# Prétraitement de documents anciens

H. Boulehmi, B. Seddik, A. Kricha, N. Essoukri Ben Amara

# Prétraitement de documents anciens

Héla Boulehmi[1] Bassem Seddik[2] Anis Kricha[3] Najoua Essoukri Ben Amara [4]

[1]Ecole Nationale d'Ingénieurs de Tunis
[2]Institut Supérieur des Sciences Appliquées et Technologies de Sousse
[3]Ecole Nationale d'Ingénieurs de Monastir
[4]Ecole Nationale d'Ingénieurs de Sousse

hela.boulehmi@yahoo.fr Anis.kricha@topnet.tn
bassemmaster@gmail.com Najoua.Benamara@eniso.rnu.tn

**Résumé** : *La problématique liée au prétraitement des documents anciens reste encore d'actualité. Des approches multiples ont été proposées pour le rehaussement de la qualité des documents anciens cependant la présence simultanée, dans un même document, de deux ou plusieurs types de bruits laisse le problème encore ouvert.*

*Dans ce papier, nous proposons une approche de prétraitement basée sur une analyse hybride texturale et statistique permettant d'identifier d'abord les défauts éventuels qu'elle présente (acidité de l'encre, recto/verso…) et d'appliquer par la suite le prétraitement approprié moyennant l'exploration de la morphologie mathématique multi-échelle.*

*Les différentes expérimentations réalisées ont été validées sur des images de manuscrits et imprimés anciens arabes en niveaux de gris, issues de la Bibliothèque Nationale de Tunis.*

**Mots-clés** : Documents anciens, caractéristiques statistiques et texturales, prétraitement, morphologie mathématique multi-échelle.

## 1 Introduction

Alors que les documents anciens sont d'une valeur inestimable, ils montrent souvent des dégradations plus ou moins importantes à cause des conditions de préservation et risquent ainsi de devenir inexploitables. Généralement parlant les défauts liés aux documents anciens peuvent être regroupés en deux grandes familles : ceux qui sont intrinsèques aux documents (humidité, acidité de l'encre, verso visible à partir du recto, pliure…) et ceux qui sont produits par la chaîne de numérisation (faible contraste, variation d'éclairage…) [KRI 06, GHA 06, DRI 07]. Pour remédier à ces problèmes plusieurs projets de numérisation de documents anciens ont vu le jour [KRI06]. De multiples techniques ont été développées pour le prétraitement des documents anciens [LIK 04, KRI 06, DEB 00], mais aucune d'elles n'est « universelle ». Elles sont généralement spécifiques à un seul défaut et font souvent apparaître des effets secondaires sur d'autres éléments de l'image considérée. En outre, nous avons mené une étude sur un grand nombre d'images issues de documents anciens, qui a montré qu'il est plutôt rare qu'un problème existe seul dans un document. Très souvent, dans le même document coexistent deux ou plusieurs défauts à la fois, ce qui nécessite des techniques de prétraitement plus élaborées. Dans ce papier, nous proposons une approche de prétraitement de documents anciens basée essentiellement sur l'exploration de la texture et la morphologie mathématique multi-échelle.

Dans la section suivante, nous présentons la chaîne développée. Les principaux tests et résultats enregistrés sont présentés dans la troisième section.

## 2 Chaîne proposée

La Figure 1 donne le schéma bloc de la chaîne de traitement proposée, notre objectif étant de développer une approche de prétraitement d'un défaut lié aux documents anciens dans son contexte (qui peut inclure d'autres types de défauts). Il s'agit d'identifier d'abord le (ou les) défaut présent dans une image de document ancien, puis d'appliquer le prétraitement approprié.

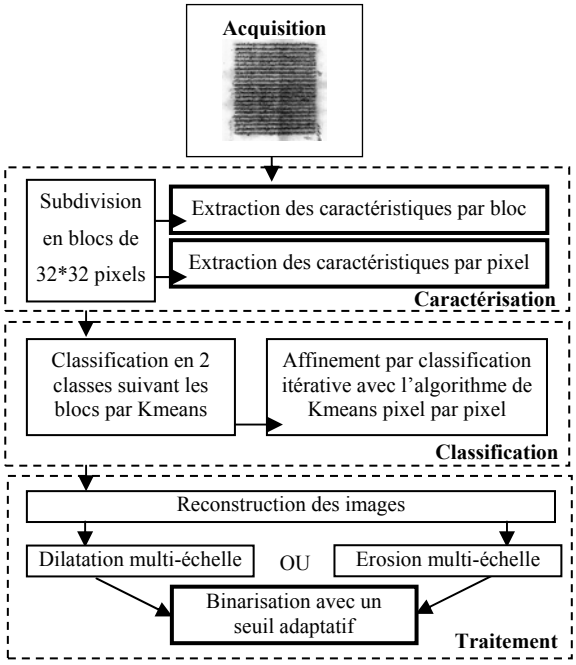



FIG. 1- Schéma bloc de la chaîne de prétraitement proposée

Au niveau de la phase de caractérisation et compte tenu des différents travaux développés au sein de notre équipe [HEN 08, KRI 07, BEN 08, GHA 06], nous avons opté pour l'usage de primitives de deux types : textural et statistique issus respectivement de la décomposition en ondelettes et de la matrice de cooccurrence associée à l'image du document.

De la transformée en ondelettes, nous avons retenu la moyenne et l'écart-type de la matrice d'approximation, les écart-types des matrices associées aux détails horizontaux, verticaux et diagonaux.

Comme caractéristiques statistiques, nous avons choisi des métriques de premier ordre et d'autres liées à la matrice de cooccurrence de niveaux de gris de l'image considérée. De cette matrice nous avons extrait les paramètres liés à la corrélation, l'énergie, l'entropie, l'homogénéité et au contraste [HIR 08].

Pour le traitement des images, nous avons opté pour la morphologie mathématique multi-échelle [MAN 05, RIV 08]. Pour un pixel donné, la taille de l'élément structurant varie en fonction des niveaux de gris de tous ses pixels voisins. Un seuil de binarisation adaptatif est alors appliqué.

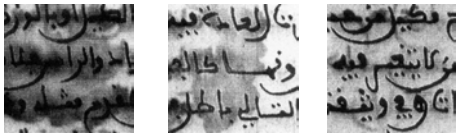
Nous présentons dans la section suivante les résultats de la phase de prétraitement proposée.

# 3 Expérimentations et résultats

Les différents tests ont été effectués sur des images issues de documents anciens provenant de la Bibliothèque Nationale de Tunis [BNT]. Les documents considérés sont de taille 1600*2240, numérisés à une résolution de 300 ppp sous format "TIF" et couvrent la majorité des problèmes liés aux documents anciens (acidité, humidité, recto/verso, faible contraste).

Au niveau caractérisation, nous avons retenu comme ondelette la « bior1.3 » qui a été prouvée efficace dans [GHA 06]. Au niveau binarisation, un élément structurant de taille variable a été adopté. Pour un pixel donné, la taille de l'élément structurant, initialisée à 3x3 pixels, varie en fonction des niveaux de gris de tous ses pixels voisins.

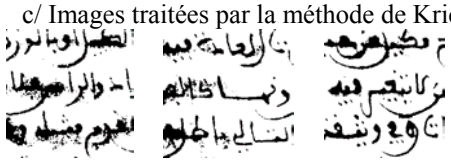
La figure 3 illustre sur trois images, les résultats de notre approche de prétraitement des défauts. Une comparaison avec les méthodes d'OTSU [PEC 01] et Kricha [KRI 06], montre des résultats prometteurs.



a/ Images originales

b/ Images traitées par la méthode d'Otsu

c/ Images traitées par la méthode de Kricha



d/ Images traitées par l'approche proposée

FIG. 3 – Comparaison des résultats de l'approche développée avec les méthodes d'OTSU et Kricha.

# 4 Conclusion et perspectives

Dans ce papier nous avons adressé la problématique liée au prétraitement des documents anciens. La présence, souvent simultanée, de plusieurs défauts dans le même document rend la tâche de prétraitement non triviale. Le défaut est souvent considéré de manière isolée sans tenir compte de son contexte. Nous avons proposé pour cela, une approche basée sur la classification en première étape et la morphologie mathématique multi-échelle ensuite pour le débruitage d'images de documents anciens.

Les différentes expérimentations réalisées sur un nombre important d'images de documents anciens, montrent l'efficacité de notre approche.

# 5 Bibliographie

[BEN 08] BEN ABDEJLIL J., KRICHA A., ESSOUKRI BEN AMARA N., Exploring a new wavelet in image processing, *accepté ISIE*, 2008.

[BNT ] http://www.bnt.nat.tn/

[DEB 00] BOUCHE R., Présentation du projet européen Debora, projet no LB 5608/A, *document distribué lors du CIFED 2000 Lyon*, 2000.

[DRI 07] DRIRA F., Contribution à la restauration des images de documents anciens, thèse de doctorat, I.N.S.A de Lyon, 2007.

[GHA 06] KRICHA A., GHARDALLOU LASMAR A., ESSOUKRI BEN AMARA N., Une Technique de Prétraitement de Documents Anciens, JTEA, 2006.

[HEN 08] HENCHIRI F., KRICHA A., ESSOUKRI BEN AMARA N., Une approche de segmentation d'images composites issues de documents ancien, accepté JTEA, 2008.

[HIR 08] HIREMATH P.S., SHIVASHANKAR S., Wavelet based co-occurrence histogram features for texture classification with an application to script identification in a document image, *ELSEVIER*, 2008.

[KRI 06] KRICHA A., ESSOUKRI BEN AMARA N., Segmentation fond/texte des documents anciens, *GEI, 2006*.

[KRI 07] KRICHA A., ESSOUKRI BEN AMARA N., Contribution au tatouage de documents anciens, *JTEA*, 2007.

[LIK 04] LIKFORMAN-SULEM L., Apport du traitement des images à la numérisation des documents manuscrits anciens, *Numéro spécial de Document Numérique*, 2004.

[MAN 05] MANZANERA A., Cours de morphologie mathématique, *UPMC/ Master IAD*, 2005.

[PEC 01] PECAUD C., LYNCEE J.L., seuillage et segmentation d'images texturées, *DESS GIE*, 2001.

[RIV 08] RIVAS-ARAIZ E.A., MENDIOLA-SANTIBANEZ J.D., HERRERA-RUIZ G., Morphological multiscale fingerprints from connected transformations, *ELSEVIER*, 2007.