons l'analyse de ces documents en cherchant à extraire les informations de dates, noms et prénoms des personnes concernées.

Depuis septembre 2019, l'arrivée dans l'équipe de mon collègue Yann Soullard, travaillant spécifiquement sur l'utilisation des réseaux de neurones profonds pour la reconnaissance d'écriture va me permettre de me focaliser davantage sur ce point. Nous avons co-encadré un stagiaire de master recherche, Killian Barrère, qui a réalisé une étude sur l'impact de méta-paramètres dans l'entraînement de réseaux de neurones pour la reconnaissance d'écriture. Killian va démarrer une thèse, encadrée par Bertrand Coüasnon, Yann Soullard, et moi même, sur l'intérêt des mécanismes d'attention dans les réseaux de neurones.

En effet, parmi les innovations en apprentissage profond apparues ces dernières années, les mécanismes d'attention ont obtenu beaucoup de succès dans le domaine de la vision par ordinateur. Les mécanismes d'attention permettent à un réseau de neurones d'apprendre à focaliser son attention sur des caractéristiques d'intérêt pour la décision. Des premières architectures de réseaux de neurones s'appuyant sur des mécanismes d'attention ont récemment émergé pour des applications en reconnaissance d'écriture [BLM17 ; YHM18] avec succès. L'utilisation de l'attention permet de tendre vers la lecture humaine, en focalisant l'attention du reconnaisseur à des emplacements appropriés pour la décision tout en atténuant la présence de bruit dans les images qui pourraient perturber la reconnaissance (rature, souligné, ...). Nous souhaitons continuer à explorer cette piste, pour analyser notamment des pages de textes manuscrits, dans le cadre de la thèse de Killian Barrère.

Dans le domaine de la reconnaissance d'écriture, nous avons également été contactés par la société AriadNext. AriadNext s'intéresse, entre autres, à l'analyse de documents d'identités. Nous prévoyons de monter une thèse CIFRE, autour de la reconnaissance conjointe de la structure de documents et de la reconnaissance d'écriture. L'objectif est ici de voir s'il est possible d'utiliser les mécanismes d'attention, et la connaissance à priori sur des modèles de documents, pour analyser ces documents en une seule passe.

7.2 Collaboration entre réseaux de neurones et règles logiques

Pour ce qui est de la reconnaissance de la structure d'images de documents, les systèmes à base de réseaux de neurones sont maintenant capables de produire des cartes de probabilités per-

mettant la détection et la segmentation d'objets. Des réseaux de type U-net [Grü+18] [ASK18] sont appliqués avec succès pour la détection de lignes de texte, des bords de page... Des réseaux prévus pour la détection d'objets tels que Mask R-CNN [He+17], RetinaNet [Lin+17] ou YOLO [RF18] obtiennent également des résultats intéressants sur certaines tâches d'analyse de documents.

Ces systèmes se montrent très performants lorsqu'il est possible d'apprendre localement une configuration de la page, qui permet de prendre une décision en fonction d'un contexte. Cependant, les cartes de probabilités fournies par ces systèmes nécessitent souvent un post-traitement : il faut en effet convertir des sorties avec des zones de probabilités, en objets représentés en général par leurs boites englobantes. La plupart des méthodes utilisent des traitements à base de seuils. C'est ici qu'il me semble pertinent d'intégrer des règles sur le contenu logique du document pour valider la présence et les frontières des objets recherchés.

De plus, les systèmes de reconnaissance à base de réseaux de neurones profonds génèrent souvent des résultats imparfaits, a fortiori lorsqu'ils sont entraînés avec peu de données. Ainsi, tel système générera des faux-positifs tandis que tel autre oubliera des données. Je pense qu'un système à base de règles doit pouvoir permettre de combiner les résultats de plusieurs systèmes de reconnaissance à base de réseaux de neurones, pour pallier les limites de chacun. Nous avons commencé à amorcer cette piste dans les travaux de Solène Tarride. Je souhaite continuer à creuser plus largement cette piste ; c'est la suite logique de mes travaux sur la combinaison de données.

Enfin, les réseaux de neurones nécessitent des données étiquetées en grande quantité, et ont un pouvoir de généralisation parfois limité. Lorsque les règles logiques sont faciles à écrire, un système à base de description grammaticale peut être plus facile à adapter à une nouvelle problématique, en particulier s'il n'existe pas de données annotées.

Ces éléments sont en faveur de la poursuite des travaux avec des systèmes à base de règles, s'ils sont utilisés dans les cas énoncés ci-dessus, pour compléter et combiner de manière intelligente des résultats de réseaux de neurones.

7.3 Apprentissage avec peu de données

Les réseaux de neurones profonds ont récemment prouvé leur efficacité, pour peu que des données étiquetées soient disponibles. Dans le contexte des documents anciens, cet étiquetage de données est souvent fastidieux : à la main, de longues heures d'annotations sont nécessaires. Le temps d'annotation est du même ordre que celui d'une traduction humaine directe.

Une des méthodes les plus utilisées pour étiqueter davantage de données est de réaliser une augmentation de données en modifiant légèrement les données proposées en apprentissage pour en créer des variantes. De nombreuses méthodes sont disponibles dans la littérature, qui proposent diverses déformations afin d'enrichir les corpus d'apprentissage.

Je souhaiterais creuser deux autres pistes.

La première est l'utilisation d'une description grammaticale rudimentaire pour générer des données approximatives, qu'il sera ensuite plus facile de corriger manuellement pour obtenir une annotation définitive.

Par exemple, nous avons récemment expérimenté un cas dans l'analyse de l'annuaire allemand Handbuch. La grammaire était souvent capable de localiser des titres de rubriques grâce à la présence de " :", mais oubliait quelques éléments. Nous avons utilisé la grammaire pour générer massivement des données sur un grand nombre de pages. Une passe manuelle simplifiée a ensuite permis de corriger le corpus en étiquetant uniquement les titres de rubriques manquants. Ces titres étiquetés ont été utilisés pour entraîner un réseau de neurones, et les prédictions du réseau de neurone sont maintenant prise en entrée, comme un nouveau calque.

Je souhaiterais travailler à formaliser davantage un tel comportement, pour exploiter les résultats de l'analyse grammaticale comme base d'une annotation. En effet, annoter un corpus de

documents entièrement à la main est coûteux, surtout pour une annotation de la structure de documents, qui requiert de dessiner des boites englobantes de contenus à reconnaître. Il est plus facile, pour l'annotation de la structure de documents, de corriger des boites pré-existantes : déplacer un bord, supprimer une boite ou en ajouter une. Il serait possible d'imaginer que la description grammaticale de documents permette de générer une vérité terrain pour les 80% de cas "faciles", et que l'annotateur n'ait qu'à rajouter l'annotation des 20% de cas particuliers restants.

La seconde piste est la génération semi-automatique d'une pseudo-vérité terrain. Nous avons exploré cette piste dans le cadre de la thèse de Cérès Carton. Il s'agit d'utiliser une méthode de clustering automatique pour regrouper les images ayant des critères similaires, et présenter à l'utilisateur seulement quelques exemples d'image de chaque cluster. L'utilisateur, en annotant un exemple issu du cluster génère alors une annotation pour toutes les images appartenant à ce cluster. Il s'agit donc d'une pseudo-vérité terrain, mais qui peut être suffisante pour entraîner des mécanismes d'apprentissage.

Depuis la thèse de Cérès Carton, nous n'avons pas eu l'occasion d'expérimenter ces travaux sur un autre corpus : je souhaiterais poursuivre dans cette direction.

7.4 Mise en place d'une boucle de reconnaissance

A long terme, j'aimerais collaborer avec des chercheurs d'autres disciplines pour mettre en place un système complet de reconnaissance, jusqu'à l'extraction d'informations et leur exploitation dans le domaine des humanités numériques.

On peut considérer qu'un processus classique de reconnaissance d'images de documents met en jeu trois briques (figure 7.1) : une brique de reconnaissance de la structure des documents, puis une brique de la reconnaissance de l'écriture, et une brique d'extraction d'information.

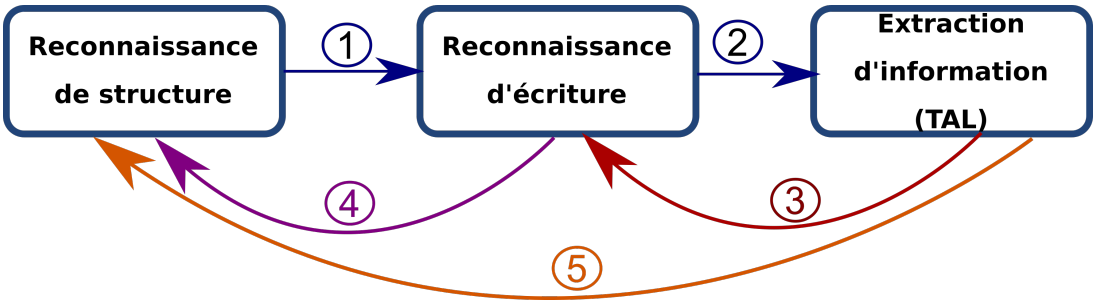


FIGURE 7.1 – Création d'une boucle dans le processus de reconnaissance : rajouter les connexions 4 et 5 pour permettre au système de reconnaissance de structure de documents de prendre en compte des informations provenant des résultats de la reconnaissance de caractères, mais aussi de l'analyse des contenus textuels.

Classiquement, les appels aux trois briques sont réalisés de manière successive, en suivant les flèches notées (1) et (2) sur le schéma, sans croisement possible de l'information. Or, les mécanismes d'interprétation humaine sont capables de fusionner les trois sources d'information pour fiabiliser une reconnaissance de structure. Il me semble donc important de mettre en place les retours d'information matérialisés par les flèches (3), (4) et (5).

La flèche (3) représente le fait d'utiliser l'interprétation des informations extraites pour fiabiliser ou remettre en cause la phase de reconnaissance d'écriture. Ceci existe déjà avec l'utilisation de dictionnaires et de modèles de langages pour la reconnaissance d'écriture. En revanche, dans les systèmes actuels, le retour d'information se fait assez peu jusqu'à la brique de reconnaissance de la structure.

La flèche (4) matérialise la possibilité d'intégrer les retours de la reconnaissance d'écriture pour fiabiliser ou remettre en cause la reconnaissance de la structure. En effet, la reconnaissance

d'écriture doit pouvoir nourrir l'interprétation de la structure. C'est d'ailleurs la conclusion du projet Maurdor, qui n'a pas eu de suite à l'heure actuelle, mais pour lequel une refonte du système se ferait non pas de manière séquentielle mais itérative.

Nous avons amorcé des travaux dans ce sens dans le cadre de la thèse de Camille Guerry (section 4.4.2), dans laquelle nous mettons en place une stratégie d'analyse qui valide la reconnaissance de la structure en fonction de la reconnaissance d'écriture et du contenu des autres pages de la collection. Je souhaite continuer à explorer à court terme ces interactions.

Dans mes projets de recherche à moyen terme, je souhaiterais mettre en place un lien correspondant à la flèche (5), qui consiste à exploiter les résultats de système d'information pour enrichir les résultats de la reconnaissance de la structure. Nous pourrions pour cela continuer d'utiliser le principe de la combinaison de données hétérogènes, en prenant en compte des données de plus haut niveau, issues de l'interprétation sémantique des documents.

Il existe une frontière parfois trop opaque entre certaines équipes focalisées sur les images de documents, et les équipes qui se concentrent sur l'analyse textuelle (TAL, traitement automatique de la langue). A long terme, je souhaiterais créer davantage de collaboration autour de ces deux thématiques, en travaillant davantage avec des collègues issus du traitement de la langue naturelle. Je pense notamment à ma collègue Fabienne Moreau, du département MIASHS, qui effectue sa recherche dans le laboratoire LIDILE de Rennes 2.

Les domaines d'application ne manquent pas pour l'ensemble de ces travaux. Dans le contexte de mon affectation à l'Université Rennes 2, je pourrais notamment collaborer avec des enseignants-chercheurs dans d'autres disciplines : histoire, géographie, sciences sociales, humanités numériques. La difficulté d'un tel travail de collaboration est de réussir à distinguer le besoin d'ingénierie des collègues, et nos intérêts de recherche : malgré l'attrait d'un travail sur des documents avec un réel besoin, il ne faut pas être simplement « l'ingénieur informaticien » des collègues en humanités numériques. C'est cette difficulté qui a tendance à freiner les collaborations, mais l'essor des techniques de reconnaissance d'écriture devrait permettre de faire émerger des nouvelles problématiques sur des documents plus complexes.

Chapitre 8

Conclusion

Ce manuscrit présente une synthèse de mes travaux de recherche depuis 2009, au sein de l'équipe Intuidoc. J'ai focalisé mes travaux sur la reconnaissance d'images de documents, et surtout l'analyse de la structure de pages. Mes travaux se basent sur l'idée de combiner des données venant de sources hétérogènes, en étant guidé par une connaissance a priori sur les documents. Ainsi, nous avons pu combiner des indices visuels issus de plusieurs niveaux de l'image, des résultats de classifieurs, des résultats d'OCR ou provenant d'une analyse de la collection. Plus récemment, nous avons intégré des résultats d'analyse avec des réseaux de neurones profonds.

La méthode grammaticale DMOS a été la base de ces travaux, notamment grâce à la possibilité d'utiliser le formalisme de calque perceptif. Il est ainsi possible avec la méthode DMOS de décrire de manière tri-dimensionnelle le contenu des documents à reconnaître : deux dimensions permettent de localiser l'analyse dans l'image de documents, la troisième dimension vient de la possibilité de choisir dans quel calque piocher les indices visuel à analyser.

Mes travaux se sont toujours appuyés sur des besoins réels : soit des compétitions internationales, soit des projets. J'ai ainsi pris part à plusieurs projets : le projet ANR Mobisketch, le projet Maurdor, le projet de maturation sur la presse ancienne, le projet IAT, le projet ANR HBDEX, le projet européen EurHisFirm, le partenariat avec la startup Doptim. Je n'ai pas encore eu l'occasion de porter de projets en mon nom propre, principalement à cause de contraintes administratives (absence de laboratoire d'informatique à Rennes 2). J'ai toutefois participé activement à l'ensemble de ces projets, et il a été envisagé que je porte le partenariat avec Doptim.

Pour tous ces travaux, j'ai eu l'opportunité de réaliser des co-encadrements : 4 doctorants (dont 2 ont soutenu), 4 ingénieurs, 1 post-doctorant ainsi que des stagiaires. Prochainement, 2 thèses devraient démarrer également.

Mes perspectives sont à court terme de me pencher davantage sur le domaine de la reconnaissance d'écriture manuscrite, pour étudier comment il est possible de relier la reconnaissance de la structure et du contenu textuel. Je souhaite me pencher également sur l'analyse des documents à base de réseaux de neurones profonds, et surtout la manière de réaliser rapidement une annotation de corpus. Enfin, je souhaite à plus long terme développer des interactions avec les collègues du traitement automatique de la langue, et des humanités numériques.

Références

[Alb+19] Michele Alberti, Lars Vögtlin, Vinaychandran Pondenkandath, Mathias Seuret, Rolf Ingold et Marcus Liwicki. « Labeling, Cutting, Grouping : an Efficient Text Line Segmentation Method for Medieval Manuscripts ». In : *CoRR* abs/1906.11894 (2019). arXiv : 1906.11894.

[ASK18] S. Ares Oliveira, B. Seguin et F. Kaplan. « dhSegment : A Generic Deep-Learning Approach for Document Segmentation ». In : *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. 2018, p. 7-12.

[BLM17] T. Bluche, J. Louradour et R. Messina. « Scan, Attend and Read : End-to-End Handwritten Paragraph Recognition with MDLSTM Attention ». In : *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. T. 01. 2017, p. 1050-1055.

[Bru+14] S. Brunessaux, P. Giroux, B. Grilhères, M. Manta, M. Bodin, K. Choukri, O. Galibert et J. Kahn. « The Maurdor Project : Improving Automatic Processing of Digital Documents ». In : *2014 11th IAPR International Workshop on Document Analysis Systems*. 2014, p. 349-354. doi : 10.1109/DAS.2014.58.

[Car16] Cérès Carton. « Semi-automatic and interactive rule inference with or without ground-truth for document structure recognition ». Theses. INSA de Rennes, mars 2016. url : https://hal.inria.fr/tel-01492966.

[CLC13] Cérès Carton, Aurélie Lemaitre et Bertrand Couasnon. « Fusion of statistical and structural information for flowchart recognition ». In : *ICDAR - International Conference on Document Analysis and Recognition*. Washington, United States, 2013, p. 1242-1246. url : https://hal.inria.fr/hal-00921640.

[CLC14a] Cérès Carton, Aurélie Lemaitre et Bertrand Coüasnon. « LearnPos : a new tool for interactive learning positioning ». In : *DRR - Document Recognition and Retrieval XXI*. San Francisco, United States, 2014. url : https://hal.inria.fr/hal-00921642.

[CLC15] Cérès Carton, Aurélie Lemaitre et Bertrand Coüasnon. « Automatic and interactive rule inference without ground truth ». In : *International Conference on Document Analysis and Recognition (ICDAR)*. Nancy, France, août 2015. url : https://hal.inria.fr/hal-01197470.

[CLC14b] Cérès Carton, Aurélie Lemaitre et Bertrand B. Coüasnon. « LearnPos : un nouvel outil pour l'apprentissage interactif de positionnement ». In : *Conférence Internationale Francophone sur l'Ecrit et le Document*. Nancy, France, 2014, p. 325-340. url : https://hal.inria.fr/hal-01088810.

[CLC16] Cérès Carton, Aurélie Lemaitre et Bertrand B. Coüasnon. « Inférence semi-automatique et interactive de règles sans vérité terrain ». In : *Conférence Internationale Francophone sur l'Ecrit et le Document (CIFED'2016)*. Toulouse, France, mars 2016. url : https://hal.inria.fr/hal-01492921.

[CLC17] Cérès Carton, Aurélie Lemaitre et Bertrand B. Coüasnon. « Eyes Wide Open : an interactive learning method for the design of rule-based systems ». In : *International Journal on Document Analysis and Recognition* 20.2 (2017), p. 91-103. doi : 10.1007/s10032-017-0282-x. url : https://hal.inria.fr/hal-01493442.

[Cha13] Joseph Chazalon. « Contextual and assisted interpretation of digitized fonds : application to sales registers from the 18th century ». Theses. INSA de Rennes, jan. 2013. url : https://tel.archives-ouvertes.fr/tel-00903372.

[CC12] Joseph Chazalon et Bertrand B. Couasnon. « Iterative analysis of document collections enables efficient human-initiated interaction ». In : *DRR - Document Recognition and Retrieval XIX, Part of the IS&T/SPIE 24th Annual Symposium on Electronic Imaging*. T. 8297. San Francisco, United States, jan. 2012, p. 82970L. doi : 10.1117/12.911995. url : https://hal.inria.fr/hal-00686858.

[CCL12a] Joseph Chazalon, Bertrand B. Couasnon et Aurélie Lemaitre. « Comment introduire simplement et uniformément deux modes d'interaction asynchrones complémentaires dans un système d'analyse de documents existant ». In : *CIFED - Conférence Internationale sur l'Ecrit et le Document*. Bordeaux, France, mars 2012, p. 285-299. url : https://hal.inria.fr/hal-00686847.

[CCL10] Joseph Chazalon, Bertrand Coüasnon et Aurélie Lemaitre. « Mémoire visuelle pour l'analyse de documents structurés ». In : *Colloque International Francophone sur l'Ecrit et le Document (CIFED2010)*. Actes du 11ème Colloque International Francophone sur l'Ecrit et le Document. Sousse, Tunisia, mars 2010. url : https://hal.inria.fr/inria-00545886.

[CCL11] Joseph Chazalon, Bertrand Coüasnon et Aurélie Lemaitre. « Iterative Analysis of Pages in Document Collections for Efficient User Interaction ». In : *Document Analysis and Recognition (ICDAR), 2011 International Conference on*. China, sept. 2011, p. 503-507. doi : 10.1109/ICDAR.2011.107. url : https://hal.archives-ouvertes.fr/hal-00644927.

[CCL12b] Joseph Chazalon, Bertrand Coüasnon et Aurélie Lemaitre. « A Simple and Uniform Way to Introduce Complimentary Asynchronous Interaction Models in an Existing Document Analysis System ». In : *DAS - 10th IAPR International Workshop on Document Analysis Systems*. Gold Coast, Australia, mars 2012, p. 399-403. doi : 10.1109/DAS.2012.11. url : https://hal.inria.fr/hal-00686853.

[Coü06] B. Coüasnon. « DMOS, a Generic Document Recognition Method : Application to Table Structure Analysis in a General and in a Specific Way ». In : *International Journal on Document Analysis and Recognition (IJDAR)* 8(2) (2006), p. 111-122.

[CL17] Bertrand Coüasnon et Aurélie Lemaitre. « DMOS, It's your turn ! » In : *1st International Workshop on Open Services and Tools for Document Analysis (ICDAR-OST)*. Kyoto, Japan, nov. 2017. url : https://hal.inria.fr/hal-01659131.

[Die+17] Markus Diem, Florian Kleber, Stefan Fiel, Tobias Gruning et Basilis Gatos. « cBAD : ICDAR2017 Competition on Baseline Detection ». In : nov. 2017, p. 1355-1360. doi : 10.1109/ICDAR.2017.222.

[DT14] David Doermann et Karl Tombre. *Handbook of Document Image Processing and Recognition*. Springer Publishing Company, Incorporated, 2014. isbn : 0857298585.

[Fle14] Sylvain Fleury. « Le rôle de l'utilisateur dans les systèmes de traitements automatiques ». 2014REN20001. Thèse de doct. 2014. url : http://www.theses.fr/2014REN20001/document.

[Fle+13a] Sylvain Fleury, Achraf Ghorbel, Aurélie Lemaitre, Eric Anquetil et Eric Jamet. « User-centred design of an interactive off-line handwritten architectural floor plan recognition ». In : *12th International Conference on Document Analysis and Recognition*. United States, août 2013, p. 1073-1077. URL : https://hal.archives-ouvertes.fr/hal-00959722.

[Fle+13b] Sylvain Fleury, Eric Jamet, Emilie Loup-Escande, Achraf Ghorbel, Aurélie Lemaitre et Eric Anquetil. « Towards Specifications for Automatic Recognition Software : An Example of a User-Centred Design ». In : *Journal of Software Engineering and Applications* 6 (2013), p. 1-4. URL : https://hal.archives-ouvertes.fr/hal-00959721.

[Fle+15] Sylvain Fleury, Éric Jamet, Achraf Ghorbel, Aurélie Lemaitre et Eric Anquetil. « Application of the Resources Model to the Supervision of an Automated Process ». In : *ACM Transactions on Computer-Human Interaction* 30 (2015), p. 103-121. DOI : 10.1080/07370024.2014.893831. URL : https://hal.inria.fr/hal-01089413.

[FJ02] Ana LN Fred et Anil K Jain. « Data clustering using evidence accumulation ». In : *Int. Conf. on Pattern recognition (ICPR)*. T. 4. 2002, p. 276-280.

[GKO14] Olivier Galibert, Juliette Kahn et Ilya Oparin. « The zonemap metric for page segmentation and area classification in scanned documents ». In : *ICIP*. IEEE, 2014, p. 2594-2598. ISBN : 978-1-4799-5751-4.

[GSL10] Basilios Gatos, Nikolaos Stamatopoulos et Georgios Louloudis. « ICFHR 2010 Handwriting Segmentation Contest ». In : *International Conference on Frontiers in Handwriting Recognition, ICFHR 2010, Kolkata, India, 16-18 November 2010*. 2010, p. 737-742. DOI : 10.1109/ICFHR.2010.120. URL : https://doi.org/10.1109/ICFHR.2010.120.

[Gho12] Achraf Ghorbel. « Interactive interpretation of structured documents : application to the retro-conversion of handwritten architectural plans ». Theses. INSA de Rennes, déc. 2012. URL : https://tel.archives-ouvertes.fr/tel-00788832.

[Gho+11a] Achraf Ghorbel, Abdullah Almaksour, Aurélie Lemaitre et Eric Anquetil. « Incremental learning for interactive sketch recognition ». In : *Ninth IAPR International Workshop on Graphics RECognition*. Séoul, South Korea, sept. 2011. URL : https://hal.inria.fr/hal-00646137.

[Gho+14] Achraf Ghorbel, Eric Anquetil, Jean Camillerapp et Aurélie Lemaitre. « IMISketch : an interactive method for sketch recognition ». In : *Pattern Recognition Letters* 35.78-90 (jan. 2014), p. 13. URL : https://hal.archives-ouvertes.fr/hal-00933697.

[GAL12] Achraf Ghorbel, Eric Anquetil et Aurélie Lemaitre. « Optimization Analysis Based On A Breadth-First Exploration For A Structural Approach Of Sketches Interpretation ». In : *10th IAPR International Workshop on Document Analysis Systems*. Gold Coast, Queensland, Australia, mars 2012. URL : https://hal.archives-ouvertes.fr/hal-00687680.

[GLA12] Achraf Ghorbel, Aurélie Lemaitre et Eric Anquetil. « Competitive hybrid exploration for off-line sketches structure recognition ». In : *International Conference on Frontiers in Handwriting Recognition (ICFHR-2012)*. Bari, Italy, sept. 2012, p. 569-574. URL : https://hal.archives-ouvertes.fr/hal-00738470.

[Gho+15] Achraf Ghorbel, Aurélie Lemaitre, Eric Anquetil, Sylvain Fleury et Eric Jamet. « Interactive interpretation of structured documents : Application to the recognition of handwritten architectural plans ». In : *Pattern Recognition* 48.8 (août 2015). DOI : 10.1016/j.patcog.2015.01.028. URL : https://hal.inria.fr/hal-01238056.

[Gho+11b] Achraf Ghorbel, Sébastien Macé, Aurélie Lemaitre et Eric Anquetil. « Interactive competitive breadth-first exploration for sketch interpretation ». In : *International Conference on Document Analysis and Recognition*. Beijing, China, sept. 2011. doi : 10.1109/ICDAR.2011.241. url : https://hal.inria.fr/hal-00646120.

[Gro+09] E. Grosicki, M. Carré, J. Brodin et E. Geoffrois. « Results of the RIMES Evaluation Campaign for Handwritten Mail Processing ». In : *2009 10th International Conference on Document Analysis and Recognition*. 2009, p. 941-945.

[Grü+18] Tobias Grüning, Gundram Leifert, Tobias Strauss et Roger Labahn. « A Two-Stage Method for Text Line Detection in Historical Documents ». In : *CoRR* abs/1802.03345 (2018). arXiv : 1802.03345. url : http://arxiv.org/abs/1802.03345.

[GCL19] Camille Guerry, Bertrand Coüasnon et Aurélie Lemaitre. « Combination of deep-learning and syntactical approaches for the interpretation of interactions between text-lines and tabular structures in handwritten documents ». In : *International Conference on Document Analysis and Recognition*. 2019.

[He+17] Kaiming He, Georgia Gkioxari, Piotr Dollár et Ross B. Girshick. « Mask R-CNN ». In : *CoRR* abs/1703.06870 (2017). arXiv : 1703.06870.

[LC13] Aurélie Lemaitre et Jean Camillerapp. « HIP 2013 FamilySearch Competition - Contribution of IRISA ». In : *HIP - ICDAR Historical Image Processing Workshop*. Washington, United States, août 2013. url : https://hal.inria.fr/hal-00854463.

[Lem+18] Aurélie Lemaitre, Jean Camillerapp, Cérès Carton et Bertrand B. Coüasnon. « A combined strategy of analysis for the localization of heterogeneous form fields in ancient pre-printed records ». In : *International Journal on Document Analysis and Recognition* 21(4).269-282 (juill. 2018). doi : 10.1007/s10032-018-0309-y. url : https://hal.inria.fr/hal-01858192.

[LCC08] Aurélie Lemaitre, Jean Camillerapp et Bertrand Coüasnon. « A generic method for structure recognition of handwritten mail documents ». In : *Document Recognition and Retrieval DRR XV*. San Jose, United States, jan. 2008. url : https://hal.inria.fr/inria-00308565.

[LCC11] Aurélie Lemaitre, Jean Camillerapp et Bertrand Coüasnon. « A perceptive method for handwritten text segmentation ». In : *Document recognition and retrieval XVIII - Electronic Imaging*. San Francisco, United States, jan. 2011, p. 7874 0C. url : https://hal.archives-ouvertes.fr/hal-00567074.

[LCC14] Aurélie Lemaitre, Jean Camillerapp et Bertrand Coüasnon. « Handwritten text segmentation using blurred image ». In : *DRR - Document Recognition and Retrieval XXI*. DRR - Document Recognition and Retrieval XXI. San Francisco, United States, jan. 2014. url : https://hal.inria.fr/hal-01087210.

[LC17] Aurélie Lemaitre et Bertrand Coüasnon. *DMOS : A vous de jouer !* Symposium International Francophone sur l'Ecrit et le Document (SIFED). Juin 2017. url : https://hal.inria.fr/hal-01534885.

[Lem+13] Aurélie Lemaitre, Harold Mouchère, Jean Camillerapp et Bertrand B. Couasnon. « Interest of syntactic knowledge for on-line flowchart recognition ». In : *Graphics Recognition New Trends and Challenges*. Sous la dir. d'Young-Bin Kwon et Jean-Marc Ogier. T. 7423. LNCS. Springer, déc. 2013, p. 89-98. doi : 10.1007/978-3-642-36824-0_9. url : https://hal.inria.fr/hal-00854450.

[Lem+11] Aurélie Lemaitre, Harold Mouchère, Jean Camillerapp et Bertrand B. Coüasnon. « Interest of Syntactic Knowledge for On-line Flowchart Recognition ». In : *International Workshop on Graphics RECognition, GREC 2011*. North Korea, sept. 2011. URL : https://hal.archives-ouvertes.fr/hal-00635457.

[Lin+17] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He et Piotr Dollár. « Focal Loss for Dense Object Detection ». In : *CoRR* abs/1708.02002 (2017). arXiv : 1708.02002.

[MA09] Sébastien Macé et Eric Anquetil. « Eager interpretation of on-line hand-drawn structured documents : The DALI methodology ». In : *Pattern Recognition*. New Frontiers in Handwriting Recognition 42.12 (déc. 2009), p. 13. DOI : 10.1016/j.patcog.2008.10.018. URL : https://hal.inria.fr/hal-01347173.

[Moy+15] B. Moysset, C. Kermorvant, C. Wolf et J. Louradour. « Paragraph text segmentation into lines with Recurrent Neural Networks ». In : *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. 2015, p. 456-460.

[OCL11] André O. Maroneze, Bertrand B. Coüasnon et Aurélie Lemaitre. « Introduction of statistical information in a syntactic analyser for document image recognition ». In : *Document recognition and Retrieval XVIII - Electronic Imaging*. San Francisco, United States, jan. 2011, p. 7874 04. URL : https://hal.archives-ouvertes.fr/hal-00567077.

[PLC14] Baptiste Poirriez, Aurélie Lemaitre et Bertrand Coüasnon. « Visual perception of unitary elements for layout analysis of unconstrained documents in heterogeneous databases ». In : *14th International Conference on Frontiers in Handwriting Recognition (ICFHR-2014)*. Crete island, Greece, sept. 2014. URL : https://hal.inria.fr/hal-01088807.

[RF18] Joseph Redmon et Ali Farhadi. « YOLOv3 : An Incremental Improvement ». In : *CoRR* abs/1804.02767 (2018). arXiv : 1804.02767.

[Ren+18] Guillaume Renton, Yann Soullard, Clement Chatelain, Sébastien Adam, Christopher Kermorvant et Thierry Paquet. « Fully convolutional network with dilated convolutions for handwritten text line segmentation ». In : *International Journal on Document Analysis and Recognition (IJDAR)* (mai 2018). DOI : 10.1007/s10032-018-0304-3.

[Ric+14] Yann Ricquebourg, Christian Raymond, Baptiste Poirriez, Aurélie Lemaitre et Bertrand Coüasnon. « Boosting bonsai trees for handwritten/printed text discrimination ». In : *Document Recognition and Retrieval (DRR)*. San Francisco, United States, fév. 2014. URL : https://hal.archives-ouvertes.fr/hal-00910718.

[Rom+18] Geoffrey Roman-Jimenez, Christian Viard-Gaudin, Adeline Granet et Harold Mouchère. « Transfer Learning for Structures Spotting in Unlabeled Handwritten Documents using Randomly Generated Documents ». In : *International Conference on Pattern Recognition Applications and Methods*. Madeira, Portugal, jan. 2018. URL : https://hal.archives-ouvertes.fr/hal-01681114.

[Sta+13] Nikolaos Stamatopoulos, Basilis Gatos, Georgios Louloudis, Umapada Pal et Alireza Alaei. « ICDAR 2013 Handwriting Segmentation Contest ». In : *Proceedings of the 2013 12th International Conference on Document Analysis and Recognition*. ICDAR '13. Washington, DC, USA : IEEE Computer Society, 2013, p. 1402-1406. ISBN : 978-0-7695-4999-6. DOI : 10.1109/ICDAR.2013.283. URL : https://doi.org/10.1109/ICDAR.2013.283.

[Tar+19] Solène Tarride, Aurélie Lemaitre, Bertrand Couasnon et Sophie Tardivel. « Signature detection as a way to recognise historical parish register structure ». In : *HIP 2019*. Sydney, Australia : ACM Press, sept. 2019, p. 54-59. doi : 10.1145/3352631.3352636. url : https://hal.archives-ouvertes.fr/hal-02433964.

[TOV20] C. Truong, L. Oudre et N. Vayatis. « Selective review of offline change point detection methods. » In : *Signal Processing, 167, 107299*. IEEE. 2020, p. 770-778.

[Wan+17] Chengcheng Wang, Harold Mouchère, Aurélie Lemaitre et Christian Viard-Gaudin. « Online flowchart understanding by combining max-margin Markov random field with grammatical analysis ». In : *International Journal on Document Analysis and Recognition* 20.2 (juin 2017), p. 123-136. doi : 10.1007/s10032-017-0284-8. url : https://hal.inria.fr/hal-01503366.

[YHM18] Mohamed Yousef, Khaled F. Hussain et Usama S. Mohammed. *Accurate, Data-Efficient, Unconstrained Text Recognition with Convolutional Neural Networks*. 2018. arXiv : 1812.11894 [cs.CV].