Extraction de formules chimiques dans des documents manuscrits composites

Nabil Ghanmi[*,**] — **Abdel Belaïd**[*]

[*] ***LORIA*** *- Campus scientifique, BP 239, Vandoeuvre-lès-Nancy, F**-54506**, France*
[**] ***eNovalys*** *- Bâtiment Pythagore, 1 Rue Jean Sapidus, ILLKIRCH, F**-67400**, France*
http://www.enovalys.com
Email: {nabil.ghanmi, abdel.belaid}@loria.fr

RÉSUMÉ. *Nous abordons dans ces travaux, le problème de la segmentation de documents de cahiers de la chimie en zones homogènes. Les documents à traiter sont manuscrits sans contraintes composés de zones de textes, de tableaux et de graphiques, représentant l'expression graphique de l'expérience réalisée. L'objectif de ce premier travail est d'extraire, dans chaque document, le bloc contenant le schéma graphique. Nous proposons une méthode d'extraction et de classification des structures élémentaires du document sur lesquels s'appuiera une technique de séparation verticale des blocs. Des descripteurs spécifiques tenant compte de la texture du texte et du graphique sont pris en compte. Des connaissances a priori sur la structure du document sont ensuite utilisées pour délimiter le bloc graphique. Les résultats expérimentaux obtenus sur une variété de documents de chimie sont de l'ordre de 92% de bonne extraction de graphique*

ABSTRACT. *In this work, we address the problem of segmentation of chemistry documents in homogeneous areas. The documents are handwritten, unconstrained and composed of text areas, tables and graphics representing the chemical formula. The goal of this first part is to extract, in each document, the block containing graphical drawings. We propose a method to extract and classify elementary structures of the document. A vertical separation of the blocks is then carried out. Specific descriptors taking into account the texture of the text and graphics are considered. A priori knowledge about the document structure is then used to delimit the graphical block containing the chemical formula. Experiments results obtained on a variety of chemistry documents are around 92% of good graphic extraction*

MOTS-CLÉS : *extraction de formules chimiques, segmentation en lignes, classification, descripteurs texturaux*

KEYWORDS: *chemical formulas extraction, Line segmentation, classification, textural features*

CIFED 2014, pp. 185–197, Nancy, 18-21 mars 2014

1. Introduction

Face au volume très important de données manuscrites contenues dans les archives de différents organismes, le besoin de systèmes permettant l'indexation, la recherche et la reconnaissance de documents numérisés croit continuellement. La rétro conversion de ces documents reste une tâche difficile car elle nécessite non seulement le développement de systèmes fiables de reconnaissance de l'écriture manuscrite mais également le développement de méthodes robustes d'extraction des différentes zones d'intérêts dans les documents. Dans cette optique, nous nous intéressons à l'extraction du bloc graphique contenant une formule chimique dans les documents de chimie. Cette phase constitue une première étape dans la chaîne d'analyse de page complète d'écriture. Elle permet d'extraire et de séparer le graphique d'une part et le texte d'autre part dans le but de donner à des outils dédiés, chacune de ces deux couches. Le texte est composé de chiffres et caractères regroupés en chaînes de caractères. Il peut être exploité par un système de reconnaissance d'écriture. Le graphique, quant à lui, est essentiellement composé de lignes, de polygones, de cercles, ... Il peut être traité par un système de reconnaissance et d'interprétation de graphique.

D'une manière générale, extraire une formule chimique dans des documents manuscrits composites est un challenge pour les raisons suivantes :

– Les éléments graphiques (lignes, cercles, polygones, ...) peuvent être de tailles quelconques et non parfaitement tracés, ce qui les rend confondables avec les éléments textuels (caractères, pseudo-mots, mots ...)

– Les formules sont généralement denses en caractères qui se superposent aux éléments graphiques, altérant ainsi les formes graphiques et rendent difficile leur détection.

– Les chaînes de caractères représentant les noms d'atomes ou de molécules, situés aux extrémités des polygones ou des liaisons peuvent être nettement détachés du schéma de la formule chimique. Ces chaînes peuvent être omises lors de la délimitation de la formule chimique.

En outre, les documents que nous traitons présentent d'autres caractéristiques particulières qui rendent l'extraction de la formule chimique encore plus difficile. En effet, ces documents présentent une large variabilité tant au niveau de l'écriture que de la qualité. La majorité de ces documents est rédigée en français et certains sont en anglais. Les documents sont écrits par différents scripteurs et il n'y a pas de fortes contraintes sur leur contenu ou leur structure (voir figure 1).

De nombreuses méthodes ont été proposées pour résoudre le problème de séparation texte/graphique dans les images de documents. Une des principales méthodes est celle de Fletcher et al. (Fletcher et Kasturi, 1998) qui se base sur un ratio lié aux dimensions des composantes connexes pour filtrer les composantes de grosses tailles probablement appartenant à des graphiques. Ensuite, ils utilisent la transformée de Hough, en considérant comme points votants les centres des boîtes englobantes des composantes connexes, pour détecter les composantes colinéaires qu'ils regroupent en chaînes de caractères. Cette approche est indépendante de la taille et du style d'écri-

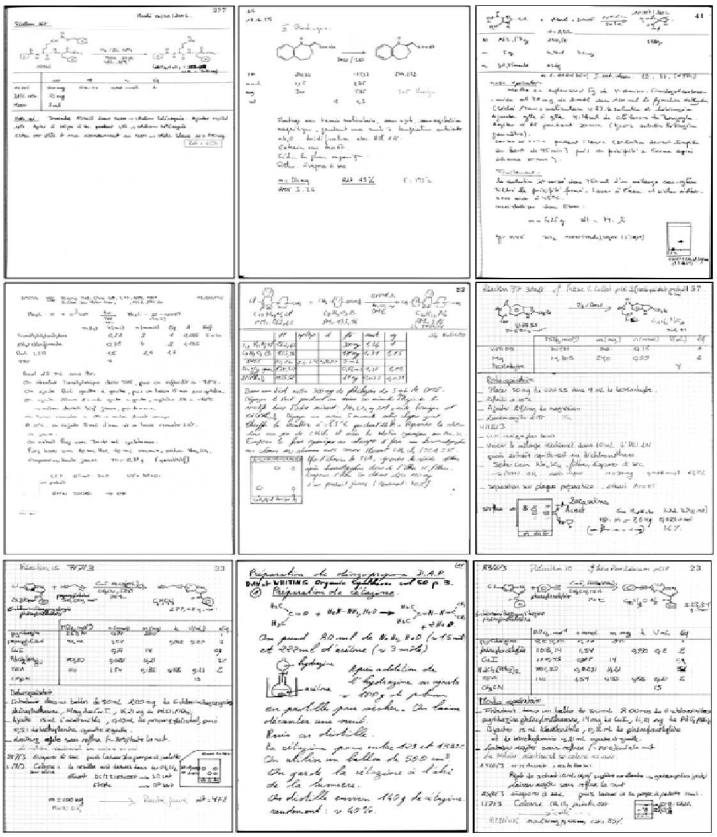


Figure 1. *Exemples de documents de travail*

ture. Elle a montré une grande robustesse même pour des documents complexes contenant du texte de différentes orientations englobées dans des graphiques de différentes formes. Le principal inconvénient de cette méthode est qu'elle est basée sur l'hypothèse que les caractères ne collent pas au graphique. Quelques améliorations (Tombre *et al.*, 2002) (Roy *et al.*, 2010) ont été proposées pour pallier cette limite. En se basant sur l'hypothèse que le texte est généralement présent sous forme de chaînes et non pas de caractères isolés et que la chaîne n'est pas entièrement collée au graphique, Tombre et al. (Tombre *et al.*, 2002) déterminent l'orientation de la chaîne à partir des boîtes englobantes des caractères déjà retrouvés pour définir une zone de recherche d'éventuels caractères collés au graphique. Les caractères appartenant à cette zone, sont ensuite détachés du graphique en segmentant le squelette et en reconstituant chaque partie à part. Récemment, Roy et al. (Roy *et al.*, 2010) ont développé une méthode basée sur les descripteurs invariants SIFT pour détecter les caractères touchant le graphique. L'idée consiste à labéliser les caractères isolés, détectés par une analyse des compo-

santes connexes (Fletcher et Kasturi, 1998), en utilisant SIFT. Le système apprend au fur et à mesure les différentes formes de chaque caractère. Ensuite, les images de ces caractères sont utilisées comme requêtes pour extraire des caractères similaires collés au graphique. Rappelons ici que nous nous intéressons à l'extraction de la formule chimique et que le problème de détection et de détachement des caractères collés au graphique ne se pose pas à ce niveau. Mais la présence de caractères touchant le graphique rend difficile cette tâche.

D'autres méthodes utilisent des techniques de filtrage morphologique pour détecter des composantes linéaires (lignes graphiques) et les séparer des autres composantes considérées comme texte. Pour extraire les chaînes de caractères dans des images de plans, Luo et al. (Luo *et al.*, 1995) utilisent des opérations morphologiques pour extraire les segments de droites. L'analyse des histogrammes de ces segments, permet de les séparer en des segments faisant partie du graphique et d'autres faisant partie du texte. Notons également l'existence d'autres travaux dans lesquels les auteurs extraient les graphiques en cherchant les lignes soit par la technique de vectorisation d'images (Dori et Wenyin, 1996) ou par une technique de propagation de distance (Kaneko, 1992).

Dans ce papier, nous présentons un système d'extraction d'une formule chimique dans un document manuscrit. L'idée de base consiste en deux étapes principales 1)une segmentation du document en zones stables et homogènes et 2)une classification permettant d'identifier la nature de la zone : texte ou graphique (formule chimique). Le système proposé est basé sur deux hypothèses principales :

– H1 : toutes les lignes d'écriture sont horizontales. Cette hypothèse est justifiée par le fait que les documents que nous traitons sont extraits d'un cahier dont les pages sont quadrillées.

– H2 : dans chaque document une et une seule formule chimique est présente.

Le reste du papier est organisé comme suit. Les différentes étapes du système seront exposées dans la section 2 : nous commençons par l'explication les opérations du prétraitement et nous présentons la méthode de segmentation adoptée et l'algorithme d'extraction de la formule chimique. Ensuite, les résultats expérimentaux sont présentés à la section 3 et nous concluons à la section 4.

2. Système proposé

Dans cet article nous proposons un système en plusieurs étapes pour extraire le schéma graphique représentant une formule chimique dans des documents manuscrits. La succession des étapes du système est illustrée dans la figure 2.

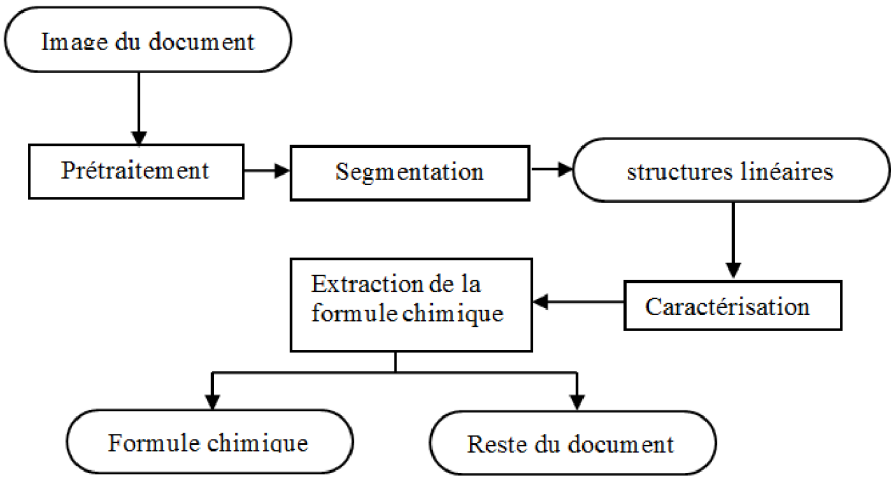


Figure 2. *Succession des étapes du système proposé*

2.1. *Prétraitement*

Dans la chaîne d'analyse d'images de documents, l'étape de pré-traitement est importante car elle influe sur toutes les étapes ultérieures. Elle permet d'améliorer la qualité des images en éliminant les défauts de mauvaise capture de documents. Ces défauts sont principalement :

– Les bordures noires.

– Le bruit impulsionnel, appelé également poivre et sel.

Le nettoyage des images de documents est constitué des opérations de filtrage suivantes :

– Suppression des composantes connexes formant les bordures en se basant sur leurs formes et leurs positions (voir figure 3)

– Suppression du bruit impulsionnel (voir figure 3) en utilisant un KFill. L'algorithme KFill utilise une fenêtre de taille $k \times k$ pixels qui doit être déplacée sur l'image entière. La suppression de bruit est effectuée en inversant les valeurs de pixels du centre de la fenêtre (régions $(k-2) \times (k-2)$ pixels). L'opération d'inversion de ces pixels est conditionnée par le nombre des composantes connexes à l'intérieur de la région et des valeurs des pixels de son contour (Al-Khaffaf *et al.*, 2008).

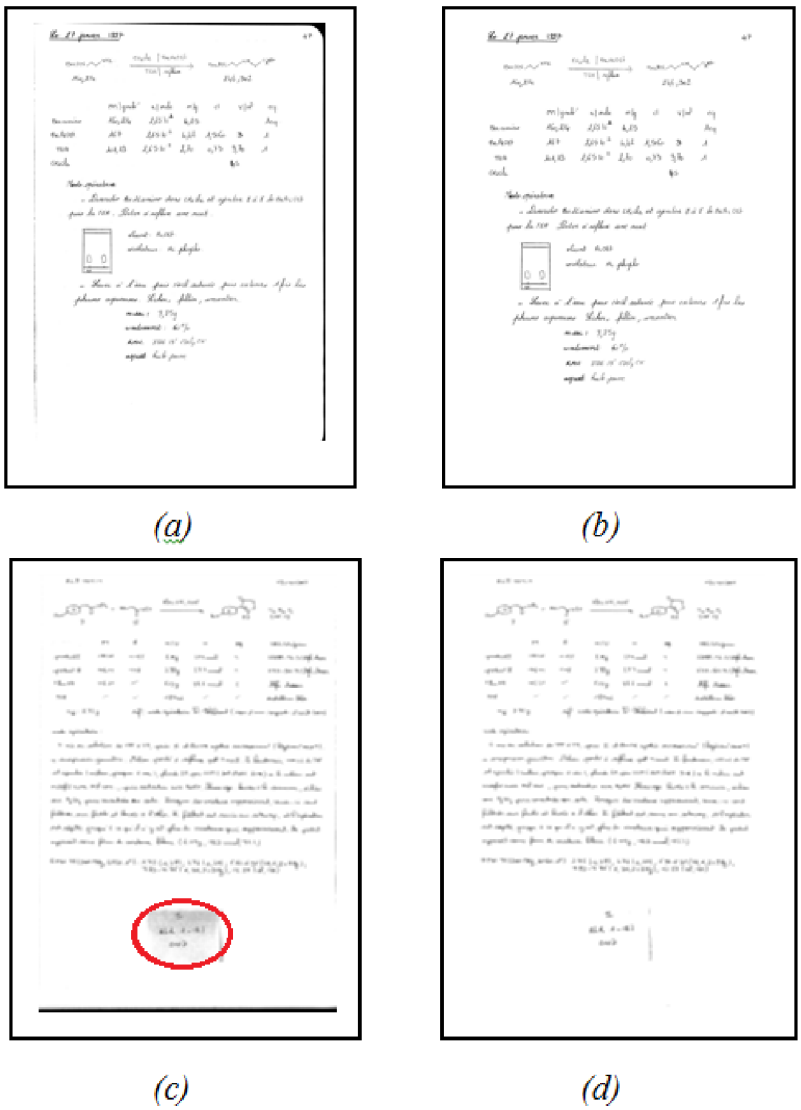


Figure 3. *Nettoyage des images de documents : (a)et (c) Images d'origine. (b) suppression des bordures noires autour de la page dans l'image (a). (d) suppression des bordures noires et du bruit poivre et sel (encerclé en rouge) dans l'image (c)*

2.2. *Segmentation des documents*

2.2.1. *Niveau de segmentation et méthode utilisée*

La structure d'un document peut être vue comme une séquence de blocs rectangulaires horizontaux contenant un schéma graphique, un tableau, plusieurs paragraphes de texte et parfois des images. Ces blocs sont verticalement séparables (voir figure 4). La notion d'alignement horizontal des composantes connexes est importante vu qu'un ensemble de composantes horizontalement alignées appartiennent à un même bloc homogène. Dans le reste de cet article, on appelle structure linéaire, l'ensemble des composantes connexes horizontalement alignées et pouvant être regroupées en se basant sur des critères de similarité, de proximité et d'orientation. L'étape de seg-

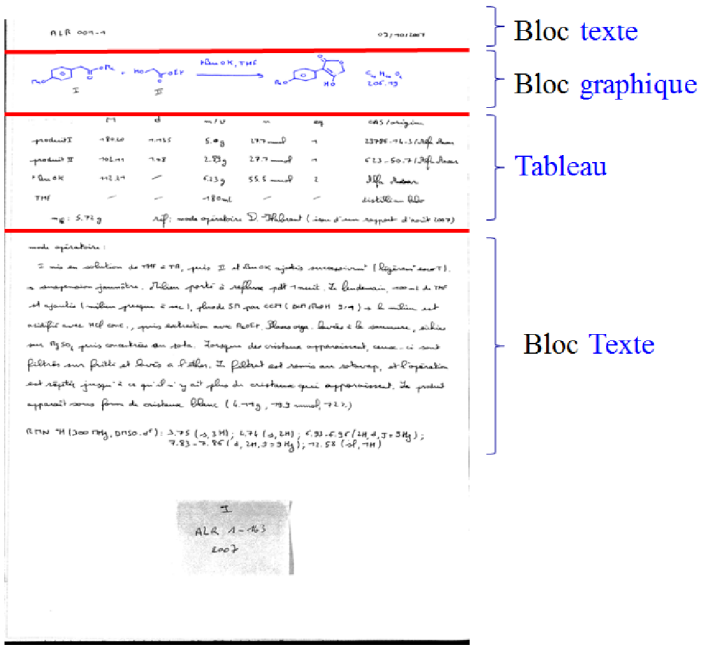


Figure 4. *Image illustrant les différentes régions contenues dans un document de chimie*

mentation consiste à subdiviser le contenu du document en zones homogènes qui vont servir pour étiqueter les parties du document en graphique ou texte. Le choix du niveau de segmentation a été guidé par la structure des documents de travail et nous avons opté pour une segmentation en structures linéaires. Pour se faire, nous avons utilisé une méthode basée sur la combinaison des composantes connexes et de la technique de lissage. Cette méthode est efficace surtout lorsqu'il s'agit de segmenter des documents contenant des textes/images alignés et séparés par des espaces blancs (Sun, 2006).

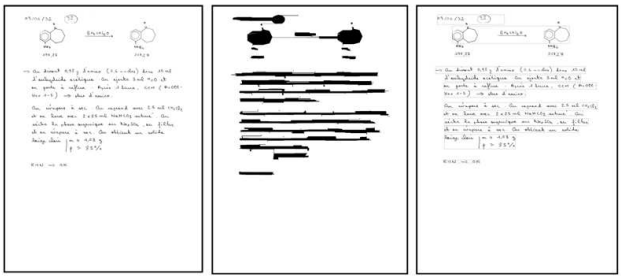
D'abord, un lissage horizontal est effectué sur l'image du document. Cet algorithme est appliqué sur une séquence binaire où les pixels blancs sont représentés par des 0, et les pixels noirs, par des 1. L'algorithme transforme une séquence binaire S_1 en une séquence S_2 selon les règles suivantes :

– Les 0 dans S_1 sont transformés en 1 dans S_2 si le nombre de 0 adjacents est inférieur ou égal à un seuil prédéfini L.

– Les 1 dans S_1 restent inchangés dans S_2.

L'algorithme permet de connecter des composantes connexes séparées par une distance inférieure à un seuil donné. Le seuil de lissage doit être suffisamment grand pour combler les espaces intra et inter composantes. Empiriquement, nous avons choisi un seuil de 200 pixels.

Ensuite, Les composantes connexes sont extraites à partir de l'image lissée. Les boîtes englobantes de ces composantes délimitent les structures linéaires dans l'image d'origine (voir figure 5).



(a) (b) (c)

Figure 5. *(a) Image nettoyée. (b) Application d'un lissage horizontal et extraction des composantes connexes dans l'image lissée. (c) Extraction de lignes dans l'image d'origine*

2.2.2. *Correction des erreurs de segmentation*

Les structures linéaires du document peuvent se toucher ou se chevaucher verticalement. Ceci peut être à l'origine d'erreur de sous-segmentation : les structures en question sont combinées en un seul bloc (voir figure 6). Nous nous sommes inspirés des travaux de Huaigu et al. (Huaigu *et al.*, 2007) pour développer un algorithme de correction de ce type d'erreur. Une connexion ou un chevauchement entre deux structures linéaires est présente sous forme d'une courte séquence horizontale de pixels noirs. L'algorithme de correction d'erreur de sous-segmentation permet de séparer les deux structures en transformant en blanc toute séquence horizontale $I(i, j_1 : j_2)$ de pixels noirs, située dans la ligne i entre les colonnes j_1 et j_2, satisfaisant la condition suivante :

$$\begin{cases} j_2 - j_1 < S_1 \\ \text{nbpixel}(i) < S_2 \end{cases}$$

S_1 est un seuil fixé à une valeur égale à 2 fois l'épaisseur du trait d'écriture. Le choix de cette valeur est expliqué par le fait que la plus longue séquence de chevauchement ou de connexion entre deux lignes adjacentes est produite par le chevauchement entre deux traits, l'un se prolongeant au-dessus de la ligne inférieure et l'autre au-dessous de la ligne supérieure. Le paramètre nbpixel(i) désigne le nombre total de pixels noirs dans la ligne i. S_2 est un seuil fixé à une valeur égale au nombre total de pixels dans la plus courte ligne contenue dans les structures linéaires extraites dans la phase précédente.

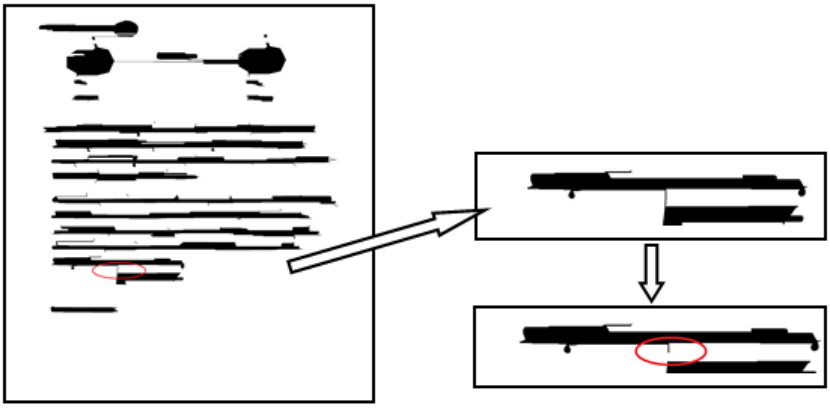


Figure 6. *Suppression d'une connexion entre deux structures linéaires*

2.3. *Caractérisation des structures linéaires et extraction de la formule chimique*

Cette étape consiste à distinguer la nature (graphique ou texte) de chacune des structures linéaires extraites dans la phase de segmentation du document. Une étape de caractérisation de ces structures est alors nécessaire pour effectuer la discrimination entre celles qui sont graphiques et celles qui sont textes, en se basant sur leurs caractéristiques. Pour choisir des descripteurs discriminants, nous nous sommes basés sur quelques caractéristiques visuelles et nous avons pu arriver aux conclusions suivantes :

Par rapport à une ligne de texte, un bloc graphique est caractérisé par :

– La présence de longues séquences de pixels noirs alignés horizontalement et/ou verticalement

– Un nombre réduit de transitions blancs-noirs horizontales, comme illustré dans la figure 7.

– La présence de plusieurs segments de droites.

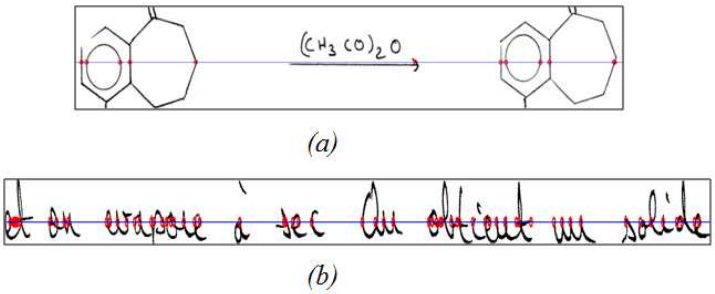


Figure 7. *En rouge, les transitions noir-blanc dans un graphique (a) et dans un texte (b)*

En se basant sur le constat décrit ci-dessus, nous avons extrait, sur chaque ligne les trois descripteurs suivants :

– mean_{rl} : la moyenne run-length horizontal et vertical. Ce descripteur est calculé à partir des deux matrices de run-length horizontale et verticale, obtenues en parcourant respectivement horizontalement et verticalement l'image de chaque ligne

– nb_{tr} : le nombre de transitions blancs-noirs horizontales,

– nb_{seg} : le nombre de segments de droites. Ce descripteur est calculé en utilisant une version probabiliste de la transformée de Hough

Pour labéliser les lignes, on calcule un score basé sur les valeurs de ces descripteurs. Ce score est décrit par la formule suivante :

$$S = \mathrm{nb}_{seg} + \mathrm{mean}_{rl} - \mathrm{nb}_{tr}$$

En se basant sur l'hypothèse que chaque document contient une seule formule chimique, la ligne qui maximise ce score correspond au bloc graphique contenant cette formule.

3. Expérimentations

le système que nous proposons ne nécessite aucun apprentissage mais il y avait quelques documents que nous avons utilisés en développement pour choisir les bons algorithmes et les paramètres inhérents. Ces documents ne sont pas utilisés dans la phase de test.

Pour évaluer les performances de notre méthode, nous avons effectué des mesures sur les résultats de chacune des deux principales étapes qui sont la segmentation en structures linéaires et l'identification du bloc graphique.

L'évaluation de la segmentation se fait suivant la mesure proposée par (Shafait *et al.*, 2008) qui nécessite une vérité terrain au niveau ligne. La vérité terrain a été réalisée à la main : chaque ligne est délimitée par un rectangle englobant, et elle est représentée par la position, la largeur et la hauteur de ce rectangle. La mesure que nous avons évaluée, pour cette étape, est le taux global d'erreur. Soit L l'ensemble de toutes les lignes dans les documents de vérité et $|L|$ le nombre de ces lignes. Le taux global d'erreur de segmentation est défini comme suit :

$$\rho = \frac{S \cup F \cup M}{|L|}$$

où S représente l'ensemble des lignes sur-segmentées. Une ligne de la vérité terrain est dite sur-segmentée si son rectangle englobant est scindé en plusieurs rectangles dans le document segmenté. F désigne l'ensemble des lignes sous-segmentées, c'est-à-dire que plusieurs rectangles englobants de lignes de la vérité terrain correspondent à un seul dans le document segmenté. M représente l'ensemble des lignes non détectées.

Une ligne de la vérité terrain est dite correctement détectée si la zone de chevauchement entre son rectangle englobant et celui de la ligne correspondant dans le document segmenté est importante. Nous considérons que ce chevauchement est important s'il représente 95% du rectangle englobant de la ligne de la vérité terrain. Nous avons utilisé 50 documents contenant 913 lignes pour l'évaluation de la segmentation. Le taux global d'erreur est de $9,5\%$. En examinant manuellement les cas d'erreurs produites dans la segmentation, nous notons que la majorité des cas sont des erreurs de sous-segmentation. Ceci est dû principalement à la présence de bruit et au chevauchement entre les lignes de textes.

La préparation de la vérité terrain pour l'évaluation de l'extraction de graphique est simple et rapide : il s'agit simplement d'étiqueter la formule chimique en graphique et le reste du document en texte. Nous avons donc évalué l'étape d'extraction de graphique sur un nombre important de document, soit 440 documents. La mesure de performance que nous avons évaluée est le taux d'extraction correcte de la formule chimique définie par le nombre de formules chimiques correctement délimitées divisés par le nombre total de formules chimiques dans l'ensemble documents. Les tests effectués ont donné un taux de $92,72\%$.

4. Conclusion

Dans cet article, nous avons proposé un système permettant d'isoler les différentes parties dans un document manuscrit. L'objectif était d'extraire une formule chimique dans un document manuscrit composite. L'idée est basée sur le fait que la formule chimique est composée principalement d'éléments graphiques, à savoir des lignes et des polygones et peut être alors distinguée du texte. La méthode est constituée d'un ensemble d'étapes permettant de nettoyer les documents, les segmenter en blocs ho-

mogènes et isoler la formule chimique en se basant sur des caractéristiques permettant une bonne discrimination de ces blocs.

Vue la qualité et la nature (manuscrit) des documents traités, une étape de prétraitement était nécessaire pour préparer les documents aux étapes qui suivent. Pour cela, un ensemble d'opérations de nettoyage basé sur le filtrage des composantes connexes et la méthode KFill a été effectué pour améliorer la qualité visuelle de ces documents et faciliter leur traitement. Pour la segmentation des documents, nous avons opté pour une segmentation en lignes, en utilisant la technique de smearing et les composantes connexes. Le choix de ce niveau de segmentation a été basé sur le fait que les différents blocs du document sont verticalement séparables et que les lignes contiennent suffisamment d'information pour être bien étiquetées. En effet, pour un niveau de segmentation plus fin, à savoir les composantes connexes ou les pseudo-mots, l'étiquetage serait plus difficile (Zheng *et al.*, 2002). Les résultats obtenus ont montré que ce niveau de segmentation est une bonne hypothèse de travail dans le cas de nos documents.

L'extraction de la formule chimique est effectuée en se basant sur un score calculé à partir d'un ensemble de descripteurs extraits sur chaque structure linéaire. Les descripteurs ont été sélectionnés pour décrire des aspects visuels permettant la discrimination entre la formule chimique et le texte.

Malgré la complexité et la diversité du contenu des documents traités, les résultats expérimentaux ont donné un taux de 92% d'extraction correcte de la formule chimique à partir d'un ensemble de 440 documents rédigés par divers scripteurs. Quelques formules chimiques n'ont pas été parfaitement délimitées à cause des composantes textuelles (noms d'atomes ou de molécules) qui sont nettement séparées du graphique et qui ont été reconnues par le système comme étant des parties des blocs texte. Nous prévoyons, à court terme, une phase de post-traitement basée sur l'analyse de voisinage de ces composantes pour affiner la délimitation de ces formules et améliorer ainsi les taux d'extraction obtenu.

5. Bibliographie

Al-Khaffaf H., Talib A. Z., Salam R. A., « Removing Salt-and-Pepper Noise from Binary Images of Engineering Drawings », *19th International Conference on Pattern Recognition (ICPR),* p. 1-4, 2008.

Dori D., Wenyin L., « Vector-Based Segmentation of Text Connected to Graphics in Engineering Drawings », *In Proceedings of 6th International SSPR Workshop, Lecture Notes in Computer Science*, vol. 1121, p. 322-331, 1996.

Fletcher L., Kasturi R., « A Robust Algorithm for Text String Separation from Mixed Text/Graphics Images », *IEEE Trans. on PAMI*, vol. 10, p. 910-918, 1998.

Huaigu C., Rohit P., Premkumar N., Ehry M., « Robust Page Segmentation Based on Smearing and Error Correction Unifying Top-down and Bottom-up Approaches », *ICDAR,* p. 392-396, 2007.

Kaneko T., « Line Structure Extraction from Line-Drawing Images », *Pattern Recognition*, vol. 25, p. 963-973, 1992.

Luo H., Again G., Dinstein I., « Directional Mathematical Morphology Approach for Line Thinning and Extraction of Character Strings from Maps and Line Drawings », *in Proceedings of 3rd ICDAR*, vol. 2423, p. 257-260, 1995.

Roy P., Pal U., Llados J., « Touching Text Character Localization in Graphical Documents Using SIFT », *Graphics Recognition. Achievements, Challenges, and Evolution, Lecture Notes in Computer Science*, vol. 6020, p. 199-211, 2010.

Shafait F., Keysers D., Breuel T., « Performance Evaluation and Benchmarking of Six-Page Segmentation Algorithms », *Pattern Analysis and Machine Intelligence*, vol. 30, p. 941-954, 2008.

Sun H. M., « Enhanced Constrained Run-Length Algorithm for Complex Layout Document Processing », *International Journal of Applied Science and Engineering,* p. 297-309, 2006.

Tombre K., Tabbone S., Pélissier L., Lamiroy B., Dosch P., « Text/Graphics Separation Revisited », *Document Analysis Systems, Lecture Notes in Computer Science*, vol. 2423, p. 200-211, 2002.

Zheng Y., Li H., Doermann D., « The Segmentation and Identification of Handwriting in Noisy Document Images », *In Proc. Document Analysis System*, p. 95-105, 2002.