Extraction de zones informatives dans des images de formulaire en couleur

Maroua Hammami, Pierre Héroux, Sébastien Adam

Extraction de zones informatives dans des images de formulaire en couleur

Maroua HAMMAMI — Pierre HÉROUX — Sébastien ADAM

LITIS, Université de Rouen
UFR des Sciences F-76800 Saint Etienne du Rouvray
maroua.hammami1@univ-rouen.fr
pierre.heroux@univ-rouen.fr
sebastien.adam@litislab.eu

RÉSUMÉ. *Nous présentons, dans cet article, une approche permettant de localiser des zones informatives dans des documents couleur, par extraction de zones rectangulaires de couleur homogène. L'objectif de ce travail est d'obtenir des ancres pour le répérage et l'extraction d'information textuelle. L'approche proposée repose sur trois étapes. La première consiste à procéder à un filtrage de bruit, tout en évitant de créer de nouvelles couleurs ou de causer des distortions des contours. Ensuite, une quantificiation de couleur est réalisée, en combinant des plans de différents espaces couleurs, pour détecter les régions homogènes dans une image. Enfin, nous détectons les composantes connexes répondant à un critère de rectangularité. L'approche est évaluée sur 134 images de document, avec des métriques de la littérature. Nous montrons dans les résultats l'importance du choix des plans de couleur utilisés.*

ABSTRACT. *In this paper, we present a new approach of locating informative zones through extracting colorimetric features from a colored document images. The aim of this study is to obtain anchors which will allow spotting and retrieving textual information. The proposed approach is lying on three major steps. First step is to apply a noise filter while making sure that no new colors will emerge and edges will be preserved. Second is to detect homogenous regions within the image by quantifying colours through combining layers of different color spaces. Finally is to detect, from the segmented image, the components satisfying the rectangularity criterion. This approach has been evaluated over 134 document images dataset. The outcomes show that the approach is reliable and efficient as we combine layers from different color spaces.*

MOTS-CLÉS : *document coleur, zones rectangulaires, espace couleur.*

KEYWORDS: *colored document, rectangular zones, color space.*

1. Introduction

Dans les documents administratifs et commerciaux tels que ceux de la figure 1, la couleur est souvent exploitée par les concepteurs pour mettre en exergue les zones informatives du document (montant global d'une facture, zone à compléter d'un formulaire ...). Cette information colorimétrique, très utile aux lecteurs humains du document, s'avère donc importante pour les systèmes de lecture automatique.

Figure 1. *Exemples de documents administratifs et commerciaux*

Paradoxalement, rares sont les travaux dans la littérature qui traitent de l'analyse de documents en couleur, notamment des formulaires. Le plus souvent, l'information colorimétrique est transformée dès le début du processus, par la conversion du document en niveau de gris (Siyuan et Baird, 2011 ; Perroud *et al.*, 2001 ; Sobottka *et al.*, 1999), voire sa binarisation (Canedo-Rodríguez *et al.*, 2012 ; Liu *et al.*, 2005). Le but est alors de ne garder que le premier plan qui contient généralement les informations textuelles.

Cependant, la projection dans un espace de dimension plus restreinte engendre naturellement une perte d'information importante, qui peut être préjudiciable aux performances des traitements ultérieurs. Dans le projet qui fait l'objet de cet article, nous proposons d'exploiter davantage l'information colorimétrique lors du traitement de formulaires en couleurs.

Plus précisément, nous décrivons dans cet article le premier étage d'un système d'analyse dont le but est d'extraire de l'information dans des documents, en exploitant

la couleur. Dans ce cadre, l'objectif de l'approche décrite dans la suite est la détection de régions informatives matérialisées dans le document par des zones rectangulaires de couleur de fond homogène (Figure 2). De tels travaux permettent d'envisager différentes pistes d'amélioration pour les traitements à effectuer sur un document. Un premier exemple repose sur la possibilité d'effectuer une binarisation locale sur la zone informative, afin d'améliorer les performances des systèmes de lecture. Un autre exemple repose sur la possibilité de déduire de ces zones un modèle du document qui permettra, par exemple, d'envisager des tâches de classification basées sur la structure.

La suite de l'article est structurée de la façon suivante. L'approche proposée est détaillée dans la section 2. Puis, les résultats obtenus sur une base de documents réels sont présentés et discutés dans la section 3. Enfin, les conclusions et les nombreuses perspectives ouvertes par ce travail sont exposées dans la section 4.

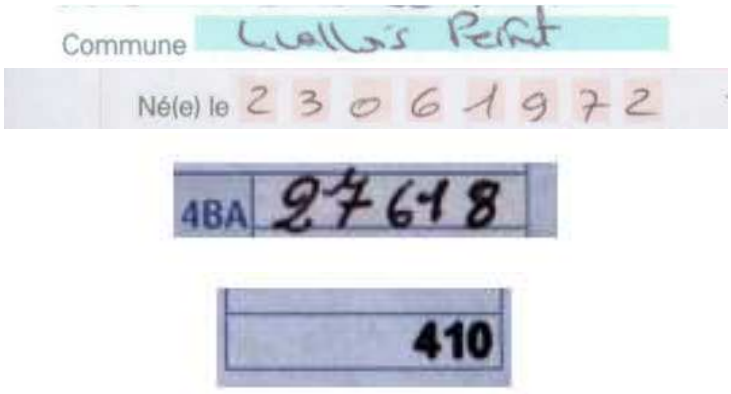

Figure 2. *Exemples de zones homogènes extraites des documents administratifs et commerciaux*

2. Approche proposée

L'approche proposée pour la détection de zones rectangulaires de couleur homogène repose sur 3 étapes. Dans un premier temps, un filtre de lissage est appliqué afin d'éliminer le bruit dû à l'impression et/ou à la numérisation. Puis, une quantification des couleurs est réalisée afin de trouver des régions homogènes présentes dans le document. Enfin, les composantes de couleur homogène de forme rectangulaire sont extraites. Chacune de ces étapes est décrite dans les sous-sections suivantes.

2.1. *Filtrage du bruit*

Afin de mieux éliminer les artefacts couleurs engendrés par l'impression et/ou la numérisation des documents en couleurs, il est nécessaire d'utiliser un filtre de

lissage de bruit. Dans ce cadre, nous appliquons le filtre de lissage EPSF proposé par Nikolaou et al. (Nikolaou et Papamarkos, 2009) qui a la particularité de préserver les contours. Il s'agit d'un filtre moyen adaptatif appliqué sur les images de document en couleur. Contrairement aux autres techniques de filtrage de bruit, EPSF réduit le bruit en le lissant par les pixels voisins sans engendrer de nouvelles couleurs comme peuvent le faire les filtres marginaux ni créer des distortions comme dans le cas des filtres vectoriels (Lambert et Montesinos, 1999). La technique consiste à appliquer un masque de dimension 3×3 sur chaque pixel de l'image. Les coefficients du masque sont donnés par l'équation suivante :

$$c_i = (1 - d_i)^p, where\ p \geq 1 \quad [1]$$

Ici, d_i est la distance de Manhattan entre le pixel central c du masque et le i^{me} pixel voisin. Sachant qu'un pixel est réprésenté par 3 valeurs R, G et B des plans de couleur de l'espace RGB, la distance d_i vaut :

$$d_i = \frac{|R_c - R_i| + |G_c - G_i| + |B_c - B_i|}{3 * 255} \quad [2]$$

Deux paramètres sont donc à fixer pour appliquer le filtre EPSF : l'exposant p dans l'équation 1 qui contrôle la "quantité" de lissage du bruit et le nombre d'itérations $iter$ pour l'application de l'EPSF.

2.2. *Quantification des couleurs*

Après l'application de l'EPSF, une classification des couleurs est réalisée pour quantifer l'image en K couches et trouver des régions de couleur. Nous utilisons l'algorithme $Kmeans$ pour cette tâche. Intuitivement, nous supposons que $K = 3$ au minimum, en supposant qu'un document contient au moins 3 couleurs : une pour l'arrière plan, une pour les zones informatives et une dernière pour le texte. La valeur de K peut être augmentée quand plus de 3 couleurs sont présentes. L'impact de la valeur de K est étudié dans les tests décrits dans la section suivante. Nous y étudions également différents choix d'espaces couleurs. La figure 3 illustre le résultat obtenu à l'issue de cette étape.

2.3. *Détection des rectangles*

Après la quantification par l'algorithme $Kmeans$, nous obtenons K couches de couleurs. Une extraction automatique des composantes connexes est alors réalisée sur chaque couche. Ensuite, les composantes sont filtrées sur un critère de rectangularité, défini comme suit :

$$\frac{S_c}{S_{BE}} \geq \theta \quad [3]$$

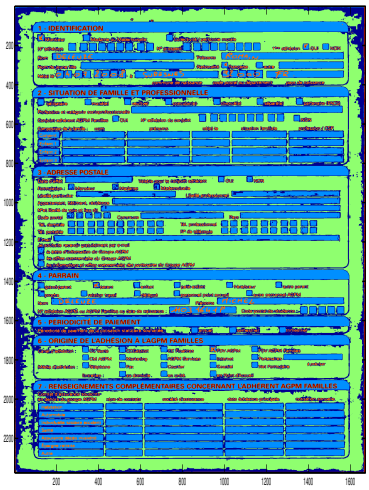

Figure 3. *Image quantifiée en prenant $K = 5$ et les plans de couleur LbC des espaces couleur Lab et LCH*

tel que S_c est la surface de la composante connexe ; le nombre de pixels qui construisent la composante, S_{BE} est la surface de la boite englobante dont la largeur et longueur sont respectivement la largeur maximale et longueur maximale de la composante connexe et $\theta \in]0, 1]$ est le seuil de rectangularité. Dans (Rosin, 1999), Rosin présente 4 méthodes pour mesurer la rectangularité et confirme que la méthode du rectangle englobant, que nous utilisons, est la meilleure pour filtrer les zones non rectangulaires. Dans nos tests, nous avons fixé le seuil θ à 70%. Nous présentons dans la figure 4 les rectangles obtenus de l'image quantifiée de la figure 3.

Dans la section suivante, nous décrivons les tests que nous avons réalisés et présentons les résultats obtenus.

3. Expériences et résultats

3.1. *Protocole et base de données*

La base de données utilisée pour les tests contient 134 documents administratifs et commerciaux en couleur répartis en 9 classes différentes en fonction du modèle de formulaire. Nous avons généré manuellement la vérité terrain contenant les régions homogènes. En moyenne, 73 zones sont identifiées par document. Certaines sont remplies alors que d'autres ne le sont pas. Dans le tableau 1, nous présentons, pour chaque

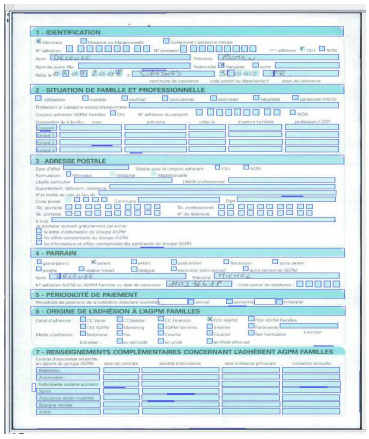

Figure 4. *Détection des rectangles*

classe, le nombre d'images ainsi que le nombre de zones identifiées dans la vérité terrain.

Classes	Nbr images	Nbr zones
1	4	162
2	11	80
4	31	59
5	29	231
6	14	21
7	12	14
8	11	166
9	9	74
12	13	80
Total	134	moy : 73

Tableau 1. *Nombre de zones homogènes dans chaque classe*

Afin d'extraire le maximum des zones homogènes, les différents paramètres utilisés par l'approche doivent être réglés. Dans cet article, nous étudions l'impact du nombre de couches $K \in [2, 8]$ et du choix de plans couleurs. Nous comparons 4 espaces couleur (RGB, HSV, LCH, Lab), sur lesquels nous appliquons l'algorithme du $Kmeans$. Nous avons également combiné différents plans issus d'espaces couleur différents, en faisant varier le nombre de plans de 1 à 3 couches. Les paramètres du filtre EPSF ainsi que le seuil de rectangularité ont été fixé de façon empirique à des valeurs par défaut ($p = 10$, $iter = 5$, $\theta = 0.7$).

3.2. *Résultats*

Dans cette section, nous présentons les résultats que nous avons obtenus. Dans les figures 6-7, nous exposons quelques résultats de différentes configurations. Les zones entourées en rouge sont celles extraites en utilisant notre approche.

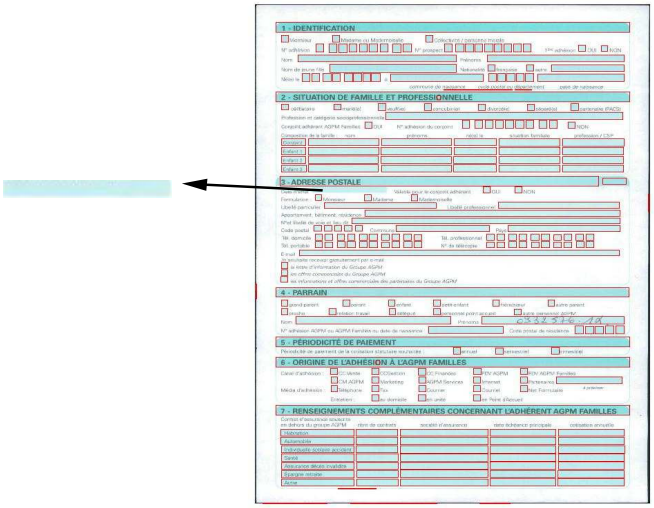

Figure 5. *Utilisation du plan de couleur C de l'espace couleur LCH et $K = 3$*

Les résultats obtenus montrent que l'approche utilisée permet l'extraction d'un nombre important de zones homogènes. Cependant, différents types d'erreur apparaissent. Certaines zones sont manquantes comme dans la figure 5, alors que d'autres ont été divisées comme dans le figure 6. Finalement, le dernier type d'erreur est dû à une sur-segmentation comme dans l'exemple de la figure 7. Afin de quantifier ces erreurs, nous avons utilisé la métrique proposée par le LNE [1] dans le cadre du projet MAURDOR [2]. Cette métrique, nommée ZoneMap, calcule une fonction d'erreur pour chaque configuration présente ci-dessous et illustrée dans la figure 8. Une zone hypothèse correspond à une zone trouvée par notre extracteur et une zone référence est une zone définie dans la vérité terrain :

1) $FalseAlarm$: Quand le rectangle apparaît dans la zone hypothèse et pas dans la zone référence.

2) $Miss$: Quand le rectangle n'apparaît que dans la zone référence.

3) $Split$: Quand la zone référence est représentée par au moins 2 zones hypothèses.

1. Laboratoire National d'Essai
2. url :http ://www.maurdor-campaign.org/

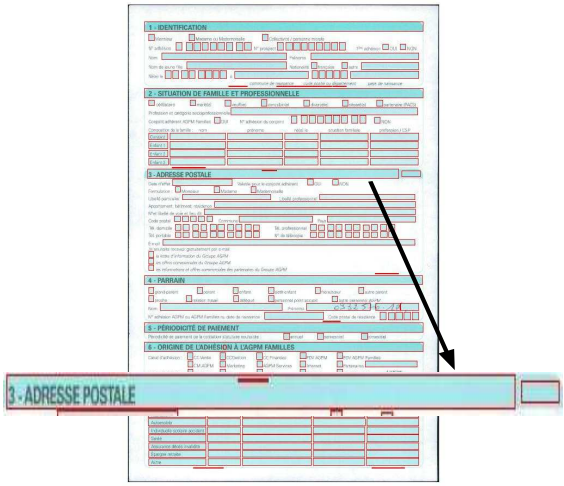

Figure 6. *Utilisation des plans de couleur* a *et* b *de l'espace couleur* Lab *et* $K = 3$

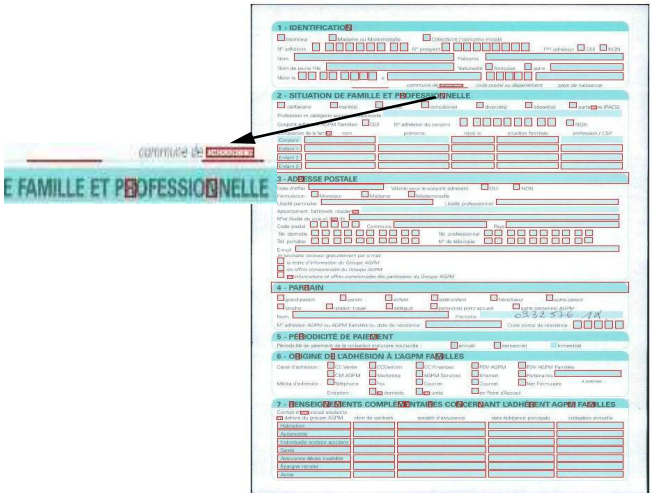

Figure 7. *Utilisation des plans de couleur de l'espace couleur* RGB *et* $K = 3$

4) $Merge$: Quand au moins 2 zones référence correspondent à une zone hypothèse.

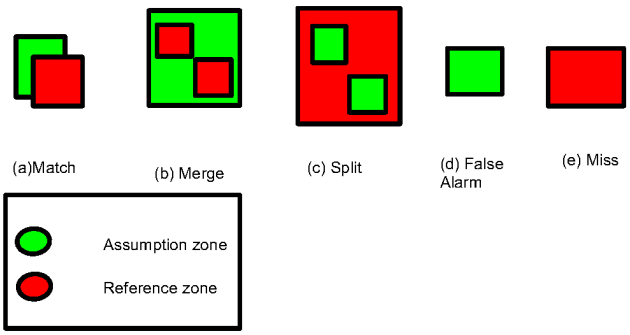

Figure 8. *Les différentes configurations de la métrique MAURDOR*

5) $Match$: Quand une zone référence est appariée à une zone hypothèse mais qu'elles ne sont pas totalement confondues.

En faisant varier les paramètres tels que définis dans la section 3.1, la métrique MAURDOR montre que notre approche donne une erreur minimale sur notre base de document pour une valeur $K = 5$ et en combinant les plans de couleur L,b et C des espaces couleurs Lab, LCH. Sur la figure 9, nous confrontons les résulats de la configuration retenue aux résultats de l'approche appliquée aux espaces couleur initiaux (RGB, HSV, LCH, Lab) ainsi qu'aux niveaux de gris. Nous remarquons que l'espace identifié donne l'erreur minimale qui est à 0.4826.

La figure 10 montre en détail les erreurs par type de configuration. Nous remarquons que contrairement à l'espace couleur HSV, notre nouvel espace LbC fait moins d'erreur de type $Miss$ et $FalseAlarm$, tandis qu'en utilisant le niveau de gris aucune zone n'est détéctée, ce qui explique la valeur 1 de l'erreur $Miss$. Ceci confirme l'hypothèse que la réduction en niveau de gris engendre une perte d'information préjudiciable aux traitements.

De la même manière, nous avons fait varier la valeur de K pour notre nouvel espace LbC. Dans la figure, nous montrons que pour ce nouvel espace, $K = 5$ est la meilleure configuration.

Afin de valider les résultats obtenus, nous avons calculé l'erreur en fonction de l'indice de Jaccard. Si A est l'ensemble des zones de référence et B est l'ensemble des zones hypothèse définies, l'erreur est définie comme suit :

$$Erreur_{Jaccard} = 1 - \frac{|A \cap B|}{|A \cup B|} \quad [4]$$

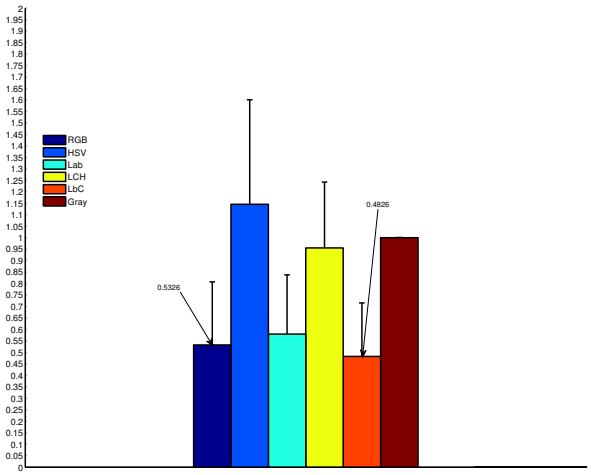

Figure 9. *Erreur totale selon la métrique MAURDOR*

L'intérêt d'utiliser l'indice de Jaccard est principalement de mesurer la qualité de notre extracteur. En effet, si la valeur de $Erreur_{Jaccard}$ est grande, celà peut être dû à 2 raisons : soit l'extracteur ne permet pas de trouver les bonnes zones, soit plusieurs fausses détections de zones ont été opérées. Inversement, si la valeur de $Erreur_{Jaccard}$ est petite, l'extracteur permet d'extraire les bonnes zones et fait moins d'erreur de fausse alarme.
Les résulats obtenus confirment ceux de la métrique MAURDOR. Nous obtenons la meilleure configuration avec $K = 5$ et les plans de couleurs L,b et C des espaces couleurs Lab et LCH. De la même manière que pour la métrique de MAURDOR, nous présentons les résultats obtenus en calculant l'erreur de l'indice de Jaccard. Dans les figures 12 et 13, notre nouvel espace est le plus performant.

4. Conclusion

Dans ce article, nous avons présenté une approche permettant d'extraire les zones homogènes de forme rectangulaire ainsi que leurs caratctéristiques colorimétriques dans des images de documents. Un filtrage est tout d'abord appliqué afin d'éliminer le bruit de numérisation et d'impression. Ensuite, nous poursuivons par une quantification des couleurs afin de trouver des régions homogènes. Enfin, une extraction des composantes connexes qui répondent à un critère de rectangularité est effectuée afin d'élimener les zones non rectangulaires. Cette approche permet de détecter au mieux

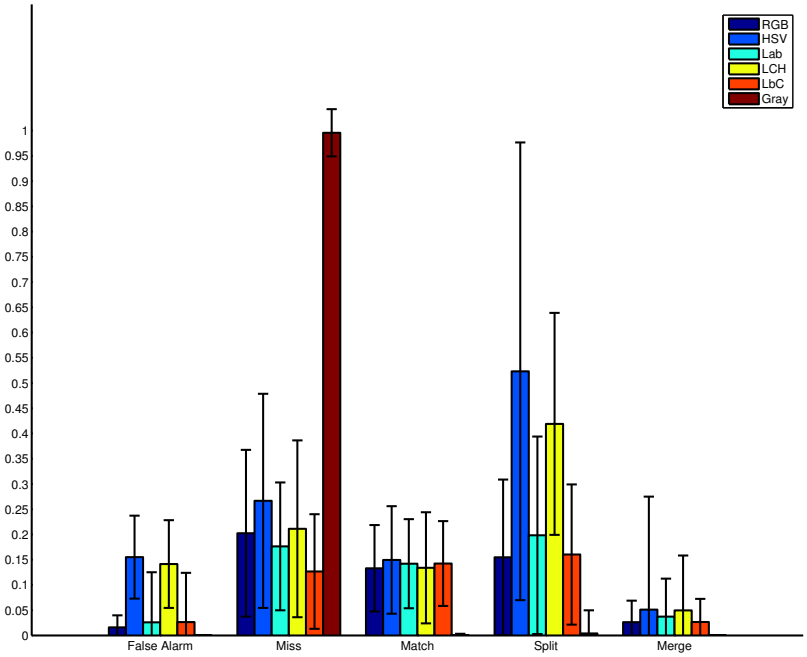

Figure 10. *Erreur MAURDOR par type de configuration*

les zones homogènes dans un document, et les résultats montrent que le choix d'un espace couleur spécifique permet d'améliorer sensiblement les résultats par rapport à l'utilisation de niveaux de gris. L'avantage de cette approche est qu'elle permet de délimiter les zones contenant l'information textuelle pour pouvoir ensuite, par exemple, binariser localement les zones informatives (voir figure 14).
En guise de perspectives, dans notre approche, plusieurs paramètres doivent être définis, certains ont été fixés de façon empirique tandis que d'autres ont été apris sur notre base de documents. Nous jugeons que le paramètre K de l'algorithme $Kmeans$ dépend plus du type et des couleurs présentes dans le document. Pour celà, il est intéressant que le nombre de couches soit variable. Nous envisageons dans ce contexte d'appliquer une nouvelle approche de classification des couleurs. De plus la métrique de MAURDOR nous permet de quantifier l'erreur d'extraction par type de configuration, cependant elle ne nous permet pas d'avoir une idée globale sur l'ensemble de configuration. Nous proposons donc d'adapter la métrique de MAUDROR de façon à avoir une quantification globale des types de configuration afin de mieux évaluer notre approche.

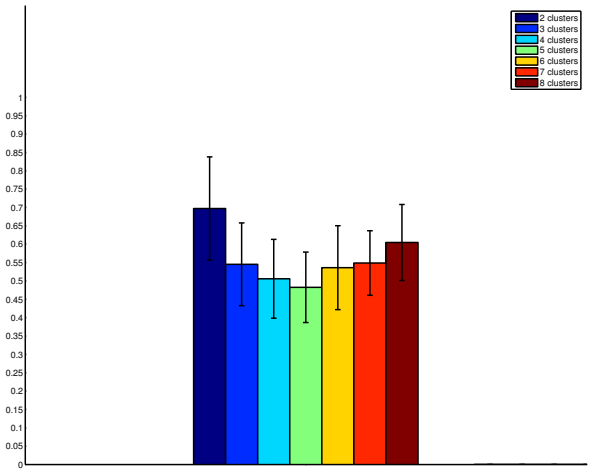

Figure 11. *Erreur selon la métrique MAURDOR pour différentes valeurs de K et le nouvel espace LbC*

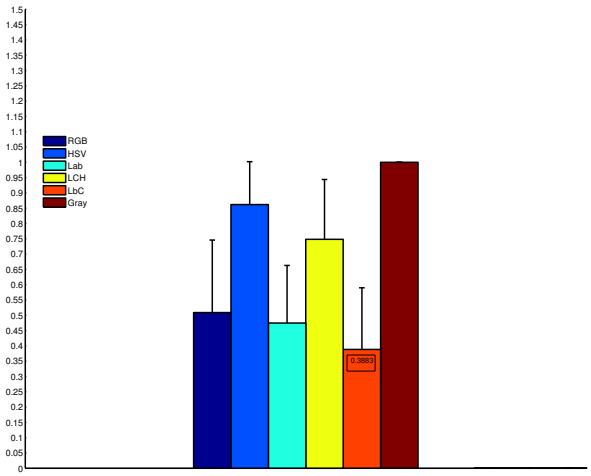

Figure 12. *Comparaison de l'erreur de Jaccard pour les différents espaces couleur à $K = 5$*

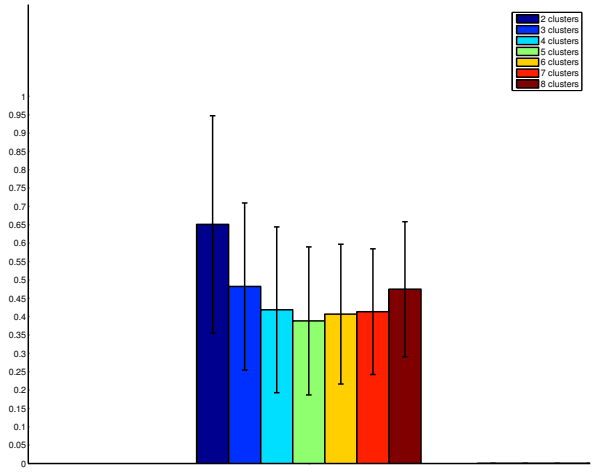

Figure 13. *Comparaison de l'erreur de Jaccard des différentes valeurs de K pour le nouvel espace couleur*

5. Bibliographie

Canedo-Rodríguez A., Hyoun K. J., Soo-Hyung K., John K., Hee K. J., Sun Y., Kiran V. S., Yolanda B., « Efficient Text Extraction Algorithm Using Color Clustering for Language Translation in Mobile Phone », *Signal and Information Processing*, vol. 3, n° 2, p. 228-237, 2012.

Lambert P., Montesinos P., « Pré-traitement des images couleur », p. 1-5, 1999.

Liu C., Wang C., Dai R., « Text Detection in Images Based on Color Texture Features », p. 40-48, 2005.

Nikolaou N., Papamarkos N., « Color reduction for complex document images », *International Journal of Imaging Systems and Technology*, vol. 19, n° 1, p. 14-26, 2009.

Perroud T., Sobottka K., Bunke H., « Text Extraction from color documents clustering approaches in three and four dimensions », p. 937-941, 2001.

Rosin P., « Measuring rectangularity », *Machine Vision and Applications*, vol. 11, n° 4, p. 191-196, 1999.

Siyuan C., Baird H., « Pixel accurate document image content extraction », *Proceedings of the 2011 ACM Symposium on Applied Computing*, vol. , n° 1, p. 245, 2011.

Sobottka K., Bunke H., Kronenberg H., « Identification of Text on Colored Book and Journal Covers », *International Conference on Document Analysis and Recognition ICDAR*, vol. , p. 57-63, 1999.

(a)

(b)

(c)

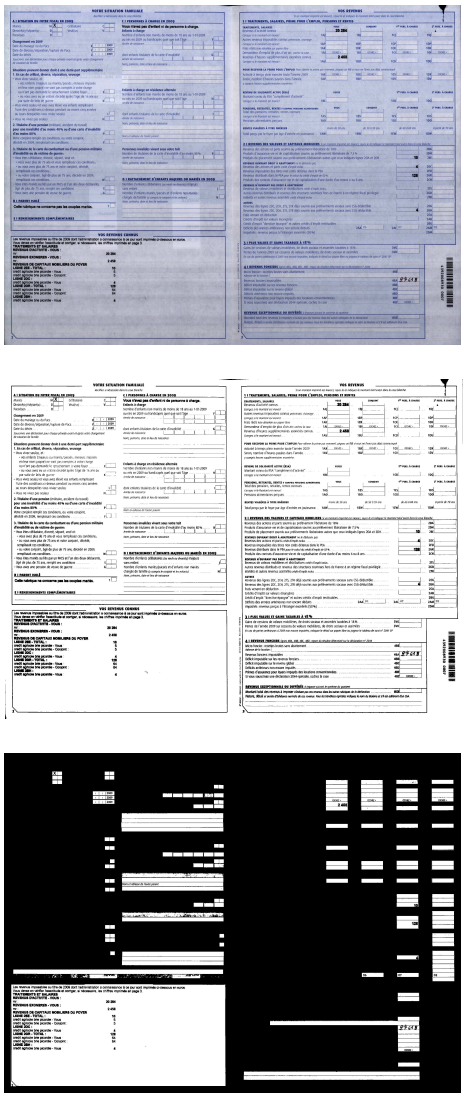

Figure 14. *Comparaison entre binarisation globale et locale ; (a) image initiale (b) binarisation globale (c) binarisation locale*