Délivrable DIGIDOC : Etat de l'art sur la génération de documents synthétiques et les méthodes de restauration d'images de documents anciens

Delalandre Mathieu, Journet Nicholas, Kieu Van Cuong,
Rabeux Vincent, Vialard Anne, Visani Muriel

28 mars 2012

1 Préambule

L'un des objectifs du projet DIGIDOC est de fournir un logiciel permettant de générer, à des fins d'évaluations de performances d'algorithmes de traitement ou d'analyse d'image, des images de documents anciens synthétiques avec leur vérité terrain associée. Il faut donc fournir une méthode permettant de créer de toutes pièces une image synthétique aussi similaire que possible que les images de documents anciens. Générer automatiquement une base d'images de documents synthétiques permet d'obtenir rapidement une grande variété d'images en jouant sur différents paramètres comme le type de fonte, les défauts présents ou la mise en page.

Cet objectif passe par la mise en place de deux processus distincts :

1. pouvoir "assembler" des éléments issus de documents anciens réels (illustrations, caractères, fonds, ...). Pour cela la principale difficulté est de trouver les bons paramètres permettant de reproduire la mise en page de documents anciens réels.
2. créer des modèles de dégradation qui seront appliqués aux documents synthétiques et permettant de reproduire les défauts les plus couramment rencontrés dans une campagne de numérisation (transparence de l'encre, déformation des caractères, courbures des feuilles, ...).

Ce délivrable est un état de l'art. Après une rapide présentation des méthodes existantes et permettant de générer des images synthétiques de

documents contemporains, nous nous focaliserons sur des méthodes de restauration d'images de documents qui, nous le pensons, peuvent être source d'inspiration pour la mise en place de modèles de dégradation.

2 Introduction sur la génération de document synthétique ou semi-synthétique

2.1 Introduction

Les travaux sur la génération d'images de documents ont d'abord porté sur la création de bases de données destinées à la validation et à l'apprentissage de systèmes de reconnaissance de caractères. Les auteurs de [16, 21] se sont intéressés à la génération d'images de texte dégradées. Etant donné un document idéal produit en utilisant LaTeX, la méthode proposée consiste à l'imprimer puis à le scanner pour en obtenir une version dégradée. Sa vérité terrain, c'est-à-dire le label et la boite englobante de chaque caractère, est obtenue en mettant en correspondance l'image idéale et l'image scannée. Les mêmes auteurs [18] ont ensuite proposé et validé un modèle de dégradation permettant d'obtenir une image de texte dégradée sans recourir à un processus physique. Ce modèle permet de contrôler l'inversion de pixels (texte / fond) et le niveau de flou.

Plus récemment, [9] ont proposé un logiciel permettant d'intégrer des spécifications lors de l'étape de génération de la vérité terrain d'images de documents contemporains. Il est ainsi possible de définir une structure logique à respecter (une DTD au format XML), des règles de formatage à appliquer à cette structure logique (feuille de style), la disposition des différents blocs proposés par la DTD (zone de texte, illustrations, ...) et enfin un ordre de lecture des différents éléments de contenu.

Dans le domaine des documents techniques, les images synthétiques sont utilisées pour évaluer la performance d'algorithmes de reconnaissance de symboles. Dans [1], des symboles tirés et retaillés aléatoirement sont positionnés dans l'image de façon à éviter les recouvrements. L'image obtenue est ensuite bruitée. Une approche similaire consistant à ajouter différents types de bruit à une image de dessin technique est proposée dans [38]. Plus récemment, [7] ont proposé une méthode de génération permettant d'obtenir un résultat plus réaliste en ajoutant des connaissances haut-niveau : les symboles sont placés sur un fond prédéfini suivant un ensemble de contraintes de position spécifiques à un domaine particulier comme l'architecture ou l'électronique.

2.2 Différentes méthodes d'acquisition ou de génération de vérité terrain document

Les travaux concernant la génération d'images ayant les caractéristiques de documents anciens sont peu nombreux. Dans [26], on trouve un modèle de dégradation permettant de simuler les effets de transparence (diffusion de l'encre du verso sur le recto). Ce modèle est ensuite utilisé dans un processus de restauration d'image. Par ailleurs, les auteurs de [32] proposent de générer des documents dégradés pour évaluer des algorithmes de binarisation. Les images dégradées sont obtenues en composant une image sans défaut (la vérité terrain) et une image de fond obtenue en scannant des pages blanches de documents du 18ème siècle. Les fonds comportent tous les défauts spécifiques aux documents anciens : variations d'intensité, taches, transparence du verso.

Une étape importante de toute démarche d'évaluation de performance est donc l'acquisition de la vérité terrain. Cette vérité terrain est mise en correspondance avec des résultats de reconnaissance d'un système donné par un algorithme de caractérisation de performance, afin d'en évaluer ses performances. Elle peut prendre différentes formes [24] mais correspond, à minima, à un résultat de reconnaissance idéal (i.e. 100 %). Dans la pratique, elle peut contenir des informations complémentaires dans des buts de caracterisation : mesures de dégradations, paramétres de transformation géométrique des objets observés, méta-données sur les documents, etc. Dans le cadre de la problématique du traitement d'images de documents à contenu textuel, différentes approches d'acquisition de vérité terrain ont été proposées dans la littérature. Ces approches ont été employées pour l'évaluation des systèmes OCR (i.e. codes ASCII des caractéres et coordonnées) mais aussi pour la segmentation de pages (structures physique et/ou logique du document). Le tableau 1 compare ces approches selon différents critéres : rapidité (temps nécessaire pour l'acquisition de la vérité terrain), robustesse (ou degré d'exactitude de la vérité terrain), type de document (réel ou synthétique), contrainte(s) de l'approche, information contenue dans la vérité terrain (niveau(x) structure et/ou OCR). Nous les détaillons dans les paragraphes suivants.

Acquisition par IHM : De nombreux travaux dans la littérature ont été proposés pour l'acquisition de vérité terrain par utilisation d'IHM [1] [37, 3, 25, 23, 29, 36, 30]. Une étude comparative des systèmes proposés est présentée dans [36]. Les IHM permettent l'édition de la vérité terrain de manière manuelle ou semi-automatique. Dans le cas semi-automatique,

1. Interface Homme-Machine

Approches	**Rapidité**	**Robustesse**	**Type**	**Contrainte(s)**	**Vérité terrain**
Acquisition par IHM	-	-	réel	aucune	structure & OCR
Transposition de documents électroniques	+	++	réel	versions électroniques	OCR
Transposition de transcriptions	+	+	réel	transcriptions	OCR
Transcriptions assistées	++	-/+	réel	aucune	OCR
Génération de documents synthétiques	+++	+++	synthétique	aucune	structure & OCR

TABLE 1 – Comparaison des approches

des algorithmes de reconnaissance de forme sont "pluggés" à l'IHM et les résultats proposés à la correction à un utilisateur [29]. Une autre manière de procéder est l'édition assistée, dans ce cas des d'algorithmes de segmentation sont pilotés en interaction utilisateur [30]. Les IHM peuvent être pourvues de différentes fonctionnalités selon les systèmes considérés : manipulation de zones, annotation sémantique, lecture mono ou multi-pages, etc. Un point important discuté dans ces différents travaux et le formalisme utilisé pour la représentation de la vérité terrain. Différents formalismes ont été proposés pour modéliser les structures physique et/ou logique des documents et y associer les informations graphiques de la couche OCR. Ce type d'approche est cependant dans la pratique peu exploitable. Le temps nécessaire à l'acquisition de la vérité terrain est trop important pour rendre l'approche applicable à des masses de données, et les bases constituées dans la littérature sont souvent "modestes" [3]. De plus, l'acquisition est majoritairement conduite par un utilisateur rendant l'approche fortement sujette aux erreurs.

Transposition de documents électroniques : De manière à automatiser l'acquisition de la vérité terrain, Kanungo a proposé dans [17] une approche permettant de transposer des documents électroniques avec leurs versions numérisées. Dans cette approche, les documents doivent Ítre initialement disponibles en version électroniques (i.e. codes ASCII et information graphique des caractères). Ces documents sont ensuite imprimés puis numérisés. Une méthode d'alignement est proposée, fonctionnant à partir de l'estimation de paramètres de transformation géométrique (décalage, rotation, mise à l'échelle) entre le document électronique et les résultats de segmentation de sa version numérisée. Cette estimation est calculée par une analyse en moindres carrés pondérés. Différentes améliorations de l'étape

d'alignement ont été proposées pour minimiser l'impact des fausses correspondances [11] ou affiner l'étape d'estimation de paramètres via des algorithmes de type "branch-and-bound" [22, 5]. Cette approche représente aujourd'hui une des meilleures possibilités pour l"acquisition de vérité terrain de manière quasi-automatique, à la fois robuste et rapide. Elle souffre cependant d'une contrainte forte : les documents doivent Ítre disponibles en version électronique. Ceci la rend donc inapplicable à des corpus anciens ou contemporains [2].

Transposition de transcriptions : Durant ces dernières années, différents systèmes ont été proposés dans la littérature afin d'adapter l'approche de Kanungo [17] aux documents anciens et contemporains[2]. Ces derniers sont discutés dans l'article [31]. L'idée de fond exploitée dans ces systèmes est l'utilisation de transcriptions de textes produites par des spécialistes afin de les transposer aux résultats de segmentation des documents numérisés. La démarche reste donc proche des systèmes présentés dans le paragraphe précédent, à la différence que les transcriptions n'incluent pas d'informations graphiques sur les caractères et peuvent contenir des erreurs de lecture et/ou d'interpretation, ou des omissions. Cette approche a majoritairement été appliquée aux textes manuscrits [31], cependant des travaux sur les documents typographiques existent [15]. Trois grands types de méthodes ont été expérimentées pour réaliser les alignements, basés sur les modèles de Markov cachés (MMC) [33], la déformation temporelle dynamique (DTW [3]) [15] et la programmation dynamique [14]. Cette approche reste cependant plus sujette au bruit que celle proposée par Kanungo [17] de part les erreurs possibles contenues dans les transcriptions et l'absence d'information graphique. De plus, les transcriptions restent difficiles et coûteuses à produire rendant l'approche difficilement applicable aux masses de données.

Transcriptions assistées : Pour générer la vérité terrain, les approches présentées précédemment reposent soit sur une forte interaction utilisateur soit sur une information connue a priori (documents électroniques, transcriptions). Différents travaux annexes ont tenté de lever ces limitations pour automatiser l'acquisition de la vérité terrain [12, 6, 13, 4]. Ces systèmes reposent sur un principe de transcription assistée par l'utilisateur. Les caractères sont segmentés en bloc à partir d'un ouvrage complet numérisé, une chaîne de "matching/clustering" est alors utilisée pour regrouper les caractères par degré de similarité. L'utilisateur interagit alors via le système pour labelliser en bloc les patterns redondants. Les systèmes diffèrent sur

2. Hors fin XX et début XXI siècle.
3. Dynamic time warping

les techniques de “matching/clustering” mises en oeuvre, mais aussi sur les problématiques de bouclage (affinage de l'appariement, prise en compte des cassures, etc.). La pertinence de cette approche reste cependant à démontrer pour une problématique d'acquisition de vérité terrain. Les résultats présentés dans la littérature [13] montrent que ces systèmes supplantent ceux OCR, sans pour autant garantir une transcription sans erreur des documents.

Generation de documents synthétiques : La dernière approche rencontrée dans la littérature est la génération de documents synthétiques [10, 8, 41]. Au sein de ces systèmes, les documents sont générés automatiquement à partir de modèles de mise en page et de fonte. La vérité terrain et les images de test deviennent automatiquement disponibles, par extraction des informations contenues dans les documents électroniques et leur rasterization. La procédure de génération reste, sur le fond, une étape principalement technique. Une des difficultés rencontrées sur cette approche est l'apprentissage des modèles de composition métiers pour la génération des documents. Cet aspect est cependant peu discuté dans la littérature.

3 Modèle de dégradation de caractères dans des images binaires

Cette section présente un résumé d'une méthode dégradation de caractères faisant référence dans la communauté document [19]. La problématique générale de cette méthode est la création de vérité terrain pour l'évaluation d'algorithmes d'OCR. La formalisation d'un modèle de dégradation ainsi que la procédure de validation de ce modèle sont réalisées dans un cadre relativement restreint :

- Les images de documents auxquelles s'intéresse cet article sont des images de texte en noir et blanc.
- Les dégradations qui peuvent être simulées sont exclusivement des défauts liés à la reproduction des documents : impression, photocopie, scan.

3.1 Modèle de dégradation

La première partie de l'article est consacrée à la définition d'un modèle de dégradation qui simule l'inversion de certains pixels ainsi que la zone de flou qui apparait autour d'un point lorsqu'un document est scanné. Ce modèle permet d'obtenir une image de document dégradé à partir de l'image d'un document idéal.

Le processus de dégradation utilise 7 paramètres $(\eta, \alpha_0, \alpha, \beta_0, \beta, k)$. Il suit les étapes suivantes :

1. calcul de la distance d de chaque pixel à la frontière des caractères,
2. inversion des pixels de texte avec la probabilité $\alpha_0 e^{-\alpha d^2} + \eta$,
3. inversion des pixels de fond avec la probabilité $\beta_0 e^{-\beta d^2} + \eta$,
4. fermeture morphologique avec un élément structurant de diamètre k.

Un pixel a d'autant plus de chances d'être inversé qu'il est proche du bord d'un caractère. Ceci permet de créer des irrégularités sur les contours des caractères. La corrélation entre points voisins est simulée par l'étape de fermeture.

FIGURE 1 – Dégradation du caractère **e**. Les paramètres utilisés sont $(\eta = 0, \alpha_0 = 1, \alpha = 1, \beta_0 = 1, \beta = 1, k = 2)$.

3.2 Validation statistique (nonparametric permutation test)

La deuxième partie de l'article propose une façon d'évaluer la représentativité des images synthétiques créées grâce au modèle de dégradation présenté ci-dessus. En d'autres termes, il s'agit de valider le processus de dégradation.

La validation s'appuie sur un jeu de données comportant des données réelles et des données synthétiques. Plus précisément, il faut rassembler N images réelles correspondant à des versions dégradées du même caractère. Ces échantillons réels seront notés $(x_i)_{i=1..N}$. Il faut également appliquer le processus de dégradation à une image idéale du caractère choisi pour en construire M versions dégradées. Ces échantillons synthétiques seront notés $(y_i)_{i=1..M}$.

On choisit une fonction de distance sur les caractères, par exemple la distance de Hamming et une fonction de distance entre ensembles de caractères, par exemple la distance MNN (Mean Nearest-Neighbor).

La validation statistique s'appuie sur l'hypothèse nulle suivante : les distributions des populations dont sont issus les échantillons (x_i) et (y_i) sont les mêmes. Pour décider si cette hypothèse peut être rejetée ou pas, les auteurs utilisent une procédure basée sur le calcul d'une distribution de distances entre ensembles créés en mélangeant de façon aléatoire les échantillons réels et synthétiques.

Cependant, si le nombre d'échantillons est suffisamment grand, l'hypothèse nulle sera toujours rejetée ce qui conduit à la conclusion que les images synthétiques "imitent mal" les images dégradées réelles. La procédure de validation sera donc plutôt utilisée pour comparer deux modèles de dégradation, c'est-à-dire pour trouver lequel simule le mieux une dégradation réelle.

L'intérêt principal de cet article réside dans la formalisation statistique de la validation d'un modèle de dégradation. A notre connaissance, cette méthode n'a pas été appliquée à d'autres modèles de dégradation. On peut avancer deux explications à cela. Premièrement, le processus de dégradation proposé est local ce qui permet de caractériser la dégradation au niveau d'un caractère. D'autres types de dégradation ne pourront pas être ainsi définies localement. Deuxièmement les images traitées sont exclusivement des images binaires, ce qui facilite la définition d'une distance entre deux images de caractères, mais qui restreint les cas d'utilisation.

4 Modèle de restauration de caractères anciens

4.1 Introduction

Les travaux présentés dans [2] portent sur la reconstruction de caractères dégradés dans les images de documents anciens [4]. Deux nouvelles techniques basées conjointement sur l'utilisation des contours actifs et d'une information structurelle est présentée. D'une part, une information relative à l'image est calculée. Elle provient de la construction d'un graphe structurel du caractère et permet, sans connaissance *a priori* de caractériser les parties dégradées d'un caractère. D'autre part, une autre approche consiste au contraire à utiliser une image référence « idéale » d'un caractère pour restaurer une instance dégradée.

4.2 Rappel sur les contours actifs

Les contours actifs (snakes) introduits par Kass, Witkin et Terwopoulos dans [20], sont des courbes continues (fermées ou non). Ils sont couramment

4. Toutes les images de cette section son extraites de [2]

utilisés pour la segmentation ou la localisation d'objets dans les images, le suivi d'objets en mouvement,... Le processus d'analyse d'un contour actif se résume en l'initialisation d'un contour (Figure 3) qui se déforme en fonction d'une énergie globale calculée sur l'image ou sur la forme même de la courbe.

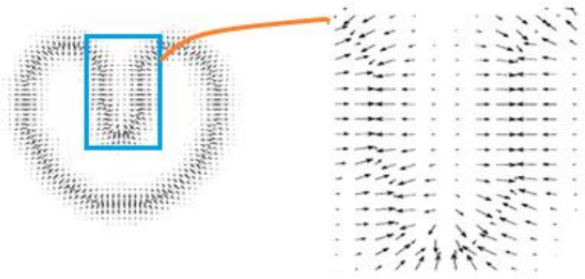


FIGURE 2 – *La force interne et externe du caractère "u"*

Soit une courbe C fonction du temps est le contour actif :

$$C = \{v(s,t) = [x(s,t), y(s,t)] \ \textit{où} \ s \in [a,b] \ \textit{et} \ t \in [0,T]\} \tag{1}$$

L'évolution du contour actif dans le temps est obtenue par minimisation de l'énergie potentielle totale E définie dans l'équation (2). Ces énergies sont divisées en deux : l'énergie interne et externe. Les deux énergies sont liées par l'équation (3) (Équation de Euler-Lagrage). L'énergie interne permet de conserver la forme de l'objet alors que la force externe a tendance à modifier la courbure. La figure 2 symbolise les directions de ces forces calculées sur une image de caractère.

$$E(C) = E_{interne}(C) + E_{externe}(C) \tag{2}$$

$$F_{interne} + F_{externe} = 0 \tag{3}$$

La force externe $F_{externe}$ qui influe directement sur le comportement du

FIGURE 3 – *Le contour initial*

snake est définie par l'utilisateur. Pour le contour actif classique, la force externe $F_{externe} = -\bigtriangledown E_{externe}$. $\bigtriangledown$ est liée au gradient de l'iamge. Le changement des énergies impacte sur la transformation de la courbe. Les méthodes de l'état de l'art se heurtent généralement à deux problèmes : la définition du contour initial et la présence de concavité dans la forme à segmenter (cf. Figure 4). Xu et al dans [35] ont ré-définit la force externe en utilisant le vecteur de densité qui s'appelle *le flux de vecteurs gradients GVF*. La direction de la force est modifiée à partir de ces informations et permet de corriger les défauts liés à la présence de concavités.

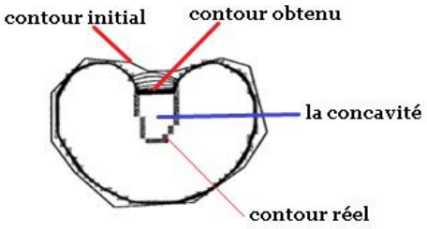


FIGURE 4 – *L'influence de la concavité sur le contour obtenu du caractère "u"*

4.3 Reconstruction de caractères dégradés par utilisation de contours actifs

La méthode de contour actif GVF est utilisée pour segmenter les images de caractères dégradés. Toute la difficulté est de pouvoir trouver la position de la partie dégradée d'un caractère pour ensuite le reconstruire.

Pour résoudre le problème de localisation d'une zone dégradée dans un caractère, l'auteur a proposé une technique utilisant la construction d'un graphe structurel de la forme *idéale*. Ce processus s'articule sur 3 étapes. D'abord, le graphe structurel (Figure 6-b) est calculé a partir d'une image *idéale* fournie par l'utilisateur (Figure 6-a). Le graphe structurel (6-b) et l'image dégradée sont ensuite combinés (Figure 6-c) dans le but de localiser la position exacte de la zone dégradée. Cette position est définie après comparaison des épaisseurs des traits.

Après avoir identifié la zone dégradée, les auteurs proposent deux méthodes pour reconstruire cette zone : l'utilisation de forces d'attraction ponctuelles et l'utilisation d'une image *idéale*. La première méthode consiste à ajouter une nouvelle force positionnée dans la cassure d'un caractère qui s'appelle *la force d'attraction ponctuelle* définie dans l'équation (4). Cette

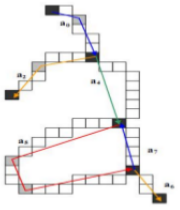


FIGURE 5 – *Exemple du graphe structurel*

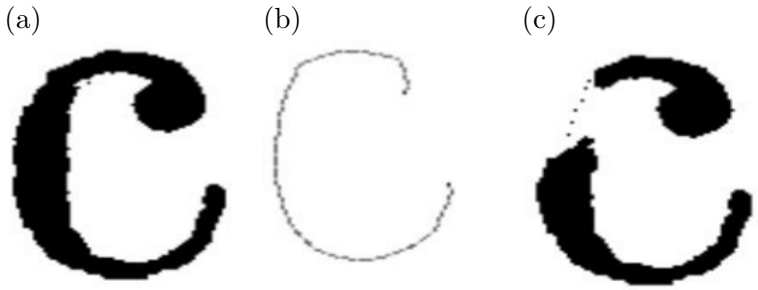


FIGURE 6 – *Extraction de graphe structurel : (a)-caractère de référence, (b)-le graphe structurel, (c)-l'image combinée*

force nécessite une adaptation utilisant le changement du rayon d'action (Figure 7). Cette force permet donc d'attirer le snake à une distance r d'un point P donné. La deuxième méthode utilise une image de référence choisie par utilisateur. Les forces dans l'image de référence sont calculées et comparées avec l'image du caractère pour trouver la zone de force de la concavité (un rectangle). Le rectangle englobant la dégradation est rempli par le rectangle correspondant dans l'image de référence.

$$F_{attraction} = -k_\alpha||P(s,t)-C(s,t)|-r^2|^2 \text{où le centre } P \text{ de la force avec rayon } r \tag{4}$$

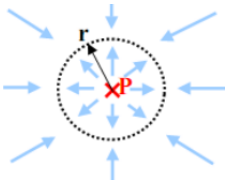


FIGURE 7 – *Le champ de forces de centre P et de rayon r*

Détails sur les 2 processus de reconstruction manuel proposés :

– Utilisation de forces d'attractions ponctuelles : sur l'image GVF, la position du centre de la force est positionnée par un utilisateur sur la zone dégradée (Figure 8-c).

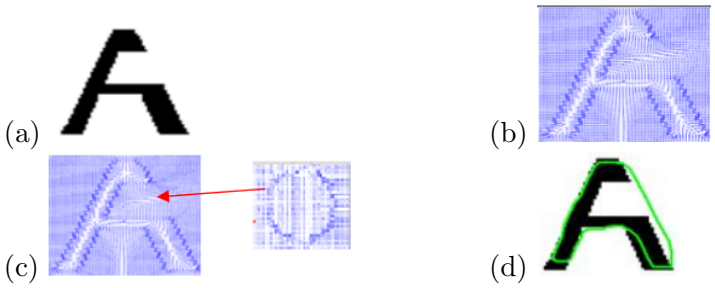


FIGURE 8 – *Reconstruction des caractères manuellement : (a)-image dégradée, (b)-le champ GVF dégradé, (c)-ajout manuel de la force d'attraction ponctuelle sur position du défaut, (d)-contour corrigé du caractère*

– Utilisation d'une image *idéale* comme référence : l'utilisateur montre la zone dégradée dans l'image d'origine (Figure 9-b). Sur cette zone, le champ GVF est remplacé par une énergie basée contours sur cette zone. Le processus est illustré dans la figure (9-c).

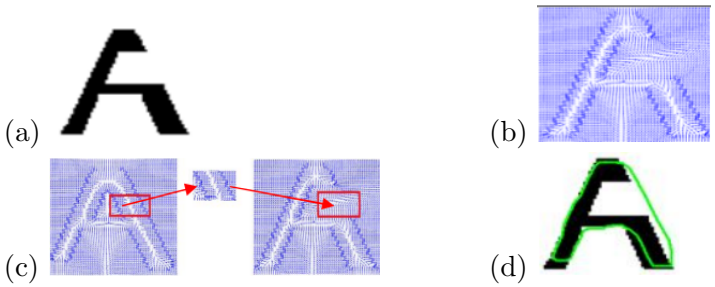


FIGURE 9 – *Reconstruction manuelle des caractères : (a)-image dégradée, (b)-le champ GVF dégradé, (c)-ajout manuel du rectangle de force extrait à la position du défaut, (d)-contour corrigé*

Détails sur le processus de reconstruction automatique des caractères utilisant le graphe structurel :

– Utilisation de forces d'attraction ponctuelles : l'image dégradée est combinée avec leur graphe structurel (figure 10-a) pour localiser automatiquement la zone dégradée. La force d'attraction ponctuelle est ajoutée par la suite à cette zone (figure 10-b) par une intégration dans

la formulation du contour actif.

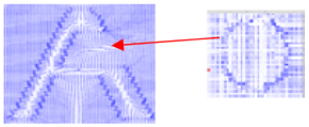
(a)  (b) 

FIGURE 10 – *Reconstruction des caractères automatiquement : (a)-image dégradée combinant leur graphe structurel, (b)-ajoutant le rectangle de force d'attraction ponctuelle à position de défaut*

– La force référencée extraite à partir de l'image *idéale* est ajoutée à cette zone (figure 11-b).

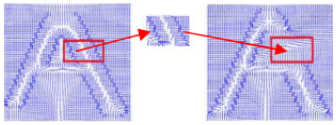
(a)  (b) 

FIGURE 11 – *Reconstruction des caractères automatiquement : (a)-image dégradée combinant leur graphe structurel,(b)-ajoutant manuellement le rectangle de force extraite à position de défaut*

(a)  (b) 

FIGURE 12 – *Exemple de caractères dégradés*

Les méthodes proposées fonctionnent relativement bien dans le cas d'apparition des trous dans les caractères. C'est un défaut couramment rencontré dans les images de documents anciens(figure 12). Nous pensons que cette technique de restauration pourrait tout à fait être adaptée dans l'optique de la mise en place d'un modèle de dégradation de caractères. Une étude précise du champ GVF d'un caractère sans défaut pourrait permettre de définir des zones où il faudrait faire disparaitre l'encre.

5 Modélisation de document anciens par équation aux dérivées partielles [27]

Afin de résoudre le problème de la transparence des documents anciens, la méthode présentée modélise plusieurs couches de dégradations présentes dans les documents. Ces couches des dégradations correspondent par exemple aux vieillissements du document, à la transparence recto verso, à l'ajout de taches, ...Cette méthode de modélisation est basée sur une diffusion des couches de dégradations vers une couche de destination, le recto. Dans cette section nous expliquerons le principe général de la méthode pour ensuite nous concentrer sur un cas particulier présenté par les auteurs : la génération de la transparence.

5.1 Problèmatique

Les dégradations des documents peuvent provenir de deux sources : du processus d'impression du fichier original ou de phénomènes physiques présents sur l'original. La plupart des dégradations résultantes d'un processus physique sont le résultat d'une sorte de diffusion sur une période de temps donnée.

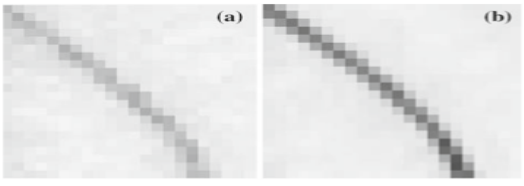


FIGURE 13 – Transparence des documents anciens : a. ligne d'encre visible par transparence depuis le recto. b. la même ligne d'encre vue depuis le verso. Image issue de [27].

Ces genres de dégradations sont omniprésentes sur les documents anciens et sur les documents de mauvaise qualité. L'état de l'art propose une méthode[40, 42] de génération de ce type de dégradations. Cette méthode modélise les dégradations avec un effet de flou et des opérateurs de transformation linéaire. Pourtant, le

processus physique de dégradation est de par sa nature non linéaire. Prenons par exemple le cas de la transparence dont un exemple est présenté en figure 13. La nature non linéaire de la transparence est visible sur cette image. En effet, certains pixels sont plus ou moins diffusés dans le papier.

La transparence de l'encre à travers un support papier est un phénomène complexe à modéliser. Beaucoup de paramètres sont à prendre en compte comme l'épaisseur d'une ligne d'encre, les caractéristiques du papier, la distribution spatiale des fibres de papier, la qualité de l'encre et ses caractéristiques de fluides. Du point de vue de la mécanique de fluides, la transparence est un flux de liquide qui se propage dans un support poreux. À une échelle microscopique, presque tous les phénomènes liés à la propagation d'un liquide dans un tel support peuvent être modélisés [39]. Malheureusement, à ce niveau d'abstraction, le temps de calcul requis rend les implémentations pratiquement impossibles. Par conséquent, un modèle équivalent, mais au niveau macroscopique est nécessaire. Ce modèle peut se baser sur la diffusion anisotropique.

5.2 Modélisation des dégradations par diffusion de couches d'informations

Un document ancien peut donc être modélisé comme la diffusion de plusieurs couches (figure 14). Ces couches sont soit des couches saines (encre du recto, encre du verso) ou des couches de dégradations (fond dégradé). Nous définissons l'opérateur $DIFF(u, s, c)$ qui représente la diffusion d'une couche source s, vers une couche de destination u avec le coefficient c. Par suite, la totalité de la méthode peut être régie par l'équation 5.

$$\frac{\partial u}{\partial t} = \sum_{i \in sources} DIFF(u, s_i, c_i) \tag{5}$$

Avec :

- u, l'image de destination,
- $sources$, l'ensemble des images à diffuser (verso, fond, ...),
- s_i, une image source,
- c_i, le coefficient de diffusion adapté à l'image s_i.

Chaque processus de diffusion permet de modéliser un type de dégradation. Par exemple, la diffusion anormale de l'encre dans le papier, les pertur-

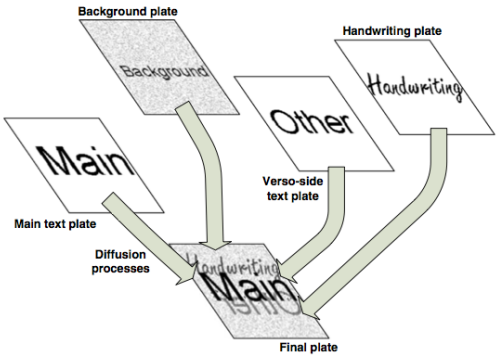


FIGURE 14 – Schéma général de la méthode : un document peut être modélisé comme la diffusion de plusieurs couches d'information (recto, verso, dégradations) vers une image de destination (le document final). Image issue de [27].

bations environnementales affectant le document, la transparence résultant de la diffusion du verso sur le recto, ... Chacune de ces diffusions prend une image source contenant les informations à diffuser, et un coefficient. C'est ce coefficient qui contrôle la diffusion et permet d'obtenir la dégradation voulue.

Prenons l'exemple de la modélisation de la transparence. Un document ancien transparent peut être vu comme la suite de plusieurs diffusions comme le montre l'équation 8.

$$\begin{aligned} \frac{\partial u}{\partial t} &= DIFF(u, s_{recto}, c_{recto}) && (6)\\ &\quad + DIFF(u, s_{fond}, c_{fond}) && (7)\\ &\quad + DIFF(u, s_{verso}, c_{verso}) && (8) \end{aligned}$$

Avec :

- u, l'image de destination (le document synthétique résultant du prcocessus),
- s_{recto}, l'image du recto sans dégradation,
- c_{recto}, le coefficient de diffusion de l'encre au recto vers l'image de destination,

- s_{fond}, l'image de fond qui possède éventuellement des dégradations (taches),
- c_{fond}, le coefficient de diffusion du fond vers l'image de destination,
- s_{verso}, l'image du verso dans dégradations,
- c_{verso}, le coefficient de diffusion du verso vers l'image de destination.

Ainsi, la spécialisation du type de dégradation est implémentée dans les coefficients de diffusion. Par suite, le coefficient c_{recto} correspond au coefficient de diffusion *classique* [28] : $c_{recto} = \frac{1}{1+(\nabla u/\sigma)^2}$. Le coefficient c_{fond} (équation 9) est quant à lui plus intéressant : il doit s'assurer que le texte ne soit pas modifié pendant sa diffusion sur le fond.

$$c_{fond} = d_{fond}(1 + tanh(u - s_{fond} - \delta_{fond}/\sigma_{fond})) \quad (9)$$

En effet, pour les pixels de texte, la différence entre u et s_{fond} est grande et la diffusion doit s'arrêter. Les paramètres δ_{fond}, σ_{fond} et d_{fond} doivent être positifs et plus petit que 1. d_{fond} défini le ratio de fond qui sera diffusé sur le recto. Les auteurs de la méthode fixent les valeurs de δ_{fond} et σ_{fond} à respectivement 0.2 et 0.3.

$$c_{verso} = \frac{d}{1 + (s-u)^2/\sigma_b^2} \frac{1}{1 + s^2/\sigma_{ink}^2} \quad (10)$$

Pour la diffusion du verso sur le recto, le coefficient c_{verso} (équation 10) doit seulement diffuser les pixels d'encre du verso sur le recto. Le paramètre d correspond au ratio de la diffusion des pixels de verso sur le recto. σ_b contrôle le degré de transparence qui apparaîtra sur le recto. Pour finir σ_{ink} est un paramètre permettant de limiter la diffusion au pixel d'encre du verso. Dans les expérimentations, σ_{ink} est fixé à 0.2.

5.3 Avantages et inconvénients

Dans les faits, ce modèle de génération de document dégradé s'avère assez pratique et simple à implémenter (si le développeur possède de bonnes bases en EDP [Equations aux Dérivés Partielles]). Dans le cadre de DIGIDOC, le modèle a permi de constituer rapidement une base de documents transparents avec leurs vérité terrain.

Parmi les inconvénients du modèle, le premier est d'ordre scientifique : aucune étude n'a été menée quand à la fidélité, par rapport à des documents

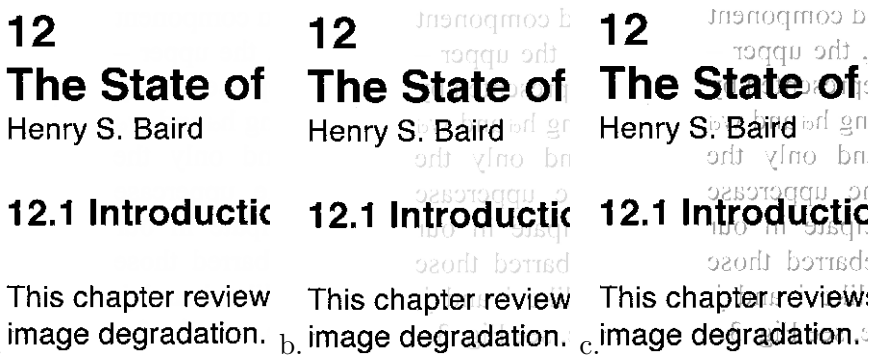


FIGURE 15 – Génération de transparence : a. l'image original, b-c. la même image avec un niveau de transparence de plus en plus élevé (l'image a. est de plus en plus dégradée).

réels, des défauts générés. Il est surement possible de montrer des corrélations ou des tendances statistiques entre ces documents synthétiques et la réponse de certains algorithmes comme l'OCR (**O**ptical **C**haracter **R**ecognition). Cela ne garantit pas que les défauts générés soient fidèles aux défauts des documents anciens.

Nous avons constaté que le modèle avait tendance à modifier un peu chaque couche lors de la diffusion. En effet l'encre est un peu lissée, et des artefacts étranges apparaissent. La présence de ces artefacts sur l'image finale peut être contrôlé en affinant la paramétrisation du modèle.

Un autre inconvénient de cette méthode est due à la complexité de sa paramétrisation. En effet, autant certains des paramètres restent peu sensibles autant d'autres sont complexes à définir. Il est par exemple nécessaire de changer les paramètres entre une phase de génération avec des documents anciens (dont le niveau de gris de l'encre varie du noir au gris claire) et une phase de génération avec des documents propres (le niveau de gris de l'encre reste très fonçée).

Le dernier problème est plus un aspect technique à prendre en compte lors de l'implémentation d'un tel modèle. En effet, même si l'article ne le mentionne pas, il est nécessaire d'utiliser des images dont le type des pixels sont des nombres à virgule flottante. Sans cela, le modèle reste fonctionnel, mais les résultats sont nettement moins satisfaisants.

6 Correction de défauts d'éclairage et de déformations géométriques

6.1 Introduction

Dans [34], l'auteur décrit les défauts les plus couramment rencontrés. Ainsi, on distingue deux " familles " de défauts :

1. Ceux inhérents aux ouvrages : détérioration du papier, transparence du verso due à l'acidité de l'encre...
2. Ceux liés à la phase de numérisation : défauts d'éclairage, problèmes de courbure et d'inclinaison de l'image...

De nombreuses solutions ont été proposées. [34] en recense plus d'une dizaine de méthodes (essentiellement développées par des industriels) et propose, lui aussi, diverses solutions pour venir à bout des problèmes évoqués précédemment. Dans la figure 16, on peut remarquer le type de corrections réalisées sur l'image après l'utilisation d'un logiciel de restauration de documents anciens. On peut notamment observer que les taches ont disparu et que les défauts de courbure ont été atténués. L'objectif est donc principalement d'améliorer le rendu visuel de la version numérique de ces documents. Ces outils déforment ou corrigent l'image afin d'obtenir un meilleur rendu. Cependant, ces traitements "endommagent" l'image initiale, ce qui n'est pas sans conséquences sur l'exploitation du contenu [5] .

(a) Image numérisée en couleur *(b) Image en niveau de gris redressée* *(c) Image restaurée*

FIGURE 16 – Exemple de corrections de défauts de numérisation avec[34]

5. Toutes le images de cette section sont extraites [34]

6.2 Correction de défauts d'éclairages

L'auteur de [34] propose une méthode de correction de défauts d'éclairage fréquemment rencontrés lors de la numérisation d'ouvrages anciens (figure 17).

FIGURE 17 – Exemple de défauts dus à un éclairage non uniforme

La méthode proposée modélise le phénomène physique apparaissant dès lors que l'ouvrage est épais et que la reliure empêche la numérisation d'une page sur un plan (figure 18).

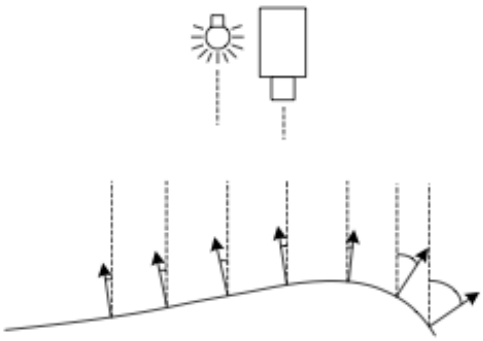


FIGURE 18 – Modèle physique de déformation adoptée par [34] expliquant la non uniformité de l'éclairage

Les auteurs proposent donc en 2 étapes de corriger ce type de défauts partant du principe que la teinte du fond de la page reste constante sur toute la page quand la luminosité varie :

1. Calcul d'un histogramme de luminosité et d'un histogramme de teinte
2. Détermination de la couleur du fond
3. Détermination de la luminosité maximale par colonne
4. Détermination de la moyenne des luminosités des pixels du fond par colonne

Pour restaurer l'image, la luminosité de chaque pixel est calculée de la manière suivante :

$$p = \frac{ln(1 - \frac{m}{M})}{ln(1 - \frac{l}{M})}$$

avec :
– l : niveau de gris du pixel à modifier
– m : niveau de gris moyen du fond
– M : niveau de gris max observé

6.3 Correction géométriques

Un autre défaut couramment rencontré, pour la même raison que pour le défaut d'éclairage, est celui du à l'épaisseur de l'ouvrage et qui engendre un défaut de courbure sur l'extrémité de l'ouvrage ou au milieu de l'ouvrage si celui-ci est scanné en double page (figure 19).

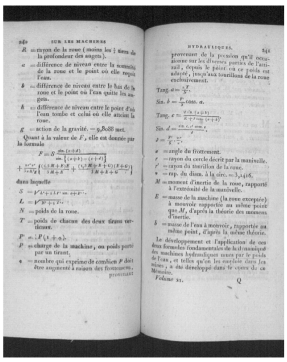


FIGURE 19 – Exemple de déformation géométrique

Pour corriger ce défaut, les auteurs présentent la méthode suivante :

1. Détection de l'inclinaison (globale) de la page :
   - Sobel vertical
   - Calcul gradient + seuillage
   - Etude de l'alignement des composantes connexes
2. Détection de la courbure des bords (Recherche *ad hoc* de surfaces noires)
3. Détection des lignes de texte

Une synthèse des résultats est effectuée pour permettre la correction de l'image. Comme l'illustre la figure 20, chaque succession de point permet d'obtenir des lignes "courbes". Une interpolation d'équation de degré n de chaque courbe selon un axe vertical et un intervalle pré-défini est calculée. Enfin, la bordure de la page sert de repère horizontal.

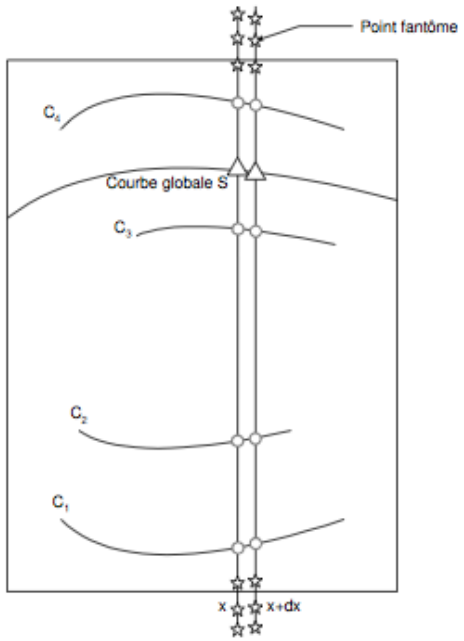


FIGURE 20 – Détail du modèle de correction géométrique

La figure 21 illustre le type de résultats obtenus par les auteurs de [34].

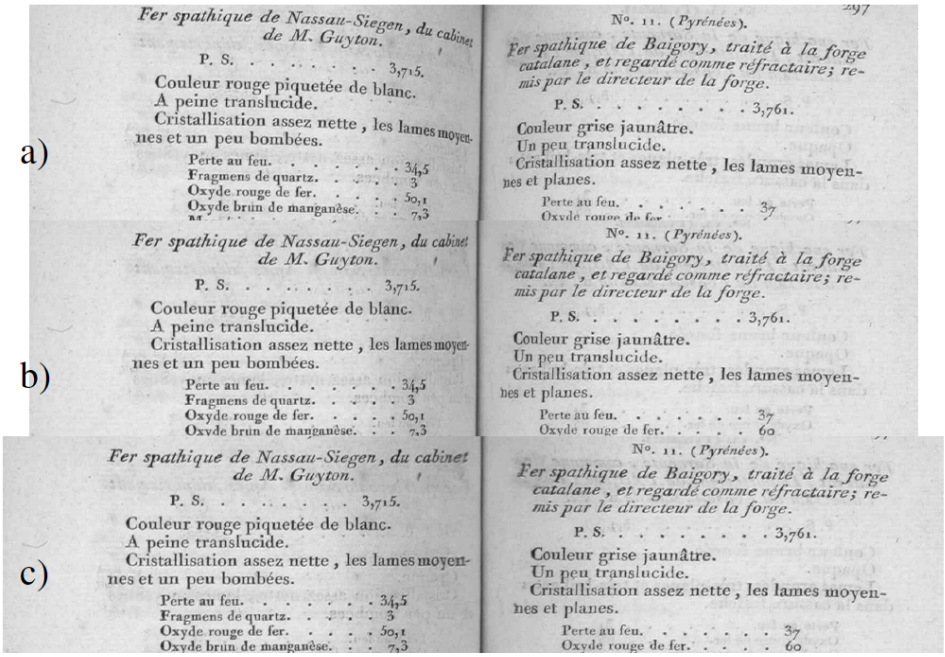


FIGURE 21 – Résultats obtenus : a) Portion de l'image originale b) Portion de l'image redressée sans correction de l'écrasement c) Portion de l'image redressée avec correction de l'écrasement

Références

[1] Selim Aksoy, Ming Ye, Michael L. Schauf, Mingzhou Song, Yalin Wang, Robert M. Haralick, Jim R. Parker, Juraj Pivovarov, Dominik Royko, Changming Sun, and Gunnar Farnebäck. Algorithm performance contest. *Pattern Recognition, International Conference on*, 4 :48–70, 2000.

[2] Bénédicte Allier. *Contribution à la Numérisation des Collections : Apports des Contours Actifs.* Thèse de doctorat, Université de Lyon, 2003.

[3] A. Antonacopoulos and H. Meng. A ground-truthing tool for layout analysis performance evaluation. In *Document Analysis Systems (DAS)*, volume 2423 of *Lecture Notes in Computer Science (LNCS)*, pages 651–660, 2002. Performance Evaluation (Layout).

[4] G. Bal, G. Agam, and O. Frieder. Interactive degraded document enhancement and ground truth generation. In *Document Recognition and*

Retrieval (DRR), volume 6815 of *SPIE Proceedings*, 2008. Performance Evaluation (OCR).

[5] J. Beusekom, F. Shafait, and T.M. Breuel. Automated ocr ground truth generation. In *Workshop on Document Analysis Systems (DAS)*, 2008. Performance Evaluation (OCR).

[6] F. Le Bourgeois and H. Emptoz. Document analysis in gray level and typography extraction using character pattern redundancies. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 177–180, 1999. Performance Evaluation (OCR).

[7] Mathieu Delalandre, Ernest Valveny, Tony Pridmore, and Dimosthenis Karatzas. Generation of synthetic documents for performance evaluation of symbol recognition & spotting systems. *International Journal on Document Analysis and Recognition*, 13 :187–207, 2010.

[8] D. Doermann and G. Zi. Groundtruth image generation from electronic text. In *Symposium on Document Image Understanding and Technology (SDIUT)*, pages 309–312, 2003. Performance Evaluation (OCR).

[9] P. Heroux, E. Barbu, S. Adam, and E. Trupin. Automatic groundtruth generation for document image analysis and understanding. In *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, pages 476 –480, 2007.

[10] P. Heroux, E. Barbu, S. Adam, and E. Trupin. Automatic ground-truth generation for document image analysis and understanding. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 476–480, 2007. Performance Evaluation (Layout).

[11] J.D. Hobby. Matching document images with ground truth. *International Journal on Document Analysis and Recognition (IJDAR)*, 1(1) :52–61, 1998. Performance Evaluation (OCR).

[12] J.D. Hobby and H Tin Kam. Enhancing degraded document images via bitmap clustering and averaging. In *International Document Analysis and Recognition (ICDAR)*, volume 1, pages 394–400, 1997. Performance Evaluation (OCR).

[13] M. Huanfeng and D. Doermann. Adaptive ocr with limited user feedback. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 814–818, 2005. Performance Evaluation (OCR).

[14] C. Huang and S.N. Srihari. Mapping transcripts to handwritten text. In *International Workshop on Frontiers in Handwriting Recognition (IWFHR)*, pages 15–20, 2006. Performance Evaluation (Handwritting).

[15] C.V. Jawahar and A. Kumar. Content-level annotation of large collection of printed document images. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 799–803, 2007. Performance Evaluation (Handwritting).

[16] T. Kanungo and R. Haralick. Pattern analysis and machine intelligence, ieee transactions on. *An automatic closed-loop methodology for generating character groundtruth for scanned documents*, 21(2) :179–183, 1999.

[17] T. Kanungo and R.M. Haralick. An automatic closed-loop methodology for generating character groundtruth for scanned documents. *Pattern Analysis and Machine Intelligence (PAMI)*, 21(2) :179–183, 1999. Performance Evaluation (OCR).

[18] T. Kanungo, R.M. Haralick, H.S. Baird, W. Stuezle, and D. Madigan. A statistical, nonparametric methodology for document degradation model validation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(11) :1209 – 1223, 2000.

[19] T. Kanungo, R.M. Haralick, h.S. Baird, W. Stuezle, and D. Madigan and. A statistical, nonparametric methodology for document degradation model validation. *Pattern Analysis and Machine Intelligence (PAMI)*, 22(11) :1209– 1223, 2000. Performance Evaluation (Degradation Model).

[20] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes : Active contour models. *International Journal of computer vision*, 1(4) :321–331, 1988.

[21] Doe-Wan Kim and Tapas Kanungo. Attributed point matching for automatic groundtruth generation. *International Journal on Document Analysis and Recognition*, 5 :47–66, 2002.

[22] D.W. Kim and T. Kanungo. Attributed point matching for automatic groundtruth generation. *International Journal on Document Analysis and Recognition (IJDAR)*, 5(1) :47–66, 2002. Performance Evaluation (OCR).

[23] C.H. Lee and T. Kanungo. The architecture of trueviz : A groundtruth / metadata editing and visualizing toolkit. *Pattern Recognition (PR)*, 36(3) :811–825, 2003. Performance Evaluation (Layout).

[24] D. Lopresti and G. Nagy. Issues in ground-truthing graphic documents. In *Workshop on Graphics Recognition (GREC)*, volume 2390 of *Lecture Notes in Computer Science (LNCS)*, pages 46–66, 2002. Performance Evaluation (GREC).

[25] S. Mao and T. Kanungo. Software architecture of pset : a page segmentation evaluation toolkit. *International Journal on Doucment Analysis and Recognition (IJDAR)*, 4 :205–217, 2002. Performance Evaluation (Layout).

[26] Reza Moghaddam and Mohamed Cheriet. Low quality document image modeling and enhancement. *International Journal on Document Analysis and Recognition*, 11 :183–201, 2009.

[27] R.F. Moghaddam and M. Cheriet. Low quality document image modeling and enhancement. *International journal on document analysis and recognition*, 11(4) :183–201, 2009.

[28] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 12(7) :629–639, 1990.

[29] M. Rigamonti, O. Hitz, and R. Ingold. A framework for cooperative and interactive analisys of technical documents. In *Workshop on Graphics Recognition (GREC)*, pages 407–414, 2003. Performance Evaluation (Layout).

[30] E. Saund and J. Lin. Pixlabeler : User interface for pixel-level labeling of elements in document images. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 646–650, 2009. Performance Evaluation (Layout).

[31] N. Stamatopoulos, G. Louloudis, and B. Gatos. Efficient transcript mapping to ease the creation of document image segmentation ground truth with text-image alignment. In *International Conference on Frontiers in Handwriting Recognition (IWFHR)*, pages 226–231, 2010. Performance Evaluation (Layout).

[32] P. Stathis, E. Kavallieratou, and N. Papamarkos. An evaluation technique for binarization algorithms. *Journal of Universal Computer Science*, 14(8) :3011–3030, 2008.

[33] A. Toselli, V. Romero, and E. Vidal. Viterbi based alignment between text images and their transcripts. In *Workshop on Language Technology for Cultural Heritage Data (LaTeCH)*, pages 9–16, 2007. Performance Evaluation (Handwritting).

[34] Trinh. *De la numérisation à la consultation de documents anciens*. PhD thesis, Université De Lyon, 2003.

[35] Chenyang Xu and Jerry L. Prince. Snakes, shapes, and gradient vector flow. In *IEEE Transaction on image processing*, volume 7, pages 359–369. IEEE Signal Processing Society, 1998.

[36] S. Yacoub, V. Saxena, and S. Sami. Perfectdoc : A ground truthing environment for complex documents. In *International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 452–456, 2005. Performance Evaluation (Layout).

[37] B.A. Yanikoglu and L. Vincent. Pink panther : a complete environment for ground-truthing and benchmarking document page segmentation. *Pattern Recognition (PR)*, 31(9) :1191–1204, 1998. Performance Evaluation (Layout).

[38] Jian Zhai, Liu Wenyin, Dov Dori, and Qing Li. A line drawings degradation model for performance characterization. *Document Analysis and Recognition, International Conference on*, 2 :1020–1024, 2003.

[39] D. Zhang. *Stochastic methods for flow in porous media : coping with uncertainties*. Academic Pr, 2002.

[40] G. Zi. Groundtruth generation and document image degradation. Technical report, DTIC Document, 2005.

[41] G. Zi and D. Doermann. Groundtruth image generation from electronic text. In *Symposium on Document Image Understanding Technology (SDIUT)*, pages 309–312, 2003. Performance Evaluation (OCR).

[42] G. Zi and D. Doermann. Document image ground truth generation from electronic text. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 2, pages 663–666. IEEE, 2004.