

Extraction de la structure physique de courrier : Application dédiée à un système automatique de tri postal

GACEB Djamel , EGLIN Véronique , LEBOURGEOIS Frank , EMPTOZ Hubert
Bât. Jules Verne, 20, av. Albert Einstein, 69621 Villeurbanne, France.

djamel.gaceb1@insa-lyon.fr

veronique.eglin@insa-lyon.fr

flebourg@insa-lyon.fr

hubert.emptoz@insa-lyon.fr

Résumé. Les systèmes de tri de la poste diffusent chaque jour plusieurs tonnes de courriers postaux possédant des structures hétérogènes. Statistiquement, on constate que la cause principale de rejet de courriers est liée à l'échec de la tâche de localisation du bloc-adresse, et plus particulièrement, à la phase d'extraction de la structure physique. Les méthodes de segmentation ascendantes et descendantes apportent des connaissances différentes qu'il ne faut pas négliger lorsque l'on désire augmenter la robustesse. Contrairement aux méthodes classiques qui comportent plusieurs sources d'erreurs, les méthodes mixtes rassemblent les deux stratégies. Leur point fort repose sur le fait qu'elles se servent des avantages de l'une pour combler les inconvénients de l'autre. Partant de ce principe, notre proposition consiste à utiliser une stratégie de segmentation mixte plus adaptée aux courriers postaux. Les étapes de haut niveau reposent en partie sur la coloration hiérarchique des graphes. Elles permettent de synthétiser, par l'intermédiaire d'une organisation pyramidale des données, la gestion des règles composées régissant la décomposition en composantes connexes des zones d'intérêt. L'évaluation des performances de notre approche a été réalisée sur un corpus de 10000 images d'enveloppes (jugées difficiles). Les temps de traitement et le taux de rejet ont été considérablement réduits.

Mots-clefs: Extraction de la structure physique, Binarisation, coloration des graphes, traitement en temps réel, tri de courriers

Introduction

Les machines de tri automatique de courriers traitent jusqu'à 17 plis par seconde, ce qui nécessite que le bloc adresse soit rapidement et précisément identifiable par l'OCR, et que les lignes d'adresse soient correctement organisées [1][2][3]. Après l'acquisition de l'image de l'enveloppe par une caméra CCD, trois modules principaux participent à la tâche de localisation du bloc adresse : segmentation de la structure physique de l'image d'enveloppe, extraction des caractéristiques et interprétation des blocs (voir Figure 1). La phase de la segmentation de la structure physique a un grand impact sur la performance de système de tri en entier. D'une manière générale, elle désigne la séparation de l'image de l'enveloppe en ses éléments constitutifs contenant des données homogènes afin de les identifier isolément. Ces éléments sont souvent espacés et forment des blocs géométriques élémentaires, à base de rectangle dans la grande majorité des cas. La définition de la segmentation est donc assez proche du sens littéral du mot analyse. On parle de sur-segmentation lorsque l'élément constitutif est lui-même fragmenté, et de sous-segmentation lorsque plusieurs éléments constitutifs n'ont pas pu être isolés.

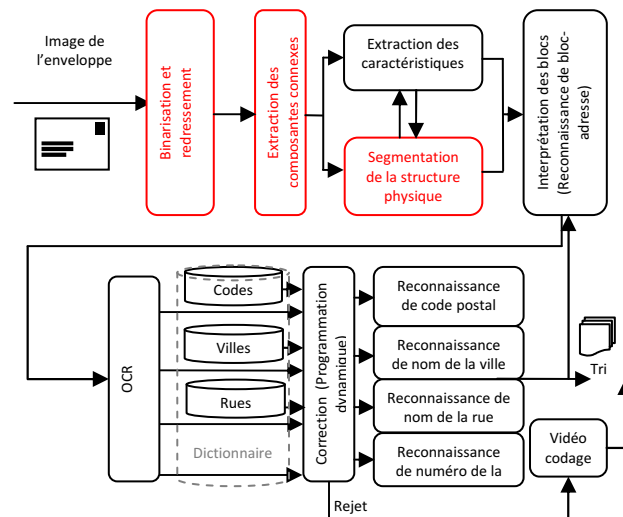


Figure 1 : Architecture modulaire de système de tri, la phase de segmentation de la structure physique est représentée en rouge.

Du point de vue de la robustesse, on remarque que les techniques de segmentation classiques se heurtent en général à plusieurs contraintes (Figure 2):

- Images dégradées (pliage),
- Très grande variété de courriers (qualité, couleur et texture de papier différent),
- Contraintes de temps réel (temps de traitement limité),
- Inclinaisons variables des zones textuelles sur l'enveloppe,
- Espacement non uniforme entre les caractères, les lignes et les blocs de texte,
- Obligation de résultats,
- Résolutions spatiales des images élevées (300 dpi),
- Présence d'éléments parasites autour du bloc-adresse (logos, annotations,...),
- Superposition de couches d'informations (tampon, notes manuscrites, ...).



Figure 2 : Très grande variété de courriers

Compte tenu de ces limites, nous proposons dans cet article une nouvelle méthode d'extraction de la structure physique appliquée aux images de courriers postaux. L'originalité de la technique que nous décrivons ici repose sur une représentation multi-résolution, faisant largement appel à l'utilisation de la théorie des graphes.

Notre objectif consiste à augmenter les performances de chaque étape de segmentation et sa cohérence avec les autres étapes pour réduire au maximum le rejet des courriers et les temps de traitement. Cette méthode a été développée pour être intégrée à un système de tri automatique de courriers postaux.

Le reste de cet article est organisé comme suit:

La deuxième partie cite les différentes méthodes de segmentation existantes. Nous y présentons les travaux précédents et les limites fixées. Dans une troisième partie, nous présentons les aspects formels de la coloration de graphes. Nous décrivons ainsi, dans la quatrième partie, l'application de la coloration au problème de segmentation. Les résultats obtenus sont alors commentés.

1. Différentes stratégies de segmentation

La littérature se réfère généralement à trois stratégies de segmentation. Les méthodes de segmentation ascendantes et descendantes apportent des connaissances différentes qu'il ne faut pas négliger lorsque l'on désire augmenter la robustesse. Contrairement aux méthodes classiques qui comportent plusieurs sources d'erreurs, les méthodes mixtes rassemblent les deux stratégies en même temps. Leur point fort repose sur le fait qu'elles se servent des avantages de l'une pour combler les inconvénients de l'autre [4]. Ces quelques remarques nous permettent de dresser une liste de différents types d'erreurs dont il est indispensable de tenir compte lors de la conception d'algorithmes d'extraction de la structure physique :

- Fusion horizontale et/ou verticale de lignes ou de blocs de texte,
- Découpage horizontal et/ou vertical de lignes ou de blocs de texte,
- Fusion et/ou confusion de texte avec des composantes graphiques ou parasites,
- Non détection de blocs ou de lignes de texte

Les méthodes de segmentation par fusion (regroupement progressif des composantes connexes, RLSA, segmentation par changement d'échelle, méthode de gradients cumulés) sont plus utilisées par les stratégies ascendantes (bottom-up) [5] [6], alors que les méthodes de segmentation par découpage (projection de profils, segmentation par l'analyse des espaces, transformée de Hough) sont plus adaptées aux stratégies descendantes (top-down) [7]. D'autres méthodes dites mixtes profitent des deux stratégies à la fois [7].

Déforges et Barba dans [6] ont présenté une méthode ascendante générique basée sur une description en multi-résolution de l'image de document utilisée pour localiser le bloc-adresse. Une structure presque similaire a été utilisée par Wang [8] pour distinguer les blocs de textes des blocs de graphiques, et les représenter dans un modèle structurel. Shi et Govindaraju dans [7] ont proposé un algorithme basé sur l'application de « fuzzy directional runlength ».

Les méthodes de segmentation de la structure physique que nous avons citées, utilisent toutes des structures de données complexes. La gestion des critères et des connaissances devient difficilement contrôlable face à la grande variabilité sur les enveloppes à trier. Partant de ces constatations, notre proposition consiste à utiliser une stratégie de segmentation mixte plus adaptée aux courriers postaux. Les étapes de haut niveau reposent en partie sur la coloration hiérarchique des graphes, permettant de synthétiser, par l'intermédiaire d'une organisation pyramidale des données, la gestion de règles composées régissant l'interprétation de la

décomposition en composantes connexes des zones d'intérêt. A ce jour, aucun travail dans ce domaine ne s'est servi de la puissance de cet outil.

2. Aspects formels de la coloration des graphes

Différents problèmes pratiques de segmentation peuvent être modélisés par la coloration des graphes. La forme générale de ces applications nécessite la formation d'un graphe par les nœuds (sommets) qui représentent les objets d'intérêt et les arcs qui définissent les relations entre ces objets.

Une question naturelle est alors de déterminer quel est le nombre minimum de classes nécessaire pour décomposer un ensemble d'objets (composantes connexes, lignes, blocs) en plusieurs sous ensembles homogènes. Cette question peut se formuler en termes de coloration de graphe. On veut par exemple regrouper un ensemble des composantes connexes en plusieurs lignes de texte homogènes sans savoir a priori leur nombre. Pour cela, il suffit de représenter chaque composante $c(i)$ par un sommet v_i et d'ajouter une arête $E(v_i, v_j)$ entre chaque paire de composantes suffisamment différents. Le graphe fini $G = (V, E)$ est défini par l'ensemble fini $V = \{v_1, v_2, \dots, v_n\}$ ($|V| = n$) dont les éléments sont appelés sommets, et par l'ensemble fini $E = \{e_1, e_2, \dots, e_m\}$ ($|E| = m$) dont les éléments sont appelés arêtes. La coloration des sommets du graphe $G(V, E)$ consiste à affecter à tous ses sommets une couleur de telle sorte que deux sommets adjacents (dissemblables) ne portent pas la même couleur. Ces couleurs vont correspondre aux différentes lignes de texte. Une coloration avec k couleurs est donc une partition de l'ensemble des sommets en k sous-ensembles homogènes. Le nombre de couleurs utilisées pour colorer le graphe G de n sommets est appelé nombre chromatique $\chi(G) \leq n$ qui représente le plus petit entier k pour lequel il existe une partition de V en k sous-ensembles homogènes [9].

Sur le graphe G de la figure3, dont l'ensemble de 11 formes différentes V représenté par les sommets $\{x_1, \dots, x_{11}\}$, on a eu besoin de quatre couleurs pour colorer les 11 sommets de sorte que deux sommets adjacents ont des couleurs différentes.

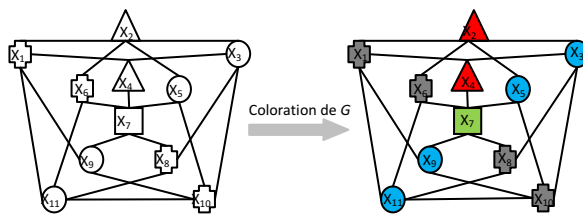


Figure 3 : Coloration de graphe G de 11 sommets par 4 couleurs (▲, ●, ■ et +).

La matrice d'adjacence Ma de G est la matrice carrée symétrique 11×11 définie comme suit :

$$Ma = \begin{matrix} & \begin{matrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 & x_9 & x_{10} & x_{11} \end{matrix} \\ \begin{matrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \\ x_9 \\ x_{10} \\ x_{11} \end{matrix} & \begin{pmatrix} & & & & & & & & & & \\ & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ & & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ & & & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ & & & & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ & & & & & 0 & 1 & 0 & 0 & 1 & 0 \\ & & & & & & 1 & 0 & 0 & 0 & 1 \\ & & & & & & & 1 & 1 & 0 & 0 \\ & & & & & & & & 0 & 0 & 1 \\ & & & & & & & & & 1 & 0 \\ & & & & & & & & & & 1 \\ & & & & & & & & & & & 1 \end{pmatrix} \end{matrix}$$

Une consultation complète de Ma prend un temps : $\Gamma = O[0.5 \times (|V| \times |V|) - |V|]$

La plupart des évaluations de $\chi(G)$ proviennent d'algorithmes de coloriage. Il en existe beaucoup, et pour ne pas nous noyer avec cette question, nous nous limiterons à citer l'étude comparative effectuée par Paschos [9]. Dans notre étude, nous sommes particulièrement intéressés à l'algorithme distribué de coloration proposé par Effantin et Kheddouci dans [10], [11].

3. Application de la coloration à notre problème

L'objectif d'une technique de segmentation repose sur sa stratégie de décision qui définit la meilleure façon d'isoler un bloc afin qu'il soit reconnu comme tel par le module de reconnaissance du bloc-adresse. Les techniques d'extraction de la structure physique que nous avons étudiées ne peuvent pas produire de blocs assez homogènes sur des enveloppes dites difficiles. Par conséquent, les connaissances livrées par les descripteurs des blocs hétérogènes (contenant des éléments parasites) ne peuvent pas être suffisamment discriminantes.

Le schéma de la figure 5 représente les différentes étapes de notre méthode d'extraction de la structure physique.

Afin d'améliorer la robustesse et la précision de la segmentation, il a fallu choisir un outil encore plus évolué. L'idée est d'utiliser une stratégie mixte de segmentation faisant largement appel à toute la richesse de la structure pyramidale. Notre méthode portera principalement sur la puissance de la méthode de coloration des graphes pour regrouper les composantes connexes en lignes de texte puis les lignes en blocs homogènes.

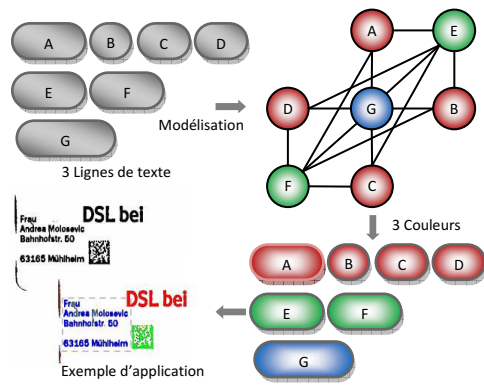


Figure 4 : Exemple d'application de la coloration des graphes.

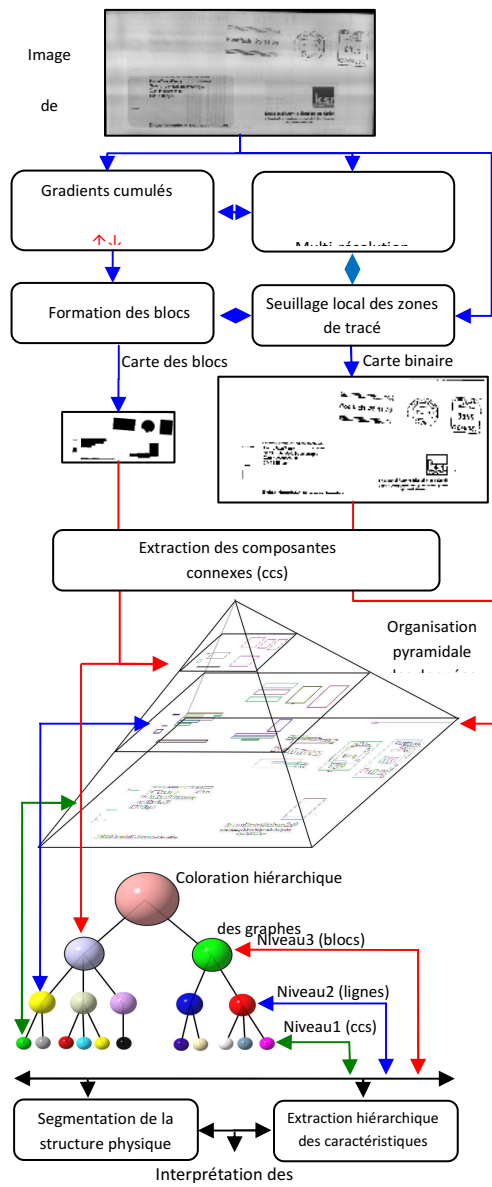


Figure 5 : Diagramme fonctionnel de la méthode

3.1. Binarisation et détection de composantes connexes

La séparation entre l'étape de binarisation et celle de localisation des zones de texte augmente le temps du calcul et conduit à une sur-segmentation du bruit et de la texture de papier sur des zones vides de l'image. En effet, aucune des méthodes classiques (globale [12] ou locale [13]) ne remplit efficacement toutes les conditions imposées. Nous avons pu optimiser cette étape en appliquant un seuillage local uniquement à proximité des zones de texte (Figure 6) que nous avons localisées par la méthode des gradients cumulés avec la multirésolution et la morphologie mathématique [14].

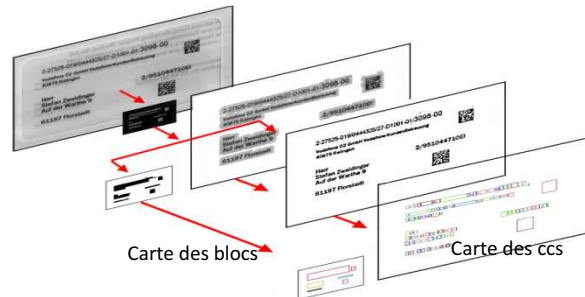


Figure 6 : Seuillage mixte et détection des composantes connexes.

Nous détectons ainsi les composantes connexes (CCs) de la carte des blocs qui sont utilisées ensuite pour guider l'extraction des CCs de premier plan [15]. La méthode utilisée est inspirée des études de Pavlidis sur la structure LAG (Line Adjacency Graph) [16]. Elle consiste à mettre en relation les séquences de pixels noirs de deux lignes consécutives de la fenêtre d'analyse.

Nous allons utiliser dans ce qui va suivre les notations suivantes : Les ensembles

$$C = \{c_i(x_i^d, y_i^d, x_i^f, y_i^f)\}_{i=0 \dots N_c},$$

$L = \{l_i(x_i^d, y_i^d, x_i^f, y_i^f)\}_{i=0 \dots N}$ et $B = \{b_i(x_i^d, y_i^d, x_i^f, y_i^f)\}_{i=0 \dots N_b}$, indiquent les listes des CCs de la carte binaire respectivement de la carte des lignes et de la carte des blocs.

$(x_i^d, y_i^d, x_i^f, y_i^f) \mid x_i^d < x_i^f, y_i^d < y_i^f$ sont les coordonnées des boîtes circonscrites. Le premier niveau de la pyramide est ainsi formé par les éléments de l'ensemble C , et le dernier par les éléments de l'ensemble B . N_c et N_b représentent les nombres des CCs à chaque niveau.

3.2. Extraction de la structure physique

Soit G un graphe non orienté à trois niveaux hiérarchiques, indépendants mais cohérents défini par la relation suivante: $G(V, E) = \bigcup_{k=1}^3 G_k(V_k, E_{k>Sk})$, tel que:

$$V_k = \begin{cases} V_1 = \{v_i^1 \Leftrightarrow \text{Descripteur}[c_i(x_i^d, y_i^d, x_i^f, y_i^f)]\}_{i=1..Nc} & \text{si } k=1 \\ V_2 = \{v_i^2 \Leftrightarrow \text{Descripteur}[l_i(x_i^d, y_i^d, x_i^f, y_i^f)]\}_{i=1..Nl} & \text{si } k=2 \\ V_3 = \{v_i^3 \Leftrightarrow \text{Descripteur}[b_i(x_i^d, y_i^d, x_i^f, y_i^f)]\}_{i=1..Nb} & \text{si } k=3 \end{cases} \quad (1)$$

Avec $V_{k=1,2 \text{ et } 3}$ sont les ensembles finis des sommets des graphes G_k . Ces sommets sont représentés par les descripteurs des éléments constitutifs de chaque niveau k de la pyramide (figureure4). Les caractéristiques sont ensuite sélectionnées de telle façon qu'elles puissent mettre en évidence les différences entre les composantes. $E_{k>Sk}$ est l'ensemble fini des arêtes représenté par les paires de sommets adjacents. Deux sommets sont donc adjacents si et seulement si la distance entre leurs descripteurs $d(v_i, v_j)$ est strictement supérieure à un seuil S (le mécanisme d'optimisation des seuils S_k est détaillé dans [3]). Cette définition peut être illustrée par la relation suivante:

$$E_{k>Sk}[v_i^k, v_j^k] = \begin{cases} 1 & \text{si } d(v_i^k, v_j^k) > s_k \\ 0 & \text{sinon} \end{cases} \quad (2)$$

La coloration hiérarchique de graphe G est utilisée ici pour décomposer l'ensemble des sommets de chaque niveau k en sous-ensembles homogènes. Pour cela, elle se focalise principalement sur les dissemblances des éléments constitutifs (représentés par des sommets) de même niveau de la pyramide de données. Elle utilise, de ce fait, une stratégie mixte de progression dans la hiérarchie, avec laquelle les couleurs d'un niveau participent à la formation et la description des sommets d'un autre niveau (Figureure 7).

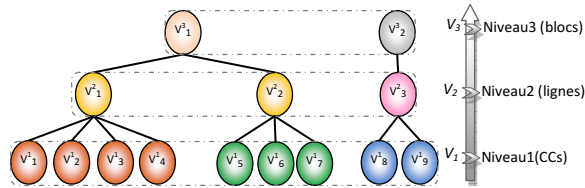


Figure7 : Progression de la coloration de graphe dans la hiérarchie.

Nous pouvons résumer les étapes de coloration d'un graphe $G_{k=1,2 \text{ et } 3}$ par l'algorithme suivant [11]:

Algorithme: Coloration (G_k)

Début

Si $col_k(i) \neq \emptyset$ Alors

Soit $M = N_{col}^k(i) \cup \{col_k(i)\}$;

$q = 0$;

Pour chaque sommet $j \in N_{adj}^k(i)$

Sachant que $col_k(j) := \emptyset$ Faire

$q = \min \{r | r > q, r \notin M \text{ Et } r \notin col_k(j)\}$;

Si $q \leq \Delta + 1$ Alors $col_k(j) := q$;

Sinon $col_k(j) := \min \{r | r \notin N_{col}^k(j)\}$;

FinSi

FinFaire

FinSi

Fin.

avec $col_k(i)$ la couleur de sommet $i \in V_k$, $N_{adj}^k(i)$ l'ensemble de sommets adjacents au sommet i . $N_{col}(i)$ est l'ensemble des couleurs des sommets $N_{adj}^k(i)$, $d_{eg}^k(i) = |N_{adj}^k(i)|$ son degré, et :

$$\Delta_k = \max \{d_{eg}^k(i) | i \in V_k\} \quad (3)$$

Notre méthode de segmentation de la structure physique se base, au départ, sur la construction de G_1 qui représente les CCs de premier niveau de la pyramide (Formule1). Cette construction doit être guidée par la carte initiale des blocs pour augmenter la vitesse de coloration. Nous colorons G_1 par la fonction $Coloration(G_1)$, et nous préservons que les sommets de G_1 qui appartient à l'ensemble des couleurs textuelles puis nous colorons à nouveau $G_1(V_1, E_{1>S1})$. Les couleurs obtenues, représentent les lignes de texte et forment le graphe $G_2(V_2, E_{2>S2})$. Une coloration de G_2 est ainsi appliquée pour former $G_3(V_3, E_{3>S3})$ qui représente les blocs textuels. Une dernière coloration de G_3 est nécessaire, conséquemment, pour fusionner les blocs qui répondent aux mêmes critères d'alignement et construire une carte finale de blocs homogènes.

Les différentes étapes de segmentation sont ainsi illustrées par l'exemple de la figure suivante :

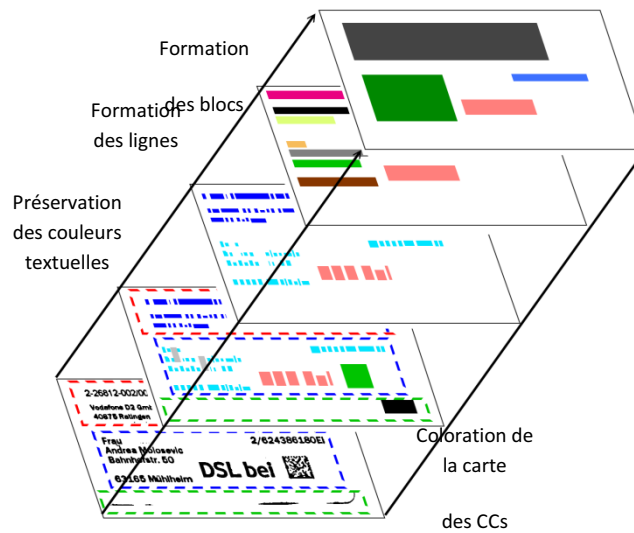


Figure 8 : Exemple de différentes phases de la coloration.

4. Expérimentation

L'évaluation des performances de notre approche a été réalisée sur un corpus de 10000 images d'enveloppes (jugées complexes), plus de 95 % de blocs-adresse ont été correctement segmentés par notre méthode, contre 60% par la méthode RLSA et 30% par la méthode de projection des profils. L'analyse des résultats montre que la coloration a réduit considérablement les erreurs de sur-segmentation et de sous-segmentation introduits par les méthodes RLSA et projection des profils (Figure 9 et Figure 10). Ces performances peuvent être justifiées par la puissance de notre méthode d'isoler les éléments textuels (caractères, lignes, blocs) et rejeter les éléments parasites.

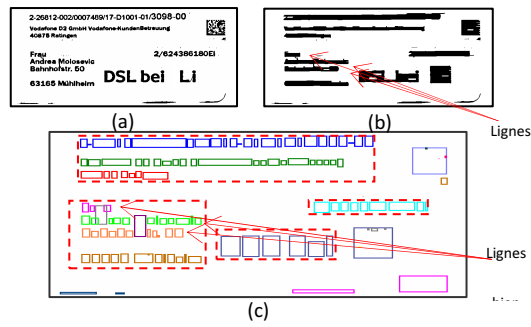


Figure 9 : (a) Image binaire, (b) émergence des lignes par la méthode RLSA, (c) Perception des lignes par la coloration.

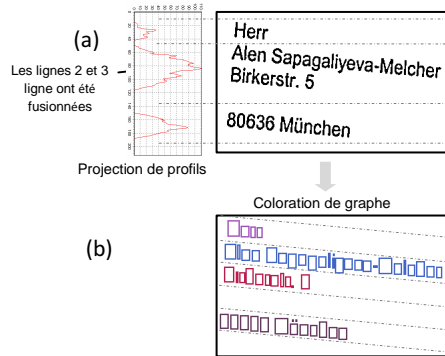


Figure 10 : Segmentation en lignes de texte, (a) méthode de projection des profils, (b) coloration hiérarchique.

L'augmentation de la cohérence entre les différentes phases de segmentation a conduit également à une réduction considérable du temps de traitement. L'implantation des algorithmes a été étudiée en même temps. Les courbes de la figure suivante montrent la faisabilité d'une solution temps réel.

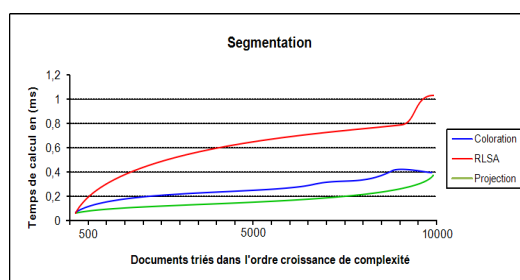


Figure 10 : Comparatif des temps de calcul de différentes méthodes de segmentation.

Conclusion

Nous avons présenté une nouvelle méthode de segmentation de la structure physique basée sur la coloration hiérarchique des graphes. Notre méthode utilise une stratégie mixte de segmentation et une architecture pyramidale des données. Nous avons vu que la coloration hiérarchique de graphe montre une grande robustesse aux composantes parasites considérées comme facteurs d'erreur des méthodes classiques de segmentation. Grâce au nombre restreint de règles dont elle dispose, cette nouvelle technique répond à une large variabilité de courriers postaux. De plus, nous avons pu augmenter les cohérences entre les différentes phases de segmentation et réduire les temps de calcul.

Références

- [1] Wang Ching-Huei, P.W. Palumbo, S.N. Srihari, Object recognition in visually complex environments: an architecture for locating address blocks on mail pieces, Pattern Recognition, 1988., 9th International Conference, IEEE, 1988, vol.1, Pages: 365 – 367.
- [2] C Viard-Gaudin, D Barba, A multi-resolution approach to extract the address block on flat mail pieces, ICASSP-91, International Conference, vol.4, Pages: 2701 - 2704 .
- [3] DJ. Gaceb, V. Eglin, F Lebourgeois, H. Emptoz. Address block localization based on graph theory. Document Recognition and Retrieval XIV, SPIE Int. Soc. Opt. Eng ed. San Jose (USA, Californie). p. 12. 2008.
- [4] Rémy Mullot, Livre: Les documents écrits de la numérisation à l'indexation par le contenu, Editeur : Hermes science Publication, Nov 2006, 365 pages, ISBN-10 : 2746211432.

- [5] D. Drivas, A. Amin, Page segmentation and classification utilising a bottom-up approach, Document Analysis and Recognition, ICDAR.1995, Pages : 610 - 614 vol.2.
- [6] Olivier Déforges and Dominique Barba, A fast multiresolution text-line and non text line structures extraction and discrimination scheme for document image analysis, ICPR 94, Pages: 134-138.
- [7] Z. Shi; Venu Govindaraju, Line separation for complex document images using fuzzy runlength, Document Image Analysis for Libraries, DIAL 2004, Proceedings, First International Workshop on 2004 Pages:306 – 312
- [8] Shin-Ywan Wang; Yagasaki, T; Block selection: a method for segmenting a page image of various editing styles, ICDAR.1995, Vol 1, Pages:128 - 133 vol.1
- [9] PASCHOS Vangelis, livre , Optimisation combinatoire5: problèmes paradigmatiques et nouvelles problématiques, Lavoisier, France, 2007, Pages : 270.
- [10] Brice Effantin and Hamamache Kheddouci , The b-chromatic number of power graphs, DMTCS 2003, Vol.6, pp. 45-54.
- [11] Brice Effantin and Hamamache Kheddouci, a distributed algorithm for a b-coloring of a graph, ISPA'2006, Serrento, Italy, 2006 (IEEE).
- [12] N. Otsu, A threshold selection method from grey-level histogram, IEEE trans system, man and cybernetics, vol 9, pp. 62-66, 1979.
- [13] J. Sauvola, and al. Adaptive Document Binarization, ICDAR'97, vol 1, 1997, Pages: 147–152.
- [14] Djamel Gaceb, Frank Lebourgeois, Véronique Eglin, Hubert Emptoz, Contribution to the automatic recognition of business documents, 6p, IWFHR 2006, La Baule, France.
- [15] Regentova, E.; Latifi, S.; Shulan Deng; Dongsheng Yao, An Algorithm with Reduced Operations for Connected Components Detection in ITU-T Group 3/4 Coded Images. Pattern Analysis and Machine Intelligence, IEEE , Vol. 24, Pages:1039 – 1047.
- [16] Z. Pavlidis and J. Zhou, A Page Segmentation and Classification, CVGIP92, vol.54, no. 6, Pages: 484-496.