Génération d'images semi-synthétiques de documents anciens à des fins d'évaluation de performances et d'apprentissage

Van Cuong Kieu, Mehri Maroua, Vincent Rabeux, Nicholas Journet, Muriel Visani

Génération d'images semi-synthétiques de documents anciens à des fins d'évaluation de performances et d'apprentissage

Van cuong Kieu[*,] — Maroua Mehri[**] — Vincent Rabeux[*] — Nicholas Journet[*] — Muriel Visani[**]**

[*] *Laboratoire Bordelais de Recherche en Informatique - LaBRI, University of Bordeaux I, Bordeaux, France*
[**] *Laboratoire Informatique, Image et Interaction - L3i, University of La Rochelle, La Rochelle, France*

RÉSUMÉ. *Dans cet article, nous étudions comment des données semi-synthétiques permettent d'évaluer finement les performances d'algorithmes ou de fournir des données d'apprentissage à un système de traitement ou d'analyse d'images de documents. Les images semi-synthétiques que nous générons reproduisent fidèlement les défauts des documents anciens liés aux moyens d'impression anciens ou à la dégradation de l'encre des caractères. La première expérimentation réalisée dans cet article vise à comparer les performances de différents descripteurs texture dans l'optique d'une segmentation d'images. La seconde expérience met en évidence le fait que l'utilisation d'images semi-synthétiques permet d'enrichir quantitativement et qualitativement une base d'apprentissage utilisée par une méthode de prédiction de résultats de binarisation d'images de documents et d'améliorer les résultats de 15%.*

ABSTRACT. *In this article, we study the advantages of using semi-synthetic data for evaluating and re-training document image analysis systems. We focus on semi-synthetic data that reproduce defects commonly encountered in old document images having an impact on texts and graphics. First, semi-synthetic images are used to efficiently evaluate and compare the performances of three different texture-based segmentation approaches in an image segmentation system. Second, these images are added into the training set to improve about 15% of the accuracy of a binarisation prediction system.*

MOTS-CLÉS : *images semi-synthétique , évaluation de performances, ré-apprentissage, modèles de dégradation.*

KEYWORDS: *Semi-synthetic Document Image, Performance Evaluation, Re-training, Degradation Models.*

---

1. Introduction

Généralement, la constitution d'une base d'apprentissage ou d'une base de test nécessite la création d'une vérité terrain générée manuellement. De ce fait, il faut pouvoir être en mesure de fournir un volume conséquent et varié d'images avec sa vérité terrain associée. Souvent choisie, l'option de la création manuelle de telles bases se heurte rapidement au problème de la pénibilité de la tâche et par conséquent du faible volume de données générées. Dans les années 90, pour s'abstraire de ce problème, l'idée de créer des images de documents synthétiques a émergé. Il existe deux approches permettant de générer des images synthétiques de documents. La première, et la plus ancienne, est celle consistant à modifier une image réelle selon un modèle prédéfini (dans ce cas on parle plutôt de données semi-synthétiques). Il y a déjà près de 15 ans, une publication de référence (Kanungo *et al.*, 1993) détaillait un modèle permettant de dégrader des caractères et un autre simulant les déformations du papier. Ces modèles sont appliqués sur des images réelles et permettent de générer des documents semi-synthétiques utilisés dans le cadre de travaux sur la reconnaissance de caractères. Le second type d'approche est celui qui consiste à générer des images totalement synthétiques à partir d'une liste de spécifications précises données par l'utilisateur (Héroux *et al.*, 2007). Il est ainsi possible de définir une structure logique à respecter, des règles de formatage à appliquer à cette structure logique (feuille de style) ou encore la disposition des différents blocs (zone de texte, illustrations, ...) et enfin un ordre de lecture des différents éléments de contenu. Si plusieurs propositions ont été faites ces dernières années pour générer des images de documents synthétiques ou semi-synthétiques, peu de tests ont été réalisés. Plusieurs questions restent donc en suspens : est-il plus judicieux d'utiliser les données synthétiques ou semi-synthétiques seules ou en combinaison avec les données réelles ? Dans ce second scénario, quel volume de données synthétiques ou semi-synthétiques est-il utile d'intégrer ? A quel point peut-on dégrader une image originale pour générer des données synthétiques qui restent suffisamment réalistes pour supporter l'évaluation de performances ou apporter une amélioration de l'apprentissage ?

Dans le cadre de nos travaux présentés dans (Kieu *et al.*, 2012) et (Kieu *et al.*, 2013a), nous avons proposé deux modèles de dégradation d'images de documents. Après un état de l'art relatif à ce domaine de recherche, nous détaillerons comment nous sommes en mesure de pouvoir produire, à la carte, des dégradations sur des images réelles. Enfin, les deux dernières sections présenteront chacune une campagne de tests mettant en évidence l'utilité qu'il y a à intégrer ces images semi-synthétiques pour une évaluation de performances précise ou pour générer des données d'apprentissage variées.

2. État de l'art

Dès les premiers travaux sur la reconnaissance de caractères (Phillips, 1968), il a été identifié que la présence de défauts dans les images de documents impactait les performances d'algorithmes de traitement ou d'analyse d'images. De ce fait, une

partie significative des travaux relatifs à la génération d'images de documents synthétiques s'est intéressée à la modélisation des défauts les plus couramment observés. Les auteurs de (Baird, 2000), (Lins, 2009), et (Ardizzone *et al.*, 2009) catégorisent une partie de ces défauts. Certains défauts peuvent être dûs au document physique lui-même mais il est également possible qu'ils apparaissent (ou s'accentuent) lors de l'étape de numérisation. Ces auteurs les classent également selon leurs spécificités visuelles : bruit local (*e.g.* caractères), bruit global (*e.g.* illumination), ou bruit diffus (*e.g.* transparence).

Dans cet état de l'art, nous détaillons plusieurs modèles de dégradation et comment ils sont utilisés dans le cadre d'une évaluation de performances ou pour la génération de données d'apprentissage.

2.1. *Modèle de dégradation*

2.1.1. *Dégradation globale*

L'auteur de (Baird, 1990) présente un modèle de dégradation pouvant simuler des bruits globaux apparaissant lors du processus de numérisation d'une image (Figure 1-a). Ce modèle se compose de dix paramètres qui permettent de modéliser quatre types de défauts : la rotation, la mise à l'échelle, la translation et l'erreur d'échantillonnage des couleurs induite par l'acquisition utilisant certains scanners, en particulier les plus anciens.

Les auteurs de (Kanungo *et al.*, 1993) présentent un modèle de déformation globale qui permet de simuler la distorsion du papier et l'effet de la l'illumination sur la page apparaissant lors de la numérisation d'un ouvrage avec une reliure épaisse (Figure 1-b). Les auteurs de (Liang *et al.*, 2008) ont également modélisé ce défaut (Figure 1-d). Dans ce modèle, la forme de la page est considérée comme une surface courbe en 3D pouvant se déplier sur un plan 2D. Dans (Kieu *et al.*, 2013a), les auteurs proposent également un modèle de déformation d'une page de document. Ce modèle s'adapte particulièrement bien au contexte des documents anciens, puisqu'il reproduit les distorsions locales telles que les plis, les trous, les abrasions présentes dans les vieux documents (Figure 1-c).

2.1.2. *Dégradation locale*

Les auteurs de (Kanungo *et al.*, 1993) présentent un modèle de bruit local qui permet de dégrader le contour de caractères dans les images binaires (Figure 1-e). Ce modèle ajoute du bruit poivre et sel sur les contours des caractères. Une opération morphologique de fermeture est ensuite appliquée pour lisser le contour. Egalement dédié aux images binaires, le modèle "hard pencil noise" proposé par (Jian Zhai et Li, 2003) permet de reproduire les lignes blanches apparaissant près des formes (Figure 1-f). L'article (Kieu *et al.*, 2012) propose un modèle de bruit local capable de reproduire, en niveaux de gris, les défauts des caractères (encre qui bave, tâches sombres/claires proches des caractères Figure 2).

Figure 1. *Exemples de dégradations des modèles existants : (a) Bruit global (Baird, 1990), (b) Distorsion globale 2D (Kanungo* et al.*, 1993), (c) Distorsion global 3D (Kieu* et al.*, 2013a), (d) Distorsion globale 2D (Liang* et al.*, 2008), (e) Bruit local (Kanungo* et al.*, 1993), (f) Source d'image de (Jian Zhai et Li, 2003), (g) Modèle de transparence, (h) Source d'image de (Curtis* et al.*, 1997)*

2.1.3. *Dégradation diffuse*

L'auteur de (Moghaddam R.F., 2009) propose un modèle permettant de simuler l'apparition de l'encre du recto sur le verso (Figure 1-g). Ce modèle se base sur un processus de diffusion dont l'idée principale est d'exécuter itérativement des opérateurs de diffusion. Un opérateur représente un processus de diffusion d'encre de la source (verso) à la cible (recto) en niveau de gris. (Curtis *et al.*, 1997) ont proposé un modèle de diffusion d'encre simulant la diffusion d'un liquide tombant sur une surface plane (Figure 1-h).

2.2. *Bilan sur des expériences utilisant des données synthétiques*

La majorité des modèles présentés précédemment intègrent des paramètres permettant de générer de manière très fine une grande variété de bruits dans les images semi-synthétiques. La première utilité est donc de pouvoir évaluer très précisément un système d'analyse d'images de documents. Par exemple, les auteurs de (Jenkins et Kanai, 1994) utilisent des images synthétiques afin de tester un logiciel d'OCR. Ils concluent que les règles typographiques, le bruit présent, et la nature du texte agissent sur les performances des OCRs. Les auteurs de (Delalandre *et al.*, 2010) intègrent le modèle de bruit de (Kanungo *et al.*, 1993) et un modèle de distorsion en 2D (rotation, translation) dans un générateur de documents architecturaux synthétiques. Trois

bases de données synthétiques sont créées. La base de symboles se compose de 1600 images contenant 15046 symboles. La base d'images de documents architecturaux se compose de 1000 images contenant 28065 symboles. Et enfin, la base d'images de plans électroniques est composée de 1000 images et de ses 14100 symboles associés. Des images sont choisies aléatoirement à partir de ces bases pour tester un moteur de détection de symboles. La variabilité du bruit ajoutée dans les documents synthétiques agit significativement sur la performance de ce moteur. Concrètement, sa performance diminue quand le nombre d'images synthétiques augmente dans la base de test.

Dans la carde de la compétition ICDAR/GREC en 2013, les auteurs de (Visani *et al.*, 2013) ont généré une base de 6000 images semi-synthétiques de documents musicaux sur lesquels ils déforment le papier et génèrent des discontinuités entre les différentes formes composant ces partitions. Ils évaluent ainsi plusieurs méthodes ayant pour objectif de supprimer les lignes de portées musicales. De manière globale, la performance des méthodes testées est d'autant plus faible que le bruit dégradant les formes (Kieu *et al.*, 2012) et celui déformant la page (Kieu *et al.*, 2013a) sont importants. De manière individuelle, ces tests ont mis également en évidence certaines méthodes plus robustes à un bruit qu'à l'autre. Concrètement, sur neuf méthodes soumises, les performances ont chuté d'environ de 4% quand le niveau de bruit local augmente. Cette méthode est plus robuste sur des distorsions locales (pli, trou, petite courbure) que sur des distorsions globales (longue courbure). Dans le cas où nous combinons les deux défauts, les performances de ces méthodes chutent de 6% en moyenne.

Il est également possible d'utiliser des images de documents synthétiques pour enrichir une base d'apprentissage. L'ambition est d'améliorer cette étape en générant un nombre suffisant et varié d'images représentatives de la réalité. Par exemple, en s'appuyant sur un modèle capable de générer des chiffres manuscrits déformés, (Mori *et al.*, 2000) génèrent trois ensembles de 200, 500, et 1000 images semi-synthétiques de chiffres manuscrits qui peuvent être ajoutées dans la base d'apprentissage. Ils observent que la performance globale diminue du fait de la présence d'images générées se trouvant être non réalistes. Par conséquent, ces images "trop" synthétiques sont supprimées dans la base d'apprentissage. Ceci permet d'améliorer de 0.2% le taux de leur moteur de reconnaissance lorsque les tests sont réalisés sur une base composée d'images réelles.

Sur le même modèle, les auteurs de (Varga et Bunke, 2003) ont dégradé leur base de 5000 lignes de texte selon quatre niveaux. Ces quatre types d'images synthétiques sont introduits dans la base d'apprentissage d'un moteur de reconnaissance d'écritures manuscrites basé sur un modèle de Markov caché. Leur travail montre une amélioration moyenne de 0.5% du taux de reconnaissance. Une conclusion intéressante sur ces tests est que les images générées en appliquant un faible niveau de dégradation sont celles ayant permis d'améliorer le plus les performances globales.

Récemment, deux bases de tests d'images utilisées pour des compétitions de segmentation (SaintGall et Parzival) ont été dégradées afin d'enrichir la base d'apprentissage d'un moteur de reconnaissance de documents manuscrits historiques (Fischer

et al., 2013). Ce travail montre une amélioration de 1% du taux de reconnaissance quand on ajoute des images semi-synthétiques dans la base d'apprentissage.

En dépit de ces travaux, si l'on compare avec les tests réalisés avec des bases d'images réelles annotées manuellement, les tests utilisant des images synthétiques sont finalement assez rares. De plus, ils ne permettent pas toujours de comprendre si oui ou non il est intéressant d'utiliser des données semi-synthétiques. C'est pourquoi, dans cet article nous proposons d'étudier plus en détail l'intérêt de telles données. L'originalité de cet article est que nous avons proposé aux auteurs des méthodes (Rabeux *et al.*, 2013) et (Mehri *et al.*, 2013) de réaliser des tests avec nos images semi-synthétiques. Notre rôle a été par la suite d'analyser dans quelle proportion et selon quelle méthodologie il était possible d'améliorer ces deux applications.

3. Modèle de dégradation utilisé pour générer des images semi-synthétiques

Les tests réalisés dans cet article sont effectués sur des images de document semi-synthétiques générées à partir du modèle présenté dans (Kieu *et al.*, 2012). Ce modèle est capable de dégrader des régions de texte en reproduisant les bruits couramment observés dans les documents réels. Plus précisément, nous catégorisons les bruits que nous sommes en mesure de générer de trois manières différentes. La Figure 2.a-e regroupe des défauts issus d'images réelles alors que les exemples Figure 2.f-j sont des défauts synthétiquement ajoutés à une image réelle.

La première catégorie de défauts que notre modèle reproduit est celle regroupant les défauts assimilés aux taches sombres ou claires. Comme on peut le voir sur la Figure 2.a-b, elles sont généralement localisées à proximité (à l'intérieur ou à l'extérieur) de la bordure d'un caractère sans jamais le toucher. Le second type de défaut que nous pouvons générer se traduit par la présence de zones touchant le bord d'un caractère mais ne créant pas de discontinuité (si la tache est blanche) ou ne fusionnant pas deux caractères (si la tache est noire). Sur la Figure 2.c-d, on peut voir qu'un ensemble de pixels est venu s'ajouter à un caractère. Enfin, nous sommes en mesure de générer des discontinuités dans les caractères. Sur la Figure 2.e on voit clairement que notre modèle a permis de couper le trait du caractère "a" en deux.

Le processus permettant la génération de ces trois catégories de défauts est résumé par la Figure 3. Tout d'abord à l'étape Figure 3.A, un algorithme pseudo-aléatoire permet de placer les points où seront localisées les futures dégradations. Nous appelons ces points, des "points de dégradation". Lors de l'étape Figure 3.B une ellipse est dessinée avec des dimensions calculées en fonction de la position du point par rapport aux contours du caractère le plus proche et une direction déterminée en fonction du gradient local dans l'image au point de dégradation. Lors de l'étape Figure 3.C chaque pixel à l'intérieur d'une ellipse est modifié pour donner l'impression que cette zone est dégradée.

Dans (Kieu *et al.*, 2013b) nous avons présenté une évolution de ce processus de génération permettant à un utilisateur de fixer lui-même une quantité de dégradation à

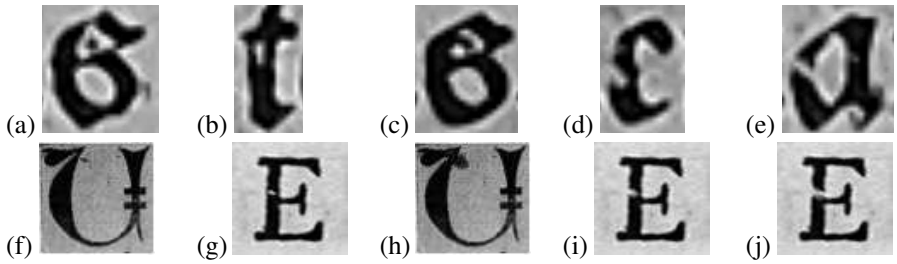

Figure 2. *Exemples de trois types du bruit dans les documents réels : (a)/(b) deux taches sombre/claires non connectées au bord d'un caractère ; (c)/(d) deux taches sombre/claires touchant un caractère ; (e) un caractère coupé. (f-j) défauts similaires générés synthétiquement*

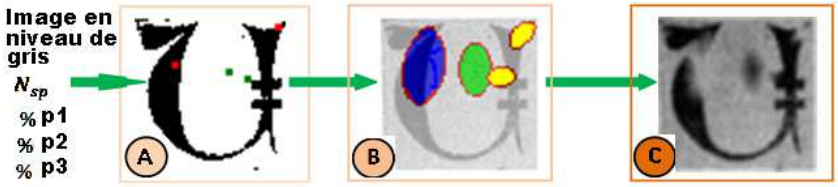

Figure 3. *Les trois étapes principales de notre modèle de dégradation : (A) sélection de point de dégradation, (B) Classification de point de dégradation, (C) génération de bruit local*

générer. Notre modèle permet également d'intégrer le souhait de l'utilisateur en termes de génération de types de défauts (cf Figure 2.f-i). La difficulté est ainsi d'attribuer au bon "point de dégradation", le type de défaut à générer. Une mauvaise affectation peut impliquer des résultats visuellement trop synthétiques. Par exemple, se servir d'un "point de dégradation" éloigné d'un caractère pour fabriquer un défaut de type Figure 2.c-d, a pour conséquence de générer une tache noire trop grosse. Nous proposons donc une heuristique pour éviter des taches trop grosses (non-réalistes).

L'heuristique est détaillée et justifiée dans (Kieu *et al.*, 2013b). Soit $p1$, $p2$, $p3$ la répartition souhaitée des trois défauts que nous sommes en mesure de générer ($p1$ + $p2$ + $p3$ = N_{sp}, où N_{sp} est le nombre total de "points de dégradation" souhaité par l'utilisateur) sur l'image réelle (Figure. 4-a), cette heuristique est la suivante :

1) Les "points de dégradation" sont calculés et classés par ordre croissant de distance à son plus proche caractère (Figure.4-b)

2) Les $p1$ premiers "points de dégradation" seront utilisés pour générer des dégradations déconnectant un caractère (Figure.4-c)

3) Les $p2$ points suivants seront utilisés pour générer des dégradations dont la caractéristique est d'être une tache (claire ou sombre) connectée à un caractère (Figure.4-d).

4) Les derniers points permettent de générer des petites taches à l'intérieur ou à l'extérieur d'un caractère (Figure.4-e).

(a) (b) (c) C_1 choisi (d) C_2 choisi (e) C_3, C_4 choisis

Figure 4. *Exemple illustrant les différentes étapes de notre heuristique affectant chaque "point de dégradation" à un type de dégradation à générer. (a) image originale, (b) sélection de 4 "points de dégradation", (c-d-e) dégradations successives de l'image*

Une fois l'affectation de chaque point à un type de défaut effectué, la direction et la taille de la région elliptique sont calculées en fonction du gradient à ce point, du type de défaut et de la distance au caractère le plus proche. Les pixels de l'ellipse sont dégradés en combinant un filtre moyenneur directionnel et un flou Gaussien. Les Figure 5.b, c, et d présentent trois images semi-synthétiques dégradées à partir de l'image Figure 5.a. Chacune contient seulement un type de bruit local.

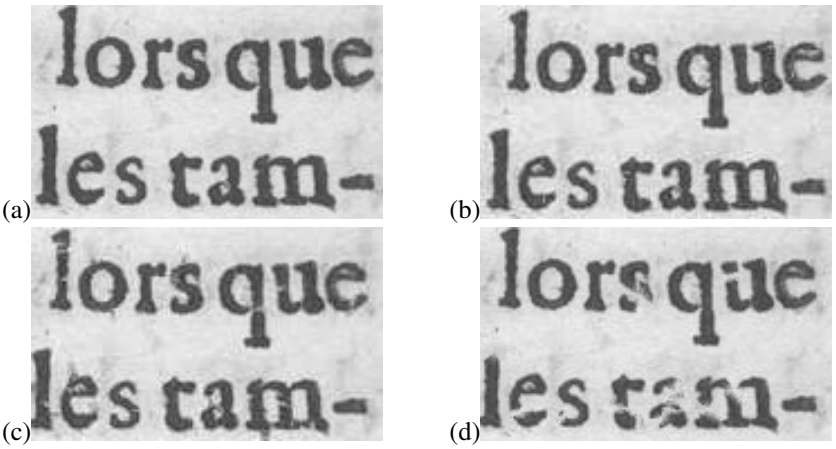

Figure 5. *Exemples de trois types de bruit. (a) image originale ; (b) des taches sombres/claires non connectées au bord d'un caractère, (c) des taches sombres/claires touchant des caractères, (d) des discontinuités*

4. Utilisation d'images de documents semi-synthétiques pour l'évaluation de performances

4.1. *Contexte*

L'objectif de cette section est de tester la robustesse du système de segmentation d'images présenté dans (Mehri *et al.*, 2013). En générant une grande variété d'images avec des caractères et zones de texte plus ou moins déformés, nous sommes en mesure d'identifier les cas de figure pour lesquels la méthode (Mehri *et al.*, 2013) est performante ou non.

La méthode présentée dans (Mehri *et al.*, 2013) se situe dans le contexte de la segmentation d'images de documents. L'hypothèse avancée par les auteurs est qu'il est possible de s'abstraire du manque d'information à disposition sur les images à analyser (mise en page, taille du texte, langue, ...) en utilisant une approche texture multi-résolution. Ainsi, sans connaissance *a priori* sur ces images, ils proposent d'appliquer un protocole en deux temps permettant de segmenter les pixels d'une image. Tout d'abord, en utilisant une fenêtre glissante de taille 4×4 pixels, un grand nombre d'indices texture est calculé pour chaque zone de l'image. Les indices texture calculés sont : l'autocorrelation, la matrice de co-occurrence et la réponse au filtre de Gabor. En répétant cette étape avec des fenêtres de tailles 4×4, 16×16 et 32×32 il est possible de capturer une information multi-résolution sur les textures présentes dans une image de document. A la fin de cette première étape, chaque pixel est décrit par un vecteur de caractéristiques qui est normalisé.

La seconde étape consiste à utiliser les informations calculées précédemment pour classer chaque pixel de l'image. Pour cela, un algorithme de "consensus clustering" est utilisé afin d'estimer le nombre de clusters présents dans un groupe d'images. Une fois ce nombre trouvé, une classification ascendante hiérarchique est appliquée. Elle permet d'affecter un label à chaque pixel des images. La figure 6 permet de montrer que cette méthode à tendance à regrouper des zones homogènes de pixels. Si le nombre de clusters est fixé à 3, alors on observe un regroupement des pixels en groupes texte/illustration/fond. Lorsque le nombre de clusters est augmenté, cette méthode à tendance soit à séparer le cluster de texte en fonction des fontes utilisées, soit de séparer les illustrations selon leurs caractéristiques visuelles.

Sur la base de tests réalisés avec un peu plus de 200 images sélectionnées dans la base Gallica (*http ://gallica.bnf.fr/*), les auteurs de (Mehri *et al.*, 2013) ont montré que leur méthode était robuste à la variété des contenus et des mises en page des documents anciens. Dans le contexte de l'analyse d'images de documents anciens les auteurs souhaitaient savoir également dans quelle mesure leur méthode était robuste aux bruits présents sur les zones de texte.

Figure 6. *Résultats illustrant la qualité de la segmentation d'images*

4.2. *Protocole expérimental*

Sur la base des images utilisées dans les expérimentations originales, nous avons généré 150 images semi-synthétiques en ajoutant divers niveaux de notre modèle de déformation de caractères. Nous avons décidé de diviser cet ensemble en 6 sous-ensembles. Les trois premiers (*Is*, *Os*, et *Ds*) correspondent à des sous-ensembles composés d'images dégradées selon un nombre fixe de "points de dégradation" et en générant seulement l'un des 3 bruits. Ce choix vise à évaluer, pour un nombre fixe de "points de dégradation", si l'algorithme de classification est sensible au type de défaut généré. Le choix du nombre de points à utiliser a été manuellement fixé pour générer des dégradations réalistes. Ainsi, chaque image d'*Is* a été dégradée en générant des taches sombres/claires non connectées à la bordure d'un caractère. Chaque image d'*Os* a été dégradée en générant des dégradations connectées à un caractère. Enfin, *Ds* contient uniquement des images avec des points de dégradation déconnectant un caractère. Les trois derniers sous-ensembles (*Ld*, *Md*, et *Hd*) tendent à évaluer la méthode, non plus en fonction du type de dégradation, mais uniquement de la quantité de dégradation présente. Ainsi, chacune des images est composée à part égale des trois types de "points de dégradation". D'un sous-ensemble à l'autre, nous faisons augmenter le nombre de "points de dégradation" et le paramètre jouant sur la taille de l'ellipse générée. Les sous-ensembles *Ld* (bas niveau), *Md* (moyen niveau), et *Hd* (haut niveau) sont générés respectivement avec 1000, 1500 et 2500 "points de dégradation" et des ellipses de taille croissante. La vérité terrain utilisée pour évaluer les performances est celle qui a été saisie manuellement à l'aide de l'environnement GEDI[1].

4.3. *Analyse des résultats*

La figure 7 permet d'illustrer le type de résultats obtenus sur la base dégradée. Ces résultats ont été obtenus avec le descripteur de texture Gabor, mais représente

1. http ://gedigroundtruth.sourceforge.net/

bien également les résultats obtenus avec les deux autres descripteurs. L'algorithme de classification des pixels est globalement robuste à n'importe quel type de dégradation. On observe une légère diminution de la qualité de la segmentation quand des caractères sont coupés en plusieurs fragments. La texture des caractères ainsi "cassés" est parfois assimilée à la texture d'une illustration Figure 7.c. Les tests réalisés sur les sous-ensembles *Ld*, *Md*, et *Hd* montrent par contre clairement que la méthode peine à discerner des textures de type texte ou caractère lorsque les dégradations sont trop importantes (Figure 7.e-f). De manière générale cela intervient quand les dégradations sont telles qu'elles déforment un caractère de près de la moitié de sa surface (suppression ou ajout de pixels).

(a) *Is* (b) *Os* (c) *Ds* (d) *Ld* (e) *Md* (f) *Hd*

Figure 7. *Résultats de classification de pixels sur des bases semi-synthétiques dégradées différemment. Les images correspondent à des zones extraites des images qui sont issues de : Is (a), Os (b), Ds (c), Ld (d), Md (e), et Hd (f).*

La figure 8 résume les résultats obtenus après une analyse quantitative réalisée sur l'ensemble des images de la base semi-synthétique. Nous avons décidé d'utiliser les mêmes métriques que celles utilisées par les auteurs de (Mehri *et al.*, 2013) pour les tests réalisés sur des images réelles. Pour chacun des 6 sous-ensembles d'images et les images réelles (non dégradées) nous comparons la vérité terrain et les résultats de la segmentation automatique sur la base de 5 métriques : Précision (P), rappel (R), Classification accuracy (CA), Silhouette Width (SW), et purity per block (PPB). Ces résultats confirment qu'en règle générale les trois descripteurs sont résistants à n'importe quel type de bruit local puisque les résultats sont quasiment au même niveau que ceux obtenus sur les images originales.

Les tests effectués sur les trois ensembles *Ld*, *Md*, et *Hd* confirment également que les résultats chutent dès lors que la dégradation est de plus en plus importante. En moyenne, les performances ont chuté de 1% entre la base originale et les images issues de *Ld* et *Md*. Les tests effectués sur les images les plus dégradées montrent que les performances chutent en moyenne de 4%.

5. Amélioration d'une étape d'apprentissage avec des images semi-synthétiques

Dans cette section, les documents semi-synthétiques sont utilisés pour enrichir une base d'apprentissage à l'origine composée uniquement de documents réels. L'ambition de cette section est de montrer qu'il est possible d'utiliser nos modèles pour générer de la vérité terrain et *in fine* d'améliorer l'apprentissage. L'algorithme testé

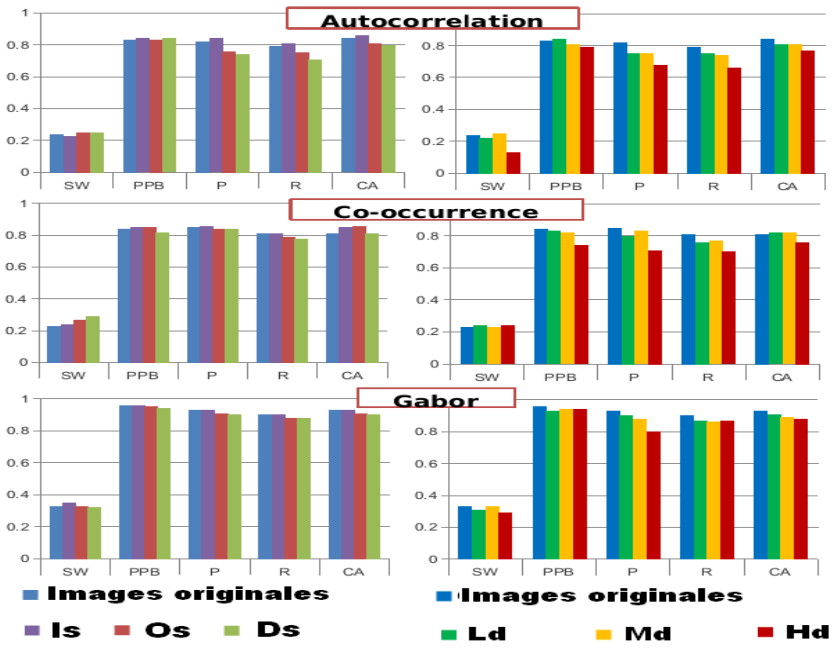

Figure 8. *Les résultats statistiques de trois descripteurs obtenus avec : deux mesures de clustering (Silhouette Width (SW), Purity Per Block (PPB)) et trois mesures de supervisée (Précision (P), Rappel (R), et le taux de classification (CA)).*

est présenté dans (Rabeux *et al.*, 2013). Il permet de prédire, pour n'importe quelle image de document, l'erreur que produirait l'application de 11 algorithmes différents de binarisation. De ce fait, l'algorithme prédictif détaillé dans (Rabeux *et al.*, 2013) permet de choisir la méthode de binarisation (parmi 11) la plus adaptée à un document donné.

La création d'une base d'apprentissage est un processus extrêmement long dans le contexte de l'évaluation de performances d'algorithmes de binarisation. En effet, la saisie de la vérité terrain ne peut être faite que manuellement et nécessite de devoir segmenter chaque pixel d'une image. Ainsi, la base DIBCO (Pratikakis *et al.*, 2011) qui est la base plus utilisée pour l'évaluation de performances d'algorithmes de binarisation contient moins de 50 images. Utiliser une base aussi peu fournie peut rendre difficile l'utilisation de méthode utilisant une validation statistique. Ainsi, pouvoir enrichir une base réelle annotée manuellement avec des images semi-synthétiques permettrait de palier à ce problème. C'est justement l'objectif de cette expérimentation. Nous chercherons plus particulièrement à évaluer le biais introduit par la quantité de données ajoutées à la base d'origine.

5.1. *Prédiction de taux d'erreur de binarisation*

L'erreur de binarisation, pour un document et un algorithme donné est très corrélée à l'état de dégradation de l'image. Les auteurs de (Rabeux *et al.*, 2013) créent pour chaque méthode de binarisation une fonction de prédiction $E_m = f_m(Q_I)$, avec la méthode de binarisation testée, E_m le taux d'erreur prédit, et Q_I un vecteur caractérisant la qualité de l'image. Q_I est mesuré sur la base du calcul de caractéristiques globales telles que la distribution de l'histogramme de niveaux de gris, la corrélation entre la moyenne des niveaux de gris des pixels considérés comme étant des pixels de dégradation et celle des pixels appartenant à la couche d'encre ou du fond. Q_I est également calculé à partir de caractéristiques mesurant la localisation et la forme des dégradations. Un ensemble d'apprentissage (et sa vérité terrain associée) est nécessaire pour générer chaque fonction f_m. Les 11 algorithmes testés sont Bernsen, Kittler, Li, Niblack, Ramesh, Ridler, Shanbag, Kapur, Otsu, Sauvola, et White. ils sont référencés dans (Rabeux *et al.*, 2013).

5.2. *Protocole de test et analyse des résultats*

Les images de documents semi-synthétiques sont générées à partir de 30% de la base originale (manuellement annotée) utilisée par (Rabeux *et al.*, 2013). Ces 30% d'images réelles ne sont ensuite plus utilisées dans la suite de nos tests. La vérité terrain d'une image réelle est utilisée comme vérité terrain de l'image semi-synthétique qu'elle a permis de générer. Soit T_s l'ensemble d'images semi-synthétiques utilisé pour l'apprentissage. Les 70% d'images réelles sont utilisés pour cet apprentissage selon le découpage suivant : l'ensemble T_o correspond à 50% de ces images sélectionnées aléatoirement et est utilisé pour compléter la base d'apprentissage ; l'ensemble V correspond aux 50% et est utilisé pour l'étape de validation statistique. Enfin, nous définissons T comme étant l'union de T_o et T_s.

Puisque le processus de sélection des images générant ces ensembles est aléatoire, les performances calculées lors de l'étape de validation statistique peuvent varier. Ainsi, afin d'obtenir une évaluation objective de l'intérêt qu'il y a à ajouter des images semi-synthétiques à une base composée uniquement d'images réelles, l'ensemble du processus est répété plusieurs fois. Afin de répondre également à la question du nombre adéquat d'images semi-synthétiques à ajouter, nous effectuons des tests où le nombre d'images ajoutées augmente. Enfin, l'ensemble de ce protocole est fait une seconde fois, mais cette fois-ci en utilisant uniquement un bruit classique poivre et sel pour la génération des images semi-synthétiques. Ceci nous permet de savoir si le modèle de bruit présenté par (Kieu *et al.*, 2013b) est pertinent et s'il permet de générer des images plus pertinentes qu'avec un modèle de dégradation classique.

Les résultats répertoriés dans la Figure 9-a montrent l'erreur moyenne de prédiction obtenue avec la méthode de binarisation de Kapur. Ils illustrent la tendance observée pour chaque modèle de prédiction. Ainsi, plus le nombre d'images semi-synthétiques ajouté à l'ensemble d'apprentissage est important et plus l'er-

reur moyenne diminue. On observe que ce taux d'erreur converge à partir de 25 images (soit 50% d'images semi-synthétiques par rapport à la base réelle initiale). En moyenne, l'utilisation d'images de documents semi-synthétiques a permis de diminuer de 15% le taux d'erreur de prédiction.

A titre de comparaison, les résultats obtenus avec des ensembles d'apprentissage composés d'images dégradées avec un simple bruit poivre et sel sont visibles dans la Figure 9-b. On observe clairement que le taux d'erreur ne converge pas spécialement vers une valeur plus faible. Par exemple pour 6 images synthétiques le taux d'erreur est équivalent au taux d'erreur obtenu avec 0 image synthétique. En moyenne les résultats obtenus avec notre modèle de dégradation sont meilleurs que ceux obtenus avec le bruit poivre et sel. Selon nous, cela vient du fait que nos déformations sont plus réalistes que celles obtenues avec le bruit poivre et sel. De ce fait, l'apprentissage correspond mieux aux types de dégradations réelles apparaissant dans les images utilisées pour l'étape de validation statistique.

Nous avons également réalisé un test de student afin de vérifier que le réapprentissage avec des images semi-synthétiques ajoutées à une base d'images réelles permet d'améliorer significativement la méthode de prédiction. Tout d'abord, la méthode de prédiction est entrainée avec une base de 14 images réelles et testée sur une base réelle de 36 images. En parallèle, la méthode de prédiction est entrainée avec une base de 14 images réelles et 12 images semi-synthétiques puis testée avec une base réelle de 36 images. Le résultat du test de student montre que notre hypothèse est acceptée (t-test(70) = 3.648, p-value < 0.01).

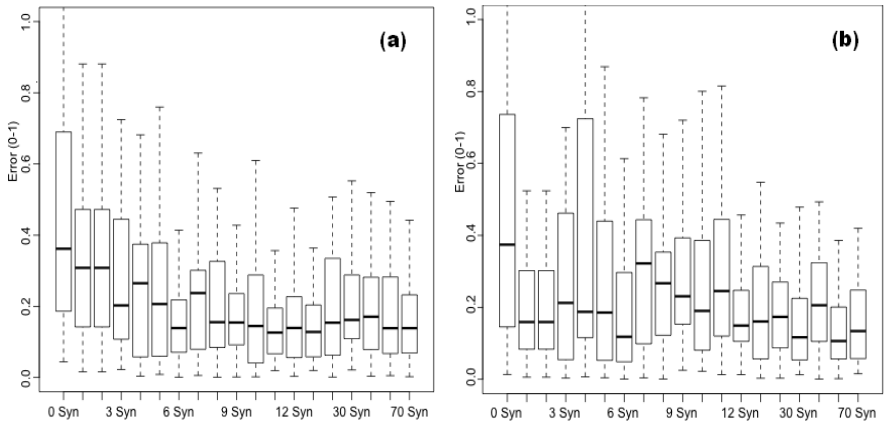

Figure 9. *Le taux d'erreur de la méthode de binarisation Kapur testé avec des images semi-synthétiques dégradées par (a) le modèle de bruit local et (b) le modèle de bruit poivre et sel*

6. Conclusion

Cet article présente deux applications différentes utilisant des images semi-synthétiques de documents anciens. Sur la base de notre modèle de dégradation nous avons été en mesure de mesurer l'intérêt de telles données dans le cadre de l'évaluation de performances et du ré-apprentissage. Le modèle appliqué permet d'altérer un document réel, d'y introduire des taches sombres ou claires, et de modifier la connexion des caractères ou des illustrations. La première expérimentation a permis de montrer qu'une méthode de segmentation d'images était robuste à la plupart des défauts présents dans les documents anciens. Les performances baissent uniquement lorsque la quantité de dégradation générée est significativement augmentée. La seconde expérience a permis de mettre en évidence l'intérêt qu'il y a à utiliser des images de documents semi-synthétiques lors d'une étape de ré-apprentissage. Nous avons en effet montré qu'introduire environ 50% d'images semi-synthétiques dans une base contenant des images réelles permet d'améliorer significativement une méthode prédisant les résultats de différents algorithmes de binarisation.

Remerciements

Ce travail s'inscrit dans la cadre du projet DIGIDOC financé par ANR (Agence Nationale de la Recherche française).

7. Bibliographie

Ardizzone E., Dindo H., Mazzola G., *Recent Advances in Signal Processing*, InTech, Italy, chapter Content-Based Image Retrieval as Validation for Defect Detection in Old Photos, p. 546-556, 2009.

Baird H. S., « Document Image Defect Models », *IAPR workshop on Syntatic and Structural Pattern Recognition*, Murray Hill, NJ, p. 13-15, Jun., 1990.

Baird H. S., « The State of the Art of Document Image Degradation Modeling », *In Proc. of 4 th DAS, Rio de Janeiro*, Rio de Janeiro, Brazil, p. 1-16, 2000.

Curtis C. J., Anderson S. E., Seims J. E., Fleischer K. W., Salesin D. H., « Computer-generated Watercolor », *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '97, New York, NY, USA, p. 421-430, 1997.

Delalandre M., Valveny E., Pridmore T., Karatzas D., « Generation of Synthetic Documents for Performance Evaluation of Symbol Recognition & Spotting Systems », *IJDAR*, vol. 13, n^{o} 3, p. 187-207, September, 2010.

Fischer A., Visani M., Kieu V. C., Suen C. Y., « Generation of Learning Samples for Historical Handwriting Recognition Using Image Degradation », *Proc. of the 2nd HIP*, p. 73-79, 2013.

Héroux P., Barbu E., Adam S., Trupin E., « Automatic Ground-truth Generation for Document Image Analysis and Understanding », *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, vol. 1, IEEE, p. 476-480, 2007.

Jenkins F., Kanai J., « Use of Synthesized Images to Evaluate the Performance of Optical Character Recognition Devices and Algorithms », *Proc. of SPIE, Document Recognition 1994*, vol. 2181, San Jose, CA, USA, 1994.

Jian Zhai Liu Wenyin D. D., Li Q., « A Line Drawings Degradation Model for Performance Characterization », *Proc. 7th ICDAR*, Edinburgh, Scotland, p. 1020-1024, August, 2003.

Kanungo T., Haralick R. M., Phillips I., « Global and Local Document Degradation Models », *Proc. of the ICDAR*, Tsukuba Science City, Japan, p. 730-734, Oct., 1993.

Kieu V., Journet N., Visani M., Mullot R., Domenger J. P., « Semi-synthetic Document Image Generation Using Texture Mapping on Scanned 3D Document Shapes », *Pro. of the 12th ICDAR, Accepted paper*, 2013a.

Kieu V., Visani M., Journet N., Domenger J. P., Mullot R., « A Character Degradation Model for Grayscale Ancient Document Images », *Proc. of the ICPR*, Tsukuba Science City, Japan, p. 685-688, Nov., 2012.

Kieu V., Visani M., Journet N., Mullot R., Domenger J. P., « An Efficient Parametrization of Character Degradation Model for Semi-synthetic Image Generation », *Pro. of the 2nd HIP*, Washington DC, USA, p. 29-35, 2013b.

Liang J., DeMenthon D., Doermann D. S., « Geometric Rectification of Camera-Captured Document Images », *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, n° 4, p. 591-605, 2008.

Lins R., « A Taxonomy for Noise in Images of Paper Documents - The Physical Noises », *Image Analysis and Recognition*, vol. 5627, Springer Berlin Heidelberg, p. 844-854, 2009.

Mehri M., Gomez-Krämer P., Héroux P., Boucher A., Mullot R., « Texture Feature Evaluation for Segmentation of Historical Document Images », *Pro. of the 2nd HIP*, Washington DC, USA, p. 102-109, 2013.

Moghaddam R.F. C. M., « Low Quality Document Image Modeling and Enhancement », *IJDAR*, vol. 11, Springer, Berlin, Heidelberg, p. 183-201, March, 2009.

Mori M., Suzuki A., Shio A., Ohtsuka S., « Generating New Samples from Handwritten Numerals Based on Point Correspondence », *Proc. 7th Int. Workshop on Frontiers in Handwriting Recognition*, Amsterdam, Netherlands, p. 281-290, 2000.

Phillips A., *Computer peripherals and typesetting*, Her majesty's stationery office, 1968.

Pratikakis I., Gatos B., Ntirogiannis K., « ICDAR 2011 Document Image Binarization Contest (DIBCO 2011) », *Pro. of the 20th ICDAR*, p. 1506-1510, 2011.

Rabeux V., Journet N., Domenger P., « Document Recto-verso Registration Using a Dynamic Time Warping Algorithm », *Proc. of the ICDAR*, Beijing, China, p. 1230-1234, Nov., 2011.

Rabeux V., Journet N., Vialard A., Domenger J.-P., « Quality Evaluation of Degraded Document Images for Binarization Result Prediction », *IJDAR*, vol. 1, p. 1-13, 2013.

Varga T., Bunke H., « Effects of Training Set Expansion in Handwriting Recognition Using Synthetic Data », *Proc. 11th Conf. of the Int. Graphonomics Society*, Scottsdale, AZ, USA, p. 200-203, Nov., 2003.

Visani M., Kieu V. C., Fornés A., Journet N., « Music Scores Competition : Staff Removal », *Pro. of the 12th ICDAR, Accepted paper*, Washington, DC, USA, 2013.