Extraction de formules à partir de documents mathématiques

Formulas extraction from mathematical documents

A. KACEM[1] A. BELAID[2] M. BEN AHMED[3]

[1] ENSI-RIADI
[2]LORIA-CNRS
[3] ENSI-RIADI

77 Rue de Carthage, Cité Mohamed Ali, 2040 Radès-Tunis TUNISIE
afef.kacem@isg.rnu.tn

Résumé

Cet article propose une méthode d'extraction automatique de formules à partir des images des documents mathématiques sans passer par un système de reconnaissance optique de caractères. L'extraction se fait d'abord par repérage des symboles les plus significatifs d'une formule, puis extension aux symboles avoisinants par l'utilisation de règles contextuelles, jusqu'à la délimitation totale de l'espace de la formule. L'étiquetage est réalisé à partir de modèles crées lors d'une phase d'apprentissage utilisant la logique floue. Le taux d'étiquetage primaire des composantes connexes est de l'ordre de 95.3%. Mais leur étiquetage secondaire accroît ce taux d'environ 4%. Les résultats obtenus montrent l'applicabilité de notre système puisque 95% des formules mathématiques sont bien extraites des documents imprimés de bonne qualité. Cet article synthétise le travail effectué, pose les problèmes rencontrés et présente les résultats obtenus.

Mots clef

Extraction des formules mathématiques, Apprentissage et Modélisation des symboles mathématiques, Logique floue, Etiquetage, Classification, Segmentation du document, Analyse contextuelle.

Abstract

This paper describes a method for the automatic extraction of formulas from images of mathematical documents without using optical character recognition system. Formula extraction is first done by location of its most significant symbols, then extension to adjoining symbols using contextual rules until delimitation of the whole formula space. Mathematical symbols labelling is realised from models created at the learning stage using fuzzy logic. From the experiments, we found that the average rate of primary labelling of mathematical symbols is about 95.3% and their secondary labelling can improve the rate about of 4%. The obtained results have demonstrated the applicability of our system since 95% of mathematical formulas are well extracted from documents printed with high quality. This paper reviews the current efforts to develop such a system, presents the encountered problems and summarises the obtained results.

Keywords

Mathematical formula extraction, Training, Fuzzy logic, Symbol modelling and training, Labelling, Classification, Document segmentation, Contextual analysis.

1 Introduction

Un document mathématique contient du texte et des formules mathématiques. Contrairement au texte qui a une structure linéaire, les formules obéissent à des règles de structure spécifiques qui échappent à un lecteur optique. Afin de restituer aux formules la structure planaire, deux solutions sont souvent proposées : reconnaissance de caractères puis restructuration ou bien étiquetage puis reconnaissance. La première solution suppose que le lecteur optique réussit à segmenter les formules et est capable de fournir l'emplacement de chaque caractère. La seconde facilite le travail puisqu'elle segmente la formule en caractères avant de les reconnaître individuellement. Cette méthode évite les procédures de segmentation trop généralistes des systèmes de reconnaissance optique de caractères. Etant donné le peu de succès de la première méthode, nous avons expérimenté la seconde en opérant une segmentation adaptative du document. L'idée est d'effectuer un étiquetage en plusieurs étapes : extraction des lignes, repérage des formules isolées, du texte puis délimitation des formules à l'intérieur du texte.

Il convient de noter qu'une formule mathématique se distingue sur plusieurs points d'un texte conventionnel : bidimensionnalité, changement fréquent de fontes, symboles avec des tailles variables, des conventions de notation différentes suivant les sources. En plus, si l'on tient compte des problèmes plus génériques de bruit, de césure de ligne, la reconnaissance de formule offre un challenge. Par ailleurs, il existe deux catégories de formules : celles qui sont isolées du texte et d'autres qui y sont insérées ce qui complique évidemment leur reconnaissance. Mais, avant de pouvoir reconnaître une formule, il faut bien la localiser et l'extraire. Notre problème revient en fait, à comment distinguer et délimiter les formules à partir des données graphiques et leurs positions dans la page.

2 Système EXTRAFOR

Fig.1. montre l'architecture générale du système EXTRAFOR proposé pour l'extraction automatique des formules mathématiques. Il fonctionne comme suit : Etant donné l'image du document, le système extrait un ensemble de composantes connexes (CCXs). Puis, en utilisant une liste de paramètres déduits des rectangles englobant les CCXs, le système attribut à chacune d'elles une étiquette en fonction du rôle qu'elle peut jouer dans la composition de la formule. Cet étiquetage primaire des CCXs permet une segmentation globale du document par extraction des lignes d'image et leur classification en lignes de texte ou en lignes de formules isolées du texte. Pour les formules insérées dans le texte, une segmentation locale est nécessaire. On procède par étiquetage secondaire des CCXs afin de lever certaines ambiguïtés observées lors de leur étiquetage primaire. Une fois, les opérateurs mathématiques sont bien identifiés, on analyse puis on étend leur contexte de manière à séparer les formules du texte pur. Nous allons décrire dans la suite chacune de ses étapes.

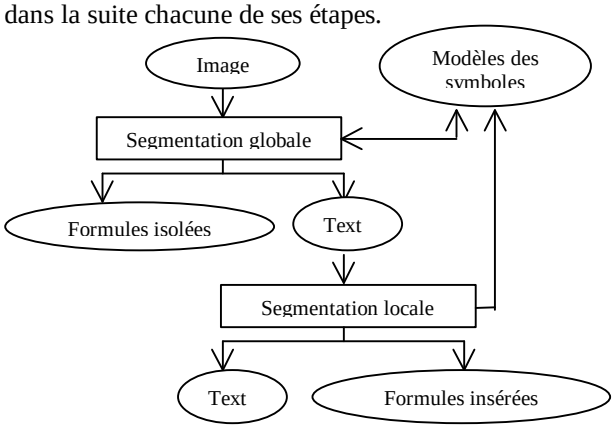

Fig.1. Architecture générale du système EXTRAFOR

3 Segmentation globale

Il s'agit du premier niveau de segmentation du document où il faut isoler les formules placées hors texte des lignes textuelles. Le document est, d'abord, segmenté en CCXs. Un étiquetage primaire des CCXs est ensuite effectué. Puis, les CCXs, horizontalement adjacentes, sont groupées dans une même ligne. Les lignes d'image, ainsi extraites, sont par la suite étiquetées en lignes de formules isolées du texte à la base de leur morphologie et mise en page. Les lignes restantes consistent en un mélange de texte pur et du texte contenant des formules mathématiques.

3.1 Extraction des CCXs

Lors de cette étape, le document est numérisé, son image est redressée et ses CCXs sont extraites. Ces CCXs constituent, en fait, la donnée de base à partir de laquelle notre système va démarrer son analyse. Chaque CCX est décrite par les coordonnées des coins, supérieur gauche (X_{min}, Y_{min}) et inférieur droit (X_{max}, Y_{max}) de son rectangle englobant et le nombre de ses pixels noirs (NPN). A partir de la hauteur : H(CCX) et la longueur : L(CCX) de la CCX, les paramètres suivants sont déterminés :

- Ratio : R(CCX)=L(CCX)/H(CCX),
- Surface : S(CCX) = L(CCX)*H(CCX),
- Densité : D(CCX)= NPN(CCX)/S(CCX).

Après extraction des CCXs, il convient de restreindre les traitements ultérieurs à celles susceptibles d'êtres parmi les opérateurs des formules mathématiques. Ainsi, on peut améliorer la précision et la vitesse de leur extraction. L'opération de filtrage des CCXs se

base sur le calcul de surface et de ratio afin d'écarter les CCXs trop petites assimilées à du bruit, des signes diacritiques ou de ponctuation et les CCXs très larges qui correspondent aux graphiques et aux séparateurs verticaux ou horizontaux puisqu'il est improbable qu'ils fassent partie des formules mathématiques.

3.2 Etiquetage primaire

Il s'agit de repérer les CCXs représentant des opérateurs mathématiques. En effet, lors de cette étape, une étiquette est attribuée à chacune d'elle en fonction du rôle qu'elle peut jouer dans la composition de la formule. Nous avons considéré une formule mathématique (FM) comme étant un ensemble d'opérandes et d'opérateurs explicites ou implicites. Les opérateurs explicites sont représentés par des symboles mathématiques tels que : $\sum$, $\int$, Π, etc. Tandis que les opérateurs implicites sont indiqués par l'arrangement spatial de leurs opérandes comme les indices et les exposants. Soit SM, l'ensemble d'opérateurs explicites, alors SM={SF, SI, SR, BFH, GDV, PD, OB}. Cela inclut les symboles les plus caractéristiques des formules mathématiques comme les signes fonctionnels (SF : sommation ou produit), les signes d'intégrale (SI), les racines (SR), les barres de fraction horizontales (BFH), les petits délimiteurs (PD), les grands délimiteurs verticaux (GDV) et les opérateurs binaires (OB).

Nous avons considéré que les signes moins dans l'ensemble OB car les autres opérateurs tels que +, *, /, <, > peuvent être facilement confus avec les caractères du texte. Pour la même raison, nous avons exclu de la liste des symboles mathématiques d'autres opérateurs comme $\forall$, $\exists$ et les lettres grecques α, β, χ, δ, etc. L'ensemble d'opérateurs implicites OI={ID, EX} contient les indices et les exposants.

L'étiquetage est réalisé à partir de modèles crées lors de la phase d'apprentissage du système. Nous allons présenter la phase de modélisation de symboles mathématiques utilisée lors du premier niveau de segmentation pour identifier les opérateurs explicites. Nous décrivons par la suite notre approche pour repérer les opérateurs implicites lors du second niveau de segmentation du document.

Avant d'entamer la phase de modélisation de symboles mathématiques, nous avons essayé de dégager quelques caractéristiques permettant de les distinguer des caractères du texte. Nous avons constaté que cette distinction peut se faire à la base de la topographie (position par rapport à la bande centrale), morphologie (ratio) et typographie (surface et densité) de leurs CCXs (Voir Table 1).

Table.1. : Classification des CCXs

Topographie	Morphologie	Typographie	SM
Débordante	Carré, Grande	Agrandie, Normale	SF
Débordante, Ascendante	Large	Agrandie	SR
Débordante	Très étendue	Agrandie	SI
Centrée	Très allongée	Dense, Très dense, Agrandie, Normale	BFH
Débordante	Très étendue	Agrandie	GDV
Ascendante	Etendue	Normale	PD
Centrée	Allongée	Très dense, réduite	OB
Descendante, Profonde	Quelconque	Densité quelconque, Surface normale, réduite	ID
Ascendante, Haute	Quelconque	Densité quelconque, Surface normale, réduite	EX

Il est évident que les indices et les exposants peuvent être de morphologie et de densité quelconque puisque tout caractère peut être indice ou exposant.

3.2.1 Modélisation des symboles mathématiques

Pour une classification effective des symboles mathématiques, le système devra tout d'abord passer par une phase dite d'apprentissage. Cette phase consiste à faire analyser au système le plus grand nombre possible de symboles tirés de différents documents, afin d'extraire les plages de valeurs de chacun des paramètres de classification : ratio, densité et surface. Nous avons considéré une méthode de classification supervisée ce qui signifie que nous déterminons les classes auxquelles un symbole peut appartenir et nous fournissons un ensemble de symboles dont nous connaissons la classe, appelé par ensemble d'apprentissage. Pour chaque instance de symbole, les valeurs des paramètres ratio, surface et densité sont calculées, observées et seules leur bornes inférieures et supérieures sont retenues.

Désignons par
- P : La liste des paramètres. P={R, S, D}.
- $BI_P(SM)$: Borne inférieure du symbole SM selon le paramètre P. Ainsi, $BI_P(SM) = Min(P(SM_i))_{i=1,\ldots,E}$
- $BS_P(SM)$: Borne supérieure du symbole SM selon le paramètre P. Ainsi, $BS_P(SM) = Max(P(SM_i))_{i=1,\ldots,E}$ où E : taille de l'échantillon.

3.2.2 Résultats d'apprentissage

Pour apprendre notre système à identifier les symboles mathématiques, nous avons étudié 175 SF, 57 SR, 69 SI, 92 BFH, 116 GDV, 205 PD et 140 OB. L'inégalité des tailles des échantillons reflète la variabilité des fréquences d'occurrence de ces différents types de symboles dans les documents mathématiques. Les résultats suivants ont été obtenus (Voir table 2). Seules les bornes inférieures et supérieures des valeurs des trois paramètres ont été conservées.

Table 2 : Résultats d'apprentissage

	$R(SM)_{i=1,...,E}$		$S(SM)_{i=1,...,E}$		$D(SM_i)_{i=1,...,E}$	
SM	$BI_R(SM)$	$BS_R(SM)$	$BI_S(SM)$	$BS_S(SM)$	$BI_D(SM)$	$BS_D(SM)$
SF	0.30	1.87	41	815	0.23	0.48
SR	1.14	9.05	300	10000	0.05	0.2
SI	0.18	0.79	138	1846	0.10	0.29
BFH	9.12	100	29	1324	0.14	1
GDV	0.06	0.30	72	1011	0.14	0.70
PD	0.07	0.47	24	207	0.22	0.93
OB	4.56	15.39	9	26	0.62	1

Les valeurs des ratios et surfaces des symboles mathématiques ont été normalisées respectivement selon le ratio le plus grand et la surface la plus large.

3.2.3 Classification floue

Lors de la phase d'apprentissage, nous avons montré comment, pour chaque paramètre et chaque type de symbole, nous retenons une valeur minimale et une valeur maximale. Un problème se poserait lors de la phase d'étiquetage, dans laquelle la mesure d'un paramètre appartenait ou n'appartenait pas à un intervalle, c'est-à-dire avait un degré d'appartenance à l'intervalle de zéro ou de un. Si maintenant au lieu de ne conserver que les bornes inférieure et supérieure de chaque intervalle, on conserve l'ensemble des valeurs mesurées, on peut alors constituer les histogrammes correspondants. L'abscisse représente alors l'ensemble des classes de valeurs possibles, c'est à dire l'ensemble des valeurs mesurées découpé en intervalles de largeur régulière. Tandis que l'ordonnée indique la fréquence relative, autrement dit, le nombre de mesures appartenant à une classe, divisé par le nombre total des mesures. L'ordonnée peut être vue alors comme étant le degré d'appartenance d'une classe à un ensemble flou. Ce degré, noté $\mu_{SM}(CCX_i)$, varie entre 0 et 1. Les histogrammes générés doivent être représentatifs, donc le plus proche que possible d'une fonction continue.

Pour identifier un symbole mathématique, étant donné sa CCX, les valeurs de chaque paramètre sont calculées. En consultant les histogrammes pour chaque type de symbole, on déduit chaque fois le degré d'appartenance de la composante candidate à un type de symbole selon un paramètre. Nous prenons, par la suite, le minimum des degrés d'appartenance de la composante à un type de symbole selon les trois paramètres : ratio, densité et surface (puisqu'il s'agit d'une conjonction de paramètres). Nous gardons, enfin, le maximum des degrés d'appartenance de la composante en question aux différents types de symboles mathématiques (disjonction de symboles mathématiques). Notons par :

- $\mu_{SM,P}(CCX_i)$: Le degré d'appartenance de CCX_i à un type de symbole SM selon un paramètre P.
- $\mu_{SM}(CCX_i)$: Le degré d'appartenance de CCX_i à un type particulier de SM, défini comme suit : $\mu_{SM}(CCX_i) = Max(Min(\mu_{SM,R}(CCX_i),\mu_{SM,S}(CCX_i), \mu_{SM,D}(CCX_i))) = Max(\mu_{SF}(CCX_i),\mu_{SR}(CCX_i), \mu_{SI}(CCX_i), \mu_{BFH}(CCX_i), \mu_{GDV}(CCX_i), \mu_{PD}(CCX_i), \mu_{OB}(CCX_i))$.

Table.3. présente un exemple d'identification d'un petit délimiteur ayant un ratio égal à 0.27, une densité de 0.32 et une surface de 418.

Table3. : Exemple d'identification d'un petit délimiteur.

SM	$m_{SM,R}$**(CCX)**	$m_{SM,D}$**(CCX)**	$m_{SM,S}$**(CCX)**	m_{SM} **(CCX)**
SF	0.04	0.16	0.22	0.04
SR	0	0	0	0
SI	0.40	0	0	0
BFH	0	0.11	0.76	0
GDV	0	0.36	0.20	0
PD	0.35	0.44	0.49	0.35
OB	0	0	0	0
			m_{SM}**(CCX)**	**0.35**
			SM(CX)	**PD**

La composante s'identifie bien à un petit délimiteur bien qu'il puisse y' avoir une confusion avec la classe des symboles fonctionnels. Mais, il est clair que $\mu_{PD}(CCX) > \mu_{SF}(CCX)$.

Le recours à la logique floue par le calcul des degrés d'appartenance aux différentes classes de symboles mathématiques nous a été utile pour traduire la non uniformité de la répartition des mesures dans les classes et lever certaines ambiguïtés pouvant être observées lors de l'étiquetage primaire des CCXs.

3.2.4 Résultats d'étiquetage primaire

Pour avoir une idée sur le taux d'étiquetage primaire des symboles mathématiques, nous avons formé un échantillon de test composé de 110 SF, 12 SR, 45 SI, 56 BFH, 93 GDV, 104 PD et 40 OB. Après classification floue de l'échantillon, nous avons calculé le nombre de symboles bien étiquetés, de symboles mal étiquetés et de symboles non étiquetés. Nous avons trouvé que le taux moyen d'étiquetage primaire est de l'ordre de 95.3% (Voir Table.4.).

Table.4. : Résultats d'étiquetage primaire

	SF	SR	SI	BFH	GDV	PD	OB	Non étiqueté
SF	100%	0%	0%	0%	0%	0%	0%	0%
SR	0%	84%	0%	0%	0%	0%	0%	16%
SI	0%	0%	100%	0%	0%	0%	0%	0%
BFH	0%	0%	0%	92%	0%	0%	3%	5%
GDV	0%	0%	2%	0%	96%	0%	0%	2%
PD	1%	0%	2%	0%	2%	95%	0%	0%
OB	0%	0%	0%	0%	0%	0%	100%	0%

Nous avons, noté quelques ambiguïtés entre les caractères, les opérateurs arithmétiques et les symboles fonctionnels ou d'intégrale et les petits délimiteurs (Voir Fig.2.). Nous allons voir comment remédier à cela lors de l'étiquetage secondaire.

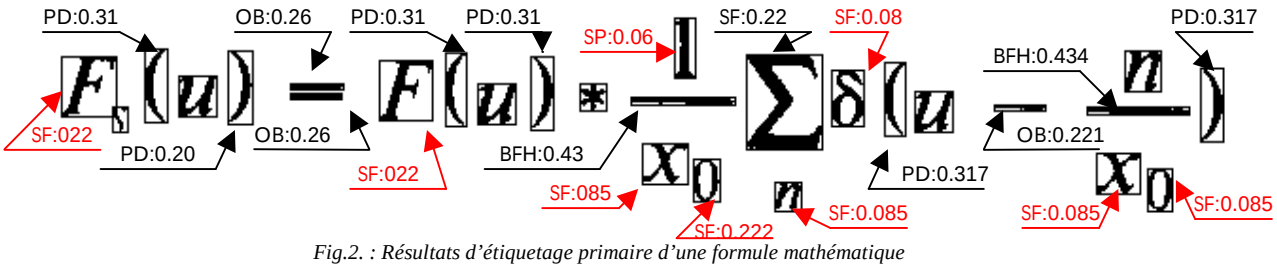

Fig.2. : Résultats d'étiquetage primaire d'une formule mathématique

3.3 Extraction des lignes

Cette étape consiste à grouper dans une même ligne les CCXs horizontalement adjacentes. Les composantes sont d'abord triées selon leur ordonnée Y_{min} croissante. Les coordonnées de la ligne, initialement égales à ceux de la CCX la plus haute à gauche, sont mises à jour à chaque fois à partir des composantes ayant une intersection commune au niveau de leur hauteur. Une fois qu'un ensemble de CCXs est associé à une ligne particulière, elles sont ordonnées selon leurs abscisses X_{min} croissante.

Soit IMG=$\{LIN_j\}_{j=1,\ldots,L}$ où L: nombre de lignes. $LIN_j=\{CCX_{i,j}/Y_{min}(CCX_{i,j})\geq Y_{min}(LIN_j)$et $Y_{max}(CCX_{i,j}) \leq Y_{max}(LIN_j)\}_{i=1,\ldots,nc}$ avec nc: nombre de CCXs appartenant à LIN_j. A présent, $CCX_{i,j}$ désigne la $i^{ème}$ composante de la $j^{ème}$ ligne. Elle est décrite par $X_{min}(CCX_{i,j})$,$Y_{min}(CCX_{i,j})$,$X_{max}(CCX_{i,j})$, $Y_{max}(CCX_{i,j})$ ainsi que son étiquette $SM(CCX_{i,j})$et son degré d'appartenance à une classe de symbole mathématique noté $\mu_{SM}(CCX_{i,j})$.

Après extraction des lignes, il convient d'assembler les CCXs des formules non linéaires comme les expressions fractionnelles, fonctionnelles ou intégrales dont les CCXs peuvent être séparées à la suite de l'étape d'extraction des lignes comme le montre la figure Fig.3. Pour cela, il faut détecter la présence des symboles fonctionnels ou d'intégration et fusionner leur limite inférieure et supérieure, de même pour les numérateurs et dénominateurs des barres de fraction.

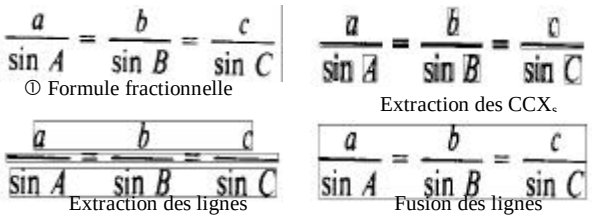

Fig.3. Exemple de fusion des lignes de CCXs

3.4 Classification des lignes

Après extraction des lignes, il est possible de repérer les formules placées hors texte à la base de la hauteur de leurs lignes généralement supérieure à la hauteur moyenne des lignes. En plus, ces formules isolées du texte sont souvent centrées, autrement dit, les distances qui les séparent des marges gauches et droites de l'image sont presque égales.

Soit HML, la hauteur moyenne des lignes, MG et MD respectivement la marge gauche et marge droite de l'image et FIS, TXT les deux étiquettes possibles de lignes, alors la séparation entre les lignes se fait comme suit :
Si($Y_{max}(LIN_j)$)>HML ou 0,5<($X_{min}(LIN_j)$-MG)/(MD-$X_{max}(LIN_j)$)<1,5) alors $LIN_j \leftarrow FIS_j$ sinon $LIN_j \leftarrow TXT_j$.
Le problème d'extraction automatique des formules isolées du texte est ainsi résolu ce qui restreint les traitements ultérieurs aux lignes textuelles pouvant contenir des formules mathématiques. De plus, nous avons décidé d'abandonner les formules trop linéaires puisqu'elles peuvent être reconnues par tout système de reconnaissance optique de caractères.

4. Segmentation locale du texte

Cela concerne les lignes textuelles. Elle vise l'extraction des formules insérées dans le texte. Pour les formules isolées, nous nous contentons de faire un étiquetage primaire des CCXs pour vérifier la présence de quelques indices permettant la fusion des lignes mal séparées tandis que pour les formules enfouies dans le texte, un étiquetage secondaire, plus fin est nécessaire à cause de l'écrasement de la formule dans le texte.

4.1 Etiquetage secondaire

Lors de cet étiquetage, nous avons tenu compte des positions des CCXs appartenant à une même ligne par rapport à la bande centrale afin de lever certaines ambiguïtés observées lors de leur étiquetage primaire. En effet, cette classification topographique des CCXs permet de distinguer les symboles fonctionnels des caractères ou chiffres de même les symboles d'intégrales des barres de fraction obliques puisque les symboles fonctionnels et d'intégrales sont des composantes débordantes alors que les caractères et les chiffres ne le sont pas. Cet étiquetage permet aussi la détection des indices et les exposants comme étant des CCXs profondes ou hautes.

4.1.1 Classification topographique

Cette classification propose six catégories de CCXs. Soit HLM: hauteur locale moyenne des CCXs appartenant à une même ligne et notons par $T(CCX_{ij})$, la classe topographique de la composante $CCX_{i,j}$. Elle est déterminée comme suit :

-$T(CCX_{i,j})$ = « Débordante » si $Y_{min}(CCX_{i,j}) < Y_{min}(BC_j)$-HLM et $Y_{max}(CCX_{i,j}) > Y_{max}(BC_j)$+HLM

-$T(CCX_{i,j})$ = « Ascendante » si $Y_{min}(CCX_{i,j}) < Y_{min}(BC_j)$-HLM et $Y_{max}(CCX_{i,j}) \leq Y_{max}(BC_j)$

-$T(CCX_{i,j})$ = « Descendante si $Y_{min}(CCX_{i,j}) \geq Y_{min}(BC_j)$+HLM et $Y_{max}(CCX_{i,j}) > Y_{max}(BC_j)$

-$T(CCX_{i,j})$ = «Centrée » si $Y_{min}(CCX_{i,j}) \geq Y_{min}(BC_j)$ – HLM et $Y_{max}(CCX_{i,j}) \leq Y_{max}(BC_j)$+HLM

-$T(CCX_{i,j})$ = «Haute » si $(Y_{min}(CCX_{i,j}) + Y_{max}(CCX_{i,j}))/2 \leq Y_{min}(BC_j)$ -HLM

-$T(CCX_{i,j})$ = «Profonde » si $(Y_{min}(CCX_{i,j}) + Y_{max}(CCX_{i,j}))/2 \geq Y_{max}(BC_j)$ +HLM

Le calcul des coordonnées de la bande centrale se fait en projetant horizontalement les ordonnées (y_{min}, y_{max}) des rectangles englobant les $CCX_{i,j}$ appartenant à une même ligne LIN_j. Ainsi, l'ordonnée inférieure : $y_{min}(BC_j)$ et supérieure : $Y_{max}(BC_j)$ de la bande centrale correspondent respectivement à la valeur maximale des projections des y_{min} et des y_{max} des CCXs.

Toutefois, quelques ambiguïtés persistent entre quelques caractères comme 'l', 't' et les petits délimiteurs puisque tous sont des composantes ascendantes. Pour éviter ce genre d'ambiguïtés, nous avons fixé un seuil de degré d'appartenance égal à 0.20 à la classe des petits délimiteurs.

Une fois que des symboles mathématiques sont bien étiquetés, les valeurs de leurs paramètres (ratio, surface et densité de leurs CCXs, sont prises en compte dans les histogrammes générés pour actualiser les degrés d'appartenance aux différentes classes ce qui rend incrémentale la phase l'apprentissage des symboles mathématiques.

Concernant les opérateurs implicites incluant les indices et les exposants, il est à remarquer que la classification topographique ne classifie pas toujours les indices comme étant des composantes profondes ou hautes. Parfois, les indices peuvent avoir des CCXs descendantes ou ascendantes. Pour traiter ces cas, une phase d'apprentissage des opérateurs implicites est nécessaire.

4.1.2 Apprentissage des opérateurs implicites

Pour pouvoir identifier les indices et les exposants, nous avons considéré deux autres paramètres :

- La taille relative des CCXs, représentée par le paramètre : X = HD/HG, où HD est la hauteur de la composante droite alors que HG est la hauteur de la composante gauche.
- La position relative des CCXs, représentée par le paramètre : Y = D/HG, où D est la distance qui sépare le sommet de la composante droite du bas de la composante gauche.

Table.5 montre les résultats obtenus à la suite de la phase d'apprentissage des opérateurs implicites selon les paramètres taille et position relative des CCXs successives. Soit F={X,Y} la liste des paramètres concernant les opérateurs implicites et OI={ID, EX}, la liste d'opérateurs implicites alors $BI_F(OI) = Min(F(OI_i))_{i=1,\ldots,S}$ tandis que $BS_F(OI_i))_{i=1,\ldots,S}$ avec S est la taille de l'échantillon (S=100).

Table.5. : Résultats d'apprentissage des opérateurs implicites

OI	$BI_X(OI)$	$BS_X(OI)$	$BI_Y(OI)$	$BS_Y(OI)$
ID	0.11	1.27	-0.21	0.76
EX	0.15	1.62	1.03	2.82

4.1.3 Classification d'opérateurs mathématiques

L'algorithme d'étiquetage secondaire fonctionne comme suit :

```
Si (T(CCX_{i,j}) = «Débordante » et (SM(CCX_{i,j}) ⊄{SF, SI, GDV}))
        Alors OP(CCX_{i,j}) = ∅
        Sinon OP(CCXi,j)=SM(CCXi,j)
Sinon
Si (T(CCX_{i,j}) = « Centrée » et (SM(CCX_{i,j}) ⊄{BFH, OB}))
        Alors SM(CCX_{i,j}) = ∅
        Sinon OP(CCXi,j)=SM(CCXi,j)
Sinon
Si (T(CCX_{i,j}) = « Ascendante »)
        Si (SM(CCX_{i ,j} ⊄{PD})
                Alors SM(CCX_{i,j}) = ∅
                Sinon OP(CCXi,j)=SM(CCXi,j)
Sinon
        Si (OI(CCXi,j) = {EX} et μ_OI(CCX_{i,j})>0.5)
                Alors OP(CCXi,j)=OI(CCXi,j)
                Sinon OP(CCXi,j)= ∅
        Fin si
Sinon
Si (T(CCX_{i,j}) = «Profonde»)
        Alors   OP(CCX_{i,j})={ID}
Sinon
Si (T(CCX_{i,j}) = «Haute»)
        Alors   OP(CCX_{i,j})={EX}
Sinon
Si (T(CCX_{i,j}) = « Descendante »)
        Si (OI(CCXi,j) = {ID} et μ_OI(CCX_{i,j})>0.5)
                Alors OP(CCXi,j)=OI(CCXi,j)
                Sinon OP(CCXi,j)= ∅
        Fin si
Fin si
```

Pour les opérateurs implicites, seuls ayant un degré d'appartenance supérieure à 0.5 sont retenus.

Nous montrons au niveau de la Table.6 l'apport de l'étiquetage secondaire sur un exemple d'une formule insérée dans le texte donnée en Fig.4.

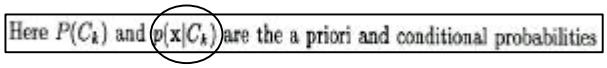

Fig.4. Exemple de formule insérée dans le texte

$SM(CCX_{i,j})$	$\mu_{SM}(CCX_{i,j})$	$T(CCX_{i,j})$	$OI(CCX_{i,j})$	$\mu_{OI}(CCX_{i,j})$	$OP(CCX_{i,j})$	$\mu_{OP}(CCX_{i,j})$
SF	0.22	Descendante	ID	0.04		
PD, GDV	0.35, 0.02	Ascendante	EX	0.37	PD	0.35
		Centrée				
		Ascendante	EX	0.39		
SF	0.30	Ascendante				
		Descendante	ID	0.51	ID	0.51
PD	0.35	Ascendante	EX	0.08	PD	0.35

Table.6. Apport de l'étiquetage secondaire

4.2 Analyse contextuelle locale

Nous allons montrer comment utiliser cet étiquetage pour localiser des formules complètes insérées dans le texte. Une formule est vue à présent comme étant un ensemble de sous expressions ayant la possibilité de s'étendre à droite et à gauche. Initialement, ces sous expressions ne sont en fait que les opérateurs mathématiques de la formule. Ensuite, par fusion successive ascendante, ces opérateurs vont inclure leurs opérandes ainsi que les sous expressions voisines de manière à séparer les formules du texte pur. Quelques heuristiques, ont été appliquées pour étendre le contexte des opérateurs mathématiques, identifiés à la suite de l'étiquetage secondaire des CCXs, et former les sous expressions correspondantes.

Selon $op(CCX_{i,j})$ **Faire**

- ID/EX : grouper son voisin le plus proche.
- SR : joindre toutes les CCXs incluses dans la racine.
- BFH : joindre les CCXs en dessus et en dessous de la barre.
- SF/SI : inclure leurs limites inférieures et supérieures, ainsi que la première composante de leur sous expression.
- GDV : inclure toutes les CCXs délimités par une paire de grands délimiteurs.
- PD : inclure les CCXs délimités par une paire de petits délimiteurs s'ils renferment un opérateur mathématique ou un nombre réduit de caractères.
  Joindre aussi les CCXs qui sont plus proches à gauche de la formule parenthèsée qu'aux CCXs qui les précèdent.
- OB : grouper les CCXs en chevauchement ou en adjacence à droite et à gauche avec le signe moins.

Fin Selon

4.3 Extension du contexte

Il s'agit d'appliquer des règles de fusion afin d'assembler les sous expressions ainsi trouvées.

- Deux formules, horizontalement adjacentes ou en chevauchement, constituent une seule formule.
- Deux formules, séparées par un nombre réduit de caractères (inférieure ou égale à 5), forment une seule formule.

5 Résultats d'extraction des formules

Nous avons développé un prototype pour la segmentation automatique des documents mathématiques sur un PC de type Pentium II. L'environnement logiciel adopté pour la réalisation du système ExtraFor est le Microsoft C/C++ Version7 sous l'environnement graphique Windows 95.

Nous avons mené des expériences d'extraction automatique des formules sur une variété de documents mathématiques (près de 50 documents à complexité variable). Les images des documents sont scannées à une résolution de 300dpi. Pour apprendre notre système à classifier les opérateurs mathématiques, nous avons utilisé un échantillon de 854 symboles mathématiques et 200 opérateurs implicites. Pour l'évaluation de notre classificateur, nous avons utilisé un échantillon de test composé de 460 symboles mathématiques et une centaine de formules insérées dans le texte. Les résultats obtenus indiquent qu'environ 95% des formules sont bien extraites d'images de documents mathématiques (Voir Fig.5)

Les principales erreurs sont dues aux confusions des indices et des exposants avec les signes diacritiques ou de ponctuation, les signes moins avec les traits d'union ainsi que les petits délimiteurs avec la lettre 'l' ou le chiffre'1' (Voir Fig.5).

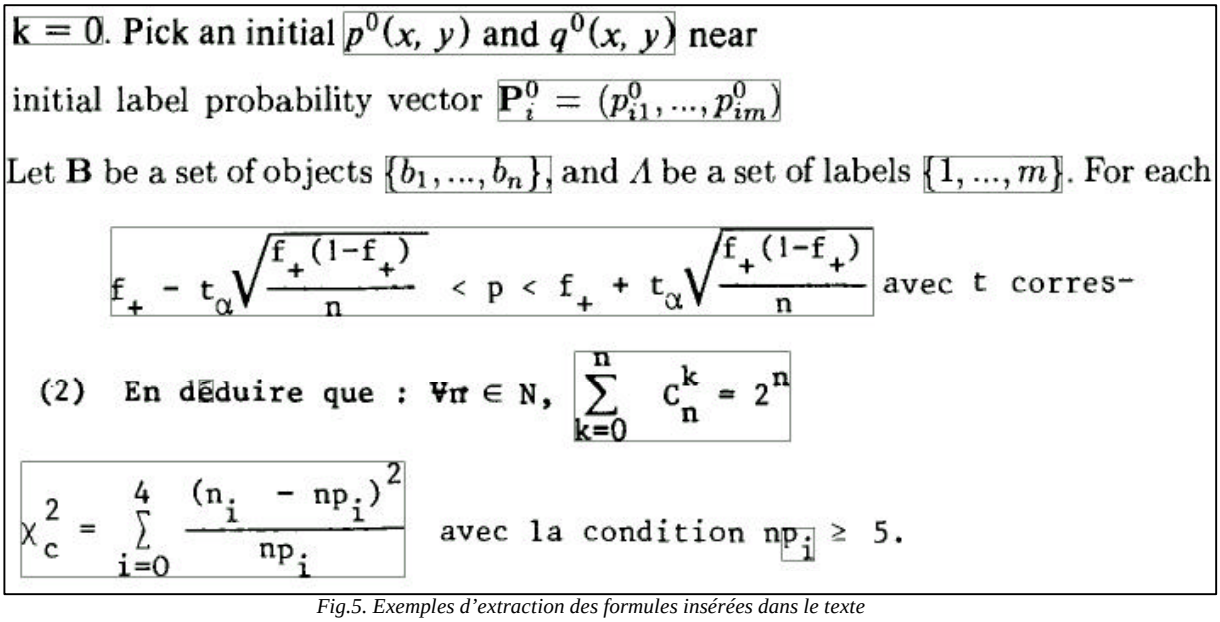

Fig.5. Exemples d'extraction des formules insérées dans le texte

6 Conclusion

Dans cet article, nous avons proposé une méthode permettant d'extraire les formules à partir de l'image d'un document mathématique d'une manière automatique et sans passer par un système de reconnaissance optique des caractères. Nous avons montré que l'introduction de la logique floue au niveau de la phase d'apprentissage du système nous a permet d'étiqueter les CCXs. Cet étiquetage nous a été utile pour identifier les symboles et par conséquent leurs formules par une analyse contextuelle de leurs CCXs. Ainsi, nous avons pu les séparer des autres composantes rédactionnelles du document.

Afin d'améliorer la méthode présentée, nous envisageons des recherches ultérieures concernant notamment les aspects suivants : Le traitement d'alignements plus complexes de symboles et l'élaboration de tests d'efficacité et de performance de cette méthode à l'aide d'une plus large base de données de formules mathématiques.

Références

[1] R.H. Anderson, "Two-Dimensional Mathematical Notation", *Syntactic Pattern Recognition Applications*, K.S. Fu, Ed. Springer Verlag, NewYork , 1977, pp. 147-177.

[2] A. Belaïd, and J. P. Haton, "A syntactic Approach for Handwritten Mathematical Formula Recognition", *IEEE Trans. PAMI, vol.6. n°1*, 1984, pp. 105-111.

[3] A. Grbavec, and D. Blostein, "Mathematics Recognition Using Graph Rewriting", *ICDAR'93*, France, 1995, pp.417-421.

[4] A. Grbavec, and D. Blostein, *Handbook of character recognition and document image analysis*, world scientific publishing company, 1997, pp. 557-582.

[5] H. J. Lee, and M. C. Lee, "Understanding Mathematical Expression in a Printed Document", *ICDAR'93*, Japan, 1993, pp. 502-505.

[6] M. Okamoto, and B. Miao, "Recognition of Mathematical Expressions by Using the Layout Structures of Symbols", *ICDAR'91*, France, 1991, pp. 242-250.

[7] H. M. Twaakyondo, and M. Okamoto, "Structure Analysis and Recognition of Mathematical Expressions", *ICDAR'95*, Canada, 1995, pp. 430-437.

[8] J. Ha, R. M. Haralick, and I. T. Phillips, "Understanding mathematical expressions from document images", *ICDAR'95*, Canada, 1995, pp. 956-959.

[9] Z. X. Wang, and C. Faure, "Structural analysis of handwritten mathematical expressions", *ICPR'88*, Washington, 1988, pp. 32-34.

[10] S. Lavirotte, and L. Pottier, "Optical formula recognition", *ICDAR'97*, Germany, 1997, pp. 357-361.

[11] K. Inoue, R. Miyazaki, and M. Suzuku, "Optical recognition of printed mathematical documents", *ATCM'98*, 1998, pp.

[12] H. J. Lee, J. S. Wang, "Design of mathematical expression recognition system", *ICDAR'95*, Japan,1995, pp.1084-1087.

[13] R. Fateman, T. Tokuyasu, B. Berman and N. Mitchell, "Optical Character Recognition and Parsing of Typeset Mathematics", *J. of Visual Commun. And Image Representation* vol 7 no. 1, March 1996, pp. 2-15.

[14] A. Kacem, A. Belaïd, and M. Ben Ahmed, "EXTARFOR : Automatic EXTRAction of mathematical FORmulas", *ICDAR'99*, Inde, 1999, pp. 527-530.