Projet ANR DigiDoc

WP7 : Qualité OCR

Etat de l'art sur la caractérisation d'un document à OCRiser

Auteurs : Geneviève Cron – Ahmed Ben Salah – Nicolas Ragot – Kamel Ait-Mohand – Thierry Paquet

Table des matières

Eléments ayant une influence sur la qualité et l'efficacité des OCRs

La numérisation de documents patrimoniaux est devenue une des activités prioritaires des institutions culturelles du XIXème siècle, notamment les grandes bibliothèques. Ce procédé permet la conservation par voie numérique et la diffusion d'œuvres écrites au-delà des murs et des horaires d'ouverture des bibliothèques. L'accès au contenu textuel des documents ainsi numérisés est possible grâce à une recherche sur les informations présentes dans les notices.

Depuis le début des années 2000, la numérisation s'accompagne d'un processus de conversion en mode texte, dit OCR pour Optical Character Recognition. Pourquoi faire cette conversion ? Elle permet avant tout l'accès au contenu "plein texte", c'est-à-dire au lien qu'il existe entre un mot et la liste des documents qui le contiennent. En ce sens, l'"OCR-isation" des documents numérisés permet l'indexation des documents "par le contenu". La conversion peut aussi être utile dans d'autres contextes comme la production d'Ebooks, la citation de parties de texte, la généalogie, la phonémisation.

Le choix de la solution de numérisation doit prendre en compte la qualité du document original ainsi que toutes les étapes de traitement des documents depuis l'acquisition, la conversion du contenu jusqu'à la correction et la mise en exploitation du document final.

Plusieurs facteurs agissent sur la qualité finale des résultats de conversion de l'OCR :

- Des caractéristiques de l'œuvre (contenu textuel, illustrations, présence de formules mathématiques,…) et de son édition (éditeur, date d'édition)
- Des caractéristiques du papier, de l'impression, de la qualité de conservation, de l'encre, de l'encrage, de la fonte;
- Des caractéristiques de la numérisation (qualité du scan, paramètres de numériseur) et de l'image numérique.

Ce compte rendu présente les différents critères qui agissent sur la qualité des résultats de l'OCR et les méthodes d'évaluation de ces résultats qui existent dans la littérature.

1. CRITERES CONCERNANT L'ŒUVRE

1.1. Composition du texte

Indiscutablement, les OCRs sont calibrés pour reconnaître du texte dactylographié, des mots. Selon [4], tous les éléments non purement textuels peuvent perturber la reconnaissance : images, tableaux, formules mathématiques et chimiques, chiffres, hiéroglyphes, annotation manuscrites, éléments graphiques sont autant d'éléments perturbateurs pour les OCRs.

1.2. Langues et alphabets

1.2.1. La langue à reconnaître

Les accents

Les premiers OCRs ont été développés pour des archives du XXème siècle. Ces documents, en anglais, ne comportent quasiment aucun signe diacritique (seul le point sur le "i" est un accent. Lorsque les OCRs ont commencé à travailler sur des langues étrangères, comme le français, sont apparus les problèmes de reconnaissance des accents. Les accents ont la spécificité d'être beaucoup plus petits qu'un caractère. Or les OCRs fonctionnent tous de la même manière, ils commencent par former une image noir et blanc (binarisation) puis recherchent ensuite les éléments connexes (dites "composantes connexes") qui pourraient être susceptibles d'être des caractères, notamment par leur forme (rapport hauteur /largeur), leur taille (hauteur, largeur). Ces deux étapes sont autant d'occasion pour les algorithmes de faire disparaître les accents : la binarisation peut "effacer des éléments très petits"; la recherche de composantes connexes de taille d'un caractère éliminera presque sûrement les accents trop petits.
La fréquence d'accents dans une langue rend donc le français et l'espagnol et a fortiori les langues slaves plus difficiles à reconnaître que l'anglais.

La longueur des mots

Une des étapes d'un OCR est la "mise en mot". Celle ci consiste à séparer une ligne en mots. Elle se base sur un algorithme statistique qui cherche parmi les espaces entre caractères ceux qui sont les plus grands, et cherche à déterminer s'ils sont significativement plus grands. Ce procédé est souvent optimisé pour une longueur moyenne de mot proche de celle de l'anglais, c'est-à-dire environ 6 caractères. L'allemand et le finnois sont plutôt autour de 9 caractères par mot en moyenne.

Le nombre de symboles

L'alphabet latin contient 26 caractères, mais à peut près 80 caractères supplémentaires sont utilisés dans des documents non techniques. Ces caractères sont des signes de ponctuation, des caractères majuscules, des caractères minuscules et des symboles spéciaux. Tous ces caractères sont incorporés dans le code *ASCII* (American Standard Code for Information Interchange) qui réunit 96 caractères adoptés par les sociétés industrielles Américaine. Le code *Uniforme (unicode)* regroupe tous les caractères et les symboles imprimables de toutes les

langues. Même les symboles spéciaux qui figurent dans certaines publications spéciales comme les dictionnaires, les textes scientifiques sont inclus dans le code *Uniforme*.

1.2.2. Nombre de langues et alphabets

Les OCRs travaillent mieux [4] avec une seule langue par unité documentaire. Même s'il est possible d'ajouter un dictionnaire d'une autre langue, cette opération peut causer des dégâts considérables sur le reste de la reconnaissance. Notamment le nombre d'erreurs et d'ambiguïtés du texte augmente.
De même, la présence d'alphabets non latins peut être préjudiciable à la reconnaissance globale du texte, et affecter même la qualité de la segmentation.
Selon [5], les langues anciennes sont mal connues par les OCRs qui travaillent généralement avec des nouveaux dictionnaires.

1.3. Références et citations

Les textes comportant des références et citations utilisent souvent les notes en bas de page. Ces notes sont des petites écritures en bas de la page, qui servent à expliquer ou référencer des termes spéciaux. La taille de ces polices est gênante pour l'OCR qui aura du mal à estimer la taille moyenne de la police sur la page.
Par ailleurs, la structure du document, pour peut qu'elle soit aussi une des tâches de l'OCR, en est grandement complexifiée.

1.4. Ponctuation

Dans les textes narratifs et descriptifs environ 60% des signets de ponctuation sont des virgules et des points. Dans les périodiques scientifiques, les points sont beaucoup plus nombreux que les virgules. En effet, les points sont utilisés dans les nombres réels, dans les abréviations et dans les fonctions scientifiques ce qui augmente leur nombre. Selon [2], la fréquence des virgules a subit des diminutions au cours de ces dernières siècles. Compte tenu de leurs apparences dans le texte, les points et les virgules sont souvent similaires. En effet leurs petites tailles empêchent les méthodes de reconnaissance de forme de faire la distinction entre les deux.
Les traits d'union « - » et les trait de type « _ » sont relativement courants. Certains systèmes de reconnaissance optique de caractères ne les distinguent pas. On peut trouver dans les textes scientifiques d'autres signes comme les guillemets (pour désigner les citations), les apostrophes et les parenthèses. Dans certaines polices de caractère les apostrophes et les guillemets se distinguent des virgules et des points à travers leur position par rapport la ligne de base. Dans la langue française et espagnole, les guillemets sont utilisés comme des marques de citation.
La fréquence des signes de ponctuation dans les textes varie avec l'auteur et le type des textes, par exemple dans la bible de King James et dans les premiers folios de Shakespeare, plus de 10% des signes de ponctuation sont les deux points « : » par contre dans les textes récents le pourcentage de « : » ne dépasse pas 5%.
Dans les échantillons ISRI (*Informatique Science Research Institute, cf. 3*), la ponctuation et les symboles spéciaux représentent environ 5% du texte. Parmi les

symboles les plus gênants pour les OCRs, on trouve les virgules, les points, les guillemets et les apostrophes.
Les problèmes posés par la ponctuation sont les mêmes que ceux posés par les accents : ils sont trop petits et risquent d'être éliminés dans les pré-traitements.

2. CRITERES CONCERNANT L'OUVRAGE

2.1. Structure textuelle du document

La structure textuelle du document définit l'agencement physique des blocs sur la page. Les textes présentés sous forme de plusieurs colonnes (comme les textes de journal) ou entourés par des images rendent l'opération de reconnaissance des caractères complexes et provoquent beaucoup d'erreurs dans le résultat de l'OCR [4].

LINGUISTIC PUBLICATIONS.
OF
TRÜBNER & CO.,
57 AND 59, LUDGATE HILL, LONDON, E.C.

विद्या नाम नरस्य रूपमधिकं प्रच्छन्नगुप्तं धन
विद्या भोगकरी यशःसुखकरी विद्या गुरूणां गुरुः।
विद्या बन्धुजनो विदेशगमने विद्या परं दैवतं
विद्या राजसु पूजिता न हि धनं विद्याविहीनः पशुः॥

Messrs. TRÜBNER & Co. respectfully solicit orders for all classes of Publications connected with the History, Antiquities, Geography, and Languages of the East, published abroad. Messrs. Trübner & Co. have established agencies in all parts of the East, of Europe, and America, and are thus enabled to furnish such publications with as little delay as possible, and at moderate prices.

AHLWARDT.—The Divans of the Six Ancient Arabic Poets, Ennâbiga, Antara, Tarafa, Zohair, 'Alkama, and Imrulkais; chiefly according to the MSS. of Paris, Gotha, and Leyden, and the collection of their Fragments; with a complete list of various readings of the Text. Edited by W. Ahlwardt, Prof. of Oriental Languages in the University of Greifswald. 8vo, pp. xxx. and 340. 16s.

ALABASTER.—The Wheel of the Law: Buddhism illustrated from Siamese Sources by the Modern Buddhist. a Life of Buddha, and an account of the Phra Bat. By Henry Alabaster, Esq. Demy 8vo. pp. lviii. and 304. 14s.

BALLANTYNE.—Elements of Hindi and Braj Bhâkhâ Grammar. By the late James R. Ballantyne, LL.D. and Editors, revised and corrected. Cr. 8vo. cl., pp. xii. 32.

——— First Lessons in Sanskrit Grammar: together with an Introduction to the Hitopadésa, and Edition, and Impression. By James R. Ballantyne, LL.D. 8vo. cl., pp. viii. and 110. 5s. 6d.

BEAL.—Travels of Fah Hian and Sung-Yun, Buddhist Pilgrims from China to India (400 A.D. and 518 A.D.). Translated from the Chinese by S. Beal, B.A., Trinity College, Cambridge. Cr. 8vo. cl. pp. lxxxiii. and 232, with Map. 10s. 6d.

——— A Catena of Buddhist Scriptures from the Chinese. By S. Beal, B.A., etc. 8vo. cl. pp. xii. and 438. 15s.

——— The Romantic Legend of Sâkya Buddha. From the Chinese-Sanscrit by the Rev. Samuel Beal. Cr. 8vo. cl. pp. 460. 10s.

BEAMES.—A Comparative Grammar of the Modern Aryan Languages of India (as will Hindi, Panjâbi, Sindhi, Gujarati, Marâthi, Oriya, and Bengali). By John Beames, Bengal C.S., M.R.A.S., etc.

Vol. I. On Sounds. 8vo. cl. pp. xvi. and 360. 16s.—Vol. II. The Noun and the Pronoun. 8vo. cl. pp. xii. and 346. 16s.

Figure 1 : Exemple d'un texte présenté sous forme de deux colonnes.

2.2. Date d'édition

La date d'édition a un impact majeur sur la qualité de la conversion OCR. D'ailleurs, la Bibliothèque Nationale de France limite, dans ses marchés, la conversion OCR à des documents postérieurs à 1650. Pendant longtemps, cette limite était 1750. Les raisons sont multiples : fontes utilisées, langues anciennes

non connue des OCRs, mise en page complexe, défauts d'impression, œuvre mal conservée, caractères cassés, morceaux de pages manquants.
Une expérience a été faite sur 924 documents sélectionnés de façon aléatoire à partir d'une liste de 2000 documents obtenus à travers une requête lancée aux bases de données de la BnF. Cette expérience analyse le comportement des taux d'OCR suivant la date d'édition. Compte tenu des politiques de numérisation, le nombre de documents par période ne peut être normalisé. On voit en effet sur la Figure 2 (courbe bleue) qu'il n'y a presque aucun document OCR-isé avant 1700 (en raison de leur difficulté de traitement) et aussi très peu au XX^{ème} siècle (œuvres sous droits). L'essentiel des documents numérisés sont donc extraits des collections de la fin du XVIII^{ème} siècle et de la totalité du XIX^{ème} siècle.

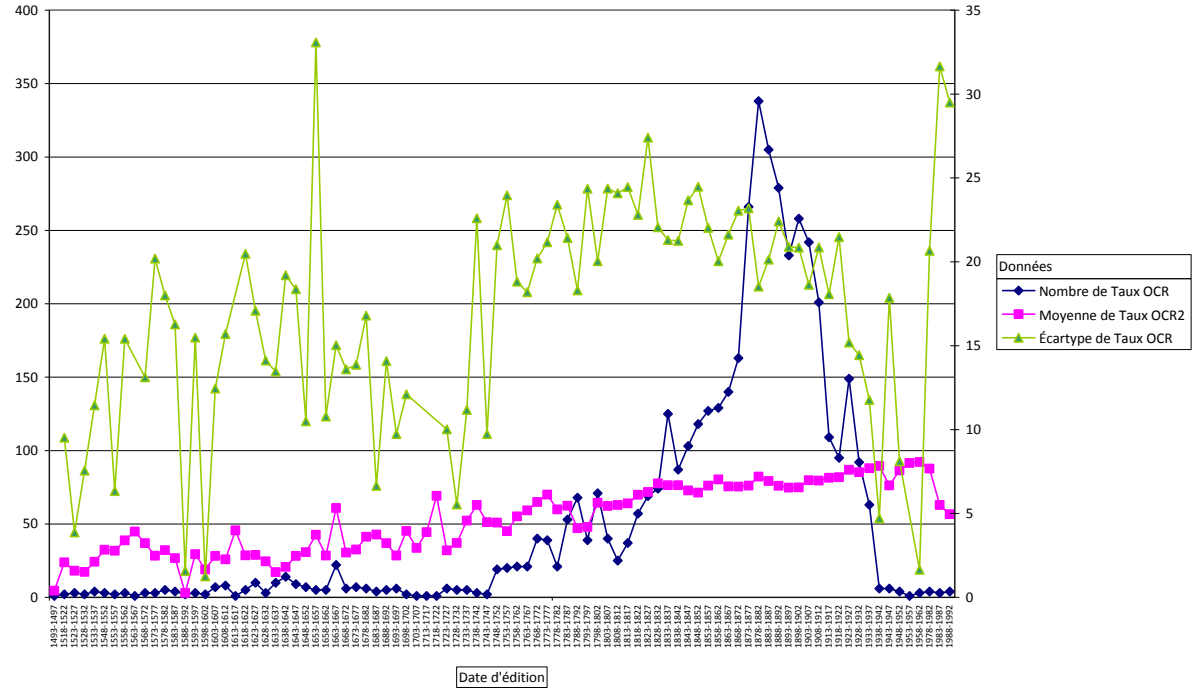


Figure 2 : Date édition, nombre de documents, taux OCR moyen et écart type

Les résultats bruts sont présentés sur la Figure 3. On peut voir que les caractéristiques ont un lien mais, qu'une corrélation est difficile à expliciter.
La Figure 4 propose de regrouper les résultats sur l'ensemble des documents édités sur une période de 5 ans. A ce niveau, une corrélation est plus visible.

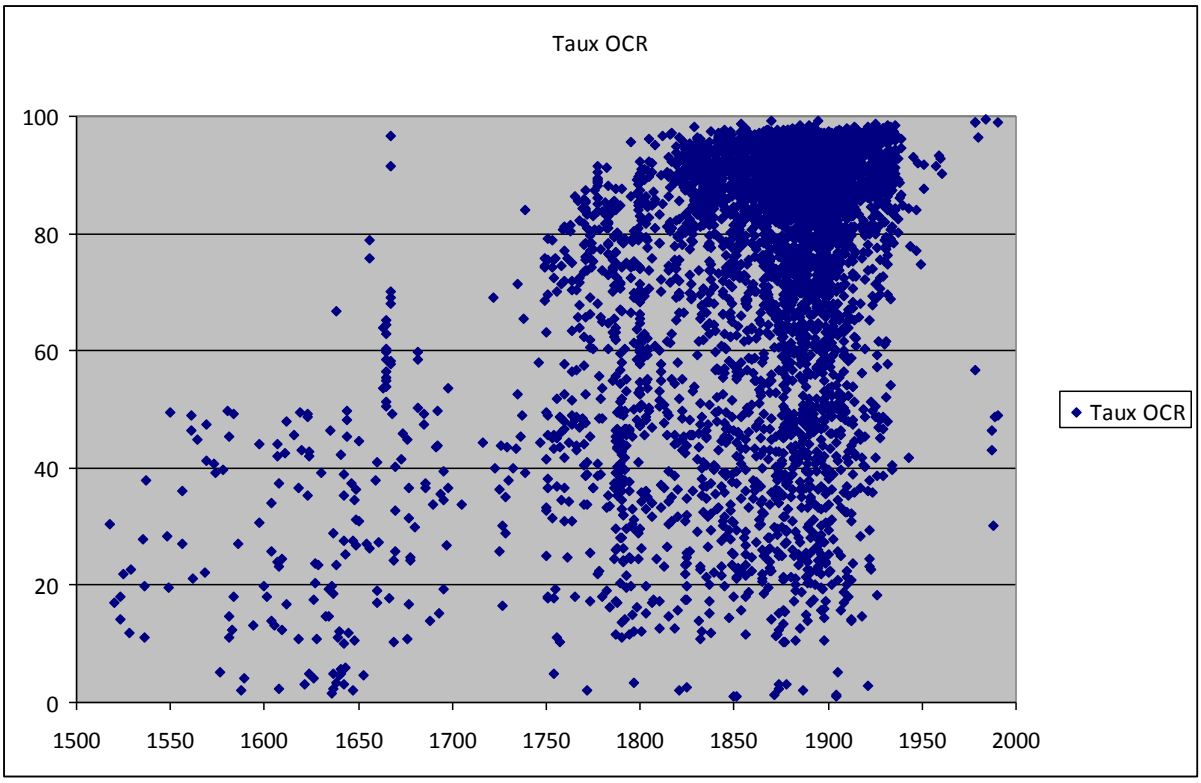


Figure 3 : Taux OCR moyen des documents vs leur date d'édition

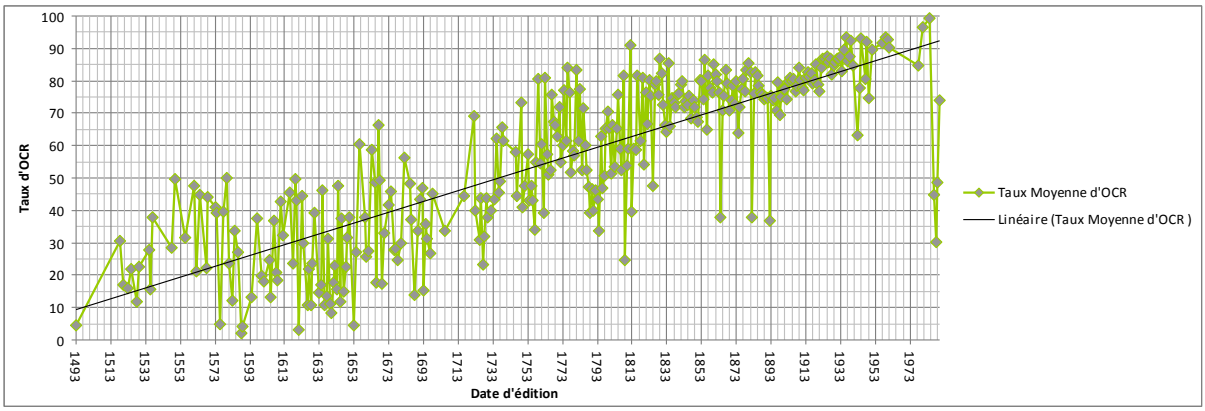


Figure 4 Taux OCR moyen des documents d'une année donnée (moyenne sur les documents)

2.3. Ouverture

L'ouverture du document est l'angle maximal entre deux pages. Lorsque la reliure originale ou de conservation est très serrée, ou si le dépliant est très dur : il est alors difficile de mettre le document à plat.

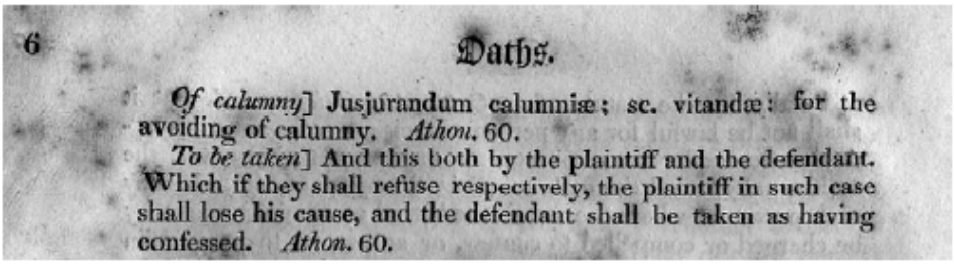
La numérisation sera alors réalisée sur des numériseurs particuliers qui scannent sans mettre à plat. Ceci peut alors poser des problèmes de mise au point et de zone noire dans la reliure, cette zone étant moins éclairée. Il existe aussi des numériseurs qui descendent dans les marges du document qui n'est pas ouvert complètement. Dans tous les cas, l'OCR peinera à réaliser une bonne conversion [4].

2.4. Marges de petit fond

Les marges sont les espaces blancs qui séparent l'écriture de l'extrémité de la page. Les marges d'un écrit font partie de la présentation, plus exactement de l'ordonnance, c'est-à-dire de la répartition de l'écrit sur la page. Les ouvrages contenant des pages dont le texte est pris dans la reliure donnent des images du texte incomplet. Les effets sont les mêmes que ceux liés à l'ouverture.

2.5. Qualité de papier

La qualité de 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. Si il existe des tâches ou/et des rousseurs sur les pages, le document numérisé risque d'être illisible. Les tâches proviennent de la dégradation du papier à travers les siècles. La**Erreur ! Source du renvoi introuvable.** Figure 5 montre un exemple de tâches sombres, elles peuvent aussi être claires.


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 5 : Tâches noires sur l'image de la page

Les tâches claires génèrent des caractères cassés de formes variables. Les tâches sombres génèrent des caractères liés avec des trous bouchés.

Les caractères cassés provoquent des erreurs dans les résultats de l'OCR [9] car les différentes parties du caractère sont considérées comme des formes distinctes. Si le papier est très transparent au point que l'écriture qui se trouve sur une face de la page apparaisse dans l'autre face, le processus de conversion par OCR devient difficilement exploitable [4].

2.6. Typographie

Définition: La typographie regroupe tous les critères des caractères (italiques, soulignés, avec ombre du fond) et les polices de caractère inhabituels et les caractères ont petits ou grands taille.

La typographie est l'art de conception de communication à travers les textes imprimés. Une bonne typographie est généralement transparente de façon à ce qu'elle transporte le message de façon claire. Mais beaucoup de polices de caractère utilisées dans l'impression restent dérivées de la calligraphie médiévale. Or il est bien connu que les typographies gothiques rendent le processus de conversion délicat et non fiable. De plus, l'utilisateur de l'OCR n'a aucun contrôle sur la typographie du texte et il se trouve souvent confronté à des mises en page, des

polices de caractère et des tailles de caractères qui ne sont pas idéales pour les OCRs.

2.6.1. Composition et variantes d'une même police

Les considérations esthétiques dictent une certaine cohérence dans la forme des lettres et des chiffres. Une police de caractère regroupe tous les symboles constitués par des formes connexes. Les exemples populaires des polices sont *Times New Roman*, *Bookman*, *Bodoni*, *Futura*. Mais les polices de caractères se distinguent aussi par leur constructeur. Chaque caractère offre plusieurs variantes stylistiques, par exemple *italique* et *les petit majuscule*, et *le poids* (*léger, moyen, gras, Extrabold*), dont l'utilisation est guidé par des conventions de mise en page. En typographie classique, une police de caractère est un ensemble complet de lettres et autres symboles, chaque caractère a une forme et une taille unique. La taille des caractères est mesurée en points de l'imprimante: un point est 1/72 de pouce. Pour des raisons de lisibilité, les éditeurs utilisent souvent une taille de texte entre 9 points et 12 points. Dans les échantillons de l'ISRI 95% des textes sont en taille 12 points.

Les petites écritures sont utilisées pour les notes de bas de la page par contre les grandes écritures sont utilisées pour les titres et les sous titres. La taille en point des caractères représente uniquement la taille du texte principal. La hauteur des caractères varie suivant la police. Les distances relatives entre *ascender line, base line* et *the descender line* sont utilisés par les OCRs pour déterminer la taille des caractères. La hauteur d'un caractère de taille 10 points numérisé avec une résolution de 300ppp est d'environ 24 pixels.
La taille horizontale d'un caractère est appelée « *set* » et elle est aussi mesurée en point. Les largeurs des caractères varient dans les polices normales avec un rapport de 3:1 de « m » jusqu'au « i ». La variation de largeur des caractères augmente la lisibilité des mots car leurs formes deviennent beaucoup plus différentiables.

Le *Kerning* est un processus particulier. Pour certains types de caractères, les espaces blancs ne suffisent pas pour séparer les caractères. Une paire de caractères est dit « *kerned* » si on ne peut pas les séparer verticalement. Les caractères italiques sont toujours « *kerned* ».
Dans plusieurs polices de caractères, certaines paires de lettres, comme fi, fl, ffi, ffl, sont regroupées dans une seule forme appelée « *ligature* ». Les ligatures sont considérées comme des cas extrêmes de *Kerning*. En effet, les formes composant les ligatures ne peuventt pas être facilement décomposées en formes significatives. Dans ce cas, la segmentation caractère peut être perturbée par la présence de ligatures.

2.6.2. Ambiguïtés liées à la police

Tous les systèmes d'OCR reconnaissent les caractères à travers leur forme. Cependant, la forme est un concept difficile à saisir. L'article [2] considère la forme comme une classe d'équivalence induite sur un contexte particulier. Cette

définition ne reflète pas l'idée que la forme est une propriété globale qui doit être invariante. Dans la reconnaissance des 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.

Les ambiguïtés peuvent également venir de l'alphabet utilisé. Ainsi, dans l'alphabet latin, il n'y a que peu de caractères qui présentent des formes très différentes : comme par exemple « A », « a » et « ɑ». Par contre, en arabe, la forme d'une lettre dépend parfois de ses voisins selon une grammaire bien formée graphiquement.

D'autre part, les méthodes de conversion utilisées dans le moteur de l'OCR doivent dépendre de l'époque de l'édition du document. En effet, d'après les études de Gorge Nagy, jusqu'au premier quart du 20ème siècle, les majuscules ont été rarement utilisées sauf au début des phrases. Les chiffres arabes apparaissent rarement dans les textes, du coup, la distinction entre « l,1 », « I,1 » et « 0,o » ne pose pas beaucoup de problème dans la conversion des documents de cette époque. Les symboles spéciaux, qui compliquent généralement l'opération de conversion des textes, tel que « $ » et « % », sont des inventions relativement modernes. Dans ce genre de documents, les ambiguïtés entre symboles sont plus fréquentes. Alors que ces confusions entre les formes similaires sont facilement résolues par le lecteur humain, parce qu'il ne traite pas chaque caractère isolement, la difficulté est bien plus grande pours les OCRs qui travaillent souvent dans un premier temps sans contexte et sans sémantique. C'est la raison pour laquelle, des polices de caractère ont été spécialement conçues pour les OCRs, tel qu'OCR-A et OCR-B. Elles contiennent des caractéristiques spéciales qui permettent la distinction entre les symboles similaires comme « O, 0 » et « 1,l ».

Enfin, dans un document présentant plusieurs polices, il est probable qu'un symbole similaire corresponde à deux caractères différents présentés par deux polices différentes. Du coup nous pouvons voir ici l'intérêt de l'analyse des caractères par classe de police de caractère.

2.7. Les défauts d'impression

Les défauts de l'impression couvrent tous les artéfacts qui aberrent la clarté et la lisibilité de l'écriture. Ils sont directement liés aux techniques d'impressions et peuvent donc varier selon la date d'édition*. Ainsi, alors que ces défauts pouvaient être fréquents avec les méthodes reposant sur l'encrage de tampons/caractères, ils tendent à se réduire avec les techniques récentes très précises. Voici quelques problèmes fréquents :

- les tâches blanches qui se trouvent sur les caractères donnent l'effet de caractères cassés ;

- à cause de problèmes dans les rubans des imprimantes on peut trouver des lignes blanches sur les phrases de textes ce qui gène le processus de reconnaissance de forme dans le moteur de l'OCR ;
- l'encrage faible ou de mauvaise qualité de certains documents peut dégrader le contraste entre l'écriture et le fond (papier) ce qui agit sur la lisibilité du document. Au contraire, l'encrage intensif du document peut nous donner des caractères épais avec des formes non uniformes et des trous bouchés (comme pour le « e » ou le « a »), ce qui complique la tâche de l'OCR [5] ;
- l'espace non uniforme entre les colonnes, les lignes, les mots et les caractères peut être une source d'erreur récurrente dans la tâche de segmentation des blocs de différents niveaux.

* Evolution des techniques d'impressions :
Au début, seuls les caractères/tampons métalliques sont utilisés dans l'impression. Puis, un siècle après Gutenburg, les caractères sont découpés et composés manuellement. Après la première guerre mondiale ce processus fastidieux a été remplacé par des machines à clavier comme les systèmes Monotype et Linotype. La lithographie est inventée depuis 1798, mais récemment cette technique a été utilisée pour les livres précieux. Comme l'indique son nom, la lithographie est utilisée originalement pour affiner certain types de pierre d'encre. Mais désormais, cette technique désigne tous les types d'impression plats sans caractères émergents. La lithographie a permis de faciliter le processus de production des illustrations dans les imprimeries de presse. « Ink-on-stone » utilise généralement des plaques plastiques ou métalliques pour obtenir une bonne qualité de couleur dans l'impression. Alors que la lithographie nécessite encore des caractères métalliques pour construire les plaques, la *photocomposition* est accomplie par un faisceau de lumineux commandé par ordinateur d'une machine de photocomposition. Ces techniques sont aujourd'hui remplacées par des techniques d'impression directe appelée xérographique. Les imprimantes *Lazer* utilisent des résolutions d'impression qui vont de 300ppp jusqu'à 600ppp.

2.8. Orientation de l'écriture dans la page

Le sens de l'écriture sur la page est supposé horizontal pour la plupart des OCRs. Une tolérance est souvent admise (autour de 12°) pour gérer des documents imprimés ou numérisés légèrement en oblique.
Certains documents comportent des textes écrits sur l'axe vertical, notamment dans le cas de tableaux ou de légendes.
Si l'OCR est paramétré pour être capable de détecter des textes verticaux, il sera de toute façon perturbé par la gestion d'hypothèses de blocs de texte d'orientation différentes.

Le traitement de l'OCR est facile lorsque les lignes de l'écriture sont horizontales. Si ce n'est pas le cas, il faut estimer l'inclinaison du texte. Ce traitement supplémentaire n'est pas toujours présent dans l'OCR. Par ailleurs, ci ce traitement n'est pas parfait, la reconnaissance sera dégradée.

3. CRITERES LIES A LA NUMERISATION

Outre les critères intrinsèquement liés au document et à la façon dont il a été généré, un bon nombre de phénomènes ayant une influence directe sur les performances des OCRs proviennent de la façon dont ils ont été numérisés, le résultat de cette numérisation étant le point de départ de l'OCR.

3.1. Bruit noir dans l'image du document

Le bruit est un signal aléatoire causé par le capteur du scanner (voir Figure 6). Le bruit apparu sur l'image à cause des défauts de capteur de scanner dégrade la qualité de l'image du document et gène les algorithmes de reconnaissances des formes intégrés dans les moteurs de l'OCR [9].

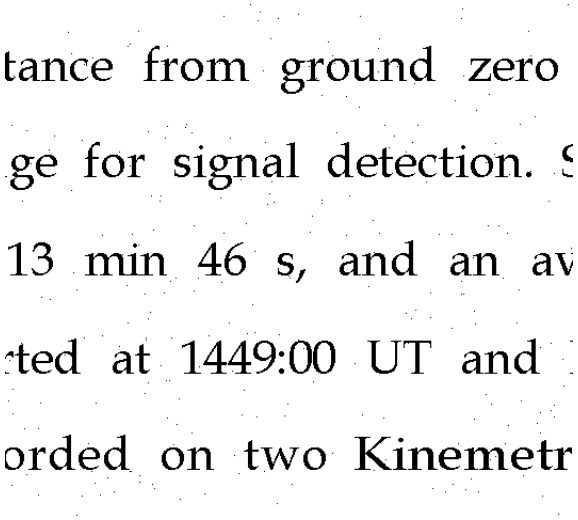


Figure 6 : Bruit noir

3.2. Dynamique de l'image

La dynamique d'une image correspond au codage des couleurs dans un pixel :

- pour une image binaire, le codage peut prendre deux valeurs 0 ou 1, soit 1 bit par pixel (bpp)
- pour une image en niveau de gris, le codage peut prendre toute valeur entière entre 0 (noir) et 255 (blanc), soit 8bpp
- pour une image couleur, chaque canal (bleu, rouge, vert) peut prendre toute valeur entière entre 0 (noir) et 255 (blanc), soit 3 x 8bpp = 24 bpp

La BnF a largement dans un premier temps numérisé en noir et blanc. Depuis quelques années, elle numérise les pages contenant des illustrations et la presse en niveau de gris [11]. Aujourd'hui en 2012, toutes les images sont produites au moins en niveaux de gris, voire en couleur.

Une étude [1] réalisée pendant le projet Impact prouve que le niveau de gris permet d'améliorer notablement le taux de reconnaissance de l'OCR. Le passage à la couleur n'apporte pas beaucoup plus.

Pour information, aucun OCR commercial ne fonctionne aujourd'hui avec un traitement en niveau de gris de reconnaissance. En général, la binarisation de l'image intervient très tôt, et la dynamique de niveaux de gris n'est exploitée que dans l'intelligence de la binarisation. Quelques expériences de segmentation caractère en niveau de gris ont été élaborées au moins dans le domaine postal.

3.3. Zone sombre coté reliure

Cette caractéristique est obtenue lorsque la reliure du document est trop dure, de façon que l'ouverture de livre ne se fasse pas à 180°. Le processus de numérisation de ces documents donne des images avec des intensités lumineuses variables comme par exemple sur la Figure 7. L'étape de binarisation de l'OCR fera alors probablement disparaître les zones sombres [10] et empêchera la reconnaissance d'être efficace.

Figure 7 : Zone sombre coté reliure

3.4. Résolution

Idéalement, le choix de la résolution de l'acquisition dépend de la qualité du contenu en termes typographiques. La difficulté souvent rencontrée est de pouvoir adapter la résolution aux différentes tailles de caractères et épaisseurs graphiques présent dans le document.
Le mauvais choix de la résolution peut conduire [5], en sous échantillonnage, à un manque d'information et en sur échantillonnages, à un surplus d'information, souvent traduit par la présence d'un bruit plus abondant.

Dans l'étude [1], il a été noté que le logiciel FineReader 10 donne des résultats optimaux pour une numérisation en 300 à 400 DPI. Les résultats se dégradent pour des résolutions plus élevées.

3.5. Compression

L'effet de la compression aura forcément un effet négatif sur le résultat des OCRs. Ceci d'autant plus que la compression est optimisée pour réduire les effets visuels de la perte, mais pas forcément les effets algorithmiques [11]. Pour garder les informations de l'image, il est préférable d'utiliser un format d'image non propriétaire qui utilise un algorithme de compression sans perte.

3.6. Inclinaison

L'inclinaison des pages du document numérisé est l'angle entre l'horizontale du document et l'horizontale du document numérisé. Les effets [3] sont les mêmes que ceux de l'orientation du document source.

3.7. Contraste et luminosité

La luminosité permet de jouer sur l'éclairage du document à capturer : de plus clair à plus sombre. Le contraste permet de varier l'accentuation ou l'atténuation des transitions Noir/Blanc. Ces deux paramètres sont souvent corrélés entre eux et jouent un grand rôle dans la qualité de reconnaissance.
Les expériences [11] ont montré qu'on peut passer d'un taux de reconnaissance de 0% à 99% par un changement léger de paramètre de luminosité.
Il est impossible de trouver un réglage idéal sur toutes les images du document. En effet, suivant l'état de l'hétérogénéité des fonds à traiter, il peut s'avérer impossible de mettre en œuvre un processus industriel automatique.

4. EVALUATION DE LA QUALITE DES OCRS

Nous nous basons sur les travaux de l'ISRI (*Informatique Science Research Institute*).

Depuis 1992, l'ISRI a introduit plusieurs tests [6] [7] [8] sur les systèmes d'OCRs commerciaux pour déterminer la précision de conversion de ces moteurs. Chaque test a été élaboré de façon automatique sous contrôle machine et sans aucune intervention humaine.

Au départ, les documents utilisés au cours de ces tests sont des documents administratifs de l'*US Department of Energy*. Puis cette collection de documents de test a été enrichie par d'autres documents ayant des structures et des langues différents (comme des journaux espagnols, des lettres et des fax).
Les critères testés dans ces travaux sont :

- La précision des caractères.
- Les valeurs de confiance.
- La vitesse moyenne de reconnaissance de caractère.
- La précision par classe de caractère.
- Les effets de résolution de l'image.
- La qualité de la page du document.
- La précision de reconnaissance des mots.
- La précision de reconnaissance des phrases.
- La détection automatique des blocs textuels
- La labellisation des zones dans l'image du document.
- L'efficacité du marquage des caractères.

Tous les travaux d'évaluations réalisés au sein de l'ISRI ont permis d'identifier environ 280 paramètres de dégradations dont quelques uns sont présentés dans [2]. Ces paramètres sont organisés en 4 classes :

- ✓ **Les défauts de numérisation et d'impression :** Impression épaisse, Impression fine, défauts de luminosité, caractères écartés, lignes d'impressions courbées.
- ✓ **Les symboles similaires :** Les symboles verticaux similaires, les autres symboles similaires.
- ✓ **Ponctuation :** les symboles spéciaux, les virgules.
- ✓ **Typographie :** Les caractères Italiques, les caractères soulignés, les caractères avec ombre du fond, les polices de caractères inhabituels, les caractères en petits ou grands taille

Prédiction des performances des systèmes d'OCR

1. INTRODUCTION

La prédiction de performances consiste à évaluer le taux de reconnaissance qu'est susceptible d'obtenir un système d'OCR sur une image spécifique. Cette prédiction est basée sur l'analyse des caractéristiques de l'image du document. Cette problématique prend naturellement sa place dans les programmes de numérisation de masses de documents, tels que ceux actuellement en cours dans les grandes bibliothèques publiques. Par exemple, la Bibliothèque Nationale de France dispose de collections contenant des millions d'ouvrages, ce qui rend nécessaire une sélection des documents en amont de la chaîne de traitement. Cette sélection est par contre inutile pour un organisme qui dispose de collections beaucoup plus restreintes (quelques milliers d'ouvrages) et qui peut se permettre de numériser et OCRiser la totalité de sa collection de documents.
Ainsi, même si les techniques de recherche d'information (utilisées pour l'accès aux plateformes numériques de ces bibliothèques) sont assez peu sensibles aux erreurs d'OCR[1], il n'en reste pas moins nécessaire de procéder à une correction manuelle des textes OCRisés, surtout si l'on veut disposer de transcriptions exactes du contenu des documents. Cette correction manuelle n'est cependant pas toujours techniquement faisable. L'OCRisation de certains documents anciens ou très dégradés donne un nombre élevé d'erreurs dont la correction peut être très coûteuse, parfois même plus coûteuse qu'une transcription intégralement manuelle du document.
La Bibliothèque Nationale de France a par exemple défini pour ses campagnes de numérisation un seuil acceptable sur le taux de reconnaissance, fixé à 85% au niveau mots, qui permet de déterminer les ouvrages pour lesquels une correction manuelle du texte OCRisé est jugée économiquement acceptable. Ce seuil oblige à procéder à une pré-sélection des documents à OCRiser de manière à ne retenir que les documents susceptibles d'une OCRisation correcte (produisant un résultat supérieur à ce seuil). La méthodologie de cette pré-sélection est sujette à caution car elle ne fait appel qu'à des critères visuels et elle est effectuée de manière subjective par des bibliothécaires qui n'ont pas d'interaction ni de retour possible par rapport aux résultats de l'OCRisation future, et qui n'ont donc pas les moyens de juger de l'efficacité de leurs critères de sélection.
Dans ce contexte, la prédiction prendrait place en amont de la chaîne de traitement des documents, permettant de sélectionner, en fonction du taux de reconnaissance

[1] Un taux de reconnaissance de mots supérieur à 80% étant suffisant pour obtenir de bons résultats en termes de rappel et de précision, voir [Doe98]

estimé, les documents dont le traitement par un système d'OCR est économiquement rentable, ou du moins utile. Rappelons enfin qu'il existe d'autres applications de la prédiction, moins courantes mais ayant donné lieu à quelques travaux de recherche, concernent la sélection automatique du système de restauration d'images le plus approprié au document [12] ou du système d'OCR le plus performant sur le type de document détecté [13].

Dans la suite, nous présentons dans un premier temps les mesures de qualités couramment utilisées pour pouvoir prédire les résultats d'un OCR. Dans un second temps, nous montrons les résultats obtenus ainsi que les usages de ces mesures.

2. MESURES DE QUALITE POUR LA PREDICTION

Peu de travaux de recherche se sont intéressés à la prédiction des performances de l'OCR. Les travaux existants se sont principalement dirigés vers une caractérisation de la qualité des images. La qualité d'une image étant matérialisée par un ensemble de descripteurs définis dans une optique de quantification des sources de dégradation dont l'impact sur les performances des OCRs est universellement reconnu (cf. première partir du document). La Figure 8 illustre quelques unes de ces cas caractéristiques, au niveau image, qui réduisent les performances des OCRs.

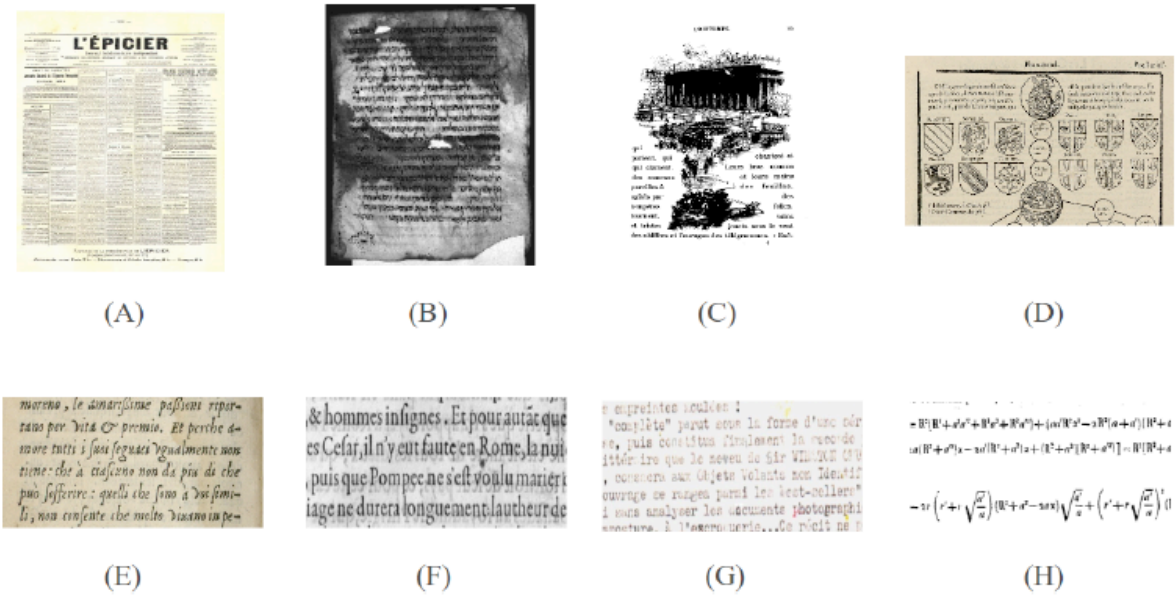


Figure 8 : quelques cas représentatifs d'images difficiles : A) mise en page complexe B) Support de mauvaise qualité C) et D) textes et graphismes enchevêtrés E) « style » difficile à traiter (italique et caractères inhabituels) F) apparition du verso par transparence G) impression de mauvaise qualité (ici caractères effacés et fragmentés) H) équations mathématiques

2.1. Qualité des caractères

Dans le premier de ces travaux [14], deux mesures principales de la qualité des images ont été définies. Ces mesures sont :

White Speckle Factor (WSF). WSF permet de quantifier le nombre de boucles de petite taille (composantes connexes "blanches" dont la largeur et la hauteur sont inférieures à 3 pixels, appelées white speckle dans l'article) présentes sur l'image (équation 2.1). Le nombre de ces boucles étant proportionnel à l'épaisseur de trait des caractères, il est, d'après les auteurs, censé être corrélé au taux de caractères fusionnés. Ces deux phénomènes (occlusion de boucles et fusions de caractères) ont une influence négative sur les performances des OCRs (figure 2.1).

$$WSF = \frac{\#\text{ CC blanches de taille inférieure à } 3 \times 3}{\#\text{ CC blanches}}$$

Figure 2.1: Occlusions de boucles sur des images de caractères (le premier e présente un white speckle. La boucle du second e présente une occlusion complète [14].

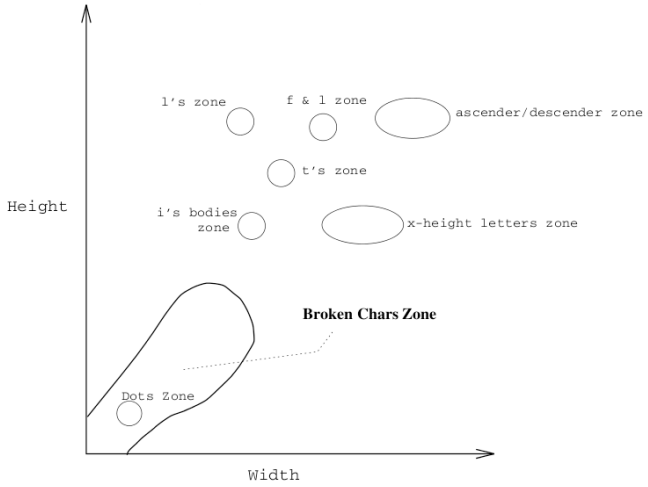


Figure 2.2: Répartition des CC dans le plan (hauteur, largeur). Les CC de caractères fragmentés sont confondus avec les points diacritiques dans une zone proche de l'origine [14].

Broken Character Factor (BCF). Une autre cause fréquente d'erreurs dans les systèmes d'OCR est la présence de caractères fragmentés. La mesure BCF a été définie afin de quantifier le taux de fragmentation des caractères (figure 2.2). Les différentes composantes connexes de l'image sont d'abord recensées et réparties selon leurs dimensions (largeur, hauteur) dans le plan défini par les axes (largeur, hauteur). Les CC de caractères fragmentés seront concentrées dans une zone proche de l'origine de ce plan. BCF mesure le pourcentage de CC présentes dans cette zone.

Dans l'article [15], les auteurs ajoutent aux deux mesures BCF et WSF de l'article précédent un ensemble de caractéristiques supplémentaires :

Vertical Brokenness Factor (VBF) :

$$VBF = \frac{\#\text{ CC blanches}}{\#\text{ CC noires}}$$

Stroke Thickness Factor (STF) : mesure l'épaisseur moyenne des traits, normalisée par la taille de la police (estimée ici par la hauteur des caractères) ;

Black Density Factor (BDF), calculé par le rapport :

$$BDF = \frac{\#\text{ CC épaisses}}{\#\text{ CC noires}},$$

une composante connexe (CC) épaisse étant une CC dont la densité est supérieure à 0, 75 ;

Touching Character Factor (TCF) :

$$TCF = \frac{\#\text{ caractères fusionnés potentiels}}{\#\text{ CC noires}},$$

un caractère fusionné potentiel étant une CC dans laquelle sont incluses deux ou plusieurs CC blanches. L'inclusion de plusieurs CC blanches, étant rare pour un caractère unique (B ou 8 par exemple), correspond avec une forte probabilité à une fusion de caractères.

Dans [13], un système initialement conçu pour l'identification du script utilisé sur le document a été réutilisé, après un apprentissage spécifique, pour déterminer le niveau de dégradation du document. Ce module d'identification du script extrait des caractéristiques décrivant les formes des glyphes présents dans le document. Ces caractéristiques sont extraites des images de mots (segmentés) et incluent notamment des moments Cartésiens centrés et normalisés, des moments de Hu, des caractéristiques statistiques (compacité, élongation, convexité...) et des caractéristiques de co-occurence [23].

2.2. Autres mesures de qualité

Dans [12], les auteurs ajoutent aux mesures précédentes deux nouvelles mesures : la taille de police estimée et le Small Speckle Factor (SSF) qui mesure la quantité de bruit (ou de tâches) présente sur le fond du document.

Dans [17] les auteurs montrent qu'une analyse de la nature du bruit présent sur les images de document peut permettre de prédire avec une certaine précision les performances d'un système d'OCR. Il est donc possible de supposer qu'une estimation des paramètres d'un modèle de dégradation connu, directement à partir des images [18], peut permettre de classer les différentes qualités (homogènes) d'images pour lesquelles il sera alors possible de prédire un taux de reconnaissance OCR. Ainsi, les descripteurs prédents ont été évalués dans l'article [21]. Dans cet article, les auteurs créent des images de mots dégradées par application du modèle de dégradation de Baird [22] (cf. 1.7.4.2). La pertinence de ces différentes mesures

est évaluée sur ces images par leur corrélation avec certains paramètres du modèle de dégradation. Les résultats montrent un certain nombre de relations fortes entre certaines métriques et quelques paramètres du modèle de Baird :

- STF est fortement corrélé au seuil de binarisation
- SSF est fortement corrélé au bruit
- TCF est assez fortement corrélé au seuil de binarisation
- WSF est fortement corrélé au bruit
  BCF a une corrélation relativement faible avec les paramètres du modèle de Baird.

Dans un article plus récent [24], les auteurs définissent un ensemble de 6 mesures permettant de quantifier l'intensité de la transparence du support (qui cause l'apparition du verso sur une image de document, produisant à la binarisation une quantité importante d'information bruitée) et son influence sur la qualité du document. Ces mesures quantifient l'intensité de la transparence, sa quantité, sa position relative par rapport aux caractères du recto.

3. PREDICTION DE PERFORMANCES ET EVALUATION DES MESURES

A notre connaissance, très peu d'études ont été faites sur la prédiction des performances des OCRs en se basant sur les méta-données de l'œuvre ou de l'ouvrage. L'étude effectuée dans [27] en est un exemple représentatif faite sur le corpus de la BnF à partir des dates d'édition, de la langue, du format du document, etc. Elle montre clairement, comme énoncé plus haut, que ces informations ont une importance et influent sur les résultats. Cependant l'intensité des relations entre les critères et les taux OCR reste faible et ne permet pas de faire de la prédiction efficace. C'est pourquoi les études se sont basées essentiellement sur les mesures visuelles issues directement du processus de numérisation, base sur laquelle fonctionne l'OCR.

Ainsi, les mesures de [14] sont utilisées pour classer des documents de la base DOE de l'ISRI en deux catégories : "bon" (TR > 90%) et "mauvais" (TR <= 90%) déterminées d'après la médiane TR des taux de reconnaissance obtenus par 6 systèmes différents. La classification est effectuée soit en appliquant un ensemble de règles sur les valeurs des mesures utilisées, soit en calculant la distance de Mahalanobis entre l'échantillon à classer et les moyennes des deux classes. Le système est paramétré de sorte à bien classer la plus grande proportion possible de "bons" documents, au détriment de la classification des mauvais. Les résultats obtenus montrent que cette méthode, dont les résultats sur les bons documents sont intéressants, est peu performante en moyenne (taux de classification correcte de 86%), principalement à cause du faible pouvoir discriminant des mesures utilisées et de leur manque de robustesse (sensibilité à la taille de police, par exemple).

La méthode utilisée dans [15] est locale, c'est à dire qu'elle estime les caractéristiques sur un ensemble de zones de l'image, le taux prédit étant le taux moyen sur l'ensemble de ces zones. La classification est effectuée par un réseau de neurones PMC selon le même protocole à deux classes (bon, mauvais) que la méthode de [14]. Cette méthode améliore nettement les résultats de la méthode précédente (elle classe correctement la quasi-totalité des bons documents) mais elle présente le même défaut qu'elle : la majorité des "mauvais" documents sont mal classés.

Dans [12], les mesures définies dans [14] sont utilisées afin de caractériser la qualité d'images de documents dans le but de sélectionner la méthode de restauration corrective la plus appropriée à chacune. Pour s'assurer de la pertinence des mesures utilisées, les auteurs ont calculé la corrélation entre ces mesures et les taux d'OCR observés. Le taux de corrélation obtenu est particulièrement élevé pour WSF et TCF et très faible pour la taille de police. Une méthode semblable est utilisée dans l'article [16] où un ensemble de 4 mesures (STF, TCF, SSF et taille de police) est appliqué à une tâche d'évaluation de la qualité des images destinée à sélectionner le filtre (destiné à rehausser la qualité des images) le plus approprié pour le traitement de chaque image.

Dans [17] les auteurs réalisent une partition de l'espace de dégradation (défini par les différents paramètres d'un modèle de dégradation [20]) en différentes zones dans le but d'améliorer l'apprentissage des OCRs en créant des systèmes spécifiques à chaque zone de l'espace de dégradation. On peut donc supposer qu'une approche de ce type est transposable à la tâche de prédiction des performances : les performances d'un système d'OCR dans une de ces zones étant connues, il est donc possible de prédire le taux de reconnaissance d'un document en fonction de la zone de l'espace de dégradation dans laquelle il est inclus.

Dans [13], un classifieur de type un k-nn est utilisé pour déterminer à quelle classe appartient une image. Trois classes ont été définies selon le type de dégradation : caractères fusionnés, fragmentés ou "propres". Ce système est appliqué pour diriger le document vers le système d'OCR le plus approprié aux images présentant le type de dégradation identifié. Ce système de détection de la dégradation donne de bons résultats dans la classification des caractères fusionnés et propres mais des résultats médiocres sur les caractères fragmentés.

Les mesures de [24] sont utilisées pour prédire les performances de deux systèmes d'OCR sur des documents anciens (souvent sujets à ce type de dégradation). La pertinence de ces mesures a été évaluée en appliquant une technique de régression linéaire [25]. Les résultats obtenus montrent une très forte corrélation entre ces mesures et les taux de reconnaissance des systèmes d'OCR (les coefficients de corrélation linéaires obtenus sont respectivement de 1.06 et de 0.99, très proches de la valeur optimale de 1).

Conclusion

Ce que nous constatons principalement à travers cette étude, c'est qu'il existe une grande variété de facteurs qui peuvent influencer les OCRs et dégrader leurs performances. La difficulté de la prise en compte de ces mesures provient donc à la fois de leurs diversités mais aussi de leurs interactions. Pour construire un système de prédiction fiable et robuste, il est donc indispensable de tenir compte, pendant le processus de prédiction, de toutes ces sources possibles de dégradation.

Références

[1] Apostolos Antonacopoulos, «The effect of scanning parameters on OCR quality», *Impact Final Conference*, 24-25 october 2011.

[2] George Nagy, Thomas A. Nartker, Stephen V.Rice, «Optical Character Recognition: An illustrated guide to the frontier », Procs. Document Recogntion and Retrieval VII, SPIE Vol. 3967, 58-69, 2000.

[3] Rémy Mullot, « Les documents écrits de la numérisation à l'indexation par le contenu», Lavoisier, 31 mai 2006, ISBN 2-7462-1143-2.

[4] Shalev Vayness, Catherine Lerouge, « Charte de Traitement "OCR brut et HQ, ALTO" » documentation interne de la Bibliothèque nationale de France, 25/02/2008.

[5] 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 , March/April 2009, vol. 15 no 3/4.

[6] Stephen V. Rice, Frank R. Jenkins, and Thomas A.Nartker, « The Third Annual Test of OCR Accuracy ». Annual Report of ISRI, 1994.

[7] Stephen V. Rice, Frank R. Jenkins, and Thomas A.Nartker, « The Fourth Annual Test of OCR Accuracy », Annual Report of ISRI, 1995.

[8] Stephen V. Rice, Frank R. Jenkins, and Thomas A.Nartker, « The Fifth Annual Test of OCR Accuracy », Annual Report of ISRI, 1996.

[9] Roger T. Hartley, Kathleen Crumpton, « Quality of OCR for degraded text images », DL '99 Proceedings of the fourth ACM conference on Digital libraries, 1999.

[10] Tapas Kanungo, « Document Degradation Models and a Methodology for degradation Model Validation», Thèse de doctorat, University of Washington, ????

[11] Marie-Elise Fréon, « Chaîne de numérisation Document préparatoire », Documentation interne de la Bibliothèque nationale de France, Service Numérisation.

[12] Cannon M., Hochberg J., Kelly P., « Quality assessment and restoration of typewritten document images », International Journal on Document Analysis and Recognition, vol. 2, p. 80-89, 1999.

[13] Ablavsky V., Pollak J., Snorrason M., Stevens M., « OCR Accuracy Prediction as a Script Identification Problem », in , D. Doermann (ed.), Proceedings of the 2003 Symposium on Document Image Understanding Technology, University of Maryland/UMIACS, 2003.

[14] Blando L. R., Kanai J., Nartker T. A., « Prediction of OCR accuracy using simple image features », Proceedings of the Third International Conference on Document Analysis and Re- cognition, ICDAR'95, p. 319-, 1995.

[15] Gonzalez J., Kanai J., Nartker T. A., « Prediction of OCR Accuracy Using a Neural Network », Proceedings of the International Workshop on Document Analysis Systems, p. 323-337, 1996.

[16] Souza A., Cheriet M., Naoi S., Suen C. Y., « Automatic Filter Selection Using Image Quality Assessment », Proceedings of the Seventh International Conference on Document Analysis and Recognition - Volume 1, ICDAR'03, p. 508-, 2003.

[17] Hartley R. T., Crumpton K., « Quality of OCR for degraded text images », Proceedings of the fourth ACM conference on Digital libraries, DL'99, p. 228-229, 1999.

[18] Yam H. S., Barney Smith E. H., « Estimating Degradation Model Parameters from Character Images », Proceedings of the Seventh International Conference on Document Analysis and Recognition, ICDAR'03, p. 710-, 2003.

[19] Smith E. H. B., Andersen T., « Text Degradations and OCR Training », Proceedings of the Eighth International Conference on Document Analysis and Recognition, ICDAR'05, p. 834-838, 2005.

[20] Barney Smith E. H., Andersen T., « Partitioning of the degradation space for OCR training », Proceedings of SPIE, the International Society for Optical Engineering, 2006.

[21] Reed D. K., Barney Smith E. H., « Correlating degradation models and image quality metrics », vol. 6815 of Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, January, 2008.

[22] Baird H. S., « Document image analysis », IEEE Computer Society Press, Los Alamitos, CA, USA, chapter Document image defect models, p. 315-325, 1995.

[23] Haralick R. M., Shanmugam K., Dinstein I., « Textural Features for Image Classification », IEEE Transactions on Systems, Man, and Cybernetics, vol. 3, n° 6, p. 610-621, November, 1973.

[24] Rabeux V., Journet N., Domenger J. P., « Ancient documents bleed-through evaluation and its application for predicting OCR error rates », vol. 7874, SPIE, p. 78740Q, 2011.

[25] Montgomery D. C., Peck E. A., Vining G. G., « Introduction to Linear Regression Analysis, Solutions Manual » (Wiley Series in Probability and Statistics), Wiley-Interscience, 2007.

[26] Garrison P., Davis D. L., Andersen T. L., Barney Smith E. H., « Study of style effects on OCR errors in the MEDLINE database », in , E. H. Barney Smith & K. Taghva (ed.), Proceedings of SPIE, the International Society for Optical Engineering, vol. 5676 of Society of Photo- Optical Instrumentation Engineers, p. 28-36, December, 2004.

[27] A. Ben Salah, G. Cron, T. Paquet and N. Ragot, "Prediction of Selection Decision of Document Using Bibliographic Data at the National Library of France (BnF)", IS&T's Archiving 2012 Conference, pp. 135-140, June 12-15, Denmark , 2012.