Implication de la b-coloration de graphes pour la reconnaissance automatique du type de document

Djamel Gaceb, Véronique Eglin, Frank Lebourgeois, Hubert Emptoz

Implication de la b-coloration de graphes pour la reconnaissance automatique du type de document

Djamel GACEB Véronique EGLIN Frank LEBOURGEOIS Hubert EMPTOZ

LIRIS INSA de Lyon
LIRIS UMR 5205CNRS, INSA de Lyon 69621 Villeurbanne Cedex

{djamel.gaceb1,veronique.eglin,flebourg,hubert.emptoz}@insa-lyon.fr

Résumé *: La société CESA souhaite réduire à tout prix les taux de rejets et d'erreurs de ses systèmes de vision existants en introduisant la reconnaissance du type de document dans une étape préliminaire. Cette opération importante va diriger les autres étapes du processus de tri automatique de documents. Une fois le document identifié, un module de lecture va alors savoir quel type de traitement faire en fonction du contenu pour extraire les informations, les contrôler et prendre rapidement la décision nécessaire. Nous proposons dans cet article une méthode originale de reconnaissance du type de documents adaptée aux documents structurés présentant de grande variabilité de mise en forme. A ce titre, nous introduisons la coloration de graphe dans la phase d'analyse de la structure physique et la b-coloration dans la phase de classification des documents. Les expérimentations qui en sont faites ont montré une très grande fiabilité de la méthode, des réponses en temps réel garanties ainsi une grande robustesse aux différentes contraintes liées à l'application de tri qui concerne la société.*

Mots-clés *: Analyse de la structure physique, classification des documents, b-coloration des graphes, tri automatique de documents.*

1 Introduction

Le traitement automatique de documents contribue à créer une grande valeur ajoutée à l'entreprise, à valoriser son patrimoine documentaire, à le rendre plus accessible, à proposer de nouveaux services ou encore à améliorer son propre processus organisationnel.

Le tri automatique de documents, en particulier, peut permettre un gain de temps et d'argent énorme pour les organisations. Les tendances les plus visibles aujourd'hui se développent dans l'amélioration des systèmes de vision où les performances ont longtemps été associées à celles de l'OCR. Mais de notables évolutions technologiques ont permis d'affiner la précision et la pertinence de la reconnaissance permettant non seulement de prendre en compte des écritures différentes (imprimées ou manuscrites) mais aussi de procéder à une reconnaissance intelligente du document. Tout système de reconnaissance de documents nécessite ainsi l'introduction de connaissances liées au type de document à reconnaître [MUL 06]. Dans la plupart de ces systèmes, la connaissance est embarquée dans le code qui devient de ce fait difficile à adapter à de nouveaux types de documents.

La reconnaissance automatique des documents (RAD) sert à classifier les documents selon leur typologie (figure 1). La connaissance délivrée par cette étape permet de cibler les informations pertinentes au tri et choisir un jeu de traitements plus adapté au contenu.

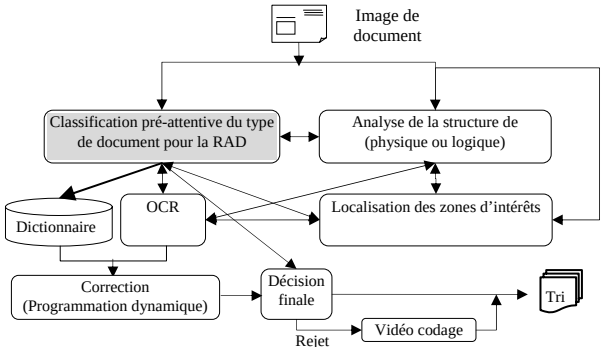

FIG. 1 – SITUATION DU MODULE DE LA RAD DANS UN SYSTÈME DE TRI

Les résultats obtenus à ce jour en matière de tri automatique sont loin d'être parfaits. L'introduction de la RAD dans une application de tri reste un problème complexe qui bute sur des difficultés encore non résolues et faisant actuellement encore l'objet de nombreuses recherches qui doivent s'adapter à certaines contraintes. Voici celles qui nous semblent être les plus significatives:

- une très grande variété de documents (de structures variées avec des contenus textuels manuscrits ou imprimés, sur des supports papiers dont la qualité, les couleurs et la texture peuvent être très différentes),
- un fonctionnement en temps réel (quelques fractions de secondes doivent suffire à la reconnaissance),
- la maîtrise de la qualité des résultats (le système doit être le plus performant possible pour éviter le coûteux traitement manuel).
- le type de document doit être identifié automatiquement malgré les aléas de numérisation (rotations, décalages, plissements du papier),
- résolution spatiale des images élevée (300 dpi),
- superposition de couches d'informations (tampons, notes manuscrites, …).

Dans ce contexte et pour satisfaire l'ensemble de ces contraintes, nous proposons dans cet article une architecture flexible de RAD basée sur la coloration hiérarchique des graphes. A ce jour, aucun travail dans ce domaine ne s'est servi de la puissance de cet outil.

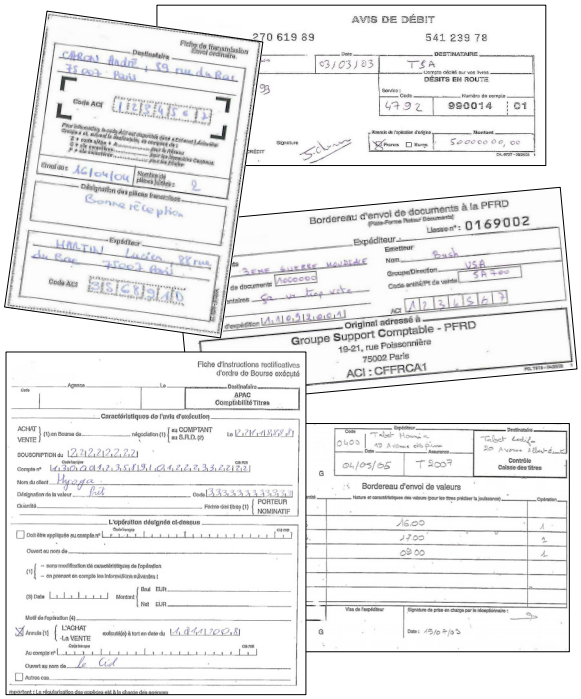

FIG. 2 - UNE TRÈS GRANDE VARIÉTÉ DE DOCUMENTS

Le reste de cet article est organisé comme suit : la partie 2 cite les différentes stratégies de classification de documents existantes. Nous y présentons les travaux existants et les limites par rapport à notre application. Dans une troisième partie, nous présentons les aspects formels de la coloration de graphes. Enfin, nous décrivons dans la quatrième partie l'implication de la coloration dans la RAD. Les résultats obtenus sont finalement commentés à la fin de l'article.

2 Méthodes de classification

2.1 Différentes stratégies

Un document peut-être vu comme un une organisation d'objets (de symboles textuels et graphiques de toutes sortes) ayant une disposition aléatoire ou structurée. La classification des documents consiste à regrouper divers documents en sous-ensembles homogènes sur la base des représentations descriptives de leurs structures physiques ou de leurs contenus textuels de telle sorte que les observations dans les mêmes classes de documents soient le plus similaire possible, et que les observations dans des classes différentes soient le plus dissemblable possible. Il existe de nombreuses méthodes de classification automatiques de documents supervisées et non supervisées qui dépendent de l'application concernée. La stratégie de décision de chacune peut être inspirée des bases de connaissances, des nuées dynamiques (K-Means), des chaînes de Markov, des arbres de décision, des isomorphismes de graphes, des machines à support de vecteurs, des réseaux de neurones, mais également des méthodes statistiques etc. [CAR 04]. Ces méthodes utilisent une représentation sous plusieurs aspects:

– description de l'image seulement et/ ou,
– description de la structure physique et/ou,
– description de la structure logique et/ou,
– description du contenu textuel.

Les systèmes de RAD se basant sur la structure logique du document [EGL03] ou son contenu textuel [MOH07] sont lents et difficiles à mettre en œuvre dans une application de tri qui doit fonctionner en temps réel. Par ailleurs, la quantité d'informations apportée par une simple description de l'image de document sans analyser sa structure physique ne peut pas être discriminante sur des documents présentant une grande variabilité de mise en forme [HER98]. Ces contraintes impliquent d'avoir tout à la fois une représentation descriptive simple et discriminante permettant de classer rapidement tous les documents susceptibles d'apparaitre dans une chaine de tri. Afin de s'adapter au mieux aux exigences de rapidité et d'efficacité imposées par notre application, l'approche que nous proposons dans cet article est basée sur la description de la structure physique des pages.

2.2 Méthodes basées sur la description de la structure physique des documents

La plupart des méthodes basées sur ce principe utilisent une représentation hiérarchique des éléments physiques sous forme de zones (blocs ou lignes de texte, graphiques, grilles, cases à cocher, tableaux …). Cette représentation permet de mettre facilement en relation les différents éléments constitutifs d'un document.

Héroux et al. [HER 98] représentent chaque document par un arbre, où les nœuds sont formés à partir de blocs issus de l'analyse de la structure physique. Ils utilisent ensuite une mise en correspondance hiérarchique entre les arbres pour regrouper les documents en classes homogènes. Esposito et al [ESP 00] projettent les attributs et les relations entre les blocs dans un langage du premier ordre. Ce langage est utilisé avec certaines règles par la phase d'apprentissage. Cesarini propose dans [CES01] un algorithme de construction de l'arbre X-Y basé sur une stratégie de segmentation descendante. Le document est découpé récursivement selon les directions horizontale ou verticale d'après des zones homogènes de séparations. Le résultat est alors un arbre où chaque nœud représente une zone de l'image. Les techniques d'apprentissage par réseaux de neurones de type PMC (perceptron multi-couches) consistent à minimiser un critère d'erreur en adaptant l'ensemble de poids du réseau représentant les modèles. Les arbres X-Y produits par les algorithmes descendants de segmentation sont également très utilisés, mais ils induisent des risques d'être insuffisamment discriminants à cause de la rotation des images. Baldi [BAL 03] et Diligenti [DIL 03] propose une extension de l'arbre X-Y en arbre XYM. Baldi projette les distances d'édition entre les arbres dans un espace de K plus-proches-voisins, alors que Diligenti l'utilise pour construire un modèle d'arbre de Markov caché

(HTMM). D'autres travaux de classification portant sur la théorie des graphes a été proposée par Bagdanov et al. dans [BAG 03]. La technique est principalement basée sur la construction de FOGGs (First Order Gaussian Graphs) où des probabilités sur les nœuds et sur les sommets ont été utilisées lors de l'apprentissage pour créer les modèles de reconnaissance.

2.3 Besoin d'un outil plus adapté

Les méthodes de classification de documents que nous avons citées, utilisent toutes des structures de données complexes au niveau de la classification, aussi bien qu'au niveau de l'analyse de la structure physique. Elles nécessitent souvent la construction d'une base contenant un grand nombre d'exemples où la gestion des critères et des connaissances devient difficilement contrôlable face à la grande variabilité sur les documents à trier. Afin de répondre au mieux aux besoins du système industriel de notre entreprise partenaire, il a fallu choisir un outil efficace garantissant des résultats cohérents par rapport aux exigences des applications temps réel. C'est la raison pour laquelle, nous avons choisi de mettre en place une architecture complète basée sur la coloration des graphes.

3 Aspects formels de la coloration des graphes

Une question naturelle est alors de déterminer quel est le nombre minimal de classes nécessaire pour décomposer un ensemble de n objets (composantes connexes, lignes, blocs ou documents) en plusieurs sous ensembles homogènes. Cette question peut se formuler en termes de coloration de graphe. Pour cela, il suffit de représenter chaque objet i par un sommet v_i et d'ajouter une arête $E(v_i,v_j)$ entre chaque paire d'objets pour lesquels on peut assurer qu'ils ne peuvent appartenir au même regroupement. Le graphe fini $G = (V, E)$ est défini par l'ensemble fini $V = \{v_1, v_2, ..., v_n\}$ $(|V| = n)$ dont les éléments sont appelés sommets, et par l'ensemble fini $E= \{e_1, e_2, ..., e_m\}$ $(|E| = m)$ dont les éléments sont appelés arêtes.

3.1 La coloration

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 doivent pas porter la même couleur. Ces couleurs vont correspondre aux différentes classes d'objets.

Le nombre de couleurs utilisées pour colorer le graphe G est appelé nombre chromatique $\chi(G) \leq n$. Ce nombre représente le plus petit entier k pour lequel il existe une partition de l'ensemble V en k sous-ensembles homogènes [PAS07].

Sur le graphe G de la figure 3, nous avons représenté un ensemble de 11 formes différentes V représenté par leurs sommets {x1,...,x11}. Nous avons eu besoin de quatre couleurs différentes pour colorer les 11 sommets de sorte que deux sommets adjacents aient des couleurs différentes.

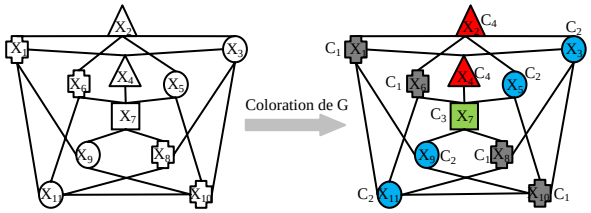

FIG. 3 - COLORATION DE GRAPHE G DE 11 SOMMETS PAR 4 COULEURS (C_1, C_2, C_3 ET C_4).

3.2 La b-coloration

La coloration est appelée b-coloration, si pour chaque couleur C_i, il existe au moins un sommet v_i coloré C_i dont le voisinage est coloré par toutes les autres couleurs. Le sommet v_i est dit sommet dominant pour la couleur C_i. L'exemple de la figure 4 présente la possibilité de b-colorer les sommets d'une classe de couleur à l'aide des autres couleurs.

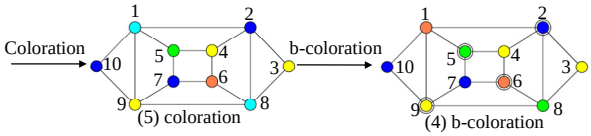

FIG. 4 - EXEMPLE DE B-COLORATION, LES SOMMETS 2, 5,6 ET 9 SONT DES SOMMETS DOMINANTS

Le nombre b-chromatique d'un graphe G, noté $b(G)$, est le nombre entier maximal de couleurs k_b tel que G peut avoir une b-coloration par les k_b couleurs.

3.3 Les algorithmes adoptés

La plupart des évaluations de $\chi(G)$ et de $b(G)$ proviennent d'algorithmes de coloriage. Il en existe beaucoup, et pour ne pas entrer dans une description exhaustive de ces différentes techniques, nous nous limiterons à citer l'étude comparative effectuée par Paschos [PAS07]. Plus de détails sur l'approximation du nombre b-chromatique ont été proposés par Corteel dans [COR05]. Dans notre étude, nous nous sommes particulièrement intéressés aux algorithmes distribués de coloration et de b-coloration proposés par Effantin et Kheddouci dans [EFF03], [EFF06].Tous ces algorithmes ont été efficacement introduits dans les travaux d'Elghazel[ELG06] qui propose une nouvelle méthode non supervisée de classification des données médicales basée sur la b-coloration de graphe où le nombre des classes n'est pas connu à l'avance. Sur la même base de données, la comparaison de cette méthode avec la méthode de classification hiérarchique agglomérative, l'approche du Hansen et la classification de DRG, a montré que cette technique offre une vraie représentation des classes par les individus dominants et garantit une meilleure disparité interclasse.

4 Implication de la coloration de graphes dans la RAD

Nous présentons dans cette section les différentes étapes de notre méthode de RAD (figure 5).

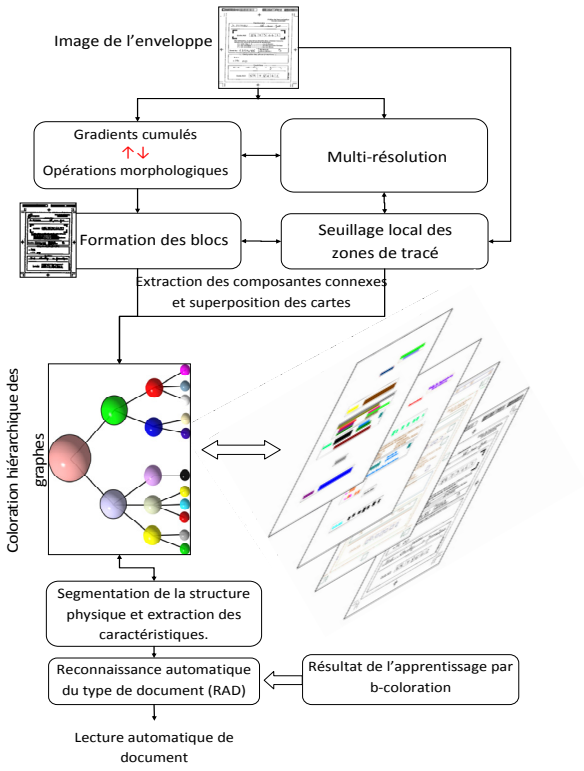

FIG. 5 - DIAGRAMME FONCTIONNEL DE LA MÉTHODE

4.1 Analyse de la structure physique

4.1.1 Binarisation et détection des composantes connexes

La binarisation est appliquée dans la première étape du processus de RAD et a un impact très fort sur les performances du système de tri. La séparation entre l'étape de binarisation et celle de localisation de tracé 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 de document. En effet, aucune des méthodes classiques (globale ou locale) 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 employant conjointement la multi-résolution et la morphologie mathématique [GAC06].

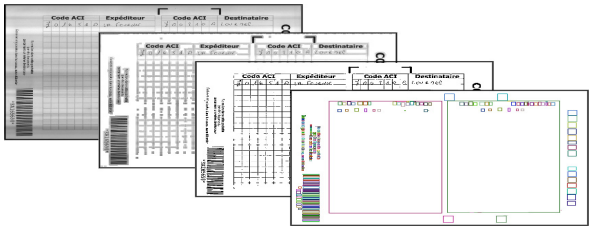

FIG. 6 – SEUILLAGE LOCAL À PROXIMITÉ DES ZONES DE TEXTE ET DÉTECTION DES COMPOSANTES CONNEXES.

Nous détectons ainsi les composantes connexes (CCs) de la carte des blocs qui sont utilisées ensuite pour guider l'extraction des composantes de premier plan. La méthode utilisée est inspirée des études de Pavlidis sur la structure LAG (Line Adjacency Graph) [PAV92].

4.1.2 Segmentation de la structure physique par coloration hiérarchique des composantes

L'idée est d'utiliser une stratégie pyramidale de segmentation qui porte principalement sur la puissance de la méthode de coloration de graphes (algorithme1) permettant de distinguer les éléments pertinents et de les regrouper en ensembles homogènes tout en rejetant les éléments parasites. Le principe consiste donc à distinguer, par coloration des CCs, toutes les zones textuelles des zones non textuelles, puis de regrouper les CCs des zones textuelles en lignes. Nous avons déjà présenté en détail le principe de notre méthode de segmentation dans [GAC08].

Algorithme1: Coloration (G_k)

Début
 Si $col_k(i) \neq \varnothing$ Alors
  Soit $M = N^k_{col}(i) \cup \{col_k(i)\}$; $q = 0$;
  Pour chaque sommet $j \in N^k_{adj}(i)$
   Sachant que $col_k(j) := \varnothing$ Faire
   q =min $\{r / r > q, r \notin M$ Et $r \notin col_k(j)\}$;
    Si $q \leq \Delta + 1$ Alors $col_k(j) := q$;
    Sinon $col_k(j)$:=min$\{r / r \notin N^k_{col}(j)\}$;
 FinSi, FinFaire, FinSi
Fin.

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

4.2 Représentation des documents

L'extraction des caractéristiques a pour objectif de minimiser la quantité d'informations nécessaire à la séparation des documents. Pour cela, nous déterminons à partir de la structure physique d'un document un certain nombre de caractéristiques : 15 globales, relatives au document dans son ensemble, et 20 locales, propres à une ligne de texte. Nous utilisons ainsi deux types de représentations :

1. Représentation structurelle : chaque document j est représenté par une séquence ordonnée de n lignes de texte : $Rs(j)=(L^j_1,L^j_2,...,L^j_n)$ où la *tème* ligne L_t est représentée par un vecteur de p caractéristiques locales avec $L_t=(x^t_1,x^t_2,...,x^t_p)$.
2. Représentation vectorielle globale : chaque document j est représenté par un vecteur de m caractéristiques globales, avec $Rv(j)=(y^j_1,y^j_2,...,y^j_m)$.

4.3 Mesures de distances

Pour comparer deux documents, on utilise la combinaison de deux distances (D_{Rv} sur Rv^m et D_{Rs} sur Rs^n) donnée par la formule suivante:

$$DT=\gamma D_{Rv}+(1-\gamma)D_{Rs} \quad \text{avec} \quad \gamma=\{k = \arg\max_{0\leq k\leq 1}(\psi_k)\}$$

La valeur de γ doit être choisie de manière à maximiser la qualité de classification Ψ [GAC08]. Si deux documents sont séparés par une faible distance DT alors ils se ressemblent.

La distance euclidienne D_{Rv} entre deux documents, représentés par les descripteurs *Rv(i)* et *Rv(j)*, se calcule de la façon suivante :

$$D_{Rv}[Rv(i), Rv(j)] = \left[\sum_{k=1}^{m} \left| y_k^i - y_k^j \right|^{\alpha}\right]^{\frac{1}{\alpha}} \text{ avec } \alpha = 2$$

La distance D_{RS} réalise un mapping spacial entre les séquences *Rs(i)* de *ni* lignes et *Rs(j) de nj* lignes, elle est appelée la *Warping Function*. L'ajustement non-linéaire entre *Rs(i)* et *Rs(j)* peut être représenté par un chemin : $C=c_1,c_2,\ldots,c$ avec $c_k=(i_k,j_k)$ (figure 7).

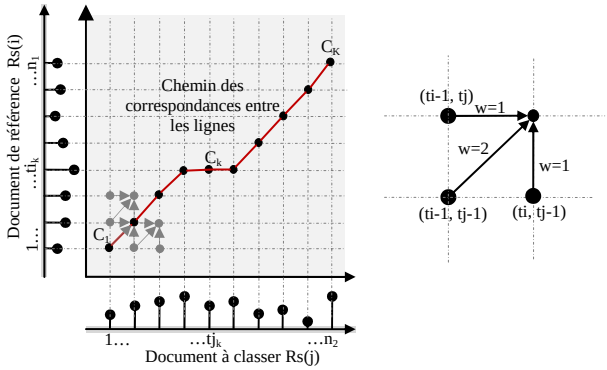

FIG. 7 –PRINCIPE DE LA FONCTION DE DÉFORMATION

La somme pondérée des erreurs le long de la Warping Function C est:

$$D(c) = \frac{\sum_{k=1}^{K} d(c_k)\cdot w_k}{\sum_{k=1}^{K} w_k} \text{ avec } d(c_k) = d(L_t^i, L_t^j) = \sqrt{\sum_{t=1}^{p}\left[x_t^i(i) - x_t^i(j)\right]^2}$$

les coefficient de pondération $w_k = ti_k - ti_{k-1} + tj_k - tj_{k-1}$ et donc $\sum_{k=1}^{K} w_k = n_i + n_j$

Dans ce cas, le problème à résoudre devient:

$$D_{Rs}[Rs(i), RS(j)] = \frac{1}{n_i + n_j} \min_{C} \sum_{k=1}^{K} d(c_k)\cdot w_k$$

Le nombre de chemins possibles croit exponentiellement avec les nombre de lignes dans les documents à comparer. Ce problème peut être résolu de manière efficace par un algorithme de comparaison dynamique qui va rapidement mettre en correspondance optimale les lignes de deux documents. Le principe est qu'au lieu d'étudier tous les chemins possibles, il est possible de trouver la solution optimale en étudiant le problème localement (figures 7 et 8).

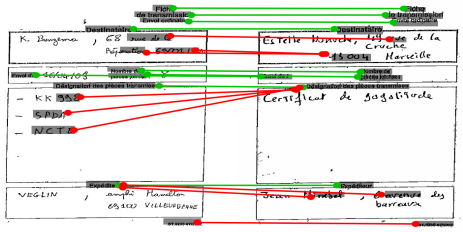

FIG. 8 –COMPARAISON DYNAMIQUE DES DOCUMENTS

4.4 Classification de documents

Nous représentons un ensemble *R* de *N* documents dans un graphe *G(V, E)* où chaque sommet correspond à un document. Deux sommets v_i et v_j sont alors adjacents si et seulement si la distance *DT* entre les documents *i* et *j* est supérieure strictement à un seuil S_{DT}. Le mécanisme d'optimisation de ce seuil est détaillé dans [GAC08]. L'adjacence entre les sommets peut être donnée par :

$$E[v_i, v_j] = \begin{cases} 1 & \text{si } DT(v_i, v_j) > S_{DT} \\ 0 & \text{sinon} \end{cases}$$

Pour décomposer l'ensemble *R* en sous-ensembles homogènes, nous colorons le graphe *G* puis nous appliquons l'algorithme suivant [EFF06]:

Algorithme 2: b-coloration(G)

```
Début
    Répéter jusqu'à (NDm = ϕ),
     q= max{k/k∈ NDm}; L= L\{q}; NDm= L\Dm;
     Pour chaque sommet v_{i/c(vi)=q} faire
      K= {k/k∈ L and k∉ Nc(vi)};
      c(vj)={c/dist(vi,c)=min_{k∈K}(DT(vi,k))};
     FinPour
     Pour chaque sommet v_{j/c(vj)} ∈ NDm faire
      Actualiser (Nc(vj));
      Si Nc(vj)=L\{c(vj)} alors Add(c(vj),Dm);
      FinSi, FinPour
Fin.
```

Avec *NDm*, est l'ensemble des couleurs non dominantes, Dm est l'ensemble des couleurs dominantes, et $C(v_i)$ est la couleurs de sommet *i*. La b-coloration garantit une grande disparité interclasse et offre une vraie représentation des classes par les sommets dominants.

4.5 Mécanismes d'apprentissage

Dans cette étape on fournit à la machine d'apprentissage une base *R* de *N=512* documents répartis en *14* classes. L'algorithme d'apprentissage utilise donc la technique de classification exposée dans la section 4.4. La b-coloration délivre automatiquement un jeu de N^* sommets dominants (représentants des classes) $R^* = \left\{R_1^*, \ldots, R_{N^*}^*\right\}$ qui seront utilisés pour reconnaître le type du document en temps réel.

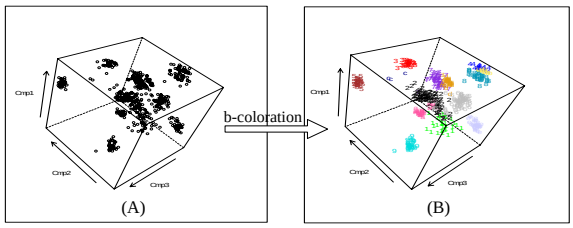

FIG. 9 – (A) REPRÉSENTATION DE 512 DOCUMENTS DANS L'ESPACE DE CARACTÉRISTIQUES, (B) ÉMERGENCE DES 14 CLASSES PAR B-COLORATION.

4.6 Reconnaissance du type de document

Étant donné un document d'entrée *T(i)*, l'objectif du système de reconnaissance est de comparer sa description avec celles de tous les représentants des classes (sommets dominants) de R^* issues de la phase d'apprentissage. L'algorithme d'appariement reconnaît en temps réel le type de document *T(i)* à partir du type le plus proche dans R^* de la façon suivante :

$$Type[T(i)] = \begin{cases} \text{Rejet si } \arg\min_{k=1\ldots N^*}\left(DT\left[T(i), R_k^*\right]\right) > S_{DT} \\ Type(R_k^* \mid \arg\min_{k=1\ldots N^*}\left(DT\left[T(i), R_k^*\right]\right) \text{ sinon} \end{cases}$$

Le seuil d'adjacence S_{DT} permet aussi de délimiter les connaissances du classifieur pour rejeter les documents qu'il n'a pas appris à reconnaître.

5 Expérimentation

Nous avons utilisé l'indice de KAPPA pour évaluer la précision de la classification de 512 documents de la base d'apprentissage par les trois méthodes (KMeans, SVM et b-coloration). Plus cet indice est proche de 100%, plus la classification est correcte. L'histogramme suivant montre que la b-coloration montre une meilleure classification.

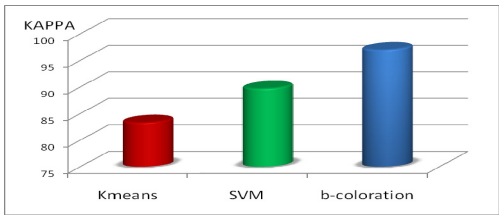

FIG. 10 – COMPARATIVE DES MÉTHODES DE CLASSIFICATION

Nous avons testé également les trois classifieurs avec une base de test de 576 documents répartis en 14 classes dont le type a été appris et 2 classes dont le type n'a pas été appris. Les courbes suivantes montrent leurs taux de reconnaissance sur les 14 classes connues et leurs taux de rejet sur les 2 classes inconnues. La b-coloration donne de meilleures performances aussi bien au niveau de la reconnaissance qu'au niveau des rejets.

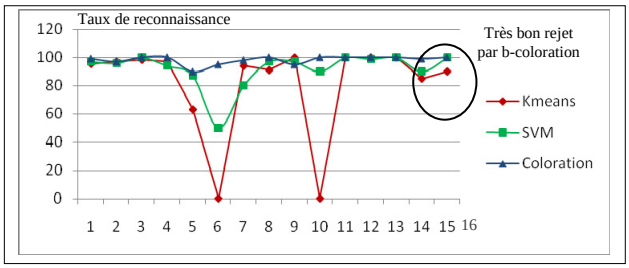

FIG. 11 – COMPARATIVE DES TROIS CLASSIFIEURS

6 Conclusion

Nous avons présenté une nouvelle méthode de reconnaissance du type de documents basée sur la coloration hiérarchique des graphes. Notre méthode utilise une représentation issue de la description de la structure physique de documents. La coloration hiérarchique de graphe a été introduite dans la phase de segmentation pour en augmenter la robustesse aux composantes parasites considérées comme facteurs d'erreur des méthodes classiques de segmentation. La b-coloration a été introduite dans la phase d'apprentissage. Grâce au nombre restreint de règles dont elle dispose, cette nouvelle technique répond à une large variabilité de documents et offre une vraie représentation des classes par les documents dominants et garantit une meilleure disparité interclasses. De plus, nous avons pu augmenter les cohérences entre les différentes phases de la RAD et réduire les temps de calcul.

En perspective de ce travail, nous comptons étendre le principe de cette méthode pour effectuer un apprentissage incrémental sur les documents rejetés. Cette étape va permettre au système de tri de classer de nouveaux documents.

Références

[PAV92] Pavlidis Z. and J. Zhou, A Page Segmentation and Classification, CVGIP92, vol.54, no. 6, pp. 484-496.

[HER 98] HÉROUX P. and al, Classification method study for automatic form class identification, *the 14th ICPR*, Brisbane, Australia, 1998, pp. 926-929.

[ESP 00] ESPOSITO F. and al, Machine learning for intelligent processing of printed documents. *J. Intell. Inf. Syst*, 2000, pp. 175-198.

[CES 01] CESARINI F. and al, Encoding of modified X–Y trees for document classification, *the 6th International Conference on Document Analysis and Recognition*, Seattle, USA, 2001, pp. 1131-1136.

[DIL 03] DILIGENTI M. and al, Hidden Tree Markov Models for document image classification, *IEEE Trans. Pattern Anal. Mach. Intell.* 25(4), 2003, pp. 519 -523.

[BAL 03] BALDI S. and al, Using tree-grammars for training set expansion in page classification, *the 7th International Conference on Document Analysis and Recognition*, Scotland, 2003, pp. 829-833.

[BAG 03] BAGDANOV A.D. and Worring M.: First order Gaussian graphs for efficient structure classification, Pattern Recognit. 36(6), 2003, pp.1311–1324.

[EGL 03] EGLIN V. and BRES S., Document page similarity based on layout visual saliency: application to query by example and document classification, *the 7th ICDAR*, Scotland, 2003, pp. 1208-1212.

[EFF 03] EFFANTIN B. and KHEDDOUCI H., The b-chromatic number of power graphs, *DMTCS* 2003, Vol.6, pp. 45-54.

[CAR 04] CARMAGNAC F. and al , Une stratégie originale de classiffication basée sur le calcul de distances avec sélection de caractéristiques, application à la classiffcation d'images de document, Actes du 14ème congrès francophone AFRIF-AFIA, 2004.

[COR 05] CORTEEL S. and al, On approximating the b-chromatic number, *Discrete Applied Mathematics archive*, 2005, Vol146 , pp. 106-110.

[EFF 06] EFFANTIN B. and KHEDDOUCI H., a distributed algorithm for a b-coloring of a graph, IEEE *ISPA'2006*, Serrento, Italy, 2006.

[ELG 06] ELGHAZEL H. and al, A New Clustering Approach for Symbolic Data: Algorithms and Application to Healthcare Data, *BDA 2006*, Lille, France.

[MUL 06] MULLOT R., Livre: Les documents écrits de la numérisation à l'indexation par le contenu, *Hermes science Publication*, 2006, pp. 365.

[MOH 07] MOHAMED H.K., Automatic documents classification, *IEEE International Conference, Computer Engineering & Systems*, 2007, pp. 33-37.

[PAS 07] PASCHOS V., livre, Optimisation combinatoire5: problèmes paradigmatiques et nouvelles problématiques, *Lavoisier*, France, 2007, pp. 270.

[GAC 06] GACEB DJ. et al, Contribution to the automatic recognition of business documents, *IWFHR*, La Baule, France, 2006, pp.6.

[GAC 08] Gaceb DJ et al, Address block localization based on graph theory. *DRR XIV*, SPIE Int. Soc. Opt. Eng ed. San Jose (USA, Californie). 2008, pp.12.