A character degradation model for grayscale ancient document images

Van Cuong Kieu, Jean-Philippe Domenger, Rémy Mullot, Nicholas Journet, Muriel Visani

Génération d'images semi-synthétiques de documents anciens

Van Cuong KIEU[1]

Laboratoire Bordelais de Recherche en Informatique - Université Bordeaux 1

RÉSUMÉ. *Cet article présente un logiciel de génération d'images semi-synthétiques de documents anciens et de la vérité terrain associée. Ce travail s'inscrit dans le cadre de la génération de données pour l'évaluation de performances d'algorithmes d'analyse d'images de documents. Ce logiciel permet à un utilisateur de spécifier le contenu et la mise en page des images à générer (choix des fontes, illustrations, interlignes). Plusieurs modèles de dégradations ont été intégrés dans ce logiciel. Il est ainsi possible de générer des images contenant les défauts les plus couramment observés dans les ouvrages anciens (transparence, dégradation des caractères et pliures d'une page). En générant une grande variété de documents, il est ainsi possible d'évaluer la robustesse d'un algorithme vis-à-vis de ces dégradations.*

ABSTRACT. *This paper presents a software dedicated to semisynthetic old document image generation and its associated ground truth. This work is thus part of the data generation for document processing algorithm performance evaluation. This software allows a user to specify the content and layout of images to generate (fonts, illustrations, line spacing). Several degradation models were incorporated into the software. It is thus possible to generate images containing defects most commonly seen in old books (ink transparency, character degradation). By generating a huge variety of documents, it is possible to assess the robustness of an algorithm according to these degradations.*

MOTS-CLÉS : *modèle de dégradation d'images de documents, génération de vérité terrain, évaluation de performance, bases de données synthétiques.*

KEYWORDS: *model of document image degradation, ground truth generation, performance evaluation, synthetic databases.*

1. Encadrants : J.P.Domenger*, R.Mullot**, N.Journet*, M.Visani**

* Laboratoire LaBRI, Université Bordeaux I, 351 cours de la libération 33405 Talence Cedex,

** Laboratoire L3i, Université de La Rochelle Dépt. Informatique, avenue Michel Crépeau 17042 La Rochelle Cedex 1, France

1. Introduction

Les compétitions organisées récemment dans les conférences en analyse d'images de documents (ICDAR2011, GREC2011) témoignent de l'importance de la problématique de l'évaluation de performances d'algorithmes d'analyse d'images de document. Si certaines de ces bases (et leur vérité terrain associée) sont générées de manière automatique (M.Delalandre *et al.* 2007) (L.Yang *et al.* 2006), la grande majorité d'entre elles ont nécessité la réalisation d'une étape de saisie manuelle de la vérité terrain. Les auteurs de (Antonacopoulos *et al.* 2006) présentent une solution logicielle au travers de la plate-forme Aletheia. Elle permet à un ensemble d'experts d'annoter précisément le contenu de chaque image d'une base. L'originalité d'Aletheia, est qu'elle propose tout un ensemble d'outils permettant de simplifier et donc d'accélérer cette phase d'annotation. Néanmoins, comme le soulignent les auteurs, ce choix de générer manuellement la vérité terrain d'images de documents réels reste fastidieux en termes de temps de travail. Il se pose également le problème de la subjectivité d'une annotation et donc de la validité des informations saisies manuellement.

Si les travaux les plus avancés traitant de la génération d'images semi-synthétiques sont ceux liés à l'analyse de symboles (M.Delalandre *et al.* 2007) ou de graphiques (L.Yang *et al.* 2006), il existe en revanche peu de propositions sur la génération de documents contemporains et aucune pour les documents anciens. Dans (Pavlos Stahis *et al.* 2008) les auteurs proposent de tester la qualité de divers algorithmes de binarisation sur des images de documents contemporains sur lesquels ont été artificiellement superposé des fonds (bruités) issus de documents anciens. Les auteurs de (Zi *et al.* 2004) et (Heroux *et al.* 2007) proposent une approche dans laquelle il est possible, pour l'utilisateur, de spécifier le contenu de chaque document généré. Le générateur proposé par (Heroux *et al.* 2007) se base sur une DTD fournie par l'utilisateur. Ce dernier peut ensuite utiliser des outils tels que word ou LaTeXpour générer une grande variété d'images. Le système proposé par (Zi *et al.* 2004), repose sur l'analyse de fichiers de configurations décrivant de manière littérale le contenu de chaque image (fichier ttf, position des blocs, taille du texte,...). Les auteurs proposent ensuite de dégrader physiquement une image en l'imprimant (depuis une imprimante, un fax,...) et de la numériser de nouveau afin de simuler un état de dégradation varié.

Ces travaux sur la génération semi-synthétique d'images de documents anciens ont été amorcés par (Journet *et al.* 2010). Cet article présente les évolutions de ce travail et les perspectives qui seront abordées lors des 3 prochaines années. La première partie de cet article décrit la méthodologie que nous proposons pour une génération d'images de document anciens auxquelles sont associées les données de vérité terrain. Dans une seconde partie, nous détaillons comment sont extraites les données servant à générer ensuite des documents factices. Nous détaillons également les modèles de dégradation que nous avons implémenté et qui permettent de générer des images de documents anciens tentant de reproduire les défauts les plus couramment rencontrés. Enfin, nous détaillons le type de vérité terrain qui est associé à chaque image synthétique.

2. Processus de génération de documents semi-synthétiques

La figure 1 illustre l'ensemble des étapes nécessaires au processus de génération et d'annotation automatique d'images de documents semi-synthétiques. Ce processus se décompose en 5 étapes : (1) Extraction de données sources : des éléments (jeux de caractères, images de fonds, illustrations, ...) sont extraits manuellement sous la forme d'imagettes à partir de documents réels. A chaque imagette est associé un ensemble d'informations (label, position et taille dans l'image d'origine, ...). (2) Modèles de dégradation : il est possible de générer, à partir d'une image source, tout un ensemble de défauts simulant ceux observés dans les images réelles (transparence, courbure de page, dégradation des caractères). (3) Paramétrage des données à générer : un utilisateur peut définir de manière précise le type d'images à générer (jeux de caractères, mise en page, défauts présents, niveaux de dégradations, ...). Les étapes (4) et (5) sont celles permettant de générer les images (4) et la vérité terrain XML associée (5).

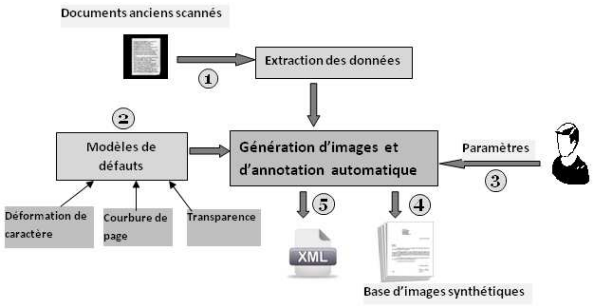

Figure 1. *Modèle de génération et d'annotation automatique de documents anciens.*

3. Modèle de génération et d'annotation d'images semi-synthétique

3.1. *Extraction de données*

Cette étape d'extraction de données réelles a pour objectif de constituer une base d'imagettes qui servira à construire des documents factices. Dans la version actuelle du logiciel, l'utilisateur délimite à la souris les éléments qu'il souhaite voir plus tard apparaître dans les documents synthétiques (différents caractères, illustrations, fonds, ...). Une perspective concernant cette étape d'extraction est de fournir à l'utilisateur des outils d'extraction semi-automatiques. En effet, à l'instar des auteurs de (Antonacopoulos *et al.* 2006) nous pensons qu'il faut simplifier cette tâche fastidieuse. Nous pensons par exemple mettre à disposition de l'utilisateur des images dont les caractères, lignes, illustrations sont déjà segmentées afin d'accélérer cette phase de détourage des composantes. Une autre piste sur laquelle nous nous pencherons est celle de la validité des données extraites. En effet les caractères, illustrations ou fonds extraits par l'utilisateur possèdent déjà, dans certains cas, de fortes dégradations engendrant donc des différences entre les divers exemplaires d'un même caractère ou d'une

même illustration extraits sur une image réelle. Nous pensons mettre à disposition de l'utilisateur des outils lui permettant, s'il le souhaite, de restaurer certaines formes avant qu'il ne les extraient. Nous pensons par exemple dans le cas des caractères segmentés en plusieurs composantes après binarisation, utiliser l'approche proposée par (Allier 2003). Ainsi, l'utilisateur pourra extraire des caractères ne présentant pas de défauts.

3.2. *Modèles de défauts*

Actuellement, le générateur intègre trois modèles de dégradation de documents (transparence, dégradation des caractères et déformation de l'extrémité de la page). Pour chacun d'entre eux, nous sommes partis de modèles de la littérature permettant de restaurer des images dégradées. Nous les avons adaptés afin de générer ces défauts. Le premier défaut que nous sommes en mesure de générer est celui de la transparence de l'encre du verso d'une feuille de document sur la page du recto. Ce défaut est souvent dû à l'acidité de l'encre qui, par capillarité, traverse l'épaisseur de la feuille. Il peut être également dû à de mauvais réglages de lumières lors de l'étape de numérisation. Nous avons adapté le modèle de (Moghaddam R.F. 2009) pour lequel il faut spécifier deux images (une correspondant au recto et l'autre au verso) et un paramètre correspondant à un coefficient de diffusion des pixels d'une image vers l'autre.

Le deuxième modèle est celui permettant de simuler la dégradation de l'encre et qui touche essentiellement les contours des caractères. Nous avons adapté le modèle de (T. Kanungo 1994) basé sur l'hypothèse que la probabilité pour un pixel de s'éclaircir (voire de disparaitre) est lié à la distance le séparant de la frontière de la forme à laquelle il appartient.

Enfin, nous avons adapté le modèle présenté par (T. Kanungo 1994) permettant de simuler l'apparition plus ou moins marquée de la reliure lors de l'étape de numérisation. Le modèle permet d'intégrer l'épaisseur de cette reliure et de ce fait l'impression de courbure plus ou moins forte qui est perçue à l'extrémité de l'image. La figure 2 illustre quelques exemples d'images de documents synthétiques générés avec les modèles de défauts proposés par (T. Kanungo 1994) et (Moghaddam R.F. 2009).

L'ajout de nouveaux modèles de dégradation inspirés de modèles de restauration est un de nos objectifs. Nous pensons tout d'abord travailler sur l'adaptation de modèle de (Allier 2003) simulant l'apparition de ruptures de connexité d'un caractère. D'autres modèles peuvent être adaptés pour simuler des effets de flous, de sur (ou sous) éclairage, après de mauvais réglages de l'appareil de numérisation (Thrin 2003), (Jian Zhai 2003).

3.3. *Paramètres et vérité-terrain*

Cette dernière étape regroupe les points $3, 4, 5$ de la figure 1. Elle consiste pour un utilisateur, à formuler à l'aide d'une interface graphique l'emplacement de blocs de

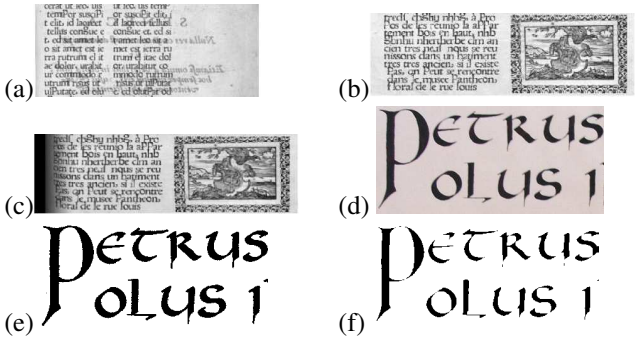

Figure 2. *Images obtenues après application de modèles de déformation : (a) image présentant une défaut de transparence ; (b) image sans déformation ; (c) image simulant une déformation due à une reliure (d) image d'origine ; (e) image binarisée ; (f) image avec déformation des caractères*

texte et d'illustration. Pour chaque bloc, il est possible de définir la fonte à utiliser, les règles de position des caractères les uns par rapport aux autres... Pour la page dans sa globalité, il est possible de définir le fond utilisé ainsi que les modèles de défauts à appliquer. Il est possible, à l'aide d'une interface graphique, de créer des documents anciens semi-synthétiques. Pour chaque image est généré un fichier XML dans lequel est stocké l'ensemble des informations contenues dans l'image. Ces informations sont de deux niveaux différents. Le premier est lié à la structure de la page générée (positions des caractères et leur label, espace interligne, nom de la fonte utilisée, ...). le deuxième niveau d'information est lié aux défauts (locaux ou globaux) utilisés pour générer l'image.

Le principal travail lié à cette étape de paramétrage et de génération d'images concerne les mécanismes de génération automatique des pages s'appuyant sur les variations envisagées par l'utilisateur et ce, en suivant des scénarios de génération proposés à l'utilisateur. Pour réaliser ces images, nous souhaitons utiliser les informations extraites lors de l'étape 1 (extraction des imagettes) afin d'extraire automatiquement des règles qui serviront lors de l'étape de génération. Nous travaillons actuellement sur un module permettant d'apprendre les règles de disposition des caractères les uns par rapport aux autres (valeur de la ligne de base, de l'espace interligne,...). L'objectif est donc d'éviter à l'utilisateur de définir lui-même tout un ensemble de paramètres et malgré tout de pouvoir disposer d'imagettes dégradées par un large spectre de défauts.

4. Conclusion

Cet article présente la nouvelle version d'un logiciel de génération semi-synthétiques d'images de documents anciens. Ces images, associées à leur vérité terrain, peuvent être utilisées dans le cadre de l'évaluation de performances d'algo-

rithmes de traitement ou d'analyse d'images de documents anciens. En positionnant des imagettes de caractères et d'illustrations issues de documents anciens réels, il est possible de générer des images factices très similaires à des documents réels. L'application de modèles de dégradation sur ces images (transparence de l'encre, déformation des caractères, apparition de la reliure sur l'extrémité d'une page) permet de reproduire l'apparition de défauts observés sur des documents réels.

5. Bibliographie

Allier B., Contribution à la Numèrisation des Collections : Apports des Contours Actifs, Thèse de doctorat, Université de Lyon, 2003.

Antonacopoulos A., Karatzas D., Bridson D., « Ground truth for layout analysis performance evaluation », *Document Analysis Systems VII*, vol. VII, Springer, Nelson, New Zealand, p. 302-311, February, 2006.

Heroux P., Barbu E., Adam S., Trupin E., « Automatic Ground-truth Generation for Document Image Analysis and Understanding », *Document Analysis and Recognition, ICDAR 2007. Ninth International Conference on*, Curitiba, State of Parana, Brazil, p. 476-480, Sept, 2007.

Jian Zhai Liu Wenyin D. D. Q. L., « A Line Drawings Degradation Model for Performance Characterization », *In Proceedings. Seventh International Conference on Document Analysis and Recognition*, IEEE Computer Society, Edinburgh, Scotland, UK, p. 1020-1024, August, 2003.

Journet N., Vialard A., Domenger J.-P., « Analyse de fontes anciennes : de la génération de données synthétiques à la reconnaissance », *Colloque International Francophone sur l'Ecrit et le Document*, Tunisie, p. 51-66, 2010.

L.Yang W., C.Tan, « Semi-automatic ground truth generation for chart image recognition », *In Workshop on Document Analysis Systems (DAS)*, Nelson, New Zealand, p. 324-335, February, 2006.

M.Delalandre T.Pridmore E. E., H.Locteau, « Building synthetic graphical documents for performance evaluation », *In Workshop on Graphics Recognition*, Grec, p. 84-87, 2007.

Moghaddam R.F. C. M., « Low quality document image modeling and enhancement », *Int. J.Doc. Anal. Recognit*, vol. 11, Springer, Berlin, Heidelberg, p. 183-201, March, 2009.

Pavlos Stahis E. K., paramarkos N., « An Evaluation Technique for Binarization Algorithms », *Journal of Universal Computer Science*, vol. 14, n 18, p. 3011-3030, 2008.

T. Kanungo R. M. Haralick I. P., « Non-linear local and global document degradation models », *Journal of Imaging Systems and Technology*, vol. 5, n 4, p. 220-30, 1994.

Thrin E., De la numerisation a la consultation des documents anciens, Thèse de doctorat, Université de Lyon, 2003.

Zi G., Doermann D., « Document image ground truth generation from electronic text », *Proceedings of the 17th International Conference on Pattern Recognition*, IEEE Computer Society, Cambridge, UK, p. 663-666, August, 2004.