Maîtrise de la qualité des transcriptions numériques dans les projets de numérisation de masse

Ahmed Ben Salah

Université de Rouen

N° attribué par la bibliothèque
␣␣␣␣␣␣␣␣␣␣␣␣␣␣

THÈSE

pour obtenir le grade de
Docteur de Université de Rouen
Spécialité : **Science Informatique**
préparée au laboratoire **LITIS**
dans le cadre de l'École Doctorale **SPMII**

présentée et soutenue publiquement
par

Ahmed BEN SALAH

le 11 juillet 2014

Titre:

Maîtrise de la qualité des transcriptions numériques dans les projets de numérisation de masse

Directeur de thèse: **Thierry Paquet**
Encadrant de thèse: **Nicolas Ragot**

Jury

Pr. Jean-Philippe Domenger,	Rapporteur
Pr. Rémy Mullot,	Rapporteur
Pr. Jean-Marc Ogier,	Examinateur
M. Laurent Duplouy,	Invité
M. Thierry Pardé,	Invité
Pr. Thierry Paquet,	Directeur de thèse
Dr. Nicolas Ragot,	Encadrant de thèse

Laboratoire d'Informatique, du Traitement de l'Information et des Systèmes

Remerciements

Ce travail a été réalisé au sein du service numérisation de la Bibliothèque nationale de France (BnF) et du laboratoire d'Informatique, du Traitement de l'Information et des Systèmes (LITIS) de l'université de Rouen dans le cadre du Plan Triennal de recherche de la BnF.

Je remercie Monsieur Laurent Duplouy, chef du service de numérisation de la BnF de m'avoir accueilli dans son service et de son soutien durant la réalisation de ce travail de recherche.

J'exprime toute ma gratitude à Monsieur Thierry Paquet, Professeur à l'Université de Rouen, et à Monsieur Nicolas Ragot, Maître de Conférences à l'Université François Rabelais, pour avoir proposé et encadré ce sujet. Je les remercie de m'avoir encouragé et soutenu pendant cette période ; je leur suis très reconnaissant de l'intérêt qu'ils ont portés à mon travail, de même que pour leur aide et leur disponibilité. Grâce à eux, j'ai découvert un domaine de recherche qui aujourd'hui me passionne. Je souhaite exprimer mes plus vifs remercîments à mes tuteurs dans le service de numérisation de la BnF, Yohann Le Tallec, Jean-Philippe Moreux et Geneviève Cron qui m'ont suivi durant la réalisation de ce travail et qui m'ont aidé énormément dans la conception de mes travaux.

J'adresse toute ma reconnaissance à Monsieur Jean-Philippe Domenger, Professeur de l'Université Bordeaux 1, et à Monsieur Rémy Mullot Professeur à l'université de La Rochelle pour avoir accepté de rapporter sur mon mémoire.

Merci à Inoannis Anagnostopoulos pour le partage du bureau et pour son soutien dans mon travail. Merci aussi aux relecteurs et correcteurs de cette thèse : Nathalie Leborgne et Isabelle Musi-Colloc'h.

Merci enfin à toutes les personnes qui m'ont entouré et aidé (directement ou indirectement) durant ces quatre années.

Je dédie cette thèse à mes parents,
aux autres membres de ma famille
et à mes amis
qui m'ont soutenu au long de ce parcours

Résumé : Ce travail s'intéresse au contrôle des résultats de transcriptions numériques produites automatiquement par des logiciels de reconnaissance optique de caractères (OCR), lors de la réalisation de projets de numérisation de masse de documents. Le but de nos travaux est de concevoir un système de contrôle des résultats d'OCR suffisamment robuste pour être performant sur l'ensemble des documents numérisés à la BnF. Cette collection est composée de documents anciens dont les particularités les rendent difficiles à traiter par les OCR, même les plus performants. Nous avons conçu un système de détection des mots omis dans les transcriptions, ainsi qu'une méthode d'estimation des taux de reconnaissance des caractères. Le contexte applicatif exclu de recourir à une vérité terrain pour évaluer les performances. Nous essayons donc de les prédire. Pour cela nous proposons différents descripteurs qui permettent de caractériser les résultats des transcriptions. Cette caractérisation intervient à deux niveaux. Elle permet d'une part de caractériser la segmentation des documents à l'aide de descripteurs de textures, et d'autres part de caractériser les textes produits en ayant recours à un second OCR qui joue le rôle d'une référence relative. Dans les deux cas, les descripteurs choisis permettent de s'adapter aux propriétés des corpus à contrôler. L'adaptation est également assurée par une étape d'apprentissage des étages de décision ou de prédiction qui interviennent dans le système. Nous avons évalué nos systèmes de contrôle sur des bases d'images réelles sélectionnées dans les collections documentaires de la BnF. Le système détecte $84,15\%$ des mots omis par l'OCR avec une précision de $94,73\%$. Les expérimentations réalisées ont également permis de montrer que 80% des documents présentant un taux de reconnaissance mots inférieur à 98% sont détectés avec une précision de 92%. On peut également détecter automatiquement 45% des documents présentant un taux de reconnaissance inférieur à 70% avec une précision supérieure à 92%.
Mots clés : erreur de segmentation, texture, classification, reconnaissance de caractères, prédiction de performances.

Table des matières

Table des figures

Liste des tableaux

Introduction

Les projets de numérisation de masse permettent aux services d'archives et aux bibliothèques nationales de préserver les collections tout en fournissant aux utilisateurs un meilleur accès aux documents via l'internet. Plusieurs plateformes numériques dans le monde offrent aujourd'hui un accès à des millions de documents (comme la bibliothèque numérique de la BnF, « Gallica [1] », la bibliothèque numérique des Etats-Unis, « DPLA », etc.).

La production de documents numériques nécessite un certain nombre d'actions dont la sélection physique des documents, leur numérisation au moyen de scanners, puis l'indexation des versions numériques par des moteurs de recherche.

Du fait de l'importance du contenu des documents patrimoniaux, les internautes sont devenus de plus en plus exigeants. Les métadonnées bibliographiques ne reflètent pas de façon exacte le contenu des documents et ne suffisent pas à l'indexation précise des contenus. L'utilisation des seules métadonnées bibliographiques par un moteur de recherche fausse les résultats des recherches et décourage les utilisateurs des bibliothèques numériques.

Pour faciliter la consultation des collections numériques, l'indexation des documents doit être réalisée sur les textes transcrits (indexation plein texte). Du fait de la quantité d'ouvrages qu'il faut transcrire, l'étape de transcription est réalisée de façon automatique grâce à des systèmes de reconnaissance optique de caractères (en anglais *Optical Character Recognition* ou *OCR*).

La technologie actuelle mise en œuvre par les systèmes de reconnaissance de caractères est basée sur la segmentation des composants textuels en caractères ou mots qui sont ensuite reconnus individuellement en utilisant des méthodes de reconnaissance de formes.

Les systèmes d'OCR sont devenus aujourd'hui de plus en plus des systèmes experts composés de plusieurs modules de traitement (pré-traitement, segmentation des composants de la page, reconnaissance de caractères, post-traitement). Chaque module tente de fournir les meilleurs résultats au module suivant. A cause de cette architecture séquentielle de traitement, les erreurs qui apparaissent dans les résultats d'un module de traitement ont des répercussions directes sur les résultats des modules suivants.

Les systèmes de reconnaissance actuels sont très efficaces sur une grande variété de documents récents. Cependant, sur des documents dégradés ou anciens, les performances des OCR deviennent médiocres. L'application des OCR sur des collections de qualité variable génère donc souvent des erreurs de segmentation et de reconnaissance de caractères. En effet, la dégradation des pages des documents, les défauts d'impression ainsi que la présence de lexiques particuliers (ancien français, vocabulaires scientifiques, etc.) dans les documents à reconnaître influent énormément sur la qualité des documents numériques qui

1. Bibliothèque numérique de la BnF : http ://gallica.bnf.fr/

sont produits au bout de la chaîne de traitement. Les erreurs que l'on peut trouver dans les résultats de l'OCR se décomposent en deux catégories : les défauts de segmentation et les défauts de reconnaissance.

Une part importante de la recherche actuelle porte sur l'amélioration des méthodes de reconnaissance de caractères appliquées au contexte des documents anciens et patrimoniaux. On notera cependant que si la majorité des efforts sont réalisés dans cette direction, très peu portent sur la prédiction et l'évaluation précise des performances des outils d'OCR. Or, dès lors que l'on considère un processus de numérisation de masse, pouvoir prédire en amont si un OCR sera capable de traiter un (lot de) document(s) ou non et contrôler en aval de la chaîne de numérisation que le résultat fourni est conforme aux exigences des spécifications, est un problème fondamental au regard des volumes traités dans les projets de numérisation de masse et donc des coûts (humains, financiers) et enfin des exigences de qualité (que ce soit pour des raisons de conservation ou de diffusion).

A l'occasion de cette étude, nous nous sommes intéressés à caractériser plusieurs étapes du processus de numérisation à la Bibliothèque nationale de France, afin de dégager des pistes d'amélioration et de les quantifier de manière rigoureuse. Dans un premier temps nous nous sommes intéressés aux critères qui agissent sur la qualité des résultats de reconnaissance des caractères afin de déterminer les sources susceptibles de défauts de segmentation et de reconnaissance des mots. Dans un second temps, nous nous sommes focalisés sur l'étape de contrôle des résultats de reconnaissance de caractères. Du fait des quantités très importantes d'ouvrages numérisés (30 000 pages par jour), il devient nécessaire d'automatiser en partie le processus de contrôle des données produites par les prestataires de numérisation.

Ce travail a été réalisé dans le cadre du Plan triennal de recherche mené par la Bibliothèque nationale de France, dont le but est d'offrir au Service de numérisation de la BnF les moyens et les outils nécessaires pour maîtriser sa mission de numérisation des documents. Notre travail a pour finalité d'être intégré dans la chaîne d'entrée des documents de la BnF.

L'objectif de notre travail est la création d'un système de contrôle des résultats des systèmes d'OCR. Notre système doit être générique et suffisamment robuste pour s'adapter à la variabilité de la collection documentaire de la BnF. Pour atteindre cet objectif, nous mettons l'accent sur l'utilisation d'un certain nombre de techniques génériques les plus appropriées à cette diversité de documents.

Ce document est organisé en deux parties. Dans la première partie, nous commençons par la présentation du contexte de notre travail, qui est relié aux projets de numérisation de masse de la BnF. Ensuite, nous présentons un état de l'art sur la chaîne de traitement des systèmes de reconnaissance de caractères, notamment les différentes opérations de préparation des images de caractères, le module d'analyse et de reconnaissance de caractères et enfin les procédures de contrôle des résultats de reconnaissance de caractères.

Dans la seconde partie de ce document, nous présentons notre contribution en ce qui concerne le contrôle des résultats de reconnaissance de caractères. Cette partie est composée de deux chapitres. Dans le premier chapitre, nous proposons une approche adaptative de vérification des résultats de segmentation des structures physiques des documents. La variabilité de la collection documentaire de la BnF nous a amené à appliquer une approche locale pour détecter les composants omis par l'OCR. Cette approche utilise des caracté-

ristiques génériques de texture pour modéliser les éléments de la page (texte, illustration et fond de page). La localité de notre approche se traduit par l'utilisation des composants déjà détectés par l'OCR sur un document ou un échantillon de pages pour construire les classifieurs de notre approche et les entraîner à la volée sur les documents en cours d'analyse. Cela permet d'adapter au mieux notre approche aux caractéristiques typographiques des documents traités (polices de caractères, tailles de caractères, défauts d'impression, etc.).

Dans le deuxième chapitre, nous présentons une deuxième contribution qui nous a permis de contrôler les résultats de reconnaissance de caractères à travers une procédure d'estimation locale du taux de reconnaissance des caractères. Nous avons aligné les résultats de reconnaissance des prestataires avec les résultats de reconnaissance d'un OCR tiers (appelé « OCR de validation ») qui va jouer le rôle d'une référence relative pour caractériser les résultats de reconnaissance des prestataires. Les taux de désaccord entre les deux OCR ainsi que les profils des couples de désaccords sont des caractéristiques indépendantes des caractéristiques typographiques et linguistiques des documents. Cela nous permet d'avoir une description générique applicable sur la totalité des types de documents de la BnF. De plus, pour renforcer l'aspect adaptatif de notre approche, nous avons appliqué une approche locale d'apprentissage qui utilise un échantillon de pages de la collection des documents à vérifier, pour former les estimateurs des taux de reconnaissance avec les caractéristiques locales des documents à vérifier.

L'évaluation de nos méthodes est réalisée sur des bases d'images réelles provenant des derniers marchés de numérisation de masse de la BnF. Afin de s'approcher du contexte variable de la collection documentaire de la BnF, nous avons regroupé dans les bases de validation des documents anciens et des documents récents. L'évaluation des deux approches de contrôle a montré des performances intéressantes dans la localisation des mots omis et dans l'estimation des taux de reconnaissance de caractères.

Enfin, nous finissons cette partie par une conclusion générale dans laquelle nous présentons un bilan du travail qui a été réalisé et des éventuelles améliorations qui peuvent être portées sur nos travaux de recherche.

Première partie

Contexte et état de l'art

Chapitre 1

Contexte : processus de numérisation des documents

De nombreuses institutions culturelles et archives mondiales ont actuellement recourt à la numérisation de masse de leurs collections de documents afin de les conserver et de les communiquer. Depuis 1991, la Bibliothèque nationale de France (BnF) a entrepris plusieurs projets de numérisation afin de préserver ses documents et les diffuser.

Le processus de numérisation à la BnF commence par une étape de programmation annuelle des documents à numériser et par une phase de sélection physique des documents. Ces deux opérations vont permettre de préparer les collections de documents avant le démarrage des projets de numérisation. Ensuite, tous les documents sélectionnés sont envoyés aux prestataires de numérisation qui procèdent à l'opération d'acquisition et de transcription automatique des pages à l'aide d'un OCR.

La capture numérique des pages est la deuxième opération réalisée dans le cadre des projets de numérisation de masse. Elle désigne le processus permettant de convertir l'information analogique contenue dans des pages en un signal numérique bi-dimensionnel représenté dans une matrice appelée image. Pour assurer la qualité des documents numériques, l'opération d'acquisition des images des documents doit prendre en considération les processus techniques mis en œuvre dans la procédure de conversion numérique des documents analogiques, ainsi que les attributs des documents sources eux-mêmes : taille et présentation, niveau de détail, gamme de tons, et présence ou non de couleur.

La reconnaissance des caractères est une étape cruciale dans la procédure de numérisation des documents. Elle permet de convertir les formes des caractères contenus dans les images en texte électronique (codé en formats ASCII, UTF-8 ou XML pour les documents structurés). La reconnaissance des caractères est réalisée grâce à des techniques d'analyse d'image et de reconnaissance de formes. Les résultats de ce traitement vont permettre la création de documents numériques et leur réutilisation par différents systèmes (publication de livres numériques, indexation dans les moteurs de recherche, etc.)

Enfin, une étape de contrôle de qualité est réalisée à la fin de la chaîne de numérisation des documents pour vérifier la qualité des images des documents et la qualité des résultats de la transcription des caractères.

Le travail présenté dans le cadre de cette thèse est réalisé dans le cadre du projet de recherche Plan triennal mené par la BnF. Ce projet vise à étudier les différents outils et méthodes permettant à la BnF de mieux maîtriser la qualité de ces documents numériques.

Pour cela, nous allons essayer dans cette partie de répondre aux questions suivantes : *Quel est le rôle de chaque étape dans la procédure de numérisation des documents ? Ou se réalise chaque étape ? Quand les étapes sont-elles effectuées ? Comment sont-elles appliquées ? Quel sont les facteurs dans chaque étape qui influent sur la qualité des documents ?*

Pour répondre à toutes ces questions, nous allons détailler dans ce qui suit les différentes étapes de réalisation des projets de numérisation de masse. Nous commençons donc par la présentation de l'opération de sélection des documents. Ensuite, nous exposons la procédure de numérisation des documents. Enfin, nous terminons cette partie par une présentation de l'opération de contrôle des documents numériques réalisées par la BnF et par les prestataires de numérisation. Dans chaque sous-partie, nous allons tenter de déterminer les critères de chaque opération ayant une influence sur la qualité des documents numériques.

1 La sélection des corpus

1.1 Introduction

La réalisation des projets de numérisation de masse commence toujours par une procédure de sélection des collections documentaires à numériser. L'objectif de cette procédure est d'assurer la qualité intellectuelle et physique des bibliothèques numériques. Cette procédure s'effectue au moyen de trois types d'examens portant sur :

- les droits d'auteur,
- les données bibliographiques des documents,
- les caractéristiques physiques des documents.

Pour les institutions culturelles, les paramètres les plus influents quant à la décision de sélection des documents sont généralement les statuts juridiques des documents à numériser (libres de droit ou sous droit d'auteur) ainsi que la valeur de leurs données bibliographiques (la disponibilité des informations bibliographiques des documents telles que le nom de l'auteur, le titre du document, sa date d'édition, etc.). En revanche, les caractéristiques physiques des documents ne sont pas d'une importance capitale dans la décision de sélection des documents. En fait, il arrive même qu'en présence de certains défauts physiques dans les documents, les institutions culturelles sélectionnent ces documents dans leurs projets de numérisation de masse du fait de leur valeur intellectuelle, ce qui engendre par la suite des lacunes dans les résultats de l'opération de reconnaissance des caractères.

En fait, dans les résultats de l'OCR, les défauts sont causés majoritairement soit par des défauts physiques dans les documents traités, soit par des défauts liés à la présence d'artefacts dans l'image. On peut décomposer les critères qui influent sur la qualité des résultats de reconnaissance automatique des caractères en trois catégories :

- *Critères de l'œuvre* qui concerne le contenu textuel du document, comme la présence de formules mathématiques et d'illustrations, le nombre de langues utilisées, etc.
- *Critères de l'ouvrage* qui concernent la qualité du papier et de l'encrage des pages, le type des caractères et la mise en page utilisés dans le document, etc.
- *Critères liés à la numérisation* qui couvrent la qualité de la numérisation (fidélité à l'original, netteté, bruit de numérisation, orientation, etc.) ainsi que les paramètres de numérisation (résolution, profondeur colorimétrique, etc.).

L'opération de sélection des documents est une tâche très importante dans les projets de numérisation de masse. D'une part, parce que la qualité des résultats de conversion de l'image est très liée aux caractéristiques des documents physiques et d'autre part parce qu'elle doit permettre d'optimiser le coût de la numérisation des documents. En effet, il sera coûteux de demander un taux OCR élevé sur des documents pour lesquels l'OCR aura de mauvais résultats. Inversement on ratera le but à atteindre si l'on rejette des documents que l'OCR pourrait traiter facilement avec un excellent taux de reconnaissance. La décision de sélection doit donc anticiper le résultat qui sera obtenu par l'OCR. La figure 1.1 représente la qualité des décisions de sélection selon trois plages de taux de reconnaissance automatique (les taux de reconnaissance inférieurs à 60%, les taux de reconnaissance entre 60% et 98% et les taux de reconnaissance supérieurs à 98%). Ces trois plages de valeurs supposent que l'intervalle 60%-98% est atteint automatiquement par l'OCR sans correction manuelle coûteuse et qu'un taux supérieur à 98% ne peut être atteint que par l'intervention d'une étape de correction manuelle. Selon cette figure, une très bonne décision de sélection est donc réalisée soit lorsqu'on décide de rejeter des documents qui produiront un taux de reconnaissance inférieur à 60%, soit lorsqu'on sélectionne des documents qui peuvent avoir une qualité de reconnaissance automatique comprise entre 60% et 98%, soit enfin lorsqu'on sélectionne des documents qui peuvent conduire à un taux de reconnaissance supérieur à 98% dans la classe des documents à haute qualité.

Réciproquement, une mauvaise décision de sélection est réalisée lorsque des documents de bonne qualité (taux de reconnaissance d'OCR supérieur à 98%) sont affectés à la classe des documents de haute qualité, car cette décision entraîne une perte économique puisqu'il est en principe possible d'obtenir une haute qualité de reconnaissance d'une manière automatique sans recourir à une procédure de correction manuelle. Une très mauvaise décision est également réalisée lorsque les services de sélection rejettent des documents de très bonne qualité. Cela signifie que l'opération de sélection des documents est trop stricte. Par conséquent, une opération fiable de sélection doit réaliser au moins des bonnes décisions. Pour cela, il faut prendre en compte tous les critères qui ont une influence sur la qualité des résultats de l'OCR. Nous détaillerons dans ce qui suit chacun de ces critères pour montrer leur impact sur la qualité des documents numériques.

1.2 Critères concernant l'œuvre

Certains critères de l'œuvre ont un impact négatif sur la qualité des résultats de reconnaissance des caractères. Les systèmes commerciaux de reconnaissance de caractères sont calibrés pour reconnaître du texte contemporain dactylographié ; ils utilisent pour cela des ressources linguistiques telles que des dictionnaires qui sont adaptés à la langue contemporaine. De ce fait ces ressources ne sont pas adaptées à des œuvres plus anciennes et peuvent même dégrader la reconnaissance des caractères dans ces situations.

Selon [Vay], tous les éléments non purement textuels peuvent perturber l'opération de reconnaissance des caractères. Par exemple les annotations, les tableaux, les formules mathématiques et chimiques et les chiffres sont autant d'éléments perturbateurs pour les OCR. Certaines langues et alphabets ont un impact sur la qualité des résultats de reconnaissance des caractères. En effet, les premiers OCR du XX^e siècle développés pour les archives traitent des documents en anglais qui ne comportent quasiment aucun signe

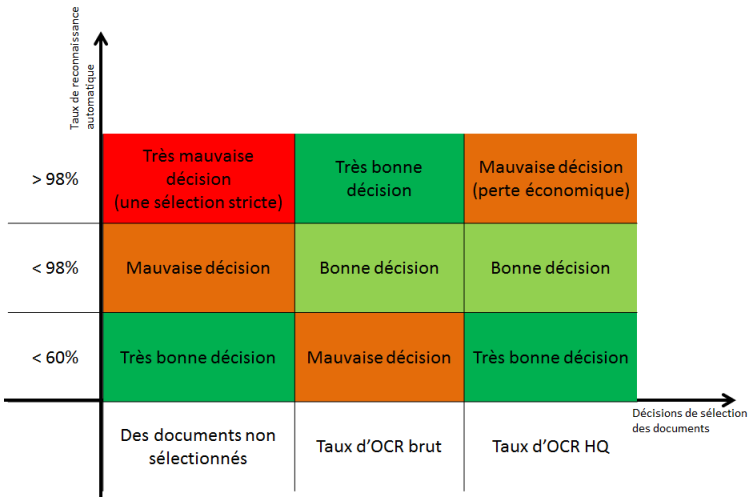


FIGURE 1.1 – L'évaluation de la décision de sélection des documents en fonction du taux de reconnaissance automatique des caractères

diacritique (excepté le point sur le i). Lorsque les OCR ont commencé à travailler sur d'autres langues, comme le français, les problèmes de reconnaissance des accents sont apparus. Les accents ont la spécificité d'être beaucoup plus petits qu'un caractère. Certains traitements de suppression de bruit les considèrent comme du bruit, ce qui élimine tous les accents trop petits. La fréquence importante d'accents rend donc le français, l'espagnol et a fortiori les langues slaves (écrites avec l'alphabet latin) plus difficiles à reconnaître que l'anglais.

La longueur moyenne des mots est un paramètre qui caractérise chaque langue. Certains systèmes de reconnaissance de caractères utilisent ce paramètre pour optimiser la procédure de segmentation des mots dans les phrases. La longueur moyenne des mots de la langue anglaise est d'environ six caractères. Alors que la longueur moyenne des mots pour l'allemand et le finnois est plutôt proche de neuf caractères. Si on n'optimise pas ce paramètre avant le lancement de la procédure d'analyse des images des documents, des erreurs de segmentation peuvent survenir.

D'après [Vay], les systèmes de reconnaissance de caractères travaillent mieux avec une seule langue par unité documentaire. Même s'il est possible d'ajouter un dictionnaire d'une autre langue, cet ajout peut engendrer des erreurs considérables sur le reste des résultats de reconnaissance. De même, la présence d'alphabets non latins peut être préjudiciable à la reconnaissance globale du texte et même affecter la qualité de la segmentation.

Les textes comportant des références et des citations utilisent souvent les notes de bas de page et le style italique. Les notes de bas de page sont généralement composées avec une petite taille de police. La variation des tailles de police sur une page gène énormément l'opération de reconnaissance des caractères qui aura du mal à estimer la taille moyenne de la police de la page. De plus, le style italique engendre souvent des erreurs de reconnaissance de caractères.

Dans les textes narratifs et descriptifs, la majorité des signes de ponctuation sont des points et des virgules. Dans les documents scientifiques, les points sont beaucoup plus présents que les virgules car ils sont utilisés dans l'écriture des nombres réels, dans les abréviations et dans les fonctions scientifiques. Compte tenu de leur apparence dans

le texte, les points et les virgules sont souvent similaires, ce qui empêche les méthodes de reconnaissance de formes de faire la distinction entre les deux. Les traits d'union «-» et les traits de soulignement «_» sont relativement courants. Certains systèmes de reconnaissance optique de caractères ne les distinguent pas. On peut trouver dans les textes d'autres signes typographiques tels que guillemets, apostrophes et parenthèses. Dans certaines polices de caractères, les apostrophes et les guillemets se distinguent des virgules et des points grâce à leur position par rapport à la ligne de base. Dans les langues française et espagnole, les guillemets sont utilisés comme des marques de citation. Les problèmes posés par les ponctuations sont les mêmes que ceux posés par les accents : ils sont de petite taille et risquent donc d'être supprimés lors du processus de traitement du bruit.

1.3 Critères concernant l'ouvrage

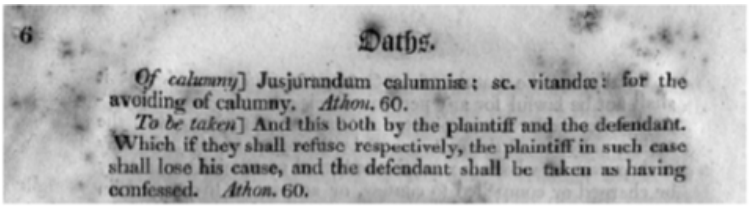
Les critères de l'ouvrage désignent l'ensemble des caractéristiques physiques du document telle que la qualité de papier, le type d'encrage, la taille des marges, etc. Ces critères ont un impact direct sur la qualité des résultats de l'OCR. En effet, les systèmes de reconnaissance de caractères utilisent les images des pages de document pour transcrire le contenu textuel des documents. Si le document original possède des défauts physiques (par exemple des tâches d'encre), ces défauts vont apparaître sur les images, ce qui engendrera des erreurs de reconnaissance et de segmentation. Nous pouvons décomposer les critères de l'ouvrage qui ont un impact sur la qualité de reconnaissance en trois classes :

- Critères liés au support : reliure, ouverture du document, papier, format, etc.
- Critères liés à la mise en page : marges, nombre des colonnes, orientation de l'écriture, etc.
- Critères liés à la typographie et à l'impression : intensité d'encrage, lacunes d'encrage, police des caractères, taille des police, présence de l'italique, etc.

L'ouverture du document définit l'angle maximal entre deux pages. La plupart des documents anciens possèdent des reliures très serrées. Cela rend difficile la procédure de mise à plat des documents pour les numériser avec les procédures classiques. Par conséquent, la numérisation de ces documents est réalisée sur des numériseurs particuliers qui scannent les pages sans mise à plat. Cette méthode de numérisation engendre l'apparition de zones noires dans les parties de l'image qui sont proches de la reliure (cf. figure 1.2) ce qui peut altérer le contraste de l'image de la page et engendrer la perte des éléments proches de la reliure.

FIGURE 1.2 – Exemples de zone sombre côté reliure

La qualité du papier est définie dans notre contexte par sa capacité à présenter les informations de façon claire. L'acidité des pages peut morceler et effriter les pages du document. S'il existe des taches ou/et des jaunissements sur les pages, le document numérisé risque d'être illisible. Les taches proviennent de la dégradation du papier à travers le temps. La figure 1.3 montre un exemple d'image d'un document qui contient des taches sombres et des taches claires. Les taches claires génèrent des caractères fragmentés et des formes variables. Par conséquent, les systèmes de reconnaissance de caractères peuvent considérer les différentes parties du caractère comme des éléments distincts, ce qui engendre des erreurs de segmentation et de reconnaissance. De plus, les taches sombres engendrent des ambiguïtés sur les formes de caractères et peuvent boucher des boucles de caractères.


6 Oaths.

Of calumny] Jusjurandum calumniæ; sc. vitandæ: for the avoiding of calumny. *Athon.* 60.

To be taken] And this both by the plaintiff and the defendant. Which if they shall refuse respectively, the plaintiff in such case shall lose his cause, and the defendant shall be taken as having confessed. *Athon.* 60.

FIGURE 1.3 – Exemple de document avec un papier dégradé

La transparence du papier provoque aussi des erreurs de reconnaissance de caractères. En effet, si le papier est très transparent au point que l'écriture qui se trouve sur une face apparaisse sur l'autre face, le processus de conversion par l'OCR devient difficile.

Critères liés à la mise en page

La structure textuelle du document définit l'agencement physique des blocs sur la page. Les textes présentés sous formes de colonnes (comme les journaux de presse) ou habillés par des motifs graphiques rendent l'opération de reconnaissance de caractères complexe et provoquent beaucoup d'erreurs dans les résultats de l'OCR [Vay] (cf. figure 1.4).


LINGUISTIC PUBLICATIONS.

TRÜBNER & CO.,

FIGURE 1.4 – Exemples de documents à plusieurs colonnes : l'image à gauche représente un document multilingue composé en double colonne. L'image à droite est une page de presse avec des publicités encadrées par des bordures.

Les marges sont les espaces blancs qui séparent la zone de texte des bords de la page. Les ouvrages présentant de faibles marges ou des éléments textuels proches de la reliure sont généralement mal traités par les systèmes de reconnaissance de caractères. Les types de défauts obtenus sont semblables à ceux liés à une ouverture réduite du document.

L'orientation de l'écriture dans la page est par convention horizontale pour la plupart

des systèmes de reconnaissances de caractères latins. Une tolérance d'environ 12° est admise pour gérer les documents imprimés ou numérisés légèrement en oblique. L'existence de blocs textuels verticaux dans les documents, notamment dans les légendes et les tableaux, perturbe énormément l'opération de conversion des documents et engendre des erreurs de segmentation et de reconnaissance des caractères. Pour éviter ces erreurs, une procédure de gestion d'hypothèses d'orientation des blocs textuels doit être appliquée. Ce traitement supplémentaire n'est pas toujours présent dans les systèmes de reconnaissance de caractères. Par ailleurs, si ce traitement n'est pas parfait, la reconnaissance sera dégradée.

Critères liés à la typographie et à l'impression

La typographie est l'art de la communication à travers les textes imprimés. Elle regroupe tous les critères tels que les styles, les polices et les tailles des caractères utilisés pour composer les textes du document. Une bonne typographie doit permettre un accès aisé et clair au lecteur, elle ne doit pas perturber la lecture par des ornementations trop prégnantes.

La typographie des documents joue un rôle déterminant dans la complexité d'un document par rapport à un processus de transcription automatique. Ainsi, lorsque certaines polices de caractères utilisées dans l'impression sont dérivées de la calligraphie médiévale, par exemple, les typographies gothiques, celles-ci rendent le processus de conversion délicat.

Les systèmes de reconnaissance de caractères reconnaissent les caractères à travers leurs formes. Cependant, certaines polices de caractère peuvent engendrer des ambiguïtés quant aux formes des caractères. En effet, pour que les caractères soient bien reconnus, leurs formes doivent être suffisamment invariantes. Dans la reconnaissance de caractères, la notion d'invariance est généralement très reliée à son champ d'application, c'est-à-dire suivant les caractères traités, une simple modification peut être invariante comme elle peut être variante. Prenons l'exemple des caractères **Q** et **O** : le jambage qui se trouve audessous du cercle distingue les deux caractères. Par contre, la présence de l'empattement dans l'extrémité supérieure du **C** en police Times New Roman ne devrait pas générer de différence avec la forme de caractère **C** en police Arial. Les ambiguïtés peuvent également venir de l'alphabet lui-même. Ainsi, dans l'alphabet latin, certains caractères minuscules ont la même forme que certains chiffres, par exemple **l** et **1** , le **O** et **0**. Dans l'alphabet arabe, la forme d'une lettre dépend parfois de ses voisines selon une grammaire bien formée graphiquement. La présence de symboles spéciaux (comme par exemple **$** , **%** , **§**, σ , etc.) dans les rapports techniques et les articles scientifiques compliquent généralement l'opération de conversion des textes. En effet, l'ambiguïté entre les formes des caractères spéciaux est beaucoup plus fréquente que pour les caractères normaux. Les documents présentant plusieurs polices de caractères sont aussi plus difficiles à convertir que les documents présentant une seule police. En effet, il est possible qu'un symbole similaire corresponde à deux caractères différents présents dans deux polices différentes.

Les défauts d'impression couvrent tous les artefacts qui nuisent à la clarté de l'écriture et altèrent les formes des caractères. Par exemple, les lignes blanches sur les éléments textuels causées par des défauts du ruban d'impression ou l'encrage délicat du document produisent un effet de caractères cassés, ce qui gêne le processus de reconnaissance de

caractères. De même, le faible encrage des documents dégrade le contraste entre les caractères imprimés et le fond (papier) ce qui agit sur la lisibilité du document. Au contraire, un encrage trop appuyé du document peut épaissir les caractères et provoquer des formes non uniformes et des boucles bouchées (comme pour les **a** , **e** et **o**). De plus, les documents anciens sont caractérisés par des espaces non uniformes entre les mots et les caractères des mots, ce qui peut générer des problèmes de segmentation des blocs de différents niveaux.

1.4 Critères liés à la numérisation

Outre les critères intrinsèques liés au document et à la façon dont il a été imprimé, le mode d'acquisition des images des documents a une influence sur la qualité des résultats de reconnaissance des caractères.

Le bruit de numérisation est l'un des artefacts qui gêne énormément les algorithmes de conversion des documents. Le bruit est un signal aléatoire qui apparaît sur l'image à cause des défauts des capteurs du scanner. L'apparition de bruit de numérisation entraîne des déformations dans les formes des caractères, ce qui génère des erreurs dans les résultats de l'OCR.

La dynamique de l'image joue un rôle déterminant dans la qualité de l'image. Une étude présentée dans [Ant11] a montré que le passage en niveau de gris permet d'améliorer notablement le taux de reconnaissance de l'OCR. La binarisation de l'image des documents réduit l'information de couleur de l'image. Mais si le seuil de binarisation n'est pas approprié, des artefacts peuvent apparaître sur les images du document, ce qui engendre des erreurs de reconnaissance de caractères.

Selon [Hol09], le mauvais choix de la résolution des images des documents peut conduire soit à un manque d'information (en sous-échantillonnage), soit à un surplus d'information (en sur-échantillonnage) qui se traduit souvent par la présence d'un bruit plus abondant. Les modèles des systèmes de reconnaissance de caractères sont obtenus avec une résolution d'image définie au préalable. Pour caractériser les formes de caractères, les OCR commencent toujours par l'adaptation de la taille des images des caractères à la taille des images de modèle. Cette opération d'adaptation peut produire des images pixélisées si la résolution de l'image des documents est inférieure à la taille des images des modèles de caractères et si on n'utilise pas un algorithme de ré-échantillonnage adapté.

La compression abusive de l'image a forcément un effet négatif sur les résultats de l'OCR. En effet, la compression avec perte engendre l'apparition d'artefacts visuels qui gênent énormément la procédure de conversion des documents. Par conséquent, pour garder les informations de l'image intactes, il est préférable d'utiliser un format d'image non propriétaire qui utilise un algorithme de compression sans perte.

Le contraste est le paramètre qui permet de faire varier l'accentuation ou l'atténuation des transitions de noir/blanc. La luminosité permet de jouer sur l'éclairage du document à capturer. Ces deux paramètres sont souvent corrélés et ils jouent un rôle très important dans la qualité des résultats de reconnaissance. Des expériences réalisées dans [Fré10] ont montré que l'on peut passer d'un taux de reconnaissance de 99% à 0% en variant légèrement le contraste et la luminosité des images du document.

1.5 Conclusion

De manière générale, les systèmes de reconnaissance des caractères atteignent des bonnes performances sur des documents récents caractérisés par un bon état physique et des propriétés typographiques standards. Au contraire, les performances décroissent significativement sur les documents anciens et patrimoniaux qui font l'objet des projets de numérisation de masse de la BnF. Ceci influe sur l'intégrité des documents numériques de la bibliothèque numérique « Gallica » de la BnF.

Pour maîtriser la qualité des résultats de l'OCR, on peut se baser sur les critères de l'œuvre et de l'ouvrage lors de la procédure de sélection des documents. Cependant, ce démarche n'est pas toujours aisée. En effet, l'accumulation des défauts physiques ainsi que leur distribution aléatoire dans les documents rendent la procédure préalable d'estimation des qualités d'OCR difficile. Prenons l'exemple des dates d'édition des documents, qui peuvent donner une indication sur l'état physique et typographique des documents. Plus les documents sont anciens, plus leur état physique est médiocre. Ceci conduit, d'après ce que nous avons présenté précédemment, à des erreurs de segmentation et de reconnaissance de mots dans les résultats de l'OCR. Pour s'assurer de cette hypothèse, on a présenté dans la figure 1.5 la distribution de 924 documents en fonction des taux de reconnaissance des caractères et des dates d'édition. Les documents de cette analyse sont sélectionnés aléatoirement à partir de la base des documents de la BnF.

D'après cette figure, on constate que conformément à l'hypothèse précédente, les documents édités entre le XVI^e siècle et le $XVII^e$ siècle ont des taux de reconnaissance très médiocres (inférieurs à 60%). Ces taux s'améliorent de manière significative en passant au XX^e siècle. Cependant, l'étendue des taux de reconnaissance des documents édités après 1700 est très importante (variation entre 5% et 99%). Pour affiner cette analyse, on a étudié aussi la distribution de ces documents en fonction des courbes de la figure 1.6. Ces courbes illustrent les variations du taux moyen de reconnaissance, la variation des taux de reconnaissance et le nombre des documents en fonction de la date d'édition. D'après cette figure, nous constatons que les taux de reconnaissance suivent une allure croissante en passant du XV^e siècle au XX^e siècle. Cette constation prouve l'existence d'une relation entre les dates d'édition et les taux de reconnaissance des caractères. Cependant, d'après la courbe de l'écart-type des taux de reconnaissance des caractères (courbe verte), nous constatons que des variations importantes des taux de reconnaissance sont enregistrées pour les documents qui ont des taux de reconnaissance moyens supérieurs à 60%. Cela signifie que les taux de reconnaissance de ces documents sont très variables.

En plus des critères de l'œuvre et de l'ouvrage, la BnF se base aussi sur la valeur intellectuelle du document pour déterminer la qualité des résultats de l'OCR qu'elle veut obtenir. Cette caractéristique n'a aucune lien avec les critères de l'œuvre et de l'ouvrage qui influent sur la qualité des résultats de l'OCR, alors qu'elle est la caractéristique principale de l'opération de sélection des documents de la BnF. Par conséquent, à partir de cette analyse, on peut déduire que la sélection des documents en fonction des critères de l'œuvre et de l'ouvrage n'est pas une procédure aisée à mettre en place dans la chaîne de numérisation de la BnF. Afin de déterminer notre champ d'action, nous présentons dans la section suivante les opérations principales de numérisation de masse adoptée par les acteurs de numérisation.

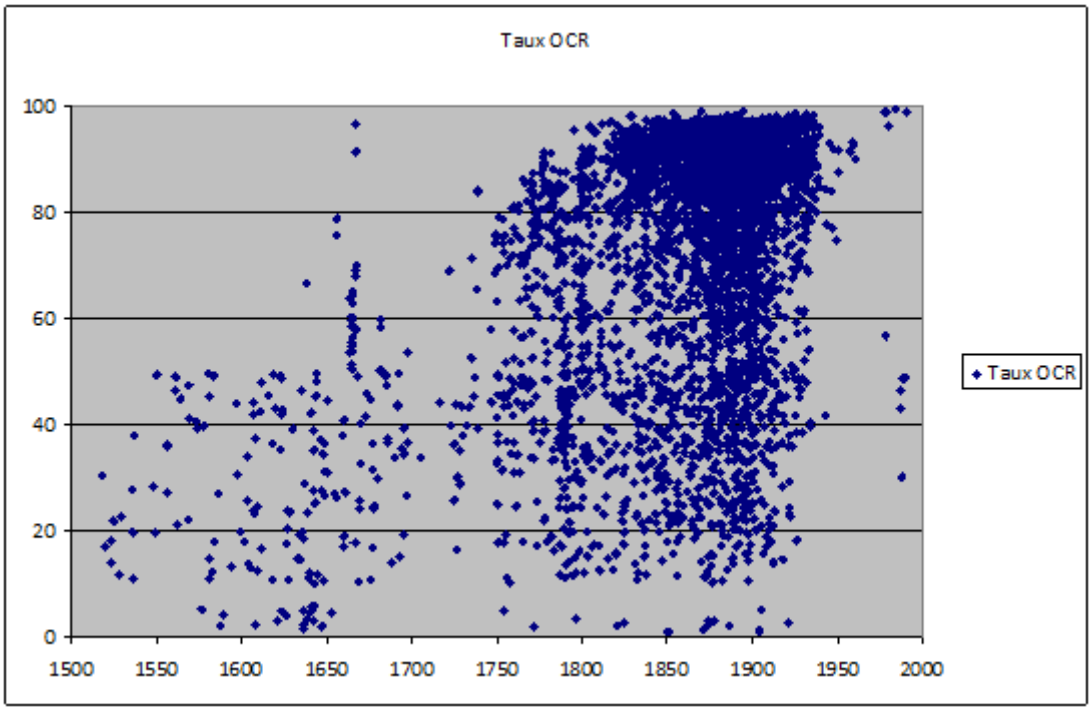


FIGURE 1.5 – Taux OCR moyen des documents vs leur date dédition

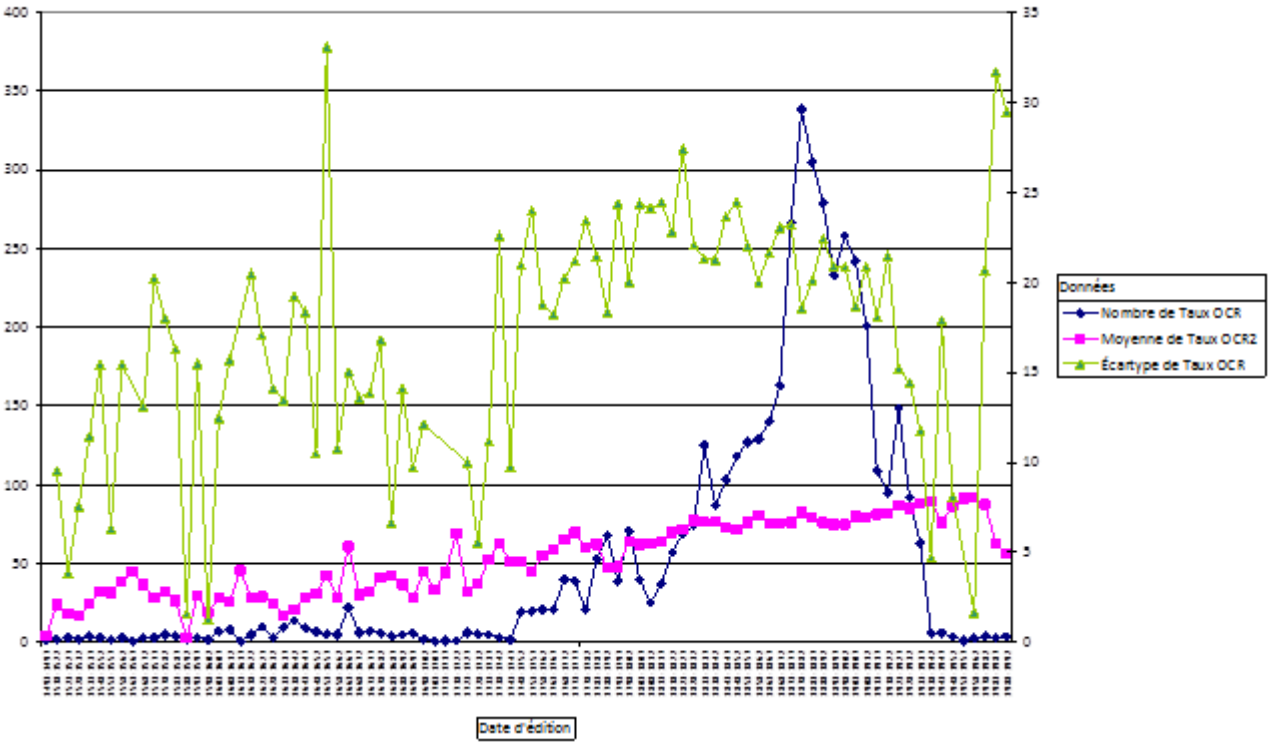


FIGURE 1.6 – Date édition, nombre de documents, taux OCR moyen et écart type

2 La chaîne de numérisation / OCRisation

2.1 Introduction

Après la procédure de sélection des collections des documents à numériser, la BnF procède à la production des documents numériques. Cette opération est réalisée en majorité par des prestataires externes. Elle commence toujours par l'acquisition des images des documents, laquelle nécessite un certain nombre de paramétrages qui permettent de produire des images lisibles pour les lecteurs et utilisables par les systèmes de reconnaissance de caractères.

Le choix du format d'image, de la dynamique de couleur et de la résolution d'image dépend des besoins de numérisation et du type des documents à numériser. En effet, les images des documents produites pour être communiquées sur le Web sont très différentes des images produites pour l'archivage numérique. Nous avons montré en évoquant les critères de sélection des documents (cf. section 1) que la qualité des résultats de reconnaissance de caractères est très dépendante de la qualité des images des documents reproduits. Par conséquent, il faut choisir les bons paramètres d'acquisition pour garantir l'intégrité des bibliothèques numériques et la qualité de l'opération de conversion textuelle des images des documents.

Dans ce qui suit, nous commençons au début par une présentation de l'opération de numérisation des documents en fonction des politiques de numérisation des documents. Après nous présentons la procédure de conversion des documents en spécifiant les exigences de qualité de la BnF.

2.2 Numérisation des documents

Numérisation faite pour l'archivage

Pour des fins de préservation, la numérisation des documents doit être réalisée de manière à être la plus fidèle possible au document original. Certaines politiques de préservation exigent même la conservation de la texture des pages des documents, ce qui engendre des spécifications techniques particulières.

Les images destinées à l'archivage doivent être non compressées, sans artefacts numériques résultant d'une compression avec perte. De plus, elles doivent être prises avec une résolution suffisante pour être employées dans des chaînes d'impression. A la BnF, afin de garantir une représentation suffisante du contenu textuel des documents, tous les documents traités dans le cadre des projets de numérisation de masse sont numérisés avec une résolution d'au moins 300 dpi. Certains documents dans les collections de la BnF sont numérisés avec des résolutions de 600 dpi et de 900 dpi ou plus pour donner une représentation plus fine du document, notamment pour les documents de petite taille (pièces de monnaie).

La colorimétrie des images des documents varie selon le type des documents traités. Par exemple, pour les manuscrits anciens, la numérisation couleur sera nécessaire dans la plupart des cas même lorsque ceux-ci ne comportent pas d'enluminures ou d'illustrations, les textes pouvant comporter des caractères et/ou des soulignés de différentes couleurs ; par ailleurs le fond de page comporte aussi des couleurs contrastées en fonction de l'histoire du document. Par conséquent, il est possible de distinguer les documents considérés

comme des **trésors** irremplaçables qu'il faudra numériser avec une très haute qualité et en couleurs, de ceux que l'on peut se contenter de traiter en niveaux de gris et avec une résolution moins importante. Pour les manuscrits et les imprimés récents qui ne possèdent pas d'illustrations, on pourrait se contenter de les numériser en niveaux de gris, mais il n'est pas possible de vérifier de façon exhaustive avant la numérisation d'un document qu'aucune de ses pages ne comporte pas d'ajout en couleur voire au crayon. Dans le marché de reproduction des originaux numériques, les deux standards de profondeur utilisés pour spécifier la plage de dynamique des couleurs de l'image :

- 8 bits (28) = 256 tons (pour la numérisation en niveaux de gris)
- 24 bits (2^{24}) = 16, 7 millions de tons (pour la numérisation en couleur RVB).

La numérisation pour la préservation peut être effectuée à des coûts raisonnables, étant donné la cadence des scanners de dernière génération, et avec une qualité uniforme. Les originaux numériques prennent plus de place mémoire puisqu'ils ne sont pas compressés. En fait, la BnF n'utilise aucune compression pour produire les originaux numériques des documents pour être le plus fidèle possible aux caractéristiques du document original. Ceci engendre des couts de stockage important pour conserver la totalité de base des documents numériques de la BnF. Aujourd'hui, la BnF utilise un espace de stockage proche de quatre pétaoctets. Par contre, dans la pratique le coût le plus important dans un processus de conversion de document est le temps humain qui est proportionnellement lié à la masse physique des documents stockés. Ce coût peut être jusqu'à 100 fois plus élevé par page que celui du stockage. Par ailleurs, les coûts de stockage informatisé ont été divisés par 2 tous les trois ans sur les vingt dernières années et des progrès sont encore à venir dans ce domaine ; il faut cependant prendre en compte les frais de gestion de l'archivage (surveillance des supports, sauvegardes, migrations...).

Numérisation faite pour la consultation

La diffusion des documents numériques par internet nécessite des images avec des spécifications techniques adaptées aux opérations d'échange distant de données. La taille des images de document doit être raisonnable afin d'assurer une consultation fluide des documents numériques sur internet tout en conservant la lisibilité des images des pages.

Les originaux numériques produits en format non compressé constitueront le fichier maître pour l'archivage à partir duquel on produira les images de consultation. Les images de consultation des documents doivent être compressées afin d'alléger leur taille. Le choix du mode de compression (**avec perte** ou **sans perte**) dépend de la qualité de rendu d'image souhaitée et du taux de perte toléré. Si la taille des images n'est pas assez réduite en utilisant la compression sans perte, on peut procéder à l'abandon d'une partie de l'information contenue dans l'image sans que cela soit perceptible à l'œil. Cette compression peut être menée soit avec un logiciel spécial (par exemple : Kakadu [1]), soit à travers le matériel de numérisation (scanners) qui assure directement la compression, ce qui est particulièrement utile lors de la création d'un grand nombre de fichiers ou de très gros fichiers.

Les formats d'images utilisés pour la consultation doivent permettre la transmission et l'échange entre une grande variété de plateformes et d'applications. Les formats JFIF,

1. Logiciel de compression d'image : http ://www.kakadusoftware.com/

GIF, JPEG 2000 et PNG sont très utilisés pour la publication des images des documents sur le Web et ils sont compatibles avec la plupart des navigateurs web.

Les documents numériques de Gallica ont généralement une résolution de 300 dpi. La compression minimale utilisée pour générer les images de documents de consultation est égale à 60% pour les manuscrits et à 75% pour les imprimés. Pour compresser les images de Gallica, la BnF utilise l'algorithme de compression de format JPEG2000 qui permet de produire des images en plusieurs taux de compression et avec plusieurs résolutions d'image.

2.3 La conversion textuelle des images des documents « OCRisation »

La transcription du contenu des documents en format textuel est une opération cruciale dans les projets de numérisation de masse. Elle assure la réutilisation du contenu textuel des documents par d'autres systèmes tels que les moteurs de recherche ou les générateurs de fichiers ePub ou PDF. Si les techniques de reconnaissance de l'écriture manuscrite ne sont pas assez mûres pour qu'elles soient appliquées dans des chaînes de production de masse, les techniques de reconnaissance des textes imprimés offrent des bonnes performances sur le texte imprimé de bonne qualité lorsqu'on utilise les meilleures solutions du marché, parmi lesquelles on peut citer ABBYY, Readiris, Tesseract ou OmniPage. Dans cette partie, nous allons nous limiter à une présentation générique des systèmes de reconnaissance optique des caractères en mettant uniquement l'accent sur les exigences de la BnF.

Les OCR sont des programmes informatiques qui assurent la segmentation de la page numérisée afin de déterminer les zones comportant du texte, des tableaux et des illustrations. Ensuite chacune des zones textuelles identifiées est elle-même segmentée finement, par ligne, puis par mot et caractères. Les zones des tableaux subissent aussi une opération de segmentation plus fine qui permet de segmenter et d'identifier les lignes et les colonnes du tableau. L'ensemble des opérations de segmentation des documents permet par la suite d'identifier la structure de la page.

Ensuite, tous les mots segmentés sont renvoyés au moteur de reconnaissance de caractères pour identifier les classes de ses caractères. Une fois que tous les caractères de la page ont été identifiés, l'OCR procède à la réalisation des opérations de post-traitement qui permet de valider les résultats de reconnaissance de formes ou de les corriger dans une certaine mesure en utilisant un dictionnaire ou un modèle de langage. Une présentation plus détaillée de ces étapes est exposée dans le chapitre 2 de cette partie.

Actuellement, la BnF demande trois qualités de conversion des textes dans ses projets de numérisation de masse :

- Une qualité brute obtenue automatiquement sans aucune opération de correction manuelle et dans laquelle les documents numériques ont des taux de reconnaissance de caractères qui varient entre 60% et 98,5%.
- Une qualité garantie à au moins 98,5% obtenue soit automatiquement, soit à travers une assistance humaine (post correction) si le taux de 98,5% exigé par la BnF n'est pas assuré (dans le dernier marché de numérisation, 80% des documents convertis sont demandés avec une qualité garantie),
- Une qualité supérieure obtenue toujours avec une assistance humaine. Cette qualité supérieure correspond à un taux de reconnaissance supérieur à 99,9%. (dans le dernier marché de numérisation, 20% des documents convertis sont demandés avec une qualité supérieure).

Les taux de reconnaissance fournis par les OCR sont calculés sur les régions lisibles de la page. Cela signifie que les mots omis et illisible (très difficile à reconnaître) sont exclus de la procédure de calcul des taux de confiance. Par conséquent, les qualités réelles des résultats de reconnaissance sont significativement plus faibles que celles qui ont été estimées automatique par l'OCR.

2.4 Conclusion

L'opération de capture et de conversion textuelle des documents est réalisée majoritairement par des prestataires externes. Seuls les documents précieux sont numérisés et océrisés en interne dans les ateliers de numérisation de la BnF. Cette politique de numérisation a été mise en place afin d'assurer une production de documents numériques en masse.

La BnF précise ses exigences de qualité dans les cahiers des charges. Le choix des paramètres des scanners et les algorithmes de compression est généralement discuté avec le prestataire pendant la phase de test du projet de numérisation de masse. Cependant, la BnF n'intervient ni dans l'opération de numérisation des documents ni dans l'opération de conversion textuelle des images, ce qui est contraignant pour notre travail car on ne peut pas agir à ce niveau pour maîtriser la qualité des documents numériques de la BnF.

3 Le contrôle

3.1 Introduction

Bien que les performances des OCR soient très bonnes sur des collections de documents récents, les résultats fournis par les prestataires sur les collections patrimoniales de la BnF englobent plusieurs types de défauts tels que des erreurs d'identification de la structure des fichiers, des erreurs d'identification de la structure physique de la page et des erreurs de reconnaissance de caractères. Cela exige une opération de contrôle qui doit être réalisée conjointement par la BnF et les prestataires de numérisation de masse avant la mise en consultation des documents dans Gallica.

Dans cette partie, nous allons exposer l'ensemble des opérations de contrôle réalisées sur les documents numériques de la BnF.

3.2 Opérations de contrôle réalisées par la BnF

Le service informatique (DSI) de la BnF prend en charge l'intégration des documents numériques envoyés par les prestataires dans les serveurs internes de la bibliothèque. Par ailleurs la chaîne d'entrée BnF mise en œuvre par le DSI effectue un certain nombre de contrôles techniques qui sont réalisés sur les fichiers désignés RefNum et TagTiff à travers des outils automatiques de vérification de structure physique. Ces examens cherchent à vérifier et valider la structure de ces fichiers. Si la composition du fichier n'est pas conforme avec les normes des fichiers RefNum ou TagTiff, les documents numériques sont rejetés.

En parallèle, le service de numérisation assure un contrôle sur les documents numériques. Les examens réalisés cherchent à vérifier la conformité des informations bibliographiques qui se trouvent dans le fichier RefNum avec les données catalogue du document.

Le service valide également dans le cadre de ce contrôle les tables des matières et index, ainsi que l'ordre des images des pages du document. Les documents numériques qui contiennent des défauts sont rejetés partiellement. Tous les documents numériques admis ou rejetés partiellement lors de ces contrôles sont vérifiés de nouveau par le même service après nouvelle livraison par le prestataire.

D'autre part, le service vérifie la conformité des propriétés de l'image avec les exigences mentionnées dans le cahier de charges et dans la charte de traitement. Les erreurs détectées sont classées en deux catégories :

- *erreurs mineures* : erreurs de pagination, changement d'échelle au sein de l'ouvrage, ombre portée qui ne gêne pas la lecture, etc.
- *erreurs majeures* : flou, image non contrastée, bruit de numérisation, etc.

3.3 Opérations de contrôle réalisées par les prestataires de numérisation

En plus du contrôle des images des pages, les prestataires de numérisation assure un contrôle sur les résultats de reconnaissance des caractères. Les lots de contrôle regroupent des documents numériques par type de niveau de qualité exigée, par genre de document (périodique, monographie, etc.), par période et par filière (techniques, scientifiques, juridiques, etc.).

Les opérations de contrôle concernent les résultats de segmentation des images et les résultats de reconnaissance des caractères. Les lots contrôlés sont constitués selon un processus d'échantillonnage qui respecte la norme ISO 2859-1. L'échantillonnage regroupe les documents numériques par lot de documents en fonction du nombre de pages de chaque document ; le nombre des pages contrôlées par lot varie entre 1201 et 3200 pages.

Le contrôle des résultats de segmentation est réalisé par des opérateurs humains à travers une application dédiée qui offre la possibilité de visualiser les images des pages superposées avec les boîtes englobantes des éléments de la page. Par conséquent, le contrôle réalisé sur les résultats de segmentation des documents numériques est un contrôle visuel. Ce contrôle est réalisé à plusieurs niveaux (mots, ligne, paragraphes) selon les exigences de qualités de la BnF.

L'ordre de lecture est également contrôlé visuellement. En cas de constat d'erreur, un compteur d'erreurs est incrémenté en fonction du type de l'erreur constatée. Comme dans les contrôles réalisés par la BnF, on peut distinguer deux types d'erreurs de segmentation :

- *erreurs mineures* : Oubli de bloc, Mauvais type de bloc, La page n'est pas une page blanche, Ordre de lecture non respecté, etc.
- *erreurs majeures* : Mauvais positionnement de bloc, Recouvrement excessif, etc.

Tous les identifiants des blocs qui possèdent des erreurs de segmentation sont également enregistrés afin de permettre la validation des erreurs. En effet, pour pallier aux erreurs humaines, une deuxième opération de contrôle est effectuée par un opérateur expert afin de vérifier systématiquement toutes les pages en erreur ainsi qu'un échantillon des pages sans erreur.

Les opérations de contrôle des résultats de l'OCR concernent aussi la qualité de la transcription des caractères des documents. Cette opération concerne la vérification d'une part des documents numériques HQ dans lesquels la BnF exige un taux de reconnaissance de caractères supérieur à 99,9% et les documents avec une qualité garantie dans lesquels

les taux de reconnaissance des caractères doivent être supérieurs à 98,5%. Les documents en qualité « Brute » ne sont pas contrôlés.

Pour réaliser la procédure de contrôle, une procédure de tirage aléatoire de mots est réalisée des lots qui ont été employés pour le contrôle de segmentation. Le nombre de mots sélectionnés par page s'élève à cinq mots au maximum. Les mots tirés dans la procédure doivent contenir au moins un caractère latin ISO Latin-1. Les mots composés de chiffres arabes, de ponctuations, d'opérateurs mathématiques, etc. sont exclus de l'opération de contrôle. Par contre si les collections numérisées contiennent des langues étrangères ou du latin, la collection des mots échantillonnés doit contenir obligatoirement des mots qui appartiennent à ces langues.

Les opérations de contrôle sont réalisées par un opérateur humain en utilisant un outil de vérification qui assure la superposition entre les images des mots et leurs transcriptions textuelles. A chaque erreur de reconnaissance l'opérateur procède à la saisie du mot qu'il voit puis de l'identifiant du mot ainsi et enfin l'identifiant de la page traitée. A chaque opération de saisie l'outil de vérification comptabilise une erreur de reconnaissance des caractères. La comparaison entre le mot saisi et le mot inscrit dans le fichier ALTO se fait caractère par caractère et obéit à un algorithme qui tient compte des ponctuations, des accents et des chiffres arabes.

3.4 Conclusion

Le processus de contrôle des documents réalisé à la BnF que nous venons de décrire nous amène à constater quelques insuffisances. D'une part, nous remarquons l'absence de toute procédure de contrôle de qualité de la transcription automatique des caractères. En effet, le contrôle se fonde sur les taux de reconnaissance estimés par l'OCR. Il est donc surestimé car il ne peut comptabiliser les zones de texte omises. D'autre part, l'opération de contrôle des documents est réalisée par des prestataires de numérisation externes sur des échantillons de documents qui regroupent quelques pages d'un document et quelques mots dans une page, ce qui peut être insuffisant pour détecter toutes les erreurs susceptibles de se produire lors de l'application de l'OCR. Enfin, sur les documents patrimoniaux, le comportement de l'OCR n'est pas toujours stable. En effet, il arrive que la cause des erreurs de reconnaissance des caractères et de segmentation ne soit difficile à identifier.

L'opération de contrôle de qualité ne détecte pas non plus certains types de défauts dans les résultats de l'OCR. Par exemple, les erreurs d'omission des éléments de la page ne sont pas contrôlées. Or de telles erreurs d'omission sont connues : pour arriver à une certaine qualité de reconnaissance, les prestataires ont tendance à rejeter les blocs de texte présentant trop d'incertitude, c'est à dire trop de mots reconnus avec une confiance trop faible.

Les taux de reconnaissance de caractères annoncés dans les documents de Gallica sont mesurés en utilisant des taux de confiance calculés automatiquement par les OCR. Du fait du mode de fonctionnement des OCR en boîte noire, la méthode de mesure de ces taux de confiance n'est pas connue par les utilisateurs. De plus, les contrôleurs de la BnF ont constaté que généralement ces taux sont surestimés. Pour pouvoir corriger les estimations des taux de reconnaissance des caractères, il serait nécessaire de parcourir manuellement tous les documents fournis par les prestataires, ce qui est impossible dans le cadre d'un projet de numérisation de masse.

Par conséquent, pour mieux maîtriser la qualité des documents numérique de la BnF, nous pouvons tenter d'améliorer l'opération de contrôle des documents en proposant des procédures automatiques de contrôle de la qualité de la segmentation des éléments de la page d'une part et de contrôle des taux de reconnaissances des caractères d'autre part. C'est précisément ce que nous avons étudié au cours de cette thèse.

Chapitre 2

Etat de l'art : système de reconnaissance de caractères

1 Introduction

La reconnaissance optique de caractères est le procédé qui permet de transcrire les images de textes en des textes électroniques codés selon une norme bien définie et connue de tous les systèmes informatiques (par exemple ASCII, ISO, UTF etc...). Dans le contexte des projets de numérisation de masse, l'opération de numérisation des documents n'a de sens que si l'on peut consulter de manière efficace les fonds numérisés. Pour un utilisateur, le mode de consultation le plus naturel est bien sûr la recherche en mode texte, en formulant sa requête au clavier. Pour offrir cette possibilité il est donc nécessaire de procéder à la transcription des documents numérisés afin de les indexer ensuite en mode texte grâce à un moteur de recherche standard.

Les premiers systèmes de gestion de documents numériques évitaient de mettre en oeuvre une transcription des documents en mode texte et se limitaient à indexer les documents numériques en utilisant les informations bibliographiques enregistrées dans des métadonnées. Cependant, ces dernières ne reflètent pas l'ensemble du contenu et ne permettent donc pas de procéder à des recherches plein texte. En effet, avec le développement des tablettes électroniques et le déploiement des connexions sans fil dans les lieux publics, de nouveaux besoins de lecture et de consultation des documents numériques ont émergés. Aujourd'hui, les lecteurs ne se contentent pas de rechercher les articles à travers des noms des journaux et les dates de leurs publications par exemple, mais ils souhaitent affiner leurs recherches à travers des mots clés qui sont contenus dans l'article. C'est la raison pour laquelle, plusieurs bibliothèques dans le monde proposent aujourd'hui à leurs lecteurs des outils de consultation sophistiqués qui offrent même des modes de consultation fondés sur l'utilisation de certains formats particuliers tels que l'EPUB ou le PDF. Ces nouvelles fonctionnalités ne sont réalisables qu'avec la disponibilité de la transcription textuelle du contenu des images des documents et la structure originale de la page.

Les systèmes de reconnaissance de caractères (**OCR**) sont des logiciels fondés sur des algorithmes permettant de passer d'un signal contenant des informations textuelles (images de caractères) à une forme de texte électronique codé au format ASCII, UTF-8, Unicode ou structurée sous au format XML.

Bien que les systèmes d'OCR actuels soient arrivés à des performances de transcrip-

tion excellentes avec les documents récents édités dans des modes standards, la qualité de la transcription automatique des caractères décroît énormément sur des documents anciens ou patrimoniaux. Cela s'explique par les propriétés particulières de ces documents et par les défauts physiques et typographiques qui peuvent apparaître sur les pages de ces collections anciennes. Cela représente un problème fondamental pour les projets de numérisation patrimoniale de masse puisque la volumétrie des données interdit de contrôler manuellement tous les documents produits par les prestataires d'une part, et que d'autre part il n'existe pas d'outils automatiques appropriés qui pourraient vérifier les résultats de l'OCR.

Bien que l'objectif de notre travail ne concerne pas l'amélioration des résultats de l'OCR, il nous semble important de détailler les opérations de l'OCR appliquées lors de la procédure de conversion textuelle de l'image. Ceci nous permet de déterminer les sources potentielles des défauts de reconnaissance et de segmentation des éléments de la page. Dans le chapitre précédent, nous avons présenté la démarche générale de la procédure de reconnaissance de caractères. Dans ce chapitre, nous allons décrire en détail la procédure de conversion des caractères en présentant les opérations de pré-traitements, de segmentation, de description et de classification des caractères. Nous évoquerons également l'ensemble des techniques utilisées dans la littérature pour évaluer et contrôler les résultats de l'OCR.

2 Chaîne de traitement d'un système d'OCR

Les systèmes d'OCR sont devenus de plus en plus des systèmes experts composés de plusieurs briques algorithmiques permettant d'effectuer des traitements successifs. Chaque brique réalise une tâche bien définie dans le processus de reconnaissance des caractères.

La chaîne de transcription des images commence généralement par une étape de prétraitements qui vise à préparer les images de documents aux algorithmes d'analyse et de reconnaissance de caractères. Ensuite une procédure de segmentation des éléments de la page est employée pour déterminer les éléments graphiques et textuels de la page. Puis, afin de caractériser les formes des caractères par des signatures les plus invariables possibles, le système extrait un certain nombre de caractéristiques sur les images des éléments textuels. Cette signature va être envoyée au moteur de reconnaissance de caractères afin de déterminer l'identité du caractère auquel elle correspond. A la fin de cette procédure, une étape de post traitements est appliquée sur les résultats de reconnaissance de caractères afin de filtrer les mots incorrects. Cette opération est réalisée généralement grâce à l'utilisation d'un dictionnaire.

Pour chaque étape dans le système d'OCR, il existe dans la littérature un nombre important d'implémentations. Nous listerons ci-dessous ces opérations dans l'ordre séquentiel habituel suivant :

- **Pré-traitements**
  - Correction de contraste
  - Débruitage
  - Binarisation d'image
  - Redressement d'images
- **Segmentation d'image**
  - Séparation texte/graphique

  - Analyse structurale et fonctionnelle
  - Extraction des lignes
  - Extraction des caractères
  - Normalisation
- **Reconnaissance des formes de caractères**
  - Extraction des descripteurs des formes
  - Classification
- **Post-traitement**
  - Application d'un dictionnaire ou d'un modèle de langue

Afin de comprendre le mode de fonctionnement des OCR et déterminer les sources potentielles d'erreurs, nous détaillons ces différentes étapes dans les paragraphes qui suivent.

2.1 Pré-traitements

Les images résultantes du processus de numérisation des documents ne sont pas sans défauts ni sans spécificités qui peuvent rendre le mécanisme de reconnaissance de texte délicat voire inopérant. Afin de rendre possible l'extraction et l'identification des éléments textuels, il est souvent nécessaire de recourir à une série de traitements qui sont effectués en amont de la chaîne afin d'améliorer la qualité des images. Nous détaillons dans les sous-parties suivantes les principaux traitements utilisés dans la littérature.

Correction de contraste

Lors du processus d'acquisition et malgré les réglages du scanner réalisés par l'opérateur, il peut arriver que les images produites possèdent un défaut de contraste. Celui-ci peut-être dû à un changement d'éclairage dans la pièce, à une variation de teinte ou à des propriétés réflectives du papier, etc. Or, la plupart des systèmes d'analyse de documents travaillent sur des images binaires. Si l'image du document est mal contrastée, l'algorithme de binarisation échoue à s'adapter à l'image qui lui est soumise pour trouver le bon paramétrage de binarisation.

La correction gamma [PV00] est l'une des techniques les plus réputées dans la procédure d'amélioration du contraste de l'image. Elle consiste à modifier par une fonction de puissance les intensités de luminance des pixels. Plusieurs travaux dans la littérature ont essayé d'estimer la valeur de gamma adéquate pour améliorer le contraste des images [Far01].

D'autres algorithmes traitent l'histogramme des intensités des pixels pour corriger le contraste de l'image. Ces algorithmes essayent d'égaliser l'histogramme de l'image afin d'ajuster les niveaux de gris de l'image. Un état de l'art sur ces transformations est présenté dans [Kau11].

Débruitage

Les scanners et les caméras ont des sources de bruit liées à la détection et à l'amplification du signal. Les conditions de prise de vue ainsi que les défauts physiques des documents sont une source de bruit de numérisation sur les images du document.

La nature de ce bruit peut être définie comme une fluctuation involontaire du nombre de pixels noirs de manière aléatoire sur des zones de pixels blancs. Cela peut engendrer

une déformation dans les formes de caractères et l'apparition de composants connexes dans les zones d'intérêts (paragraphes) ce qui perturbe énormément la reconnaissance de caractères.

Le bruit est donc un attribut important dans le système d'imagerie numérique et dans la procédure d'analyse d'images de document. C'est la raison pour laquelle il est important de l'éliminer avant de procéder à la segmentation et à la reconnaissance des caractères. Dans la littérature, il existe deux familles de méthodes du débruitage :

- Les méthodes qui utilisent un filtrage **spatial** (filtre moyenneur, filtre médian, etc.) ou fréquentiel,
- Les méthodes qui utilisent des opérateurs morphologiques (dilatation et érosion).

Redressement d'inclinaison

Du fait de la procédure de capture par scanner, parfois très rapide, les images des documents peuvent comporter une certaine inclinaison qui engendre par la suite des problèmes dans la procédure de segmentation des caractères. En fait, certains algorithmes de segmentation font l'hypothèse d'une image horizontale. Par conséquent, il est important de redresser l'inclinaison des images des documents avant de procéder à l'opération d'analyse.

Généralement, les systèmes d'OCR utilisent un algorithme de rotation basé sur la détection automatique des bords de pages pour déterminer l'angle d'inclinaison. Ces algorithmes ne sont pas totalement fiables, surtout quand les pages présentent des contours irréguliers, une couleur de fond qui se rapproche de la couleur de l'écriture, ou une mise en forme particulière qui induit des erreurs de détection des bords de page. On peut décomposer les approches proposées dans la littérature en quatre catégories :

- Algorithmes utilisant la transformée de Hough [NGP07], [KS12a],
- Algorithmes procédant par projection horizontale de l'image [Hou83],
- Algorithmes de regroupement des composantes connexes voisines [HYR86], [O'G93], [LT03]
- Algorithmes utilisant la transformée de Fourrier [Pos86]

De plus, si la rotation d'une image est une opération triviale lorsqu'elle est orthogonale (90°, 180° ou 270°), elle l'est moins dans la plupart des cas (cf. figure 2.1), lorsque l'angle d'inclinaison est compris entre 0° et 90°. En effet, la rotation dans ce cas implique le calcul par interpolation des pixels résultant, sur une matrice qui doit être élargie.

Selon [CWM11], si l'on veut conserver l'image la plus fidèle à l'originale, on doit calculer l'intensité des pixels résultat en utilisant l'une des méthodes d'interpolation suivantes :

- **Plus proche voisin :** chaque pixel résultat prend la valeur du pixel d'origine le plus proche géométriquement.
- **Interpolation linéaire :** chaque pixel résultat prend la valeur moyenne des deux pixels les plus proches
- **Interpolation bilinéaire :** chaque pixel résultat prend la moyenne pondérée des quatre pixels les plus proches (les deux plus proches pixels comptent plus que les deux pixels les plus éloignés)
- **Interpolation bicubique :** une approximation polynomiale est utilisée pour déterminer la valeur du pixel résultat en prenant en compte les seize pixels les plus proches.

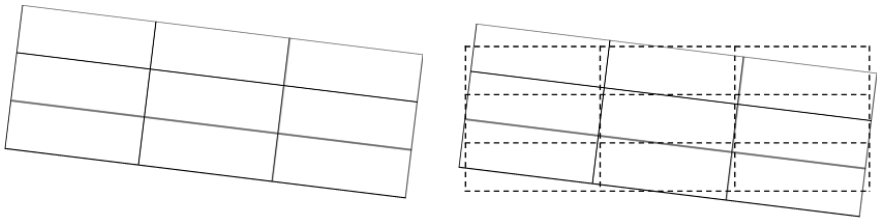


FIGURE 2.1 – Opération de redressement de l'image : l'image à gauche représente l'image avant l'opération de redressement et l'image à droite représente le résultat de l'opération de redressement.

Binarisation

La plupart des systèmes d'OCR procèdent à une binarisation car la plupart des méthodes de segmentation de la littérature travaillent généralement sur des images binaires. La binarisation est la procédure qui permet de transformer une image couleur ou en niveau de gris, issue d'un scanner, en une image binaire. Dans l'image binaire le fond de la page est généralement représenté par des pixels blancs alors que les caractères et les graphiques sont représentés en noir. On peut considérer la binarisation comme une opération de classification qui affecte chaque pixel de l'image à l'une des deux classes, avant plan, fond du document.

La difficulté essentielle de l'opération de binarisation réside dans la sélection d'un seuil optimal, qui réalise le meilleur compromis entre une perte d'information et l'introduction de bruit. Généralement, le choix de ce seuil dépend de plusieurs facteurs comme les conditions d'éclairage, l'état du papier, la noirceur de l'encre. Dans les documents patrimoniaux, ces paramètres peuvent varier à plusieurs niveaux d'un document à l'autre, d'une page à l'autre d'un même document, au sein d'une même page, si la couleur ou l'état du papier change.

L'application d'un seuil constant sur l'ensemble des images fournit donc en général des résultats médiocres non exploitables par les autres étapes du système d'OCR. Pour pallier à ce problème, plusieurs méthodes dans la littérature ont été proposées pour sélectionner la valeur du seuil optimal pour l'image considérée. On peut regrouper ces méthodes en deux catégories :

- Les méthodes de binarisation globale qui utilisent un seul unique seuil pour binariser [Ots79], [Ng06],
- Les méthodes de binarisation locale qui utilisent des seuils multiples pour binariser l'image [TW03].

2.2 Analyse d'images de documents

Une fois toutes les opérations de préparation des images réalisées, on procède à l'identification des éléments de la page. Dans l'image d'un document, l'information pertinente est présente dans deux types de zones. Des zones textuelles qui regroupent l'ensemble des paragraphes, des lignes, des mots et des caractères de la page, et des zones graphiques comme les illustrations, les séparateurs graphiques entre les paragraphes, les lettrines et les figures. La transcription automatique du contenu des documents nécessite de localiser les zones textuelles. Qui plus est, il peut être intéressant d'analyser certains éléments

graphiques (comme les séparateurs, les structures tabulaires etc.) car ils organisent les éléments textuels et permettent d'en donner une signification particulière (i.e. Table des matières, Titre, Légende, etc.).

On trouve dans la littérature différentes méthodes qui s'intéressent à la localisation des textes dans les images. Cette tâche élémentaire constitue l'un des maillons essentiel en analyse d'images de documents sur lequel sont bâties la plupart des méthodes d'analyse de la structure physique. Mais il existe par ailleurs un grand nombre de méthodes qui ont été proposées pour détecter des informations textuelles dans des images de scènes naturelles.

Dans les paragraphes qui suivent nous présentons ces différentes méthodes en gardant à l'esprit que l'objectif ultime de nos travaux n'est pas de procéder à l'extraction de la structure physique des documents mais bien de détecter les éléments textuels qui ont été omis par l'OCR. Notre présentation débute par un rappel des principales méthodes utilisées pour l'analyse de la structure physique des documents. Puis nous développons les principales approches proposées dans la littérature pour détecter les informations textuelles.

L'analyse de la structure des pages

L'analyse de la structure de la page est l'opération de localisation et d'identification des différentes zones qui composent l'image de la page. La structure d'un document peut être décrite à deux niveaux d'analyse. Le premier niveau d'analyse concerne la structure physique du document. Elle est définie par les caractéristiques typographiques (police de caractère, couleurs, style de caractère, etc.) et la mise en forme de la page (interlignes, alignements, nombre de colonnes, etc.). Selon [Mul06], on peut décomposer les travaux de la littérature en trois catégories :

1. Les approches de segmentation descendantes traitent progressivement les éléments de la page en passant du niveau page au niveau caractères. Ce type d'approches repose sur un découpage récursif de l'image en utilisant un modèle de page. Les approches descendantes sont rapides dans la procédure de segmentation des éléments de l'image de la page puisqu'elles reposent sur des heuristiques et des paramètres de segmentation prédéfinis. Par contre, ces méthodes présentent des limites dans le traitement des documents hétérogènes. Les algorithmes X-Y cut de [NKK+88] et [HHP95] sont des approches de segmentation descendante.
2. Les approches de segmentation ascendantes regroupent récursivement des composants élémentaires de la page (pixels, caractères, composantes connexes) pour passer du niveau caractères au niveau page. Ce type d'approche repose sur des méthodes de fusion de blocs de proche en proche. L'inconvénient principal de ce type d'approche réside dans le temps de traitement considérable nécessaire pour traiter un grand nombre d'objets élémentaires pour reconstruire la structure physique de la page. Par contre, ces méthodes sont applicables sur une grande variété d'images avec des règles de fusion élémentaires. L'algorithme RLSA (Run Length Smearing Algorithm) [WW82] ainsi que d'autres méthodes basées sur le diagramme de Voronoï [KSI98] appartiennent aux approches de segmentation ascendantes.
3. Les approches de segmentation mixtes utilisent des algorithmes de fusion et de découpage en même temps pour tirer avantages des approches ascendantes et descendantes. En fait, pour accélérer le processus de segmentation de la structure physique de la

page, les approches mixtes sélectionnent les objets élémentaires avant de les utiliser dans la procédure de construction de la structure physique de la page. De plus, les approches mixtes n'utilisent pas de règles de division strictes qui s'appliquent uniquement à une catégorie spécifique de documents. Cela rend ces approches moins contraintes que les approches purement descendantes.

Le deuxième niveau d'analyse concerne l'identification de la structure logique des éléments de la page obtenus lors de l'analyse de la structure physique de la page. Ce niveau d'analyse exige l'identification des fonctions de chaque élément de la page en se basant sur des conventions typographiques (comme la taille des éléments de la page et l'alignement des éléments textuels pour définir les titres et les sous titres, etc.) et sur l'interprétation visuelle des éléments de la page (comme le type de texture des éléments et la taille des éléments qui sert à identifier les éléments textuels et les éléments graphiques, etc.). Ensuite, on analyse les relations qui existent entre les différents éléments de la structure physique de la page (comme l'ordre de lecture, les liens entre les titres des articles et les articles, etc.) pour déterminer leur structure logique et leurs niveaux hiérarchiques.

A partir de ces deux niveaux d'analyse, il existe deux stratégies pour identifier la structure de la page. La première stratégie procède séquentiellement en commençant par la détermination de la structure physique du document puis en procédant à l'identification de la structure séquentielle et fonctionnelle de l'image. La deuxième stratégie d'approche analyse simultanément la structure physique et logique de la page en utilisant des algorithmes qui permettent à la fois de localiser les éléments de la page et d'identifier leur nature.

Quelle que soit la stratégie de segmentation, l'efficacité des algorithmes dépend toujours de la régularité des caractéristiques des éléments textuels (comme la largeur, la hauteur, les alignements, etc.). En pratique, la régularité des éléments textuels de la page n'est pas toujours respectée. En effet, certaines images de documents possèdent par exemple des défauts de courbure ou d'impression qui perturbent l'alignement des lignes de textes par exemple.

D'autre part, certains documents difficiles possèdent des espaces variables entre les caractères d'un même mot, ce qui rend l'opération de segmentation des caractères compliquée. Certaines erreurs de fusion ou de fragmentation des mots sont également causées par la présence de certains couples de caractères liés dans les documents médiévaux ou par des défauts d'impression qui fusionnent certains caractères. Pour résoudre ce problème certains systèmes de reconnaissance ont adopté une approche de segmentation basée sur une combinaison entre la segmentation et la reconnaissance, ce qui permet de combiner plusieurs hypothèses de segmentation pour choisir la segmentation la plus pertinente.

Localisation des éléments textuels

Quel que soit l'objectif applicatif visé, analyse d'images de documents ou analyse d'images de scènes naturelles, les principes mis en œuvre pour identifier et localiser les éléments textuels dans les images sont assez semblables. Selon [JKJ04], nous pouvons décomposer les méthodes de localisation d'éléments textuels en deux classes :

- Les méthodes basées sur l'utilisation de forts a priori.
- Les méthodes basées sur les caractéristiques de texture.

La première catégorie d'approches exploite les propriétés essentielles des documents pour faciliter et orienter la détection des zones de textes. La seconde catégorie d'approches ne fait pas les mêmes hypothèses restrictives et permet de proposer des méthodes plus génériques. Nous détaillerons dans les paragraphes suivants les caractéristiques des méthodes de chaque classe.

(a) Approches utilisant de forts a priori

Comme les méthodes de segmentation de documents, les approches de localisation suivent généralement une stratégie ascendante ou descendante. Les stratégies ascendantes ou guidées par les données commencent par la détection des éléments textuels élémentaires comme les composants connexes ou les contours. Puis, elles regroupent ces éléments en utilisant des règles de fusion pour construire hiérarchiquement des mots, des lignes et des paragraphes. L'état de l'art de [JKJ04] décompose les approches ascendantes en deux sous-classes d'approches :

– Les approches qui reposent sur les traitements en composants connexes
– Les approches qui se basent sur les traitements des contours

Les approches à base de composants connexes utilisent des algorithmes de regroupement récursif des composantes élémentaires jusqu'à la détection de la totalité des éléments de la page. Des analyses géométriques sont appliquées par la suite pour labéliser les éléments de la page, fusionner les éléments de même nature, filtrer le bruit de détection et définir les frontières des éléments textuels.

Zhong et al, ont présenté aussi dans [ZKJ95] une méthode de localisation des textes qui se base sur une analyse en composants connexes. La méthode proposée utilise des hypothèses sur les couleurs des éléments de l'image ainsi que leur localisation pour détecter les zones textuelles potentielles dans l'image. Ensuite, les auteurs emploient des heuristiques sur la surface, le diamètre, l'alignement du texte et sur les espaces entre les éléments textuels pour filtrer le bruit de détection. La méthode proposée par Kim dans [Kim96] adopte une démarche semblable à celle du Zhong puisqu'elle regroupe les pixels de l'image en utilisant leurs couleurs pour séparer les composants non textuels du reste de l'image. Les lignes textuelles horizontales ainsi que les segments textuels sont extraits à travers une analyse itérative basée sur des projections horizontals. Une phase de post-traitement est effectuée pour fusionner les éléments textuels en utilisant des heuristiques. A cause de l'utilisation de plusieurs seuils de décision cette méthode n'est pas fiable dans un contexte générique où il faut localiser des textes dans différentes polices et à différentes tailles. Les expérimentations de cette méthode réalisées sur une base composée de 50 images ont montré que le taux de localisation est de 87%.

La chaîne des traitements proposée par Messelodi et Modena[MM99] regroupe aussi trois étages de traitements séquentiels qui assurent : **(i)** l'extraction des objets élémentaires, **(ii)** le filtrage des objets et **(iii)** la sélection des éléments textuels des lignes. Des étapes de pré-traitements comme la réduction de bruit, l'amélioration du contraste et la requantification de l'image sont réalisés pour améliorer la qualité de l'image. Puis, une analyse en composantes connexes est réalisée pour localiser les éléments textuels de la page. Différents filtres, basés sur les caractéristiques locales des composantes connexes comme leurs tailles, leurs surfaces, leurs rapports hauteurs / largeurs, etc. sont employés pour filtrer les éléments non textuels de l'image. D'après les auteurs, cette méthode est très dépendante des valeurs des seuils définies. L'algorithme de localisation du texte commence

par le traitement récursif des régions de l'image jusqu'à la satisfaction d'un certain nombre de critères comme l'adjacence des éléments textuels et l'alignement des éléments textuels définis par l'utilisateur. Cette approche est applicable sur différents types de documents qui peuvent contenir différents types de polices et différents styles de texte. Cette approche est capable aussi de sélectionner des lignes textuelles définies avec plusieurs inclinaisons dans la même page. L'évaluation de cette approche est réalisée sur une base d'images composée de 100 livres et conduit à un taux de localisation des composants textuels de 91,2%.

Les approches qui se basent sur les contours exploitent le fort contraste entre les éléments textuels et le fond de la page pour localiser le texte dans l'image. Généralement dans ce genre d'approches, les contours des éléments textuels sont identifiés en premier lieu en utilisant un filtre (par exemple l'opérateur de Canny). Les éléments détectés peuvent ensuite être fusionnés en utilisant des règles qui se basent sur des caractéristiques de proximité et sur des opérateurs morphologiques pour regrouper ensemble les éléments de la page. La dernière étape de traitement applique des heuristiques pour filtrer le bruit résiduel.

Smith et Kanade ont appliqué dans [SK95] un filtre différentiel horizontal de taille 3×3 pour déterminer les contours horizontaux dans l'image. Une opération de filtrage est appliquée pour éliminer le bruit (les petits contours) et fusionner les contours adjacents. Les éléments issus de cette étape de traitement sont ensuite renvoyés à la procédure de création des boîtes englobantes. Une deuxième opération de filtrage basée sur des heuristiques sur les caractéristiques internes des boîtes englobantes (la taille, le rapport entre la hauteur et la largeur et le facteur de remplissage) est employée pour filtrer les éléments non-textuels. Finalement, les histogrammes d'intensité de chaque élément issu de la deuxième opération de filtrage sont examinés pour former des groupes de boîtes qui ont des caractéristiques de texture et de forme similaires.

Les stratégies de localisation descendantes ou approches guidées par un modèle procèdent à l'inverse des stratégies ascendantes en partant de l'image dans sa globalité pour arriver à des représentations de plus en plus locales du contenu de la page. Par conséquent, ces approches commencent par l'analyse des grandes composantes de l'image (comme les paragraphes) pour localiser les éléments élémentaires (comme les caractères) de la page. Les approches appartenant à cette catégorie emploient des règles de découpage prédéfinies pour décomposer les zones de l'image de la page.

L'approche la plus connue qui appartient à cette catégorie est l'algorithme XY-Cut introduit par Georges Nagy dans [Nag86] pour segmenter les différents composants de l'image de la page. Son principe se base sur une opération de découpage récursif en utilisant deux projections horizontales et verticales d'éléments textuels pour passer de niveau paragraphes au niveau caractères. Le découpage XY est bien adapté à des documents qui ont une structure physique fixe tels que les journaux, les formulaires, les ouvrages, etc. Cependant, cette approche échouera sur des documents caractérisés par des structures physiques variables.

(b) Approche texture

Dans la section précédente, nous avons présenté des approches de segmentation et de détection d'éléments textuels qui se basent sur des connaissances a priori pour réaliser leurs tâches. Dans cette partie, nous allons présenter une famille de descripteurs génériques qui

se base sur les caractéristiques de texture de l'image pour segmenter l'image.

Les approches textures permettent d'extraire des éléments textuels sans mobiliser de connaissances *a priori*, sur les documents étudiés. L'état de l'art proposé par Tuceryan [TJ98] décompose les descripteurs de texture en quatre familles. On distingue parmi elles les méthodes statistiques, les méthodes géométriques, les méthodes à base de modèles probabilistes et les méthodes fréquentielles. Dans les sous-sections, nous présentons quelques méthodes de ces familles.

i. Descripteurs statistiques

La méthode des matrices de co-occurrence des niveaux de gris (GLMC) proposée par Haralick dans [HSD73] représente l'une des méthodes les plus connues parmi les méthodes statistiques. La GLMC est une matrice qui indique le nombre d'apparition des couples de pixels ayant un niveau $I(i, j)$ selon une direction et un déplacement donnés ($d = (d_x, d_y)$). Les valeurs de la matrice de co-occurrence permettent de caractériser la régularité, la répétitivité et le contraste des textures.

Une autre méthode de caractérisation de texture développée par Laws [Wie80] se base sur l'utilisation de 25 convolutions spatiales prédéterminées pour construire 25 versions de l'image traitée. Ces filtres de convolution sont calculés à partir de 5 masques simples suivants :

$$L5 = [14641] \quad E5 = [-1-2021] \quad S5 = [-1020-1] \quad W5 = [120-21] \quad R5 = [1-46-41] \tag{2.1}$$

En multipliant ces filtres entre eux, nous obtenons les 25 filtres bi-dimensionnels. L'extraction des caractéristiques de textures se réalise en effectuant les quatre étapes suivantes :

1. Application des 25 filtres sur l'image pour décrire les pixels avec un vecteur de caractéristiques composé de 25 valeurs. F_k (m,n) avec k=1....25 et m, n sont les coordonnées des pixels de l'image.
2. Pour chaque résultat de filtrage, on calcule une énergie de texture sur une région de taille de 15×15 : $E_k(m, n) = \sum_{j=n-7}^{n+7} \sum_{j=m-7}^{m+7} |F_k(i, j)|$
3. Normalisation des valeurs des matrices de fréquence par le produit des deux masques
4. Enfin, on procède au calcul des caractéristiques de la texture de l'image comme par exemple le rapport de l'énergie des segments horizontaux et verticaux.

L'approche proposée dans [Ros99] utilise la matrice de longueur de plage pour segmenter les éléments textuels de l'image. En effet, cette matrice des informations semblables à celui de GLMC qui permet de rechercher des successions (plages) de pixels selon un niveau de gris et un angle précis. Des propriétés de texture comme la taille des plages, leurs fréquences et leurs répartitions sont déterminées grâce à l'utilisation de cette méthode.

ii. Descripteurs géométriques

D'autres descripteurs appartiennent à la classe des approches géométriques qui caractérisent les textures de l'image à travers la description des formes élémentaires et des relations géométriques qui relient ces différents composants pour former la texture. Plusieurs méthodes géométriques dans la littérature ont été proposées. L'état de l'art présenté dans [Egl08] présente quelques méthodes. Dans [Tuc94], les auteurs ont développé une approche de segmentation utilisant les moments géométriques. Cette méthode est appliquée sur des images de documents afin d'extraire les différents composants de la page. Dans le

même cas d'utilisation, Khedekar et al. [KRSG03] ont développé une méthode de séparation texte/graphique qui se base sur la construction des histogrammes horizontaux pour segmenter les documents hébreux.

Dans [Che96], l'auteur analyse les blocs de l'image prédécoupés dans le but de les classer soit en tant que texte, soit en tant que graphique. Les critères analysés sont extraits à travers la projection des pixels selon différents angles. Journet et al. proposent dans [Egl08] une approche multi-résolutions de segmentation qui repose sur la caractérisation de la direction principale de la texture. En effet, contrairement aux régions graphiques, à différentes résolutions les textures des régions textuelles possèdent une seule direction principale. Pour déterminer les directions principales de la texture, les auteurs ont construit des histogrammes des directions (rose des vents) qui utilisent qui utilisent les réponses de la fonction d'autocorrélation (cf. figure 2.2).

L'approche proposée extrait trois caractéristiques de directions qui sont la direction principale, la variance des directions et l'amplitude la fonction d'autocorrélation selon la direction principale. Ces caractéristiques sont calculées sur trois fenêtres glissantes (de tailles 128×128, 64×64 et 32×32) qui parcourt l'intégralité de l'image. Cette approche est très couteuse en temps de traitement avec des images à haute résolution. Par contre, elle est très fiable puisque 83% des éléments graphiques et 92% des éléments textuels sont correctement détectés.

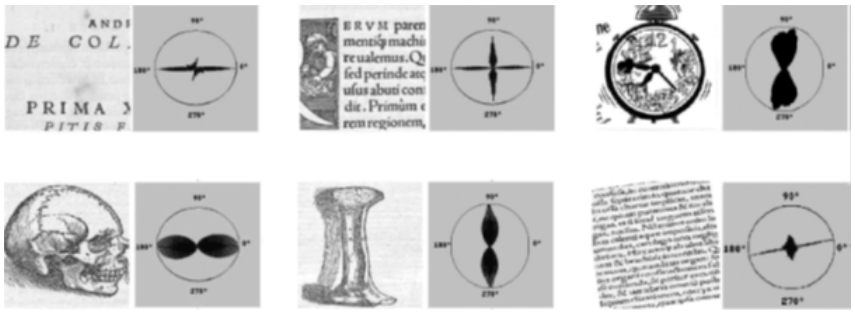


FIGURE 2.2 – Exemple d'histogramme des directions (roses des directions) pour différents types de contenus.

iii. Descripteurs fréquentiels

A une échelle globale, le texte peut être considéré comme de la texture qui a des propriétés différentes de celles du fond de l'image. Les approche basées sur une étude des textures dans le domaine des fréquences comme les filtres de Gabor, les ondelettes, la transformée de Fourrier etc. peuvent être utilisées pour détecter la texture des régions textuelles.

Sin et al. ont utilisé dans [SKC02] des caractéristiques fréquentielles comme le nombre des pixels verticaux et horizontaux des contours des caractères et le spectre de Fourrier pour détecter les régions textuelles dans des images de scènes réelles. En se basant sur l'hypothèse que les régions textuelles sont incluses dans des zones rectangulaires, l'approche proposée commence par la détection de ces zones en utilisant le transformé de Hough.

Mao et al. proposent dans [MCLS02] une approche à base de texture pour localiser les éléments textuels dans les images des scènes naturelles. L'approche proposée se base sur la décomposition en ondelettes pour définir la variation de l'énergie locale de l'image à différentes échelles. L'image binaire, produite à travers une opération de seuillage locale de

variation d'énergie, est analysée en utilisant des filtres sur les composants connexes. Ces filtres se basent sur un certain nombre d'heuristiques comme le rapport largeur / hauteur et la taille des éléments de la page. Toutes les régions textuelles détectées à différentes échelles sont par la suite regroupées ensemble pour produire les images des éléments textuels.

iv. Méthodes à base de modèles statistiques

Malgré le fait que les approches texture ne reposent pas sur des a priori aussi forts que les approches précédentes, ces systèmes restent sensibles aux types de polices des caractères, à leurs tailles et style. En fait, il est difficile de concevoir un filtre capable de caractériser toutes les textures textuelles possibles. Pour surmonter cette difficulté, plusieurs travaux de la littérature ont traité la problématique en intégrant une étape d'apprentissage qui permet de générer automatiquement les filtres de texture adéquate avec les caractéristiques des éléments des images traitées.

Certains travaux [JZ96] [JK96] [JB92] utilisent des méthodes d'apprentissage pour former des filtres ou des classifieurs de texture qui permettent de séparer les éléments textuels des éléments graphiques. Dans l'état de l'art de [TJ98], les auteurs définissent les méthodes de localisation de texture à base de modèles comme étant « *celles qui se basent sur la construction d'un modèle d'image permettant non seulement de caractériser les textures de l'image mais aussi d'en générer* ».

Les travaux réalisés par Jung dans [Jun01] et par Jeong et al. Dans [JJKK99] rentrent dans cette catégorie d'approches. Les auteurs ont appliqué des méthodes d'extraction de texture qui emploient un réseau de neurones pour former un ensemble discriminant de masques de texture. La formation des masques est réalisée avec le critère de minimisation des erreurs de classification des textures de l'image. La méthode distingue deux classes : les régions textuelles et les régions non-textuelles. Le réseau de neurone voit en entrée le pixel courant et son voisinage et pour les trois composantes de couleur (Rouge, Vert et Bleu). Contrairement aux approches précédentes, aucune phase d'extraction de caractéristiques n'est mise en œuvre. C'est le réseau de neurones qui se charge de construire les caractéristiques discriminantes à partir des informations brutes qui lui sont fournies en entrée. En fait, les auteurs combinent les réponses obtenues sur chaque bande de couleur en utilisant un réseau de neurones arbitraire. Finalement dans l'étape de production des résultats, les boîtes englobantes des éléments textuels sont générées en utilisant les projections perspectives de ces éléments par rapport à l'axe horizontal et vertical. L'évaluation de cette méthode sur une base composée de 950 images de taille 320×240, a permis d'obtenir un taux de détection de 92,2% des éléments textuels avec un temps de traitement de 11,3 secondes.

Les champs de Markov et les méthodes fractales entrent également parmi les outils de cette catégorie d'approches qui ont été largement appliquées. D'une part, la dimension des fractales est un outil rigoureux pour mesurer la répétitivité spatiale et à différente dimension d'un motif. Le travail le plus marquant est celui qui a été développé dans [CCMV03] et qui utilise la loi de puissance (loi de Zipf) pour localiser les zones d'intérêt (les zones textuelles) dans une image naturelle.

Dans [NKPH05] les auteurs ont montré l'intérêt d'utiliser des champs de Markov pour segmenter des images de document compliqués comme des manuscrits de Flaubert qui ont la particularité de contenir de nombreuses hachures et ratures, ce qui rend l'utilisation des approches classiques peu performante. L'approche proposée dans ce travail a permis de

classifier les zones d'intérêt de l'image du document en plusieurs classes (ligne de texte, rayures, notes de marge, etc.).

Conclusion

Les méthodes de détection des informations textuelles fondées sur de forts a priori nécessitent un paramétrage et une configuration complexes. Cela rend difficile leur adaptation à d'autres corpus, et peut les rendre inopérantes dans un contexte fortement variable. Dans le contexte de projets de numérisation de masse tels que ceux auxquels la BnF est confrontée, les corpus à traiter sont très variables (presse, ouvrage, publicité, formulaire, etc.). Les espaces entre les différents composants textuels sont très différents, la mise en page de ces documents est également très variable. Il semble donc évident qu'il serait vain de tenter de trouver des règles heuristiques, ou des paramétrages pouvant convenir à la totalité de ces documents. il est donc nécessaire de développer une approche générique pour traiter le problème de détection des mots manqués.

Les approches textures ont un avantage principal qui se situe dans leur plus grande généricité grâce à leur capacité à décrire la variabilité des textures qui peuvent être rencontrées sur des corpus hétérogènes. De plus, le fait que ces descripteurs utilisent des informations de bas niveau permet de s'affranchir d'un bon nombre d'a priori. Qui plus est, les approches textures sont tout à fait compatibles avec des images en niveau de gris, ce qui permet de s'affranchir d'une étape de binarisation. Comme nous le verrons dans la suite de ce document, c'est donc assez naturellement que nous nous sommes orientés vers ce type d'approche pour développer une méthode de détection des zones textuelles omises par les OCR.

2.3 Reconnaissance de caractères

Dans la section précédente, nous avons exposé les différentes approches proposées dans la littérature pour segmenter et analyser les composants des images des pages. Ceci nous a permis de faire l'état des lieux de ces approches et de préciser les limites de chaque technique de segmentation. Cependant, les défauts de reconnaissance des caractères ne sont pas causés exclusivement par des erreurs de segmentation. En effet, les éléments textuels obtenus lors de la procédure d'identification de structure de la page sont ensuite utilisés par des systèmes de reconnaissance de caractères pour assigner à chaque forme la classe du caractère qui lui correspond.

Des erreurs de reconnaissance de caractères peuvent apparaître à ce niveau ce qui engendre des défauts dans les résultats de l'OCR. Par conséquent, même si l'objectif de cette thèse ne concerne pas l'amélioration des résultats de reconnaissance des caractères, il nous semble important d'étudier cette procédure afin de diagnostiquer les sources d'erreur de reconnaissance.

Généralement dans la littérature, trois étages de traitement composent la procédure de reconnaissance de caractères :

1. La première étape est la procédure de préparation des images des caractères qui permet de réduire la variabilité des formes des caractères dans l'image par rapport aux formes des modèles des caractères appris par les classifieurs.

2. La deuxième étape concerne la description des formes des caractères. Pour cela plusieurs familles de descripteurs ont été proposées dans la littérature. L'objectif de toutes ces méthodes est la production d'une signature unique, ou du moins la moins variable possible, pour chaque classe de caractère.
3. La troisième étape concerne les procédures d'apprentissage et de classification des images des caractères. Dans la littérature, plusieurs classifieurs et combinaisons de classifieurs ont été proposés. L'objectif de cette opération est de fournir les meilleurs résultats possibles de transcription automatique des images de caractères.

Nous allons détailler dans les parties suivantes l'ensemble de ces étapes de traitement en présentant à chaque fois l'objectif de chaque opération et les limites de ses résultats. Pour cela, nous suivont l'ordre chronologique des traitements que nous venons de citer.

Normalisation des éléments textuels segmentés

Une fois l'analyse de structure effectuée, les éléments textuels détectés doivent être envoyés au moteur de reconnaissance. Cependant, ceux-ci ont généralement des caractéristiques de taille et d'inclinaison variables qui peuvent engendrer des erreurs de reconnaissance de caractères. Pour remédier à ce problème, les systèmes de reconnaissance de caractères normalisent les tailles des lignes, des mots et des caractères afin de supprimer certaines des différences de **styles** de manière à obtenir des données standardisées. La normalisation peut inclure les opérations suivantes :

Normalisation de la hauteur qui permet d'adapter la hauteur des éléments textuels avec les modèles définis *a priori* par les systèmes de reconnaissance de caractères. Cette opération est capitale pour les systèmes à fenêtre glissante comme les modèles de Markov cachés, car elle détermine quelle portion de caractères sera inclue dans la fenêtre à un instant t.

Normalisation de l'inclinaison des caractères qui concerne à la fois le redressement de la ligne de base des lignes de texte et le redressement vertical des caractères. Dans les deux cas, le principe de redressement de l'inclinaison est basé sur deux opérations. La première consiste à détecter l'angle de l'inclinaison θ par rapport à l'axe vertical ou horizontal de la page. La deuxième concerne la correction de l'inclinaison des caractères en effectuant une transformation inverse en utilisant l'angle α. La littérature propose différentes méthodes de correction d'inclinaison. Parmi ces méthodes, on trouve celles qui emploient la projection horizontale des caractères pour déterminer l'orientation de l'écriture, les méthodes fondées sur la détection des traits verticaux des caractères et les méthodes basées sur l'analyse des contours.

Normalisation de l'épaisseur ou squelettisation. Il s'agit d'un traitement qui permet de transformer l'image d'un caractère où les traits sont plus ou moins épais en fonction de la police et de la graisse, en une image ne comportant que des traits d'épaisseur unité. La représentation simplifiée d'une forme par l'image de son squelette peut faciliter l'opération de classification. Elle permet également d'extraire plus facilement des informations sur la structure de la forme, telles que la présence d'occlusions, de points de jonction. Plusieurs algorithmes de squelettisation ont été développés. On peut classer ces méthodes en quatre catégories :

– Amincissement topologique qui consiste en une opération récursive afin de retirer des points du contour de la forme tout en préservant ses caractéristiques topologiques.

- Extraction de la carte de distance qui associe à chaque pixel de l'objet sa distance au point le plus proche de contour. Les maxima locaux de la carte de distance constituent le squelette de la forme du caractère.
- Simulation du front enflammé est le premier algorithme d'amincissement développé par Blum pour obtenir le squelette [Blu67]. Cet algorithme simule la propagation uniforme et à vitesse constante d'un feu allumé simultanément sur les contours de la forme. Le squelette est alors l'ensemble des points où les fronts allumés se rencontrent.
- Calcul analytique qui assimile le problème de formation de squelette à un problème géométrique. Des outils géométriques tels que le diagramme de Voronoï ou la polygonisation des contours ont été employés.

Une évaluation des algorithmes d'amincissement sur des applications d'OCR a été proposée dans [LS95]. Les différentes opérations de normalisation des caractères doivent être accompagnées de traitements de lissage et de correction d'image pour atténuer les effets de discontinuité qui peuvent se produire suite à une opération de normalisation.

Caractérisation des formes de caractères

Après la normalisation des formes de caractères, les classifieurs utilisent généralement des descripteurs de formes pour classer les images de caractères. Le moyen le plus intuitif pour décrire les formes des caractères est d'utiliser les intensités des pixels de l'image. En effet, certains classifieurs appliquent une méthode de comparaison par appariement pour comparer les images des caractères à reconnaître avec les images de modèles au patron des caractères. Les réseaux de neurones qui ont la capacité à construire des représentations intermédiaires à partir des données fournies en entrées du classifieur sont des systèmes de classification qui se comportent très bien sur des caractéristiques aussi élémentaires que les pixels. Cependant en général, cette caractéristique reste peu informative et c'est la raison pour laquelle d'autres méthodes de description des formes ont été proposées dans la littérature.

Le principe de base utilisé pour décrire des formes consiste à construire un jeu de caractéristiques qui minimise les variances des caractéristiques qui appartiennent à la même classe de caractères et qui maximise les variances interclasses de caractères. En effet, plus la distance entre les classes de caractères est importante, plus les représentations sont séparables. Ce qui offre une meilleure qualité de classification.

La taille des vecteurs de caractéristiques influe sur les performances du classifieur. En effet, plus la dimension de l'espace de caractéristiques est importante, plus le nombre de paramètres du classifieur est grand ce qui augmente la complexité du système et le nombre de données d'apprentissage nécessaire pour entraîner les classifieurs. Par conséquent, la meilleure combinaison de caractéristiques est celle qui est composée d'un nombre réduit de caractéristiques tout en conservant l'information discriminante.

Les caractéristiques choisies pour caractériser les formes de caractères sont très nombreuses. Le choix des méthodes de caractérisation dépend de la nature des formes à caractériser et de l'algorithme de classification utilisé. On peut distinguer trois groupes principaux de caractéristiques :

- Caractéristiques de bas niveaux ou de nature statistiques
- Caractéristiques structurelles

– Transformées et développement en série

Certaines transformations peuvent entraîner des ambiguïtés et des difficultés dans la procédure de reconnaissance. Pour éviter ce genre de problèmes, il est préférable de construire des caractéristiques invariantes par rapport à ces transformations.

(a) Les caractéristiques de bas niveau

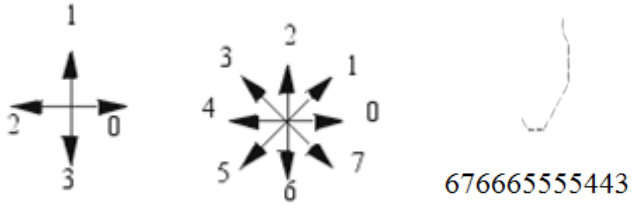
Les caractéristiques de bas niveau sont des mesures élémentaires appliquées sur les images. L'extraction de ces caractéristiques est rapide et peu coûteuse en puissance de calcul. De plus, ces caractéristiques sont généralement tolérantes à la présence de bruit, aux déformations et aux variations de styles. De même, certaines d'entre elles sont invariantes par rapport à la rotation et à la translation. Les caractéristiques de bas niveaux sont généralement calculées dans différentes zones de l'image et les valeurs obtenues sont concaténées par la suite pour produire un vecteur décrivant la forme complète.

L'idée la plus simple consiste à faire une comparaison **template matching** entre l'image à caractériser avec les images des masques de caractères définis au préalable. Dans la littérature plusieurs types de masques ont été proposés [MSY92]. On distingue les masques binaires [BK83] composés par des images bitmaps du même type que l'image de caractères à caractériser, les masques statistiques [Sch00] composés par des poids assignés aux pixels de l'image en fonction de l'importance du pixels dans la composition de la forme du caractère et les masques ternaires [MSY92] définis par une zone d'extension maximale dans laquelle tous les pixels à classifier doivent être inclus et une zone de réduction extrême dans laquelle les pixels du noyau du caractère à classifier doivent être entièrement contenus.

Les descripteurs géométriques sont considérés aussi comme des descripteurs de bas niveau puisqu'ils se basent sur des mesures géométriques comme l'élongation, la compacité, le taux de remplissage. Les différents moments géométriques permettent de dégager un certain nombre d'informations comme la surface avec le moment d'ordre 0, le centre de gravité avec les moments d'ordre 1, l'axe d'inertie avec les moments centrés d'ordre 2, etc. Les moments de Zernike, invariants à la rotation et au changement d'échelle [iDTJT96], sont aussi souvent utilisés par les algorithmes de reconnaissance de caractères pour caractériser les formes [KS02].

D'autres descripteurs bas niveau utilisent les profils des caractères obtenus à partir des projections horizontales et verticales pour décrire les caractères. La projection permet de compter le nombre de pixels qui forme un caractère selon une direction bien déterminée (cf. figure 2.3). Cette caractéristique ne dépend ni de la police de caractère ni du style de l'écriture ce qui donne une description générique des formes. Plusieurs travaux ont utilisé cette caractéristique soit pour reconnaître les caractères soit pour identifier l'écriture manuscrite [CP98] [DR02] [PSWK07].

FIGURE 2.3 – Résultats de l'opération de projection perspective horizontale et verticale



(a) Voisinage à quatre points (b) Voisinage à huit points (c) Exemple de code de Freeman

FIGURE 2.4 – Exemple de Code de Freeman obtenu sur le squelette de caractère arabe , le point de départ est le point supérieur de forme.

(b) Les caractéristiques structurelles

Au contraire des caractéristiques de bas niveau, les caractéristiques structurelles produisent des descriptions sous formes de graphes, d'images ou d'autre valeurs non-scalaires pour caractériser les formes de caractères de la page.

Des primitives extraites sur le squelette ainsi que la description des relations entre ces primitives élémentaires sont généralement utilisées pour décrire la structure des caractères. Les informations qui sont extraites sont par exemple les concavités, les jonctions. Par contre, la squelettisation n'est pas toujours la meilleure solution pour l'extraction de caractéristiques puisque certains caractères ne sont distinguables qu'en prenant en compte les variations d'épaisseur de leurs tracés. Certaines méthodes [SB12] analysent les plages blanches qui entourent les caractères ou qui sont contenues à l'intérieur des formes (par exemple les occlusions). Les concavités, les ouvertures, le sens de l'ouverture ainsi que les occlusions sont des éléments fondamentaux dans la description de la topologie des caractères. De plus, la position ainsi que la taille des surfaces blanches donnent aussi une indication sur l'identité de la forme du caractère.

D'autres méthodes utilisent les contours pour décrire les formes des caractères. Le code de Freeman est parmi les méthodes les plus connues dans cette famille de descripteur. Il décrit l'agencement géométrique des pixels qui composent le contour d'une forme X en utilisant son 4-voisinage ou son 8-voisinage. L'objectif du code de Freeman [Lon98], [ZL04] est de coder le contour d'un objet en utilisant une chaîne de codant qui donne la position relative du point suivant du contour par rapport à la position du point courant. Selon le type de voisinage, la valeur de codant peut varier entre 0 et 4 pour $V_4(i, j)$ ou entre 0 et 8 pour $V_8(i, j)$. Ainsi en utilisant 8 directions, le codant 0 signifie que le pixel suivant sur le contour est situé à droite du pixel courant et le codant 2 signifie que le pixel suivant sur le contour est situé en haut du pixel courant (cf. figure 2.4).

(c) Transformées et développements en séries

Les transformées de l'image sont des techniques utilisées pour produire des caractéristiques invariantes aux déformations et à la rotation des formes des caractères. Ces caractéristiques utilisent les contours des caractères pour décrire leur forme ce qui explique leur sensibilité au bruit qui affecte les contours. L'utilisation des coefficients de la transformée permet de réduire la taille des données nécessaires pour représenter les caractères.

Dans la littérature, plusieurs transformées ont été proposées [], parmi ces transformées on trouve :

- La transformée de Fourier fournit une description qui se base sur le contour extérieur de la forme du caractère. Le descripteur de Fourier est très utilisé dans la littérature [LSJ93], [A.94], puisqu'il fournit des caractéristiques invariantes à la translation, à la rotation et au changement d'échelle.
- La transformée de Hough détermine l'orientation et la courbure d'une forme. Les caractéristiques de Hough sont invariantes par rapport aux bruits et aux distorsions de forme de caractères [CHC89].
- Les descripteurs de Gabor sont utilisés dans plusieurs travaux [LKF05], [WDL05], [HGF01]. Ce descripteur imite les caractéristiques du traitement visuel humain pour caractériser l'image par une description multi-fréquentielle et multi-orientée. Il se base sur l'utilisation de plusieurs filtres fréquentiels pour décrire plusieurs tailles et plusieurs orientations de la forme. Les réponses de ces filtres permettent de construire une carte d'orientations qui va décrire le caractère.
- La transformée en ondelette [Kit05] est un outil de traitement d'image qui permet d'analyser à plusieurs échelles les propriétés locales de l'image. L'utilisation de cette technique a donné lieu à de nombreuses applications dans des domaines très variés tels que le codage vidéo, la compression des images, et la vision par ordinateur. [Nab13] applique des fonctions d'orientation périodique sur les résultats de la transformée en ondelettes pour extraire des caractéristiques liées aux contours des formes des objets de l'image. [Kit05] a montré que l'utilisation de la transformée en ondelettes permet de produire des caractéristiques invariantes par rapport à la translation, la rotation et la transformation affine.
- La transformée en curvelet est une variante de la transformée en ondelettes. D'après [JEBE07], elle est très adaptée pour caractériser les écritures puisqu'elle permet de modéliser la courbure et la direction des formes à différentes échelles. Ce qui la rend invariante aux différentes tailles de caractères. La différenciation entre les différentes classes de caractères est réalisée en utilisant la courbure et la direction principale des formes des caractères.

(d) Reconnaissance de caractères

Une fois que les formes présentes dans les images sont transformées en une représentation dans l'espace de caractéristiques, on procède à l'opération de reconnaissance. Cette opération consiste à affecter au vecteur de caractéristiques du caractère inconnu la classe qui lui correspond parmi un ensemble de classes connues.

Généralement, les classifieurs fournissent une liste de solutions possibles triées par degré de ressemblance. En effet, en acceptant de maintenir une certaine ambiguïté sur les réponses du classifieur, on peut diminuer d'autant le taux d'erreur en utilisant une procédure de reconnaissance basée sur une interaction entre les classifieurs et des modèles de langage.

L'utilisation de connaissances linguistiques pendant l'opération de reconnaissance de caractères permet de valider les résultats de reconnaissance des formes de caractères fournis par le classifieur. Ce mode de fonctionnement est très utilisé dans des applications à vocabulaire limité comme la reconnaissance de contenu de formulaires, le traitement des chèques et le tri des lettres postales. Cependant, avec des applications à vocabulaire large, la complexité de l'opération d'alignement des mots croît avec la taille du lexique ce qui peut engendrer des temps de traitement considérables. Par conséquent, l'utilisation d'une

stratégie de reconnaissance couplée avec des connaissances lexicales doit être généralement être précédée par des approches de réduction du lexique [GK96], [MG93]. Cependant, avec l'accroissement des puissances de calcul, il est de moins en moins couteux de procéder à ces traitements conjoints. C'est notamment le cas en traitement de la parole et en reconnaissance de l'écriture manuscrite, et ces approches pourraient à l'avenir se généraliser au traitement des documents imprimés.

Plusieurs classifieurs ont été proposés dans la littérature pour classer les formes de caractères. On peut décomposer ces approches en trois classes :

- Les approches de reconnaissance structurelle (Analyse syntaxique, Arbre de décision, Chaîne de Markov) modélisent le problème de reconnaissance des caractères à travers des graphes. Le principe de ces méthodes consiste à déterminer des sous-problèmes, à leur trouver une solution, puis à combiner une de ces solutions pour résoudre le problème globalement. Ces approches sont généralement efficaces en termes calculatoires.
- Les approches syntaxiques utilisent des algorithmes issus de la théorie des langages formels pour déterminer la ressemblance entre deux formes de caractères. Un langage formel est un alphabet de symbole structuré dans une grammaire **G** . Les approches syntaxiques reposent sur un principe de base qui considère que deux formes de caractères semblables ont nécessairement une structure commune qui peut être représentée par une grammaire. Un automate associé à la grammaire **G** permet par conséquent de modéliser les différentes formes des caractères.
- Les approches de reconnaissance statistiques (K plus proches voisins, classifieurs Bayésiens, réseaux de neurones, machines à vecteurs supports) utilisent des descriptions par partition de l'espace des caractéristiques pour calculer la probabilité de ressemblance entre deux formes de caractères.

Ces différentes techniques diffèrent d'une part par le principe de modélisation des classes (par les frontières pour les modèles discriminants, par les densités pour les modèles génératifs) et par les critères utilisés pour guider l'apprentissage (taux d'erreur, score de vraisemblance, distance, etc.). Le choix d'un classifieur dépend de la nature des données à classer, de la dimension de l'espace des caractéristiques, du temps de classification, de la taille de la base d'apprentissage et du temps d'apprentissage. Un bon apprentissage doit permettre de réaliser le minimum d'erreurs de reconnaissances.

Une opération de post-traitement basée sur des connaissances de niveau supérieur (linguistique, pragmatique, etc.) est généralement nécessaire pour surmonter les problèmes de reconnaissance de forme. Nous détaillerons dans les parties suivantes les différentes opérations de préparation des données d'apprentissage et de classification ainsi que les méthodes d'apprentissage des formes.

Apprentissage de l'extraction de caractéristiques

Certaines approches de classification incorporent, en plus de l'étape de classification, des étapes de détection de caractéristiques. Les caractéristiques présentées précédemment correspondent à une modélisation humaine de l'information discriminante entre les caractères. Cette modélisation correspond à des formes de caractères bien formés et aisément reconnaissables. Dans la pratique ces conditions ne sont pas toujours vérifiées.

Pour tenter de trouver des caractéristiques plus robustes et exploiter au maximum

l'information discriminante disponible dans les données, des systèmes mettent en place des classifieurs à base de réseaux de neurones. Les travaux les plus représentatifs de ce type d'approche sont les systèmes proposés par Yan Le Cun au cours des années 1990. Ces systèmes emploient plusieurs couches de type TDNN (Times Delay Neural Network) chacune étant spécialisée dans la détection d'une information spécifique provenant de la couche précédente [LJB+95].

Le processus d'apprentissage de ces systèmes commence par l'optimisation des caractéristiques. La première couche cachée sert à l'extraction automatique des caractéristiques à partir de l'image. Puis grâce à la rétro-propagation du gradient de l'erreur sur les couches suivantes du réseau, l'extraction des caractéristiques est optimisée.

L'analyse en composant principale (ACP) est aussi un outil statistique qui permet de transformer les représentations des individus (dans notre cas les formes de caractères) l'espace des caractéristiques obtenu lors de la procédure de description des formes vers un sous-espace de caractéristiques réduit mais conservant les propriétés de séparabilité des classes. Pour obtenir la meilleure représentation du nuage des individus, cette approche choisit dans le nouvel espace de caractéristiques les premières dimensions qui maximisent l'inertie du nuage d'individus. Selon l'intensité de l'inertie du nuage des individus sur un plan factoriel, l'ACP permet de réduire les distances intra-classes et de maximiser les distances inter-classes. Ce qui permet d'améliorer la séparabilité des représentations des classes des individus.

Apprentissage et classification des formes de caractères

L'entraînement d'un classifieur est une procédure obligatoire avant toute opération de classification. La procédure d'apprentissage est réalisée à partir d'une base de données d'images de caractères servant de référence. La mise en place d'un classifieur omnifonte robuste à la variabilité des styles et des polices de caractères nécessite la constitution d'une base d'apprentissage de très grande taille.

La variabilité des formes de caractères ainsi que le nombre d'exemples qui composent la base d'apprentissage sont des paramètres aussi importants que le choix du classifieur. L'expérience a montré en effet que la nature de la base d'apprentissage, sa taille, sa diversité ont un rôle déterminant sur les performances finales du classifieur obtenu. Pour limiter le biais introduit par le choix d'une base d'apprentissage spécifique, des procédures de validation croisée ont été proposées pour valider les paramètres des systèmes de reconnaissance. Ces procédures nécessitent la décomposition de la base de caractères en différentes sous bases de test et d'apprentissage. Dans la littérature il y a au moins deux façons pour mettre en oeuvre une procédure de validation croisée :

- La sélection aléatoire des sous-ensembles d'apprentissage et de test permet de sélectionner aléatoirement par exemple 60% des caractères de la base de validation pour former la base d'apprentissage et 40% des caractères pour composer la base de validation. Puis les performances sont moyennées sur l'ensemble des expériences réalisées.
- K-répétition de l'opération de validation permet de diviser la base en k échantillons puis de sélectionner un des k ensembles pour tester le classifieur et les $k-1$ sous-ensembles restants pour l'apprentissage. Cette opération est répétée k fois et les performances finales sont les moyennes des différentes performances calculées à chaque

itération.

La procédure de validation croisée est généralement utilisée avec des approches de sélection de modèles (comme par exemple « gridsearch », « algorithme génétique ») qui permettent d'optimiser le choix des hyper paramètres des algorithmes d'apprentissage. Différents critères comme l'erreur quadratique moyenne, les taux de précision et de rappel sont utilisées par les approches de sélection de modèles pour optimiser les hyperparamètres.

Les approches classiques utilisent généralement un seul classifieur pour classer les formes des caractères. Cette stratégie de reconnaissance ne peut pas garantir des résultats parfaits malgré les taux de reconnaissance excellents obtenus par certains systèmes. Pour améliorer les performances de reconnaissance, il est souvent utile de combiner les résultats de plusieurs systèmes de reconnaissance de caractères.

Les travaux les plus marquants sont ceux développés par [BN00] et [Gun04]. L'objectif de cette stratégie est de tirer parti des avantages de chaque système de reconnaissance afin d'améliorer les résultats de reconnaissance. Une étude menée par S.V. Rice dans [vRjktaN94] a montré que la combinaison de plusieurs OCR ayant des taux de reconnaissance individuels de l'ordre de 97% a permis d'éliminer 50% des erreurs de reconnaissance.

Abby Fine Reader emploi dans sa version 10 une combinaison de plusieurs classifieurs pour déterminer les identités des caractères de l'image. Ces classifieurs se basent sur des caractéristiques des contours, des caractéristiques structurelles, des templates de caractères ainsi que des caractéristiques de trait distinctif pour classer les formes des caractères. Les résultats obtenus après l'utilisation de cette combinaison sont nettement meilleurs que les résultats obtenus avec la version 9 de cet OCR qui n'utilise qu'un seul classifieur.

2.4 Post-traitements

Les opérations de post-traitement interviennent quand le processus de reconnaissance des caractères aboutit à la génération de la transcription textuelle. Certains systèmes de reconnaissance proposent une liste d'hypothèses pour chaque forme de caractère, cette liste est généralement triée par ordre décroissant selon leurs scores de vraisemblance.

Le but principal de cette opération est d'améliorer les résultats des systèmes de reconnaissance de caractères en appliquant des corrections orthographiques et morphologiques sur les mots reconnus par le classifieur.

Certaines erreurs de reconnaissance sont causées par des déformations des formes ou par des défauts d'impression. Ces erreurs sont inévitables. Par conséquent, pour corriger ces erreurs, on doit combiner les informations visuelles issues de la forme des caractères avec des informations pragmatiques et linguistiques qui permettent d'améliorer les résultats des classifieurs n'exploitant que la forme (parfois dégradée) des caractères.

Les connaissances pragmatiques concernent les informations liées à la nature des documents traités et permettent de privilégier les lexiques, les modèles de langue, ou bien encore de tenir compte de la localisation connue a priori de certaines informations dans les documents, telles que des dates ou des adresses dans les en-têtes des documents. Alors que les connaissances linguistiques font appel à des contraintes linguistiques et grammaticales liées à l'agencement des caractères pour une langue donnée et à la formation des phrases. Nous détaillons dans les paragraphes suivants les caractéristiques de chaque catégorie de connaissance.

Connaissances pragmatiques

Certaines règles pragmatiques liées au contexte des documents à reconnaître peuvent être utilisées pour améliorer les résultats de reconnaissance.

Par exemple, dans un système de traitement automatique de formulaire les emplacements des données à reconnaître ainsi que le nombre des caractères possibles de chaque case est connu. De plus, les réponses possibles aux questions des formulaires sont comprises dans un champ lexical réduit. On peut utiliser les mots de ce champ lexical pour les confronter avec les résultats de la reconnaissance de caractères.

D'autres connaissances pragmatiques liées à l'agencement des caractères dans les phrases permettent de lever l'ambiguïté sur la forme des caractères. Par exemple, certains caractères majuscules ont la même forme que les caractères minuscules (comme x et X, o et O, v et V), la position des caractères dans la phrase permet de déterminer si le caractère est en majuscule ou en minuscule. Cette règle permet de lever l'ambiguïté des formes de ces caractères et de corriger les confusions entre les caractères majuscules et minuscules.

D'autres caractères ont les mêmes formes que les chiffres (comme 0 et O, l et 1). Cette ambiguïté conduit souvent à des erreurs dans les résultats de reconnaissance. Comme les séquences alphanumériques n'apparaissent que dans des champs de formulaires spécifiques, on peut très souvent lever cette ambiguïté et corriger les confusions entre les chiffres et les caractères.

Connaissances linguistiques

Les connaissances linguistiques représentent une source d'information intéressante pour corriger les erreurs commises par les classifieurs. La nature des informations linguistiques diffère selon le type d'application visée. Certaines applications de traitement de courrier et de chèque utilisent un vocabulaire limité. L'utilisation d'un dictionnaire de mots limités qui englobe par exemple, les montants littéraux, les codes postaux et les noms des communes permet de corriger les erreurs de reconnaissance des caractères.

D'autres applications, comme les systèmes de reconnaissance de cartes géographiques, utilisent un vocabulaire étendu (par exemple les noms de toutes les rues d'un pays) qui peut être réduit dynamiquement au cours de l'exécution de l'algorithme pour corriger les erreurs de reconnaissance de caractères.

Les OCR commerciaux des documents de presses ou monographies utilisent des dictionnaires de langue très étendus pour corriger les erreurs de reconnaissance. Ces dictionnaires regroupent généralement les mots singuliers et pluriels courants ainsi que la conjugaison des verbes. Ce ne sont donc pas des dictionnaires à proprement parler mais plutôt des listes de mots possibles : des lexiques.

L'utilisation des connaissances linguistiques comme une étape de post-traitement permet de valider *a posteriori* les séquences de mots reconnues sans aucune interaction avec les classifieurs. L'opération de correction des mots en utilisant des lexiques se traduit par le remplacement de tous les mots qui n'apparaissent pas dans les dictionnaires par les mots les plus proches du dictionnaire selon une distance d'édition.

On voit dans ce cas apparaître les limites de l'utilisation de lexiques. Les inconvénients principaux de l'utilisation de lexiques résident dans la nécessité d'avoir des lexiques exhaustifs pour les différentes catégories des documents traités. En effet, les abréviations

techniques, les termes scientifiques, l'ensemble des entités nommées ne sont pas prédictibles et enregistrés dans les dictionnaires standards. L'utilisation abusive de lexiques inadaptés sur des documents soumis à l'OCR peut introduire des erreurs sur les mots reconnus. Le traitement des mots hors lexiques nécessite donc des traitements spécifiques, et constitue une problématique à part entière.

D'autres analyses linguistiques utilisent les formulations grammaticales d'une langue pour corriger les défauts de reconnaissance de caractères. La première application de ces approches a été réalisée dans le domaine de l'édition afin de signaler les erreurs grammaticales aux éditeurs avant l'impression des documents.

Certains systèmes de reconnaissance commerciaux ont imité la même approche pour décomposer en premier lieu les éléments des phrases reconnus selon leurs catégories grammaticales (sujet, verbe, nom, adjectif, etc.). Puis ensuite vérifier les résultats de reconnaissance de caractères en suivant un certain nombre de règles grammaticales telles que les règles d'accord et de conjugaison. Par contre, l'inconvénient majeur de ces systèmes réside dans le temps de traitement important nécessaire pour vérifier les résultats de reconnaissance des caractères.

2.5 Conclusion

Nous avons présenté dans ce chapitre les différentes étapes employées par un classifieur de forme de caractères pour identifier les classes des caractères de la page. Bien que les techniques employées dans chaque étage de traitement aient déjà atteint depuis quelques années un certain niveau de maturité et de performance, la qualité des résultats de ces approches reste étroitement liée aux contextes de leurs domaines d'application. En effet, dans les projets de numérisation de masse, les prestataires de numérisation utilisent des systèmes d'OCR standards qui ne peuvent pas être toujours spécifiquement adaptés aux propriétés des documents patrimoniaux analysés.

De plus, l'agencement séquentiel des différents étages de traitement rend l'identification de la source des erreurs de reconnaissance difficile. En effet, les défauts de normalisation des formes des caractères ont un impact direct sur les résultats des descripteurs des formes. Ce qui cause des erreurs de classification des caractères.

D'autre part, des erreurs de reconnaissance des mots peuvent être introduites si le dictionnaire employé dans la phase de post-traitements n'est pas adapté à la langue du document traité. Par conséquent, une procédure de vérification des résultats de reconnaissance est obligatoire pour garantir l'intégrité des documents numériques. Dans le chapitre suivant, nous passons en revue les différentes procédures d'évaluation et de vérification des résultats de segmentation et de reconnaissance.

3 Contrôle et évaluation des résultats de reconnaissance

Les contraintes liées à la production de documents numériques telles que la volumétrie, la cadence de numérisation et surtout la haute qualité des résultats de reconnaissance exigée en bout de chaîne de numérisation, rendent nécessaire l'utilisation d'une méthode d'évaluation des résultats de reconnaissance. L'une des principales difficultés dans les projets de numérisation de masse est de savoir évaluer les résultats de reconnaissance de caractères puisque par définition on ne dispose pas de la vérité terrain.

Les résultats des systèmes de reconnaissance de caractères sont utilisés par différentes applications. Les moteurs de recherche utilisent la transcription textuelle du document pour l'indexer. D'autres applications apparues récemment utilisent les résultats de l'OCR comme un format pivot à partir duquel sont produits plusieurs types de documents électroniques (dans des formats spécifiques : EPUB, PDF, etc.) qui sont ensuite accessibles par différentes plateformes électroniques de consultation (Smartphone, Tablette, E-Book etc.). Ces nouveaux modes d'utilisation des documents numériques sont de plus en plus exigeants en termes de qualité des transcriptions. La Bibliothèque nationale de France, par exemple, exige un taux de reconnaissance supérieur à 98,5% pour tous les documents convertis par un OCR. Cela signifie qu'au plus deux erreurs tous les cent mots sont permises dans les résultats de reconnaissance. Des études internes réalisées à la BnF ont montré qu'à partir de cinq erreurs pour cent mots, la lecture devient fastidieuse.

Pour garantir une telle qualité il faut pouvoir mettre en place des procédures de contrôle des transcriptions fournies par les OCR. En pratique, cela se traduit souvent par la mise en place de contrôles et de corrections systématiques par des opérateurs, car à ce niveau de qualité les procédures de contrôle automatique atteignent leurs limites. Cela restreint de fait la quantité de documents qu'il est possible de produire en un temps donné, et augmente de surcroit le coût de production. De manière concomitante ce niveau d'exigence complexifie les procédures de passation des marchés de numérisation de la BnF auprès des prestataires.

L'évaluation des techniques de numérisation est une question qui occupe de plus en plus la communauté scientifique qui s'intéresse d'une part à élaborer les critères d'évaluation les plus appropriés aux spécificités des problèmes rencontrés, et d'autre part favorise la constitution de corpus annotés nécessaires à la mise en place de campagnes d'évaluation rigoureuses accessibles aux chercheurs. Dans la littérature, il y a deux façons d'évaluer les systèmes de reconnaissance de caractères. L'approche dite « boîte noire » ou encore « Goal Directed » considère le système d'OCR comme un tout indivisible et se focalise sur le contrôle des résultats finaux des systèmes de reconnaissance de caractères. Ce type d'évaluation évalue l'influence de chacun des modules de l'OCR (pré-traitement, segmentation, reconnaissance ou post-traitement) au regard de sa contribution aux performances finales du système. Ainsi par exemple, deux algorithmes de pré-traitement ne sont pas comparés par rapport aux améliorations qu'ils sont sensés apporter dans les images qui leurs sont fournies en entrée, mais par rapport à l'amélioration relative du taux de reconnaissance de caractères qu'ils produisent dans le système d'OCR dans lequel ils sont placés. La deuxième approche d'évaluation est l'approche dite « boîte blanche » qui évalue séparément les différents modules qui forment les systèmes de reconnaissance. Par rapport à l'approche boîte noire, elle ne nécessite pas de disposer d'un système complet de reconnaissance pour être mise en œuvre. Mais en revanche elle nécessite l'utilisation d'un critère d'évaluation parfois complexe, voire impossible à exprimer rigoureusement, notamment sur des étages de pré-traitement. Par exemple, comment traduire l'amélioration apportée localement sur un pixel de l'image à l'amélioration globale de l'image. De plus, cette évaluation boîte blanche nécessite de disposer d'une vérité terrain pour chaque type de module de la chaîne de traitement, ce qui freine la constitution de corpus annotés de taille significative, étant donnés les niveaux de détails important qu'il faut considérer. Enfin, une dernière difficulté liée à l'évaluation de type boîte blanche réside dans la néces-

sité d'accéder aux résultats intermédiaires de chaque étage de traitement ce qui n'est pas toujours possible avec les systèmes de reconnaissance de caractères commerciaux.

Nous détaillons dans les paragraphes qui suivent les différents types d'erreurs que nous pourrons rencontrer dans les résultats de l'OCR ainsi que les différentes méthodes proposées dans la littérature pour détecter leur présence. Nous nous intéresserons principalement aux méthodes d'évaluation de la segmentation, et aux méthodes d'évaluation de la reconnaissance de caractères fondées sur l'utilisation d'une référence que l'on désigne généralement sous le terme de « vérité terrain ». Ces approches ne sont hélas pas transposables au problème qui nous intéresse car elles utilisent une vérité terrain. Ce constat nous amène à examiner dans la dernière partie de ce chapitre, les approchent qui cherchent à qualifier des résultats de reconnaissance sans vérité terrain.

3.1 Typologies des erreurs et métriques pour l'évaluation de performances des OCR

Evaluation des résultats de segmentation

(a) Erreurs de segmentation

La segmentation est l'opération de décomposition du document en unités structurelles, telles que les régions textuelles ou les graphiques. Une mauvaise décomposition du document conduit généralement à des erreurs dans l'ordre de lecture de la page, dans la classification des différentes composantes de la page, voire des erreurs de reconnaissance des caractères. Selon [SMJ+00], on peut décomposer les erreurs de segmentation en huit classes :

- La fusion horizontale de régions textuelles est une erreur qui entraîne la fusion de deux blocs textuels. Cette erreur de segmentation induit des erreurs dans l'ordre de lecture du document, où l'ordre de lecture a été modifié suite à une erreur de fusion horizontale des lignes du document traité. Dans d'autre cas, la fusion horizontale des mots peut causer des erreurs de reconnaissance puisque l'OCR considère les deux mots fusionnés comme un seul mot.
- La fusion verticale de régions textuelles est une erreur qui cause le regroupement vertical de deux éléments adjacents de la page. Cette erreur n'entraîne pas toujours d'altération dans l'ordre de lecture du document.
- La scission horizontale de régions textuelles est une erreur de segmentation qui entraîne la division d'un élément de la page horizontalement. Cette famille d'erreur biaise l'ordre de lecture du document.
- La scission verticale de régions textuelles est l'erreur de segmentation qui divise un élément de la page verticalement en plusieurs éléments. Cette erreur n'entraîne pas d'altérations dans l'ordre de lecture du document par contre on peut avoir des caractères fragmentés ce qui engendre des erreurs de reconnaissance de caractères.
- L'omission d'une région textuelle est l'erreur qui est provoquée par la classification des éléments considérés comme des zones de fond ou des zones graphiques.
- La confusion entre graphique/bruit et texte est l'erreur de segmentation qui conduit presque systématiquement (sauf rejet de la part du classifieur) à des insertions de caractères. Cette erreur est obtenue généralement sur des graphiques et des illustrations formés avec des traits qui possèdent des caractéristiques assez proches des

caractéristiques du texte.
- La fusion horizontale des éléments textuels avec des éléments graphiques ou bruit. Cette erreur est très fréquente dans les documents qui possèdent une mise en page compliquée. Elle entraîne généralement l'omission des caractères impliqués qui sont classés à tort comme éléments graphiques et donc pas soumis à la reconnaissance.
- La fusion verticale entre graphique et bruit, cette erreur de segmentation conduit au même genre d'erreur que dans le cas précédent.

(b) Critères d'évaluation des résultats de segmentation

Plusieurs articles [KNRN93] [KRNN95] proposent des études sur l'évaluation de la segmentation. Ces études décomposent les méthodes d'évaluation en deux catégories. La première catégorie de méthodes utilise une comparaison au niveau pixels pour mesurer la qualité des résultats de segmentation tandis que la deuxième catégorie évalue la qualité de la segmentation à travers l'alignement des résultats de reconnaissance des caractères. Nous les détaillons ci-dessous.

i. Approche image

Cette approche d'évaluation se base sur les coordonnées des éléments de la page décrits par des polygones pour vérifier l'affiliation de chaque pixel de l'image du document segmenté par un OCR aux régions correspondantes dans l'image de référence [SMJ+00] [RV94]. L'image de référence est constituée par différentes zones informatives validées manuellement par un opérateur humain. La difficulté de l'approche image est de savoir quelles régions de la vérité-terrain correspondent à quelles régions détectées par l'OCR. Pour surmonter cette difficulté plusieurs schémas de fusion-scission de régions dans les deux directions principales ont été mises en œuvre : correspondance un-à-un (one-to-one match), un-à-plusieurs (one-to-many match), plusieurs-à-plusieurs (many-to-many match). Ensuite, un score de recouvrement des régions est calculé pour qualifier la qualité de la segmentation. S. Randriamasy propose dans [RV94] un algorithme qualitatif d'évaluation composé par deux étapes :

1. Une première étape d'identification des opérations de fusion et de scission verticales et horizontales des régions qui permet d'apparier les résultats de segmentation de l'OCR avec la structure exacte de la page.
2. Une deuxième étape de qualification des erreurs de segmentation qui se réalise à travers un ensemble de règles de décision qui permettent de juger du niveau d'exactitude de chaque modification réalisée dans la première étape.

A chaque opération de fusion ou de scission, on associe un coût (ou une pénalité). La valeur de la pénalité varie selon l'importance de l'erreur de segmentation :
- Pénalité faible associée à une erreur mineure qui se produit généralement dans les zones de fond, par exemple une scission d'une zone de fond.
- Pénalité moyenne associée à une erreur acceptable qui se produit soit sur des illustrations soit sur des graphiques. Cette erreur cause la scission des représentations graphiques de la page, par contre elle n'engendre pas une dégradation dans le contenu textuel de la transcription automatique de la page.
- Pénalité importante associée à une erreur de segmentation majeure qui se produit sur des éléments textuels et qui cause une altération du contenu textuel. Par exemple, une scission d'un mot en plusieurs mots distincts.

En fonction du nombre de pixels appartenant à chaque région mal segmentée et du type de confusion réalisée on détermine le score final de la segmentation de la page par rapport à la vérité terrain. Pour évaluer quantitativement les résultats de segmentation, les auteurs ont défini deux mesures :

- Le score de la page qui est ratio de qualité reflétant les effets de segmentation sur les régions et leurs contenus

$$PS = w_{set}[w_{reg} \times (S_{reg}(GTF(I)) + S_{reg}(S(I))) + w_{pix} \times (S_{pix}(GTF(I)) + S_{pix}(S(I)))] \tag{2.2}$$

  Avec :
  - S_{reg} le ratio du nombre des régions bien détectées sur le nombre total des régions de la page,
  - S_{pix} le ratio du nombre des pixels bien détectés sur le nombre total des pixels,
  - $GTF(I)$ l'ensemble des régions de la vérité terrain de l'image I,
  - et $S(I)$ l'ensemble des régions obtenues lors de la procédure de segmentation automatique de l'image I.
- Le coût de la page désigne le nombre d'opérations de scission, de fusion et de reclassification nécessaires pour corriger la segmentation et se rapprocher du résultat de segmentation exacte de la page.

$$CP = w_{err} \times O_{err} + w_{acc} \times O_{acc} \tag{2.3}$$

  Avec O_{err} : le nombre des mauvaises opérations de segmentation et O_{acc} : le nombre des bonnes opérations de segmentation.
  $w_{set} = 0,5$ $w_{pix} = 0,5$ $w_{reg} = 0,5$ $w_{err} = 1$ $w_{reg} = 0,5$

D'autres métriques ont été proposées dans la littérature pour mesurer la qualité de la segmentation :

- Dans [NS95] trois métriques différentes évaluent la classification des blocs, la fusion des blocs et la justesse de l'ordre de lecture.
- Dans les compétitions ICDAR une mesure appelée f-measure est appliquée pour qualifier la qualité de la segmentation. Cette mesure est définie par des scores de précision **p** et de rappel **r** basés sur le taux de recouvrement entre les régions détectées par l'OCR et les régions de la vérité terrain (cf. [LPS+05]).

$$f - mesure = \frac{r \times p}{r \times p + (1 - \alpha) \times p} \quad avec(0 < \alpha < 1) \tag{2.4}$$

- Une métrique plus complexe, appelée PRImA-measure, a été utilisée dans la compétition ICDAR 2009 qui permet de qualifier la qualité de la segmentation d'une manière plus fine. Cette métrique calcule la fréquence et détermine l'effet de cinq situations de mauvaise segmentation (zone entièrement mal étiquetée, zone correctement étiquetée, zone partiellement localisée, zones fusionnées, zone totalement ou partiellement omise) en fonction du contexte du document et de l'application visée (cf. [AAP09]).

L'avantage de l'évaluation par une métrique fondée sur l'image réside dans sa capacité à prendre en compte la typologie des erreurs rencontrées en pratique. Cependant, cette

méthode est conçue pour tester des algorithmes de segmentation dans un contexte expérimental académique. Elle nécessite de disposer précisément des informations de localisation des différents blocs textuels et graphiques et donc de disposer d'une vérité terrain difficile à constituer.

ii. Approche textuelle

D'autres approches de type « boite noirs » développées dans la littérature se basent sur les sorties textuelles de l'OCR pour mesurer la qualité des résultats de segmentation des documents. En effet les erreurs de reconnaissance et de segmentation sont incluses dans les résultats de l'OCR. Le principe de ces méthodes d'évaluation est de distinguer les erreurs qui sont dues au module de segmentation et les erreurs qui sont dues au module de reconnaissance. Le score qui est ensuite calculé en faisant l'hypothèse que les erreurs de segmentation et de reconnaissance sont indépendantes, ce qui n'est pas tout à fait correct évidemment.

Le principe de cette évaluation consiste à appliquer une même reconnaissance de caractères sur les blocs de textes détectés par la méthode de segmentation que l'on cherche à évaluer et sur les blocs de texte de la vérité terrain. Les premiers résultats de reconnaissance regroupent les défauts de segmentation et de reconnaissance alors que les seonds résultats contiennent uniquement des défauts de reconnaissance des caractères.

La qualité des résultats de segmentation est évaluée à travers des mesures de coût de correction basées sur le principe de la distance d'édition qui permet de passer des résultats de la transcription automatique du document (zonage automatique + reconnaissance automatique des caractères) aux résultats de la transcription semi-automatique (zonage manuel + reconnaissance automatique des caractères) ([vRjktaN94]).

L'application de cette méthode nécessite une procédure de mise en correspondance entre les éléments textuels des deux résultats d'OCR. La mesure de l'erreur de segmentation est réalisée en comparant ces deux textes pour calculer un coût d'édition par des algorithmes de mise en correspondance des chaînes de caractères. [KRNN95] propose une autre métrique qui consiste à calculer le nombre des opérations d'édition qui permettent de transformer les textes de deuxième résultat de reconnaissance aux textes de premier résultat de reconnaissance.

L'application de cette méthode d'évaluation ne nécessite ni un format particulier pour représenter les zones, ni des traitements d'image spécifiques de mise en correspondance des pixels d'image. Cependant cette évaluation nécessite de disposer de la vérité terrain au niveau texte, ce qui est souvent plus fastidieux à constituer qu'une vérité de segmentation en blocs. Et il faut par ailleurs disposer d'un système de reconnaissance de caractères.

Evaluation des erreurs de reconnaissance

(a) Erreurs de reconnaissance de caractères

En plus des erreurs de segmentation des éléments de la page, des erreurs de reconnaissance de caractères peuvent apparaître à différents niveau dans les résultats de l'OCR. Elles peuvent toucher un seul caractère ou plusieurs caractères par mot.

Ces erreurs de reconnaissance de caractères sont dues à différents défauts. Les défauts de numérisation (comme par exemple le bruit de numérisation, les défauts de courbure) peuvent engendrer des déformations sur les formes des caractères. La binarisation des

images peut causer des troncatures qui altèrent les formes des caractères ce qui engendre des erreurs de reconnaissance. Finalement, les post-traitements peuvent être l'origine de certaines erreurs de reconnaissance de caractères. En effet, l'utilisation d'un dictionnaire lexique inapproprié peut causer des modifications dans les résultats de reconnaissance. Par exemple, dans l'ancien français « ie » référence au pronom personnel « je », l'utilisation d'un dictionnaire du Français récent engendre le remplacement de tous les « ie » par « je ». Alors que les projets de numérisation du patrimoine ont pour objectif la reproduction fidèle des textes. On peut décomposer les erreurs de reconnaissance de caractères en quatre types :

- Une confusion qui se traduit par le remplacement d'un caractère par un autre, elle se rencontre généralement avec les caractères qui ont des formes proches par exemple : « 0, O et o », « 1 et l », «s et 5 », « n et h », etc.
- Une suppression qui survient par l'omission d'un caractère en le considérant comme un bruit, comme un élément graphique, ou comme un élément appartenant à l'arrière-plan.
- Un rejet qui survient lors du remplacement d'un caractère par ~. Ce genre d'erreur survient lorsque le score de confiance de l'hypothèse de reconnaissance est trop faible
- Une insertion qui survient sur des caractères de forme ambiguë propise à la sur-segmentation tels que « m » qui peut être reconnu comme « rn », « d » qui peut être reconnu comme « cl », etc.

(b) Erreurs de reconnaissance de mots

Les erreurs de reconnaissance de mots peuvent être d'une part le résultat d'une mauvaise interprétation des éléments de la page lors de l'opération de segmentation et d'autre part le résultat d'une erreur de reconnaissance de caractères. Généralement, les erreurs d'estimation de la largeur moyenne des mots ainsi que des espaces inter-mots et inter-caractères conduisent soit à des erreurs de scission de mots, soit à des erreurs de fusion de mots.

De plus, les erreurs de reconnaissance de mots peuvent être obtenues suite à l'utilisation d'un dictionnaire inapproprié dans la procédure de post-traitement des résultats de reconnaissance. Ce genre d'erreur se traduit généralement par le remplacement total des mots absents du dictionnaire par les mots les plus proches. Les entités nommés (ex. les noms propres, les noms de villes et des pays, etc.) sont absents dans les dictionnaires, ils sont donc fréquemment mal reconnus, ores ils représentent très souvent des mots à travers lesquels les utilisateurs formulent leurs requêtes. De plus, dans les documents multilingues, les mots en langues étrangères sont généralement mal reconnus si le système d'OCR employé utilise un seul dictionnaire. Ce genre de scénario d'erreur est très fréquent dans les projets de numérisation de masse.

Les défauts physiques comme le surlignage, le fond grisé, la courbure des lignes ont également un impact important sur les erreurs de reconnaissance mots. Ce sont essentiellement les étapes de pré-traitements qu'il faut mettre en cause ici.

(c) Critères d'évaluation des résultats de reconnaissance de caractères

Etant donné des erreurs de reconnaissance des caractères et des mots présentées dans les sections précédentes, plusieurs travaux dans la littérature ont essayé de les qualifier par des mesures et des critères d'évaluation. La totalité des procédures d'évaluations proposées emploient des références absolues (vérité terrain) de mots ou de caractères pour les aligner

avec les résultats de reconnaissance des OCR. Ceci permet de dégager des mesures de qualité de reconnaissance.

Nagy [Nag95a] énumère trois méthodes dites explicites et deux méthodes dites implicites pour réaliser l'opération d'évaluation.

1. Les méthodes *explicites* se basent sur une opération de comparaison directe entre les résultats de reconnaissance automatique et les textes de la vérité terrain. Il faut pour cela disposer d'une vérité terrain qui indique la localisation des lignes de texte et leur transcription ASCII dans les images. Cette vérité terrain peut être obtenue de trois façons différentes :
   - La première procédure correspond à la démarche naturelle. On dispose des images scannées des pages et on doit construire la vérité terrain. Pour cela on procède à l'OCR-isation des images pour obtenir la transcription automatique des caractères. Ensuite, on corrige manuellement les résultats de l'OCR pour créer la vérité terrain. Cette procédure est finalement couteuse en moyens humains pour produire une quantité importante de données annotées,
   - Dans la deuxième procédure on suit la démarche inverse. On dispose des fichiers textuels de vérité terrain et on cherche à produire des images de documents contenant ces textes. Pour obtenir les images , on génère des images artificielles de documents. Pour se rapprocher le plus possible des images réelles, on applique ensuite des modèles de dégradation sur les images synthétisées,
   - En suivant la même démarche, on peut également créer des images synthétiques comportant des dégradation a réalistes à travers un traitement spécifique qui consiste à imprimer les documents synthétisés puis à les photocopier plusieurs fois (par photocopie des photocopies). L'ensemble des documents papier obtenus est enfin scanné pour produire des images de documents présentant des dégradations plus ou moins fortes.
2. Les méthodes *implicites* sont des méthodes d'évaluation indirectes qui essayent de qualifier la qualité des résultats de reconnaissance de caractères sans l'utilisation de la vérité terrain. On mesure l'effet des erreurs de reconnaissance des caractères sur les résultats des systèmes qui utilisent la transcription automatique des documents pour réaliser d'autres tâches de plus haut niveau (ex. les systèmes de tri du courrier, les systèmes de lecture de chèques, etc.).

A la différence de la décomposition de Nagy, Belaïd et al. proposent dans [Mul06] une autre décomposition des méthodes d'évaluation basée sur la précision de l'opération de vérification des résultats d'OCR. En fait, selon cette décomposition, on peut répartir les méthodes d'évaluation en deux familles :

1. Les méthodes d'évaluation globale
2. Les méthodes d'évaluation locale

Les méthodes globales considèrent les systèmes de reconnaissance de caractères comme des boîtes noires dont seules ses entrés et ses sorties sont accessibles à l'utilisateur. Les résultats d'OCR sont exportés en format texte brut. Les positions des mots et les caractères ne figurent pas dans les résultats de l'OCR. L'évaluation commence donc par une étape de mise en correspondance des mots qui permet de mettre en correspondance les mots reconnus par l'OCR avec les mots de la vérité terrain. Ensuite, en utilisant des opérations

d'édition (ajout, suppression et substitutions), on essaye de mettre en correspondance les caractères reconnus par l'OCR avec les caractères de la vérité terrain. Après ces opérations de mise en correspondance, on procède au calcul des distances d'édition qui indiquent la fiabilité de l'opération de reconnaissance des caractères. On estime ainsi un taux de reconnaissance de caractères ou de mots à partir d'une vérité terrain qui ne précise pas la position exacte des mots et des caractères dans les lignes.

L'application de ce type d'opération d'évaluation est généralement rapide. Par contre, les résultats d'évaluations produits ne sont pas très précis. Ils ne permettent pas de déterminer exactement les sources d'erreur de reconnaissance. Par exemple, l'erreur de substitution est quantifiée de la même façon sur la totalité du document. Or, en pratique on peut distinguer deux causes d'erreurs de substitution. La première cause est liée à l'utilisation mal appropriée de dictionnaires et la deuxième cause est liée à l'ambigüité des formes de caractères. Par conséquent, les techniques d'évaluation globales ne permettent pas de différentier ces deux types d'erreurs.

Les méthodes d'évaluation locales résolvent le problème des méthodes d'évaluation globales. En utilisant des descriptions plus fines de résultats de reconnaissance (comme les taux de confiance par caractères, les positions des caractères, les degrés de conformité des mots avec un dictionnaire de langue, etc.), les approches d'évaluation locale catégorisent les erreurs de reconnaissance -. Ce genre d'évaluation est très précis puisqu'elle permet de maîtriser l'opération de contrôle soit en fonction des taux de confiance de l'OCR, soit en fonction de l'endroit où se situent les mots (par exemple les zones de courbure de l'image).

L'évaluation locale permet donc de comprendre la cause des erreurs de reconnaissance. Par contre, l'inconvénient principal de ces méthodes reste dans le temps de traitement important nécessaire pour évaluer les résultats de l'OCR. Les résultats sont également beaucoup plus fastidieux à analyser.

i. Métriques pour l'évaluation explicites

Pour qualifier la qualité des résultats de reconnaissance de caractères, Belaïd définit dans (Mullot, 2006) plusieurs métriques. La mesure la plus intuitive est le taux d'erreur global Γerr qui donne le pourcentage de caractères incorrects dans les résultats de reconnaissance :

$$\Gamma err = 100 \times \frac{n_e}{n_c} \tag{2.5}$$

Où n_e est le nombre d'erreurs produites et n_c le nombre total de caractères dans la vérité terrain.

Certains systèmes de reconnaissance remplacent les caractères douteux par des « ˜ ». Par conséquent, on peut compter le nombre de « ˜ » dans les résultats de reconnaissance pour compléter le taux d'erreurs global par le taux de rejet qui représente le pourcentage de caractères rejetés par le système évalué. Ce taux est exprimé par :

$$\Gamma_{rej} = 100 \times \frac{n_r}{n_c} \tag{2.6}$$

Avec n_r le nombre des caractères rejetés.

Le taux de reconnaissance est par conséquent le ratio du nombre de caractères bien reconnus sur le nombre total de caractères qu'il faut reconnaître. La formule suivante donne le taux de reconnaissance de caractères :

$$\Gamma_{rec} = 100 \times \frac{n_c - n_r - n_e}{n_c} \tag{2.7}$$

La somme des trois taux précédents doit vérifier l'égalité suivante : $\Gamma err + \Gamma rej + \Gamma rec = 100\%$.

Le taux de confiance ou le taux de fiabilité défini par l'équation 2.8 est le taux d'erreur obtenu sur les résultats qui ne sont pas rejetés.

$$\Gamma_{conf} = 100 \times \frac{(n_c - n_r - n_e)}{(n_c - n_r)} = \frac{\Gamma_{rec}}{(\Gamma_{rec} + \Gamma_{err})} \tag{2.8}$$

Dans le contexte des projets de numérisation de masse, les prestataires de numérisation emploient le taux de confiance pour estimer la qualité des résultats de reconnaissance. Or on sait qu'en augmentant les cas de rejet en utilisant par exemple la confiance fournie par le classifieur, le nombre d'erreurs a tendance à diminuer. [Cho70] a montré que le taux d'erreur est approximativement l'inverse du taux de rejet $\Gamma_{err} = \frac{1}{\Gamma rej}$. De ce fait, il est possible d'augmenter le taux de confiance en augmentant le rejet du classifieur.

D'après Belaïd et al [Mul06], il existe au moins deux moyens plus au moins fins pour contrôler le taux de rejet :

1. Si le taux de confiance des caractères obtenus lors de l'opération de transcription automatique est disponible, on pourra contrôler le taux de rejet en fixant un seuil de confiance à partir duquel on peut rejeter les caractères. Selon Belaïd et al. si on se permet de rejeter un taux important de caractères de l'ordre de 3% à 5%, le taux de d'erreur reste relativement faible de l'ordre de 1 pour cent à 1 pour mille. Cependant, le rejet intensif des caractères rend l'opération de correction manuelle coûteuse. Les prestataires de numérisation considèrent souvent qu'au-delà d'un taux de confiance de 70%, l'utilisation de l'OCR est inutile car la ressaisie manuelle des documents devient plus rentable,
2. Si la fonction d'apprentissage des classifieurs est accessible, on peut déterminer les taux de rejet en fonction des taux de recouvrement entre les distributions des classes de caractères. Cela signifie qu'il est possible de fixer un taux de rejet par classe de façon à rejeter les caractères les plus fréquemment confondus.

Les résultats des systèmes de reconnaissance de caractères commerciaux ne contiennent que les caractères reconnus avec les taux de confiance calculés au niveau mot. Les paramètres de rejets ainsi que la fonction de calcul des taux de confiance ne sont donc pas disponibles au niveau des caractères. Par conséquent, pour fixer un taux de rejet, on doit soit utiliser les taux de confiance fournis par les OCR au niveau mots, sans savoir à quoi ils correspondent ni comment ils sont calculés ; soit utiliser une opération de post-traitement en utilisant des ressources linguistiques qui permettent d'estimer des taux de confiance au niveau mots ou caractères afin de fixer par la suite le taux de rejet adéquat. Il est d'ailleurs vraisemblable que les taux de confiance mots (Word Confidence) fournis par les OCR, utilisent des techniques de ce type sans qu'elles ne soient pour autant communiquées.

ii. Métrique pour l'évaluation implicite

L'estimation des performances des systèmes de reconnaissance des caractères en calculant le coût des opérations d'édition permet de comparer indirectement les résultats de reconnaissance des avec la vérité terrain. En effet, en utilisant la distance d'édition nous

calculons le nombre minimal d'opérations d'édition qui permet de transformer un texte reconnu par un système OCR à un texte de la vérité terrain.

L'utilisation des opérations d'éditions offre deux avantages principaux dans la procédure d'évaluation des résultats de l'OCR. D'une part, elle permet de déterminer d'une manière fidèle la qualité des résultats de reconnaissance ; d'autre part grâce aux opérations d'édition, on peut prévoir les types d'erreurs commises par les systèmes de reconnaissance qui sont en cours d'évaluation. Cependant, cette approche nécessite une bonne estimation des coûts d'édition, ce qui n'est pas une tâche facile. Idéalement cette estimation devrait simuler le coût pour un opérateur humain d'une correction complète d'un document. Qui plus est, cette approche repose sur l'existence d'une vérité terrain, ce qui ne sera jamais le cas dans le cadre de projets de numérisation de masse.

Une deuxième approche d'évaluation implicite a été présentée dans [Nag95a] qui permet d'évaluer les résultats de l'OCR en fonction de l'impact des erreurs de reconnaissance sur les rendus des systèmes qui se base sur la transcription automatique des caractères. En effet, les résultats de l'OCR ne sont pas toujours destinés à l'utilisation humaine. Ainsi, dans plusieurs domaine d'application les résultats de reconnaissance des caractères servent comme des données d'entré pour plusieurs systèmes (comme les systèmes de tri des courriers postaux, les systèmes de lecture des chèques bancaires, etc.). En se basant sur ce principe plusieurs méthodes d'évaluation implicites ont été proposées. Chacune d'entre eux mesurent l'effet des erreurs de reconnaissance des caractères sur le comportement du système étudié. Il s'agit donc de méthodes d'évaluation dirigées par les buts (goal directed evaluation).

D'autre part, selon les cas d'utilisation possibles des résultats de reconnaissance des caractères, plusieurs métriques ont été proposées. Par exemple, dans le contexte des systèmes de tri du courrier, la métrique utilisée pour qualifier les résultats de reconnaissance des caractères est le nombre de caractères erronés qui causent le renvoi du courrier à une fausse adresse.

Compagnes d'évaluation

L'institut ISRI (Institut de Recherche en Sciences de l'Information du Névada) a procédé au cours des années 1990 à une série de campagnes d'évaluation de systèmes d'OCR sur des documents imprimés de natures différentes ([RJN96] , [RJN95] , [RJN94]). D'après les résultats de ces campagnes, on constate que les systèmes de reconnaissance de caractères de l'époque avaient des performances assez élevées sur des documents de bonne qualité (cf. tableaux de la figures 2.5 et 2.6). On constate aussi d'après ces résultats que les performances des systèmes de reconnaissance varient d'un type de document à un autre. Elles étaient plus importantes sur des documents à structure simple comme les lettres, les ouvrages et les documents juridiques, que sur les documents techniques, les journaux et les magazines dont la structure est complexe.

Belaïd a évalué dans [CBd05] cinq systèmes de reconnaissance commerciaux (AbbyFine Reader, Omni Page, TextBridge et Type Reader) sur 200 pages obtenues à partir du journal officiel de la communauté européenne numérisées avec une résolution de 150 et 300 dpi ainsi que sur un ensemble de huit pages anciennes à double colonnes sélectionnées à partir du dictionnaire de Trévoux du XVIIème siècle et numérisées avec une résolution de 300 dpi. D'après les résultats de l'évaluation présentés dans le tableau 2.7, on constate que

Table 3c: English Newspaper Sample

	300 dpi Binary			300 dpi 8-bit Gray Scale		
	Errors	% Accuracy	Failures	Errors	% Accuracy	Failures
Caere OCR	5,079	98.97	none	7,478	98.48	none
EDT ImageReader	—	—	3 / 1.74	—	—	—
HP Labs OCR	6,432	98.69	none	5,125	98.96	none
INM NeuroTalker	47,773	90.29	none	—	—	—
Ligature CharacterEyes Pro	11,230	97.72	none	—	—	—
MAXSOFT-OCRON Recore	7,002	98.58	none	—	—	—
Recognita OCR	10,495	97.87	none	—	—	—
XIS OCR Engine	5,513	98.88	none	—	—	—

FIGURE 2.5 – Performance des systèmes de reconnaissance testés dans [RJN95]

	DOE Sample		Magazine Sample	
	# Misrec.	% Accuracy	# Misrec.	% Accuracy
Caere OCR	7,028	94.12	5,406	95.27
Calera WordScan	3,215	97.31	4,546	96.02
EDT ImageReader	11,199	90.63	9,153	92.00
ExperVision RTK	3,925	96.72	3,575	96.87
Recognita Plus DTK	12,385	89.64	7,690	93.28
XIS OCR Engine	4,693	96.07	6,486	94.33

Table 11: Word Accuracy

FIGURE 2.6 – Performance des systèmes de reconnaissance testés dans [RJN94]

les documents anciens présentent des taux d'erreurs plus importants que les documents récents.

Méthodes	%Reconnaissance	%Rejet	%Erreur
150ppp			
OCR 1 (Omnipage 12)	96.63	0.02	3.35
OCR 2 (FineReader 7)	71.79	0.10	28.11
Vote Majoritaire	69.24	27.69	3.07
Espace de connaissances (BKS)	98.02	0.00	1.98
Réseau de neurones (MLP)	97.49	0.00	2.51
300ppp			
OCR 1 (Omnipage 12)	96.96	0.00	3.04
OCR 2 (FineReader 7)	99.61	0.00	0.39
Vote Majoritaire	96.86	2.61	0.53
Espace de connaissances (BKS)	99.88	0.00	0.12
Réseau de neurones (MLP)	99.35	0.00	0.65

FIGURE 2.7 – Résultats de la procédure d'évaluation de Belaïd [CBd05]

Depuis 2003, la conférence ICDAR organise régulièrement des compétitions d'évaluations permettant de tester les nouvelles approches de segmentation et de reconnaissance de caractères. Les performances des méthodes testées sont rapportées dans la figure 2.8 et comparées avec des résultats de segmentation de deux logiciels de référence (Abby Fine Reader et OCRopus). Les résultats montrent que les méthodes de segmentation récentes ont une robustesse acceptable sur des documents simples avec une mise en page régulière. Cette robustesse décroît sur des documents avec des structures complexes ou imbriquées. On remarque aussi la difficulté qu'ont ces systèmes à localiser les informations non textuelles.

	Non-text	Text	Overall
DICE	19.58	43.16	39.12
Franhofer	56.93	80.15	76.45
REGIM-ENIS	44.70	9.14	16.08
Tesseract	45.85	72.06	68.44
FineReader	21.44	57.00	54.90
OCRopus	29.26	31.03	32.98
BESUS	35.63	37.68	35.85
Tsinghua1	34.91	76.12	67.59
Tsinghua2	36.58	72.93	65.70

FIGURE 2.8 – Performances des systèmes analysés dans le cadre de la compétition de *ICDAR 2009*

Conclusion

Les méthodes d'évaluation à base de vérité terrain sont très utilisées dans le contexte académique pour faire des expériences sur les approches de traitement des documents qui sont en cours de réalisation. Cependant, dans le contexte des projets de numérisation de masse, l'objectif n'est pas de tester de nouvelles méthodes de reconnaissance mais de s'assurer que les résultats de la numérisation, y compris la transcription des textes vers un format électronique standard, sont d'un niveau suffisant. Il est donc bien clair que la vérité terrain (le texte de référence) n'est pas disponible pour les collections traitées, puisque c'est précisément l'objectif de la numérisation automatique que de produire le texte le plus proche possible de la référence dont on ne dispose pas. Il n'est pas non plus imaginable de procéder à une saisie manuelle totale ou partielle des textes pour des raisons de coût. C'est d'ailleurs l'un des enjeux de la numérisation que de pouvoir produire des transcriptions à très faible coût grâce aux techniques de numérisation et d'OCR. Par conséquent, les méthodes d'évaluation de la littérature ne semblent pas adaptées aux besoins de contrôle des projets de numérisation de masse. En revanche elles constituent de très bonnes références en ce qui concerne les métriques utilisées pour constituer des critères quantitatifs permettant d'évaluer la qualité des données produites.

3.2 Les approches de contrôle des décisions des systèmes de reconnaissance

Les systèmes de reconnaissance de caractères et de parole sont loin d'être parfaits. En effet, leurs résultats dépendent énormément de la présence de bruit dans le signal et de l'adéquation des paramètres des classifieurs avec la nature des données à reconnaître. Cette variabilité des résultats de reconnaissance pose un certain nombre de problèmes lorsque le système de reconnaissance s'inscrit dans une chaîne complète de traitement où le résultat de la reconnaissance est ensuite exploité pour interroger une base de données par exemple.

Dans de telles situations il peut être très pénalisant de requêter le système sur des données erronées. On préfère annuler la suite des traitements dès qu'une incohérence, ou une incertitude trop importante est détectée dans les résultats de reconnaissance. On cherche alors à mettre en place des techniques de contrôle des décisions de reconnaissance qui se fondent sur l'élaboration d'une mesure de confiance dans la réponse des classifieurs.

De nombreux travaux ont traité ce problème dans la littérature. Nous pouvons décomposer les méthodes de contrôle des décisions de reconnaissance en deux familles : les

approches basées sur des méthodes de classification et les approches basées sur des opérations de post-traitement.

Les systèmes de reconnaissance de parole formulent le problème de reconnaissance des phonèmes de la même façon que les systèmes de reconnaissance de caractères. En effet, dans les deux on fait face à un problème de reconnaissance de formes complexe qui repose sur une première étape de représentation des données (des phonèmes ou des caractères) avant de les reconnaître dans un second temps en utilisant des modèles statistiques de classification. Par conséquent, nous ne distinguons pas les applications envisagées (parole ou écriture) dans les paragraphes qui suivent, et nous présentons les méthodes utilisées pour estimer le taux de confiance sur les résultats de reconnaissance, qu'il s'agisse de parole ou d'écriture.

Approche de vérification basé sur des méthodes de post-traitement

Comme on l'a vu dans la section 2.4 de cette première partie, les systèmes d'OCR valident les résultats de reconnaissance des caractères en utilisant des dictionnaires et des modèles de langage. Conformément à ce principe, les connaissances lexicales peuvent être modélisées par des modèles de langage statistiques pour calculer des taux de confiance sur des mots ou des séquences de caractères. Ces modèles sont utilisés pour donner aux séquences de caractères ou aux mots reconnus leur probabilité de conformité à un langage donné. Ce qui permet d'identifier les mots ou séquences de caractères incorrects dans les résultats de reconnaissance et guider les opérations de corrections manuelle ou automatique.

L'analyse la plus intuitive pour estimer le taux de confiance est celui qui emploi un modèle n-grammes de mots ou de caractères comme modèle de langage pour calculer un taux de confiance des mots ou des séquences de caractères. Plusieurs techniques sont possibles à ce niveau comme par exemple les modèles statistiques de n-grammes estimés par des techniques dites de repli (LMBB « *Langage Model Back-off Behavior* ») pour estimer les taux de confiance sur les mots. Selon [Mau06], en reconnaissance de parole/caractère les erreurs de reconnaissance sont propagées sur les mots voisins. Ceci signifie que lorsqu'un mot est mal reconnu, les mots qui l'entourent sont souvent affectés par des erreurs de reconnaissance. Par conséquent, il est important d'intégrer des mesures de confiance basées sur des informations concernant une séquence de mots plutôt que des mots isolés. La mesure de LMBB effectue ce traitement en associant l'ordre du n-gramme le plus élevé à un mot.

Le problème principal des méthodes de vérification linguistique réside dans la richesse des modèles de langage utilisés pour vérifier les résultats de reconnaissance. Le calcul des modèles de langage s'effectue sur des documents d'apprentissage. Par conséquent, les n-grammes de mot/caractères sont fortement liés au type et au champ lexical des documents d'apprentissage. Par exemple, si la base de documents d'apprentissage du modèle de langage est composée essentiellement de documents de médecine, le lexique comportera majoritairement des noms de maladies et des noms de médicaments. Pourtant, en dehors de ce contexte, ces mots sont très minoritaires dans les textes.

L'utilisation d'importantes bases de documents d'apprentissage résout en partie ce problème, par contre ceci peut engendrer des incohérences dans le modèle de langage obtenu lorsque l'on mélange des documents de différentes natures (de langue moderne avec des

documents en langue ancienne par exemple). En effet, une étude interne réalisée à la BnF a montré que l'utilisation de plusieurs modèles de langage lors des post-traitements peut accroitre les taux d'erreurs de reconnaissance des caractères si les ressources linguistiques ne sont pas adaptées.

Par conséquent, pour contrôler les résultats de l'OCR avec des modèles de langage, il faut utiliser pour chaque langue et chaque type de document le modèle de langage qui lui convient. Or ce mode de fonctionnement est inapproprié avec les caractéristiques des projets de numérisation de masse. En effet, la BnF programme dans ces projets, la numérisation de documents qui peuvent provenir de différents domaines (juridiques, religieux, administratifs, scientifiques, etc.). L'adaptation des modèles de langage à cette variabilité est une difficulté pour les méthodes qui utilisent des informations linguistiques pour estimer les taux de confiance. Dans le contexte de notre étude, de telles approches semblent délicates à mettre en oeuvre.

Approches de vérification basée sur des méthodes de classification

Dans la littérature, plusieurs méthodes ont été développées pour qualifier la qualité des résultats de reconnaissance. Ces méthodes ont analysé les scores de reconnaissance des éléments détectés, le temps de traitement des données, l'uniformité des caractéristiques des résultats de reconnaissance. De façon générale, le principe de ces méthodes se base sur la détection des éléments incorrects ou douteux dans les résultats de reconnaissance pour estimer le taux de reconnaissance des caractères. Par conséquent, il s'agit bien d'un problème de classification de forme de caractère dans lequel on accepte ou rejette le résultat de reconnaissance. Jiang dans [Jia05] décompose les méthodes d'estimation du taux de confiance qui appartiennent à cette famille d'approche en trois catégories :

1. Les méthodes qui exploitent les sorties des classifieurs (ex. scores de vraisemblance, probabilité a posteriori, N-meilleures hypothèses, etc.) et des informations linguistiques (ex. score de model de langage, comportement repli du modèle de langage) pour construire une signature pour chacune des deux classes. Un classifieur est alors utilisé pour donner une décision d'acceptation ou de rejet,
2. Les méthodes qui calculent la probabilité a posteriori des classes pour qualifier l'exactitude des résultats de reconnaissance. En fait, les algorithmes classiques de reconnaissance de formes formulent le problème de reconnaissance avec des règles de décisions basées sur la maximisation de probabilité a posteriori (cf. équation 2.12) des caractères, afin de déterminer la séquence de mots la plus probable. Le défaut principal de ces méthodes réside dans la détermination de la probabilité de l'observation (l'évidence) qui nécessite en théorie de connaître toutes les situations d'apparition de chaque observation ce qui est difficile voire impossible en pratique.
3. Le problème de l'estimation du taux de confiance peut être formulé comme un problème de test d'hypothèse statistique. Selon [Jia05], la vérification de l'énonciation est une opération de post-traitement qui permet d'examiner l'exactitude des résultats de reconnaissance en s'appuyant sur deux hypothèses complémentaires, l'hypothèse nulle H_0 qui représente l'hypothèse d'une reconnaissance correcte et l'hypothèse alternative H_1 qui représente les cas où la reconnaissance est incorrecte. Un modèle de Markov caché est utilisé pour représenter chacune des deux hypothèses. Ensuite,

on confronte les deux hypothèses H_0 et H_1 en évaluant le rapport de vraisemblance par rapport à un seuil de décision τ (cf. équation 2.14). La difficulté dans ce genre de méthode réside dans la formulation de l'hypothèse alternative : le rejet.

(a) Méthodes basées sur une combinaison de caractéristiques

i. Caractérisation des résultats de reconnaissance

Plusieurs travaux dans la littérature posent le problème du contrôle des résultats de reconnaissance comme un problème de reconnaissance de formes. En suivant ce schéma on cherche donc à caractériser les résultats de reconnaissance pour tenter ensuite de distinguer les résultats corrects et incorrects à l'aide d'un classifieur. Ces caractéristiques peuvent être de différentes natures (syntaxique, lexicale, sémantique, typographique, etc.). D'après [Jia05], on trouve dans la littérature un ensemble de caractéristiques qui décrit les résultats de reconnaissance et que l'on peut regrouper de la manière suivante :

- Le score de vraisemblance normalisé obtenu sur chaque image de caractère
- Les N-meilleurs résultats de reconnaissance.
- La stabilité des résultats de reconnaissance mesurée à travers le nombre d'hypothèses alternatives générées après post-traitement linguistique réalisé avec un modèle de langage.
- Durée de transition entre les états de treillis de HMM, la durée de reconnaissance des caractères et la durée de reconnaissance des mots.
- Des mesures reliées au modèle de langue (LM) comme le score de LM, LM back-off behavior, etc.
- La probabilité *a posteriori*

Selon [Jia05], aucun des descripteurs de la littérature n'est idéal. C'est la raison pour laquelle plusieurs travaux ont essayé de combiner ces caractéristiques pour améliorer la performance des estimateurs des taux de confiance.

P.Xiu propose dans [XB12] une approche de calcul du taux de confiance qui combine une caractéristique image issue d'un modèle iconique avec une caractéristique linguistique obtenue à partir d'un modèle de langage. Ce travail se base sur le constat de [SBZ03] qui considère qu'une image de document est représentée par deux modèles : un modèle iconique et un modèle linguistique. Dans un document isogène (où les caractères sont de forme homogène, même police), les erreurs de reconnaissance de caractères sont le résultat d'une imperfection dans au moins l'un des deux modèles. Par conséquent, en combinant les réponses des deux modèles, on peut déterminer si le caractère a été bien reconnu ou non.

En reconnaissance de parole, Zhang présente dans [ZR] une étude sur les caractéristiques des résultats de reconnaissance des caractères. Dans cette étude plusieurs combinaisons de caractéristiques acoustiques, et d'hypothèses (de modèle de langage, de treillis de mots et de liste de N-meilleur résultats) ont été employées, pour calculer les scores de confiance sur les éléments reconnus. Les meilleurs résultats sont obtenus avec des systèmes qui combinent la probabilité *a posteriori* des caractères reconnus avec l'information mutuelle des bigrammes de mots définie par le ratio de fréquence de bigrammes de mot dans un dictionnaire, sur le produit des fréquences des caractères qui constituent les bigrammes (cf. équation 2.9).

$$MI(x_i, y_j) = \log(\frac{P(x_i, x_j)}{(P(x_i)P(y_i)}) \tag{2.9}$$

ii. Prédiction de qualité de reconnaissance

D'autres prédicteurs peuvent être utilisées pour calculer un score de confiance dans la décision de reconnaissance. Des modèles linéaires de combinaison de caractéristiques ont été utilisés dans [Suk94] [SL96], ou dans [GIY97]. Siu et Gish rapportent dans [SG99] différents scores de confiance utilisés à ce niveau :

1. La régression logistique, qui se fonde sur une combinaison linéaire des caractéristiques

$$p\left(c_i = \frac{1}{X}\right) = \frac{\exp\left(\sum_j \beta_j x_{ij}\right)}{1 + \exp\left(\sum_j \beta_j x_{ij}\right)} \tag{2.10}$$

   extension non linéaire du modèle de régression logistique qui applique une transformation non-linéaire sur les caractéristiques avant leurs combinaisons (cf. équation 2.11).

$$p\left(c_i = \frac{1}{X}\right) = \frac{\exp\left(\sum_j g_j(x_{ij})\right)}{1 + \exp\left(\sum_j g_j(x_{ij})\right)} \tag{2.11}$$

où $p(c_i = \frac{1}{X})$ est la probabilité que le $i^{ème}$ caractère soit correctement reconnu, x_i est le vecteur de caractéristiques du i^{eme} caractère $x_i \in$ X et x_{ij}est le j^{eme} composant de x_i. Le coefficient β sont les paramètres de la régression qui sont déterminés lors d'une phase d'apprentissage. g_j est une fonction non linéaire qui permet de transformer les caractéristiques d'entrée.

D'autres approches d'estimation du taux de confiance ont utilisé un modèle de mélanges gaussiens [Chi92] pour représenter la distribution des caractéristiques des résultats de reconnaissance corrects et incorrects. [MM91], [WBR+97], [SSPH+01] ont appliqué des réseaux de neurones pour estimer le taux de confiance des résultats de reconnaissance. Les approches proposées dans ces travaux permettent d'estimer des taux de confiance au niveau mot.

Dans [SSPH+01], un perceptron multicouche (MLP) a été utilisé pour combiner les réponses des descripteurs. L'évaluation de cette approche a été réalisée sur trois niveaux différents : niveau mot, niveau énoncé et niveau contextuel. Les résultats des expérimentations ont montré que l'approche proposée est capable de rejeter 53,2% des mots incorrects, 53,2% de bruit d'énoncé et 50,1% des concepts incorrects.

D'autres approches d'estimation des performances des systèmes de reconnaissance ont utilisé d'autres familles de classifieur ; par exemple [EGJM95], [NRE97] ont appliqué des arbres pour classer les résultats de reconnaissance en classe de caractères corrects et incorrects. Généralement, les paramètres des différents modèles combinatoires des caractéristiques sont estimés à partir de certaines procédures d'entraînement discriminatives basées sur des critères de discrimination tels que l'entropie croisée, le taux d'erreur de classification [WBR+97].

(b) Méthodes basées sur la probabilité a posteriori

Les méthodes de reconnaissance des caractères sont des approches de classification de forme qui appliquent des règles de décision à base d'algorithmes de maximisation de

probabilité *a posteriori* pour déterminer la séquence des mots $\hat{W}$ la plus probable qui correspond aux formes des mots traités. Cette séquence des mots correspond généralement à la probabilité a posteriori maximale *a posteriori* maximale p(W|X) (cf. équation 2.12).

$$\hat{W} = argmax_{W \in \sum} \quad p(W|X) = argmax_{W \in \sum} \frac{p(X|W)p(W)}{p(X)} \tag{2.12}$$

Avec $\sum$ désignant l'ensemble des mots à reconnaître, $p(W)$ est la probabilité de séquence des mots W obtenue à travers des modèles de langage. $p(X)$ est la probabilité de l'observation X et $p(X|W)$ est la probabilité de l'apparition avec la séquence des mots W. Théoriquement, $p(W|X)$ la probabilité *a posteriori* $p(W|X)$ est une bonne métrique pour mesurer la confiance de la décision de reconnaissance W. Dans la pratique, les systèmes de reconnaissance de caractères ignorent toujours la probabilité de l'observation car elle n'affecte pas l'ordre des hypothèses, donc la décision. Cependant, cette simplification rend les scores des décodeurs inadéquats pour juger la pertinence des résultats de reconnaissance. Théoriquement, la probabilité de l'observation est définie par la fonction suivante :

$$p(X) = \sum_{H} p(X, H) = \sum_{H} p(H)p(X|H). \tag{2.13}$$

H désigne toutes les apparitions possibles du mot X. La somme doit être effectuée sur toutes les hypothèses d'apparition de l'observation X qui correspondent à toutes les combinaisons des mots, des caractères et du bruit. De toute évidence, sans aucune autre condition, il est impossible d'énumérer et de modéliser toutes les hypothèses d'apparition de l'observation X ce qui rend le calcul exact de la probabilité d'observation impossible.

Dans la littérature, plusieurs travaux ont traité la problématique de l'estimation de la probabilité d'observation p(X). Ces travaux sont décomposés dans [Jia05] en deux catégories :

- Les méthodes à base de remplissage [You94] représentent la première famille des méthodes. Elles estiment la probabilité de l'observation en prenant en compte des modèles de fond qui regroupent tous les phonèmes à reconnaître. Selon [Jia05], ces approches sont simples à mettre en œuvre et elles réalisent généralement des bonnes performances dans l'estimation du taux de confiance.
- Les méthodes à base de treillis constituent la deuxième famille des méthodes. Elles essayent d'estimer la probabilité d'observation à travers le parcours du treillis de la chaîne de Markov cachée en utilisant l'algorithme de forward-backward [KS97]. D'après [Jia05], l'utilisation du treillis donne la meilleure estimation de la probabilité a posteriori. En effet, grâce à sa capacité de représentation de toutes les hypothèses alternatives des résultats de reconnaissance, l'utilisation de la fonction devient envisageable et par conséquent l'approximation de la probabilité *a posteriori* est meilleure. Cependant, la génération des graphes des mots pour le calcul de la probabilité *a posteriori* est généralement une tâche compliquée et coûteuse en terme de temps de calcul. Une alternative des graphes des mots consiste en l'utilisation des listes de n-meilleurs résultats de reconnaissance pour approximer la probabilité d'observation p(X).

(c) Méthodes basées sur le test d'hypothèse

Certaines approches de vérification des résultats de reconnaissance formulent le problème d'estimation du score de confiance en utilisant un test d'hypothèse. Pour une image

de caractère X donnée, ces approches supposent qu'elle appartient à la classe des caractères W. Ce résultat est généré par un modèle de Markov caché λ_w. Dans le cadre des méthodes de test d'hypothèse, on commence généralement par la définition de deux hypothèses complémentaires :

1. H0 est l'hypothèse nulle qui suppose que l'image de caractère X est correctement reconnue par le modèle HMM λ_w.
2. H1 est l'hypothèse alternative qui suppose que l'image de caractère X est mal reconnue par un modèle HMM $\lambda_{w'}$ autre que λ_w.

Ensuite, on compare l'hypothèse nulle par rapport à l'hypothèse alternative pour déterminer si nous devons accepter les résultats de reconnaissance ou les rejeter. Selon Neyman-Pearson Lemma, en respectant certaines conditions, la solution optimale est obtenue en utilisant le test de ratio vraisemblance (LRT).

$$LRT = \frac{p(X|H_0)^(H_0)}{p(X|X_1)_(H_1)} \geq \tau \tag{2.14}$$

τ est le seuil de la décision critique. Selon [Jia05], la vérification des résultats de reconnaissance basée sur l'application du LRT fournit une bonne formulation théorique pour estimer les taux de confiance des résultats de reconnaissance. D'après [Lee01], le LRT peut être transformé en un score de confiance en utilisant une fonction de normalisation appropriée. La modélisation de l'hypothèse alternative représente la difficulté principale du LRT. En effet, la distribution des données de l'hypothèse alternative est inconnue.

Pour surmonter cette difficulté, plusieurs travaux ont été proposés dans la littérature ([RLJ97], [SSLJ97], [JD01]). Selon [Jia05], les approches proposées dans ces travaux appliquent la même modélisation par modèles de Markov cachés, utilisés pour modéliser les données de l'hypothèse nulle, pour simuler la distribution des données de l'hypothèse alternative. On construit de cette façon des modèles génériques de remplissage (filler model), des anti-modèles d'hypothèses spécifiques, des modèles des données synthétiques ou une combinaison de l'ensemble de ces modèles pour formuler l'hypothèse alternative et évaluer son score de confiance.

4 Conclusion

L'objectif du travail effectué dans le cadre de notre projet de recherche est l'amélioration de la maîtrise des projets de numérisation de masse de la BnF. Pour atteindre cet objectif, il est important de comprendre le déroulement des projets de numérisation de masse. Dans cette première partie, nous avons détaillé les différentes étapes de réalisation des projets de numérisation de masse. A la BnF, nous pouvons décomposer la procédure de réalisation des projets de numérisation en trois phases :

1. Phase de préparation des projets de numérisation qui englobe la procédure de préparation des cahiers de charges et l'opération de sélection des documents
2. Phase de production des documents numériques qui englobe la procédure de segmentation des images et de reconnaissance des caractères
3. Phase de contrôle de qualités des documents numériques

Pour décrire la première phase, nous avons commencé par la présentation des critères utilisés pour sélectionner les documents. Ces critères sont décomposés selon leurs natures en trois catégories : des critères liés à l'ouvrage, des critères liés à l'œuvre et des critères liés à la procédure de numérisation. Nous avons montré également que la qualité des documents numériques dépend énormément des critères de sélection des documents physiques.

Dans la première partie du deuxième chapitre, nous avons détaillé l'ensemble des opérations qui permettent d'acquérir les images du document. Nous avons présenté également les paramètres qui permettent d'adapter les résultats de l'opération de numérisation selon les objectifs des projets de numérisations de masse. En fait, on peut décomposer les besoins de numérisation en deux catégories :

1. la numérisation pour la conservation qui cherche à produire des originaux numériques qui peuvent être utilisés par la suite pour produire des formats pivots numériques des documents originaux.
2. la numérisation pour la communication qui cherche à produire des documents numériques consultables facilement via internet.

La deuxième partie de ce chapitre s'intéresse aux systèmes de reconnaissance de caractères. Pour cela, nous avons commencé par la présentation des différents traitements réalisés par l'OCR tels que les pré-traitements, la segmentation des éléments de la page, la reconnaissance des formes des caractères et les post-traitements. Nous avons exposé également les techniques appliquées pour effectuer ces traitements. Ces techniques ont généralement de bonnes performances sur des collections documentaires éditées avec des règles standards. Cependant, sur des documents historiques et anciens, les performances de ces systèmes de reconnaissance ont une tendance moins bonnes voir médiocres.

Le contrôle de la qualité des documents numériques représente un enjeu capital pour les archives et les institutions culturelles. En effet, les bibliothèques numériques soulèvent depuis une dizaine d'années des inquiétudes concernant l'intégrité des données numériques qu'elles agrègent. Les défauts de segmentation et de reconnaissance de caractères biaisent énormément l'intégrité de ses collections numériques. De plus, à cause de la volumétrie importante des projets de numérisation de masse, le contrôle des résultats de conversion automatique des caractères est une tâche difficile et très coûteuse en termes de temps de traitement. Dans la littérature, le problème de vérification des résultats de reconnaissance des caractères dans le contexte des projets de numérisation de masse n'a quasiment pas été étudié par la communauté scientifique alors qu'il représente la première priorité des acteurs des projets de numérisation de masse. Par conséquent, nous avons orienté vers la résolution de ce problème en proposant des solutions originales qui s'adaptent parfaitement aux besoins de numérisation des projets de numérisation de masse.

Dans la deuxième partie de cette thèse, nous présentons notre contribution en ce qui concerne le sujet de contrôle des résultats d'OCR. Dans le contexte des projets de numérisation de masse, la difficulté dans la procédure de contrôle de qualité réside dans l'absence de la transcription exacte des documents numériques (vérité terrain). Ceci rend la procédure de contrôle de qualité classique basé sur l'utilisation de la vérité terrain inutilisable dans notre contexte. Pour pallier à cette difficulté, nous avons proposé deux approches de vérification des résultats d'OCR.

la première approche s'intéresse à la détection des mots omis dans les résultats de l'OCR pour vérifier la qualité des documents numériques issus de la procédure de numé-

risation de masse. Dans la section 3 du chapitre « Contexte », nous avons montré que les erreurs d'omission des mots ne sont pas couvertes ni par la procédure de contrôle de la BnF ni par la procédure de contrôle des prestataires de numérisation.

Pour résoudre ce problème, nous avons proposé une approche de détection de texte, générique dans sa démarche et adaptative dans ces traitements, pour localiser les mots omis dans les images du document. Cette approche s'adapte parfaitement avec la variabilité des propriétés typographiques et physiques des documents de la BnF. Ce qui la rend utilisable dans la chaine de contrôle de la BnF. Les résultats fournis pas cette approche permettent d'une part de rejeter les documents possédant des taux d'omission importants et d'autre part de rectifier les taux de reconnaissance des mots présentés dans les fichiers ALTO de chaque document numérique de la BnF.

Dans un deuxième temps, nous nous sommes intéressés au problème de contrôle des résultats de reconnaissance des caractères. En effet, les systèmes d'OCR fournissent dans les résultats de transcription automatique des caractères un score de confiance appelé « Word confidence » qui permet à la BnF de juger de la qualité de ces documents numériques.

La formule de calcul de ces scores n'est pas transparente. De plus, les contrôleurs de la BnF ont constaté qu'elle surestime généralement le vrai taux de reconnaissance. Par conséquent, pour pallier à ce problème, nous avons développé dans la deuxième partie de notre contribution une approche du taux de reconnaissance des caractères qui nous permet de qualifier correctement les résultats de l'OCR. Pour cela, nous avons employé des réponses de deux descripteurs génériques qui se basent sur l'opération d'alignement des résultats de reconnaissance de caractères et sur le taux d'isogénie des formes de caractères appartenant aux mêmes classes de caractères de l'image.

De plus, nous avons employé une procédure d'apprentissage adaptative qui nous permet d'adapter l'estimateur du taux de reconnaissance en fonction des pages à vérifier. L'utilisation de ces techniques a permis d'adapter notre approche à la variabilité des propriétés typographiques et physiques des collections documentaires de la BnF.

Deuxième partie

Contribution

Chapitre 3

Contrôles des résultats de segmentation

1 Introduction

Dans les projets de numérisation de masse de la BnF, les résultats des systèmes d'OCR sont sauvegardés dans des fichiers XML selon le format ALTO [1]. Les fichiers ALTO de la BnF contiennent, en plus des transcriptions réalisées par l'OCR, les positions des éléments de la page [ndF13]. Les technologies de reconnaissance de caractères ont atteint depuis quelques années un certain niveau de maturité technique, qui leur permet d'obtenir de bonnes performances, surtout sur des documents récents édités avec des règles d'édition standard. Cependant, la collection documentaire de la BnF regroupe en plus des documents récents des documents anciens, des journaux et des manuscrits qui se caractérisent par des typographies anciennes et par des mises en page complexes, ce qui conduit à des défauts considérables dans les résultats de segmentation de la structure physique des pages.

Dans le chapitre « Etat de l'art », nous avons présenté les différents types d'erreurs de segmentation des pages. On peut décomposer ces erreurs selon leur nature en quatre classes : fusion de blocs horizontalement et verticalement, scission de blocs horizontalement et verticalement, omission de composants de la page et confusion d'éléments détectés dans la page. Les défauts de segmentation de type fusion, scission et confusion conduisent généralement à des erreurs de reconnaissance de mots. A la BnF, les résultats des systèmes d'OCR ne sont contrôlés qu'à travers les taux de reconnaissance de mots estimés par les OCR. De ce fait, les erreurs d'omission de mots ne sont pas analysées et par conséquent elles échappent à la procédure de contrôle des résultats d'OCR.

A cause de fonctionnement en boîte noire des OCR commerciaux utilisés dans les projets de numérisation de masse de la BnF, l'origine des erreurs d'omission de mots reste toujours ambiguë. En effet, elles apparaissent à différents niveaux (caractères, mots, phrases, paragraphes) dans les résultats de l'OCR. Parfois, ces erreurs sont causées par des défauts physiques ou typographiques qui rendent les caractéristiques des éléments textuels des pages semblables à celles du bruit. Par contre dans d'autre cas, elles apparaissent dans des régions qui sont bien imprimées et qui ne possèdent aucun défaut physique apparent.

Pour vérifier l'existence des erreurs d'omission d'éléments textuels dans les résultats

1. http ://www.loc.gov/standards/alto/

d'OCR, nous avons développé dans cette partie une approche locale de détection d'éléments textuels omis. La localité des traitements de notre approche provient de l'utilisation de caractéristiques des éléments détectés dans la page à vérifier pour chercher des éléments similaires dans les régions identifiées comme du fond et donc censées d'être vides de tout élément textuel ou graphique.

Dans la suite de ce chapitre, nous commençons par l'analyse des erreurs d'omission dans les pages. Ensuite, nous présentons la méthodologie de notre approche ainsi que les différentes techniques qui nous ont permis de caractériser et reconnaître les différents composants omis dans la page. Enfin, nous terminons par l'évaluation de l'approche proposée avant d'évoquer les perspectives d'amélioration.

2 Les erreurs d'omission des éléments de la page

Bien que les performances des systèmes d'OCR aient atteint un niveau de performance important, la mise en production de ces technologies sur des documents techniques et patrimoniaux conduit trop souvent à des erreurs de segmentation. Dans la première partie de cette thèse, nous avons énuméré les différentes erreurs que l'on peut rencontrer dans les résultats de segmentation d'un document. Ces erreurs peuvent se produire à différents niveaux dans la page (paragraphe, ligne, mot et caractères).

Les erreurs d'omission des mots peuvent se reproduire à différents niveaux dans le document. Parfois, on a des caractères qui manquent dans les mots (cf. figure 3.1b) et parfois on a même des blocs textuels qui sont oubliés par l'OCR (cf. figure 3.1c). Certains éléments omis (cf. figure 3.1a) sont pourtant bien délimités et bien contrastés. La raison de leur omission n'est pas toujours bien comprise du fait du mode de fonctionnement en boîte noire des OCR. De plus, l'architecture des systèmes de reconnaissance de caractères composés de plusieurs étapes de traitements organisés de façon séquentielle tend à cumuler et même amplifier les erreurs à chaque étape.

Par exemple, si le contraste entre les éléments textuels et le fond de la page n'est pas significatif, les algorithmes de binarisation peuvent causer des suppressions drastiques d'éléments textuels qui possèdent des intensités de niveaux de gris clair. D'autre part, pour atteindre les seuils de taux de reconnaissance exigés dans les cahiers des charges de la BnF, les prestataires de numérisation peuvent paramétrer leurs OCR de façons à ce qu'ils rejettent les éléments reconnus avec un faible taux de reconnaissance. Par conséquent, ils seront automatiquement supprimés, et donc omis dans le calcul d'estimation des performances.

De ce fait, une procédure de contrôle de résultats des OCR est obligatoire pour contrôler ce genre d'erreur. Actuellement, l'ensemble des opérations de contrôle est réalisé par des opérateurs humains sur des échantillons limités de pages. Elles commencent généralement par une première étape de vérification de la qualité des résultats de segmentation qui se poursuit par une étape de contrôle des résultats de reconnaissance des caractères.

Cependant, l'application de cette procédure de contrôle possède deux inconvénients majeurs. D'une part l'utilisation d'un échantillon de quelques centaines de pages ne peut pas refléter la qualité des résultats de l'OCR obtenus sur les millions de page des projets de numérisation de masse. D'autre part, certains défauts comme l'omission des contenus textuels ne sont pas détectés par cette opération de contrôle. En effet, dans les campagnes de

(a) Plusieurs elements omis (b) Un caractères omis (c) Un bloc omis

FIGURE 3.1 – Exemples des résultats d'OCR fournis par des prestataires ; les rectangles rouges représentent les boîtes englobantes des mots, le rectangle vert qui délimite la lettrine représente l'élément graphique ; (a) Exemple de résultat d'OCR qui regroupe des mots partiellement détectés, des mots entièrement omis et des illustrations complétement oubliées. (b) Exemple de caractère manquant dans les résultats d'OCR, (c) Exemple d'un résultat d'OCR avec omission d'un bloc textuel et confusion d'un bloc de texte en graphique.

contrôle manuel, seuls les éléments reconnus par l'OCR sont vérifiés. Des études réalisées au sein du service de numérisation de la BnF ont montré que généralement les transcriptions automatiques de documents anciens ont tendance à contenir beaucoup d'erreurs d'omission, cela dans le but d'optimiser le taux de reconnaissance de caractères en maximisant le rejet de tous les résultats de reconnaissance douteux. Mais cette stratégie souvent mise en œuvre par les prestataires de numérisation met en péril l'intégrité des archives et des bibliothèques numériques et biaise les résultats des systèmes tiers qui se basent sur les sorties des OCR pour effectuer des opérations de plus haut niveau comme l'indexation des pages, la détection des entités nommées, l'analyse syntaxique et sémantique des documents, etc.

Pour détecter la présence de ce genre d'erreurs dans les résultats de l'OCR, nous proposons ici une approche [SRP13] qui détecte les éléments omis dans la page en utilisant des techniques de localisation et de reconnaissance des éléments textuels dans les images (cf. état de l'art dans la première partie section 2.2). Nous détaillons dans les paragraphes suivants les différents composants de notre approche.

3 Méthodologie

Intuitivement, pour détecter les éléments omis de la page, nous pouvons appliquer une procédure de segmentation traditionnelle qui permet d'identifier et localiser les éléments de la page puis de comparer les résultats de segmentation de notre algorithme avec les résultats de segmentation de l'OCR du prestataire pour détecter les éléments qui ne coïncident pas.

Cependant cette démarche est difficile à mettre en œuvre. Une première raison est la variabilité de la collection documentaire de la BnF, puisqu'elle regroupe différents types de

documents (des ouvrages, des journaux, des articles scientifiques, etc.) édités avec plusieurs technologies d'impression (manuscrits, typographies anciennes et typographies récentes). La figure 3.2 représente des exemples de page de la collection documentaire de la BnF. Cette variabilité implique le développement d'une approche de segmentation (ou détection) d'éléments omis qui permet de couvrir la totalité de la collection documentaire de la BnF. D'autre part, la segmentation des éléments des pages qui proviennent de ce genre collection a déjà fait l'objet de recherche avancées [LSZT07] et [JKJ04]. Les algorithmes développés dans ces travaux sont généralement pris en compte par les systèmes d'OCR commerciaux. Par conséquent, il semble difficile de réaliser des résultats de segmentation meilleurs que ceux qui sont déjà obtenus par les OCR si ce n'est pas sur un seul type de document. L'objectif de notre traitement n'est donc pas de développer une nouvelle méthode de segmentation d'images de documents, mais de vérifier l'existence d'éléments omis par l'OCR. Par conséquent, nous ne cherchons pas à segmenter tous les éléments de la page pour déterminer ceux qui ont été omis et à localiser de manière précise les éléments de la page.

Traditionnellement, les algorithmes de détection d'éléments textuels utilisent une base d'apprentissage composée de plusieurs exemples d'images (avec texte et sans texte) afin d'entraîner le détecteur avec les caractéristiques des caractères à localiser dans l'image. Or cette procédure est incompatible avec notre contexte. En effet, dans la pratique, il est presque impossible de concevoir un descripteur de texte capable de décrire les images de tous les types des documents. Par contre, si nous essayons d'adapter nos descripteurs aux propriétés locales des images de document traitées, nous pouvons espérer obtenir une méthode de détection générique adaptable à tous types de corpus. C'est cette stratégie qui a guidé notre réflexion.

Selon les contrôleurs de la BnF, les résultats de segmentation des prestations de numérisation dans le cadre de projets de numérisation de masse souffrent peu de confusion entre les éléments textuels et les éléments graphiques. De plus selon [XB12], la typographie des caractères au sein d'une page est généralement isogène. Cela signifie que les éléments textuels omis par l'OCR ont souvent des caractéristiques typographiques similaires à ceux qui ont été détectés. D'autre part, les éléments omis par l'OCR sont généralement confondus avec des régions de l'arrière-plan de la page. En effet, les éléments textuels inclus dans les illustrations (les tampons, les publicités, etc.) ne sont pas considérés comme du texte. Cette dernière constation permet de nous focaliser sur l'analyse des régions étiquetées comme arrière-plan par l'OCR pour identifier les éléments textuels omis par l'OCR.

En se basant sur ces observations, nous pouvons utiliser les caractéristiques des éléments textuels présents dans les fichiers ALTO pour rechercher des éléments similaires dans les régions classées par l'OCR comme arrière-plan. Ceci permet d'adapter les classifieurs de pixels de notre approche avec les caractéristiques locales des éléments textuels de la page et par conséquent de rendre notre approche totalement adaptative. De plus, afin de garantir une description générique des différentes régions de la page, nous avons opté pour l'utilisation de caractéristiques de type texture. En effet, dans l'état de l'art nous avons montré que l'utilisation de caractéristiques de type texture permet de caractériser les éléments de l'image sans aucune connaissance *a priori* ce qui est tout à fait en adéquation avec les orientations que nous avons souhaité donner à cette étude.

Une texture est caractérisée par sa direction, sa régularité et son degré de contraste

entre pixels foncés et pixels clairs. Par conséquent, l'algorithme de détection des éléments textuels que nous proposons va exploiter ses propriétés en utilisant des descripteurs d'orientations, de régularité et de contraste. Ensuite, en se basant sur les résultats de segmentation de l'OCR, les régions de l'avant-plan sont utilisées pour apprendre les caractéristiques de la page alors que les régions de l'arrière-plan sont quant à elles analysées afin de se prononcer sur leur éventuelle similitude avec les autres éléments de la page ou du document. Dans le cas d'une similitude importante avec les autres éléments textuels ou graphiques détectés par l'OCR, la région d'arrière-plan considérée est réattribuée à la classe appropriée.

L'analyse et la détection étant réalisées au niveau des pixels, un traitement à base d'analyse en composantes connexes est appliqué pour passer du niveau pixel au niveau mot et pour construire l'image des enveloppes des mots détectés. Enfin, un filtre basé sur la taille des éléments détectés est appliqué pour filtrer le bruit de détection. Les éléments textuels très petits sont supprimés.

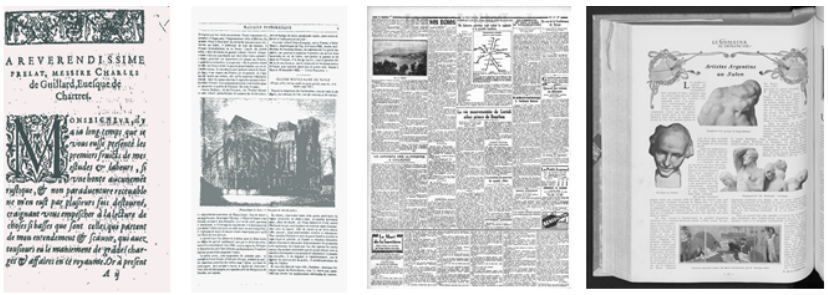


FIGURE 3.2 – Exemples de pages de documents traitées dans le cadre des projets de numérisation de masse.

La figure 3.3 présente l'organisation des différents traitements de notre approche. Nous décomposons ceux-ci en quatre étapes :

1. Caractérisation des textures de l'image de la page.
2. Apprentissages des détecteurs d'éléments textuels et graphiques
3. Classification des pixels de fond
4. Passage du niveau pixel au niveau composantes connexes

L'ensemble de ces opérations est réalisé à chaque opération de vérification de page ou de document. Nous détaillerons dans les sous-parties suivantes les caractéristiques de chaque étape de traitement.

3.1 Caractérisation des résultats de segmentation

A l'instar des méthodes de segmentation des pages, nous proposons une démarche axée sur l'extraction et l'apprentissage de caractéristiques locales des éléments textuels et graphiques pour détecter les éléments omis dans la page. Pour cela, nous avons appliqué un ensemble de descripteurs génériques pour caractériser les éléments de la page. Face aux propriétés des images du corpus de la BnF, nous avons porté notre choix sur des descripteurs bas niveau.

L'agencement des mots sur la page ainsi que leurs formes rend l'apparence globale des éléments textuels semblable à celle d'une texture. Nous avons vu dans le chapitre «

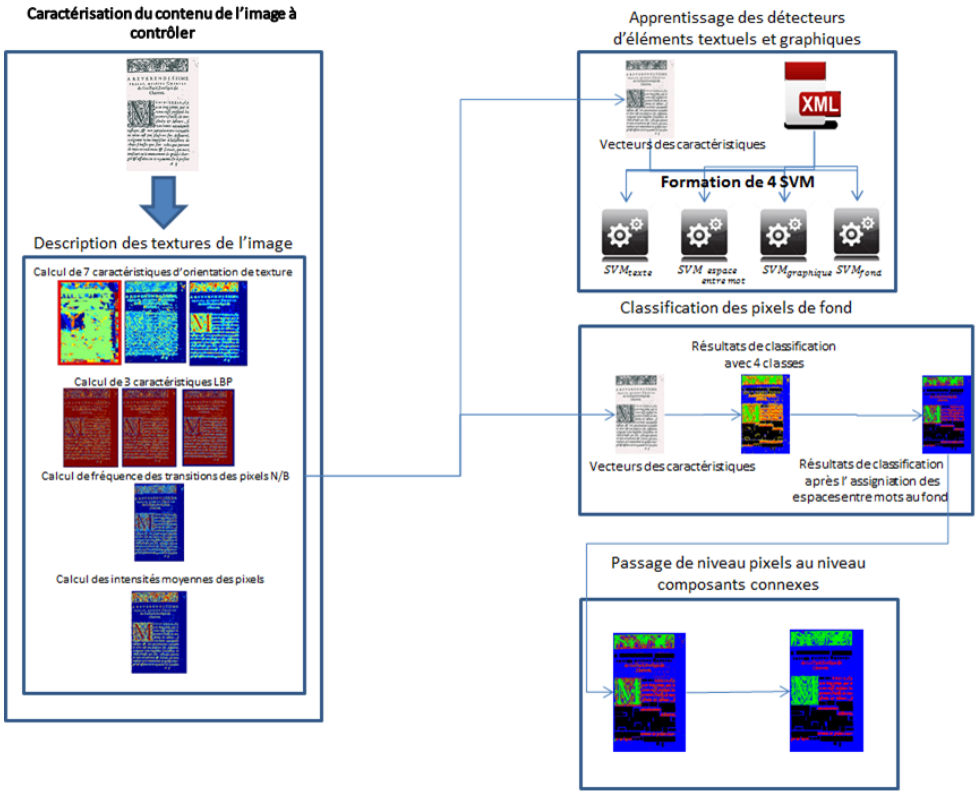


FIGURE 3.3 – Organisation des étapes de l'approche de détection des éléments omis.

Etat de l'art » que les descripteurs de texture sont souvent appliqués sur des collections de documents très variables. En effet, ces méthodes dites « bas niveau » permettent de décrire les formes sans introduire de connaissance préalable relatives aux caractéristiques physiques et typographiques du document et à la mise en page de ces documents. Par conséquent, l'application de ce genre de descripteurs est tout à fait adaptée au contexte de notre travail.

Les régions textuelles et graphiques ont généralement des apparences différentes. Par exemple, les régions textuelles présentent une orientation dominante selon la direction horizontale alors que les régions comportant des illustrations peuvent présenter plusieurs orientations importantes. De plus, selon [Egl08] dans un traitement multirésolution, la direction principale de la texture dans les régions textuelles est la même à différentes échelles (cf. figure 3.4). Par contre, la direction principale des textures des régions graphiques est variable d'une échelle à une autre (cf. figure 3.3). Par conséquent, nous pouvons utiliser cette caractéristique pour décrire les textures des éléments textuels de l'image.

Dans notre approche nous avons caractérisé les textures des images de la page en utilisant trois familles de caractéristiques :

1. La première famille est liée à l'orientation principale du texte. Nous avons pour cela mesuré le consensus des orientations principales de la texture à différentes échelles, l'intensité médiane des orientations de la texture et la variation des orientations de la texture dans une région. Cette famille de caractéristiques permet de caractériser la plupart des éléments de la page.
2. La deuxième famille de caractéristiques décrit la régularité de la texture de l'image

pour différencier les régions lisses des régions textuelles.

3. La troisième famille de caractéristiques est liée à la fréquence de transition entre pixels clairs et pixels foncés dans une fenêtre de taille prédéfinie.

A partir de ces caractéristiques, nous avons formé pour chaque pixel de l'image un vecteur de caractéristiques composé de 12 mesures. Ces vecteurs de caractéristiques vont représenter d'une part les données d'apprentissage qui vont servir à entraîner les classifieurs, et d'autre part les données à classifier pour rechercher des éléments textuels omis par les OCR.

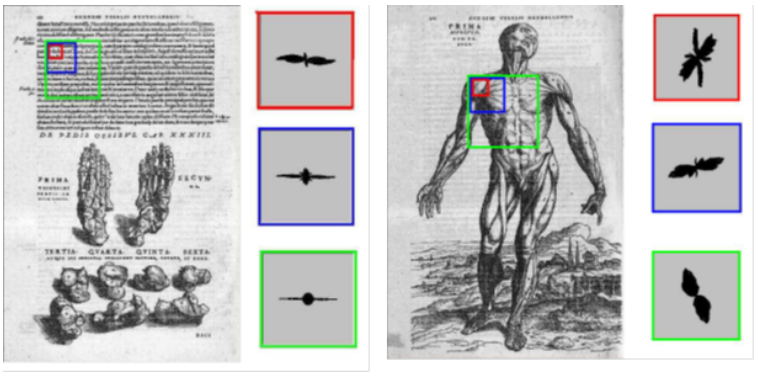


FIGURE 3.4 – Directions des textures des régions textuelles et des régions graphiques à différents échelles selon (Journet, Ramel, Mullot, & Eglin, 2008)

Caractérisation de l'orientation des textures

L'orientation est l'une des caractéristiques visuelles importantes impliquées dans la vision pré-attentive. Cette propriété décrit parfaitement les textures des régions textuelles sans qu'elle soit dépendante de la typographie. Grace à ses propriétés discriminantes, le modèle d'Itti [IK00] utilise l'orientation des éléments textuels pour caractériser les points de saillance dans les images naturelles. Dans [Egl08], les auteurs ont construit un certain nombre de descripteurs de texture qui se basent sur les orientations. Dans notre travail, nous allons nous inspirer de ces descripteurs afin de construire trois descripteurs de texture rendant compte de cette propriété. Le premier descripteur décrit la variation de l'orientation principale de la texture à différentes échelles dans une fenêtre glissante, le deuxième descripteur décrit l'intensité de l'orientation principale à une échelle et le troisième décrit l'importance des différentes orientations de texture obtenues sur une région de l'image.

Dans la suite, nous commencerons cette partie par la présentation de la transformée de Radon qui nous a permis de calculer les caractéristiques d'orientation principale de texture. Ensuite, nous présenterons les différents descripteurs d'orientation que nous avons employés pour décrire les textures de nos images.

(a) Détection de l'orientation principale de la texture

Dans la littérature, plusieurs travaux ont recouru à des filtres directionnels de type Gabor ou ondelette pour détecter l'orientation principale de la texture. Cependant ce type de filtre nécessite le choix du banc de filtres approprié. Pour cela il faut faire l'hypothèse de l'homogénéité des régions textuelles, ce qui n'est pas toujours le cas dans notre contexte.

Notre ligne de conduite étant le développement d'une approche générique qui s'adapte à la totalité des documents de la BnF, nous avons donc écarté les approches de type filtres puisqu'il est difficile de choisir le banc des filtres. De plus, il est parfois nécessaire de binariser les images avant d'appliquer ces descripteurs pour obtenir des descriptions fiables, ce qui peut engendrer des défauts supplémentaires sur les images des pages et une perte considérable d'information.

Ainsi, nous avons choisi d'utiliser une approche non paramétrique qui se base sur l'utilisation de la transformée de Radon pour définir la direction principale de la texture. Le choix de la transformée de Radon est justifié par sa capacité à décrire l'orientation de la texture, la simplicité de sa mise en oeuvre et son indépendance par rapport aux opérations de paramétrages préalables, ce qui répond parfaitement aux exigences du contexte de notre travail. La transformée de Radon est un outil qui permet de tracer l'histogramme de projection des pixels selon des orientations bien déterminées. Selon [COU98], la transformée de Radon est définie par l'équation suivante :

$$f(p,\theta) = \int_{-\infty}^{+\infty} f(p\cos(\theta) - s\sin(\theta), p\sin(\theta) + s\cos(\theta)d_s \tag{3.1}$$

Où θ est l'angle de projection, p est la coordonnée du point P sur l'hyperplan de projection des pixels et s est la coordonnée du point P selon la perpendiculaire à cet hyperplan. d_s les variations élémentaires le long de cette perpendiculaire qui est l'axe d'intégration L (cf. figure 3.5). Par conséquent, à travers la transformée de Radon, on réalise une rotation pour transformer les coordonnées des points du repère cartésien $\Re(O,x,y)$ au repère $\Re_\Theta(O,p,s)$. Ensuite, on intègre l'intensité lumineuse le long de chaque colonne de l'image.

L'application de la transformée de Radon sur l'image I exige l'utilisation de la formulation discrète de la transformée de Radon définie par l'équation suivante :

$$\hat{I}_\theta(p) = \delta_s \sum_{s=-\infty}^{+\infty} I(p\cos(\theta) - s\sin(\theta), p\sin(\theta) + s\cos(\theta)) \tag{3.2}$$

Si de plus les valeurs de p et s vérifient $|p| \geq P_{max}$ et $|s| \geq S_{max}$ alors nous avons :

$$I(p_i,\theta_j) = \delta_s \sum_{s=-N}^{+N} I(p_i\cos(\theta_j) - s_k\sin(\theta_j), p_i\sin(\theta_j) + s_k\cos(\theta_j)) \tag{3.3}$$

Avec :

$$N\delta_s = S_{max}$$
$$i = 0, ...N_p; N_p\delta_p = P_{max}$$
$$j = 0.....N_\theta; \theta_{N\theta} = \pi; \theta_0 = 0$$

Notre approche détermine donc la direction principale de la texture des éléments textuels de la page en appliquant la transformée de Radon sur des parties de l'image I délimitées par une fenêtre glissante centrée sur le pixel traité. Ceci permet d'assigner à chaque pixel de l'image une signature des orientations principales de la texture.

Dans notre étude, pour nous adapter aux différentes tailles de police de caractères, nous avons appliqué trois fenêtres glissantes de taille $128 \times 128 (k = 1)$, $64 \times 64 (k = 2)$ et $32 \times 32 (k = 3)$. De plus, nous avons examiné à chaque échelle sept orientations $\Theta \in$

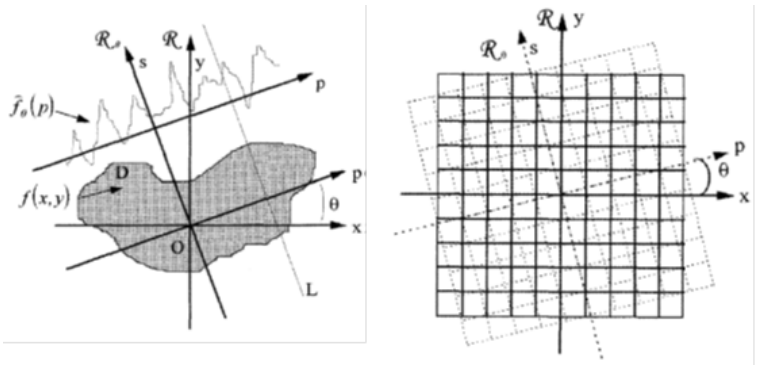
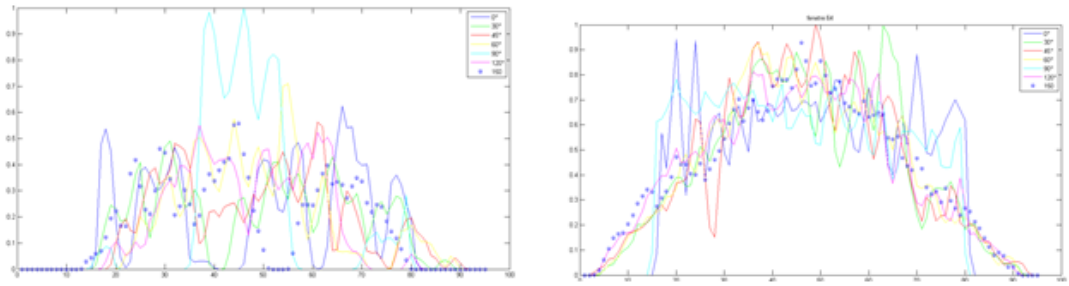


FIGURE 3.5 – Représentation de la transformée de Radon. L'image à gauche représente la projection $f_\Theta(p)$ de l'image définie dans le repère en rotation.



(a) Texte (b) Graphique

FIGURE 3.6 – Transformée de Radon sur une fenêtre de taille 64×64

$\{0°, 30°, 45°, 60°, 90°, 120°, 150°\}$ de texture. Les résultats de la transformée de Radon d'une région textuelle et d'une région graphique sont présentés dans les figures 3.6a et 3.6b. D'après ces figures, nous constatons la présence d'une orientation dominante dans les réponses de la transformée de Radon sur une zone de texte et de plusieurs orientations à des intensités importantes dans les réponses de la transformée de Radon sur une zone graphique.

La première mesure que nous proposons à partir de la transformée de Radon permet d'identifier la direction principale de la texture à chaque échelle. Cette mesure est définie par l'équation 3.4 qui détermine l'indice de la direction principale en utilisant les histogrammes des orientations de texture (cf. figure 3.6). Ceci donne un indice d'orientation dominante par échelle et pour chaque position de la fenêtre glissante. Dans notre approche, les indices des orientations de texture varient de 1 à 9. L'indice « 1 » référence l'orientation verticale « 90° » alors que l'indice « 5 » référence l'orientation horizontale « 0° ». Les 7 premiers indices référencent respectivement les orientations suivantes { 0°, 30°, 45°, 60°, 90°, 120°, 150° alors que les indices « 8 » et « 9 » ne référencent aucune orientation puisqu'ils sont utilisés respectivement dans le cas où nous avons deux orientations principales et dans le cas où nous n'avons aucune orientation de texture (région de fond).

$$\Gamma_k(i,j) = argmax_\Theta(max(R_k(\Theta))) \tag{3.4}$$

Avec $\theta \in \{0, 30°, 45°, 60°, 90°, 120°, 150°\}$ et $k = 1.....3$

Les réponses de ce descripteur obtenues sur trois fenêtres glissantes (32×32, 64×64 et

128×128) sont présentées sur la figure 3.7. D'après ces représentations, nous constatons que l'orientation principale de la texture est majoritairement horizontale sur les zones textuelles (couleur vert) ce qui est conforme avec l'orientation du texte de ces documents.

Les orientations de la texture dans la zone de lettrine et d'illustration sont multiples et très variables, ce qui signifie qu'il n'existe pas d'orientation dominante dans les textures de ces zones. De plus, à travers les réponses de ce descripteur, on constate qu'il est possible de caractériser les zones de fond en deux classes : la classe des espaces entre mots et la classe des marges des pages. Les orientations de la texture des régions espaces inter-mots sont très variables alors qu'elles sont majoritairement égales à « 8 » ou à « 9 » sur les zones de marges. Par conséquent, nous pouvons déduire que la caractéristique d'orientation de la texture est discriminante pour séparer les différents éléments qui nous intéressent dans la page.

Par contre, le défaut principal de ce descripteur réside dans les imprécision obtenues par l'utilisation d'une fenêtre de grande taille qui homogénéise la description des zones textuelles avec les espaces inter-mots. Ceci peut causer des erreurs de fusion entre les mots omis et par conséquent l'incapacité de calcul des mots omis. En effet, d'après les exemples de la figure 3.7 obtenus avec les configurations de 128×128 et 64×64, les orientations principales assignées aux pixels des espaces entre les mots sont les mêmes que les orientations obtenues sur les pixels des éléments textuels.

(b) Consensus des orientations principales à différentes échelles

Les textures des régions textuelles se caractérisent traditionnellement par une orientation horizontale. Cependant dans certains documents particuliers (par exemple : les poèmes, les textes artistiques, les documents techniques) l'orientation du texte peut changer au sein du même document. Donc l'orientation du texte ne suffit pas à elle seule pour caractériser les régions textuelles.

Pour résoudre ce problème, nous avons exploité une autre caractéristique qui rend compte de la stabilité de l'orientation principale de la texture dans les trois échelles. En effet, d'après la figure 3.7, les pixels inclus dans les régions textuelles conservent le même indice d'orientation à différentes échelles. Par conséquent, pour caractériser les textures de ces régions, nous avons examiné pour chaque pixel les indices de la direction principale de la texture obtenue à chaque échelle. La formule 3.5 représente le descripteur des variations inter-échelle de l'orientation principale de texture. Le descripteur f_1 prend la valeur 2 si la direction principale est la même dans les trois fenêtres glissantes, la valeur 1 si la direction principale est la même dans au moins deux fenêtres glissantes, et la valeur nulle si les directions principales sont différentes d'une fenêtre à une autre (cf. équation 3.5)

$$f_1(i,j) = (\Gamma_1(i,j) == \Gamma_2(i,j)) + (\Gamma_2(i,j) == \Gamma_3(i,j)) \tag{3.5}$$

En prenant en compte les hypothèses que nous avons posées dans l'introduction de cette partie, les réponses de ce descripteur doivent varier entre 0 et 1 sur les régions d'illustration et puisqu'il y a au plus deux fenêtres glissantes qui présentent la même orientation principale. Par contre, elles varient entre 1 et 2 sur les régions textuelles puisque au moins deux fenêtres glissantes dans ces régions doivent présenter le même indice. La figure 3.8 confirme ces déductions. En effet, d'après ces exemples, nous remarquons que les régions textuelles obtiennent généralement un consensus d'orientations principales de

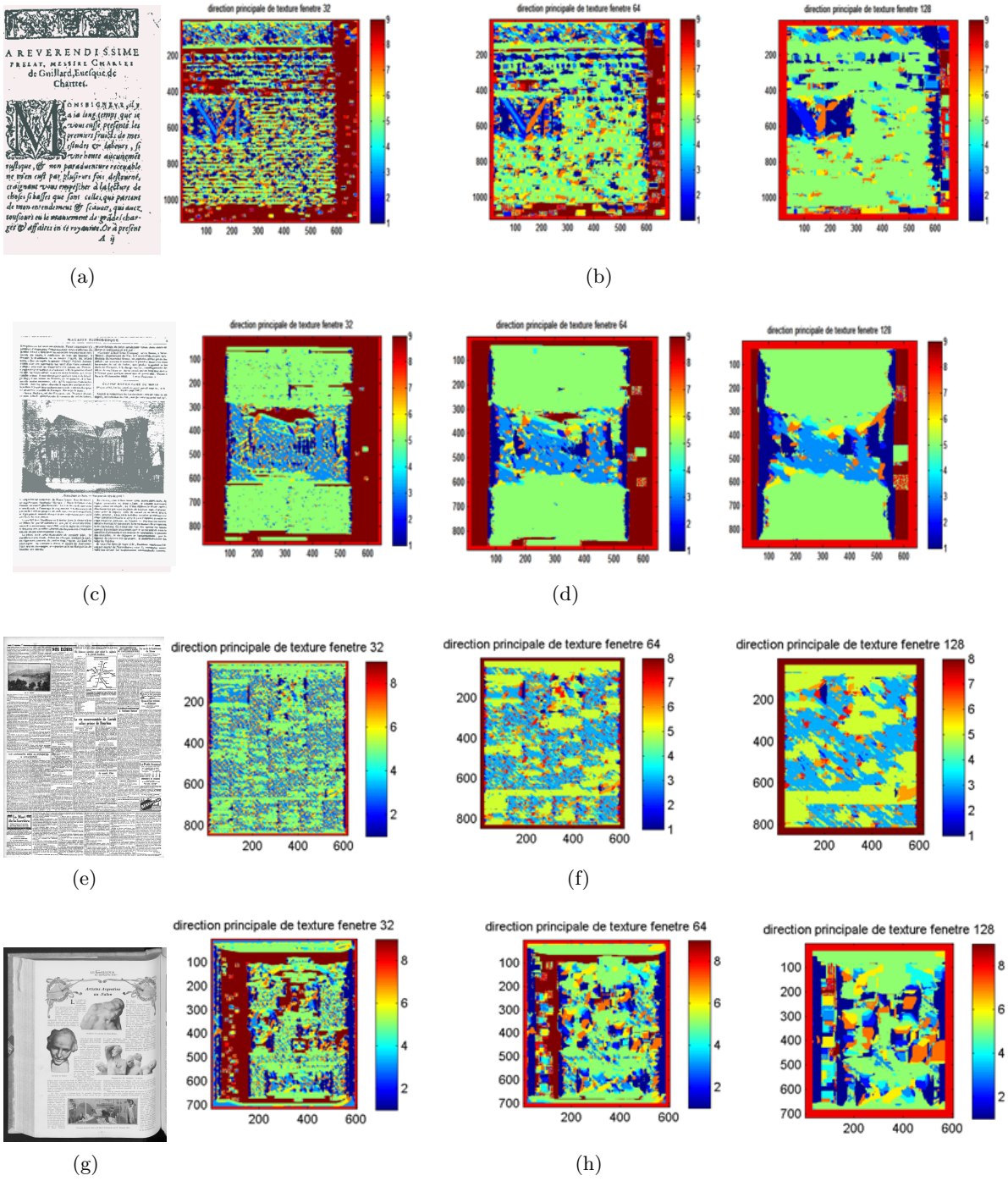


FIGURE 3.7 – Réponses de descripteur des orientations principales de textures dans les trois échelles de fenêtre glissante que nous avons utilisées.

texture obtenues lors de l'utilisation des trois fenêtres glissantes (des valeurs de 2 sont assignées dans les régions textuelles)

Les orientations des textures des régions graphiques sont très variables sur les exemples de la figure 3.8b. En effet, nous remarquons que les indices de consensus des orientations assignés aux pixels d'illustration sont variables entre 0 et 1.

Cependant, les pixels de fond dans les marges de la page ont des réponses de descripteur f_1 semblables à ceux qui se trouvent dans les régions textuelles. Selon la figure 3.8b, nous constatons que les cas de consensus des orientations (réponse égale à 2) sont même plus fréquents sur les régions de marge que sur les régions textuelles. Ceci parait contradictoire avec ce que nous avons supposé dans l'introduction de cette partie par contre en revenant à la définition de ce descripteur ce résultat est logique. En effet, d'après les exemples de la figure 3.7, nous remarquons que les indices d'orientations principales obtenus sur ces régions sont majoritairement égaux à « 8 » ou à « 9 ». Et même en changeant l'échelle de fenêtre glissante ces réponses restent les mêmes ce qui entraine des cas de consensus d'orientations principales de texture et par conséquent des réponses semblables à ceux qui ont été obtenus sur les éléments textuels.

Nous trouvons aussi des régions de pixels de fond qui englobent des réponses de descripteur f_1 égal à « 1 ». Selon la figure 3.8b, ces cas de figure sont obtenus surtout sur les zones de fond qui sont voisines soit des régions textuelles de la page, soit des bords de la page qui sont labélisés par des indices d'orientation égaux « 8 ». Par conséquent, nous pouvons déduire que les réponses de ce descripteur sont confuses.

Pour résoudre ce problème, nous avons utilisé l'information de l'orientation principale des textures pour pondérer les réponses de ce descripteur. En fait, le problème majeur de ce descripteur réside dans la similarité de réponses obtenues sur les éléments textuels et les régions d'arrière-plan. Or en regardant les images des exemples de la figure 3.7, nous remarquons que les pixels d'arrière-plan qui sont situés dans les zones des marges de la page sont caractérisés par des indices d'orientation de « 8 » et de « 9 ». Par conséquent, nous pouvons pondérer les réponses de ce descripteur en vérifiant le cas où nous avons un consensus sur les orientations principales d'indice « 8 » et « 9 ». Cette procédure de pondération est définie comme suivant :

- On pondère la réponse du descripteur de consensus d'orientation par un poids nul si les indices d'orientation principale de texture obtenus dans les trois échelles étudiées sont égaux soit à « 8 » ou à « 9 ».
- On pondère la réponse du descripteur de consensus d'orientation par un poids de « 0,5 » si les indices d'orientation principale de texture obtenus sont égaux soit à « 8 » ou à « 9 » dans au plus de deux échelles étudiées .
- On pondère la réponse du descripteur de consensus d'orientation par un poids de « 1 » si les indices d'orientation principale de texture obtenus dans les trois échelles étudiées sont différents de « 8 » et de « 9 ».

L'application de cette procédure de pondération donne les résultats illustrés dans la figure 3.8c. D'après les exemples de cette figure, nous constatons que les descriptions des pixels des marges sont différentes des descriptions des pixels des régions textuelles. Par contre, les descriptions des espaces inter-mots ainsi que les pixels des régions d'illustration sont difficiles à utiliser. Par conséquent à partir de cette analyse, nous pouvons déduire que les réponses de ce descripteur sont insuffisantes pour séparer les classes d'éléments de

la page. D'où le besoin d'enrichissement de cette description par d'autres caractéristiques.

(c) Détection de l'intensité moyenne des orientations

Une autre caractéristique riche en information est l'intensité médiane des directions de la texture qui donne une indication sur le degré d'encrage des régions de la page. Par définition, la transformée de Radon donne une information sur la projection des niveaux de gris des pixels selon une orientation θ. Dans notre cas, le calcul de la transformée de Radon a permis d'étudier l'association des niveaux de gris de l'encre selon des directions précises. Mais on peut également noter que la valeur médiane des intensités de la transformée de Radon est plus importante sur des régions d'illustration que sur des régions d'écriture ou d'arrière-plan. Le descripteur des intensités médianes de la texture est défini par l'équation 3.6

$$f_{(k+1)}(i,j) = mediane(max(R_k(\theta))) \tag{3.6}$$

Avec $\theta \in \{0°, 30°, 45°, 60°, 90°, 120°, 150°\}$ et $k = 1.....3$. Pour effectuer un traitement multi-résolution, nous avons appliqué ce descripteur sur les trois fenêtres glissantes.

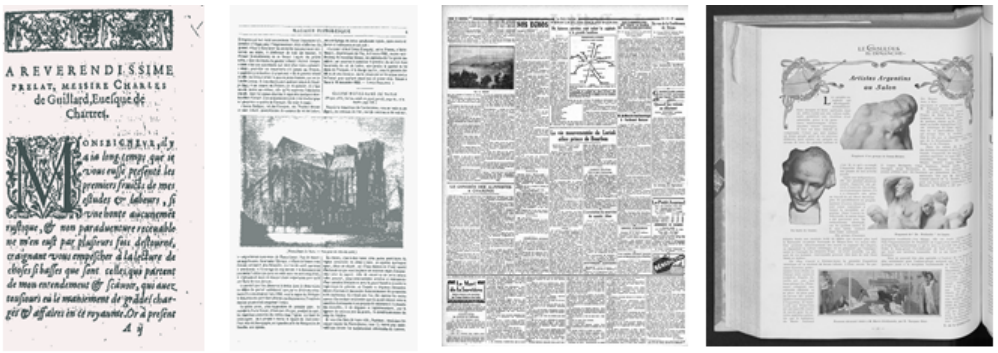
Les réponses de ce descripteur sur les exemples de la figure 3.2 sont présentées dans la figure 3.9. D'après ces représentations, nous remarquons bien que les textures de texte ont des intensités médianes d'orientations principales moins importantes que celles des textures d'illustration. Par conséquent, nous devrions pouvoir séparer les régions d'illustration des régions textuelles en utilisant les réponses de ce descripteur.

Le comportement de ce descripteur est le même sur les différentes échelles d'analyse ce qui renforce son pouvoir de discrimination. En effet, selon la figure 3.9, les pixels du fond qui se trouvent hors de la zone imprimée (c'est-à dire dans les marges de la page) sont caractérisés par des intensités médianes d'orientations principales très proches de zéro. L'analyse multi-résolution est donc utile pour ce descripteur afin d'adapter ses réponses à la variabilité des polices de caractères que l'on peut retrouver dans les collections de la BnF. Cependant, ce traitement peut causer des problèmes semblables à ceux qui ont été obtenus avec le descripteur précédent. En effet, selon la figure 3.9,nous remarquons que les descriptions des espaces inter-mots sont semblables aux descriptions des éléments textuelles obtenus avec des fenêtres glissantes de taille 64×64 et 32×32. Par contre, on constate que ce défaut apparaît moins dans les résultats obtenus avec la fenêtre glissante de taille 32×32. Par conséquent, l'utilisation de plusieurs échelles de fenêtre glissante permet de décrire les textures de l'image de manière locale et globale, ce qui permet de séparer les différents éléments de la page que nous souhaitons détecter.

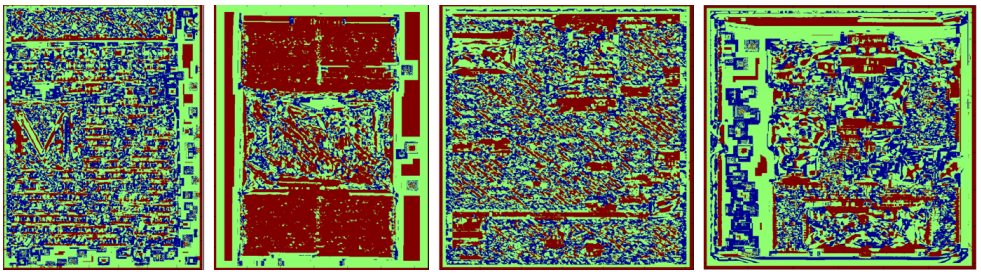
En conclusion, même si les descriptions de tous les éléments de la page semblent séparables avec ce descripteur, les régions graphiques sont nettement différente des autres classes pour les différentes échelles que nous avons considérées. Ce qui signifie que ce descripteur est très pertinent pour décrire les éléments graphiques de la page.

(d) Variation des orientations de la texture

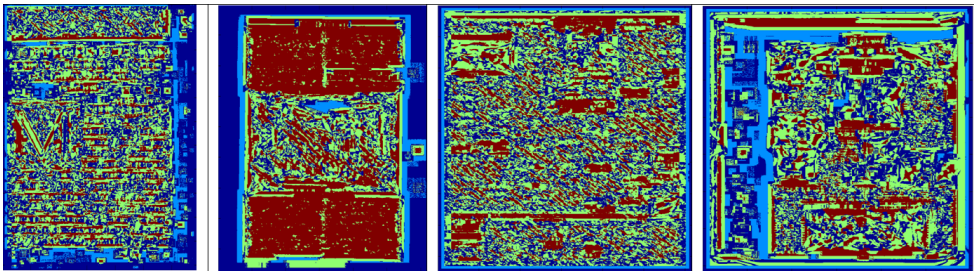
La variance des orientations de la texture représente aussi une information pertinente. En effet, selon les figures 3.6a, 3.6b la réponse de la transformée de Radon sur les éléments textuels présente généralement une seule orientation dominante, alors que sur les régions d'illustration nous remarquons que toutes les orientations sont importantes. Par conséquent, la variance des orientations de texture est plus importante sur les régions de



(a) Image originale

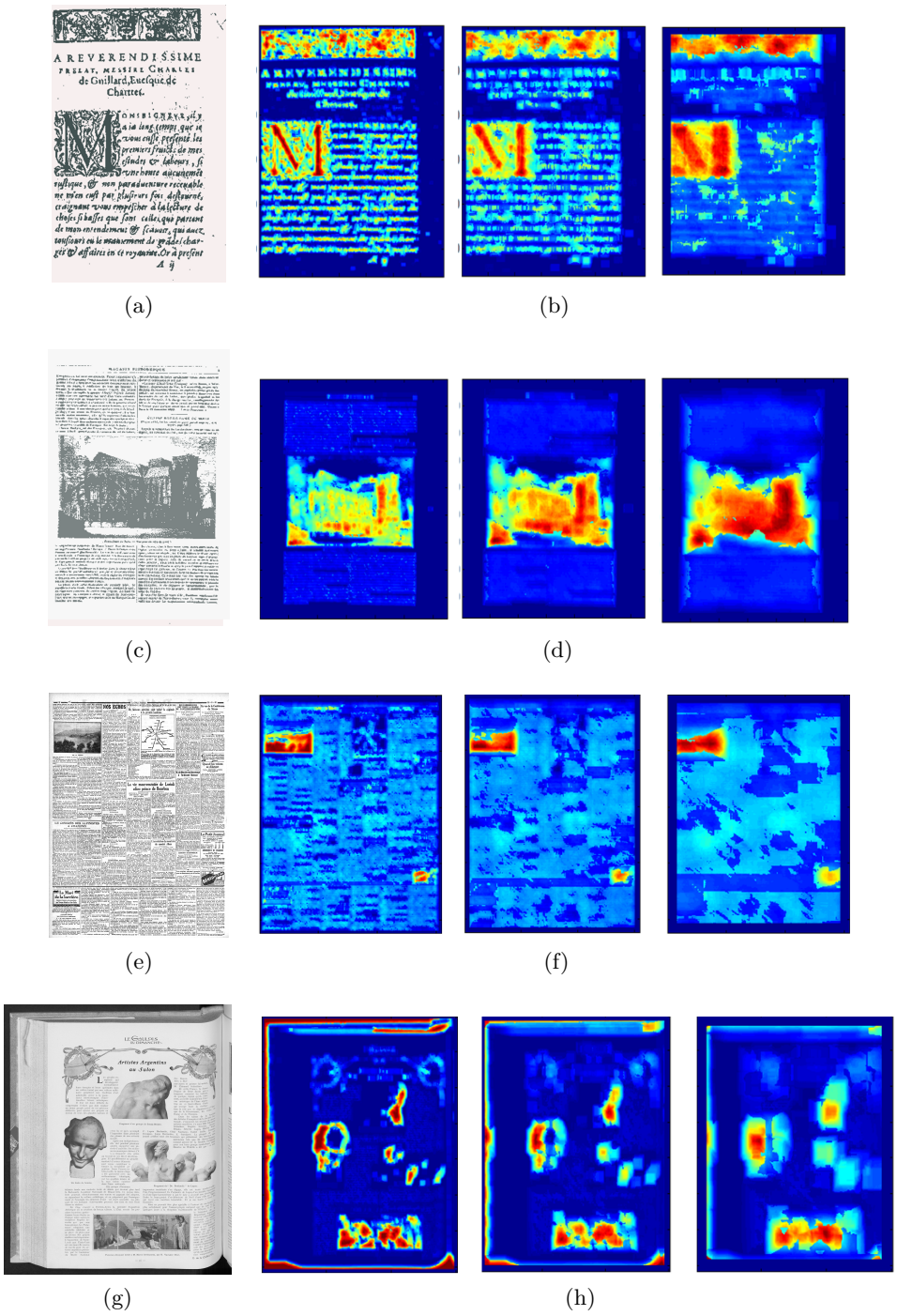


(b) Avant la pondération des orientation



(c) Aprés la pondération des orientation

FIGURE 3.8 – Réponses de descripteur f_1 qui vérifie les orientations principales des textures à différentes échelles. (a) représente les résultats de ce descripteur avant l'application de la procédure de pondération (les pixels ayant aucun consensus ($f_1(i,j) = 0$) de direction principale sont représentés par la couleur bleu, les pixels ayant au moins deux directions semblables ($f_1(i,j) = 1$) sont représentés par la couleur verte et les pixels ayant un consensus de direction principale ($f_1(i,j) = 2$) sont représentés par la couleur rouge), (b) représente les résultats de ce descripteur après l'application la procédure de pondération (0 est représenté par la couleur bleu, 0,6 est représenté par la couleur bleu ciel, 1 est représenté par la couleur verte et 2 est représenté par la couleur rouge).



(a) (b) (c) (d) (e) (f) (g) (h)

FIGURE 3.9 – Réponses de descripteur $f_{(k+1)}$ qui utilise l'intensité médiane des orientations de texture pour décrire les régions de la page.

texte que sur les régions graphiques.

Nous allons donc exploiter cette caractéristique pour décrire les textures de la page. Pour cela, nous avons développé un quatrième descripteur défini par l'équation 3.7 :

$$f_{(k+4)}(i,j) = std(max(R_k(\theta))) \tag{3.7}$$

Avec $\theta \in \{0°, 30°, 45°, 60°, 90°, 120°, 150°\}$ et $k = 1.....3$. Ce descripteur analyse donc les valeurs maximales des orientations de la texture obtenues dans une fenêtre glissante pour calculer leur variance. L'application de ce descripteur sur les images de la figure 3.1a donne les résultats représentés sur la figure 3.10. Les couleurs chaudes (proches du rouge) définissent les valeurs maximales de la variance, alors que les couleurs froides (proches du bleu) sont employées pour représenter les faibles valeurs de variance. Selon la figure 3.10, les régions d'illustration et de lettrine sont caractérisées par de faibles réponses alors que les régions textuelles présentent de fortes réponses.

D'autre part, grâce à ce descripteur, nous pouvons aussi distinguer les espaces inter-mots des éléments textuels de la page. En effet, en utilisant les réponses de ce descripteur obtenues avec une fenêtre glissante de taille 32×32, la variance des intensités d'orientations est plus faible sur les espaces inter-mots que sur les éléments textuels (cf. figure 3.10). Par contre, nous remarquons l'existence d'un problème de séparabilité lorsqu'on utilise les fenêtres glissantes de taille 64×64 et 32×32.

Les régions de fond qui se trouvent dans les marges présentent des réponses différentes de celles des espaces entre les mots. En effet, puisque les intensités d'orientations de la texture sont peu variables dans ces régions, leur variance est aussi très faible ce qui caractérise ces régions par rapport aux autres régions de la page.

Par conséquent à partir de cette analyse, nous pouvons déduire que la variance des orientations de la texture est pertinente pour séparer les régions d'illustration, de texte et du fond.

Caractérisation de la régularité de la texture

Plusieurs études ont été proposées dans la littérature [OPM02], [LSK+12] pour caractériser la régularité des textures. Parmi ces méthodes, nous trouvons les motifs binaires locaux (*Local Binary Patterns* « LBP ») [Mäe03].Ce descripteur permet de décrire la régularité d'une texture dans une fenêtre circulaire de rayon R et comportant P points de mesure. Grâce à ce descripteur, on peut définir une texture T dans un voisinage local par la distribution spatiale des niveaux de gris [AGP10]

$$T = t(g_c, g_0, ..., g_{P-1}) \tag{3.8}$$

Avec g_c le niveau de gris du pixel central et g_p $(P = 0, ..., P-1)$ les P points équidistants sur le cercle de rayon R centré sur le pixel central. Les coordonnées des P pixels sont déterminées par $(x_c + R \times \cos(\frac{2\pi p}{P}), y_c - R \times \sin(\frac{2\pi p}{P}))$ ce qui localise les voisins sur le cercle de centre (x_c, y_c) et de rayon.

La figure 3.11 représente des masques LBP exprimés avec des paramètres différents. D'après ces masques, les points de voisinage considérés par ce masque ont des coordonnées polaires par rapport au pixel central. Très souvent, les points du voisinage se retrouvent à

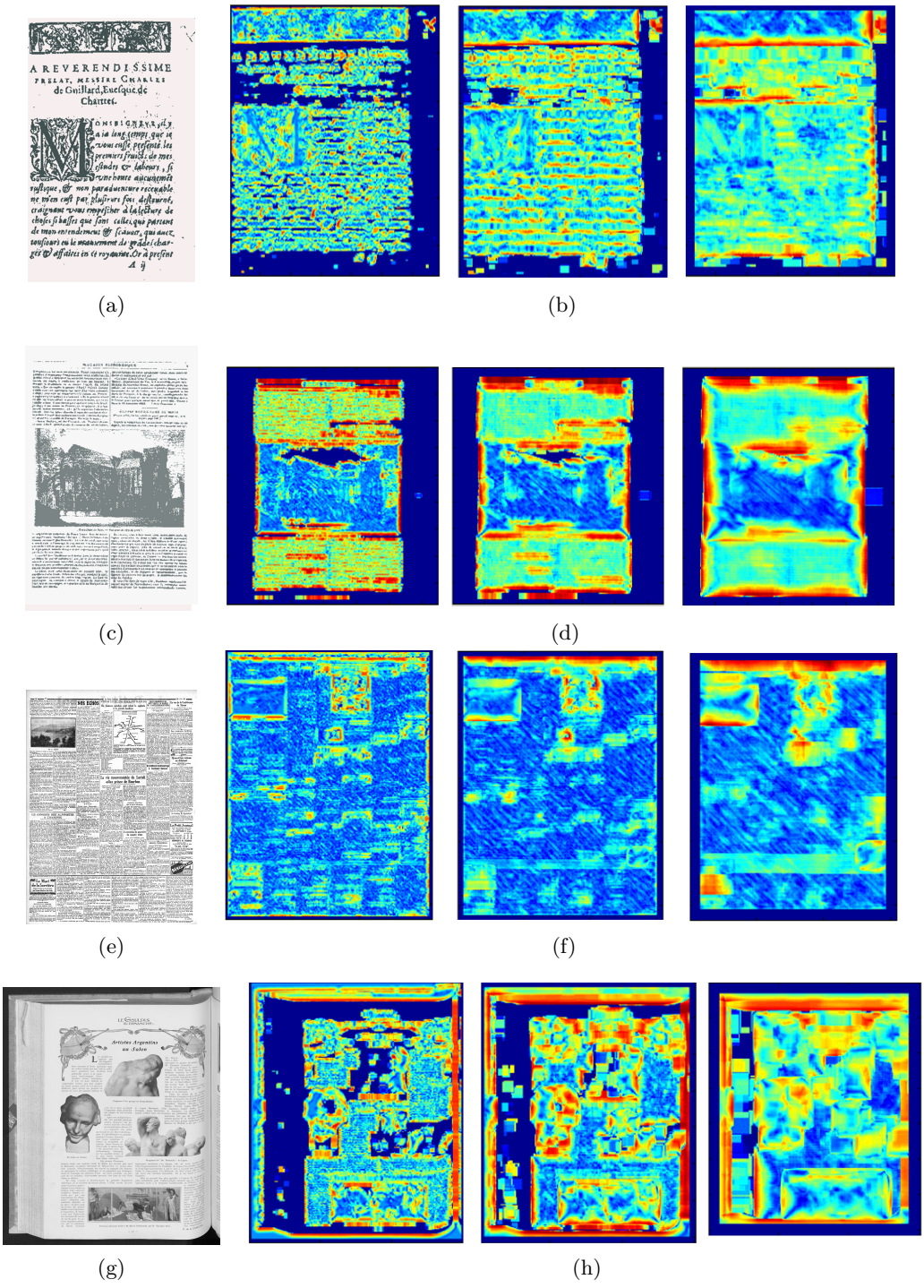


FIGURE 3.10 – Réponses de descripteur $f_{(k+4)}$ qui référence la variance des orientations des textures pour décrire les textures de la page. Les premières images de chaque exemple représentent les réponses obtenues avec une fenêtre glissante de taille 32×32, les deuxième images de chaque exemple représentent les réponses obtenues avec une fenêtre glissante de taille 64×64 et les troisièmes images représentent les réponses obtenues avec une fenêtre glissante de taille 128×128.

l'intersection de plusieurs pixels de l'image (cf. figure 3.11 configurations « $p = 12, p = 2.5$ » et « $p = 16, p = 4$ »). Dans ce cas, on associe à ces points des intensités de niveaux de gris égales à la moyenne pondérée des intensités des pixels voisins du point P considéré. Les

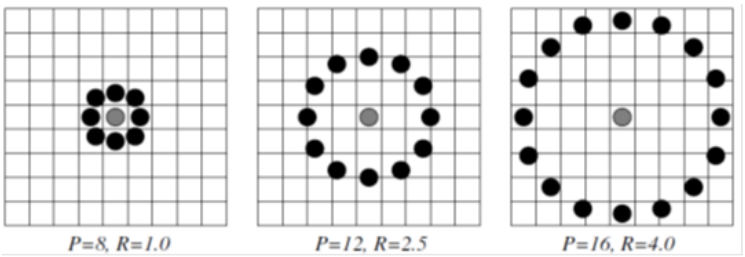


FIGURE 3.11 – L'ensemble des voisins circulaire et symétrique. Les échantillons qui ne correspondent pas à la grille des pixels sont corrigés par interpolation.

intensités des pixels voisins g_p sont ensuite soustraites de la valeur de l'intensité du pixel central g_c. Les P différences obtenues donnent une description de la distribution spatiale des niveaux de gris, et permettent donc de caractériser la texture analysée.

L'utilisation de la différence brute des intensités de pixel rend les réponses de ce descripteur variables aux défauts d'acquisition des niveaux de gris. Pour remédier à ce défaut et rendre les réponses des LBP invariantes à ce défaut, les auteurs [Mäe03] proposent d'utiliser les signes des différences pour caractériser les textures de la page au lieu de leurs valeurs absolue.

$$T = t(s(g_0 - g_c), ..., s(g_{P-1} - g_c)) \tag{3.9}$$

où s est l'opérateur LBP défini par la formule suivante :

$$s(x) \begin{cases} 1 & g_p - g_c \geq 0 \\ 0 & g_p - g_c < 0 \end{cases} \tag{3.10}$$

Chaque signe est affecté d'un poids égal à 2^p, ce qui permet de transformer les différences des intensités des pixels voisins en un code LBP qui décrit de manière unique chaque situation possible dans le voisinage considéré. Ce code caractérise donc la texture locale de l'image dans le voisinage du point (x_c, y_c) :

$$LBP_{P,R}(x_c, y_c) = \sum_{p=0}^{p-1} s(g_p - g_c)2^p \tag{3.11}$$

Selon cet opérateur, la valeur maximale des réponses des descripteurs est obtenue lorsque toutes les intensités des pixels voisins concernés par le masque LBP sont supérieures ou égales à l'intensité du pixel central. D'autre part, les réponses de ce descripteur deviennent minimales lorsqu'aucune intensité des pixels voisins n'est égale à l'intensité du pixel central. Par conséquent, nous pouvons déduire que les réponses de ce descripteur sont très dépendantes du nombre des voisins considérés par le masque de LBP.

En se basant sur ce principe, les zones uniformes plates ou peu variables (illustrations ou fond de page) sont caractérisées par des réponses LBP maximales. tandis que, les régions

textuelles caractérisées par de fortes variations de niveaux de gris sont décrites par des réponses assez variables.

Généralement, le descripteur LBP utilise ses réponses pour décrire ensuite les propriétés statistiques de la texture en plus de la description structurelle qui est obtenue localement par le descripteur LBP. Ainsi chaque code LBP peut être considéré comme un micro-texton. Les primitives locales détectées par le LBP correspondent à des motifs locaux qui illustrent des formes de tache, des régions planes, des contours, des coins, etc. Quelques exemples de motifs locaux obtenus avec une fenêtre LBP composée de huit points « $LBP_{8,R}$ » sont présentés dans la figure 3.13. Dans cette figure, les cercles blancs correspondent aux valeurs à 1 de la signature de texture alors que les valeurs nulles sont représentées par des cercles noirs.

L'analyse des motifs locaux, localisés par le descripteur LBP, peut être effectuée soit avec des approches structurelles basées sur l'étude de la distribution de répartition des micros-textons dans l'image, soit avec des approches statistiques basées sur l'examen de la répartition des étiquettes des micros-textons à travers des statistiques calculées sur une fenêtre glissante. Bien que les caractéristiques de texture des régions textuelles soient

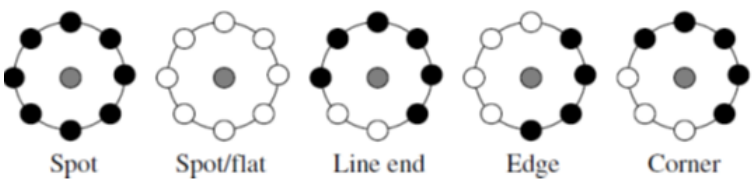


FIGURE 3.12 – Exemples de motifs locaux détectés par le descripteur LBP

bien différentes des textures des autres éléments de la page, leur régularité n'est pas assez prononcée dans les pages. Ceci rend l'utilisation des descriptions structurelles pour caractériser les régions de la page peu efficace. D'autre part, l'analyse statistique des répartitions des étiquettes des micro-textons peut être coûteuse en termes de temps de calcul pour décrire les régions de la page. Par conséquent, au lieu d'utiliser ces deux approches pour caractériser les textures de la page, nous avons décidé d'exploiter directement les réponses de l'équation 3.11 pour former les vecteurs des caractéristiques des textures de la page.

Dans notre approche, l'utilisation de decsripteur LBP permet de différencier les régions du fond ou d'illustrations caractérisées par des réponses LBP maximales, des régions textuelles caractérisées par des réponses LBP moins importantes.

Les étapes de calcul des caractéristiques de LBP sont les suivantes : (1) Au début nous commençons par le parcours de tous les pixels de l'image qui se trouvent en dehors des zones de bord de l'image pour les caractériser par rapport à leurs voisinages. (2) Pour chaque pixel nous calculons les cordonnées des points voisins ainsi que leurs intensités. Ensuite, (3) nous appliquons l'équation 3.11 pour obtenir la forme du modèle LBP de ce pixel. Enfin, (4) nous associons à chaque pixel de l'image la valeur de réponse du descripteur de LBP associée au modèle obtenu.

Pour caractériser plusieurs tailles de police de caractères dans l'image de la page, nous

avons analysé la répartition des motifs binaires à travers l'utilisation de trois configurations de masque LBP ($k' = 8(P = 16, R = 5)$, $k' = 9(P = 32, R = 10)$ et $k' = 10(P = 64, R = 20)$) qui nous a permis de calculer les réponses de ce descripteur à différentes échelles. Ce traitement permet d'adapter les descriptions aux variations de tailles de caractères.

Les résultats de l'application de ce descripteur sont présentés dans la figure 3.13. Le code de couleur de cette figure référence l'intensité des réponses du descripteur LBP. Les couleurs chaudes représentent les réponses maximales alors que les réponses minimales sont représentées par des couleurs froides. D'après la figure 3.13, nous constatons que les pixels de fond sont caractérisés par de fortes réponses du descripteur LBP, alors que dans les régions textuelles, nous remarquons une faible variation dans les intensités des réponses LBP. Les faibles intensités sont obtenues dans les régions de contours puisque les intensités des voisins sont très différentes de celle du pixel central.

Le problème de ce descripteur réside surtout dans sa sensibilité par rapport au bruit de numérisation ce qui nécessite l'adoption d'une procédure de filtrage de bruit avant l'application de ce descripteur. Dans notre approche, nous avons utilisé un filtre médian qui applique sur l'image un élément structurant de taille 3×3. Ceci permet d'éliminer tout le bruit de numérisation qui peut biaiser les réponses de ce descripteur.

Caractérisation de l'intensité des pixels

(a) Caractérisation de la fréquence des transitions encre/fond

En complément des informations liées aux orientations et à la régularité des textures, nous avons extrait des informations liées au contraste dans l'image. On peut caractériser la notion de contraste dans les images de documents par la probabilité élevée d'un passage d'un pixel foncé (encre) à un pixel clair (papier).

Cependant, certaines caractéristiques particulières liées aux méthodes d'impression peuvent engendrer des ambiguïtés. Par exemple, certains documents anciens englobent des illustrations au trait qui sont caractérisées aussi par de fortes transitions entre des pixels foncés et des pixels clairs. Par conséquent, la description de la fréquence de transition des pixels foncés aux pixels clairs obtenue sur les textures de ces éléments sera semblable à celle des textures des régions textuelles. Cela signifie qu'on ne peut pas utiliser la simple variance des intensités des pixels de l'image pour décrire les fréquences de transition des intensités de la page.

Plusieurs travaux dans la littérature [[EBE98], [All03], [CWS03]] ont été proposés pour caractériser les textures de la page en se basant sur la fréquence de transition entre les pixels clairs et les pixels foncés. Ces travaux présentent des descripteurs dédiés à l'analyse des polices et des styles de caractères particuliers ce qui rend difficile leur application sur des collections documentaires variables. De plus, la plupart des méthodes proposées dans ces travaux utilisent des images binaires, ce qui nous oblige à réaliser une binarisation qui peut causer des défauts supplémentaires et une perte considérable d'information.

Afin de caractériser les fréquences de transition, nous avons adopté un descripteur de texture inspiré de celui qui a été proposé dans [Egl08]. Dans ce travail, les auteurs ont utilisé des descripteurs non-paramétriques qui utilisent les propriétés des transitions de niveaux de gris pour caractériser différentes textures de la page et séparer le texte des illustrations.

(a) (b)

(c) (d)

(e) (f)

(g) (h)

FIGURE 3.13 – Résultats de l'application des descripteurs LBP employés avec trois configurations de masque différentes ($k' = 8(P = 16, R = 5)$, $k' = 9(P = 32, R = 10)$ et $k' = 10(P = 64, R = 20)$)

Dans notre travail, nous avons calculé un indice de transition entre les pixels d'encre et les pixels de papier en utilisant une fenêtre glissante de taille $n \times n$. Pour valider la taille de la fenêtre glissante, nous avons testé trois configurations différentes. Ensuite, pour chaque ligne de la zone analysée par la fenêtre glissante, nous avons calculé la somme des différences des niveaux de gris d'un pixel et de son voisin de gauche. Enfin, nous avons effectué la moyenne de ces sommes de différences pour mesurer la valeur moyenne des variations de niveaux de gris. Plus la valeur de la moyenne des différences est importante, plus la fréquence de passage entre les pixels clairs et les pixels foncés est importante, ce qui indique que nous traitons une région textuelle. Par contre, les faibles valeurs moyennes des différences des niveaux de gris signifient qu'il y a peu de transition entre les pixels foncés et clairs, ce qui arrive généralement sur des régions homogènes (papier ou illustration). L'équation 3.12 définit le descripteur fréquentiel de notre approche :

$$f_{11} = Avg_{i \in I'}(\sum_{j \in J'}(p_{ij} - p_{ij+1})) \tag{3.12}$$

Avec I' et J' la taille de la fenêtre d'analyse de p_{ij} le niveau de gris du pixel de coordonnées (i, j).

La figure 3.14 présente les résultats de l'application du descripteur 3.12 sur les exemples de la figure 3.2. Ces résultats sont obtenus à trois échelles différentes. Le code de couleur de ces exemples montre l'importance des intensités des réponses de ce descripteur où les faibles intensités sont représentées par des couleurs proches du bleu alors que les fortes réponses sont caractérisées par des couleurs proches du rouge. D'après ces représentations, nous remarquons que les régions de fond sont bien caractérisées par de faibles valeurs de transition entre les pixels foncés et les pixels clairs. Par contre, les régions textuelles sont caractérisées par de fortes réponses.

Les contours des éléments graphiques sont caractérisés par de fortes variations de niveau de gris ce qui entraîne des réponses importantes de ce descripteur sur ces régions. Par contre, les régions homogènes des éléments graphiques sont caractérisées par de faibles fréquences de transition. Ceci caractérise les régions d'illustration par rapport aux autres régions de la page.

Les résultats obtenus dans la figure 3.14 sont obtenus avec des fenêtres glissantes de taille respective égales à la hauteur de la médiane des boîtes englobantes des mots detectés par l'OCR ou de taille 7×7 ou 9×9. D'après ces exemples, nous constatons que plus la taille de la fenêtre est petite plus les réponses de ce descripteur sont significatives sur les régions textuelles de la page. Nous remarquons cependant que les intensités moyennes des sommes des différences sont très importantes sur les régions textuelles de l'exemple 3.14b alors qu'elles sont plus faibles sur les autres exemples. Ceci est dû à la faible taille de la police de caractères dans les documents à plusieurs colonnes et aux intensités des pixels textuels dans l'exemple 3.14b. Ce dernier accentue les fréquences de transitions obtenues sur les régions textuelles. En conclusion, pour appliquer ce descripteur sur les images de documents de la BnF, nous avons utilisé une fenêtre glissante de taille 7×7 pour calculer les différences moyennes des niveaux de gris consécutifs. Finalement, les résultats de cette opération sont présentés dans la dernière colonne de la figure 3.14.

(b) Caractérisation des intensités des pixels

Les pixels foncés appartiennent généralement à des régions encrées (textes et illustrations) alors que les pixels clairs appartiennent à des régions d'arrière-plan (papier).

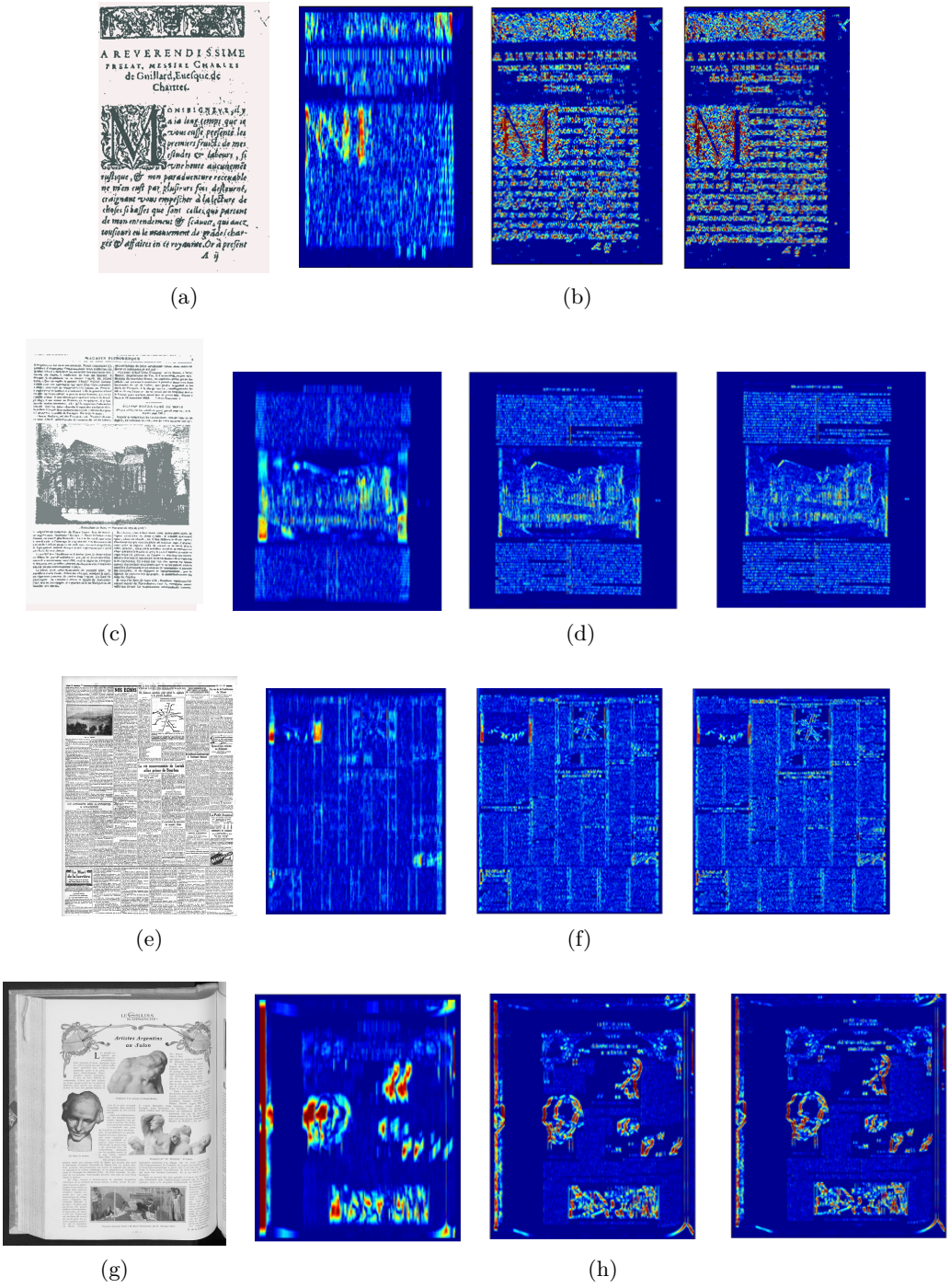


FIGURE 3.14 – Résultats de l'application du descripteur des fréquences de transition entre les pixels clairs et foncé. La première image de chaque exemple représente les résultats obtenus lors de l'utilisation d'une fenêtre glissante de taille $median(hauteur)_{mot} \times median(hauteur)_{mot}$, les deuxièmes images de chaque exemple sont obtenues en utilisant une fenêtre glissante de taille 9×9 et les troisièmes images sont obtenues en utilisant une fenêtre glissante de taille 7×7.

Cependant, certains artefacts issus de l'opération de capture des images peuvent rendre certains pixels de l'arrière-plan foncés, ce qui engendre des erreurs d'étiquetage des pixels si nous nous basons uniquement sur l'intensité des pixels.

Pour réduire l'effet du bruit, nous avons utilisé une fenêtre glissante de taille 5×5 qui nous permet de calculer l'intensité moyenne des pixels. Le choix de la taille de la fenêtre glissante a été réalisé de manière empirique. Cette taille correspond généralement à l'espace qui se trouve entre les mots. Ce descripteur est défini par l'équation 3.13. Il permet de décrire les espaces entre les éléments textuels de la page. Les résultats de ce descripteur sur les exemples de la figure 3.2 sont présentés dans la figure 3.15.

$$f_{12} = Avg(\sum_{i \in I'} \sum_{j \in J'} p_{ij}) \tag{3.13}$$

Avec I' et J' la taille de la fenêtre d'analyse de p_{ij} le niveau de gris du pixel de coordonnées (i, j)

Comme dans les figures précédentes, les fortes intensités moyennes sont représentées par des couleurs chaudes proches du rouge alors que les faibles intensités moyennes sont représentées par des couleurs froides proches du bleu. D'après les exemples de cette figure, les illustrations sont caractérisées par des fortes intensités moyennes de pixels alors que les régions du fond sont définies par des faibles intensités moyennes de pixels.

Pour les régions textuelles, les intensités de pixels sont généralement supérieures à celles des pixels de fond et inférieures à celles des pixels d'illustration. Par contre, leur intensité varie en fonction des images de la page. Par exemple, les intensités des pixels des régions textuelles de la figure 3.15.a sont proches des intensités des pixels de la lettrine ou de l'illustration. Par contre, dans la figure 3.15.d les intensités des pixels sont faibles et plutôt proches des intensités des pixels de fond. Ceci prouve bien la variabilité des caractéristiques des éléments textuels et par conséquent l'intérêt d'utiliser une approche locale de detection des mots omis.

Les espaces entre les mots sont mieux définis avec les réponses de ce descripteur puisqu'ils sont représentés avec le même code de couleur que des pixels du fond. Cela devrait donc permettre par la suite de différencier les espaces entre les mots et les éléments textuels et graphiques de l'image.

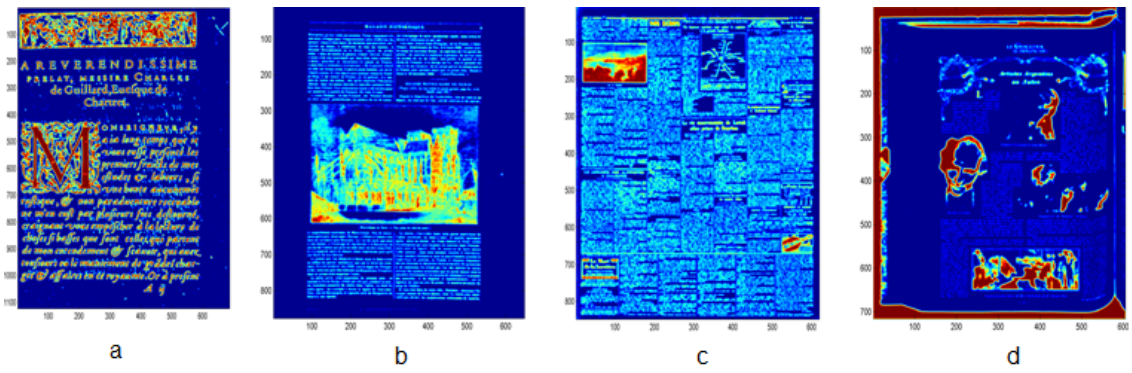


FIGURE 3.15 – Résultats de l'application du descripteur des intensités moyennes des pixels

3.2 Apprentissage et classification des pixels d'arrière-plan

Dans la section précédente, nous avons décrit l'ensemble des descripteurs qui nous permettent de caractériser les différentes régions d'intérêt de la page. Pour cela, nous avons illustré les réponses de chaque descripteur sur des pages qui proviennent de différents types de documents (presse, monographies, etc.) afin de montrer la pertinence des caractéristiques proposées.

Après la phase de caractérisation des images, nous nous intéressons maintenant à l'étape de décision. Rappelons que cette étape est adaptative à chaque page de document. Sur chaque page on utilise les descripteurs sélectionnés pour caractériser les éléments détectés par l'OCR (textes et illustrations) ; ceci permet d'adapter le classifieur aux caractéristiques typographiques et physiques du document. Cela suppose naturellement que nous faisons confiance aux résultats de segmentation fournis par l'OCR pour les classes texte et graphique, ce qui est valide en pratique. La classifieur ainsi optimisé sur le document examiné est ensuite utilisé pour analyser les autres partie de l'image qui n'appartiennent ni aux éléments textuels ni aux éléments graphiques de la page détectés par l'OCR.

A la fin de cette opération, nous obtenons des résultats de classification au niveau pixels. A la BnF, la procédure d'évaluation des résultats de l'OCR utilise des scores de confiance fondés sur la proportion de mots reconnus. Nous avons donc procédé dans une dernière étape au regroupement des pixels de texte pour former des composantes connexes qui peuvent se rapprocher de mots ou parties de mots omis. Dans les paragraphes qui suivent, nous présentons l'ensemble des étapes qui ont permis de mettre en œuvre la méthode de détection des éléments textuels omis.

Procédure d'apprentissage et de classification des pixels

A l'issue de l'opération de caractérisation des textures de l'image nous obtenons pour chaque pixel de l'image un vecteur de caractéristiques composé de 12 valeurs constituées des descripteurs mesurés à différentes échelles. La méthode de détection des éléments omis repose sur le choix d'un système de décision qui analyse les descripteurs de chaque pixel pour fournir une décision sur sa classe d'appartenance (texte / graphique / fond). Par conséquent, c'est un problème de classification supervisée pour lequel nous connaissons *a priori* les étiquettes des points permettant de construire les classes.

Dans un premier temps, nous avons sélectionné plusieurs classifieurs candidats (k-ppv, modèle de mélanges de gaussiennes, réseaux de neurones et séparateurs à vaste marge « SVM ») pour sélectionner le meilleur algorithme de classification. Pour valider les résultats de chaque classifieur, nous avons utilisé une approche de validation croisée qui a montré que les meilleurs résultats de classification sont obtenus avec un classifieur de type SVM.

Dans la littérature, il existe d'autres algorithmes de classification que nous n'avons pas testés. Ce qui a motivé le choix de ces algorithmes de classification est avant tout leur capacité de traiter des vecteurs de caractéristiques de grande dimensions et la facilité de leurs mises en oeuvre.

(a) Optimisation du classifieur

Les SVM sont des méthodes de classification binaires basées sur la théorie statistique de l'apprentissage de V. Vapnik. Les SVM sont une génération de classifieurs linéaires qui se basent sur deux principes clés qui sont la séparation des représentations des deux classes

en utilisant l'hyperplan à marge maximale, et l'augmentation de l'espace de représentation des classes en appliquant une fonction noyau sur les données d'apprentissage.

La détermination de l'hyperplan qui sépare les échantillons des deux classes est réalisée en recherchant l'hyperplan maximisant la distance aux deux classes. Les échantillons proches de la frontière entre les deux classes appelés « vecteur support » sont ensuite sélectionnés pour former l'hyperplan séparateur. Dans le cas linéairement séparable, l'équation de l'hyperplan est définie par la formule suivante :

$$h : X \to Y, \quad d'équation \quad h(x) = \langle w^*, x \rangle + w_0^* = \sum_{i=1}^{m} \alpha_i^* u_i \langle x_i x \rangle + w_0^* \tag{3.14}$$

où X est de dimension d et w un vecteur de poids et x un vecteur d'entrée

Dans le cas non linairement séparable, l'application de fonctions noyaux permet de transformer l'espace initial de représentation des données en un espace de plus grande dimension. Cette transformation non linaire des données a pour objectif de se ramener au cas linéairement séparable mais dans un espace de dimension plus grande. L'équation de l'hyperplan séparateur s'écrit alors :

$$h(x) = \sum_{i=1}^{m} \alpha_i^* \quad u_i \quad \langle \phi(x) \quad \phi(x_i) \rangle + w_0^* \tag{3.15}$$

Les noyaux doivent respecter certaines conditions. Ils doivent correspondre d'abord à un produit scalaire défini dans un espace de grande dimension. Une fonction $K : X \times X \to R$ est une fonction noyau si et seulement si elle est symétrique et définie semi-positive. Une démonstration de ce théorème peut être trouvée dans [SC08] page 118.

Dans la littérature, plusieurs fonctions noyaux ont été proposées :

- Les noyaux linéaires : $k(x, x') = x.x'$
- Les noyaux exponentiels : $k(x, x') = exp(\langle x.x' \rangle)$
- Les noyaux gaussiens (RBF) : $k(x, x') = exp \quad \frac{\|x-x'\|}{\gamma^2}$
- Les noyaux binomiaux : $k(x, x') = (1 - \langle x.x' \rangle)^{-\alpha}$
- Les noyaux polynomiaux : $k_d(x, x') = (x.x' + c)^d$ polynôme de degré d, c étant une constante.

Le choix du noyau et de ses paramètres comme la largeur de la bande est souvent un problème pratique et critique lors de la mise en oeuvre des SVM. Dans notre travail, nous avons testé plusieurs noyaux ainsi que plusieurs configurations des hyper paramètres du classifieur afin de sélectionner le classifieur optimal. Pour cela, nous avons appliqué deux techniques de validation :

1. La première est le *grid search* qui permet de valider les choix des paramètres des classifieurs SVM,
2. La deuxième technique est la validation croisée qui permet d'estimer les performances moyennes de chaque SVM en répétant plusieurs fois (4 fois en pratique) l'opération d'apprentissage et de classification. Cela signifie que nous avons décomposé les données d'apprentissage en quatre parties, de façon à ce qu'à chaque itération, nous sélectionnions une partie pour l'évaluation du classifieur et trois parties pour l'apprentissage du classifieur

En plus du choix du noyau, deux hyper paramètres supplémentaires régissent l'apprentissage d'un classifieur SVM. Il s'agit de la valeur de la marge maximale qui sépare les

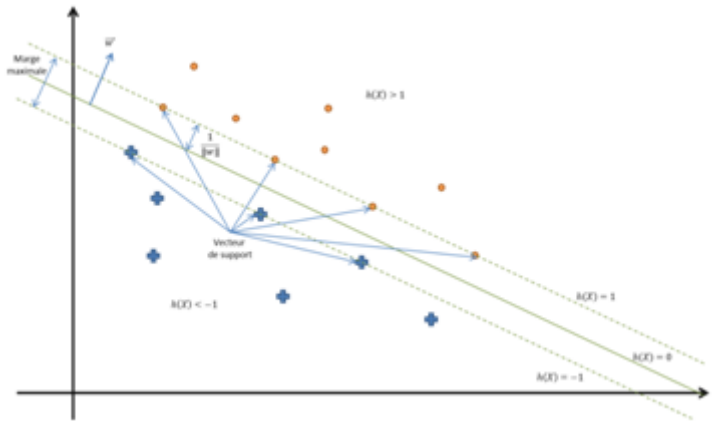


FIGURE 3.16 – L'hyperplan optimal séparant la classe des plus des classe des cercles

deux classes et du paramètre C qui permet de contrôler le compromis entre le nombre d'erreurs de classification et la taille de la marge.

Dans notre méthode, les classifieurs sont entrainés sur les données locales de chaque image. L'optimisation des paramètres des classifieurs est réalisée dans une étape préalable sur un ensemble de 50 images afin de valider le choix des hyperparamètres de la fonction SVM. Pour cela, nous avons testé les noyaux suivants : linéaire, Gaussien, binomial, et sigmoïde. Les paramètres de marges explorés dans notre procédure de *grid search* sont les suivants : $\xi = \{0.1 \quad 0.075 \quad 0.05 \quad 0.025 \quad 0.01\}$. La liste des valeurs du paramètre C est sont $C = \{2^{-4} \quad 2^{-3} \quad 2^{-2} \quad 2^{-1} \quad 1 \quad 2^{1} \quad 2^{2} \quad 2^{3} \quad 2^{4}\}$.

Les meilleurs résultats de classification ont été obtenus majortairement avec un noyeau *Gaussien*, une taille de marge $\xi = 0.05$ et un coût de tolérance d'erreurs de classification $C = 2^{-1}$. Par conséquent, nous avons opté pour cette combinaison d'hyperparametres pour former nos classifieurs.

(b) Apprentissage du classifieur pour la détection des régions omises

Les images sont décomposées en trois classes de pixels (arrière-plan, graphique, texte). Les éléments textuels et graphiques sont décrits par des boîtes englobantes définies dans le fichier ALTO. Par conséquent, les zones de texte contiennent aussi des parties non encrées qui peuvent se confondre avec des zones de l'arrière-plan. Ces situations très fréquentes montrent qu'il est impossible de distinguer ces deux classes si on se limite à un examen trop local du voisinage du pixel. Cependant, les descripteurs que nous avons choisis permettent, dans une certaine mesure, de distinguer ces situations grâce à un paramétrage approprié du voisinage considéré.

La classe des pixels d'arrière-plan regroupe des pixels (clairs) qui se trouvent dans la région imprimée et des pixels (clairs) qui se trouvent dans les marges de la page. Cependant, en utilisant les descripteurs de notre approche, les caractéristiques des pixels d'arrière-plan englobés dans les zones imprimées ont des valeurs différentes de ceux qui se trouvent dans les marges de la page. En effet, l'utilisation de fenêtres glissantes à différentes échelles pour décrire les textures de la page a lissé la description des textures obtenue dans les boîtes englobantes des mots sur les espaces qui se trouvent entre les mots, ce qui engendre un chevauchement entre la représentation des éléments textuels et la représentation de l'arrière-plan dans l'espace des caractéristiques utilisés. Pour éviter ce genre d'ambiguïtés,

nous avons décomposé la classe arrière-plan en deux sous-classes (les espaces entre mots et les régions de marge). Cela a permis de qualifier chaque type de pixels d'arrière-plan d'une manière précise.

Les données d'apprentissage de la quatrième classe (espace inter-mots) sont déterminées à partir des résultats de l'OCR. En effet, d'après le schéma XML des fichiers ALTO, les blocs textuels englobent à la fois des mots et des espaces entre les mots. Par conséquent, pour sélectionner les données d'apprentissage des pixels d'espaces entre les mots, nous procédons tous d'abord à la sélection de tous les blocs textuels de la page puis nous retranchons les boîtes englobant des mots dans ces éléments. Les zones restantes représentent les espaces entre mots. Ce traitement est réalisé sous la condition vérifiée par la BnF et qui considère les boites englobantes des mots ne sont pas collées.

En ce qui concerne les exemples d'apprentissage de l'arrière-plan, contrairement aux autres classes de pixels, on ne peut pas utiliser les régions considérées par l'OCR comme arrière-plan pour apprendre les classifieurs de notre approche. En effet, ces régions sont généralement les zones candidates pouvant contenir des éléments textuels omis par l'OCR. Par conséquent, pour entrainer le classifieur des pixels d'arrière-plan, nous avons sélectionné 10 images (cinq images binaires et cinq images en niveau de gris) pour constituer des exemples d'apprentissage du fond. Certaines images contiennent des défauts physiques (des taches et des déchirures) pour couvrir le cas où il y a des défauts dans les régions d'arrière-plan.

Après la sélection des données d'apprentissage, nous avons adopté la stratégie d'apprentissage *un contre tous* qui consiste à entraîner un SVM biclasse en utilisant les éléments d'une classe contre les autres. L'inconvénient principal réside alors dans la représentativité des exemples et des contre-exemples dans la base d'apprentissage. Pour résoudre ce problème, nous employons pour chaque classifieur une base d'apprentissage équilibrée. La première partie contient les données d'apprentissage de la classe à détecter et la deuxième partie comprend les données d'apprentissage des autres classes de la page. Les données des autres classes sont sélectionnées aléatoirement dans la page jusqu'à atteindre un effectif équilibré avec les données de la classe à reconnaitre.

L'approche de détection des éléments omis que nous avons développée englobe donc quatre classifieurs de type séparateurs à vaste marge « SVM ». Chaque classifieur est spécialisé dans la classification d'une classe d'éléments : 1. Classe des pixels de l'arrière-plan (papier) ; 2. Classe des pixels des espaces entre éléments textuels (papier + régions textuelles) ; 3. Classes des pixels des éléments textuels (mots) et 4. Classes des pixels des éléments graphiques (illustrations).

La procédure d'apprentissage adaptative que nous avons adoptée est réalisée à chaque nouvelle opération de vérification. Cela signifie que la formation des classifieurs de notre approche est effectuée sur les caractéristiques locales de chaque page vérifiée. Enfin, une fois l'opération d'apprentissage achevée, nous procédons à la classification des pixels qui se trouvent dans les régions d'arrière-plan. L'inconvénient majeur de cette procédure réside dans le temps de traitement important causé par l'apprentissage des classifieurs à chaque page à vérifier.

(c) Classification des pixels des régions de fond de l'OCR

Après l'apprentissage des classifieurs de notre approche, nous procédons à la classification des pixels considérés par l'OCR comme arrière-plan (les pixels qui n'appartiennent

ni aux boites englobantes des mots ni à celles des illustrations dans l'ALTO produit par l'OCR).

La classification des pixels se déroule en deux étapes :

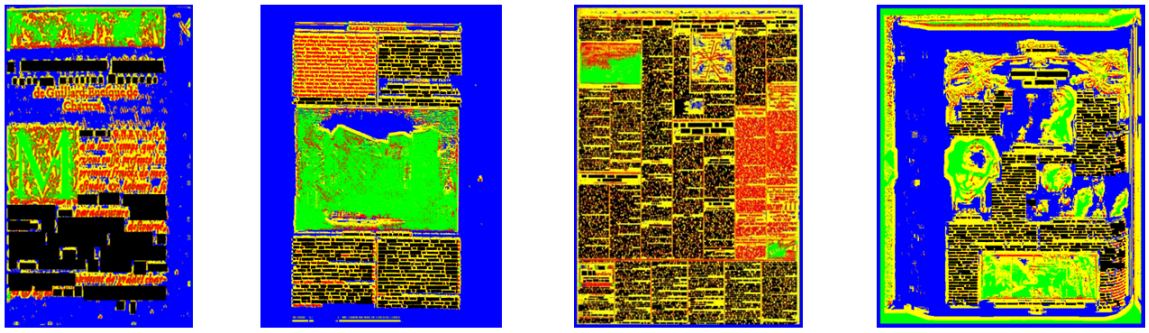
1. La première étape utilise directement les classifieurs appris précédemment pour produire un résultat préliminaire sur les régions de l'arrière-plan. Por cela, nous employons quatre classifieurs SVM pour classer les pixels d'arrière-plan dans quatre classes de pixels (texte, illustration, espace entre mots et arrière-plan).
2. La deuxième étape se base sur une pile de décisions de classification qui permet de fusionner les résultats des classifications réalisées lors de la première étape.

Après la réalisation de la première phase de classification, nous obtenons pour chaque pixel, quatre décisions de classifications. Ces décisions peuvent être utilisées directement si les quatre classifieurs sont en accord sur la classe du pixel. Par exemple, le SVM d'éléments élément textuel classifie un pixel dans la classe du texte alors que les SVM des autres composants le classifient dans la classe des contre-exemples. Cependant, dans certains cas, on peut avoir des décisions de classification contradictoires, comme par exemple des confusions entre la classe des pixels des éléments textuels et la classe des pixels des espaces entre mots. Pour résoudre ce problème nous avons employé des règles de décision qui majorent les classes des pixels textuels ou graphiques par rapport aux classes des pixels d'arrière-plan ou d'espace entre mots. Si un pixel est classé à la fois comme pixel d'espace entre mot et pixel textuel, la classe finale de ce pixel sera éléments textuels. Cette règle de majoration s'applique également sur tous les cas de confusions entre la classe des éléments textuels et les classes des régions d'arrière-plan ou des éléments graphiques. D'autre part, si un pixel est classé à la fois comme pixel graphique et pixel d'arrière-plan ou pixel d'espace entre mots, il sera classé finalement dans la classe des pixels graphiques. Les pixels de marge et d'espace entre mots appartiennent à la classe des pixels d'arrière-plan, en conséquence le problème de confusion entre ces deux sous-classes ne se pose pas.

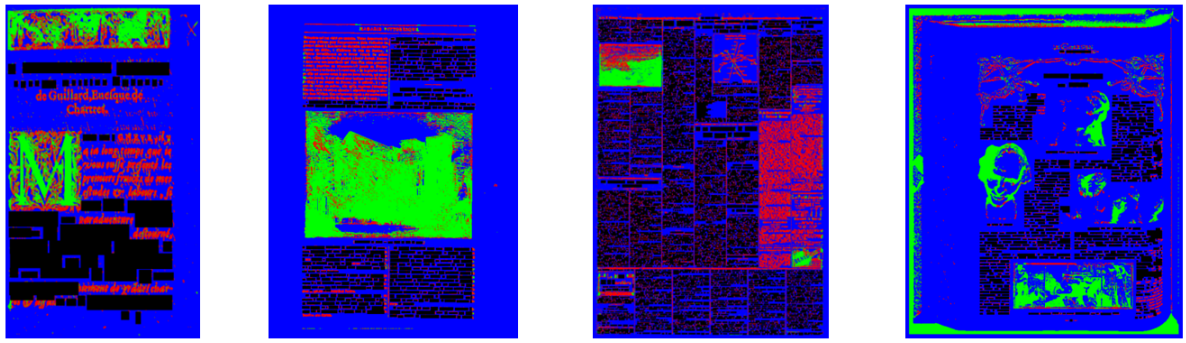
La figure 3.17 représente les résultats de l'opération de classification des pixels des images de la figure 3.2. avant et après l'opération de fusion des pixels de fond avec les pixels d'espaces entre mots. Les pixels noirs sont ceux détectés par l'OCR et que nous n'avons pas traités dans notre approche. Volontairement, nous avons classé dans ces exemples les pixels d'illustration afin de vérifier la capacité de classification de nos classifieurs. La première ligne de la figure 3.17 représente les 4 classes détectées tandis que la seconde ligne représente le résultat de la détection après fusion des classes arrière-plan et espace entre mots. Les pixels rouges, verts et bleus représentent respectivement les éléments textuels, graphiques et les régions d'arrière-plans. Les pixels jaunes sur la première ligne de résultats de la figure 3.17 représentent la classe espaces entre les mots.

D'après ces figures nous constatons que les éléments textuels et graphiques omis sont bien détectés. Par contre, nous remarquons aussi la présence de bruit de détection dans les régions d'arrière-plan ainsi que des confusions de classification entre les zones textuelles et graphiques. Ces erreurs de classification peuvent être causées par certaines incohérences dans les réponses de nos descripteurs. Pour étudier le comportement de nos descripteurs, nous présentons dans l'annexe A une analyse en composantes principales de réponses de nos descripteurs.

Afin d'éliminer ces artefacts de détection, nous appliquons une opération d'ouverture morphologique avec un élément circulaire de rayon $R = 3$. La taille de l'élément structu-



(a) Avant la fuison des pixels d'espace inter-mots avec les pixels des marges



(b) Aprés la fuison des pixels d'espace inter-mots avec les pixels des marges

FIGURE 3.17 – Résultat de classification des pixels de l'image

rant est choisie de manière empirique. Enfin nous appliquons un lissage des résultats de classification des pixels en affectant à chaque masse connexe l'étiquette majoritaire qui lui est associée. Les résultats de ces opérations de post-traitement sont présentés dans la figure 3.18. D'après ces représentations, le bruit de détection est presque totalement éliminé.

Passage du niveau pixel au niveau mot

Bien que la procédure de classification précédente classe les pixels d'arrière-plan en trois classes (classe des pixels de texte, classe des pixels d'illustration, classe des pixels

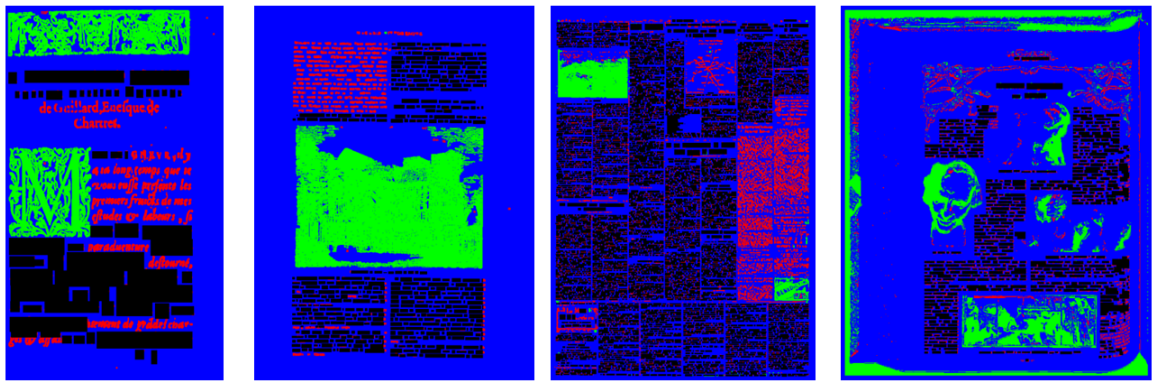


FIGURE 3.18 – Résultats finaux de la procédure de classification de notre approche

d'arrière-plan), on ne peut pas utiliser ces résultats directement pour évaluer la quatité des mots omis et donc contrôler les résultats de l'OCR. En effet, à la BnF, l'évaluation des performances des résultats de l'OCR est réalisée au niveau mots à travers l'utilisation des scores de confiance annoncés sur chaque mot reconnu. Ces scores de confiance sont utilisés par la plateforme numérique de la BnF pour estimer le taux de reconnaissance des mots obtenus sur un document.

Du fait du fonctionnement des systèmes OCR commerciaux (en boîte noire), la méthode de calcul des scores de confiance des mots est inconnue des utilisateurs. De plus, seuls les éléments détectés et reconnus par l'OCR sont utilisés pour évaluer les performances des résultats de reconnaissance des caractères. En effet, les scores de confiance sont communiqués uniquement sur les mots reconnus par l'OCR. Ce qui signifie que l'erreur d'omission des mots n'est pas prise en compte par le processus de contrôle qualité de ces systèmes.

Afin d'être compatible avec la chaîne de contrôle qualité de la BnF, il est donc nécessaire de donner une évaluation du texte omis en terme de nombre de mots équivalents omis. Pour cela, nous avons développé un post-traitement spécifique qui analyse les images fournies par le système de détection pour en sortir une estimation approximative des mots omis. Pour ce faire, nous avons regroupé les pixels connectés ou proches avec un algorithme de dilatation basé sur la transformée en distance.

Nous présentons dans ce qui suit les détails de l'algorithme de regroupement des pixels.

(a) Outil de regroupement des pixels

La notion de distance est très utilisée dans l'analyse d'image et la description des formes. Elle intervient par exemple pour la mesure de la longueur ou de l'épaisseur des objets présents dans l'image, pour la mesure de similarité entre formes, pour guider l'opération de squelettisation des formes ou encore pour calculer un diagramme de Voronoï généralisé. Dans la littérature, de nombreuses familles de distances ont été employées pour calculer la transformée en distance.

De manière générale, la transformée en distance est une représentation d'image numérique binaire qui consiste à associer à chaque pixel d'une forme discrète X dans une image I de taille $n \times n$, sa distance par rapport au point du fond le plus proche. Inversement, on peut calculer aussi la distance de chaque point du fond à l'objet le plus proche dans l'image. Un algorithme rapide de calcul de la carte des distances est présenté par Maurer dans [MQR03].

Selon [TC07], pour des raisons de fiabilité et de facilité de calcul, les distances à valeurs entières sont souvent utilisées dans les algorithmes d'analyse d'image. Parmi ces méthodes nous trouvons la distance euclidienne, les distances de Chanfrein ou les distances géodésiques.

D'après [TC07], une distance d sur un ensemble non vide (noté E) à valeur dans un sous-groupe de R (noté F) est une application $d : E \times E \rightarrow F$ vérifiant les quatre propriétés suivantes :

- (positivité) $\forall p, q \in E, d(p, q) \geq 0$;
- (définie) $\forall p, q \in E, d(p, q) = 0 \Leftrightarrow p = q$;
- (symétrie) $\forall p, q \in E, d(p, q) = d(q, p)$;
- (inégalité triangulaire) $\forall p, q \in E, d(p, q) \leq d(p, r) + d(r, p)$.

Un écart est une distance « d » qui ne vérifie pas $\forall p, q \in E, d(p, q) = 0 \Leftrightarrow p = q$;

pour tout $\forall p, q \in E$. La fonction de distance euclidienne au carré d_E^2 à valeur entière pour deux points de Z^n ne vérifie pas l'inégalité triangulaire. Par exemple, pour deux point A et B définis dans un espace Z^n avec les coordonnées $A = (1, 0, 0, ..., 0)$ et $B = (2, 0, 0, ..., 0)$. La distance euclidienne au carré $d_E^2(O, A) = d_E^2(A, B)$ et $d_E^2(O, B) = 4$ donc $d_E^2(O, B) d_E^2(O, A) + d_E^2(A, B)$. Par contre, la distance euclidienne est bien une distance puisque elle vérifie la propriété de l'inégalité triangulaire. Pour les deux points de l'exemple précédent $d(O, A) = d(A, B)$ et $d(O, B) = d(O, A) + d(A, B)$.

La distance euclidienne au carré est généralement utilisée pour classer les éléments du plus proche au plus éloigné et comme la racine carrée est une fonction monotone croissante, le classement est le même si on travaille avec la distance au carré ou la distance euclidienne (la différence entre les deux c'est qui va plus vite si on n'applique pas la racine carrée).

Dans notre travail, nous avons appliqué l'algorithme de [MQR03] sur les résultats de classification de notre approche pour convertir les résultats de détection de notre approche du niveau pixels au niveau mots avec la distance euclidienne. Nous expliquons dans la section suivante la démarche que nous avons adoptée.

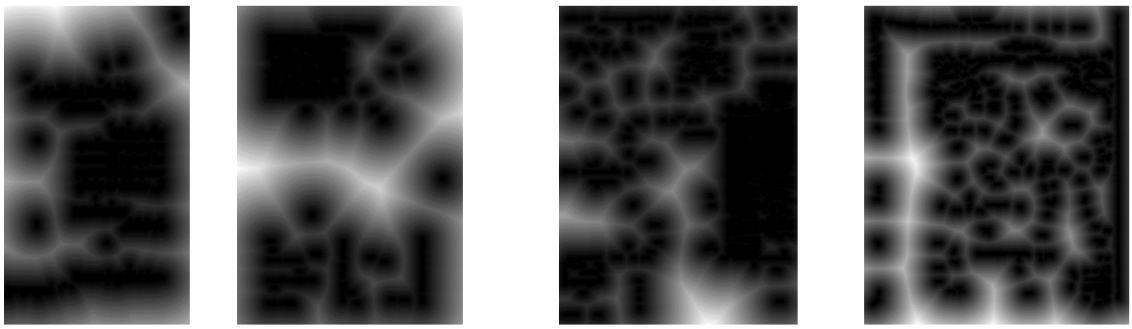
(b) Mise en œuvre de la transformée en distance

L'algorithme de regroupement que nous avons adopté commence par l'analyse des espaces qui existent entre les composantes connexes obtenues lors du processus de classification des pixels sur l'image masquée. Cette analyse assure le rassemblement de toutes les composantes de l'image qui sont très proches.

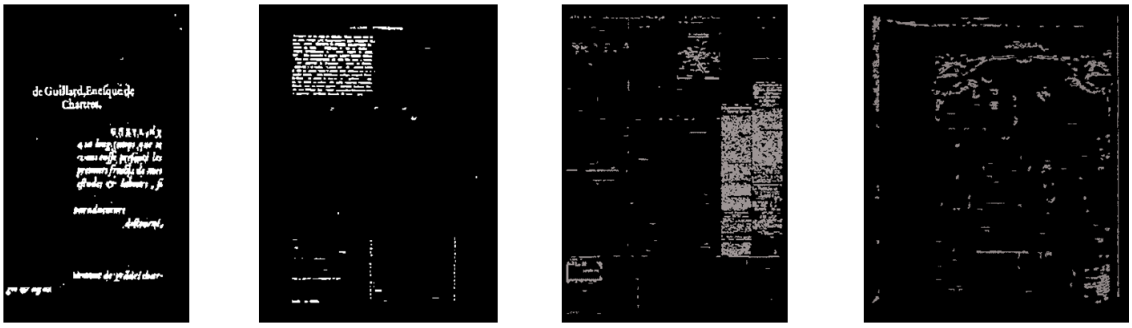
Les distances entre les caractères des mots sont généralement plus faibles que les distances qui se trouvent entre les mots. Cela signifie qu'en appliquant une stratégie de regroupement en composantes connexes basée sur une opération de comparaison des distances, nous pourrons transformer les résultats de détection des éléments omis au niveau pixels vers le niveau mots. Cette opération de comparaisons emploie les coordonnées des composantes connexes détectées par notre approche et la taille des espaces entre les mots de la page pour regrouper les pixels qui appartiennent au même mot.

Dans le fichier ALTO, l'information quant aux espaces existant entre les mots de la page est présente. Par conséquent, nous avons calculé la taille moyenne des espaces entre les mots en utilisant les données du fichier ALTO. Cette valeur moyenne sert ensuite de seuil pour regrouper les pixels et obtenir une pseudo segmentation en mots de l'image résultat de la détection des éléments omis. Cette démarche locale de calcul de seuil de regroupement rend notre algorithme adaptatif aux propriétés internes des documents, ce qui répond parfaitement aux exigences de notre contexte.

Les résultats de la transformée en distance sur l'image produite par notre détecteur d'éléments omis sont présentés sur la figure 3.19a. Sur ces figures, les pixels clairs représentent les pixels les plus éloignés des éléments omis (présentant une distance importante). Alors que les pixels foncés représentent les pixels voisins des éléments omis (présentant une faible distance). Nous allons donc rassembler tous les pixels qui ont des distances inférieures au seuil de regroupement pour transformer les résultats de notre approche de vérification du niveau pixels au niveau mots. Les éléments obtenus lors de la réalisation de cette opération représentent les enveloppes des mots omis. Une étape de filtrage permet d'éliminer toutes les petites composantes de l'image résultat obtenue après seuillage. Les résultats de cette opération sont présentés sur la figure 3.19b.



(a) Résultats transformée de distance



(b) Eléments textuels détectés

FIGURE 3.19 – Résultats de l'opération de regroupement des pixels de l'image

4 Evaluation

Dans cette section, nous présentons le protocole d'évaluation qui nous a permis de déterminer les performances de notre méthode. Pour évaluer notre approche, nous avons suivi une stratégie d'évaluation graduelle qui commence par une analyse qualitative des résultats. Ensuite, pour généraliser les résultats, nous avons réalisé une évaluation quantitative sur une base de validation composée de 165 images prises aléatoirement à partir de 50 documents. Nous avons décomposé cette base de validation en deux familles de documents : des documents contenant que du texte et des documents contenant des textuels et des graphiques. Enfin, pour tester notre approche dans un contexte de production à grande échelle, nous avons réalisé une campagne d'évaluation assistée par opérateurs humains sur des documents numérisés récemment dans le cadre des projets de numérisation de masse de la BnF. Cette base est composée de 1270 pages prises du journal officiel et 160 pages de presses. Pour qualifier les performances de notre méthode nous avons calculé les taux de précision et de rappel de détection des mots omis. Enfin pour se conforter aux spécifités du contexte de contrôle de qualité de la BnF, nous avons évalué la capacité à estimer la qualité de la segmentation des pages et le rejet des pages de mauvaise qualité.

Cette partie est organisée de la façon suivante : nous commençons par la présentation des bases de documents utilisées. Ensuite, nous exposons les métriques que nous avons utilisées. Enfin, nous présentons et commentons les expérimentations que nous avons effectuées.

4.1 Bases de validation

Dans notre processus d'évaluation, nous avons employé trois bases d'images différentes :
- Une base de validation composée de pages de documents monographiques
- Une base de pages qui proviennent du journal officiel
- Une base de pages de presse anciennes.

La première base (Base 5 siècles cf. figure 3.20) a été utilisée pour la procédure d'évaluation automatique. Elle est composée de 165 images de document sélectionnées aléatoirement à partir de 50 documents. Cette base regroupe 65 pages mixtes (contenu textuel et graphique) et 100 pages textuelles. Ceci nous a permis d'affiner notre analyse selon ces deux types de pages.

Afin de correspondre au mieux aux collections documentaires de la BnF, nous avons sélectionné des documents qui couvrent une grande variété de caractéristiques physiques et typographiques. Cette base couvre une période de cinq siècles d'impression. Nous trouvons des documents avec des propriétés variables telles que :
- la présence de lettrines,
- des mises en page à plusieurs colonnes,
- des annotations dans les marges gauches et droites de la page,
- des styles de caractères italiques et des polices de caractères anciennes,
- des tableaux et des formules mathématiques,
- des mises en page à faible marge,
- des défauts d'impression (taches, défauts d'encrage, transparence, etc.),
- des défauts physiques (papiers jaunis, trous, etc.),
- la superposition de tampons sur les textes des pages des documents.

Ces pages sont souvent difficiles à traiter par les algorithmes de segmentation à cause de leur mauvais état physique et de leurs caractéristiques typographiques, et aussi à cause de la complexité des structures physiques présentes. Les illustrations au trait ont des caractéristiques semblables à celles des régions textuelles et les OCR commerciaux ont tendance à confondre ces éléments graphiques avec les régions textuelles. Cette base comprend également des images de documents récents imprimés avec des polices de caractères standards et agencées selon des mises en page uniformes. Les éléments textuels et graphiques ont des propriétés différentes ce qui facilite normalement la séparation de ces deux classes. Les espaces entre les caractères des mots et les mots appartenant aux mêmes paragraphes sont standards et constants sur l'intégralité de l'image de la page, ce qui minimise les défauts de scission et de fusion des éléments de la page.

La deuxième base d'évaluation (Base Journal officiel) est composée de 1270 images du Journal officiel, édité au XIX^e *siècle*. Ces images sont particulières. En effet, elles présentent des structures très délicates à reconnaître puisqu'elles sont composées de trois colonnes et qu'elles présentent des polices de caractères de petite taille. De plus, ces documents sont caractérisés par des papiers jaunis, des lacunes d'encre et des textes délavés qui rendent la procédure de détection des éléments de la page très difficile pour les systèmes d'OCR commerciaux. Du côté de la reliure, en plus des problèmes de contraste, les images sont généralement courbées. Ceci déforme les formes des mots et des caractères et l'OCR tend à les confondre avec du bruit et à les éliminer brutalement.

Contrairement à la base de validation précédente, les pages de ces documents sont numérisées en couleur. De plus ces images sont OCR-isées par des prestataires de numéri-

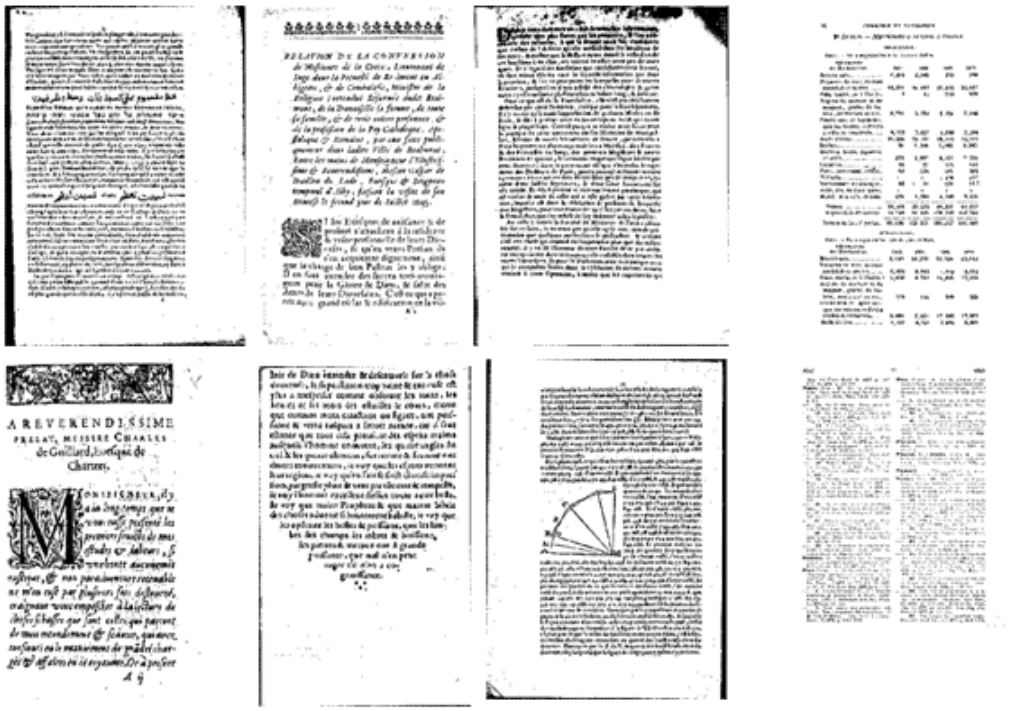


FIGURE 3.20 – Exemples de pages de documents que nous avons utilisées pour évaluer notre approche de détection d'éléments omis

sation externes dans le cadre des projets de numérisation de masse. Les résultats de l'OCR sont fournis dans des fichiers XML qui suivent le schéma des fichiers ALTO. Par contre, nous ne disposons pas de la vérité terrain de ces documents. C'est la raison pour laquelle, nous n'avons employé cette base que pour la procédure de vérification assistée par un opérateur humain. La figure 3.21a représente quelques exemples de pages qui appartiennent à cette base d'image.

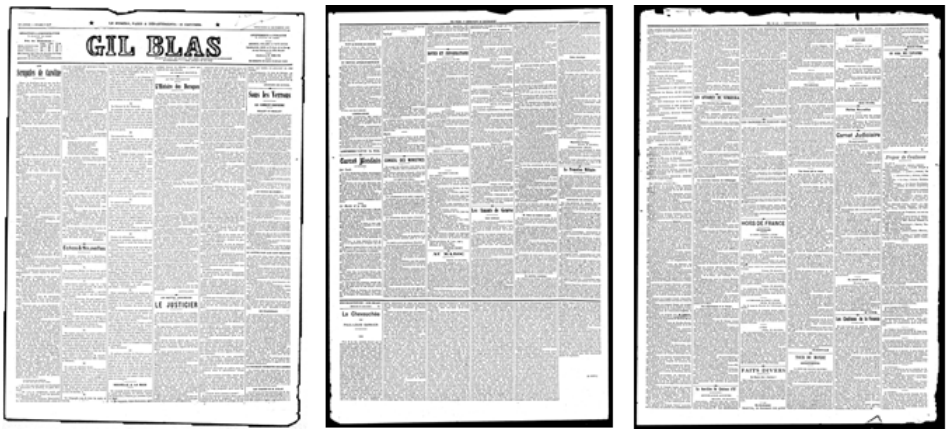
La troisième base de validation (Base Presse) est composée de 160 images binaires de presses numérisées à la résolution de 300 dpi. Ce sont des documents provenant de la presse du XIXème siècle. La figure 3.21b présente des exemples de cette base. Comme pour les images du journal officiel, cette base est très difficile à traiter par les systèmes d'OCR. D'une part, la structure en colonne de ces documents rend la segmentation délicate. D'autre part, les polices sont de petite taille. De plus l'encrage des caractères n'est pas très marqué. Certains artéfacts de binarisation font apparaitre des déformations des caractères. La méthode de numérisation de ces documents a tendance à noircir les marges. Cette base ne comporte pas non plus de vérité terrain c'est la raison pour laquelle l'analyse des résultats de détection des mots omis obtenus sur ces images est réalisée à travers la procédure d'évaluation assistée.

4.2 Métriques pour mesurer la performance

Métriques pour la détection des mots omis

Les métriques que nous avons employées reposent sur les définitions suivantes : $VTEM_{mots}$ désigne la vérité terrain des mots omis, EMD_{mot} désigne les éléments détectés par notre

(a) Journal officiel de France



(b) Presse Française

FIGURE 3.21 – Exemples de pages de documents que nous avons utilisées pour évaluer manuellement notre approche.

approche comme mots omis, $VTSP$ référence la vérité terrain de la structure physique des documents et SP désigne la structure physique des documents obtenue automatiquement par l'OCR. Pour évaluer les performances de notre approche, nous pouvons calculer le rappel et la précision des mots omis. A ce niveau, la précision donne une indication sur l'exactitude des résultats fournis. Elle est définie par le rapport du nombre d'éléments correctement détectés sur le nombre total d'éléments détectés (cf. équation 3.16).

$$Précision = \frac{card(VTEM_{mot} \cap EMD_{mot})}{card(EMD_{mot})} \tag{3.16}$$

Le rappel chiffre la capacité de notre méthode à détecter les éléments omis. Il est défini par le rapport du nombre d'éléments textuels correctement détectés sur le nombre total d'éléments omis par l'OCR dans la page (cf. équation 4.13).

$$Rappel = \frac{card(VTEM_{mot} \cap EMD_{mot})}{card(VTEM_{mot})} \tag{3.17}$$

Métrique pour l'estimation du taux de couverture de l'OCR

Comme nous l'avons montré dans la section 2 de ce chapitre, la méthode proposée ici s'intègre dans le cadre du processus de contrôle de qualité de la BnF. Nous avons donc besoin de définir une mesure de qualité qui serve les controleurs de la BnF. Pour cela, nous définissons le taux de couverture de l'OCR par le rapport du nombre de mots détectés par l'OCR sur l'estimation du nombre total de mots présents dans les documents. Le nombre total des mots $card_{mot}(TOTAL)$ est constitué par le nombre de mots dans le fichier ALTO $card_{mot}(OCR)$ ainsi que par le nombre de mots omis obtenus lors de l'application de notre approche $card_{mot}(OMIS)$. Par conséquent, nous pouvons déduire que le taux de couverture donne une indication sur la proportion du nombre des mots reconnus dans la page par l'OCR.

$$\tau_{couv} = \frac{card_{mot}(OCR)}{card_{mot}(TOTAL)} \tag{3.18}$$

$$\widehat{\tau}_{couv} = \frac{card_{mot}(OCR)}{card_{mot}(OCR) + card_{mot}(OMIS)} \tag{3.19}$$

Pour évaluer la capacité de notre approche à estimer le taux de couverture, nous avons utilisé la racine carrée de l'erreur quadratique moyenne des taux de couverture ($RETC$) définie par l'équation suivante :

$$RETC = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\tau^i_{couv} - \hat{\tau_{couv}}^i)^2} \tag{3.20}$$

où τ^i_{couv} représente le taux de couverture réel et $\hat{\tau}^i_{couv}$ est le taux de couverture estimé.

Métrique pour le rejet de pages

A partir de l'estimation du taux de couverture et en se fixant le seuil minimal de couverture admissible pour l'OCR, il est possible de réaliser un système de rejet des documents non conformes aux exigences du taux de couverture de la BnF. Ce système de rejet est alors évalué par une courbe rappel précision sur les pages rejetées pour chaque seuil de couverture fixé.

4.3 Evaluation de la détection des mots omis

Procédure de vérification avec vérité terrain

(a) Création de la vérité terrain

Afin d'évaluer les résultats de notre approche, nous avons dû constituer la vérité terrain appropriée. Les éléments omis sont généralement les composantes de la page qui existent dans le fichier vérité terrain et qui sont absentes ou partiellement détectées par l'OCR. Les coordonnées des éléments de la page ainsi que leurs tailles sont définies par des boîtes englobantes rectangulaires. Dans notre travail, nous avons utilisé les boîtes englobantes de la vérité terrain et les boîtes englobantes des résultats de l'OCR pour former la vérité terrain des éléments manqués. Cette opération est réalisée en superposant la vérité terrain de la structure de la page (VTSP) avec la structure de la page déterminée automatiquement par l'OCR (SP) pour former $VTEM_{mots} = VTSP_{mots} \neg \{VTSP_{mots} \quad \cap \quad SP_{mots}\}$.

Les éléments totalement oubliés (mots entiers) par l'OCR sont directement assignés à la liste de la vérité terrain des éléments omis $VTEM_{mots}$. Les éléments textuels partiellement omis sont analysés selon leur taille. Si la taille du fragment omis dépasse le quart de la taille de l'élément textuel considéré, la partie omise est assignée à $VTEM_{mots}$. Sinon elle n'est pas ajoutée à $VTEM_{mots}$. A la BnF, l'omission d'éléments graphiques est considérée comme une erreur mineure. Par conséquent, nous n'avons pas évalué ces omissions.

Après la création de la vérité terrain et lors de la procédure d'évaluation de notre approche, nous assignons à chaque élément textuel détecté un élément textuel de la VT en utilisant la distance euclidienne entre les mots des deux fichiers (Vérité terrain et résultats de détection). Un élément manqué par l'OCR est considéré comme bien détecté si plus de 90% de sa surface est couverte par un ou plusieurs éléments trouvés par notre approche. Les éléments non détectés ni par l'OCR ni par notre approche sont considérés comme des erreurs de détection. En conclusion, nous considérons trois classes de mots pour l'évaluation de notre détecteur :

1. Des éléments textuels omis par l'OCR et détectés par notre approche,
2. Des éléments textuels omis par l'OCR et omis par notre approche,
3. Des éléments textuels détectés incorrectement par notre approche (fausse alarme).

(b) Résultat d'évaluation de la détection des mots omis

L'évaluation quantitative est réalisée d'abord globalement sur la base « 5 SIECLES » puis spécifiquement sur les deux sous-bases qui la composent (base des images mixtes et base des images textuelles). En suivant ce critère, nous avons calculé le rappel et la précision de notre procédure de détection. Les résultats de cette évaluation sont présentés dans le tableau 3.1.

	Rappel	Précision
Base 5 siècles	$84,15\%$	$94,73\%$
Sous-base des images textuelles	$83,7\%$	$96,32\%$
Sous-base des images mixtes	84.6%	$93,14\%$

TABLE 3.1 – Evaluation des résultats de notre approche sur la base « 5 SIECLES »

On constate que notre approche détecte près de 85% des mots omis avec une précisions de près de 95%. Ces résultats apparaissent tout à fait intéressants si l'on considère la grande variabilité des documents de la base « 5 SIECLES » et également le seuil de détection très exigeant, fixé à 90% de la surface des éléments omis pour cette expérience.

La décomposition de la base d'évaluation selon la nature des pages montre que les performances de notre détecteur sont à peu près les mêmes sur les deux sous-bases d'évaluation. Selon le tableau 3.1, 84.6% des mots omis dans la base des pages mixtes sont détectés avec une précision de 93,14%. Cette précision s'améliore sur les images des pages textuelles puisque 83,7% des mots omis sont détectés avec une précision de 96,32%. On peut expliquer cette tendance par la nature de certaines illustrations (comme les lettrines, les illustrations au trait, etc.) qui ont des caractéristiques semblables à celles du texte, ce qui biaise légèrement les résultats de notre approche.

En conclusion, d'après les résultats de cette évaluation quantitative, nous pouvons déduire que les performances de notre approche sont intéressantes. Cependant malgré la variabilité des documents de notre base d'évaluation, la taille de cette base peut être insuffisante pour étudier correctement les performances de notre approche. D'où les évaluations sur les autres bases. Mais avant de présenter ces résultats, nous présentons dans ce qui suit les résultats d'évaluation de l'opération d'estimation des taux de couverture et de rejet des documents.

(c) Evaluation de l'estimateur du taux de couverture de l'OCR

La procédure de détection des mots omis que nous avons développée dispose de deux modes de fonctionnement. Elle peut être utilisée comme un outil de contrôle semi-automatique qui propose au contrôleur de la BnF les pages des documents qui contiennent le plus d'erreur d'omission de mots. Elle peut aussi être utilisée comme un prédicteur du taux de couverture réalisé par l'OCR sur les pages d'un document.

Dans cette partie, nous allons évaluer les résultats de notre approche en l'utilisant comme un estimateur du taux de couverture. Le protocole d'évaluation que nous avons employé commence donc par le calcul du taux réel de couverture des pages en se basant sur la vérité terrain d'éléments manqués ($VTEM$).

Dans un second temps, nous estimons les taux de couverture des pages en utilisant les résultats de détection des mots omis de notre approche. Enfin, une fois les données d'évaluation obtenues, nous passons au calcul de la racine de l'erreur quadratique moyenne pour évaluer la précision de nos estimations. Les valeurs de cette mesure sont présentées dans le tableau 3.2.

Selon ce tableau, nous constatons, conformément aux résultats d'évaluation précédente, que les prédictions de notre détecteur sont assez précises. En effet, les racines carrées de l'erreur quadratique moyenne obtenues sur les estimations des taux de couverture sont inférieures ou égale à 10% dans les différents intervalles de taux de couvertures étudiés. De plus, nous remarquons que les erreurs d'estimation des taux de couverture obtenus sur les taux qui sont supérieurs à 95% sont assez faibles (inférieur à 2%), ce qui réduit le risque de fausses alarmes sur ces fichiers. Par exemple, si on considére que les documents qui ont un taux de couverture inférieur à 98% de mauvaise qualité, on peut fixer le seuil de couverture minimal à 97% puisque la $RETC$ obtenue sur cet intervalle est inférieure à 1%.

Afin de visualiser le comportement de notre estimateur, nous présentons dans la figure

	[50% 60%]	[60% 70%]	[70% 80%]	[80% 90%]	[90% 95%]	[95% 96%]	[96% 97%]	[97% 98%]	[98% 99%]	[99% 100%]
RETC	0, 1042	0, 1015	0, 0895	0, 0487	0, 0383	0, 0206	0, 0148	0, 0113	0, 0066	0, 0039

TABLE 3.2 – Les racines carrées des erreurs quadratiques moyennes obtenues sur les estimations des taux de couverture

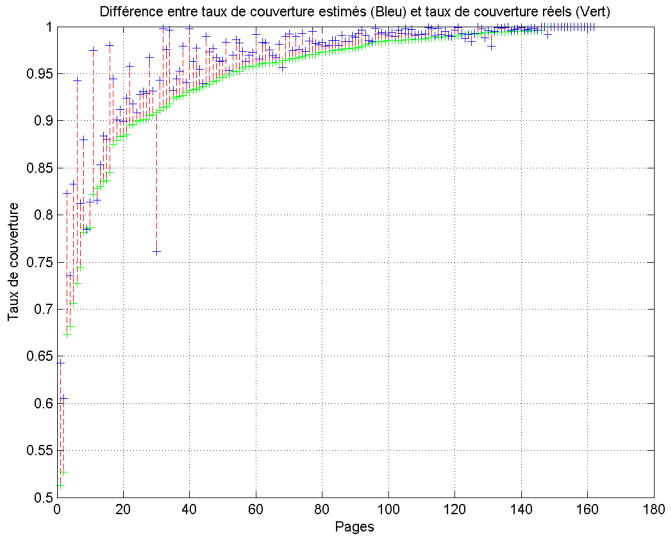


FIGURE 3.22 – Distribution simultanée des taux de couverture réels et estimés

3.22 la distribution des pages selon leurs taux de couverture réels (points verts) et estimés (points bleus). D'après cette figure, nous remarquons que les estimations des taux de couverture sont précises sur presque l'ensemble des pages traitées par notre approche. En effet, nous constatons que l'erreur d'estimation des taux de couverture est inférieure à 10% dans la plus part des cas évalués par notre approche (sauf pour les 20 premières pages). De plus, nous constatons que les estimations des taux de couverture suivent bien la variation des taux réels. En effet, lorsqu'on a une diminution dans les taux de couverture réels, on constate une diminution dans les taux de couverture estimés. Par conséquent, nous pouvons déduire que même si les estimations de notre approche sont biaisées, l'utilisation de ces taux dans une procédure de rejet automatique est envisageable.

(d) Evaluation de l'opération de rejet des documents

L'estimation des taux de couverture permet de contrôler la qualité des résultats de segmentation des pages en rejetant ceux qui ne correspondent pas aux exigences de qualité de la BnF. Pour cela, la BnF fixe dans les cahiers de charge des projets de numérisation de masse un seuil minimal de couverture des documents. On procède au rejet des pages ayant des taux de couverture inférieurs au seuil de rejet. Pour réaliser les expérimentations de cette évaluation, nous avons employé deux seuils de rejet : le premier seuil, appelé seuil effectif de rejet, est utilisé pour former la vérité terrain des pages à rejeter. En pratique, c'est le seuil fixé par la BnF. Le deuxième seuil, appelé seuil expérimental de rejet, est utilisé pour former les résultats expérimentaux de rejet des documents. En pratique, c'est le seuil adopté par notre approche.

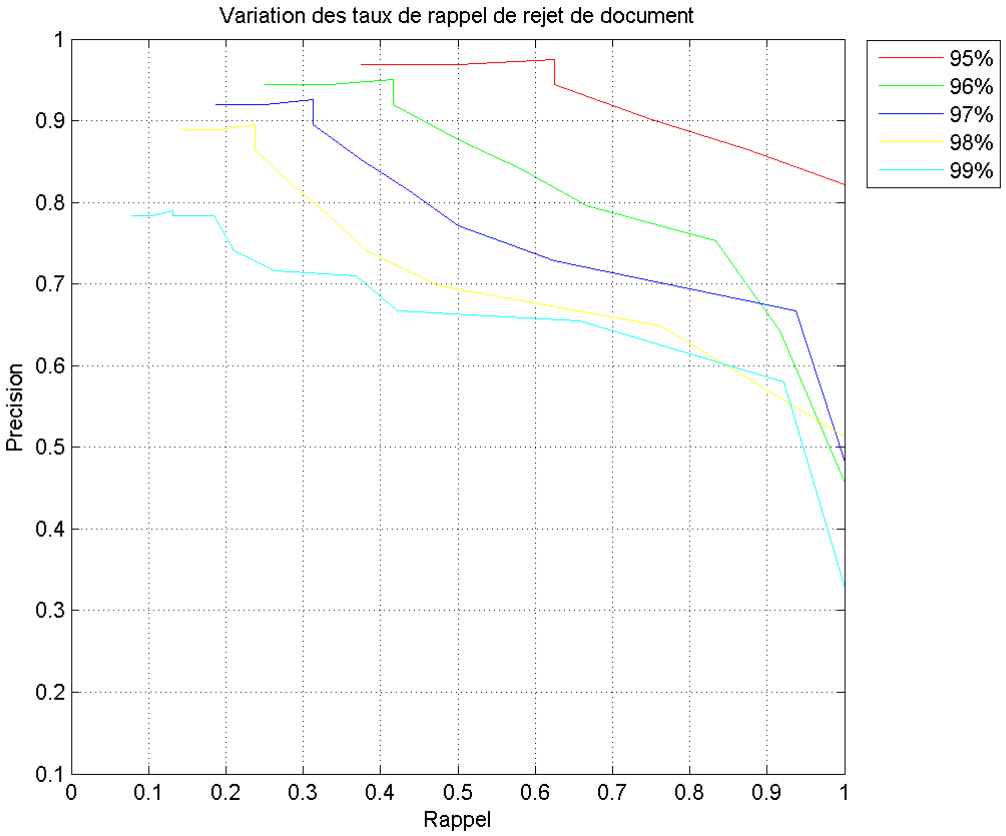


FIGURE 3.23 – Performances de notre procédure de rejet automatique des documents en variant le seuil de rejet expérimental des documents (de 80 à 99%) et pour différents niveaux de qualité exigé (seuils effectifs de 95 à 99%).

En suivant cette démarche, nous procédons, dans un premier temps, à la formation de la vérité terrain des pages à rejeter en fixant plusieurs seuils de rejet effectifs. Ceci nous permet de visualiser le comportement de notre système à différents niveaux de qualité. Ensuite, nous comparons les estimations de taux de couverture de notre approche aux seuils de rejet expérimentaux de notre procédure d'évaluation. Une fois ces deux opérations sont effectuées, nous alignons les résultats de rejet de notre approche avec la vérité terrain des résultats de rejet ce qui nous permet de calculer par la suite les performances de notre approche.

La figure 3.23 présente les allures des courbes de rappel/précision obtenus en variant les valeurs des seuils expérimentaux par rapport à la vérité terrain obtenue avec les seuils de rejet effectifs.

Les courbes de la figure 3.23 représentent la distribution des taux de rappel et de précision en variant le seuil de rejet expérimental des documents. Dans cette évaluation, les seuils effectifs de rejet varient entre 95% et 99% alors que l'intervalle des seuils expérimentaux est compris entre 80% et 99%. Ce choix d'intervalle des seuils a été fait d'une part en se basant sur les exigences de qualité de la BnF qui ne tolère pas plus de 5% de mots omis dans les projets de numérisation de masse et d'autre part afin de mesurer les performances de notre procédure de rejet des documents en prenant des seuils de rejet expérimentaux

inférieurs et supérieurs aux seuils de rejet effectifs, ce qui détermine l'influence des erreurs d'estimation.

D'après les résultats de l'opération de rejet de notre approche, nous constatons que plus le seuil de taux de couverture expérimental est petit, plus le taux de rappel de notre procédure de rejet de document est petit. Par exemple, pour un taux de couverture expérimental de 80%, nous obtenons un rappel du rejet inférieur à 10% pour le système pour lequel on souhaiterait un seuil de rejet effectif de 99%. Plus le seuil de rejet effectif est proche du seuil de rejet expérimental, plus le rappel des documents rejetés est important. En effet, pour le système correspondant à un seuil de rejet effectif souhaité de 90% (courbe rouge), le rappel minimal, obtenu avec un seuil de rejet expérimental égal à 80%, est égal à 60%.

Le rappel des systèmes de rejet des documents atteignent la valeur maximale (rappel = 100%) en augmentant le seuil de rejet expérimental à 99%. Ces résultats sont logiques puisque l'augmentation du seuil de rejet des pages entraine l'augmentation du nombre des pages rejetées ce qui augmente le rappel du rejet des documents.

Par contre, contrairement au taux de rappel, l'augmentation des seuils expérimentaux de rejet des documents entraine aussi la diminution du taux de précision. En effet, d'après la figure 3.23, nous remarquons que les estimations des taux de couverture de la page sont généralement sous-estimées. Par conséquent, l'augmentation du seuil de rejet des pages, nous exposera plus aux fausses alarmes ce qui entraîne la diminution de la précision de notre procédure de rejet des documents.

Les meilleures performances sont obtenues avec le système configuré avec un seuil de rejet effectif égal à 95%. En effet, selon la distribution des résultats de ce système, nous pouvons obtenir une précision de rejet supérieure à 90% avec un rappel égal à 70%. De plus, nous pouvons couvrir la totalité des pages à rejeter (rappel = 100%) avec une précision de rejet supérieure à 80%.

A travers cette analyse, nous constatons qu'il existe un compromis entre les taux de précision des décisions de notre approche et les taux de rappel de l'opération de rejet des pages. La meilleure configuration de taux de rejet effectif et expérimental est celle qui garantit une bonne couverture des documents à rejeter toute en conservant une bonne précision de l'opération de rejet des pages. Dans le contexte des projets de numérisation de masse, les échéances de réalisation des projets rend les erreurs de rejet des documents très couteuses. Ceci rend l'utilisation de notre système en mode de rejet automatique(précision très importante) ou semi-automatique (précision assez bonne) envisageable.

Evaluation in situ à la BnF

Une base de 165 images n'est pas suffisante pour évaluer convenablement les performances de notre approche. Nous allons présenter dans cette partie une procédure d'évaluation assistée par un opérateur réalisées sur les documents numériques produits dans le cadre des derniers projets de numérisation de masse de la BnF. Pour tester notre approche dans ce cadre d'une production de masse, nous avons mené une campagne d'évaluation au sein du service de numérisation de la BnF.

Nous commençons par la présentation de l'outil d'évaluation que nous avons développé et les protocoles expérimentaux que nous avons adopté pour évaluer les résultats de notre approche. Ensuite, nous exposons la présentation et l'analyse des résultats d'évaluation.

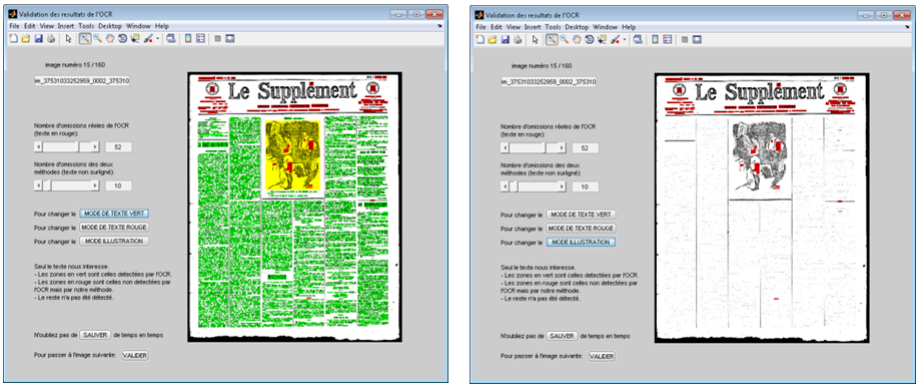


FIGURE 3.24 – Interface graphique : détection sur une image (gauche) et même image avec zones masquées (droite).

Cette campagne d'évaluation a été effectuée sur des images de documents issues de la base du journal officiel et de la base de presse pour avoir une visibilité du comportement de notre approche sur plusieurs types de documents.

(a) Outil d'évaluation et protocole d'évaluation

Pour effectuer les tests, une interface graphique a été créée pour l'évaluation. Elle permet d'obtenir les informations nécessaires pour calculer les métriques d'évaluation de notre approche.

La figure 3.24 montre l'interface graphique mise en place afin d'effectuer les tests. Elle a été pensée dans le but de compter (manuellement) le plus facilement possible les éléments textuels omis par l'OCR : zones texte omis détectées par notre méthode (en rouge, contenant des faux positifs) et zones de texte omises non détectées par notre méthode. Les zones de texte trouvées par l'OCR (en vert) sont connues grâce au fichier ALTO. Théoriquement, l'union de ces trois ensembles donne la totalité du texte de la page.

Deux sliders permettant d'indiquer les nombres de zones comptées sur chaque image ont été mis à disposition de l'utilisateur (cf. figure 3.24). Afin de rendre les tests moins fastidieux, l'application permet une sauvegarde des données déjà validées, ce qui permet de ne pas effectuer les tests en une seule fois, ceux-ci étant long.

De plus, les zones qui nous intéressent étant réparties sur l'ensemble de la page, et les images traitées ayant une grande taille, celle-ci doit être parcourue en entier afin de trouver les zones à compter. Pour faciliter cela, nous avons permis de masquer les zones déjà détectées par l'OCR (en vert), comme le montre la figure 3.24 droite, celle-ci ne nous intéressant pas. Cela rend le comptage de zones omises par notre approche (zones non surlignée) plus facile.

Nous enregistrons finalement dans des tableaux les informations de chaque image de la base : le nombre de zones détectés comme étant omises par l'OCR (résultat de notre approche), le nombre de zones réellement omises par l'OCR, le nombre d'élément oubliés par notre méthode, ainsi que le nombre de zones détectées par l'OCR.

(b) Résultats d'évaluation

Nous avons appliqué deux protocoles d'évaluation qui nous ont permis d'examiner les performances de notre approche dans deux modes de fonctionnement différents. Le premier permet de vérifier la capacité de notre approche à détecter du texte omis. Le deuxième

	Rappel	Précision
Base de presse	84, 62%	72, 74%
Base de journal officiel	81, 26%	89, 55%

TABLE 3.3 – Evaluation des résultats de notre approche sur la Base 5 siècles pour un seul de détection de 90% de surface des éléments manqués.

mode de fonctionnement assure la vérification de la capacité de rejet automatique des documents en utilisant le taux d'omission estimé de notre approche.

Sur la base de presse, notre approche est capable de détecter 84, 62% de texte omis avec une précision de 72, 74%. Sur la base de *Journal officiel*, notre approche produit un rappel de 81, 26% avec une précision de 89, 55%. Les performances de notre approche sur les deux bases de document est donc acceptable. De plus, nous remarquons que les résultats de notre approche sont sont meilleures sur les images du *Journal officiel* que sur la base de presse. Ceci peut être dû aux particularités des images de presse, qui se distinguent par des caractéristiques physiques plus détériorées que celles des images du Journal officiel. Ainsi, les espaces entre les éléments textuels sont plus ou moins clairs sur les documents de journal officiel alors qu'ils sont très restreints sur la presse. Cela peut engendrer des défauts de fusion ou de scission de composants textuels dans les résultats de la procédure de regroupement des pixels textuels de l'image. D'autre part, des défauts physiques tels que défauts de transparence, lacunes de papier ainsi que déchirures sont omniprésentes dans les documents de presse. Par contre, ils sont moins présents dans les collections des documents du journal officiel. Ceci explique bien le fait que les résultats de détection des mots omis sont plus biaisés sur les documents de presse que sur les documents du Journal officiel.

5 Intégration de notre approche à la BnF

Comme nous l'avons montré dans la section 2 de ce chapitre, les documents numériques produits passent par une opération de contrôle qualité dans laquelle la BnF vérifie la structure des fichiers ALTO ainsi que la qualité des images et celle de la transcription automatique des OCR. La méthode que nous proposons va s'intégrer dans ce processus de contrôle de qualité de la BnF. Elle permettra aux contrôleurs de la BnF de vérifier l'existence des mots omis dans les résultats d'OCR.

L'outil que nous proposons à la BnF peut être utilisé comme un moyen de contrôle automatique ou semi-automatique. L'automatisation des résultats de rejet des documents s'effectue en comparant le taux de couverture estimé de la page par rapport à un seuil de couverture fixé a priori par le service de contrôle qualité de la BnF. Ce mode de fonctionnement permet de rendre l'opération de contrôle des résultats de conversion des documents totalement automatique et par conséquent rapide. Par contre, comme vu dans la section 4.3, en suivant ce mode de fonctionnement, on s'expose plus aux faux rejets.

Pour limiter ce risque, nous proposons à la BnF une procédure de contrôle semi-automatique qui se base sur une opération de déqualification collaborative des résultats de l'OCR. Cette procédure commence par proposer aux contrôleurs de la BnF les pages de document candidates au rejet automatique (ex. figure 3.25) pour qu'elles soient validées.

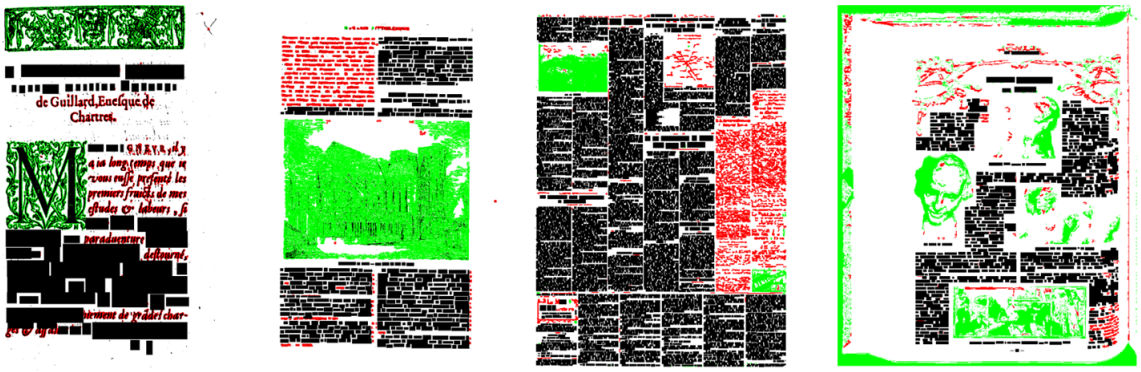


FIGURE 3.25 – Résultat de l'approche de détection d'éléments omis : pages avec un taux de couverture de l'OCR insuffisant

Pour cela, nous avons développé un outil qui permet de valider les résultats de détection automatique de notre approche, à la manière de ce qui a été décrit dans la section 4.3. La figure 3.26 présente l'interface que nous proposons pour annoter les erreurs d'omission de mots. En se basant sur le retour d'expérience obtenu lors de la campagne d'évaluation de notre approche, la conception de cette interface a été optimisée dans le but de faciliter la procédure de vérification des zones candidates de l'image.

Cette interface d'annotation des mots omis est composée de deux fenêtres, la partie à gauche présente les éléments qui ont été détectés par l'OCR. Nous pouvons également contrôler l'affichage de ces éléments selon leurs types (mots, phrases, paragraphes, illustrations). La couleur d'affichage des mots varie selon les scores de confiance assignés par l'OCR aux mots reconnus ce qui donne une visibilité sur la qualité des résultats de reconnaissance des mots. Les boites englobantes vertes représentent les mots ayant des taux de confiance supérieurs à 99%, les couleurs des boites rouges varient de plus clairs au plus foncés en fonction des taux de confiance des mots. Le rouge foncé représente les mots ayant un taux de confiance inférieur à 60% alors on utilise les boites rouges claires pour représenter les mots ayant des taux de reconnaissance supérieurs à 90%. Dans la partie droite de l'interface, nous affichons les résultats de l'opération de vérification de l'existence des mots omis. Les mots omis sont affichés en rouge, les illustrations omises sont présentées en vert alors que les régions vides (arrière-plan) sont non colorées. Les boîtes noires représentent les mots détectés par l'OCR. Cette représentation des résultats de notre détecteur rend la procédure d'annotation de comptage des mots omis intuitive.

Notre interface propose également une estimation des nombres des mots, des phrases et des paragraphes omis dans la page traitée. On peut modifier ces nombres s'ils ne correspondent pas à la réalité des éléments omis. D'autre part, pour simplifier la procédure de vérification des résultats d'OCR, notre approche commence par le tri des résultats d'OCR en fonction de leurs taux de couverture estimé. Cela signifie qu'elle propose au début de l'opération de contrôle les pages qui possèdent les plus faibles taux de couverture. Cela permet de traiter les pages des documents des plus douteuses au moins douteuses, ce qui permet de rejeter les résultats de l'OCR plus rapidement en atteignant rapidement les seuils de rejet des documents.

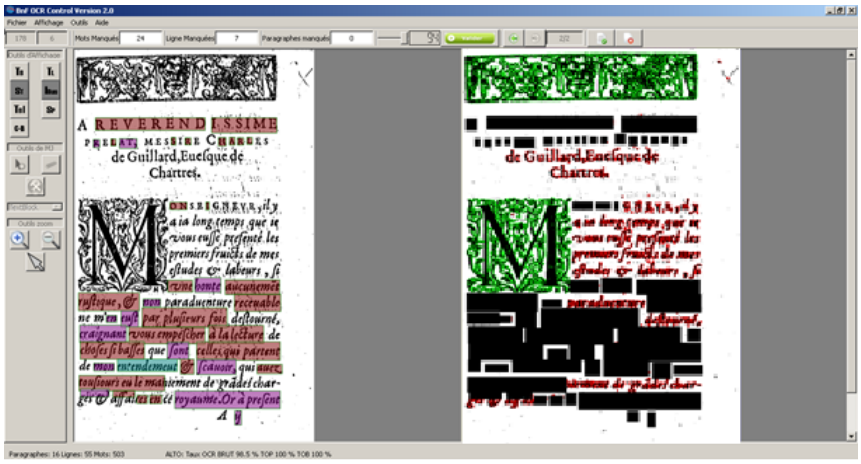


FIGURE 3.26 – Outil de vérification semi-automatique des éléments de la page omis

6 Conclusion

Nous avons présenté dans ce chapitre une approche de détection des éléments textuels omis dans les résultats de transcription automatique des documents numérisés dans le cadre des projets de numérisation de masse. Les structures physiques transcrites par les OCR commerciaux ne sont pas parfaites. Elles englobent généralement des erreurs de fusion, de scission ou de typage d'éléments de la page. Ces erreurs sont détectées par des processus de contrôle manuel ou automatique. Par contre, d'autres erreurs comme les éléments textuels omis ne sont pas couvertes par ces procédures d'estimation de qualité. D'où la nécessité d'une procédure de détection de mots omis.

La collection documentaire de la BnF est très variée. Elle englobe des documents caracterisés par des caractéristiques physiques et typographiques très variables. Cela exige l'utilisation de descripteurs génériques qui ne dépendent ni des caractéristiques typographiques des documents ni de leur mise en page. Les descripteurs de texture offrent ainsi une description générique des éléments textuels. En effet, à échelle globale, les zones textuelles ont des propriétés texturales différentes des zones d'illustration ou de fond de page. La direction principale de texture à différentes échelles est la même dans une zone de texte, alors que dans des zones d'illustration ou de fond de page, cette direction principale est variable à différentes échelles. De plus, la variance des directions principales est plus importante sur les zones d'illustration que sur les zones de texte. La fréquence de transition des pixels foncés aux pixels clairs est une caractéristique déterminante pour les zones textuelles. En effet, les zones textuelles ont généralement des fréquences de transition plus importantes que les fréquences des zones d'illustration ou de fond de page.

De plus, nous avons adopté une méthodologie adaptative pour détecter les éléments omis de la page. Cette méthodologie utilise les caractéristiques des éléments détectés par l'OCR pour chercher dans les zones de fond de page des éléments semblables à ce qui ont été détectés. Cette méthodologie se base sur l'hypothèse que l'opération de typage des éléments de la page effectuée par un OCR est assez précise de telle façon que nous pouvons utiliser ses résultats pour former des modèles de classification pour chaque classe d'éléments de page. De ce fait, en utilisant ce principe d'apprentissage, nous avons rendu notre approche adaptable aux caractéristiques typographiques des pages.

Une fois les signatures des différents pixels de l'image obtenues, elles sont renvoyées par la suite à quatre classifieurs SVM appris localement sur la page traitée. Chaque classifieur est utilisé pour prédire la classe des pixels de fond de l'image en utilisant une stratégie un contre tous. Après la classification des pixels, nous procédons à la formation des enveloppes des éléments textuels détectés afin de pouvoir intégrer les résultats de notre approche dans la procédure de contrôle qualité de la BnF.

Les résultats d'évaluation de notre approche ont montré que les performances de notre détecteur sont assez bonnes, puisque 84,15% des éléments omis dans notre base d'évaluation sont détectés. De plus, la procédure de détection est précise puisque le taux de précision de notre détecteur est égal à 94,73%. Le défaut principal de notre méthode réside dans l'utilisation du résultat du typage automatique des éléments de la page, qui peut biaiser l'opération d'apprentissage des classifieurs dans le cas où il y a des erreurs d'étiquetage. Pour résoudre ce problème et limiter l'effet de bruit dans la base d'apprentissage, nous pouvons former nos classifieurs avec des données collectées sur plusieurs pages d'un même document. Un autre problème réside dans le temps de traitement assez long (en moyenne 7 minutes par page) qui est dû au parcours de tous les pixels de l'image de la page induit par la technique de fenêtre glissante.

Chapitre 4

Contrôle des résultats de reconnaissance des caractères

1 Introduction

Au cours du chapitre précédent, nous avons présenté une approche de vérification qui détecte les mots ou portions de mots omis par l'OCR lors de la segmentation des images des pages. Outre ce premier type d'erreur, les résultats de l'OCR comportent également des erreurs de reconnaissance. Cela nécessite l'application d'une procédure de contrôle automatique des résultats de reconnaissance des caractères pour déterminer la qualité de la transcription.

A la BnF, la qualité des résultats de l'OCR est exprimée à l'échelle des mots en utilisant les scores de confiance « **word confidence** » calculés automatiquement par l'OCR. Comme nous avons déjà eu l'occasion de le mentionner, cette approche à tendance à surestimer la qualité des transcriptions produites car elle ne prend pas en compte les mots oubliés. Ainsi ce n'est pas le taux de reconnaissance de caractères qui est estimé mais plutôt la confiance dans la qualité des mots reconnus. La plupart du temps, l'OCR utilise un dictionnaire pour valider les hypothèses de reconnaissance des caractères, et le score de confiance mots est très vraisemblablement calculé en tenant compte de l'occurrence du mot dans le dictionnaire. Ainsi, les mots hors-lexique vont avoir tendance à présenter des scores de confiance plus faibles ou à être confondus avec le mot le plus proche présent dans le dictionnaire. De plus, la méthode d'estimation des taux de reconnaissance utilisée dans les projets de numérisation (fondée sur le word confidence) varie selon les prestataires et les systèmes OCR utilisés, tandis que les performances à atteindre sont définies a priori dans le cahier des charges avant le lancement des projets de numérisation. A la BnF, le contrôle manuel des résultats des prestataires est réalisé sur des échantillons prélevés aléatoirement. Par conséquent, la BnF ne maîtrise pas complètement la qualité des documents numériques intégrés dans Gallica, ce qui peut menacer l'intégrité de sa bibliothèque numérique.

Ainsi que la qualité des résultats de reconnaissance fournis par les prestataires ne peut être garantie complètement, la BnF, pour sa part, ne peut engager les moyens humains qui seraient nécessaires pour les vérifier totalement. Cette situation milite pour que des moyens de contrôle automatique soient conçus afin de traiter la masse des documents numérisés chaque mois. La difficulté à développer un tel process de contrôle automatique tient d'une part à la masse de documents à traiter, et d'autre part à la nécessiter de développer une

méthodologie qui ne dépende pas d'une vérité terrain, dont on ne peut pas disposer par définition.

Pour résoudre ce problème, nous proposons dans ce chapitre une approche automatique d'estimation des taux de reconnaissance des caractères qui permet d'estimer la qualité des résultats de l'OCR sans avoir recours à une vérité terrain ni à la détermination des erreurs de reconnaissance. L'état de l'art a permis de mettre en évidence un ensemble de méthodes de contrôle des résultats de reconnaissance. Ces méthodes caractérisent les résultats de reconnaissance à l'aide de différents critères qui constituent souvent les caractéristiques d'un système de rejet entraîné spécifiquement à détecter les erreurs de reconnaissance. L'approche que nous proposons s'inspire de ces méthodes dans le sens où nous cherchons aussi à caractériser les résultats de reconnaissance à l'aide de différents indices. Ces indices vont ensuite alimenter un système de prédiction de taux de reconnaissance des caractères à l'échelle MACRO de la page, contrairement aux systèmes de vérification de reconnaissance de la littérature qui essayent de détecter les erreurs pour qualifier la qualité de reconnaissance des caractères.

Ce chapitre va donc s'articuler comme suit : nous débutons par l'examen des causes d'erreurs de reconnaissance des caractères. Nous poursuivons en présentant la méthodologie que nous avons développée. Nous exposons les différents descripteurs que nous avons employés pour caractériser les résultats de reconnaissance de caractères ainsi que les approches de régression utilisées pour estimer les taux de reconnaissance de caractères. Enfin, nous terminons cette section par la présentation des performances de notre approche obtenues lors de la réalisation de différents scénarios d'évaluation.

2 Difficultés liées au contrôle de la reconnaissance

A la BnF, les résultats du processus de numérisation-transcription, stockés au format XML ALTO, englobent à la fois la structure physique de la page, les mots reconnus par l'OCR ainsi que les taux de confiance fournis par l'OCR. Ces taux de confiance servent à certifier que la qualité de la transcription correspond à celle demandée dans le cahier des charges, via l'estimation d'un taux de reconnaissance associé à la transcription (puisqu'aucune vérité terrain n'est disponible pour évaluer sa qualité réelle). La valeur des taux de confiance peut varier entre 0 et 1, par contre lors de la correction manuelle des résultats de reconnaissance toutes les valeurs des taux de confiance associées aux mots sont automatiquement positionnées à 1.

A la BnF, seuls les taux de confiance au niveau mot (*word confidence*) calculés par les OCR sont considérés. Ces scores représentent donc le degré d'adéquation d'un résultat de reconnaissance par rapport au modèle de langue et de formes utilisés par l'OCR. Cependant, la formule de calcul de ces taux de confiance n'est pas connue. Il est d'ailleurs intéressant de remarquer que si le standard ALTO prévoit d'associer un taux de confiance à chaque mot reconnu, elle n'en donne aucune définition. Finalement, ce score *word confidence* est le seul paramètre externalisé par l'OCR et fourni à l'utilisateur pour lui servir de paramètre de qualité du résultat de reconnaissance. Il n'est donc guère étonnant que les études faites sur les résultats de ces processus de numérisation montrent que les performances annoncées ne sont pas toujours conformes aux résultats réellement observés. D'où la nécessité de vérifier la qualité de la reconnaissance.

Nous avons montré dans la section « Etat de l'art » que les OCR sont devenus des systèmes de plus en plus complexes composés par plusieurs étages de traitements. Chaque étage applique un ensemble de traitements qui visent à améliorer au mieux la qualité des résultats de reconnaissance. Du fait de l'agencement séquentiel de ces traitements, les erreurs commises à chaque étage de traitement ont un impact direct sur les résultats de la transcription automatique des mots. Des erreurs de binarisation peuvent déformer les formes des caractères ; de même, des erreurs de segmentation peuvent causer des erreurs de scission et de fusion des mots ce qui peut causer des erreurs dans les résultats finaux de l'OCR.

Le mode de fonctionnement en « boîte noire » des OCR commerciaux rend l'analyse et la détermination des sources d'erreurs difficiles. En effet, dans certains cas, nous trouvons des séquences de caractères qui n'ont pas de sens et qui ne correspondent à aucun mot d'un lexique susceptible d'avoir été appliqué à l'étage de post-traitement de l'OCR. D'autres erreurs de reconnaissance conduisent de même à des transcriptions erronées. Ces erreurs peuvent par exemple être causées par l'application abusive d'un modèle de langage. A cela, il faut également ajouter que le comportement des OCR reste fortement dépendant des documents fournis en entrée du processus. Par conséquent, il est souvent difficile d'affirmer exactement quelle est la source d'erreur. Il semble donc vain de tenter de qualifier les sorties de l'OCR vis-à-vis d'un ensemble d'erreurs typiques observées qui pourraient être détectées par des procédures automatiques spécifiques.

Ce constat sur la relative imprévisibilité des erreurs de reconnaissance des OCR nous a amené à développer une méthodologie pour laquelle nous n'avons pas cherché à exploiter une typologie d'erreurs comme cela est fait traditionnellement dans la littérature. Nous avons au contraire retenu une méthodologie plus globale et intégrant moins d'a priori pour développer une solution d'estimation du taux de reconnaissance de l'OCR.

Par ailleurs, à la BnF, la granularité de la qualité de reconnaissance considérée dans les projets de numérisation de masse n'est pas fine. Les taux de confiance annoncés dans le fichier ALTO de la BnF sont obtenus en calculant les moyennes des taux de confiance des mots qui se trouvent dans les pages. C'est par abus de langage qu'à la BnF on désigne ce taux « taux de reconnaissance ». Ceci étant, dans les lignes qui suivent et qui décrivent les pratiques de la BnF, nous gardons cette même dénomination. La méthode que nous proposons dans cette partie s'attache à estimer le taux exact de reconnaissance des mots qui est disponible sur les ensembles de documents qui ont été mis à notre disposition pour cette étude. A la BnF, on distingue généralement deux qualités de résultats d'OCR : la qualité brute, et la qualité supérieure « HQ ». Les taux de reconnaissance estimés lors du processus de transcription automatique varient généralement entre 60% et 98%. En revanche, pour les documents en qualité HQ, les taux de reconnaissance exigés doivent être supérieurs à $99,95\%$. Le coût de la transcription des documents en qualité HQ est presque le double du cout de transcription des documents en qualité brute. A cause des contraintes financières, la majorité des documents numériques de la BnF sont transcrits en qualité brute.

Pour la BnF, et dans le cas d'une numérisation « brute », il n'y a aucune différence entre un document converti avec un taux de 60% et un autre document qui serait converti avec un taux de 98%. On voit bien que dans cette stratégie de numérisation « brute », la précision sur le taux de reconnaissance n'est pas vraiment primordiale. Ce constat nous amène à

penser qu'il n'est donc pas nécessaire de détecter exactement les erreurs de reconnaissance dans les résultats de l'OCR, (quant bien même cela serait envisageable en mettant en oeuvre un système de rejet). C'est pourquoi nous proposons dans cette partie plusieures méthodes originales d'estimation du taux de reconnaissance qui s'attachent à qualifier les résultats de l'OCR à l'échelle de la page ou du document tout en s'affranchissant du score de confiance fourni par l'OCR qui n'est pas un paramètre clairement défini.

3 Estimation du taux de reconnaissance

Nous avons donc fait le choix de développer une approche visant à évaluer la qualité de la transcriptions des caractères en utilisant un processus que nous qualifions de « macro » procédant au niveau de la page ou d'un ouvrage et permettant d'apporter une réponse qualitative quant à la qualité de la transcription (l'OCR a produit des résultats médiocres, satisfaisants, très satisfaisant, exceptionnels, etc.). Naturellement, plus cette réponse sera précise et proche du taux de reconnaissance exact et plus la méthodologie développée sera jugée pertinente.

Ainsi, nous avons cherché à construire une nouvelle méthodologie (jamais abordée dans la littérature à notre connaissance) en cherchant à caractériser du mieux possible les qualités et les limites de cette méthodologie. Au lieu de tenter de détecter les mots incorrects dans les résultats de l'OCR, nous proposons une démarche qui caractérise le comportement de l'OCR sur la page et qui exploite cette caractérisation pour déduire la performance que réaliserait un OCR sur la page. Les principes qui nous ont guidés dans la construction de notre méthodologie peuvent s'énoncer comme suit :

1. Au lieu de tenter de détecter les mots ou les caractères incorrects dans les résultats de l'OCR comme cela se fait habituellement, nous avons fait le choix de développer une procédure visant à évaluer la qualité de la transcription à un niveau macro (au niveau de la page ou d'un ouvrage), en nous basant sur les spécificités de l'ensemble des caractères.
2. Nous limitons notre approche à la caractérisation du comportement d'un OCR relativement à un corpus déterminé. En effet, de la même façon qu'il n'existe pas d'OCR universel, il nous semble impossible de construire un système qui fonctionne sur tous types de documents sans adaptation spécifique.
3. L'estimateur de qualité pourra être adapté à chaque corpus et en fonction de l'OCR prestataire utilisé. L'adaptation se fera par l'introduction d'une étape de calibration du système. Cette étape consistera en la production d'un échantillon étiqueté du corpus (production d'une vérité terrain en quantité limitée par des opérateurs) puis en l'apprentissage de la procédure de contrôle. Cette démarche s'inscrit parfaitement dans le processus que souhaite mettre en place la BnF avec ses prestataires, processus qui prévoit la fourniture de données étiquetées (vérité terrain) pour une certaine fraction des corpus traités. Nous allons donc employer cette vérité terrain pour produire des données qui serviront à l'étalonnage.
4. Nous souhaitons développer une méthodologie qui s'affranchisse le plus possible des dictionnaires et modèles de langage, car ces ressources ne sont pas toujours disponibles et il existe parfois des variations locales qui peuvent mettre en défaut les OCR

et qui pourraient donc faire de même avec une procédure de contrôle, la rendant ainsi inefficiente.

5. Nous souhaitons développer une méthodologie qui s'affranchisse également de connaissances typographiques car elles nécessitent une expertise trop longue et coûteuse.

Pour répondre à ces exigences pragmatiques, nous proposons donc d'élaborer une méthodologie visant à construire un système de vérification adapté à chaque type de corpus et en fonction des besoins des futurs marchés de la BnF. Cette méthodologie nous a amené à explorer quatre approches d'estimation du taux de reconnaissance de caractères. La présentation de ces approches est réalisée dans l'ordre chronologique que nous avons adopté pour mener nos expérimentations. La première procédure d'estimation du taux de reconnaissance s'appuie sur le principe d'isogénie des formes des caractères d'un même document ou corpus (les mêmes caractères sont représentés par les mêmes formes de polices). La deuxième procédure que nous avons explorée s'appuie sur l'utilisation d'un OCR tiers qui sert de référence relative en l'absence de vérité terrain. Ces deux estimateurs utilisent une procédure de régression polynomiale pour estimer les taux de reconnaissance des caractères. Le choix du degré du polynôme est très important pour garantir la qualité des résultats d'estimation des taux de reconnaissance. Pour optimiser ce choix nous avons utilisé l'indicateur de Mallow qui est une estimation de l'erreur quadratique moyenne de prévision. Le troisième estimateur exploite simultanément les propriétés d'isogénie et l'alignement sur un OCR tier. L'utilisation de ces deux descripteurs permet d'examiner la complémentarité des deux familles de caractéristiques. L'estimation du taux de reconnaissance est réalisée en utilisant une régression à base de vecteurs supports (*Support Vector Regression*). Enfin, le quatrième estimateur est une version modifiée du troisième estimateur. Elle emploie une procédure de sélection des documents de la base d'apprentissage qui permet de spécialiser l'estimateur pour chaque document traité. Cette procédure d'adaptation offre la possibilité de généraliser le système à de plus grandes collections de documents (par rapport à l'ensemble réduit qui a été utilisé pour les expériences réalisées dans cette thèse), en exploitant un historique actualisé des corpus déjà traités. Cette ultime approche, qui est la plus aboutie de ce travail est donc générique dans sa démarche, et adaptable aux différents besoins.

L'ensemble des procédures d'estimation que nous proposons dans ce travail utilise le schéma de traitements suivant :

1. Préparation des données d'apprentissage soit par la segmentation des caractères des mots, soit par l'utilisation d'un deuxième OCR.
2. Description des résultats de reconnaissance des caractères produits par le prestataire.
3. Apprentissage des estimateurs de taux de reconnaissances des caractères.
4. Estimation des taux de reconnaissance des caractères.

3.1 Approche par contrôle de l'isogénie des caractères

Cette approche exploite le principe d'isogénie des caractères au sein d'un même document pour déterminer dans quelle mesure les mêmes caractères reconnus présentent le même aspect. En effet, la variabilité des formes des caractères appartenant à la même classe représente une information pertinente pour estimer l'exactitude des résultats de

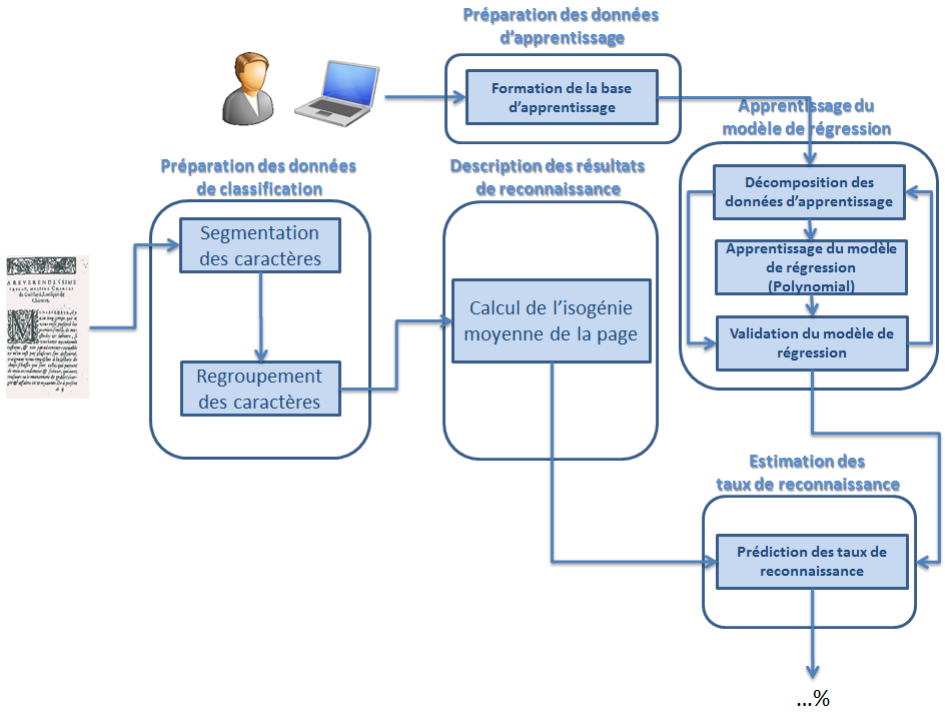


FIGURE 4.1 – Etapes des traitements de l'approche d'estimation des taux de reconnaissance de caractère qui utilisent les distances intraclasses de caractère

reconnaissance. En principe il doit y avoir une faible variabilité des formes au sein d'une même classe [SBZ03]. D'après Xiu et al. [XB12] les déformations des formes de caractères peuvent engendrer des erreurs de reconnaissance de caractères qui peuvent se répéter à différents endroit dans l'image et dans l'ensemble du document. Ceci signifie que plus la différence entre les formes des caractères appartenant à la même classe est importante plus nous risquons d'avoir des confusions dans les résultats de l'OCR.

Pour tenter d'exploiter ce principe nous avons développé une approche qui procède selon le schéma de la figure 4.1. Dans une première étape, nous exploitons les résultats de reconnaissance fournis par les prestataires et sauvegardés au format ALTO pour extraire les lignes de texte reconnues et localiser les caractères dans les lignes. Il faut souligner ici qu'une étape de détection des caractères est mise en place car la position des caractères reconnus n'est pas renseignée dans les fichiers ALTO. Seules les positions des mots sont renseignées. Pour chaque mot on cherche donc à aligner la séquence de caractères reconnus sur les pixels de l'image du mot. On utilise pour cela un système de reconnaissance développé dans le cadre du projet *NAVIDOMASS* par Kamel Ait Mohand dans le cadre de sa thèse [AM11] et qui permet, en l'utilisant dans un mode de fonctionnement dit par alignement forcé, de réaliser la segmentation en caractères recherchée.

Caractéristiques d'isogénie des formes des caractères

Dans une deuxième étape, nous exploitons l'alignement obtenu pour regrouper classe par classe les images des caractères de la page. Le degré d'isogénie de chaque classe de caractère Isg_c est alors estimé en calculant la distance de Hamming moyenne pour la classe

« c » selon la formule 4.3. Dans cette formule, la distance de Hamming est calculée sur les bitmaps des caractères qui sont préalablement ramenés à une taille normalisée (15×15). Pour appliquer cette distance, nous commencons tout d'abord par la formation de l'image de référence « a » de chaque classe de l'alphabet. Cela va nous servir à comparer les images des caractères avec leurs formes moyennes respectives de la page. Nous formons donc les images de référence en calculant la moyenne de toutes les images « b » appartenant aux mêmes classes de caractères (cf. équation 4.1).

$$a = \frac{1}{L_c} \sum_{l=1}^{L_c} b_l \tag{4.1}$$

Pour calculer la distance de Hamming entre une image de caractère b et une image de référence a, nous avons utilisé l'équation 4.2 où L_c est le nombre de caractères appartenant à la classe « c », n est le nombre de caractères de l'alphabet qu'on cherche à vérifier. L'opérateur $\oplus$ désigne le ou exclusif qui permet de déterminer les différences entre les valeurs des pixels des deux images. Par conséquent, $d(a, b)$ désigne le nombre de pixels différents entre une image de caractère et l'image de référence de sa classe.

$$\forall b \in F \qquad b = (b_i)_{i\in[0,n-1]} \quad et \quad a = (a_i)_{i\in[0,n-1]} \quad d_l(a,b) = \sum_{i=0}^{n-1} (a_i \oplus b_i) \tag{4.2}$$

$$Iseg_c = \frac{1}{L_c} \sum_{l=1}^{L_c} d_l \tag{4.3}$$

Finalement l'indice d'isogénie moyen est calculé selon la formule 4.4 sur le document en faisant la moyenne des degrés d'isogénie de chaque classe de caractère (calculées selon l'équation 4.3).

$$f_1 = \quad \frac{\sum_{c=1}^{NC} Iseg_c}{NC} \quad avec \quad NC = 26 \tag{4.4}$$

La dernière étape des traitements de ce prédicteur exploite l'indice d'isogénie pour estimer le taux de reconnaissance de la page. Pour cela, on recourt à un modèle numérique de prédiction constitué d'une régression polynomial dont les paramètres (ordre du polynôme et valeurs des coefficients) sont optimisés sur un ensemble de documents d'apprentissage et pour lesquels on dispose du taux réel de reconnaissance de caractères. Cette procédure d'apprentissage est réalisée en utilisant une procédure de validation croisée et une technique de *grid search*.

La figure 4.2 représente la distribution d'un échantillon de 230 pages le taux de reconnaissance des caractères et la distance moyenne intra-classe de caractères. D'après cette représentation, nous constatons pour la plupart des pages de cet échantillon, plus les indices d'isogénie moyen sont faibles plus les taux de reconnaissance de caractères sont importants. Ceci signifie l'existence d'une relation de corrélation entre ces deux variables. Par contre, certaines pages sont caractérisées par des faibles taux de reconnaissance et des faibles indices d'isogénie. Ces exemples sont trés difficiles à traiter et causent généralement des erreurs d'estimation considérables.

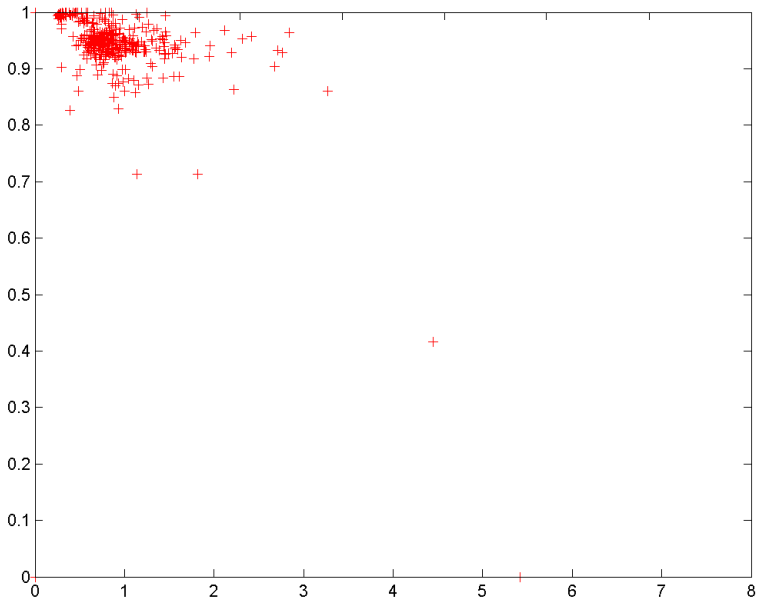


FIGURE 4.2 – Représentation de 230 pages selon le taux de reconnaissance des caractères et la distance moyenne intra-classe de caractères.L'axe d'abscisse représente les distances moyennes intra-classes de caractères, l'axe des ordonnées représente des taux de désaccord.

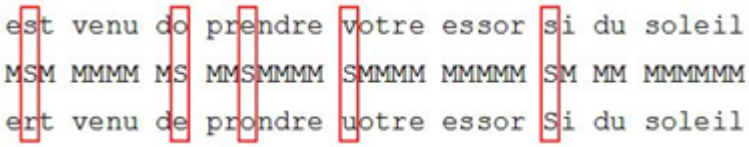


FIGURE 4.3 – Un exemple de résultat d'alignement des caractères des mots de l'OCR des prestataires (en haut) avec les caractères des mots de l'OCR de validation (en bas). Au milieu, les opérations d'éditions nécessaires (M=matching ; S=Substitution ; D=suppression). Les cadres rouges représentent les caractères en désaccord entre deux OCR.

3.2 Approche par alignement sur un second OCR

Dans cette approche nous tentons de nous rapprocher des approches classiques d'estimation de performances qui utilisent une vérité terrain. Comme nous ne disposons pas d'une telle vérité terrain, nous créons une pseudo vérité terrain qui va être fournie par un second OCR. Il devient alors possible de caractériser les résultats de reconnaissance par rapport à ceux de l'OCR auxiliaire en procédant à l'alignement des résultats de l'un sur l'autre (figure 4.3). Ces statistiques vont ainsi constituer des descripteurs à partir desquels nous allons tenter de faire une prédiction du taux de reconnaissance. Lors des expérimentations que nous avons réalisées dans cette partie, nous avons utilisé l'OCR qui a été conçus dans le cadre du projet NAVIDOMAS par Kamel Ait Mohand. Ce système procède à la reconnaissance des caractères sans utiliser de connaissances lexicales, les résultats qu'il fournit ne sont donc pas biaisés par des post-traitements linguistiques ce qui assure à notre méthode une certaine robustesse vis à vis de ces phénomènes.

La figure 4.4 présente le schéma des traitements de cette approche. La procédure commence par une étape de préparation des données qui se traduit par l'application du deuxième OCR sur les images des documents à vérifier. Pour simplifier l'opération de mise en correspondance des résultats des deux OCR, nous utilisons les structures physiques des pages obtenues par l'OCR du prestataire pour produire les transcriptions du deuxième OCR. Cela signifie que les deux résultats de reconnaissance de caractères sont obtenus à partir du même processus de segmentation des documents. L'alignement des deux OCR met en oeuvre les opérations d'édition usuelles (Ajout, Suppression, Substitution) pour produire les données d'alignement. Cette procédure est similaire à celle qui est réalisée lors de l'alignement sur la vérité terrain. L'alignement réalisé permet d'obtenir une description complète des zones de texte qui coïncident, des caractères insérés, des caractères omis ainsi que des caractères substitués. A partir de ces résultats d'alignement, nous extrayons une caractéristique globale qui permet de décrire les résultats de reconnaissance à l'échelle de la page ou du document. Cette caractéristique est basée sur le taux de désaccord par classe de caractères obtenu dans le fichier OCR de prestataire. Il est défini par l'équation 4.5. On déduit ensuite le taux de désaccord moyen sur une page qui est la moyenne des taux de désaccord par classe de caractères dans la page, (équation 4.6).

Les caractères minuscules sont plus présents que les caractères majuscules ou chiffres dans les documents de la BnF. Par conséquent, pour former la signature des désaccords de chaque page, nous avons choisi de ne traiter que les caractères minuscules pour calculer le taux moyen de désaccord.

$$p(\delta_i) = \frac{nombre \quad des \quad caractères \quad \delta_i \quad en \quad désaccords}{nombre \quad de \quad caractères \quad \delta_i} \tag{4.5}$$

avec $i = 1...26$ et $\delta_i \in \{a, b, c, ..., z\}$

$$f_2 = \frac{1}{n} \sum_{i=1}^{n} p(\delta_i) \tag{4.6}$$

La dernière étape de cette méthode exploite la caractéristique ainsi obtenue pour estimer le taux de reconnaissance de la page. Pour cela on utilise à nouveau un système de prédiction qui modélise la relation qui existe entre le taux moyen de désaccord entre les deux OCR et le taux de reconnaissance. La figure 4.5 représente la distribution d'un échantillon de 230 pages selon le taux de désaccord des OCR et de reconnaissance de caractère. Nous constatons que cette distribution suit une allure non-linaire décroissante. En effet, plus les taux de désaccord sont importants moins les taux de reconnaissance de caractères sont bons. Certaines pages présentent des taux de désaccord supérieurs à 50% alors que leurs taux de reconnaissance sont supérieurs à 90%. Ces pages vont être difficiles à traiter et elles vont causer des erreurs considérables d'estimation. Par contre, le nombre des pages présentant ce défaut est minoritaire, ce qui signifie l'existence d'une relation de corrélation entre ces deux variables. En se basant sur cette distribution, nous avons employé un modèle de régression polynomial pour modéliser cette relation. Comme dans l'approche précédente, nous avons utilisé une procédure de calibrage qui permet de déterminer le degré du polynôme de régression.

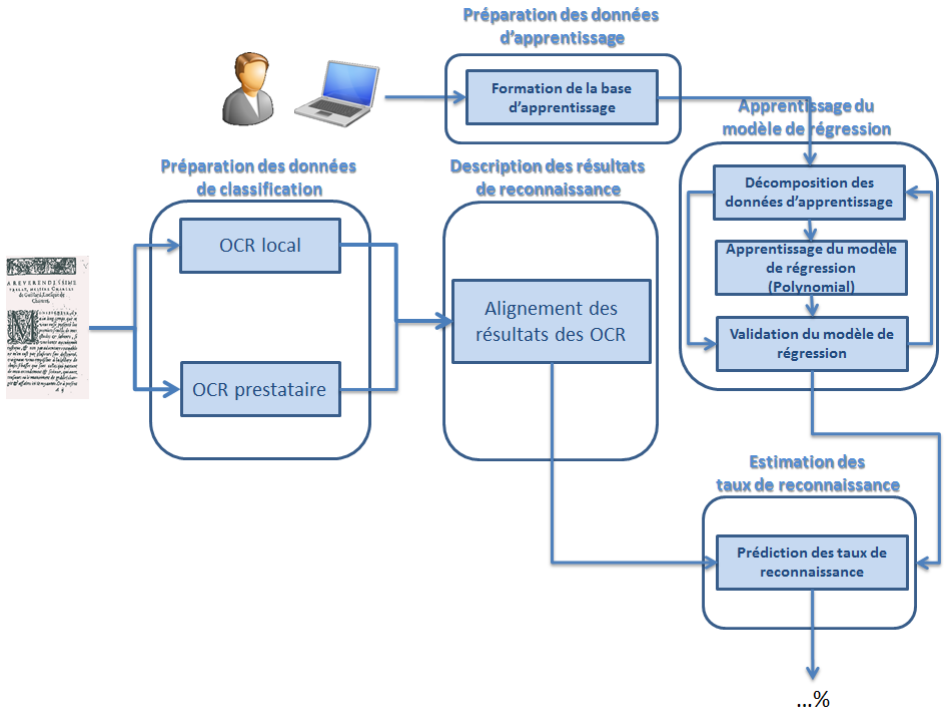


FIGURE 4.4 – Etapes des traitements de l'approche d'estimation des taux de reconnaissance de caractères qui utilise les taux moyens de désaccord

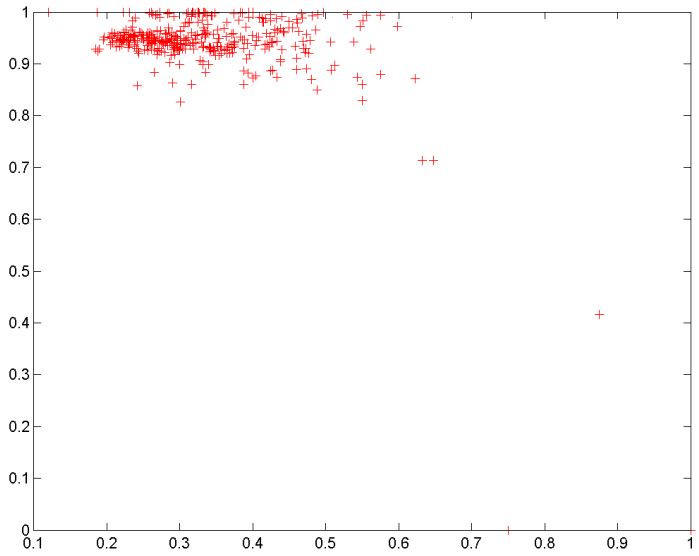


FIGURE 4.5 – Représentation d'un échantillon de 230 pages selon le taux de reconnaissance des caractères et le taux de désaccord entre l'OCR BnF et l'OCR de validation. L'axe des abscisses représente les taux moyens de désaccord, l'axe des ordonnées représente les taux de reconnaissance des caractères.

3.3 Combinaison des caractéristiques

La troisième approche que nous avons développé combine les deux approches précédentes et exploite donc les deux caractéristiques calculées : taux de désaccord et taux d'isogénie des caractères. A titre d'illustration, la figure 4.6 présente la distribution d'un échantillon de 230 pages selon leurs taux de désaccord et leurs distances moyennes intra-classe des caractères. Les pages de cet échantillon sont sélectionnées aléatoirement à partie de notre base de validation. Les couleurs des points indiquent la valeur du taux de reconnaissance des caractères. Sur cette figure, nous remarquons que les pages présentant de faibles taux de reconnaissance sont situées dans le coin supérieur droit. Les pages ayant de bons taux de reconnaissance sont concentrées en majorité autour de l'origine des deux axes, ce qui signifie que la majorité des taux de reconnaissance sont caractérisés par de faibles taux de désaccord et de faibles distances moyennes.

Les documents ayant de forts taux de reconnaissance et de forts taux de désaccord présentent de faibles distances moyennes intra-caractère. De même les documents ayant des forts taux de reconnaissance et des fortes distances moyennes sont caractérisés aussi par un faible taux de désaccord. Par conséquent, nous pouvons déduire que l'utilisation conjointe de ces deux caractéristiques résout en partie le problème des deux approches précédentes. Ceci devrait rendre les estimations de cette troisième approche plus précises.

Le régresseur polynomial simple n'est pas assez robuste pour donner des estimations fiables du taux de reconnaissance. C'est pour cela, nous nous sommes tournés vers un modèle de régression à base de vecteurs support (Support Vector Regression). Cet algorithme de prédiction est une généralisation de l'algorithme de classification SVM (Support Vector Machine) sur des problèmes d'estimation des valeurs de variables. Il opère généralement dans des espaces de caractéristiques de grandes dimensions pour former des fonctions de prédiction développés sur un sous-ensemble de vecteurs de support. La première version de l'algorithme de SVR a été proposée par Vapnick et al dans [VGS96] en 1996. Cet algorithme est caractérisé par deux notions qui sont :

1. La notion du noyau de transformation qui permet de projeter l'espace des caractéristiques des données dans un espace de plus grande dimension ;
2. La notion d'hyperplans à marge maximale qui permet de calculer une fonction de coût en utilisant les données d'apprentissage qui se trouvent en dehors d'une marge de taille ϵ

Cette méthode pose les mêmes difficultés que pour l'approche de classification SVM et qui sont liées :

- au choix des paramètres du régresseur SVR comme la taille de la marge, le type d'algorithme de régression (epsilon-SVR ou nu-SVR), le choix de la fonction noyau qui assure la transformation de l'espace des caractéristiques ;
- à la taille de l'espace de représentation des données d'apprentissage qui peut provoquer une chute de performances lorsqu'il est trop grand, même si, dans la théorie, les SVM sont censés avoir une bonne performance sur des espaces de grande taille.

Dans notre approche, le deuxième problème évoqué ne se pose pas puisque nous n'utilisons que deux variables. Par ailleurs, pour surmonter la première difficulté, nous avons adopté une procédure de recherche de paramètres qui assure l'étalonnage des modèles de régression en utilisant la technique de *grid search* qui effectue l'optimisation des hyper-

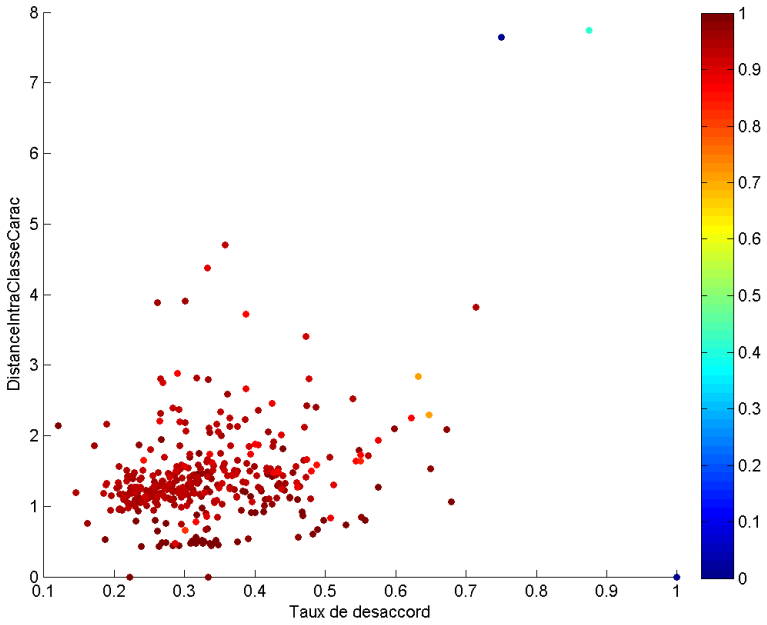


FIGURE 4.6 – Distribution de 230 pages tirées de façon aléatoire en fonction des taux de désaccord et des distances intra-classe de caractère. Les points rouges correspondent aux pages ayant des forts taux de reconnaissance. Les points bleus correspondent aux pages ayant des faibles taux de reconnaissance.

paramètres du modèle de régression en fonction d'une liste de paramètres fixée a priori. Dans notre travail, nous avons optimisé les paramètres suivants :

1. Noyau de transformation, nous avons testé les noyaux linéaire, polynomial, bayésien (RBF) et sigmoïde,
2. Taille de marge de l'hyperplan ϵ qui définit l'intervalle des erreurs d'estimation tolérées. Ceci signifie que toute erreur plus petite que ϵ ne requiert pas une valeur non nulle de l'erreur ξ_i et ne doit pas être prise en compte par la fonction objectif,
3. Le coût, C, qui contrôle le compromis entre la tolérance des erreurs d'estimation et l'utilisation d'une marge rigide pendant l'étape d'apprentissage des modèles de régression. L'utilisation de ce paramètre entraîne la création d'une marge souple pour tolérer des erreurs d'estimation. L'augmentation de la valeur de C cause l'augmentation du coût des mauvaises estimations ce qui force la création d'un modèle plus précis. Ceci peut influencer la capacité de généralisation de l'estimateur.

En plus de la technique de *grid search*, nous avons appliqué la technique de *validation croisée* qui nous permet de valider le choix des hyperparamètres du modèle de régression en répétant les expériences k fois. Pour réaliser cette procédure d'optimisation, nous avons décomposé la base d'évaluation en quatre parties de façon à ce que les documents qui appartiennent à chaque partie soient sélectionnés de façon aléatoire et exclusive. A chaque itération, nous choisissons $k-1$ portions pour l'apprentissage des modèles de régression et une portion pour le test. De plus, en parcourant les listes des valeurs des paramètres de la fonction de régression, nous calculons l'erreur quadratique moyenne obtenue avec

chaque combinaison des valeurs des paramètres. Ensuite, nous conservons la combinaison qui assure l'erreur quadratique minimale.

3.4 Approche par filtrage des données d'apprentissage

Les approches d'estimation des taux de reconnaissance précédentes utilisent une fonction de régression apprise sur l'ensemble des documents de la base d'apprentissage fournis par l'utilisateur au début de chaque opération de contrôle. Cette démarche possède deux inconvénients majeurs. D'une part, l'utilisation de la totalité des documents pendant l'apprentissage rend cette tâche complexe surtout dans le cas où la base d'apprentissage est hétérogène (ce qui est presque toujours le cas dans les projets de numérisation de masse). D'autre part, l'opération de création de la base d'apprentissage, qui doit être semblable aux images des documents à vérifier, doit être renouvelée régulièrement, idéalement pour chaque docment numérisé. Cela engendre donc un travail supplémentaire pour mettre en oeuvre la vérification. En pratique, il n'est pas évident de produire une base d'apprentissage représentative de la totalité des documents qui seront traités lors de la réalisation des projets de numérisation de masse. En effet, des difficultés techniques et économiques compliquent la production d'une base d'apprentissage conséquente et représentative de l'ensemble des pages traitées dans le cadre des projets de numérisation de masse.

Pour résoudre le problème de la représentativité des données d'apprentissage et maîtriser la phase de définition des modèles de régression en fonction des documents à vérifier, nous proposons une technique de filtrage des documents avant la réalisation de la phase d'apprentissage. L'objectif de cette procédure est de sélectionner uniquement les documents ayant des caractéristiques semblables à celles des documents traités. Ainsi, l'opération d'apprentissage du modèle de régression est réalisée pour chaque page pour laquelle une estimation du taux de reconnaissance est requise. Cet apprentissage est donc dédié à la page courante et dépendra essentiellement de ses caractéristiques. Ceci rend cette approche d'estimation parfaitement adaptable à la volée aux documents traités, bien que moins rapide. Puisque les procédures de caractérisation des résultats de l'OCR sont identiques à celles de la méthode précédente, nous allons nous contenter de présenter la procédure de sélection à la volée des documents utilisés pour entraîner le régresseur.

En suivant le principe des moteurs de recherche d'information, nous allons mettre en œuvre une méthode d'indexation des documents de la base d'apprentissage en utilisant les profils de résultats de l'OCR prestataire. Chaque page est caractérisée par une signature composée par les x confusions les plus fréquentes. Ces confusions ne portent que sur les caractères minuscules et majuscules qui sont très présents dans les documents (les chiffres sont plus rares, notamment dans les monographies et journaux). Cette restriction permet de garantir la qualité de la signature des pages pour la tâche qui nous intéresse ici.

Le nombre x de confusions utilisé pour former la signature de la page est un paramètre très important dans notre approche. En effet, d'après la figure 4.7 plus le nombre de confusions est important moins il est possible de trouver des pages similaire dans la base d'apprentissage. Ce résultat est logique puisque l'augmentation du nombre des confusions rend la requête de sélection des documents plus sélective. Dans ce cas nous obtenons des exemples plus semblables à la page à vérifier mais moins nombreux. La contrepartie est que nos prédicteurs perdront leurs capacités de généralisation. Le nombre optimal de confusions sera celui qui garantira une bonne qualité des estimations des taux de reconnaissance.

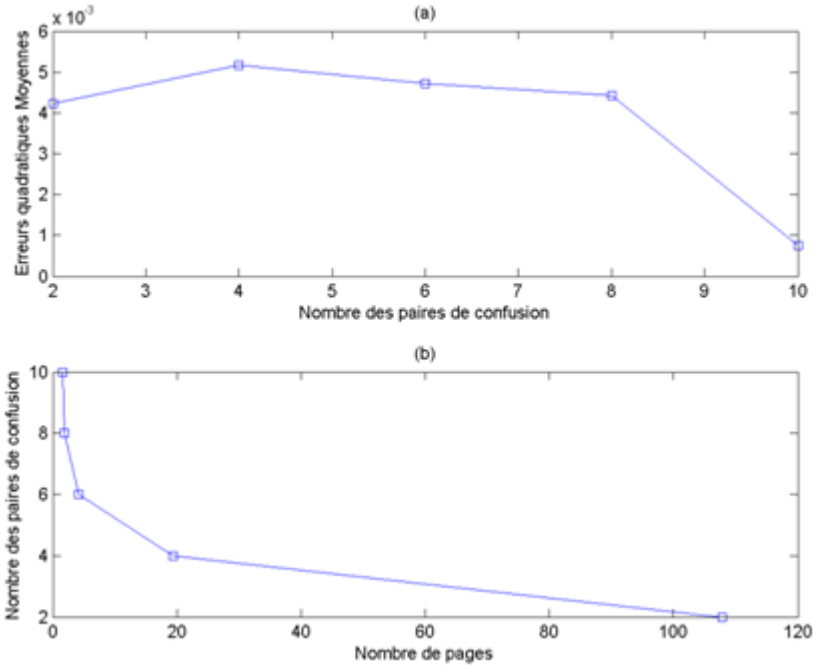


FIGURE 4.7 – Variation de l'erreur quadratique moyenne et du nombre des pages d'apprentissage en fonction du nombre des paires de désaccord qui constituent la signature de la page.

La procédure d'optimisation du nombre des confusions x est réalisée par une approche de validation croisée en 4 plis. Pour chaque itération, nous utilisons une seule partie pour tester notre approche et les trois autres pour apprendre notre modèle de régression avec une valeur du nombre de confusions x donnée. Les performances du système pour cette valeur de x sont données par la moyenne des performances sur les 4 parties. Nous avons fait varier x de 1 à 10 confusions les plus fréquentes. Finalement, un profil de page composé des trois confusions les plus fréquentes permet d'obtenir dans $97,3\%$ des pages testées une estimation du taux de reconnaissance avec une erreur quadratique moyenne égale à $4,71\%$. Par conséquent, tout au long de cette étude, nous allons employer ce nombre de paires de désaccord dans toutes les expérimentations de cette approche.

4 Evaluation

Pour évaluer nos différentes approches, nous avons employé la base « 5 SIECLES » que celle qui a été utilisée pour la détection des mots omis. Cette base de document est très particulière. D'une part elle couvre 5 siècles d'imprimerie ce qui signifie qu'elle est composée de documents anciens caractérisés par des défauts physiques et des difficultés linguistiques, et de documents récents de bonne qualité. De plus, elle englobe différents types de polices (Gothique), de mises en page (textes normaux ou à double colonnes), de langues (latin, français, allemand), de graphiques (tramés et avec traits). Toutefois, la politique de numérisation de la BnF favorisant la numérisation des documents dont le support est de bonne qualité, le nombre de documents difficiles est minoritaire par rapport

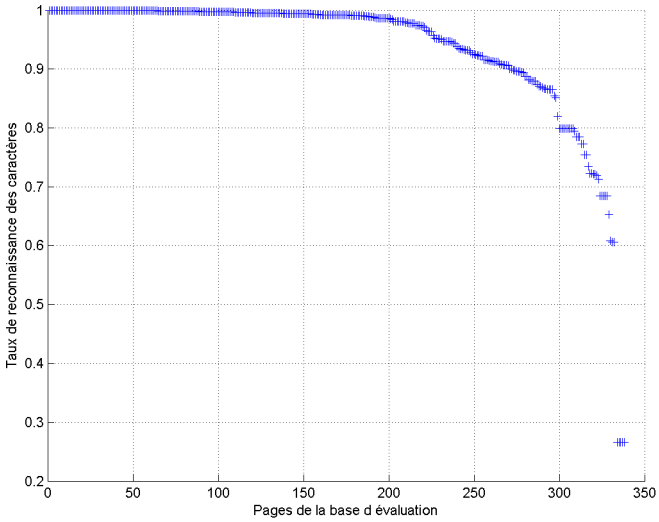


FIGURE 4.8 – Distribution des taux de reconnaissance des caractères de la base des documents d'évaluation

au nombre des documents de bonne qualité.

Les caractéristiques physiques et typographiques des documents de mauvaise qualité sont très variables. Cependant, les images des documents faciles sont généralement de meilleure qualité que les images de la base des documents difficiles. Les typographies de ces documents ainsi que leur mise en page sont généralement uniformes. La langue utilisée dans les documents récents est la langue française moderne gérée par la plupart des systèmes d'OCR commerciaux. Par contre, la langue employée dans les documents anciens est le vieux français ainsi que d'autres langues latines. De ce fait, les documents récents ne causent pas beaucoup de difficultés à l'OCR alors que la reconnaissance des caractères des documents anciens est difficile. La figure 4.8 présente simultanément la distribution des taux de reconnaissance des caractères des pages récentes et anciennes. D'après cette figure, les documents récents sont reconnus avec des taux de reconnaissance supérieur à 90%. Par contre, les taux de reconnaissance des documents difficiles sont plus faibles. En particulier, une cinquantaine ont des taux de reconnaissance inférieurs 80%.

Afin d'évaluer les résultats de nos approches, nous avons analysé les résultats d'estimation des taux de reconnaissance de nos approches en employant en même temps les bases des documents faciles et difficiles. Ceci nous a permis de déterminer le comportement et les performances de nos prédicteurs sur une base de documents hétérogène semblable à celle qui est traitée dans les projets de numérisation de masse.

Afin de déterminer la capacité de notre approche dans un contexte de contrôle des résultats de reconnaissance, nous avons également utilisé les estimations de la quatrième approche pour réaliser un système de rejet automatique des documents par rapport à un seuil de qualité fixé au préalable par la BnF.

4.1 Métriques pour l'évaluation des performances

Analyse des erreurs d'estimation des taux de reconnaissance

Pour pouvoir évaluer les performances des approches d'estimation des taux de reconnaissance, il nous semble important de rappeler ici les propriétés des fonctions d'estimation afin de déterminer par la suite les métriques adéquates pour notre contexte.

En statistique, l'estimation est la procédure permettant d'inférer la valeur d'une variable obtenue à partir d'un échantillon créé par exemple lors de la réalisation d'un sondage. L'estimation d'un paramètre de la population peut être faite de deux façons :

1. Estimation ponctuelle : définie par une valeur unique d'une statistique. Par exemple, la moyenne d'un échantillon $\overline{x}$ est une estimation ponctuelle de la moyenne d'une population μ.
2. Estimation par intervalle : consiste à calculer, à partir d'un estimateur choisi $\widehat{\theta}_n$, un intervalle dans lequel il est vraisemblable que la valeur correspondante du paramètre se trouve.

L'utilisation fréquente des estimateurs ponctuels fait que l'on souhaite qu'ils possèdent certaines propriétés. Ces propriétés sont importantes pour choisir le meilleur estimateur qui s'approche le plus possible du paramètre à estimer. Traditionnellement, on peut qualifier la qualité d'un estimateur en fonction du biais d'estimation défini par l'équation 4.7 et par la variance de l'estimation définie par l'équation 4.8.

$$B(\hat{\theta}_n) = E(\hat{\theta}_n) - \theta \tag{4.7}$$

$$VAR(\hat{\theta}_n) \quad = E[(\hat{\theta}_n - E(\hat{\theta}_n))^2] \tag{4.8}$$

L'erreur quadratique moyenne est un outil très utile dans la comparaison de plusieurs estimateurs. Elle peut être exprimée en fonction du biais d'estimation et de la variance (cf. équation 4.9). Un estimateur de bonne qualité est un estimateur sans biais et de variance très faible. La convergence de l'estimateur est aussi un indicateur assez fiable sur la qualité de l'estimateur. Un estimateur $\widehat{\theta}_n$ est convergent si sa distribution tend à se concentrer autour de la valeur inconnue à estimer en augmentant la taille de l'échantillon traité.

$$MSE(\hat{\theta}) = Biais(\hat{\theta})^2 + Var(\hat{\theta}) \tag{4.9}$$

Par conséquent, pour s'assurer de la fiabilité de nos estimateurs, il faut que ces propriétés soient vérifiées par les estimateurs que nous avons développés. Comme nous le verrons dans la section 4.2, les estimations de nos approche sont asymetriques et ne suivent pas une loi normale. Par conséquent, on ne peut donc pas déduire des moments d'ordre un et deux ni déterminer l'intervalle de confiance de nos estimateur.

Qualitativement, la forme « *piquée* » des histogrammes plutôt que « *en cloche* » est un indicateur du très bon comportement du prédicteur. Par conséquent pour évaluer la précision des estimations, nous avons analysé les erreurs d'estimation commis par nos approches en utilisant une analyse des centiles d'erreur d'estimation qui nous a permis d'étudier la distribution des erreurs d'estimation en éliminant les résultats extrêmes. La première métrique que nous avons examiné dans cette évaluation est l'étendue des intervalles d'erreurs d'estimation. Ceci nous donne une idée de l'importance des erreurs d'estimation obtenues

à chaque proportion des centiles d'erreurs. Nous avons utilisé comme deuxième indicateur la racine carrée de l'erreur quadratique moyenne :

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(f\left(x_i\right) - y_i\right)^2} \quad (4.10)$$

La $RMSE$ caractérise la « *distance* » entre les valeurs réelles « y_i » et estimées « $f\left(x_i\right)$ » des taux de reconnaissance. Plus la valeur de $RMSE$ est faible plus les estimations de nos prédicteurs sont précises.

La troisième métrique dans notre évaluation est le coefficient de corrélation quadratique (r-squared) défini par l'équation 4.11. Ce coefficient permet de mesurer l'intensité de la liaison entre deux caractères quantitatifs. C'est donc un paramètre important dans l'analyse de régression. D'après l'équation 4.11, nous constatons qu'en alignant les taux estimés avec les taux réels, nous pourrons vérifier l'existence de liaisons de corrélation entre les taux de reconnaissance exacts et estimés.

$$CCQ = \frac{\left(n\sum_{i=1}^{n} f\left(x_i\right) - y_i - \sum_{i=1}^{n} f\left(x_i\right)\sum_{i=1}^{n} y_i\right)^2}{\left(n\sum_{i=1}^{n} f\left(x_i\right)^2 - \left(\sum_{i=1}^{n} f\left(x_i\right)\right)^2\right)\left(n\sum_{i=1}^{n} {y_i}^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right)} \quad (4.11)$$

Ce coefficient prend une valeur nulle s'il n'existe pas de relation linéaire entre les résultats estimés et les résultats réels. Contrairement, plus la valeur de ce coefficient est proche de 1 plus les relations entre les résultats estimés et les résultats réels sont forts. Cet indicateur mesure donc le degré de variation des estimations de nos approches par rapport aux variations des taux de reconnaissance réels des pages.

Test de la capacité de rejet automatique des documents

En plus de l'évaluation des résultats d'estimation des taux de reconnaissance, nous avons testé aussi la capacité de rejet des documents du meilleur estimateur de taux de reconnaissance des caractères. En effet, l'objectif ultime de l'étude réalisée dans ce chapitre est de trouver les moyens adéquats qui nous permettent de contrôler les résultats de reconnaissance des caractères obtenus lors d'une procédure de numérisation de masse.

Pour cela, à l'instar de la procédure de contrôle des résultats de l'OCR de la BnF, nous avons procédé au rejet des documents ayant des taux estimés inférieurs à un seuil de reconnaissance fixé par la BnF. Pour évaluer les performances de notre approche dans ce contexte, nous avons employé d'une part deux seuils de rejet : le premier seuil est (appelé seuil de rejet effectif « $T\%$ ») fixé généralement par la BnF, le deuxième seuil est (appelé seuil de rejet expérimental « $T'\%$ ») utilisé pour rejeter les pages en se basant sur les estimations de taux de reconnaissance « τ » de notre approche. L'utilisation de ces deux seuils a permis de calibrer le paramètre de rejet de notre approche en fonction de la qualité des transcriptions que nous voulons obtenir. D'autre part, nous avons calculé la précision et le rappel de l'opération de rejet des pages. La précision à $T\%$ est définie par le ratio du nombre de pages correctement rejetées présentant un taux de reconnaissance inférieur à $T'\%$ au nombre total de pages rejetées par notre approche (cf. équation 4.12).

$$Précision_T = \frac{card\left(Pages \quad correctement \quad rejetées\right)_{\tau<(T'\% \quad \& \quad T\%)}}{card\left(Pages \quad total \quad rejetées\right)_{\tau<T'\%}} \quad (4.12)$$

Le rappel à $T\%$ est défini par le rapport du nombre de pages correctement sélectionnées présentant un taux de reconnaissance supérieur à $T'\%$ sur le nombre des pages qui doivent être rejetées dans la base de validation (cf. équation 4.13).

$$Rappel_T = \frac{card\,(Pages \quad correctement \quad rejetées)_{\tau<(T'\%\&T\%)}}{card\,(Pages \quad rejetées \quad réellement)_{\tau<T'\%}} \tag{4.13}$$

4.2 Résultats d'évaluation

Analyse des erreurs d'estimation des taux de reconnaissance

(a) Approche par contrôle de l'isogénie des caractères

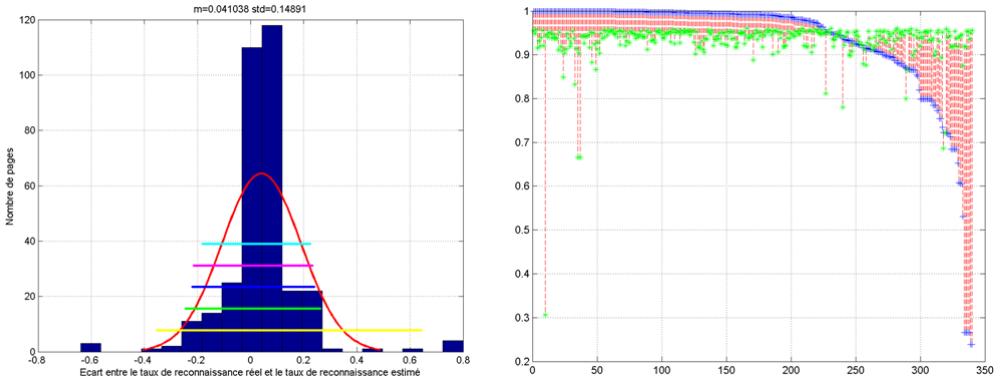
Nous commençons notre évaluation par une analyse des résultats d'estimation de chaque approche. L'histogramme de la figure 4.9a montre que l'estimateur basé sur les données d'isogénie des caractères est biaisé puisque les erreurs d'estimation sont concentrées dans un intervalle de taux de reconnaissance qui s'étend entre $-0,62 \quad (-62\%)$ et $0,7973 \quad (79,73\%)$. De plus, nous remarquons aussi que l'histogramme des erreurs d'estimation est décalé à droite de la valeur nulle. Ce qui signifie que les erreurs d'estimation de cette approche sont généralement sous-estimées. Cependant, ce n'est pas toujours le cas. En effet, selon la figure 4.9b nous remarquons que les faibles taux de reconnaissances sont surestimés. Globalement, les estimations de cette approche ne suivent pas complètement les variations des taux de reconnaissance à estimer.

Ces résultats peuvent être expliqués d'une part par la variabilité des propriétés typographiques des documents d'apprentissage qui rend la distribution des documents d'apprentissage non homogène en fonction des distances intra-classe de caractère. D'autre part, la sensibilité du modèle de régression polynomial aux taux de reconnaissance extrêmes biaise généralement la procédure de formation de la fonction de régression. Ceci engendre par la suite des erreurs considérables d'estimation du taux de reconnaissance des caractères.

Pour pallier ces problèmes, il est préférable d'utiliser cette approche sur des bases de vérification composés de documents caractérisés par des propriétés typographiques homogènes. La figure 4.10 donne la représentation simultanée des taux de reconnaissance réels et estimés obtenus sur une base de pages qui proviennent du même document. D'après cette figure, nous remarquons que les estimations des taux de reconnaissance suivent mieux les variations des taux de reconnaissance réels. De plus, nous constatons aussi la diminution des écarts entre les taux de reconnaissance réels et estimés.

Les segments horizontaux colorés dans l'histogramme de la figure 4.9 représentent les intervalles des erreurs d'estimation qui sont comprises entre le $n^{ième}$ centile et le $(100-n)^{ième}$ centile ($C_{n\%}$ et $C_{100-n\%}$) avec $n \in \{1 \quad 2 \quad 3 \quad 4 \quad 5\}$. D'après cette représentation, nous remarquons que les intervalles des erreurs d'estimation commises par cette approche sont plus ou moins importants sur l'ensemble des populations traitées.

Prenons l'exemple des erreurs d'estimation qui sont comprises entre $C_{1\%}$ et $C_{99\%}$. L'intervalle des erreurs d'estimation correspondant est très étendu puisqu'il est compris entre $-0,35$ et $0,65$. Le plus faible intervalle d'erreur d'estimation réalisé par cette approche sur la population $[C_{5\%} \quad C_{95\%}]$ est compris entre $-0,182$ et $0,226$. Cet intervalle reste toujours important ce qui signifie que les estimations de cette approche sont imprécises. Le tableau 4.1 présente les racines carrées des erreurs quadratiques moyennes et les coefficients de corrélation quadratique obtenus sur les différentes populations de centiles des erreurs de



(a) Histogrammes des erreurs d'estimation des taux de reconnaissance

(b) Distribution des taux de reconnaissance réels (points bleus) et estimés (points verts)

FIGURE 4.9 – Résultats de l'estimateur basé sur les données d'isogénie

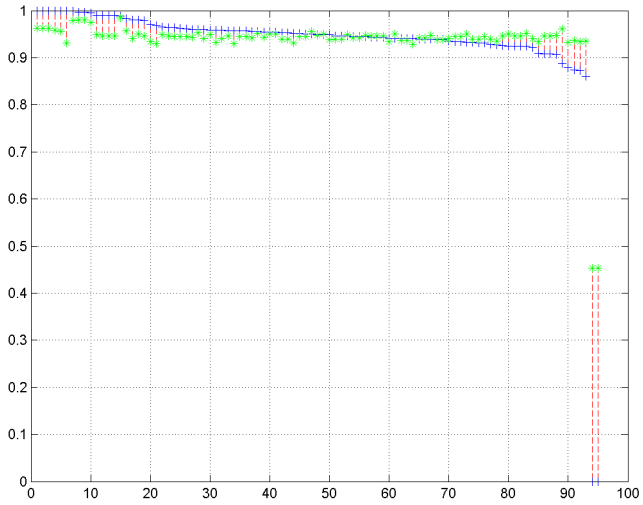


FIGURE 4.10 – Résultats de l'estimateur basé sur les données d'isogénie sur une base d'images homogène

cette approche. D'après ce tableau, nous remarquons que généralement les $RMSE$ sont importantes sur l'ensemble des populations d'erreurs d'estimation. De plus, les CCQ sont différentes de nul, ce qui signifie l'existence de relation de corrélation linaire entre les taux de reconnaissance estimés et les taux réels. Ces relations sont assez faibles puisque les valeurs de ces coefficients sont inférieures à $0,5$. Par conséquent, nous pouvons déduire que l'utilisation d'un descripteur d'isogénie est inadéquate avec une base de documents hétérogène.

(b) Approche par alignement sur un second OCR

L'utilisation directe des images des caractères peut rendre les estimations des taux de reconnaissance sensibles aux variations des polices de caractères au niveau de la page. Pour pallier ce problème, nous pouvons nous baser sur l'expertise des systèmes de reconnaissance automatique de caractères pour décrire les résultats de l'OCR des prestataires.

	$[C_{5\%} \quad C_{95\%}]$	$[C_{4\%} \quad C_{96\%}]$	$[C_{3\%} \quad C_{97\%}]$	$[C_{2\%} \quad C_{98\%}]$	$[C_{1\%} \quad C_{99\%}]$
Limite inférieur	$-0,18$	$-0,21$	$-0,22$	$-0,23$	$-0,35$
Limite supérieur	$0,22$	$0,22$	$0,23$	$0,26$	$0,65$
RMSE	$0,08$	$0,08$	$0,09$	$0,09$	$0,11$
CCQ	$0,185$	$0,183$	$0,176$	$0,17$	$0,152$

TABLE 4.1 – Résultats de l'analyse des centiles des erreurs d'estimation obtenues avec l'approche basée sur l'isogénie des caractères

	$[C_{5\%} \quad C_{95\%}]$	$[C_{4\%} \quad C_{96\%}]$	$[C_{3\%} \quad C_{97\%}]$	$[C_{2\%} \quad C_{98\%}]$	$[C_{1\%} \quad C_{99\%}]$
Limite inférieur	$-0,19$	$-0,2$	$-0,23$	$-0,24$	$-0,26$
Limite supérieur	$0,09$	$0,1$	$0,15$	$0,2$	$0,25$
RMSE	$0,0501$	$0,0525$	$0,0615$	$0,0676$	$0,0769$
CCQ	$0,427$	$0,375$	$0,3573$	$0,3568$	$0,353$

TABLE 4.2 – Résultats de l'analyse des centiles des erreurs d'estimation obtenues avec l'approche basée sur l'alignement des résultats de l'OCR

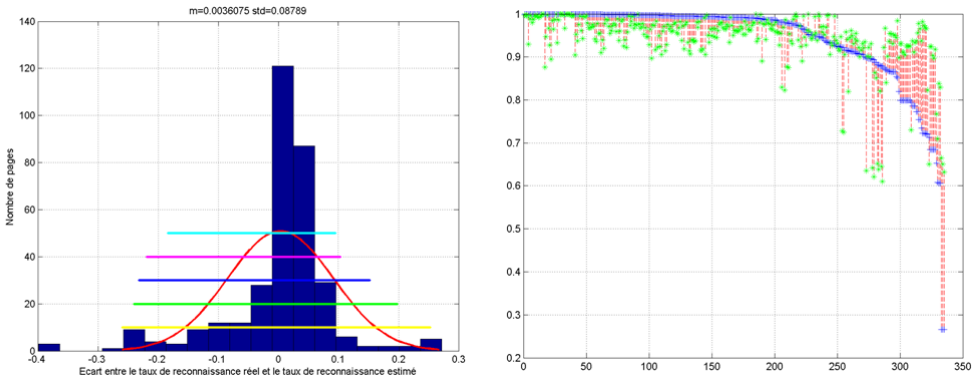
Pour s'assurer de l'intérêt de ce genre de description, nous avons évalué les performances d'estimation du deuxième estimateur. D'après l'histogramme des erreurs d'estimation, nous remarquons que les pics des erreurs de reconnaissance sont compris dans l'intervalle $[0 \quad 10\%]$. De plus, nous remarquons que les erreurs d'estimation sont concentrées autour d'une valeur très proche de zéro $(0,9\%)$ ce qui signifie que les estimations de cet estimateur sont plus précises que l'approche précédente par contre elles restent légèrement biaisées.

La représentation simultanée des taux de reconnaissance réels et estimés de la figure 4.11b confirme cette amélioration par rapport à l'approche précédente. En effet, les taux de reconnaissance estimés suivent mieux les variations des taux de reconnaissance réels bien que les erreurs d'estimation restent proportionnellement importantes sur les estimations des faibles taux de reconnaissance.

L'analyse des centiles des erreurs d'estimation conforte les résultats visuels. En effet, selon le tableau 4.2 nous remarquons que l'étendue de l'intervalle des erreurs d'estimation le plus important de cette approche est plus faible que celle qui a été obtenue avec l'approche précédente. De plus, les $RMSE$ obtenues sur l'ensemble des populations d'erreurs d'estimation sont aussi plus faibles que leurs $RMSE$ respectives obtenues avec l'approche précédente. En effet, la $RMSE$ obtenue sur les erreurs qui sont comprises entre $C_{1\%}$ et $C_{99\%}$ est égale à $0,0769$ alors qu'elle est égale à $0,11$ sur les résultats de l'approche précédente. Nous remarquons aussi que les intensités des relations de corrélation linaire entre les taux estimés et les taux réels sont beaucoup plus fortes que les relations qui existent dans les résultats de l'approche précédente.

Par conséquent, nous pouvons déduire que l'utilisation des résultats d'un OCR tiers pour décrire les résultats de reconnaissance de caractères est plus bénéfique que l'utilisation directe des données d'isogénie des formes des caractères. Cependant, la précision des estimations de ce prédicteur n'est pas encore suffisante pour contrôler correctement la qualité des résultats de reconnaissance des caractères.

(c) Approche par combinaison des caractéristiques



(a) Histogrammes des erreurs d'estimation des taux de reconnaissance

(b) Distribution des taux de reconnaissance réels (points bleus) et estimés (points verts)

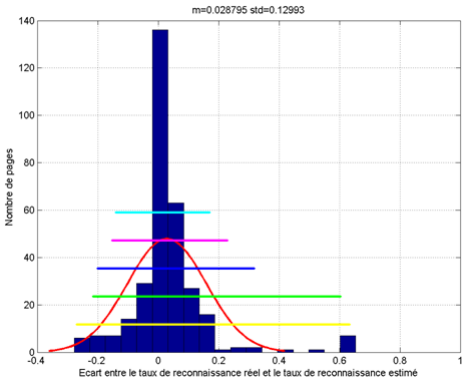
FIGURE 4.11 – Résultats de l'estimateur basé sur les données d'alignement des résultats d'OCR

La combinaison des caractéristiques d'isogénie et d'alignement d'OCR peut résoudre certains problèmes dans la description des résultats de l'OCR liés ou la dysfonctionnement de l'un des deux descripteurs. Pour déterminer l'intérêt de cette démarche sur la qualité des estimations des taux de reconnaissance, nous analysons ici les résultats de la troisième approche. L'histogramme de la figure 4.12a montre que les erreurs d'estimations de cette approche sont concentrées autour de la valeur nulle. Ce qui signifie qu'en moyenne les estimations de cette approche ne sont pas biaisées. De plus, nous remarquons aussi l'augmentation de la proportion des erreurs d'estimation qui sont comprises entre $-0,1$ et $0,1$ à $76,18\%$ contre $66,7\%$ pour l'approche qui utilise les données d'isogénie. Par contre, par rapport aux résultats de la deuxième approche, la proportion des erreurs qui appartiennent à cet intervalle est plus faible. En effet, $79,12\%$ des erreurs d'estimation obtenus avec la deuxième approche sont comprises entre $[-0,1 \quad 0,1]$.

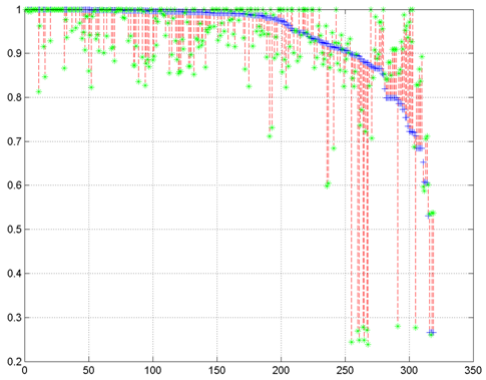
La représentation simultanée des taux de reconnaissance réels et estimés par la troisième approche confirme les résultats des évaluations précédentes. En effet d'après la figure 4.12b, nous remarquons que conformément aux résultats de la deuxième approche, les taux de reconnaissance estimés suivent correctement les variations des taux de reconnaissance réels. Cependant, le biais d'estimation obtenu dans les résultats de cette approche sur les faibles taux de reconnaissance est moins important que le biais d'estimation obtenu avec les approches précédentes sur le même type de page.

Les intervalles des centiles des erreurs d'estimation présentées dans le tableau 4.3 sont un peu différentes de ceux de l'approches précédente. En effet, sur la population des erreurs d'estimation $C_{5\%}$ et $C_{95\%}$ l'étendue de l'intervalle des erreurs d'estimation est presque égale à l'étendue de son intervalle d'erreurs respectif obtenu dans les résultats de l'approche précédente. Par contre, sur la population $C_{1\%}$ et $C_{99\%}$, nous avons un intervalle d'erreurs légèrement plus étendu. Les RMSE des erreurs d'estimation sont légérement supérieurs aux RMSE obtenues avec l'approche précédente. De plus, les CCQ de la troisième approche sont presque égales aux CCQ de l'approche précédente. En effet, pour la population d'erreur $[C_{1\%}$; $C_{99\%}$ le CCQ est égale à $0,352$ dans les résultats de la troisième approche et $0,353$ sur les résultats de la deuxième approche.

En conséquence, la précision de l'estimateur de cette approche reste insuffisante pour



(a) Histogrammes des erreurs d'estimation des taux de reconnaissance



(b) Distribution des taux de reconnaissance réels (points bleus) et estimés (points verts)

FIGURE 4.12 – Résultats de l'analyse des centiles des erreurs d'estimation obtenues avec l'approche basée sur l'utilisation simultanée des données d'isogénie et d'alignement des résultats d'OCR

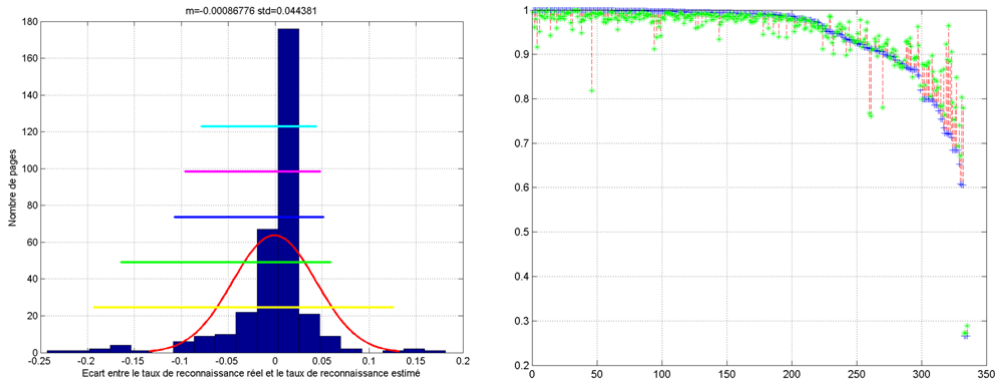
	$[C_{5\%} \quad C_{95\%}]$	$[C_{4\%} \quad C_{96\%}]$	$[C_{3\%} \quad C_{97\%}]$	$[C_{2\%} \quad C_{98\%}]$	$[C_{1\%} \quad C_{99\%}]$
Limite inférieur	$-0,19$	$-0,2$	$-0,21$	$-0,22$	$-0,29$
Limite supérieur	$0,09$	$0,11$	$0,15$	$0,2$	$0,25$
RMSE	$0,0611$	$0,0666$	$0,0702$	$0,0741$	$0,0831$
CCQ	$0,397$	$0,363$	$0,3572$	$0,3569$	$0,352$

TABLE 4.3 – Résultats de l'analyse des centiles des erreurs d'estimation obtenues dans les résultats de l'approche basée sur l'utilisation simultanée des données d'isogénie et d'alignement des résultats d'OCR

classer correctement les documents numériques appartenant à une base hétérogène de documents. En effet, selon la figure 4.12b, nous constatons que certaines pages en qualité *HQ* reçoivent des taux de reconnaissance estimés inférieurs à 95%, ce qui cause leur déqualification si on suit la procédure de contrôle de la BnF. Par conséquent, nous pouvons déduire que l'apprentissage naïf du modèle de régression ne permet pas d'estimer correctement les taux de reconnaissance de caractères. D'où la nécessité d'une procédure d'apprentissage adaptative.

(d) Approche par filtrage des données d'apprentissage

Une amélioration portant sur la troisième approche est proposée dans la quatrième approche. Cette amélioration consiste à filtrer les documents d'apprentissage avant la procédure d'apprentissage des modèles de régression. Pour déterminer l'intérêt de cette technique, nous avons évalué les estimations de quatrième approche. D'après la figure 4.13b, nous constatons que l'estimation des taux de reconnaissance qui sont supérieurs à 90% est généralement fiable puisque les distributions des taux de reconnaissance réels et estimés sont plus ou moins confondues. Nous observons aussi la présence de certains biais d'estimation en traitant les documents ayant des taux de reconnaissance inférieurs à 90%. Cependant, ces erreurs d'estimation sont plus faibles que celles qui ont été obtenues avec les prédicteurs précédents. L'histogramme 4.13a montre que la majorité des erreurs d'estimation sont comprises dans un intervalle entre $-0,05$ (-5%) et $0,05$ (5%). Ceci



(a) Histogrammes des erreurs d'estimation des taux de reconnaissance de caractères

(b) Différences entre les taux de reconnaissance réels et estimés

FIGURE 4.13 – Résultats d'estimation des taux de reconnaissance des caractères obtenus avec l'approche basée sur la sélection des données d'apprentissage

	$[C_{5\%} \quad C_{95\%}]$	$[C_{4\%} \quad C_{96\%}]$	$[C_{3\%} \quad C_{97\%}]$	$[C_{2\%} \quad C_{98\%}]$	$[C_{1\%} \quad C_{99\%}]$
Limite inférieur	$-0,06$	$-0,07$	$-0,07$	$-0,08$	$-0,12$
Limite supérieur	$0,04$	$0,05$	0.06	$0,07$	$0,08$
RMSE	$0,0186$	$0,0213$	$0,0299$	$0,0252$	$0,0292$
CCQ	$0,725$	$0,72$	$0,682$	$0,67$	$0,663$

TABLE 4.4 – Résultats de l'analyse des centiles des erreurs d'estimation obtenues dans les résultats de l'approche basée sur la sélection des données d'apprentissage

représente le plus faible intervalle d'erreurs obtenu dans les résultats des quatre approches. De plus, nous remarquons la présence d'un pic important à l'erreur d'ordre $0,1\%$ ce qui signifie que la majorité des erreurs d'estimation sont regroupées autour de cette valeur.

L'analyse des centiles des erreurs d'estimation montre que les intervalles des erreurs d'estimation sont plus étroits sur l'ensemble des populations d'erreur étudiées dans notre analyse. Par exemple, pour la population des erreurs la plus importante $C_{1\%}$ et $C_{99\%}$, l'intervalle des erreurs d'estimation s'étend entre $-0,12$ et $0,08$. Cet intervalle est le plus étroit pour cette population par rapport aux intervalles d'erreurs obtenues avec les autres approches. De plus, 92% des estimations de la quatrième approche sont réalisées avec une erreur d'estimation inférieure à 5%.

Les RMSE obtenues sur l'ensemble des populations de cette analyse sont inférieures à $0,3$ ce qui montre que les estimations de cette approche sont assez précises. De plus, les relations de corrélation linaires entre les taux estimés et les taux réels sont assez fortes puisque les valeurs de CCQ sont supérieures à 0.5 pour l'ensemble des populations de notre approche et supérieures à $0,7$ pour la population $C_{5\%}$ et $C_{95\%}$. Ceci signifie que les estimations de cette approche suivent correctement les variations des taux de reconnaissance réels.

En conclusion, à partir de l'analyse des performances d'estimation des taux de reconnaissance, nous pouvons déduire que la procédure adaptative d'entrainement d'estimateur est la plus fiable pour développer un prédicteur de taux de reconnaissance de caractère capable de qualifier correctement les résultats des projets de numérisation de masse.

Test de la capacité de rejet automatique des documents

Notre objectif final est de pouvoir contrôler les résultats des OCR prestataires de manière automatique ou semi-automatique. Pour pouvoir évaluer les performances de notre systèmes dans ce contexte, nous avons testé sa capacité de rejet et d'acceptation en se fixant différents niveaux de performance pour l'OCR (allant de 60% à 98%), ces valeurs étant les valeurs extrêmes admissibles à la BnF pour un OCR brut.

Pour qualifier la capacité de rejet des documents, nous avons donc calculé la courbe rappel-précision pour différents taux de reconnaissance souhaités. Les tests sont effectués en utilisant une procédure de validation croisée en 4 plis. Nous répétons donc 4 fois l'expérience en choisissant à chaque itération 75% des documents pour former les estimateurs et 25% des documents pour évaluer les performances en rejet. Les figures 4.14 et 4.16 présentent les différentes courbes rappel/précision moyens obtenues pour différents seuils de performance de reconnaissance souhaité. La figure 4.14 est obtenue avec le quatrième estimateur qui utilise un profil de page composé de trois premières paires de confusion pour filtrer les pages d'apprentissage. Les figures 4.16a et 4.16b présentent respectivement les résultats obtenus en utilisant une seule paire de confusion et cinq paires de confusion. Ces figures montrent l'influence de la variation du nombre des paires de confusion sur les estimations des taux de reconnaissance des caractères.

En pratique, afin d'apporter une aide au service de numérisation de la BnF dans la remontée des alarmes, il est préférable de disposer d'un système de rejet automatique le plus précis possible, de façon à ne pas solliciter l'opérateur pour trier les alarmes. Il semble en effet que par rapport à la situation actuelle où peu de documents sont vérifiés du fait des moyens humains mobilisables pour cette tâche, il faille chercher à remonter aux opérateurs le moins de fausses alarmes possible. En examinant la figure 4.15, on voit que le système offrant le meilleur rappel (80%) pour une précision en rejet de 92% est le système paramétré pour un taux de reconnaissance supérieur à 98% (courbe noire). Dans cette situation, on dispose donc d'un système qui détecte 80% des images à rejeter et qui présente moins de 10% d'erreur dans les alarmes émises.

De la même manière, si on se fixe une précision en rejet de 90% cette fois, on remarque que les trois systèmes paramétrés avec un taux de 96, 97 et 98% de reconnaissance aboutissent à un fonctionnement conduisant à plus de 50% de rappel environ. Avec le système paramétré avec un taux de 98%, on est en mesure de détecter 80% des documents présentant un taux de reconnaissance supérieur à 92% en faisant environ 10% d'erreur dans les alarmes remontées. On peut également remarquer sur la figure 4.14 que le système est capable de détecter 45% des documents présentant un taux d'erreur inférieur à 70% de reconnaissance mots avec une précision de 93%, ce qui peut constituer un système de remontées d'alarmes très fiable pour le service de la BnF.

L'augmentation du nombre des paires de confusion dans la signature des profils des pages améliore les performances de rejet des documents. Les figures 4.16a et 4.16b montrent les courbes rappel/précision de notre système de rejet de document obtenus en utilisant des profils de page composés par une seule paire de désaccord et par cinq paires de désaccord.

Selon la figure 4.16a, nous constatons que les pentes de décroissance des courbes du système qui utilise une seule paire de desaccord sont plus importantes que les pentes de décroissance des courbes de la figure 4.16b. De plus, la précision maximale obtenue dans le système qui utilise une seule paire de désaccord est inférieure à 100% alors que la

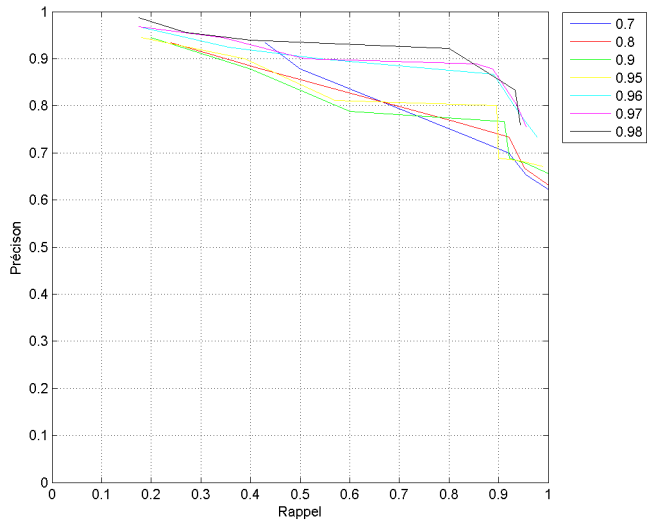


FIGURE 4.14 – Courbes rappel/précision en fonction de la valeur du seuil de rejet des documents pour le prédicteur SVR utilisant le taux de confusion entre caractères et le taux d'isogénie après une étape de sélection des documents utilisant le profil de confusion des caractères les plus fréquents (ici les trois plus fréquents)

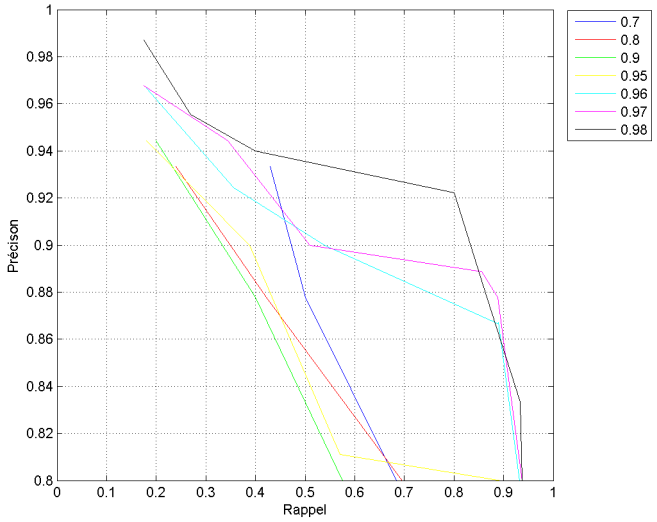
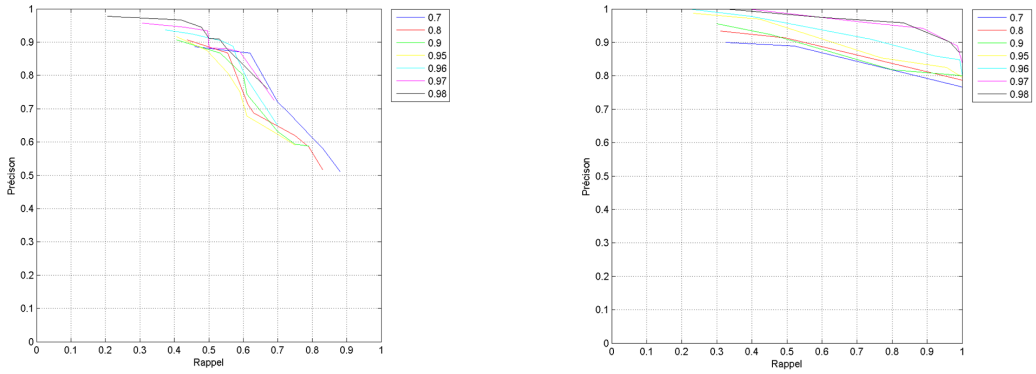


FIGURE 4.15 – Zoom sur les courbes rappel/précision avec mise en évidence des seuils de précision en rejet à 90% et 95%



(a) performances obtenues avec une seule paire de désaccord

(b) performances obtenues avec cinq paires de désaccord

FIGURE 4.16 – Courbes rappel/précision obtenues en variant le nombre des paires de désaccord qui constituent les profils des pages

précision maximale du système qui utilise des profils des pages composés par cinq paires de désaccord atteint 100% lorsqu'il est paramétré avec des seuils de rejet de 96, 97 et 98%.

Les précisions minimales des systèmes qui utilisent un seule paire de désaccord sont supérieures à 50%. Ces précisions sont assez faibles par rapport à celles qui sont obtenues avec les systèmes qui utilisent cinq paires de désaccord. Ceci signifie que plus les profils des pages sont détaillés, moins les erreurs de rejet sont fréquentes dans les résultats de notre approche.

L'utilisation de la procédure de filtrage des documents en utilisant les profils des pages est assez fiable pour réaliser un système de rejet automatique des documents. En effet, on a vu que plus les profils des documents sont riches plus l'opération de rejet est précise. Par contre, le problème principal du système configuré avec un profil de 5 pairs de désaccord est l'exigence de sa procédure de sélection des documents d'apprentissage (au moyenne 7 pages d'apprentissage par page traitée). Ceci réduit énormément la portée des traitements de notre approche sur des exemples non représentés dans la base d'apprentissage en adoptant la formulation des profils de 5 pairs de désaccord. En fait, dans l'expérimentation réalisé dans cette partie, 20% des pages d'évaluation sont traitées par le système configuré avec un profil de désaccord composé de 5 pairs de désaccord. Alors qu'avec le système qui emploi un profil de désaccord composé de 3 pairs de désaccord, la totalité des pages de validation sont traitées par notre approche ce qui correspond à une couverture maximale de notre base d'évaluation.

En se basant sur ces résultats, nous pouvons déduire que l'augmentation du nombre des pairs de désaccord dans la composition des profils de désaccord améliore considérablement la précision des estimations de notre approche. Par contre, elle limite la portée de notre approche sur la base d'évaluation.

5 Conclusion

Nous avons proposé dans cette section différentes approches visant à estimer automatiquement le taux de reconnaissance des mots dans les résulats d'OCR. L'objectif de cette

étude est de développer une procédure de contrôle qualité des transcriptions automatique qui puisse être intégrée dans la chaîne de contrôle de la BnF. Cette méthode estime les performances des résultats de reconnaissance des mots et permet de rejeter automatiquement les documents qui ont un taux de reconnaissance inférieur à un seuil de qualité défini par l'utilisateur.

L'approche proposée a été conçue pour que le système fonctionne sur une grande variété de documents. Le système a été testé sur une base de documents très hétérogène pour se confronter aux problèmes réels des projets de numérisation de masse. L'application de notre système de contrôle ne nécessite aucune connaissance experte pour être configurée (pas de connaissance linguistique ni de connaissance typographique). En conséquence, l'approche proposée est générique et peut être adaptée sur différents corpus afin d'offrir des résultats optimaux. Ceci est tout à fait en accord avec la démarche mise en œuvre par la BnF auprès de ses prestataires qui exige maintenant que lui soit fournie une petite proportion de documents transcrits sans erreur (vérité terrain) pour chaque corpus numérisé. Ainsi il sera possible de paramétrer de manière optimale le système que nous proposons en utilisant ces documents.

L'évaluation de notre approche a été réalisée sur une base d'images de documents réels composée de 345 pages qui englobe à la fois des documents difficiles (transcrits avec un faible taux de reconnaissance) et des documents faciles (transcrits avec un bon taux de reconnaissance). Les résultats de cette évaluation ont montré que le taux de confusion entre l'OCR analysé et un OCR auxiliaire utilisé comme référence, est un indicateur pertinent pour sélectionner des documents similaires en termes de difficulté vis à vis de l'OCR. La construction d'un estimateur robuste du taux de reconnaissance est alors possible en utilisant une technique de régression à base de vecteurs supports et en employant des descripteurs génriques telques l'isigénie des formes des caractères et le désaccord entre deux résultats d'OCR.

Nous avons donné une évaluation de cette approche à la fois pour qualifier sa capacité à estimer le taux de reconnaissance mots d'un document inconnu mais également en nous plaçant dans les conditions qui seront les conditions réelles d'utilisation au sein du service de numérisation de la BnF, chargé d'analyser la qualité des résultats d'OCR fournis par les prestataires. Nous avons montré dans cette dernière utilisation que le meilleur système proposé est capable de détecter 80% des documents présentant un taux de reconnaissance mots inférieur à 98% avec une précision de 92%. On peut également détecter automatiquement 45% des documents présentant un taux de reconnaissance inférieur à 70% avec une précision supérieure à 92%.

Une telle approche n'a jamais été proposée dans la littérature à notre connaissance, car la communauté des chercheurs du domaine n'a jamais étudié une telle question. On a vu néanmoins apparaître récemment des recherches sur la qualité des images de documents et plus particulièrement dans la perspective de mieux contrôler la prise de vue. C'est l'une des questions abordées dans le cadre du projet ANR DigiDoc qui s'intéresse aux futures générations de « scanners intelligents ». C'est également le cas de travaux qui s'intéressent à la prise de vue à l'aide de smart phones en situation de mobilité où il est important de détecter très tôt dans la chaîne de traitement si la qualité de l'image du document est suffisante pour les traitements qui prendront souvent place de manière déportée et différée.

Conclusion générale

L'expansion des besoins de numérisation des documents avec la mise en œuvre de plateformes numériques de consultation en ligne a généralisé l'utilisation des systèmes de reconnaissance de caractères (ou « OCR »). Les technologies actuelles des systèmes d'OCR industriels présentent généralement des performances très satisfaisantes sur la plupart des documents récents. Ces systèmes sont cependant mis en difficulté par les documents anciens et patrimoniaux, caractérisés par des dégradations de leur support physique ainsi que par la nature même de leurs polices de caractères, typographies et langues.

A cause de ces difficultés, des incohérences dans les résultats de segmentation et de reconnaissance surviennent dans les documents numériques. Or certains systèmes de consultation et d'édition se basent sur le contenu numérique des documents pour remplir leurs fonctions. La présence d'erreurs dans les résultats de l'OCR peut donc biaiser le fonctionnement de ces systèmes. C'est la raison pour laquelle les acteurs de la numérisation mettent en œuvre des procédures de vérification des résultats d'OCR dans leurs chaînes de numérisation. Mais à notre connaissance, le problème du contrôle des résultats d'OCR dans le cadre des projets de numérisation de masse n'est pas ou peu abordé dans la littérature. Les solutions adoptées par les prestataires de numérisation ainsi que par les utilisateurs des contenus numériques sont basées sur des contrôles visuels menés par des opérateurs humains (parfois épaulés par des outils linguistiques), cela sur des échantillons restreints de pages.

Nous avons donc proposé deux approches de vérification des résultats de la transcription automatique des caractères permettant de vérifier la qualité des documents numériques. La première approche traite le sujet de la vérification du problème d'omission de mots dans les résultats de l'OCR. La deuxième approche essaye de vérifier la qualité de la transcription des mots en estimant les taux de reconnaissance des caractères. Notre contribution principale a consisté dans le développement d'algorithmes efficaces destinés au traitement d'une collection documentaire très variable et qui ne nécessitent pas une vérité terrain.

Pour garantir l'aspect adaptatif de nos approches, nous avons utilisé les caractéristiques locales des documents pour réaliser les traitements. L'approche de détection des mots omis se base sur les caractéristiques des éléments textuels et graphiques détectés sur une page pour rechercher des éléments similaires dans les zones d'arrière-plan. Pour décrire les éléments de la page, nous avons utilisé des caractéristiques de texture. Ce choix a été dicté par la généricité de cette famille de caractéristiques. Les textures de l'image peuvent être décrites par leurs directions, leurs régularités et leurs fréquences de transition entre les pixels foncés et les pixels clairs.

Pour décrire les textures de l'image dans notre approche, nous avons assigné à chaque

pixel de l'image un vecteur de caractéristiques composé de 12 caractéristiques. Ensuite, nous avons employé quatre classifieurs SVM pour classer les pixels en quatre classes de pixels (classe des pixels textuels, classe des pixels d'illustration, classe des pixels d'espace entre mots et classe des pixels d'arrière-plan). Enfin, afin d'être conforme avec la procédure d'évaluation de la BnF, nous avons appliqué une analyse en composantes connexes qui nous permet de transformer les résultats de nos classifieurs du niveau pixel au niveau mot.

Trois procédures d'évaluation ont été employées pour valider cette approche. La première a été réalisée qualitativement sur un échantillon de 165 images de pages sélectionnées aléatoirement parmi 50 documents. La deuxième approche d'évaluation concerne une campagne d'évaluation réalisée dans le service de numérisation de la BnF sur deux bases de documents différentes. La première base est composée par des documents du *journal officiel* (1270 pages) et la deuxième base est composée par de 160 pages de *presse*. Les résultats d'évaluation de notre approche ont montré que les performances de notre détecteur sont assez bonnes, puisque $84,15\%$ des éléments omis dans notre base d'évaluation sont détectés avec une précision égale à $94,73\%$.

Pour contrôler la qualité des documents numériques, l'approche de vérification des résultats de reconnaissance se base sur une procédure d'estimation des taux de reconnaissance. Les systèmes d'OCR se basent généralement sur des taux de confiance exprimés au niveau du caractère ou du mot pour évaluer leurs résultats de reconnaissance. Ces taux sont calculés en mettant en correspondance les formes des caractères de l'image avec le modèle des formes de caractères de l'OCR, et le texte reconnu avec le modèle de langue de l'OCR. Cependant, ces méthodes de vérification ne sont pas conformes avec le contexte de la numérisation de masse. D'une part, la masse importante des documents rend la vérification de la totalité des résultats de l'OCR issus des projets de numérisation de masse impossible. D'autre part les estimations de qualité de reconnaissance restent fortement liées aux caractéristiques typographiques et linguistiques des documents, ce qui ne convient pas au contexte des projets de numérisation de masse, qui sont amenés à traiter une grande diversité de documents.

Pour surmonter cette difficulté, nous avons commencé par poser les deux hypothèses suivantes :

1. Dans une page isogène, les formes des caractères appartenant aux mêmes classes sont peu variables.
2. L'accord entre deux systèmes d'OCR se réalise majoritairement sur des caractères bien reconnus.

En se basant sur ces hypothèses, nous avons étudié quatre approches originales d'estimation du taux de reconnaissance, qui ne dépendent ni des caractéristiques typographiques ni des spécificités linguistiques des documents. La première approche utilise la distance moyenne intra-classe de caractères pour mesurer l'isogénie des caractères de la page ; la deuxième approche emploie le taux de désaccord moyen entre deux OCR pour caractériser les résultats de reconnaissance ; la troisième approche emploie conjointement le taux moyen de désaccord et la distance moyenne intra-classe de caractères pour décrire les résultats de l'OCR ; la quatrième approche utilise une procédure de filtrage des documents d'apprentissage pour former les modèles de régression des taux de reconnaissance.

Pour évaluer ces approches nous avons employé une base d'image composée de 350 images sélectionnées aléatoirement parmi les documents numériques de la BnF. L'éva-

luation de ces approches a montré l'intérêt de l'utilisation conjointe des caractéristiques d'isogénie des caractères, de l'alignement des résultats d'OCR et de l'application de procédure de filtrage des documents d'apprentissage. En effet, la quatrième approche affiche les meilleures performances pour les résultats d'estimation du taux reconnaissance (Erreur Quadratique Moyenne = $4,83\%$ $(0,04829)$) et pour les résultats de rejet des documents (Précision= $92,3\%$ et Rappel = $73,4\%$).

Dans les perspectives de ce travail, nous devrons tester l'algorithme de contrôle de résultats de reconnaissance de caractères dans le contexte de production de masse afin de tester ses performances à grande échelle. De plus, les différents algorithmes de vérifications conçus dans ce travail ont été développés sous forme de prototypes de test, ce qui a permis d'évaluer leurs performances et leur potentielle utilisation par le service de numérisation de la BnF. Toutefois, des travaux complémentaires d'ingénierie logicielle sont nécessaires pour permettre leur implémentation la plus efficace possible afin de répondre aux exigences d'utilisation au sein du service de numérisation de la BnF.

Annexe A

Analyse de contribution de chaque caractéristique dans chaque classe

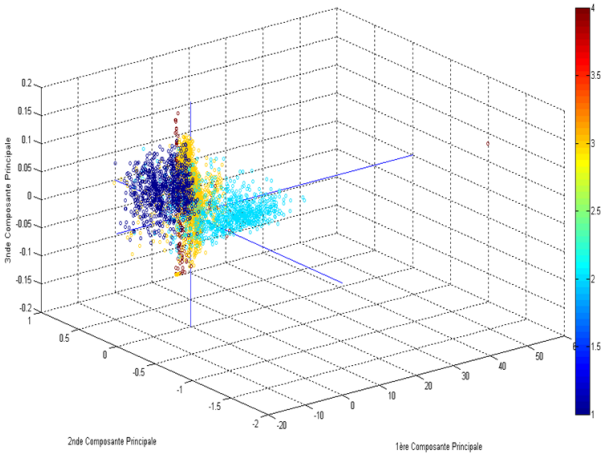
Dans le chapitre 3, nous avons exposé les descripteurs ainsi que la méthode de détection des mots omis dans les résultats d'OCR. L'évaluation de cette méthode a montré que nous pouvons atteindre des bonnes performances malgré la variabilité des caractéristiques physiques et typographiques des pages d'évaluation. Toutefois, la caractérisation des textures à différentes échelles et avec différents types de descripteurs que nous avons choisie est fondée sur des intuitions et il peut exister des redondances dans la description des textures. D'autre part, la complémentarité entre les caractéristiques de texture choisies est une propriété très importante pour garantir une description la plus complète possible des pages analysées.

C'est la raison pour laquelle, nous analysons ici la contribution des différentes caractéristiques que nous avons choisies afin de vérifier leur intérêt réel dans la méthode que nous avons développée. Pour cela, nous présentons une analyse en composantes principales sur les réponses de nos descripteurs obtenus sur deux images de document. La première image (cf. figure A.1a) illustre une page composée par une illustration au trait et des textes en colonne. La deuxième image (cf. figure A.2a) illustre une page qui n'englobe que du texte.

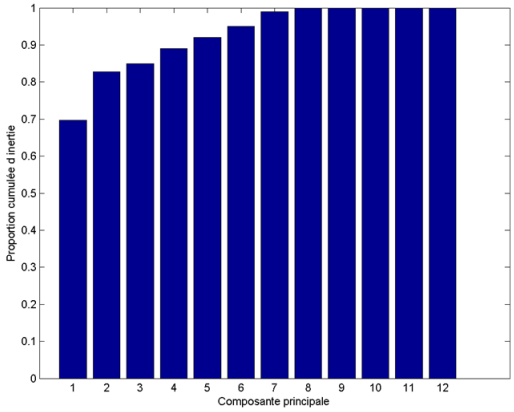
La carte des composantes principales des individus (cf. figure A.1b) présente la distribution des pixels de l'image selon les trois premiers axes principaux. Les couleurs des points ont une signification dans notre analyse. Elles correspondent aux classes des pixels des images de la figure A.1a). Les points bleus référencent les pixels textuels de l'image, les points cyans représentent les pixels des régions d'illustration, les points jaunes représentent les espaces entre les mots alors que les pixels d'arrière-plan sont représentés par des points rouges.

D'après la carte des composantes principales des individus, nous constatons que les projections des pixels textuels et graphiques sur le premier axe principal sont séparables. En effet, selon cette figure, les projections des pixels textuels sont situées dans le côté des coordonnées négatives du premier axe principal. Par contre, les projections des pixels d'illustration sont situées dans le côté des coordonnées positive de cet axe. Nous remarquons également que les représentations des pixels des régions inter-mot (points jaunes) selon le premier axe principal sont confondues avec les représentations des pixels textuels. Par conséquent, l'utilisation du premier axe principal ne permet pas de séparer ces deux classes de pixels. De plus, les projections des pixels d'arrière-plan (points rouges) selon le

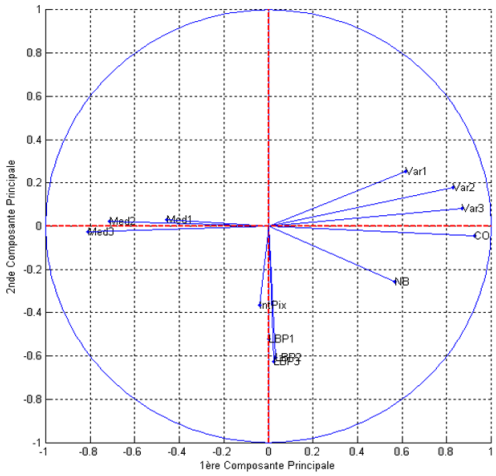
(a) Image origine



(b) Nuages des pixels



(c) Inerties des axes factoriels



(d) Cercle de corrélation entre le premier et le deuxième axe factoriel

FIGURE A.1 – Résultats de l'analyse en composantes principales sur le deuxième exemple de la figure 3.2b

premier axe sont aussi confondues sur les projections des pixels textuels et d'espace entre mots.

L'utilisation du deuxième axe principal résout une partie de ce problème. En effet, selon cet axe d'une part nous remarquons que les projections des pixels textuels ont majoritairement des coordonnées positives ; par contre les projections des pixels d'arrière-plan et d'inter-mots sont situées dans la côté négatif des coordonnées de cet axe. Par contre, même en utilisant les deux premiers axes principaux pour représenter les pixels d'arrière-plan et d'espace inter-mots, les représentations de ces deux classes de pixels restent toujours inséparables.

L'utilisation du troisième axe principal n'apporte pas une grande amélioration dans la représentation des pixels des classes de la page. Par conséquent, nous pouvons déduire que nous avons besoin d'un espace de caractéristiques supérieur à trois dimensions pour représenter les quatre classes de pixels de l'image.

Pour déterminer le nombre minimal d'axes principaux capables de représenter correctement les données de notre analyse, nous avons représenté dans la figure A.1c) l'histogramme cumulé de la proportion d'inertie de chaque composante principale par rapport à l'inertie totale du nuage des données. D'après cette figure, le premier axe principal détient 69,73% de l'inertie totale du nuage des données. Cette proportion d'inertie cumulée augmente progressivement jusqu'à une valeur proche de 100% au huitième axe principal. Par conséquent, nous pouvons déduire qu'en utilisant un espace de caractéristiques composé de huit dimensions, nous pouvons obtenir une représentation fidèle des données de test.

De plus, afin d'affiner l'interprétation des résultats de l'ACP, nous avons employé la représentation du cercle de corrélation (cf. figure A.1d) qui nous a permis d'étudier d'une part les liens qui existent entre les variables originales et les axes principaux, et d'autre part les liens qui existent entre les caractéristiques de notre approche. Dans le cercle de corrélation de la figure A.1d, nous représentons les descripteurs de texture de notre approche en utilisant les notations suivantes :

- CD : pour désigner la caractéristique qui vérifie l'orientation principale de texture à différentes échelles,
- Var_i : pour désigner les caractéristiques de variance d'orientation de texture,
- Med_i : pour désigner les caractéristiques de l'intensité médiane des orientations de texture,
- LBP_i : pour désigner les caractéristiques de descripteur LBP,
- NB : pour désigner les fréquences de passage des pixels clairs aux pixels foncés,
- $intensité \quad des \quad pixels$: pour désigner l'intensité moyenne des pixels.

Les caractéristiques d'orientation et de régularité de texture sont mesurées en utilisant trois fenêtres glissantes de taille différentes. L'indice $i \in [1 \quad 3]$ est employé pour désigner les trois fenêtres glissantes de chaque descripteur. Par conséquent, dans le cercle de corrélation, nous obtenons 12 étiquettes qui correspondent aux 12 caractéristiques de notre approche.

Le cercle de corrélation présenté dans la figure A.1d montre les coordonnées des caractéristiques de texture selon les deux premiers axes factoriels. Ces représentations montrent d'une part que les relations qui existent entre les caractéristiques de texture des images de document, et d'autre part la contribution de chaque propriété dans la formation des axes principaux.

D'après la figure A.1d, nous remarquons que la contribution des descripteurs qui ap-

partiennent à la famille d'orientations des textures est déterminante dans la formation du premier axe principal. En effet, selon cette représentation, nous remarquons que les représentations des caractéristiques de consensus d'orientations principales, de l'intensité médiane des orientations et des variations des orientations de texture sont plus au moins proches de cercle de corrélation. De plus, les angles réalisés entre les représentations de ces caractéristiques et le premier axe principal sont petits ce qui prouve le rôle important de ces caractéristiques dans la formation de cet axe.

Par contre l'intensité de la contribution de chaque caractéristique est dépendante de sa coordonnée sur le premier axe principal. Par exemple, la caractéristique de consensus de l'orientation principale de texture contribue plus que les autres caractéristiques dans la formation du premier axe. De plus, la contribution des autres descripteurs de texture (médianes et variances) dans la construction du premier axe principal est variable selon la taille de fenêtre glissante utilisée pour caractériser les textures de la page. En effet, d'après la figure A.1d, les contributions les plus importantes sont réalisées par les plus grandes tailles de fenêtre glissante. Cette contribution diminue en diminuant la taille de fenêtre glissante. Ceci prouve l'intérêt de la procédure de traitement multi-échelle que nous adoptons.

De plus, nous remarquons que la caractéristique de consensus des orientations principales de texture est corrélée positivement avec les caractéristiques des variances des orientations de texture. Par contre, elles sont décorrélées avec les descripteurs de l'intensité moyenne des orientations de texture. Ceci est conforme avec l'analyse des descripteurs que nous avons effectuée dans la section 3 de ce chapitre. En effet, les descripteurs de consensus des orientations principales de texture ainsi que celui de la variance des orientations produisent des réponses maximales sur les régions textuelles de la page. Au contraire, les réponses de descripteur de l'intensité médiane des orientations de texture sont maximales sur les régions d'illustration. Ceci explique bien la distribution opposée des représentations de ces caractéristiques selon le premier axe principal. Par conséquent, nous pouvons déduire que le premier axe principal est formé pour séparer les classes des pixels textuels et graphiques.

Les caractéristiques appartenant à la famille des descripteurs de régularité des formes de texture contribuent de manière importante à la formation du deuxième axe principal. En effet, d'après la figure A.1d les représentations de ces caractéristiques sont presque confondues avec le deuxième axe principal. De plus, les intensités de la contribution des descripteurs LBP dans la formation du deuxième axe principal sont importantes. En effet, les projections de ces caractéristique sur le deuxième axe principal donnent des cordonnées supérieures à 0.6. D'autre part, comme pour le premier axe principal, la contribution des descripteurs de LBP est dépendante de la taille du modèle LBP que nous utilisé. En effet, l'utilisation de la configuration maximale de LBP ($R = 20$ et $P = 64$) donne la valeur de contribution la plus importante. Ceci prouve aussi l'intérêt de l'utilisation de l'analyse multi-échelle que nous adopterons pour calculer les réponses de ce descripteur.

L'intensité moyenne des pixels joue aussi un rôle déterminant dans la formation de cet axe. En effet, d'une part la représentation de ce descripteur est presque confondue sur le deuxième axe principal. D'autre part, nous remarquons que cette caractéristique est corrélée positivement avec les caractéristiques de LBP. Cette représentation est logique. En effet, en retournant à l'analyse de la carte des individus, nous trouvons que le deuxième

axe principal assure une séparation entre les pixels des régions d'arrière-plan et les pixels des régions textuels. Par conséquent, nous pouvons déduire que la liaison entre ces deux caractéristiques provient du fait que les descripteurs LBP donnent des réponses maximales sur les régions de fond dans lesquelles les intensités moyennes des pixels sont minimales.

En conclusion, à partir de l'histogramme cumulé des inerties des axes principal et de l'analyse des cercles de corrélation, on peut déduire l'existence de redondances dans les signatures utilisées pour décrire les textures de l'image. En effet, l'analyse en composante principale des caractéristiques de texture montre que nous pouvons réduire la dimension des vecteurs de caractéristiques des pixels en passant de 12 caractéristiques à 8 toute en conservant 100% de l'inertie totale des données d'analyse. Par conséquent, nous pouvons réduire le nombre de caractéristiques dans la signature des textures de l'image pour décrire les classes des pixels de cette image.

Pour s'assurer de la reproductibilité de cette conclusion sur des images de natures différentes, nous avons étudié les réponses de nos descripteurs sur l'image de la figure A.2a. La page illustrée par cette figure contient des éléments textuels, des espaces entre mots et des régions d'arrière-plan. La classe des pixels d'illustration est absente dans cette figure.

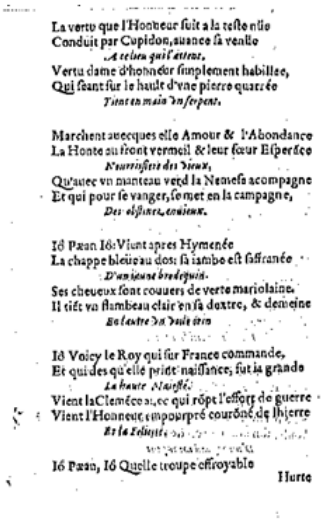
L'application de l'analyse en composantes principales sur les descriptions de texture de cette image montre que l'inertie portée par les 2 premiers axes principaux est proche de 100%. Par conséquent, il est possible de réduire l'espace des caractéristiques des pixels de l'image de 12 à 2 dimensions tout en conservant $95,99\%$ de l'inertie total du nuage des points.

La carte des individus montre la distribution des pixels de l'image selon les deux premiers axes principaux. Selon la figure A.2b, nous remarquons que les classes des pixels textuels, des pixels d'espace inter-mot et des pixels d'arrière-plan sont généralement séparables selon le premier axe principal. De plus, l'utilisation du deuxième axe principal permet de séparer les pixels d'espaces inter-mots des pixels textuels.

D'autre part, d'après le cercle de corrélation présenté dans la figure A.2d. nous remarquons que l'ensemble des caractéristiques d'orientation de texture de l'image sont corrélées négativement selon le premier axe principal. De plus, d'après la carte des individus, les pixels textuels ont généralement des cordonnées négatives selon cet axe. Par conséquent, nous pouvons déduire que l'utilisation de ces caractéristiques permet de décrire les éléments textuels de la page.

Nous remarquons aussi que les caractéristiques de LBP et d'intensité moyenne des pixels sont décorrélées avec les caractéristiques d'orientation des textures. Ces représentations sont logiques puisque les descripteurs LBP ont des réponses maximales sur les régions d'arrière-plan. De plus, l'intensité moyenne de ces pixels sont généralement proches de l'intensité nulle ce qui rend ces deux familles de caractéristiques corrélées positivement.

Les faibles tailles des fenêtres glissantes que nous avons utilisées pour calculer les réponses des descripteurs d'orientations de texture et de LBP, donnent des indications sur les espaces qui se trouvent entre les mots. Ceci apparaît clairement sur le cercle de corrélation A.2d. En effet, d'après cette figure, nous remarquons que les caractéristiques de médiane des intensités d'orientations, de variance des orientations et de LBP obtenues avec les faibles tailles de fenêtre glissante sont corrélées négativement selon le deuxième axe principal. Or d'après la carte des individus, les projections des pixels des espaces


La vertu que l'Honneur fuit a la teſte nüe
Conduit par Cupidon, auance ſa veuë
A telle qui l'attend.
Vertu dame d'honneur ſimplement habillee,
Qui ſeant ſur le haut d'vne pierre quarrée
Tient en main vn ſerpent.

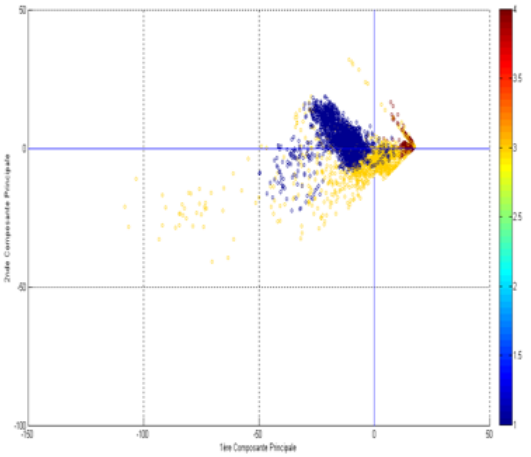
Marchent auecques elle Amour & l'Abondance
La Honte au front vermeil & leur ſœur Eſperance
Nourriſſent des Dieux,
Qu'auec vn manteau verd la Nemeſis acompagne
Et qui pour ſe vanger, ſe met en la campagne,
Des eſpines, couleurs.

Iõ Paan Iõ: Vient apres Hymenée
La chappe bleüe au dos: ſa iambe eſt ſafranée
D'vn ieune brodequin.
Ses cheueux ſont couuers de verte mariolaine.
Il tiẽt vn flambeau clair en ſa dextre, & demeine
En l'autre vn doux tein

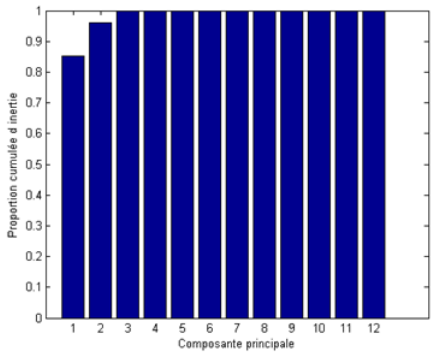
Iõ Voicy le Roy qui ſur France commande,
Et qui des qu'elle prit naiſſance, fut la grande
La haute Maieſté.
Vient la Clemẽce auec qui rõpt l'effort de guerre
Vient l'Honneur empourpré couronné de lhierre
Et la Felicité

Iõ Paan, Iõ Quelle troupe effroyable
Hurte

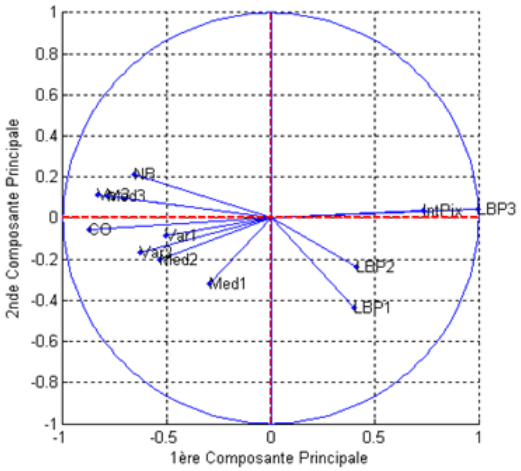
(a) Image origine



(b) Nuages des pixels



(c) Inerties des axes factoriels



(d) Cercle de corrélation entre la première et la deuxième axes factoriels

FIGURE A.2 – Résultats de l'analyse en composant principal sur un exemple de page qui contient que du texte

inter-mots sont situées dans le coté négatif de cet axe.

En conclusion, nous constatons d'après les analyses de ces deux exemples de page l'existence de descriptions redondantes dans les signatures des textures proposées. Par contre, il difficile de déterminer lesquelles de ces caractéristiques sont toujours redondantes. En effet, selon les caractéristiques physiques et typographiques de l'image, le comportement de nos descripteurs change. Par conséquent, il faudrait construire plusieurs combinaisons de descripteur pour chaque type d'image pour décrire les textures de la page et éliminer les redondances dans les signature des textures de l'image. Ceci n'est pas facilement réalisable dans le contexte de notre travail puisque la collection documentaire de la BnF est très variable. D'où la nécessité de l'utilisation de l'ensemble des caractéristiques que nous employons dans notre approche.

Bibliographie

[A.94] MAHMOUD S. A. Arabic character recognition using fourier descriptors and character contour encoding. *Pattern recognition*, 27(6) :815–824, 1994.

[AAP09] D. Bridson A. Antonacopoulos, S. Pletschacher and C. Papadopoulos. Icdar 2003 robust reading competitions : entries, results, and future directions. *International Conference on Document Analysis and Recognition (ICDAR)*, 10(2-3) :1370–1374, 2009.

[AGP10] Marios Anthimopoulos, Basilis Gatos, and Ioannis Pratikakis. A two-stage scheme for text detection in video images. *Image and Vision Computing*, 28(9) :1413 – 1426, 2010.

[All03] B. Allier. *Contribution à la numérisation des collections : apports des contours actifs.* 2003.

[AM11] Kamel Ait-Mohand. *Techniques d'adaptation de modèles markoviens. Application à la reconnaissance de documents anciens.* PhD thesis, 2011. Thè se de doctorat dirigé e par Paquet, Thierry Informatique Rouen 2011.

[Ant11] Apostolos Antonacopoulos. The effect of scanning parameters on ocr quality. In IMPACT, editor, *IMPACT FINAL CONFERENCE*, Londre, 2011.

[BK83] Satyabrata Banerji and Shovonlal Kundu. A linear mapping technique for optimizing binary templates in noise-free pattern matching. *Pattern Recognition*, 16(5) :481 – 487, 1983.

[Blu67] Harry Blum. A Transformation for Extracting New Descriptors of Shape. In Weiant Wathen-Dunn, editor, *Models for the Perception of Speech and Visual Form*, pages 362–380. MIT Press, Cambridge, 1967.

[BN00] Dennis Bahler and Laura Navarro. Methods for combining heterogeneous sets of classifiers. In *17th Natl. Conf. on Artificial Intelligence (AAAI), Workshop on New Research Problems for Machine Learning*, 2000.

[CBd05] Hubert Cecotti and Abdel Belaï d. Hybrid ocr combination for ancient documents. In Sameer Singh, Maneesha Singh, Chid Apte, and Petra Perner, editors, *Pattern Recognition and Data Mining*, volume 3686 of *Lecture Notes in Computer Science*, pages 646–653. Springer Berlin Heidelberg, 2005.

[CCMV03] Yves Caron, Harold Charpentier, Pascal Makris, and Nicole Vincent. Power law dependencies to detect regions of interest. In Ingela Nystram, Gabriella Sanniti di Baja, and Stina Svensson, editors, *DGCI*, volume 2886 of *Lecture Notes in Computer Science*, pages 495–503. Springer, 2003.

[CHC89] F.-H. Cheng, Wen-Hsing Hsu, and M.-Y. Chen. Recognition of handwritten chinese characters by modified hough transform techniques. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 11(4) :429–439, April 1989.

[Che96] Jisheng ; Komuves J. ; Haralick R.M. Chetverikov, D. ; Liang. Zone classification using texture features. In IAPR, editor, *Proceedings of the 13th International Conference on Pattern Recognition*, Vienna, 1996.

[Chi92] B. Chigier. Rejection and keyword spotting algorithms for a directory assistance city name recognition application. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 2, pages 93–96 vol.2, Mar 1992.

[Cho70] C Chow. On optimum recognition error and reject tradeoff. *Information Theory, IEEE Transactions on*, 16(1) :41–46, 1970.

[COU98] P COURMONTAGNE. Transformée de radon et filtrage : Application à la détection de sillages de mobiles marins. *TS. Traitement du signal*, 15(4) :297–307, 1998.

[CP98] B. B. Chaudhuri and U. Pal. A complete printed bangla ocr system. *Pattern Recognition*, 31(5) :531–549, 1998.

[CWM11] Thierry Claerr, Isabelle Westeel, and Michel Melot, editors. *Manuel de la numérisation*. Collection Bibliothéques. Ãd. du Cercle de la librairie, Paris, 2011.

[CWS03] Zheru Chi, Qing Wang, and Wan-Chi Siu. Hierarchical content classification and script determination for automatic document image processing. *Pattern Recognition*, 36(11) :2483 – 2500, 2003.

[DR02] D. Dhanya and A.G. Ramakrishnan. Script identification in printed bilingual documents. In Daniel Lopresti, Jianying Hu, and Ramanujan Kashi, editors, *Document Analysis Systems V*, volume 2423 of *Lecture Notes in Computer Science*, pages 13–24. Springer Berlin Heidelberg, 2002.

[EBE98] Véronique Eglin, Stéphane Bres, and Hubert Emptoz. Characterization and classification of printed text in a multiscale context. In *SSPR/SPR*, pages 960–967, 1998.

[EGJM95] Ellen Eide, Herbert Gish, Philippe Jeanrenaud, and Angela Mielke. Understanding and improving speech recognition performance through the use of diagnostic tools. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 221–224. IEEE, 1995.

[Egl08] Nicolas Journet ; Ramel Jean Yves ;Remy Mullot ; Véronique Eglin. Document image characterization using a multiresolution analysis of the texture : application to old documents. *IJDAR*, 2008.

[Far01] Hany Farid. Blind inverse gamma correction. *IEEE Transactions on Image Processing*, 10(10) :1428–1433, 2001.

[Fré10] Marie-Elise Fréon. Chaîne de numérisation document préparatoire. `http://www.bnf.fr/documents/ref_num_ocr.pdf`, 2010.

[GIY97] L. Gillick, Y. Ito, and J. Young. A probabilistic approach to confidence estimation and evaluation. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 2, pages 879–882 vol.2, Apr 1997.

[GK96] Venu Govindaraju and Ram K. Krishnamurthy. Holistic handwritten word recognition using temporal features derived from off-line images. *Pattern Recognition Letters*, 17(5) :537 – 540, 1996.

[Gun04] System of classifiers : state of the art and trends. *International journal of Pattern Recognition and Artificiel Inteligence*, 31(8) :983 – 1001, 2004.

[HGF01] Qiang Huo, Yong Ge, and Zhi-Dan Feng. High performance chinese ocr based on gabor features, discriminative feature extraction and model training. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, volume 3, pages 1517–1520 vol.3, 2001.

[HHP95] Jaekyu Ha, Robert M. Haralick, and Ihsin T. Phillips. Recursive x-y cut using bounding boxes of connected components. In *ICDAR*, pages 952–955. IEEE Computer Society, 1995.

[Hol09] Rose Holly. How good can it get ? analysing and improving ocr accuracy in large scale historic newspaper digitization programs. *National Library of Australia D-Lib Magazine*, 2009.

[Hou83] H.S. Hou. *Digital document processing.* A Wiley-Interscience publication. Wiley, 1983.

[HSD73] Robert M. Haralick, K. Shanmugam, and Its'Hak Dinstein. Textural features for image classification. *Systems, Man and Cybernetics, IEEE Transactions on*, SMC-3(6) :610–621, nov 1973.

[HYR86] Akihide Hashizume, Pen-Shu Yeh, and Azriel Rosenfeld. A method of detecting the orientation of aligned components. *Pattern Recogn. Lett.*, 4(2) :125–132, April 1986.

[iDTJT96] Øivind Due Trier, Anil K. Jain, and Torfinn Taxt. Feature extraction methods for character recognition-a survey. *Pattern Recognition*, 29(4) :641 – 662, 1996.

[IK00] Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, 40(10) :1489–1506, 2000.

[JB92] Anil K. Jain and Sushil K. Bhattacharjee. Text segmentation using gabor filters for automatic document processing. *Mach. Vis. Appl.*, 5(3) :169–184, 1992.

[JD01] Hui Jiang and Li Deng. A bayesian approach to the verification problem : Applications to speaker verification. *Speech and Audio Processing, IEEE Transactions on*, 9(8) :874–884, 2001.

[JEBE07] G. Joutel, V. Eglin, S. Bres, and H. Emptoz. Curvelets based queries for cbir application in handwriting collections. In *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, volume 2, pages 649–653, Sept 2007.

[Jia05] Hui Jiang. Confidence measures for speech recognition : A survey. *Speech communication*, 45(4) :455–470, 2005.

[JJKK99] Ki-Young Jeong, Keechul Jung, Eun Yi Kim, and Hang Joon Kim. Neural network-based text location for news video indexing. In *ICIP (3)*, pages 319–323, 1999.

[JK96] Anil K. Jain and Kalle Karu. Learning texture discrimination masks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(2) :195–205, 1996.

[JKJ04] Keechul Jung, Kwang In Kim, and Anil K. Jain. Text information extraction in images and video : a survey. *Pattern Recognition*, pages 977–997, 2004.

[Jun01] Keechul Jung. Neural network-based text location in color images, 2001.

[JZ96] Anil K. Jain and Yu Zhong. Page segmentation using texture analysis. *Pattern Recogn.*, 29(5) :743–770, May 1996.

[Kau11] Jasdeep ; Kaur Jappreet Kaur, Manpreet ; Kaur. Survey of contrast enhancement techniques based on histogram equalization. *International Journal of Advanced Computer Science and Applications*, 2011.

[Kim96] Hae-Kwang Kim. Efficient automatic text location method and content-based indexing and structuring of video database. *Journal of Visual Communication and Image Representation*, 1996.

[Kit05] Kimcheng Kith. *Contribution à la description des formes par la transformée en ondelettes*. PhD thesis, 2005. Thèse de doctorat dirigée par Zahzah, El-Hadi Informatique La Rochelle 2005.

[KNRN93] J. Kanai, T.A. Nartker, S. Rice, and G. Nagy. Performance metrics for document understanding systems. In *Document Analysis and Recognition, 1993., Proceedings of the Second International Conference on*, pages 424–427, Oct 1993.

[KRNN95] Junichi Kanai, Stephen V. Rice, Thomas A. Nartker, and George Nagy. Automated evaluation of ocr zoning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 17(1) :86–90, January 1995.

[KRSG03] Swapnil Khedekar, Vemulapati Ramanaprasad, Srirangaraj Setlur, and Venu Govindaraju. Text - image separation in devanagari documents. In *ICDAR*, pages 1265–1269. IEEE Computer Society, 2003.

[KS97] Thomas Kemp and Thomas Schaaf. Estimating confidence using word lattices. In *EuroSpeech*, 1997.

[KS02] Chao Kan and Mandyam D Srinath. Invariant character recognition with zernike and orthogonal fourier–mellin moments. *Pattern Recognition*, 35(1) :143–154, 2002.

[KS12a] Deepak Kumar and Dalwinder Singh. Modified approach of hough transform for skew detection and correction in documented images, 2012.

[KS12b] Deepak Kumar and Dalwinder Singh. Modified approach of hough transform for skew detection and correction in documented images, 2012.

[KSI98] Koichi Kise, Akinori Sato, and Motoi Iwata. Segmentation of page images using the area voronoi diagram. *Comput. Vis. Image Underst.*, 70(3) :370–382, June 1998.

[Lee01] C.-H Lee. *Statistical confidence measures and their applications.* August 2001.

[LJB+95] Y. LeCun, L. D. Jackel, L. Bottou, A. Brunot, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, U. A. Muller, E. Sackinger, P. Simard, and V. Vapnik. Comparison of learning algorithms for handwritten digit recognition. In F. Fogelman and P. Gallinari, editors, *International Conference on Artificial Neural Networks*, pages 53–60, Paris, 1995. EC2 & Cie.

[LKF05] Cheng-Lin Liu, M. Koga, and H. Fujisawa. Gabor feature extraction for character recognition : comparison with gradient feature. In *Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on*, pages 121–125 Vol. 1, Aug 2005.

[Lon98] Sven Loncaric. A survey of shape analysis techniques. *Pattern Recognition*, 31(8) :983 – 1001, 1998.

[LPS+05] SimonM. Lucas, Alex Panaretos, Luis Sosa, Anthony Tang, Shirley Wong, Robert Young, Kazuki Ashida, Hiroki Nagai, Masayuki Okamoto, Hiroaki Yamamoto, Hidetoshi Miyao, JunMin Zhu, WuWen Ou, Christian Wolf, Jean-Michel Jolion, Leon Todoran, Marcel Worring, and Xiaofan Lin. Icdar 2003 robust reading competitions : entries, results, and future directions. *International Journal of Document Analysis and Recognition (IJDAR)*, 7(2-3) :105–122, 2005.

[LS95] Louisa Lam and Ching Y. Suen. An evaluation of parallel thinning algorithms for character recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(9) :914–919, 1995.

[LSJ93] Yi Lu, Steven Schlosser, and Michael Janeczko. Fourier descriptors and handwritten digit recognition. *Machine Vision and Applications*, 6(1) :25–34, 1993.

[LSK+12] Sang-Heon Lee, Myoung-Kyu Sohn, Dong-Ju Kim, Byungmin Kim, and Hyunduk Kim. Face recognition of near-infrared images for interactive smart tv. In *Proceedings of the 27th Conference on Image and Vision Computing New Zealand*, IVCNZ '12, pages 335–339, New York, NY, USA, 2012. ACM.

[LSZT07] Laurence Likforman-Sulem, Abderrazak Zahour, and Bruno Taconet. Text line segmentation of historical documents : a survey. *International Journal of Document Analysis and Recognition (IJDAR)*, 9(2-4) :123–138, 2007.

[LT03] Yue Lu and Chew Lim Tan. A nearest-neighbor chain based approach to skew estimation in document images. *Pattern Recognition Letters*, 24(14) :2315–2323, 2003.

[Mäe03] Topi Mäenpää. *The Local binary pattern approach to texture analysis : Extenxions and applications*. Oulun yliopisto, 2003.

[Mau06] Julie Mauclair. *Mesures de confiance en traitement automatique de la parole et applications*. PhD thesis, Université du Maine, 05/12/2006 2006.

[MCLS02] Wenge Mao, Fu-Lai Chung, Kenneth K. M. Lam, and Wan-Chi Siu. Hybrid chinese/english text detection in images and video frames. In *ICPR (3)*, pages 1015–1018, 2002.

[MG93] Sriganesh Madhvanath and Venu Govindaraju. Holistic lexicon reduction. *Proceedings Int. Workshop on Frontiers in Handwriting Recognition*, pages 71–81, 1993.

[MM91] L. Mathan and Laurent Miclet. Rejection of extraneous input in speech recognition applications, using multi-layer perceptrons and the trace of hmms. In *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*, pages 93–96 vol.1, Apr 1991.

[MM99] Stefano Messelodi and Carla Maria Modena. Automatic identification and skew estimation of text lines in real scene images. *Pattern Recognition*, 32(5) :791–810, 1999.

[MQR03] Jr. Maurer, C.R., Rensheng Qi, and V. Raghavan. A linear time algorithm for computing exact euclidean distance transforms of binary images in arbitrary dimensions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(2) :265–270, Feb 2003.

[MSY92] S. Mori, C.Y. Suen, and K. Yamamoto. Historical review of ocr research and development. *Proceedings of the IEEE*, 80(7) :1029–1058, Jul 1992.

[Mul06] Remy Mullot. Les documents écrits de la numérisation à l'indexation par le contenu, November 2006. " La reconnaissance des structures" chapitre 3, traité IC2, Hermes Lavoisier, ISBN 2-7462-1143-2, pp. 87-178.

[Nab13] Adnan Abou Nabout. Object shape recognition using wavelet descriptors. *Journal of Engineering*, 2013(15), 2013.

[Nag86] S ; Stodard S D Nagy, G. ; Seth. *Document analysis with an expert system.* Elsiver science, 1986.

[Nag95a] George Nagy. Document image analysis : Automated performance evaluation. In *In Document Analysis Systems*, pages 137–156. World Scientific, 1995.

[Nag95b] George Nagy. Document image analysis : Automated performance evaluation. In *In Document Analysis Systems*, pages 137–156. World Scientific, 1995.

[ndF13] Biblothèque nationale de France. Référentiel ocr. `http://www.bnf.fr/documents/ref_num_ocr.pdf`, 2013.

[Ng06] Hui-Fuang Ng. Automatic thresholding for defect detection. *Pattern Recogn. Lett.*, 27(14) :1644–1649, October 2006.

[NGP07] Manjunath Aradhya V N, Hemantha Kumar G, and Shivakumara P. Skew detection technique for binary document images based on hough transform. *International Journal of Computer, Information Science and Engineering*, 1(8) :74 – 80, 2007.

[NKK+88] George Nagy, Junichi Kanai, Mukkai Krishnamoorthy, Mathews Thomas, and Mahesh Viswanathan. Two complementary techniques for digitized document analysis. In *In Proceedings of the ACM Conf. Document Processing Systems*, pages 169–176, 1988.

[NKPH05] Stéphane Nicolas, Yousri Kessentini, Thierry Paquet, and Laurent Heutte. Handwritten document segmentation using hidden markov random fields. In *ICDAR*, pages 212–216. IEEE Computer Society, 2005.

[NRE97] Chalapathy V Neti, Salim Roukos, and E Eide. Word-based confidence measures as a guide for stack search in speech recognition. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 2, pages 883–886. IEEE, 1997.

[NS95] D. Niyogi and S.N. Srihari. Knowledge-based derivation of document logical structure. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, volume 1, pages 472–475 vol.1, Aug 1995.

[O'G93] L. O'Gorman. The document spectrum for page layout analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15(11) :1162–1173, November 1993.

[OPM02] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7) :971–987, 2002.

[Ots79] Nobuyuki Otsu. A Threshold Selection Method from Gray-level Histograms. *IEEE Transactions on Systems, Man and Cybernetics*, 9(1) :62–66, 1979.

[Pos86] W. Postl. Detection of linear oblique structures and skew scan in digitized documents. 1986.

[PSWK07] U. Pal, N. Sharma, T. Wakabayashi, and F. Kimura. Off-line handwritten character recognition of devnagari script. In *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, volume 1, pages 496–500, Sept 2007.

[PV00] Konstantinos N. Plataniotis and Anastasios N. Venetsanopoulos. *Color Image Processing and Applications*. Springer-Verlag New York, Inc., New York, NY, USA, 2000.

[RJN94] Stephen V Rice, Frank R Jenkins, and Thomas A Nartker. The third annual test of ocr accuracy. *1994 Annual Report of ISRI, University of Nevada, Las Vegas*, pages 11–50, 1994.

[RJN95] Stephen V Rice, Frank R Jenkins, and Thomas A Nartker. The fourth annual test of ocr accuracy. *1995 Annual Report of ISRI, University of Nevada, Las Vegas*, pages 11–50, 1995.

[RJN96] Stephen V Rice, Frank R Jenkins, and Thomas A Nartker. *The fifth annual test of OCR accuracy*. Information Science Research Institute, 1996.

[RLJ97] Mazin G Rahim, Chin-Hui Lee, and Biing-Hwang Juang. Discriminative utterance verification for connected digits recognition. *Speech and Audio Processing, IEEE Transactions on*, 5(3) :266–277, 1997.

[Ros99] Christophe Rosenberger. *Mise en oeuvre d'un systÃ¨me adaptatif de segmentation d'images*. PhD thesis, Rennes 1, Grenoble, 1999. Th. : traitement du signal.

[RV94] Sabine Randriamasy and Luc Vincent. A region-based system for the automatic evaluation of page segmentation algorithms. In *Proceedings of the International Association for Pattern Recognition Workshop on Document Analysis Systems DAS94*, pages 29–41, 1994.

[SB12] SargurN. Srihari and Gregory Ball. An assessment of arabic handwriting recognition technology. In Volker MÃ¤rgner and Haikal El Abed, editors, *Guide to OCR for Arabic Scripts*, pages 3–34. Springer London, 2012.

[SBZ03] Prateek Sarkar, Henry S Baird, and Xiaohu Zhang. Training on severely degraded text-line images. In *Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on*, pages 38–43. IEEE, 2003.

[SC08] I. Steinwart and A. Christmann. *Support Vector Machines*. Information Science and Statistics. Springer, 2008.

[Sch00] Dan Schonfeld. On the relation of order-statistics filters and template matching : optimal morphological pattern recognition. *IEEE Transactions on Image Processing*, 9(5) :945–949, 2000.

[SG99] Manhung Siu and Herbert Gish. Evaluation of word confidence for speech recognition systems. *Computer Speech & Language*, 13(4) :299–319, 1999.

[SK95] Michael Smith and Takeo Kanade. Video skimming for quick browsing based on audio and image characterization. Technical Report CMU-CS-95-186, Computer Science Department, Pittsburgh, PA, July 1995.

[SKC02] B. K. Sin, S. K. Kim, and B. J. Cho. Locating characters in scene images using frequency features. In *Proceedings of the International Conference on Pattern Recognition*, volume 3, pages 489–492, 2002.

[SL96] R.A. Sukkar and Chin-Hui Lee. Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition. *Speech and Audio Processing, IEEE Transactions on*, 4(6) :420–429, Nov 1996.

[SMJ+00] AGNE S., ROGGER M., ROHRSCHNEIDER J., Lopresti Daniel P., and Jiangying Zhou. Benchmarking of document page segmentation. *SPIE proceedings series*, 3967(7) :165–171, 2000. eng.

[SRP13] Ahmed Ben Salah, Nicolas Ragot, and Thierry Paquet. Adaptive detection of missed text areas in ocr outputs : application to the automatic assessment of ocr quality in mass digitization projects. In *DRR*, 2013.

[SSLJ97] Rafid A Sukkar, Anand R Setlur, Chin-Hui Lee, and John Jacob. Verifying and correcting recognition string hypotheses using discriminative utterance verification. *Speech Communication*, 22(4) :333–342, 1997.

[SSPH+01] Rubén San-Segundo, Bryan Pellom, Kadri Hacioglu, Wayne Ward, and José M Pardo. Confidence measures for spoken dialogue systems. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, volume 1, pages 393–396. IEEE, 2001.

[Suk94] R.A. Sukkar. Rejection for connected digit recognition based on gpd segmental discrimination. pages 393–396, 1994.

[TC07] Edouard Thiel and David Coeurjolly. Distances discrètes, June 2007. Géométrie discrète et images numériques, Ouvrage collectif, Traité IC2, Hermès, Eds. David Coeurjolly, Annick Montanvert, Jean-marc Chassery.

[TJ98] Mihran Tuceryan and Anil K. Jain. Texture analysis, 1998.

[Tuc94] Mihran Tuceryan. Moment based texture segmentation, 1994.

[TW03] Salvatore Tabbone and Laurent Wendling. Multi-scale binarization of images. *Pattern Recognition Letters*, 24(1-3) :403–411, 2003. Article dans revue scientifique avec comité de lecture. A03-R-006 || tabbone03a A03-R-006 || tabbone03a.

[Vay] Catherine Vayness, Shalev Lerouge. Charte de traitement : Ocr brut et hq, alto. Technical report, Bibliothèque nationale de France.

[VGS96] Vladimir Vapnik, Steven E. Golowich, and Alex Smola. Support vector method for function approximation, regression estimation, and signal processing. In *Advances in Neural Information Processing Systems 9*, pages 281–287. MIT Press, 1996.

[vRjktaN94] Stephen v. Rice, junichi kanai, and thomas a. Nartker. An algorithm for matching ocr-generated text strings. *International Journal of Pattern Recognition and Artificial Intelligence*, 08(05) :1259–1268, 1994.

[WBR+97] Mitch Weintraub, Françoise Beaufays, Ze'ev Rivlin, Yochai Konig, and Andreas Stolcke. Neural-network based measures of confidence for word recognition. In *in Proc. ICASSP*, pages 887–890, 1997.

[WDL05] Xuewen Wang, Xiaoqing Ding, and Changsong Liu. Gabor filters-based feature extraction for character recognition. *Pattern Recognition*, 38(3) :369 – 379, 2005.

[Wie80] Thomas F. Wiene. Rapid texture identification. In SPIE, editor, *Image Processing for Missile Guidance*, San Diego, 1980.

[WW82] K. Y. Wong and F. M. Wahl. Document analysis system. *IBM Journal of Research and Development*, 26 :647–656, 1982.

[XB12] Pingping Xiu and H.S. Baird. Whole-book recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(12) :2467–2480, Dec 2012.

[You94] S.R. Young. Detecting misrecognitions and out-of-vocabulary words. In *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, volume ii, pages II/21–II/24 vol.2, Apr 1994.

[ZKJ95] Yu Zhong, Kalle Karu, and Anil K. Jain. Locating text in complex color images. *Pattern Recognition*, 28(10) :1523–1535, 1995.

[ZL04] Dengsheng Zhang and Guojun Lu. Review of shape representation and description techniques, 2004.

[ZR] Rong Zhang and Alexander I Rudnicky. Word level confidence annotation using combinations of features.