Caractérisation d'objets mathématiques et redondance graphique pour la lecture automatique de documents mathématiques

J.-Y. Toumit, S. Garcia-Salicetti, H. Emptoz
RFV – Insa de Lyon
bât. 403, 20, av. A. Einstein
69621 VILLEURBANNE CEDEX
Tel : 33.4.72.43.80.96, Fax : 33.4.72.43.80.97
e-mail : [jtoumit – sonia – emptoz] @rfv.insa-lyon.fr

Résumé

La rétroconversion des manuels scolaires est aujourd'hui un problème important pour les éditeurs ; nous travaillons actuellement dans ce contexte à la rétroconversion des documents et ouvrages de mathématiques. A notre connaissance, il n'y a que peu de travaux sur l'ensemble du document mathématique, seules des études concernant l'analyse des formules de mathématiques ont été développées à ce jour. C'est la raison pour laquelle nous posons le problème de la lecture automatique de ces documents.

Ceux-ci contiennent deux types d'informations de natures différentes : le texte et les objets mathématiques. Afin de traiter le texte plus efficacement, nous sommes conduits à séparer ces deux types d'informations ; dans cet article, nous nous intéressons particulièrement à cette étape qui peut être abordée comme un problème de segmentation multi-langages. Les méthodes classiques de segmentation ne donnant pas des résultats satisfaisants, nous avons été conduits à préconiser de nouvelles voies de segmentation physique et logique, de bas niveau.

Elles s'appuient en particulier sur la redondance graphique de caractères dans le texte, et la propagation autour de marqueurs que nous introduisons. Pour cela, une définition du texte mathématique est proposée, ainsi qu'une première classification des objets mathématiques le composant. Nous détaillons plus spécifiquement les techniques de redondance et la détection d'une certaine classe de formules mathématiques.

Ce travail est réalisé dans le cadre d'un contrat industriel avec la société PRITEC de Toulouse et un appui de l'ANVAR.

Mots-clefs

Rétroconversion, formules mathématiques, objets mathématiques, redondance graphique, segmentation

I. La lecture des documents mathématiques : un défi à relever

A. La problématique générale

La lecture automatique de documents mathématiques est à ce jour un problème très original, dans la mesure où les recherches effectuées s'attaquent plutôt au problème de la lecture automatique de formules mathématiques isolées ([1], [2]). Ainsi, le document mathématique est rarement vu dans son intégralité ([3], [4]). Aborder le document mathématique en tant que tel pose en effet plusieurs problèmes complexes, à un niveau plus global que celui de la reconnaissance de formules mathématiques: comment séparer le texte proprement mathématique du texte standard? Et, de surcroît, en amont de cette dernière question, comment définir le texte mathématique?

Ce sont les deux premières tâches que nous nous sommes données dans cet article.

B. Définitions

Notre but premier consiste à segmenter le document mathématique en texte standard et texte mathématique ; à ces fins, nous définissons le texte mathématique par rapport à la notion générique d'**objet** mathématique. L'**objet** mathématique est défini comme étant l'unité fondamentale du texte mathématique ; il a de nombreuses déclinaisons, des plus simples aux plus complexes ; ces dernières sont à leur tour composées d'**objets** agencés selon des règles propres à une certaine grammaire.

Dans la suite, nous appelons **texte mathématique** les objets mathématiques parsemés dans le document. Le texte mathématique comprend en effet des **formules**, des **abréviations** mathématiques (*sin* désigne ainsi la fonction sinus), des **signes** mathématiques particuliers (+, -, >, <, signe d'intégration, traits de

fractions...), puis des **caractères** d'origine diverses (lettres grecques, lettres latines), dont le sens au sein du texte mathématique est modulable selon des règles spatiales (**exposants, indices**). Ainsi, nous assumons que la notion d'**objet** mathématique a toutes ces multiples déclinaisons qui constituent une **hiérarchie** d'objets. Le terme de **hiérarchie** porte ici tout son sens, dans la mesure où les **objets** mathématiques, comme dans toute grammaire, ne fonctionnent pas sur le même plan.

Dans notre approche, une **formule** mathématique est à la fois considérée comme étant un **objet mathématique** et un agglomérat d'**objets** mathématiques, à structure **spatiale**, par opposition au texte standard à structure linéaire, organisés dans le plan. Remarquons que parmi ces objets mathématiques interviennent les **opérateurs** mathématiques. Nous appelons **opérateur** mathématique une entité mathématique qui a un rôle d'**agencement** d'objets mathématiques, comme le sont les opérateurs arithmétiques (qu'ils agissent sur deux objets ou sur un seul), les signes d'égalité et d'inégalité, et autres signes mathématiques (intégration, **abréviations** mathématiques, dérivées partielles...). Ces **opérateurs** ont en effet un rôle grammatical privilégié dans le texte mathématique, que nous exploitons pour effectuer l'étiquetage d'**objets** mathématiques dans le texte.

On appelle **marqueur** mathématique un **opérateur** dont la détection par des techniques bas niveau est très simple et rapide. Le rôle de ces **marqueurs** est précisément de localiser de façon efficace des "zones" mathématiques du texte.

Les **mots** du texte mathématique sont considérés comme étant les mots qui peuvent figurer au sein de **formules** mathématiques. Le **mot** est caractérisé par le fait d'être un agencement de caractères non mathématiques, constituant un **objet mathématique**. Ces **mots** sont constitués par les **abréviations** mathématiques mentionnées ci-dessus, dont *sin*, *cos*, *ln*, *sh*, *ker* sont un maigre échantillon. Nous avons ainsi constitué un **vocabulaire** de mots mathématiques d'environs 150 mots. Suite à cette définition, remarquons que le mot "fonction" n'est pas un mot mathématique.

Finalement, sont considérés comme étant des **caractères** mathématiques ceux qui peuvent figurer au sein de **formules** mathématiques. Ces **caractères** désignent aussi bien des ensembles, des variables, des fonctions, que des classes d'équivalence, et peuvent figurer dans le texte en tant qu'exposant, indice, borne, etc. Les **caractères** mathématiques ont un spectre bien plus diversifié que les **mots**, et le texte peut en être parsemé.

Ainsi, cette classification première du texte mathématique aboutit à l'agencement décrit en Figure 1, d'**objets** mathématiques de niveaux hiérarchiques différents.

Enfin, la **formule** mathématique a ses déclinaisons que nous formulons par rapport à l'**objet mathématique élémentaire**. Ce dernier est défini comme étant la plus petite entité gardant un sens mathématique lorsqu'elle est segmentée en éléments plus simples. Ainsi, les **mots** (abréviations) sont des **objets mathématiques élémentaires**.

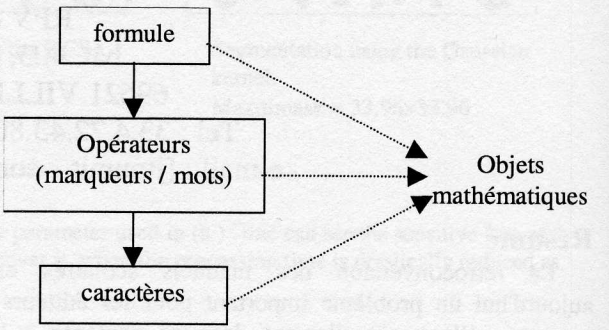

Figure 1 : agencement hiérarchique des objets mathématiques

Prenons par exemple le cas de l'objet mathématique *th(t)*. Une segmentation de ce dernier en **objets élémentaires** nous permet de retrouver l'abréviation *th*, la parenthèse ouvrante, la lettre *t* et la parenthèse fermante. Dans ce contexte, ces quatre objets sont des **objets mathématiques élémentaires**. Remarquons qu'un caractère est ou n'est pas **objet mathématique élémentaire** en vertu du contexte dans lequel il se trouve, c'est le cas des deux parenthèses.

telles que $f^{(p)}$ soit bornée

en posant $a = \psi(x)$, on a

La fonction $t \mapsto \operatorname{ch}\left(\psi\left(\frac{\tan b}{b}\right)\right)\frac{\sin b}{b}$ est définie tant

Figure 2 : Différents exemples de petites formules

Nous disposons à présent de tous les éléments pour catégoriser les **formules**. Une **petite formule** sera un agencement d'**objets mathématiques élémentaires** et d'**opérateurs** (**marqueurs/mots**), incluse dans une ligne de texte (voir Figure 2). Une **grosse formule** sera aussi un agencement d'**objets mathématiques élémentaires** et d'**opérateurs**, mais isolée du texte.

C. Stratégies

D'emblée, une stratégie de segmentation en accord avec le caractère spatial du texte mathématique s'est

imposée à nous. Le texte mathématique est composé d'**objets** mathématiques de niveaux hiérarchiques différents dont la disposition dans le plan signifie le rôle de chacun par rapport aux autres. Il est ainsi naturel que notre segmentation s'appuie sur ces significations. Nous avons donc exploité des caractéristiques bas niveau de l'organisation spatiale du texte pour effectuer une segmentation de ce dernier en éléments connexes, comme détaillé ci-dessous. Ainsi, notre approche est de ne pas effectuer de reconnaissance en amont de la segmentation.

Notre but premier est l'étiquetage, parmi les éléments issus de la segmentation, des **objets mathématiques**. A ces fins, nous avons fait le choix en un premier temps d'exploiter l'information contextuelle de type graphique, c'est-à-dire l'information contextuelle bas niveau. De plus, nous utilisons aussi l'aspect hiérarchique du texte mathématique : en effet, par propagation autour de **marqueurs**, identifiés graphiquement, nous sommes en mesure d'émettre des hypothèses sur la présence d'objets mathématiques dans leur voisinage. Lorsque ces deux stratégies sont couplées, un étiquetage efficace d'éléments du texte en objets mathématiques est possible : l'usage des marqueurs initialisant le processus, l'appariement d'objets mathématiques déjà étiquetés avec d'autres objets parsemés dans le texte permet de propager des hypothèses.

Finalement, nous avons mis en œuvre pour ce processus d'étiquetage un système d'expertise. Ce choix est dû aux nombreuses ambiguïtés auxquelles on est confronté lors de cette tâche d'étiquetage. Cela exige une approche souple et dynamique, permettant de remettre en question les hypothèses d'étiquetage émises à tout instant. Une coopération de nombreux experts remplit cette fonction dans notre système.

D. Segmentation

1. Segmentation physique

Parce qu'elles sont les plus utilisées aujourd'hui, nous avons d'emblée testé des méthodes existantes de segmentation ascendante sur des documents mathématiques. Elles donnent toutes des résultats peu satisfaisants. La principale raison de cet échec provient de la nature même des documents pour lesquels ces méthodes ont été développées. Les documents textuels standards ont en effet une organisation mono-dimensionnelle (les caractères sont organisés en lignes horizontales) et les méthodes ascendantes tirent parti de cet agencement **linéaire** des caractères pour segmenter le document. Or, comme analysé ci-dessus, le texte mathématique a une structure spatiale (fractions, intégrales, matrices, exposants, indices…) incompatible avec une telle approche de segmentation.

Il aurait été possible d'adapter ces méthodes en relâchant les contraintes de linéarité des caractères qu'elles imposent, mais cela aurait engendré de nombreux autres problèmes et alourdi considérablement la tâche. Nous avons donc opté pour une solution radicalement différente.

Nous avons recherché des méthodes de segmentation fiables, rapides, et fonctionnant en niveaux de gris sur des documents mathématiques. Nous nous sommes alors tournés vers les méthodes de segmentation descendantes, qui utilisent la multi-résolution et une vision "éloignée" ou "floue" du document. Remarquons qu'une telle approche a l'atout d'une perception **globale** et **spatiale** du document ce qui est particulièrement adapté au cas du document mathématique, comme signalé ci-dessus.

Nous avons fait le choix de travailler en niveaux de gris ; cela permet de s'affranchir de la perte d'information inhérente à la binarisation d'une image, et n'est pas réducteur dans la mesure où une image en noir et blanc peut toujours être considérée comme une image à niveaux de gris ne contenant que deux nuances de gris (le noir et le blanc).

Ainsi, nous avons développé une méthode de segmentation descendante basée sur la multi-résolution (voir [5]). Les étapes de cette méthode apparaissent en Figure 3.

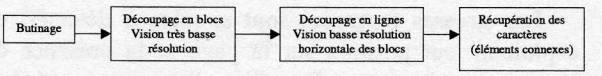

Figure 3 : étapes de notre méthode de segmentation physique

La première étape, appelée "butinage" est chargée d'évaluer rapidement des grandeurs propres au document traité (hauteur et largeur moyenne des caractères…). Ceci suppose que le document soit uniforme, ce qui est incompatible avec certaines classes de documents (sommaires de revues, etc.), mais convient à des livres ou des revues de texte mathématique.

L'étape suivante consiste à découper en blocs le document à partir d'une vision à très basse résolution de ce dernier. Ceci nous permet d'obtenir dans un premier temps les blocs de texte, puis nous extrayons les lignes en utilisant une multi-résolution directionnelle dans chaque bloc et enfin, les caractères de chaque ligne. L'utilisation d'une multi-résolution directionnelle (c'est-à-dire qui n'utilise pas la même résolution suivant les directions) permet de regrouper les caractères d'une ligne sans interférer avec les lignes avoisinantes. Notre approche de segmentation a ainsi la capacité de s'affranchir de tous les problèmes relatifs à la non linéarité des formules mathématiques lors de la segmentation du texte en lignes.

2. Séparation texte/objets mathématiques

Lorsque la segmentation physique a été achevée, il reste à effectuer la segmentation logique du document. Dans le contexte des documents mathématiques, cette segmentation logique ne se résume pas seulement à l'étude des titres et autres en-têtes, elle comprend en effet la segmentation du texte en texte standard et **objets mathématiques**.

Nous avons divisé la séparation du **texte standard** et des **objets mathématiques** en trois étapes : localisation des **grosses formules**, puis des **petites formules** et enfin des **objets mathématiques élémentaires** isolés dans le texte.

Dans la suite, le paragraphe 1 est consacré à la localisation des **grosses formules**. Quant à la détection des **petites formules** et des **objets mathématiques élémentaires**, nous utilisons, comme mentionné ci-dessus en I.C, une combinaison de stratégies, parmi lesquelles l'appariement de caractères occupe une place décisive. Nous consacrons à cette technique le paragraphe III.

II. Repérer les "grosses formules" : une première étape à franchir

A. La problématique

Les **grosses formules** sont simples à détecter, en exploitant leur position sur la page et la présence de signes caractéristiques. En effet, elles sont en général centrées et, dans presque tous les cas, n'occupent pas toute la largeur de la page ; elles comptent aussi très fréquemment de grands traits de fractions ou des matrices. Cependant, ces critères ne suffisent pas à une détection exhaustive de ces formules. Des stratégies plus fines ont donc été explorées.

B. L'espace inter-caractères : un critère important

Une étude d'un grand nombre d'ouvrages mathématiques nous a permis de conclure que, dans la grande majorité des cas, l'espacement entre les caractères diffère dans du texte simple et au sein de formules mathématiques (voir Figure 4). En effet, dans un texte standard, les caractères sont regroupés en mots, ce qui créé deux types d'espacements : un "long" qui représente un espace entre deux mots et un "court" qui représente l'espace entre deux caractères d'un même mot. Une étude statistique de ces distances inter-caractères nous permet d'extraire les mots du texte. La Figure 5 montre en effet un histogramme des distances inter-caractères sur une ligne de texte standard : on distingue un premier pic à gauche correspondant aux espaces inter-caractères à l'intérieur d'un même mot, plus nombreux que les espaces entre les mots représentés par le pic de droite.

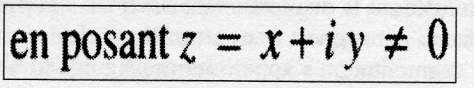

Figure 4 : espacement des caractères dans une formule mathématique

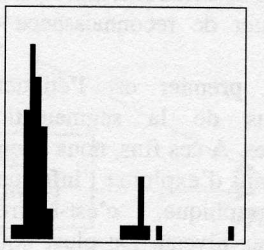

Figure 5 : histogramme des inter-caractères sur une ligne

Suivant les types d'impressions ou la résolution, cet espace critique de regroupement des caractères en mots peut varier considérablement d'un document à l'autre. Nous calculons grâce à l'histogramme des espaces inter-caractères une distance de regroupement des caractères en mots adaptée au document traité. Un exemple de regroupement en mots est montré dans la Figure 6.

Nous allons le préciser dans le cas des séries de ENGEL.

Figure 6 : exemple de regroupement en mots dans du texte standard

Les **grosses formules** ne peuvent contenir comme mot que des abréviations mathématiques. Or ces abréviations sont des mots de trois ou quatre lettres au maximum (il existe toutefois quelques abréviations assez rares de cinq ou six lettres telles que *arccos*). Nous considérons par conséquent comme "longs" des mots contenant plus de cinq lettres. Les **grosses formules** peuvent alors être détectées grâce à l'absence de mots "longs" dans une ligne. De plus, comme le nombre de caractères de la grande majorité des abréviations mathématiques ne dépasse pas trois, nous calculons la moyenne du nombre de caractères par mot sur la ligne. Si cette moyenne dépasse trois caractères, la ligne toute entière n'est pas étiquetée comme **grosse formule**, dans la mesure où une grosse formule est par définition isolée. Il existe en effet dans les documents mathématiques de nombreux mots courts qui font partie du texte standard (*pour*, *soit*, *donc*, etc.). Nous nous basons également sur la présence de fractions et d'exposants ou d'indices. Des exemples de mots détectés dans un document mathématique sont donnés Figure 7. La Figure 8 montre les **grosses formules** détectées sur une page d'un document mathématique.

On remarque sur cette figure qu'une des **grosses formules** détectées contient toutefois des mots de texte : "et donc". Ces deux mots très courts de texte ne peuvent être distingués pour l'instant des abréviations mathématiques. Ils ne seront séparés de la formule qui les suit que lorsque la reconnaissance des mots du texte aura été effectuée.

$$\varphi'(a) = \frac{a(1-\text{th}^2 a) - \text{th}\, a}{a^2}.$$

$\text{th}^2 a) - \text{th}\, a$. On a $\omega'(a) = -2a\,\text{th}\, a(1-\text{th}^2 a)$ nte sur $[0, +\infty[$ et, comme $\omega(a) = 0$, que $\omega(a)$ $), +\infty[$, donc φ est strictement décroissante.

Figure 7 : exemple de mots détectés dans une formule (abréviations mathématiques)

III. La redondance graphique des caractères : un formidable outil

Après avoir étudié la détection automatique de **grosses formules**, nous décrivons dans cette partie un des nombreux outils que nous utiliserons pour détecter les **objets mathématiques élémentaires** et les **petites formules**.

A. La problématique : quelques constatations

Comme nous l'avons déjà signalé, nous devons repérer le plus d'**objets** et de **formules mathématiques** *avant* d'avoir reconnu le texte, bien que la détection d'une partie de ces objets et formules nécessitent une phase de reconnaissance. La Figure 9 montre un exemple d'**objet mathématique** qui doit être étiqueté comme tel impérativement *avant* la reconnaissance, puisqu'il s'agit d'un caractère grec : la lettre φ. Dans ce même exemple, le φ apparaît deux fois au milieu du texte mais également à l'intérieur d'une **grosse formule**. Nous exploitons cette information contextuelle que nous appelons **redondance graphique** ; elle permet de conclure que les deux φ isolés sont en réalité le même **objet mathématique**.

Pour cela, nous devons obtenir l'information que ces trois caractères sont identiques. Comme nous ne pouvons pas les reconnaître à ce stade, nous devons nous en tenir à l'information purement graphique.

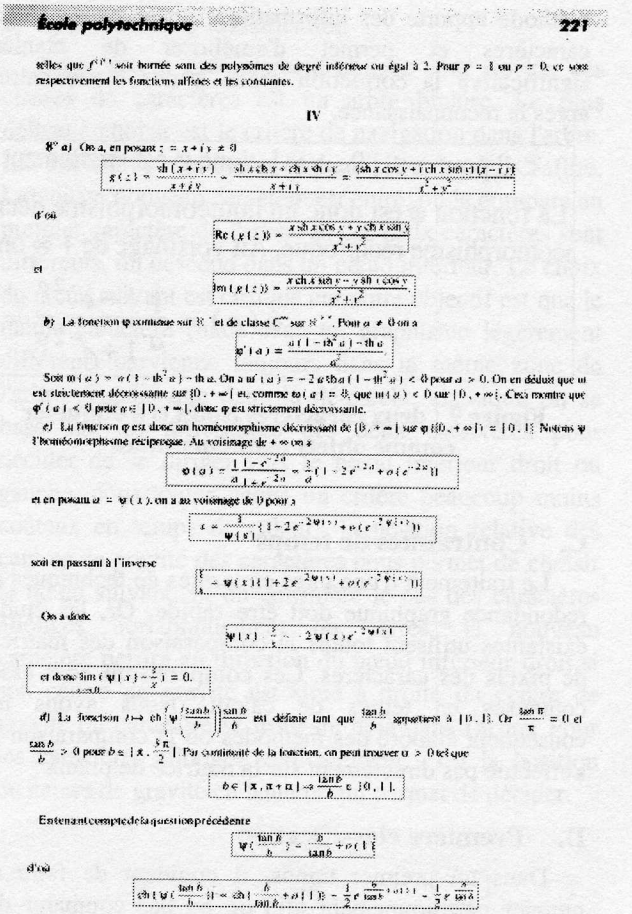

Figure 8 : exemple de grosses formules détectées

Notre but est donc de classer en utilisant les informations graphiques les caractères de la page de la manière suivante : tous les caractères identiques graphiquement doivent être regroupés en classes. Cette étape de classification doit être *rapide* et *fiable*. Comme nous utilisons la seule information graphique, il est cependant possible que deux caractères identiques soient séparés en raison de différences typographiques ou de mauvaise impression du document.

B. L'existant

Historiquement, la redondance graphique des caractères ou matching a été inventée et utilisée pour **comprimer** des fax. Depuis, d'autres recherches ([6], [7], [8]) ont montré que la redondance graphique des caractères a de nombreuses autres applications. On peut bien entendu l'utiliser pour effectuer de la **reconnaissance de caractères**. Comme elle permet de comprimer les informations graphiques du document, il y a beaucoup moins de caractères à reconnaître ce qui **accélère la phase de reconnaissance**. Enfin, cette

méthode apporte des informations sur le **contexte** des caractères et permet d'améliorer de manière significative la correction grâce à des dictionnaires, après la reconnaissance.

$) < 0$ pour $a \in]0, +\infty[$, donc φ est strictement

La fonction φ est donc un homéomorphisme décr

néomorphisme réciproque. Au voisinage de $+\infty$ on

$$\varphi(a) = \frac{1}{a}\frac{1-e^{-2a}}{1+e^{-2a}} =$$

Figure 9 : deux caractères isolés φ à étiqueter comme objets mathématiques

C. Contraintes de temps

Le traitement d'une page en termes de techniques de redondance graphique doit être rapide. Or, les études existantes utilisent toutes la comparaison des matrices de pixels des caractères. Ces comparaisons sont assez coûteuses en temps de calcul. Nous avons par conséquent élaboré des méthodes où la comparaison ne s'effectue pas directement sur la matrice de pixels.

D. Première classification

Dans un premier temps, il convient de faire un premier partitionnement afin de ne pas comparer des caractères trop différents. Cette première classification doit être très rapide puisqu'elle doit traiter tous les caractères.

Nous avons retenu deux critères très simples à calculer : la surface du rectangle englobant du caractère et sa hauteur.

Dans une première phase, nous classons les caractères par surface. De cette manière, les caractères des titres ne seront pas comparés aux caractères du texte. Cela permet également d'isoler les "petits" caractères, comme les virgules et les points.

Dans chaque classe ainsi créée, nous séparons les caractères qui n'ont pas la même hauteur. De cette manière, un "l" sera séparé d'un "m" mais un "o" restera groupé avec un "a". Afin d'adapter cette deuxième phase au document traité, nous bâtissons un histogramme des hauteurs sur la page (un exemple est présenté Figure 10). Sur cet histogramme, on distingue généralement cinq classes (de gauche à droite) : les caractères fins, les virgules, les caractères sans hampes ni jambages, les caractères avec une petite hampe ou un petit jambage puis les caractères avec une grande hampe ou un petit jambage.

Grâce à cet histogramme, nous choisissons donc les valeurs les plus adaptées au document traité (en fonction de la police de caractères utilisée), c'est-à-dire les seuils donnés par cet histogramme en vue d'une discrimination en classes.

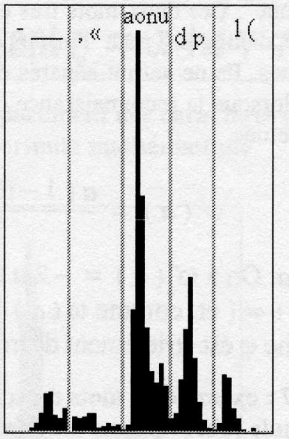

Figure 10 : histogramme des hauteurs sur une page de texte et les classes associées

E. Classification avancée

1. Comparaison de deux caractères

Nous devons ensuite comparer les caractères de taille similaire et les classer en classes de caractères identiques. Nous avons choisi de comparer les différents profils et projections des caractères entre eux. Nous calculons donc pour chaque caractère ses projections totales et partielles ainsi que ses profils. Des exemples sont donnés Figure 11.

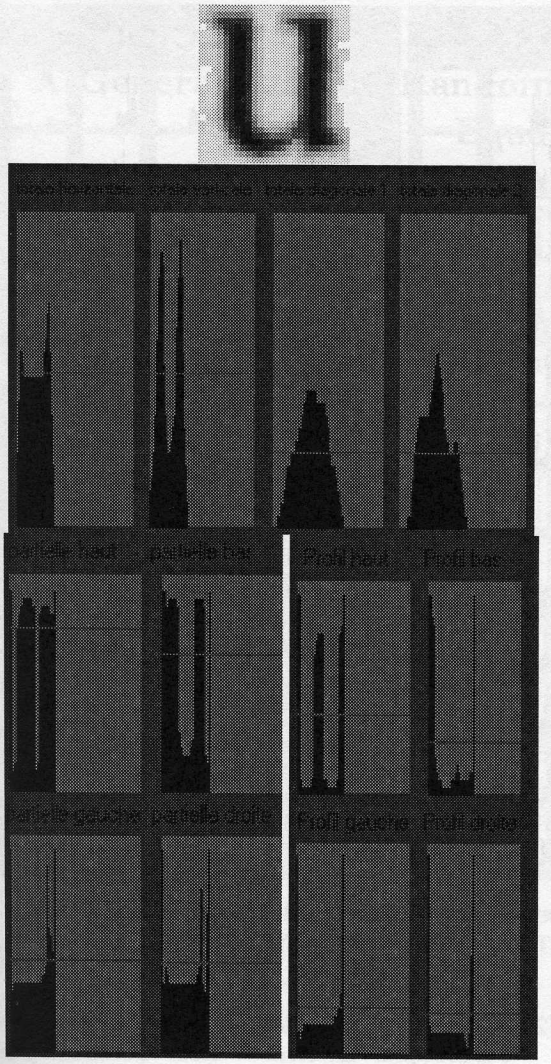

Figure 11 : un caractère et ses projections totales (à sa droite), ses projections partielles (en bas à gauche) et ses profils (en bas à droite)

La comparaison entre deux caractères s'effectue de la manière suivante : les différents profils et projections sont comparés en effectuant une différence absolue entre eux. Ensuite, plusieurs valeurs (telles que la moyenne, l'écart-type...) sont calculées à partir de ces différences et leur somme pondérée permet de décider si les caractères sont identiques ou différents. Cette pondération reflète l'importance relative de chacun des critères de caractérisation de la forme par rapport à l'objectif qui est le nôtre. La Figure 13 montre des exemples de comparaisons de caractères : un "u" et un "n" dont les profils et projections partielles seuls permettent de les distinguer ; deux "s" dont les différences de profils et de projections sont quasi nulles ; un "e" et un "c" où seuls le profil droit et la projection partielle gauche permettent de les distinguer.

2. Structure de classification

La structure généralement utilisée pour stocker ces classes de caractères est un arbre binaire. Le plus délicat à choisir est le critère de navigation dans l'arbre. Les caractères sont entrés séquentiellement dans l'arbre. Lorsqu'un nouveau caractère est entré, on le compare au premier caractère de l'arbre. Si ces deux caractères sont différents, on descend dans un nœud inférieur. Le choix du nœud suivant est critique car notre objectif est que le même caractère (mais avec un graphisme légèrement différent) parvienne toujours dans la même zone de l'arbre. Tandis que dans la démarche classique on se base sur les différences des deux caractères pour décider de se diriger vers le nœud inférieur droit ou gauche, nous avons choisi un critère beaucoup moins coûteux en temps de calcul : la position relative des centres de gravité des caractères nous permet de choisir le nœud suivant. Si on considère le cas des caractères allongés horizontalement ou carrés, le nouveau caractère partira en direction du nœud inférieur droit si son centre de gravité est situé à droite du centre de gravité du caractère se trouvant déjà dans l'arbre. Pour les caractères allongés verticalement, c'est la position du centre de gravité en hauteur qui permet de décider.

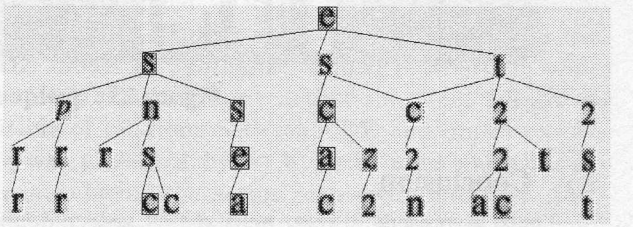

Figure 12 : exemple d'arbre généré

La Figure 12 montre un exemple d'arbre obtenu pour les caractères carrés et de surfaces proches, du texte. Nous avons bâti ici un arbre ternaire actuellement à l'étude.

F. Limites de la méthode

Nous utilisons cette technique de redondance graphique des caractères pour détecter des **petites formules** ou des **objets mathématiques élémentaires**. Certains problèmes subsistent comme les "a" (verbe avoir), les "y" (adverbe de lieu), les "s", "l", "d", etc., suivis d'un apostrophe du texte standard, qui peuvent être étiquetés comme **objet mathématique**. Pour ces cas particuliers, une reconnaissance est indispensable à un étiquetage correct.

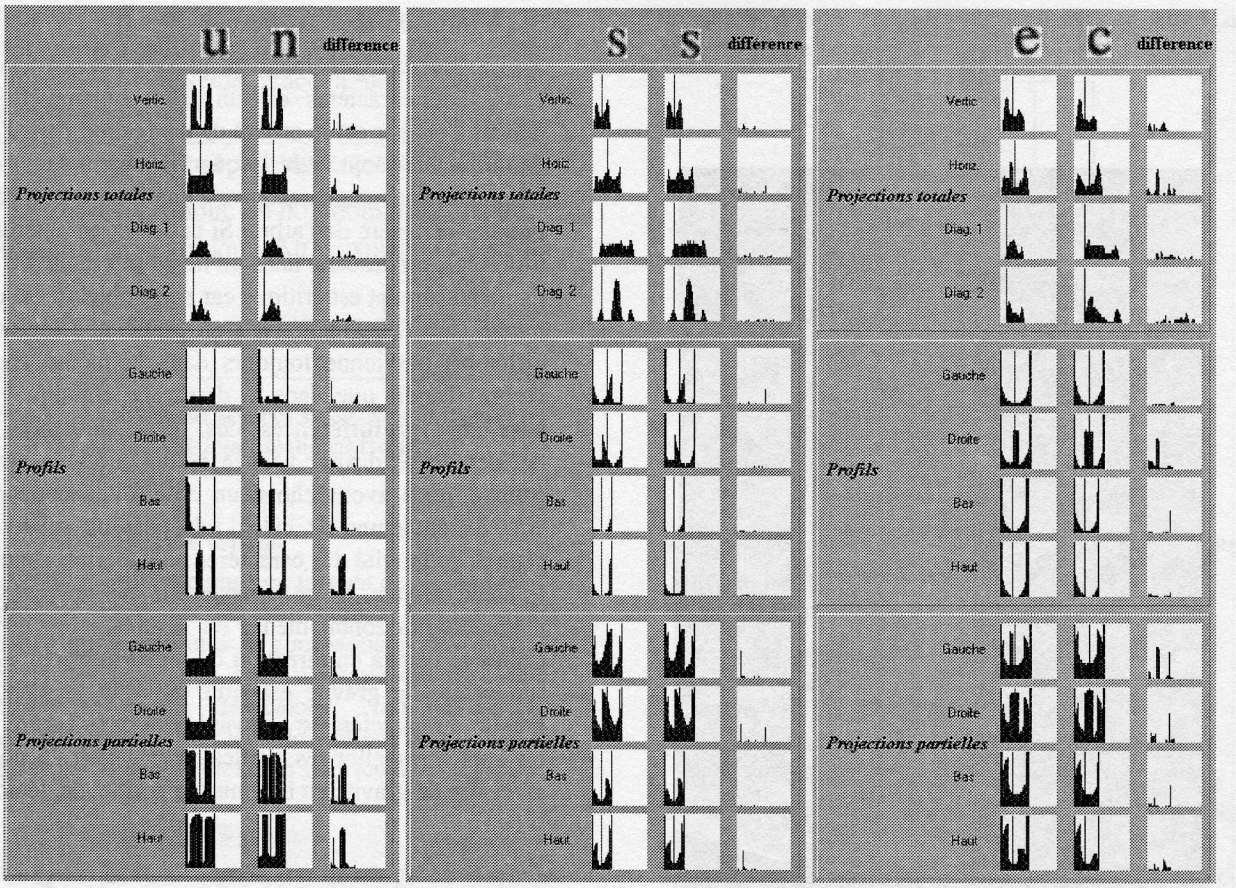

Figure 13 : quelques comparaisons de caractères

IV. Conclusion

Dans le présent article, nous avons présenté une définition du texte mathématique et une classification hiérarchique des éléments de ce dernier. En fonction d'une telle classification, des stratégies de segmentation physique et logique ont été adoptées, pour l'instant de bas niveau, et couplées à un système d'expertise. Un outil utilisable pour détecter des objets mathématiques élémentaires a également été présenté. Les limites actuelles de notre approche sont dues à l'absence de techniques de haut niveau pour segmenter logiquement le texte. Nous envisageons ainsi l'exploitation de l'information contextuelle grammaticale, la reconnaissance de mots clefs et d'abréviations. En effet, des règles d'une grammaire mathématique restent à être posées, et des retours sur segmentation sont à pratiquer suite à la reconnaissance de mots du texte. Une approche cyclique pour la lecture automatique du document mathématique sera ainsi explorée.

V. Bibliographie

[1] H.J. Lee, M.C. Lee, "Understanding Mathematical Expressions in a Printed Document", *ICDAR'93*, Tsukuba, Japon, pp. 502-505

[2] J. Ha, R. M. Haralick, I.T. Philips, "Understanding Mathematical Expressions from Document Images", *ICDAR'95*, Montréal, Canada, pp. 956-959

[3] M. Okamoto, A. Miyazawa, "An Experimental Implementation of a Document Recognition System for Papers Containing Mathematical Expressions", in *Structured Document Image Analysis*, H. S. Baird, H. Bunke, K. Yamamoto, Springer Verlag, 1992, pp. 36-63

[4] H.J. Lee, J.S. Wang, "Design of a Mathematical Expression Recognition System", *ICDAR 95*, Montréal, Canada, pp. 1084-1087

[5] J.Y. Toumit, H. Emptoz, "From the segmentation to the reading of a mathematical document", *GKPO'98*, Machine Graphics & Vision

[6] F. Lebourgeois, "Approche mixte pour la reconnaissance des documents imprimés", *Thèse de doctorat*, RFV, INSA de Lyon, 1991

[7] R.G. Casey, F.M. Wahl, K.Y. Wong, "Document analysis system", *IBM. J. Res. Develop.*, Vol. 26, n°6, Nov. 82, pp. 647-656

[8] J.Y. Toumit, H. Emptoz, "A character matching method for mathematical object detection", *RECPAD'98*, Lisbone, Portugal, p. 83