Reconnaissance de texte par ordinateur à partir des mots

Word-Based Computer Recognition of Printed Text

Charles SENAY
INRS-Télécommunications,
3, place du Commerce,
Verdun, Québec, Canada,
H3E 1H6

Charles Senay a obtenu son diplôme d'ingénieur (B. Ing.) en génie électrique de l'École Polytechnique de Montréal, Québec en 1983. De 1983 à 1986, il a travaillé pour Bell Canada dans le domaine de la commutation numérique. Il a obtenu une maîtrise (M. Sc. A-Télécommunications) à l'INRS-Télécommunications en 1990. Il travaille présentement pour Bell Canada dans le domaine des multiplexeurs intelligents.

Amar MITICHE
INRS-Télécommunications,
3, place du Commerce,
Verdun, Québec, Canada,
H3E 1H6

Amar Mitiche a obtenu une licence ès-sciences en mathématiques pures à l'Université d'Alger et une thèse de doctorat (Ph. D.) en informatique à l'Université du Texas à Austin. Son travail de thèse à l'Université du Texas à Austin concernait l'analyse et la discrimination des textures. Après un séjour au Laboratory for Image and Signal Analysis de l'université du Texas à Austin où il s'est intéressé à divers problèmes de vision par ordinateur, il est professeur à l'Institut National de La Recherche Scientifique (INRS-Télécommunications) à Montréal, Québec, Canada, depuis 1985. Il s'intéresse actuellement, surtout du point de vue méthodologique, au calcul et à l'interprétation du mouvement, la reconnaissance des formes, l'intégration de sources d'information multiples, et aux réseaux neuronaux.

SOMMAIRE

La reconnaissance de texte par ordinateur s'est traditionnellement faite à partir des caractères : on isole chacun des caractères d'un mot et, par une méthode de reconnaissance quelconque, on établit l'identité de chaque caractère. Une fois la reconnaissance de toutes les lettres d'un mot achevée, une analyse contextuelle est faite (dictionnaire, matrice de confusion, etc.). Des approches « top-down » suggèrent que la reconnaissance d'un mot peut s'établir à partir du contexte de ce mot dans la phrase. Ce contexte peut être de nature diverse : statistique, syntaxique ou sémantique.

Notre approche s'insère entre les deux approches mentionnées : on ne s'attardera ni aux caractères individuels constituant un mot, ni aux relations qui existent entre les mots d'une phrase. On s'intéressera au mot lui-même, à sa forme générale, à sa « signature graphique ». Cette signature graphique est établie à partir de caractéristiques très simples prélevées directement sur le mot (position dans le mot des ascendantes, descendantes, boucles, etc.). Chaque mot d'un dictionnaire donné est classifié selon sa signature graphique : une classe sera donc constituée de mots dont la signature graphique est identique. Pour reconnaître un mot, il s'agit donc d'extraire les caractéristiques sur le mot, de trouver dans le dictionnaire la classe à laquelle il appartient, en extraire le (ou les) mots et d'y appliquer un traitement simple pour finaliser la reconnaissance du mot.

Les résultats expérimentaux nous ont permis de démontrer que la reconnaissance de texte à partir des mots est non seulement réalisable mais qu'elle comporte des avantages : entre autres, sa grande simplicité et son efficacité dans les environnements omnifontes ainsi que son habilité à fonctionner dans les environnements bruités.

MOTS CLÉS

Vision par ordinateur, reconnaissance de formes, reconnaissance de textes, reconnaissance de mots, reconnaissance de caractères, attributs, dictionnaire.

ABSTRACT

Computer recognition of printed text has been traditionally based on characters : each character is first extracted and then recognized by one of various methods. Word recognition follows where contextual information is brought in (reference dictionary, confusion matrices, syntactic and semantic context, etc.). Our approach is based directly on the recognition of the word ; no emphasis is put on the character or on context, although these can be used. A word is represented by simple and stable features computed directly from the word (length, position of ascenders, descenders, holes, etc.). A word is recognized by matching its characteristics against those of a reference set, a hierarchically organized dictionary. A simple preferential process may be instantiated in the case of multiple matches. Experimental results have demonstrated not only the feasibility of the approach but also its advantages, simplicity, robusteness, and efficiency in an omnifont context.

KEY WORDS

Computer vision, Shape recognition, Text recognition, Character recognition, Features, Dictionary.

1. Introduction

La reconnaissance de texte par ordinateur représente toujours, après quelques décennies de recherches, un défi de taille. Bien sûr, des progrès énormes ont été accomplis : il existe aujourd'hui des systèmes de reconnaissance de texte relativement performants. Malgré cela, jamais encore les performances humaines en ce domaine n'ont été égalées. Les systèmes mécanisés sont fragiles et peu flexibles : les performances de reconnaissance peuvent chuter radicalement si l'on change la taille des caractères utilisés, si l'on utilise une fonte différente, si la qualité graphique est diminuée (ex. : photocopie), etc. Les problèmes à surmonter aujourd'hui sont bien différents de ce qu'ils étaient il y a quinze ou vingt ans. Les problèmes actuels nous incitent pratiquement à revoir les concepts utilisés jusqu'à maintenant pour faire de la reconnaissance de texte.

La reconnaissance de texte par ordinateur s'est traditionnellement faite à partir des caractères : on isole chacun d'eux, on les reconnaît individuellement et on les regroupe par la suite en mots. Cette technique est bien adaptée au contexte de reconnaissance monofonte et à une qualité graphique supérieure. Par contre, cette méthode perd de l'efficacité dans un environnement omnifonte et bruité : la grande variété de fontes disponibles complique la reconnaissance individuelle de caractères et le bruit introduit le problème de dégradation graphique des caractères (caractères mutilés, qui se touchent, etc.). Dans ce contexte, la méthode consistant à reconnaître les caractères de façon individuelle est-elle la seule possible ?

C'est précisément cet aspect que nous voulons aborder dans cette étude : nous examinerons la reconnaissance de texte fondée sur la reconnaissance directe des mots. L'élément de base à reconnaître est le mot et non plus le caractère. Nous pensons que cette nouvelle voie, peu explorée jusqu'à maintenant, est réalisable et qu'elle présente des avantages sur la méthode traditionnelle (réduction du temps de traitement, souplesse dans les environnements omnifontes et bruités, etc.).

L'utilisation de la similitude entre les mots permet d'établir une classification des mots d'un dictionnaire dont la taille et le contenu sont connus. Cette similitude graphique s'établit à partir de la position dans le mot d'attributs très simples prélevés sur le mot (ascendantes, descendantes, etc.). Les mots qui sont graphiquement similaires sont regroupés dans une même classe.

La première phase de cette étude consiste à vérifier si le nombre de mots dans chaque classe n'est pas trop élevé. Ce nombre est évalué en fonction de la taille du dictionnaire utilisé. Ce dictionnaire est constitué d'une liste de mots qui ont été regroupés en plusieurs classes ; ces classes sont déterminées par la position et le nombre d'attributs dans les mots. Nous montrerons que les résultats de cette première phase sont concluants.

La deuxième phase consiste à construire et à tester un algorithme de reconnaissance de texte à partir des mots. Le système de reconnaissance sera constitué des éléments suivants : un détecteur d'attributs, le dictionnaire de la première phase et un système d'appoint de reconnaissance de caractères individuels. Le but dans cette deuxième partie est de tester les performances systémiques de l'algorithme de reconnaissance : réduction du temps de traitement (*vs.* méthodes traditionnelles de reconnaissance), performances omnifontes et performances dans un environnement bruité.

2. Reconnaissance de texte : différentes approches

Les modélisations de systèmes sur ordinateur ont souvent été copiées sur les systèmes humains. Après tout, ces derniers sont ceux dont l'utilisation nous est la plus familière (même si nous ne savons que très peu sur leur fonctionnement réel). A titre d'exemple, la synthèse de parole par ordinateur peut se faire à partir d'une modélisation du conduit vocal par une série de filtres dont la position des pôles dépend du phonème à émettre : le déplacement des pôles de chacun des filtres correspond à la modification de la forme de chacune des parties du conduit vocal. Dans le cas de la reconnaissance d'objets, on s'inspire également des systèmes humains : la reconnaissance d'objets doit être précédée d'une identification des caractéristiques fortes de l'apparence des objets (contour externes, contours internes, textures, etc.). Avant d'aborder la reconnaissance de texte par ordinateur et de proposer notre système, il serait intéressant d'examiner quelques aspects du fonctionnement de l'être humain dans ce contexte.

Beaucoup de recherches en psychologie expérimentale ont été faites dans le domaine de la reconnaissance de textes [9], [10], [11]. Les quatre hypothèses les plus répandues sont :

a) Reconnaissance à partir des lettres : selon cette hypothèse, chacune des lettres qui compose un mot est reconnue.

b) Reconnaissance à partir des mots : cette hypothèse soutient l'idée que l'unité de reconnaissance de base est le mot (et non plus la lettre) ; on reconnaît d'un coup la forme des mots.

c) Reconnaissance à partir de groupements de lettres : celle-ci se base sur la décomposition du mot en groupes de lettres qui correspondent à des « unités phonétiques » (« spelling pattern »).

d) On s'est aperçu que les 3 modèles précédents ne représentaient pas convenablement le processus de perception humaine de textes. Des théories plus élaborées ont été émises et semblent s'approcher davantage de la réalité. L'une des principales est le « Sophisticated Guessing Model ». La théorie est la suivante : pour la reconnaissance, une première étape consiste en une extraction de caractéristiques (« features ») sur le mot ; une génération de candidats possibles suit ; finalement, le candidat qui répond le mieux aux caractéristiques est choisi. Des expériences parallèles ont montré que, pour la reconnaissance de mots, la similarité graphique est plus utilisée par l'humain que la similarité sémantique. La similarité graphique entre les mots est exprimée par un facteur (facteur de Weber) qui varie selon la position des ascendantes et des descendantes dans le mot. Ces attributs sont définis à la section 3.2.

Nous utiliserons certains aspects de cette dernière théorie dans la construction de l'algorithme de reconnaissance de mots.

Il existe deux approches possible pour faire de la reconnaissance de texte par ordinateur :

a) Reconnaissance des caractères isolés (RCI) : *chaque* caractère est segmenté, isolé ; on en extrait des caractéristiques à partir desquelles se fait la classification. L'éventail de ces caractéristiques est très grand : caractéristiques globales (transformées, moments d'inertie, etc.), caractéristiques topologiques (nœuds, boucles, arcs, etc.) [1], [2], [3], [4]. Après la reconnaissance de toutes les lettres d'un mot, on fera un post-traitement à l'aide d'une matrice de confusion et d'un dictionnaire : l'utilisation du dictionnaire permet de confirmer l'existence d'un mot ; la matrice de confusion permet d'augmenter les performances en effectuant des substitutions de lettres (ex. : *i* et *l*) lorsque la probabilité de confusion est grande. *Le mot est donc utilisé à la fin du processus de reconnaissance.* Cette approche est de type ascendante (« bottom-up »). Le schéma-bloc de la figure 1 illustre bien la nature linéaire de cette technique. De même, il apparaît évident que la courbe « Temps de reconnaissance *vs.* Nombre de caractères dans le mot » a un comportement linéaire.

b) Reconnaissance globale à partir des mots : on extrait des caractéristiques générales au niveau de chaque mot,

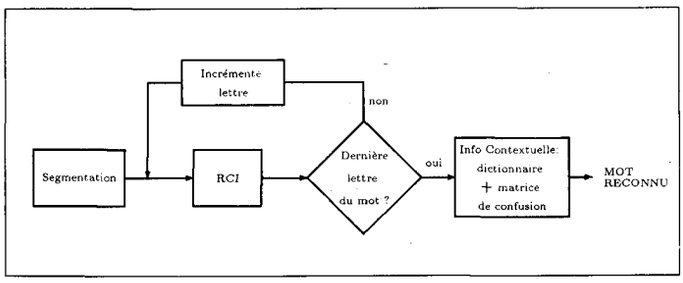

Fig. 1. — Schéma-bloc d'un système de reconnaissance traditionnel.

on génère une série d'hypothèses pour chaque mot et on passe au contexte global où l'on considère l'interaction possible entre les mots (corrélation, syntaxe et sémantique) (travaux de J. Hull [5], [6], [7], [8]). *Le mot est utilisé au début du processus de reconnaissance.* Cette approche est de type descendante (« top-down »).

Notre projet propose de *placer le mot au centre du processus de reconnaissance.* Ce projet combinera des éléments des deux approches citées en *a*) et *b*) :

— on utilisera la génération d'hypothèses à partir de caractéristiques générales tirées du mot, sauf que la corrélation ne se fait plus entre les mots (comme dans le cas de Hull) mais entre les lettres d'un même mot : la réduction du nombre d'hypothèses se fera à partir du contexte entre les lettres d'un même mot plutôt que du contexte entre les mots.

— On utilisera en système d'appoint la reconnaissance de caractères isolés (RCI) ; la reconnaissance des lettres n'est plus le centre d'intérêt de notre système mais plutôt un apport secondaire qui complètera, au besoin, la reconnaissance du mot.

3. Description du système

3.1. GÉNÉRALITÉS

Il s'agit donc de construire un système de reconnaissance qui utilise les mots comme base (Réf. *fig.* 2). On extrait des caractéristiques (attributs) sur le mot, on génère une série d'hypothèses (à partir d'un dictionnaire préalablement construit) qui auront une certaine similarité graphique avec le mot analysé. On reconnaît les caractères dans l'ordre de lecture : à chaque caractère reconnu, on réduit le nombre d'hypothèses en éliminant celles qui ne correspondent plus à la nouvelle information (lettre reconnue) sur le mot inconnu. On reconnaît les caractères jusqu'à ce que le nombre d'hypothèses soit réduit à un. En termes systémiques, on modifie la configuration du schéma-bloc de la figure 1 (systèmes de reconnaissance de caractères) où les trois phases principales de la reconnaissance de texte, le pré-traitement (segmentation), la reconnaissance de caractères isolés (RCI) et le post-traitement (information contextuelle) sont complètement séparées, indépendantes. Cette modification consiste à inclure le bloc de

post-traitement *à l'intérieur* de la boucle de rétroaction RCI, tel qu'indiqué à la figure 2. Pour bien marquer le fait que le système RCI employé dans notre système est un système secondaire (ou d'appoint) utilisé sporadiquement (et non plus systématiquement comme dans le cas RCI) nous l'appellerons RSCI (Reconnaissance Secondaire de Caractères Isolés).

Un tel système s'apparente très bien aux théories les plus répandues en terme de psychologie perceptuelle (« Sophisticated Guessing Model »), tel que mentionné à la section précédente. On peut donc s'attendre à ce que l'allure de la courbe de temps de reconnaissance *vs.* nombre de caractères dans le mot tende asymptotiquement vers une valeur fixe. Selon le nouveau schéma, il est possible que le mot analysé ne soit pas dans le dictionnaire utilisé. Aucune hypothèse ne sera générée. A ce moment, la reconnaissance du mot devra se faire caractère par caractère (RSCI). Ce comportement s'apparente au système humain : les mots inconnus à l'être humain semblent davantage scrutés au niveau des lettres qu'au niveau de leur « signature graphique ».

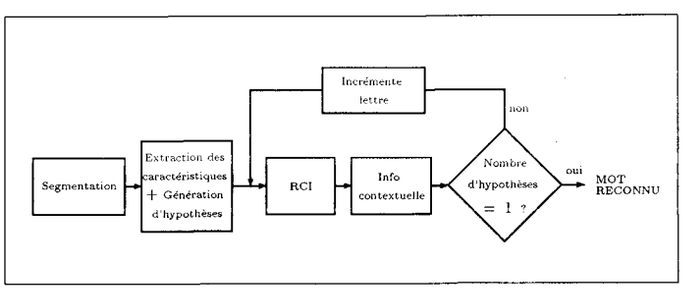

Fig. 2. — Schéma-bloc du système de reconnaissance proposé.

3.2. ATTRIBUTS DE RECONNAISSANCE

Il faut d'abord établir un choix des attributs qui seront utilisés pour la classification des mots. Ces attributs doivent posséder trois qualités : être le plus universel possible (pouvoir se retrouver sur la majorité des fontes), être relativement résistants au bruit et être facile à extraire. Le « pouvoir de discrimination » de l'attribut est aussi une qualité importante : si la classification des mots du dictionnaire selon un attribut spécifique donne un fort pourcentage des mots dans la même classe, on peut douter de l'utilité de celui-ci, même si son extraction automatique est très robuste.

Comme nous l'avons mentionné précédemment, l'humain semble utiliser la similitude graphique pour établir des hypothèses sur un mot observé ; deux de ces attributs seraient les ascendantes et les descendantes. Ces attributs sont définis comme suit (Réf. *fig.* 3) :

les ascendantes sont des caractères ayant des pels au-dessus de la ligne de col et qui sont connectés aux pels sous la ligne de col ;

les descendantes sont des caractères ayant des pels en dessous de la ligne de base et qui sont connectés aux pels au-dessus de la ligne de base.

On peut donc croire que si l'humain utilise ces attributs, on peut se fier à leur constance. Ils seront donc retenus. Les points sur les « *i* » et « *j* » ainsi que les boucles fermées des lettres (telles que celles sur le « *a* », « *b* », etc.) seront également retenus.

Un critère important est la localisation de la position des attributs dans le mot. On peut établir une localisation relative à la manière d'un pile. Une localisation absolue par une distance normalisée à partir du début du mot est aussi possible. Toutefois, la méthode adoptée sera celle de la localisation absolue des attributs à partir de la position du caractère (où l'attribut se trouve) dans le mot. Cette technique offre effectivement un potentiel intéressant au niveau des lettres liées.

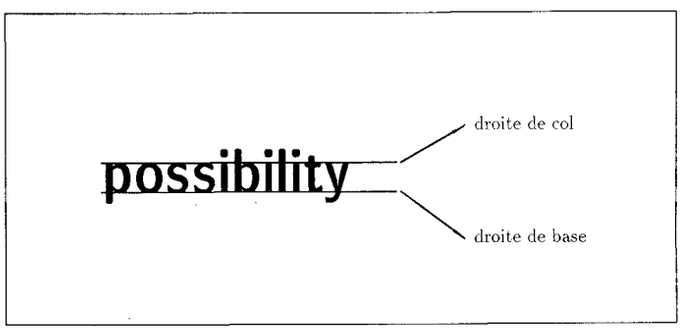

Fig. 3. — Exemple.

A titre d'exemple, le repérage des attributs du mot apparaissant à la figure 3 donne les résultats suivants :

— 1 descendante en 1re position
— 1 boucle en 1re position
— 1 boucle en 2e position
— 1 point en 5e position
— 1 ascendante en 6e position
— 1 boucle en 6e position
— 1 point en 7e position
— 1 ascendante en 8e position
— 1 point en 9e position
— 1 ascendante en 10e position
— 1 descendante en 11e position.

3.3. STRUCTURE ET UTILISATION DU DICTIONNAIRE

On adoptera un mode de classification parallèle comportant 5 structures en arbres :

— Un premier arbre à 16 nœuds terminaux permet de classifier les mots selon leur nombre de lettres. Les mots de 16 lettres ou plus seront classés dans le nœud 16.

— Les 4 autres arbres ont une structure semblable comportant 969 nœuds terminaux : à chacun de ces nœuds terminaux correspond une position bien déterminée de chaque attribut. Un exemple : au nœud 599 de l'arbre des ascendantes correspond 3 ascendantes respectivement sur le huitième, quinzième et seizième caractère du mot considéré (16 lettres).

Cette structure parallèle a été adoptée en raison de sa grande souplesse : si on veut ajouter ou retirer un attribut, il s'agit simplement d'ajouter ou de retirer l'arbre correspondant. En ce sens, la structure série apparaît comme beaucoup moins souple. La construction du dictionnaire se fait comme suit : on établit une correspondance (« mapping ») entre les lettres et les attributs (ex. : « *d* » comprend une boucle et une ascendante, « *p* » comprend une boucle et une descendante, etc.). Les mots qui constitueront le dictionnaire sont ensuite tirés d'une banque de données (dictionnaire Merriam Webster ou le « Brown Corpus »). L'analyse du mot se fait de gauche à droite ; à l'aide du « mapping », on vérifie pour chaque lettre si elle contient un ou plusieurs attributs ; si oui, on stocke la position de la lettre dans les vecteurs d'attributs appropriés. Quand la fin du mot a été atteinte, on utilise les vecteurs d'attributs pour calculer les nœuds terminaux respectifs. Le nœud de l'arbre du nombre de lettres est aussi disponible. On stocke alors l'indice-mot associé au mot analysé dans chacun des 5 nœuds terminaux obtenus. On recommence le processus pour tous les mots que l'on veut inclure dans le dictionnaire réduit. La taille du dictionnaire réduit est établie par l'utilisateur.

Quand le dernier mot a été analysé, on obtient un dictionnaire réduit qui consiste en : une liste des mots du dictionnaire réduit et leurs indices-mots associés, une liste des nœuds terminaux de l'arbre du nombre de lettres dans le mot et des indices-mots stockés dans chacun de ces nœuds, 4 listes des nœuds terminaux des 4 arbres (ascendantes, descendantes, points et boucles) et les indices-mots stockés dans chacun de ces nœuds.

3.4. GÉNÉRATION D'HYPOTHÈSES

Comme nous l'avons vu à la figure 2, le système fonctionne sur une base de génération d'hypothèses : après l'extraction des attributs sur le mot observé, on établit une liste de tous les mots du dictionnaire qui ont une similitude graphique avec le mot observé ; en d'autres termes, pour qu'un mot du dictionnaire réduit soit retenu comme hypothèse valable pour le mot observé, il faut que tous les attributs détectés sur le mot observé soient identiques (par leur nombre et par leurs positions respectives) au mot du dictionnaire réduit considéré. Cette liste constitue la génération d'hypothèses pour un mot donné. Voici comment fonctionne ce processus.

Le processus de génération d'hypothèses est l'inverse de ce qui a été fait à la construction du dictionnaire : plutôt que de stocker les indices-mots dans les nœuds terminaux, on calcule les 5 nœuds terminaux à partir des attributs extraits sur le mot à reconnaître et on prélève tous les indices-mots de ces 5 nœuds. On obtient donc 5 groupes distincts d'indices-mots. On établit l'intersection de ces 5 groupes, soit les indices-mots qui sont communs à tous les groupes (tel que prévu par l'arrangement parallèle). Puis, par correspondance mots $\leftrightarrow$ indices-mots, on transforme le groupe résultant d'indices-mots en un groupe de mots. Ce groupe (liste de mots) constitue le résultat de la génération d'hypothèses.

3.5. RÉDUCTION D'HYPOTHÈSES

La section précédente nous a permis d'émettre une liste de mots qui ont une similitude graphique avec le mot observé. On sait qu'un seul des mots de cette liste est identique au mot observé. L'algorithme suivant décrit le processus de reconnaissance du mot observé à partir d'une réduction successive du nombre d'hypothèses :

i = position de la lettre dans le mot
L = nombre de lettres dans le mot
N = nombre d'hypothèses restantes.

1. $i = 1$.
2. i-ième lettre des N hypothèses est la même ?

   oui : incrémenter i, étape 3
   non : étape 4 .

3. $i > L$?

   oui : étape 8
   non : étape 2 .

4. RSCI sur i-ième lettre du mot à reconnaître : lettre reconnue : δ_i.
5. Élimination des hypothèses n'ayant pas δ_i en i-ième position ; nombre d'hypothèses restantes = N.
6. N = 1 ou N = 0 ?

   oui : étape 7
   non : incrémenter i, étape 2 .

7. N = 0 : mot non reconnu

   N = 1 : MOT RECONNU, étape 9 .

8. Mot non reconnu.
9. Fin.

A titre d'exemple, prenons le cas où 2 hypothèses ont été générées pour un mot inconnu : « perle » et « perte ». On constate que l'algorithme précédemment décrit permet de reconnaître un mot de 5 lettres à partir de la reconnaissance d'une seule lettre (i.e. la quatrième lettre). De plus, si on emploi, pour la reconnaissance de caractères isolés des techniques d'appariement de gabarits où il faut calculer une distance entre l'inconnu et une série de modèles, on constate qu'il y aura réduction en termes de nombre de comparaisons à établir. Pour l'exemple précédent, on n'aura qu'à comparer la quatrième lettre du mot observé avec un « *l* » et un « *t* ». On ne fait que 2 comparaisons alors que dans un cas de reconnaissance de caractères isolés, on en ferait au moins 26 (lettres minuscules de l'alphabet). En plus d'économiser substantiellement du temps de calcul, on augmente les chances de reconnaissance : on a moins de chances de faire une erreur en comparant 2 modèles avec une lettre inconnue (même si ces modèles, par essence, doivent se ressembler (définition des attributs)) qu'en comparant 26 modèles (ou plus) avec une lettre inconnue. Cet aspect sera systématisé à la prochaine section (i.e. Ratio de temps de calcul).

3.6. TRAITEMENT DES LETTRES LIÉS

Les caractères liés constituent un problème majeur en reconnaissance de texte. Les caractères sont liés par insuffisance de résolution au niveau du capteur ou par dégradation générale de la qualité du document analysé (ex. : photocopie). Notre étude ne se base pas sur l'établissement d'une segmentation précise : il s'agit seulement de connaître le nombre de caractères liés et la position des attributs (ex. : si « *re* » sont liées : on peut facilement savoir que le groupe de caractères liés contient deux caractères et que la boucle du « *e* » est sur le deuxième caractère). Les hypothèses peuvent par la suite être générées. Pour la reconnaissance, on s'abstiendra d'utiliser des caractères liés. En fait, on recueille peu d'information sur les caractères liés mais, par contre, la qualité et la robustesse de cette information est grande. Le contexte sera ensuite utilisé afin de déterminer à quelles classes ces caractères liés appartiennent.

Cette technique de « séparation » permet ainsi d'éviter les problèmes reliés à la segmentation conventionnelle ainsi qu'à la reconnaissance des caractères liés : on minimise l'effet de la portion bruitée (élément de liaison) entre les caractères.

3.7. LA SAISIE OPTIQUE

Les documents à soumettre au système de reconnaissance ont été produits par imprimante laser à l'aide de 3 types de fonte (12 points). Ces documents ont été numérisés par un télécopieur à une résolution de 200 points par pouce. L'image binaire résultante constitue une version « bruitée » de l'impression laser : quelques caractères sont liés par deux ou mutilés (ex. : boucle de « *o* » ouverte).

Les images binaires associées à chacune des pages ont été traitées par un logiciel de segmentation. Ce logiciel a pour but de :

— segmenter tous les éléments (ex. : lettres, point de « *i* »),
— établir les caractéristiques graphiques du texte (position des marges, interlignes, angle du document, etc.),
— corriger l'angle (« skew ») du document,
— caractériser les caractères : position, dimensions, topologie (boucles fermées, ascendantes, descendantes, etc.),
— établir l'ordre de lecture des caractères,
— segmenter des mots.

Les résultats de la segmentation sont ensuite acheminés au système de reconnaissance de mots.

4. Les critères de performance

Les critères d'efficacité pour analyser les performances du système sont multiples. Les résultats sont analysés à travers une série de ratios de 2 classes : les ratios de temps de calcul et les ratios de fiabilité.

4.1. LES RATIOS DE TEMPS DE CALCUL

Ce type de ratio permet d'établir une comparaison entre les performances de notre système de reconnaissance de mots et les systèmes de reconnaissance de caractères. Cette comparaison se fait au niveau de la somme de travail nécessaire pour obtenir des résultats (« output ») de reconnaissance pour les mots (i.e. quelle somme de travail faut-il investir dans l'un et l'autre système pour reconnaître une page de texte ?).

De façon plus formelle, UL (Utilisation des Lettres) est le quotient du nombre de lettres du document analysé qui ont été soumises à la reconnaissance RSCI (L) et du nombre total de lettres du document analysé (N) :

$$UL = L/N .$$

NLW (Nombre de Lettres au Warping) est le nombre de lettres utilisées en moyenne pour l'anamorphose (DCW). Mentionnons que le système d'appoint pour la reconnaissance de caractères isolés fonctionne sur une base de « Template Matching » utilisant l'anamorphose spatiale (« Dynamic Contour Warping »). Pour calculer NLW, on n'utilise que les lettres qui ont été soumises à la reconnaissance RSCI. Dans l'exemple « perle, perte », *nlw* pour la quatrième lettre du mot inconnu est de 2 car il n'y a que 2 possibilités pour cette lettre : un « *l* » ou un « *t* ». Donc, soit nlw_1, nlw_2, ..., nlw_L les *nlw* pour chacune de ces lettres, on obtient donc l'expression suivante :

$$NLW = 1/L \sum_{i=1}^{L} nlw_i$$

RG (Réduction globale) est une combinaison des deux ratios précédents. Il donne une mesure en pourcentage de la réduction globale, en termes de comparaisons à faire au niveau de l'anamorphose. Pour un système RCI, le nombre de comparaisons à faire à l'anamorphose est donné par

$$C_1 = N \times 26$$

où 26 représente les 26 lettres minuscules de l'alphabet.

Pour le système utilisé ici, le nombre de comparaisons est donné par

$$C_2 = N \times UL \times NLW$$

RG est donc donné par l'expression suivante :

$$RG = C_2/C_1 = (UL \times NLW)/26 .$$

4.2. LES RATIOS D'EFFICACITÉ

Les ratios d'efficacité pour la reconnaissance RCI sont au nombre de 3 : le taux de succès, le taux d'échecs et le taux d'abstention. Pour la reconnaissance de textes à partir des mots, il faut tenir compte d'un facteur important : s'il y a abstention au niveau de la reconnaissance d'un mot une solution de rechange existe. On peut très bien passer les mots sur lesquels il y a eu des abstentions à la reconnais-

sance caractère par caractère. Pour cette raison, en plus de 3 ratios cités ci-haut, on ajoute un quatrième ratio qui est le taux de succès sans les abstentions. Donc, soit S_l, E_l, A_l respectivement le nombre de lettres reconnues avec succès, le nombre de lettres où il y a erreur de reconnaissance et le nombre de lettres où le système s'est abstenu de prendre une décision, les ratios « SL » (Succès sur les lettres), « EL » (Erreur sur les lettres), « AL » (Abstention sur les lettres) et « SLSA » (Succès sur les lettres sans les abstentions) sont donnés par

$$SL = S_l/N$$
$$EL = E_l/N$$
$$AL = A_l/N$$
$$SLSA = S_l/(N - A_l)\,.$$

Les mêmes ratios peuvent être développés pour les mots : soit M, S_m, E_m, A_m respectivement le nombre de mots analysés dans le texte, le nombre de mots reconnus avec succès, le nombre de mots où il y a erreur de reconnaissance et le nombre de mots où le système s'est abstenu de prendre une décision, les ratios « SM » (Succès sur les mots), « EM » (Erreur sur les mots), « AM » (Abstention sur les mots) et « SMSA » (Succès sur les mots sans les abstentions) sont donnés par

$$SM = S_m/M$$
$$EM = E_m/M$$
$$AM = A_m/M$$
$$SMSA = S_m/(M - A_m)\,.$$

5. Résultats et discussion

5.1. NOMBRE D'HYPOTHÈSES EN FONCTION DE LA TAILLE DU DICTIONNAIRE RÉDUIT

Cette expérience fournit le nombre d'hypothèses générées en fonction de la taille du dictionnaire réduit. On veut connaître les pourcentages PC_i pour $i = 1, ..., J$ où PC_i représente le pourcentage de mots du dictionnaire réduit qui ont i hypothèses ou moins lors de la phase de génération d'hypothèses. Le paramètre J représente le nombre maximum d'hypothèses générées pour un mot donné du dictionnaire ; PC_J est toujours de 100 %.

Quand on veut faire référence au dictionnaire du système de reconnaissance, on emploie l'expression *dictionnaire réduit :* ce dictionnaire est un *sous-ensemble* prélevé dans une banque de mots (telle que le dictionnaire Merriam Webster ou le Brown Corpus).

Les résultats obtenus respectivement pour un dictionnaire réduit de 7 000 mots tirés à partir du Brown Corpus et pour un dictionnaire réduit de 10 000 mots tirés à partir du Merriam Webster apparaissent aux figures 4 et 5.

Afin de simplifier le traitement, un seul dictionnaire réduit sera utilisé subséquemment : celui obtenu à partir du Brown Corpus. Ce choix est justifié par des raisons

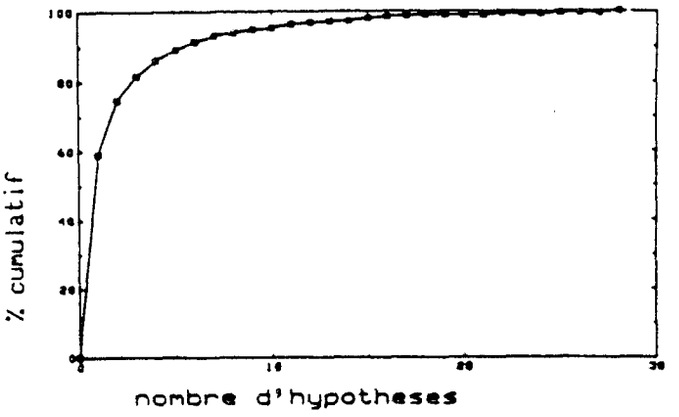

Fig. 4. — Graphique du pourcentage cumulatif des mots du dictionnaire réduit (7 000 mots) *vs.* le nombre d'hypothèses. Base de données : Brown Corpus.

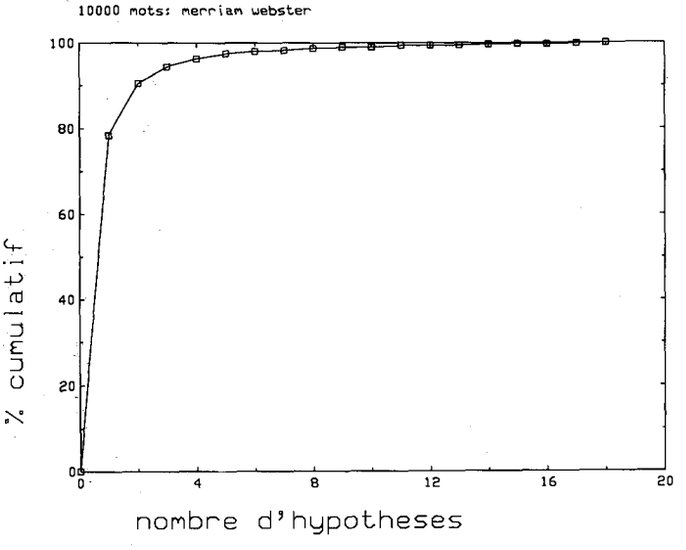

Fig. 5. — Graphique du pourcentage cumulatif des mots du dictionnaire réduit (10 000 mots) *vs.* le nombre d'hypothèses. Base de données : dictionnaire Merriam Webster.

pratiques : un texte choisi au hasard donnera, en moyenne, de meilleurs résultats de reconnaissance avec le dictionnaire réduit tiré du Brown Corpus qu'avec le dictionnaire réduit tiré du Merriam Webster à cause de la nature même du Brown Corpus (formé des mots les plus utilisés en anglais).

En observant plus en détail les résultats obtenus pour le dictionnaire tiré du Brown Corpus, on constate que les mots du dictionnaire réduit dont le nombre d'hypothèses est inférieur à 5 représentent environ 90 % des mots du dictionnaire réduit. L'expérience a été répétée pour des dictionnaires réduits de 1 000, 3 000 et 5 000 mots. Ces résultats nous permettent de visualiser la tendance de ce pourcentage en construisant les graphes suivants : pour un nombre d'hypothèses H donné, on établit le pourcentage cumulatif (le même que celui montré aux figures précéden-

tes) en fonction de la taille du dictionnaire réduit. Un exemple est donné à la figure 6 pour H = 5 : on peut y observer le comportement linéaire de la courbe. A titre d'exemple, une extrapolation linéaire pour un dictionnaire réduit de 20 000 mots donne un pourcentage cumulatif de 82,7 %, ce qui signifie que 82,7 % des mots de ce dictionnaire ont un nombre d'hypothèses inférieur ou égal à 5. Le pourcentage cumulatif réel est probablement supérieur à 82,7 % : il est plausible de penser que, plus la taille du dictionnaire augmente, plus la courbe perd son caractère linéaire pour adopter un profil asymptotique ; en effet, à cause de la nature même du dictionnaire (i.e. mots les plus utilisés classés en ordre décroissant), plus sa taille augmente, plus la signature graphique des derniers mots est exclusive, donc moins le nombre d'hypothèses générées est grand.

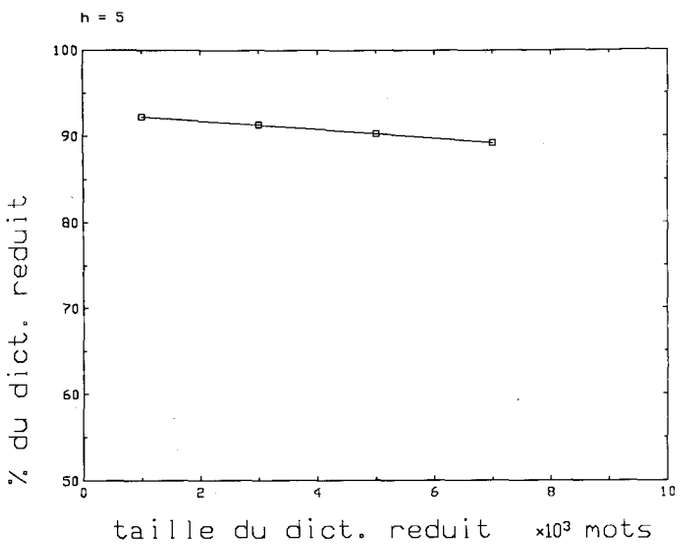

Fig. 6. — Graphique du pourcentage cumulatif en fonction de la taille du dictionnaire ; H = 5.

De façon générale, les résultats présentés dans cette section démontrent la faisabilité de l'approche : avec des attributs très simples et peu nombreux qui décrivent les mots, le nombre moyen d'hypothèses générées par mot est suffisamment restreint pour permettre la production d'un algorithme de reconnaissance de texte basé sur la reconnaissance de mots et ce, même pour un dictionnaire réduit de taille relativement élevée. Pour un dictionnaire réduit de 40 000 mots (soit près de deux tiers du dictionnaire Merriam Webster), le pourcentage des mots de ce dictionnaire réduit qui auraient 5 hypothèses ou moins serait de 72,7 %. Ces résultats démontrent clairement la faisabilité de l'approche proposée dans le cadre de ce projet.

5.2. TAUX DE RECONNAISSANCE

L'étude s'est faite comme suit : on a fait imprimer les 1 000 mots les plus fréquemment utilisés en anglais ; la fonte choisie a été la amss 12 points. L'image binaire associée a été obtenue à l'aide d'un télécopieur XeroxFax à une résolution graphique de 200 dpi. Afin d'en faciliter le traitement informatique, cette image a été segmentée en trois « fenêtres » de 400, 335 et 265 mots. La taille du dictionnaire réduit a été fixée à 1 000 mots (les 1 000 mots les plus utilisés en anglais). Les résultats de reconnaissance étant consistants d'une fenêtre à l'autre, seuls ceux de la fenêtre 1 seront présentés (réf. tableau 1).

TABLEAU 1

Résultats du test fait avec la fonte amss ; taille du dictionnaire réduit : 1 000 mots ; fenêtre 1

RATIOS	POURCENTAGES
SL	99,10
EL	0,00
AL	0,90
SLSA	100,00
SM	99,25
EM	0,00
AM	0,75
SMSA	100,00

En observant ces résultats, on note que le taux de reconnaissance sans les abstentions (SLSA) est de 100 %, ce qui veut dire qu'aucun mauvais candidat n'a été retenu comme hypothèse finale. Il y a eu abstention sur 0,90 % des lettres : *ces abstentions ne génèrent pas d'erreur ;* elles nous indiquent simplement que le système n'a pas voulu se risquer à retenir d'hypothèse finale pour certains mots. Si on veut reconnaître les mots sur lesquels il y a eu des abstentions, on n'a qu'à utiliser le système RSCI pour reconnaître tous les caractères.

5.3. EFFET DE LA TAILLE DU DICTIONNAIRE

On répète l'expérience faite à la section 5.2 en faisant varier la taille du dictionnaire. On utilisera la même image binaire qu'à la section précédente. Les valeurs que prendra N sont les suivantes : 500, 1 000, 3 000, 5 000 et 7 000 mots. Les résultats sont présentés aux tableaux 2, 3 et 4. Pour les mêmes raisons évoquées à la section précédente, seuls les résultats de la fenêtre 1 sont présentés.

On constate que, dans le cas du tableau 3, plus N augmente, plus le taux de reconnaissance (SL) et le taux d'erreur augmentent au dépend du taux d'abstention. On peut expliquer ce phénomène de la façon suivante. Si on augmente la taille du dictionnaire réduit, on génère davantage d'hypothèses pour un mot inconnu. Cette génération supplémentaire peut impliquer un taux d'erreur plus élevé. Avant de donner un exemple, notons un point important : si un mot inconnu à caractères non-liés est correctement segmenté et que ses attributs sont tous bien

TABLEAU 2
Ratios de performance pour la fonte amss

FENETRE 1 (400 mots, 1 892 lettres)					
RATIOS	TAILLE DU DICTIONNAIRE RÉDUIT (nombre de mots)				
	500	1 000	3 000	5 000	7 000
SL	99,10	99,10	99,10	99,10	99,10
EL	0,00	0,00	0,00	0,00	0,00
AL	0,90	0,90	0,90	0,90	0,90
SLSA	100,00	100,00	100,00	100,00	100,00
SM	99,25	99,25	99,25	99,25	99,25
EM	0,00	0,00	0,00	0,00	0,00
AM	0,75	0,75	0,75	0,75	0,75
SMSA	100,00	100,00	100,00	100,00	100,00

localisés et que la reconnaissance RSCI nécessaire identifie correctement tous les caractères, le mot inconnu sera identifié correctement *de façon certaine*. Rappelons aussi comment fonctionne l'algorithme de réduction du nombre d'hypothèses tel que décrit à la section 3.5 : une liste de mots ayant une forte similitude graphique avec le mot observé est produite à partir des attributs extraits sur le mot observé ; il s'agit ensuite de réduire cette liste à 1 seul candidat à partir de lettres reconnues sur le mot observé ; on analyse les mots-hypothèses de gauche à droite ; quand la lettre à la i-ième position n'est pas la même pour toutes les hypothèses, on doit reconnaître la i-ième lettre du mot observé ; cette lettre reconnue est ensuite comparée à la i-ième lettre de toutes les hypothèses ; les hypothèses dont la lettre n'est pas la même sont éliminées ; on recommence le processus avec la $i+1$ i-ième lettre et on arrête le processus quand la liste d'hypothèses est réduite à une seule ou quand la dernière lettre du mot observé a été rencontrée. Voici maintenant l'exemple : le mot à reconnaître est « force » avec deux caractères qui se touchent (le « r » et le « c ») ; la génération d'hypothèses selon un dictionnaire réduit de N_1 mots est la suivante :

have
here
lose
home
love

Les hypothèses générées ont 4 lettres car « force » contient 4 boîtes (à cause des 2 lettres liées). A l'étape de réduction d'hypothèses, la première lettre de « force » est reconnue : le « f » ne correspond à aucune des premières lettres des hypothèses ; il y a donc abstention sur le mot « force ». Supposons maintenant que la génération d'hypothèses se fasse selon un dictionnaire réduit de N_2 mots où $N_2 > N_1$. Supposons que la génération d'hypothèses est la suivante :

have
here
lose
home
love
face

A l'étape de réduction d'hypothèses, le « f » de « force » sera reconnu ; une seule hypothèse commence par un « f » : « face ». La réduction tombe donc à une seule hypothèse : elle sera donc retenue et sera identifiée au mot inconnu. Il y a donc génération d'une erreur au niveau du mot. Supposons maintenant une génération d'hypothèses faite selon un dictionnaire réduit de N_3 mots où $N_3 > N_2$. Supposons que la génération d'hypothèses soit la suivante :

have
here
lose
home
love
face
fare

A l'étape de réduction d'hypothèses, la reconnaissance du « f » réduira le nombre d'hypothèses à 2 (« face » et « fare ») ; la reconnaissance de la deuxième lettre de « force » (le « o ») ne correspond pas au « a » de « face » ou de « fare ». Il y aura donc abstention sur le mot « force ». Si, avec un dictionnaire réduit de N_4 mots (où $N_4 > N_3$), on génère les mêmes hypothèses qu'avec le dictionnaire N_3 avec, en plus, le mot « fore », il y aura à nouveau une erreur au niveau du mot (« fore » sera identifié à « force »).

Suite à cet exemple, on peut constater une sorte de « pulsation succès-erreur-abstention » de la reconnaissance en fonction de la taille du dictionnaire réduit (réf. tableau 4). Il est à noter que plus la taille du dictionnaire augmente, plus l'erreur (en termes de lettres) diminue : dans l'exemple précédent, « face » est à une distance de 2 lettres de « force » (il manque le « o » et le « r ») mais « fore » est à une distance d'une seule lettre (il ne manque que le « c »). Par contre, plus la taille du dictionnaire augmente, plus le nombre de mots soumis à cette « pulsation » augmente. Il semble donc normal que le SLSA ait tendance à diminuer si la taille du dictionnaire augmente (au profit du EL). En observant les résultats, on constate que le SL augmente si la taille du dictionnaire réduit augmente (réf. tableau 3). Cela semble normal si on considère que le EL augmente aussi avec une augmentation de la taille du dictionnaire réduit : le système prend des décisions (retient un candidat final) là où il s'abstenait pour une taille de dictionnaire plus petite ; même si le mot retenu n'est pas le bon mot (car il y a erreur), étant donné la forte similitude graphique, plusieurs lettres du mot retenu seront tout de même valides ; ces lettres biens reconnues seront donc inclues dans le calcul du SL et augmenteront celui-ci.

5.4. TAUX DE RECONNAISSANCE OMNIFONTE

Le terme « omnifonte » s'applique aux cas où aucune information spéciale n'est fournie au système sur les types

de fontes qui lui seront soumis. C'est le cas de notre système : tel que spécifié à la section 3.2, les attributs extraits sur les mots se retrouvent sur la majorité des fontes usuelles ; de plus, le système RSCI, avec ses grandes capacités d'adaptation (déformations des contours du caractère), alloue une grande variété de fontes.

On a donc répété l'expérience de la section 5.2 avec des fontes différentes (amdh et ozub) et pour des tailles de dictionnaire réduit de 500, 1 000, 3 000, 5 000 et 7 000 mots. Il est à noter que chacune des 3 fontes choisies (amss, amdh et ozub) présentent des caractéristiques graphiques exclusives : pour tester les possibilités omnifontes d'un système de reconnaissance, il est évidemment essentiel de choisir des fontes relativement différentes les unes des autres. Les résultats sont présentés aux tableaux 3 et 4. Étant donné que l'espacement et l'épaisseur des caractères varient en fonction de la fonte utilisée, il est évident que nombre total de mots par fenêtre est différent d'une fonte à l'autre.

TABLEAU 3
Ratios de performance pour la fonte amdh

FENETRE 1 (315 mots, 1 408 lettres)					
RATIOS	TAILLE DU DICTIONNAIRE RÉDUIT (nombre de mots)				
	500	1 000	3 000	5 000	7 000
SL	95,88	95,95	95,95	96,02	96,02
EL	0,14	0,07	0,07	0,28	0,28
AL	3,98	3,98	3,98	3,69	3,69
SLSA	99,85	99,93	99,93	99,71	99,71
SM	95,24	95,24	95,24	95,24	95,24
EM	0,32	0,32	0,32	0,63	0,63
AM	4,44	4,44	4,44	4,13	4,13
SMSA	99,67	99,67	99,67	99,34	99,34

TABLEAU 4
Ratios de performance pour la fonte ozub

FENETRE 1 (295 mots, 1 305 lettres)					
RATIOS	TAILLE DU DICTIONNAIRE RÉDUIT (nombre de mots)				
	500	1 000	3 000	5 000	7 000
SL	96,78	96,86	96,48	96,48	96,78
EL	0,00	0,00	0,00	0,00	0,00
AL	3,22	3,14	3,52	3,52	3,22
SLSA	100,00	100,00	100,00	100,00	100,00
SM	96,27	96,27	95,93	95,93	96,27
EM	0,00	0,00	0,00	0,00	0,00
AM	3,73	3,73	4,07	4,07	3,73
SMSA	100,00	100,00	100,00	100,00	100,00

5.5. EFFET DES LETTRES LIÉES

Une façon d'augmenter le taux de reconnaissance (SL) est de forcer le système à prendre une décision sur les boîtes à lettres liées. Il s'agit de trouver une méthode qui permette d'établir le nombre de lettres liées dans une boîte donnée ainsi que l'appartenance des attributs dans la boîte (ex. : l'ascendante détectée appartient à la première ou à la deuxième lettre dans une boîte à 2 lettres liées).

Une étude statistique des boîtes à lettres doubles (les plus fréquentes dans un texte de qualité « raisonnable ») a été faite sur une banque de données comportant 21 pages de textes divers, de qualité d'impression moyenne. Les boîtes ont été classées par ordre décroissant de fréquence d'apparition. On a évalué le « degré de séparation » des attributs : c'est une mesure empirique du degré de certitude qu'un attribut appartienne à la première ou à la deuxième lettre. Il est à noter que l'on ne parle pas ici de segmentation : la segmentation est plus complexe car il faut établir la frontière entre les 2 lettres. La séparation ne veut qu'établir l'appartenance les attributs à chacune des lettres. Voyons plus en détails ce degré de séparation.

Les boîtes qui contiennent uniquement 2 attributs du même type (ex. : la boîte à lettres liées « th » contient uniquement 2 ascendantes) sont facilement séparables : la boîte double se divise en deux boîtes avec un attribut chaque ; l'analyse de la position des attributs n'est même pas nécessaire. Ce type de boîtes à lettres liées représente 9,9 % du total des boîtes de l'étude statistique (en pondérant par la fréquence d'apparition de chacune des boîtes). Les boîtes à lettres liées qui ne contiennent aucun attribut (ex. : la boîte « rs ») ne peuvent être séparées car leur largeur est trop semblable à la largeur des grosses lettres (« m » et « w ») qui ne contiennent aucun attribut également. Elles représentent 9,5 % du total des boîtes. Le degré de séparation des boîtes restantes dépend de la position des attributs dans la boîte. Prenons 2 exemples : les boîtes à lettres liées « pe » et « rh ». La boîte « pe » ne cause pas de problème car la descendante se situe à l'extrême gauche de la lettre gauche et la boucle se situe au centre de la lettre droite ; la boîte « rh » est plus problématique car l'ascendante se situe à peu près au milieu de la boîte, ce qui diminue la certitude quant à l'appartenance de l'ascendante (à la lettre gauche ou à la lettre droite ?). Les boîtes à degré de séparation élevé (telles que « pe ») représentent 75,4 % des boîtes et les boîtes à faible degré de séparation (telles que « rh ») représentent 5,2 %. En résumé, 85,3 % des boîtes ont un degré de séparation élevé (9,9 % + 75,4 %), 5,2 % ont un degré marginal de séparation et 9,5 % n'offrent aucune possibilité de séparation.

Selon ce qui vient d'être dit, on peut s'attendre à deux

tendances au niveau des taux d'efficacité : étant donné que l'on force le système à prendre une décision sur les boîtes à lettres liées, on s'attend d'une part à ce que le taux de reconnaissance (SL) augmente ; d'autre part, on s'attend à une croissance du taux d'erreur (EL) : si la séparation était parfaite, on aurait un taux d'erreur identique à l'expérience où l'on ne fait aucun travail sur les boîtes à lettres liées. Ce serait une conversion du taux d'abstention en taux de succès. Étant donné que la séparation n'est pas parfaite et que l'algorithme ne tient pas compte de certaines situations (ex. : la boîte à lettres liées « co » produit 2 boucles (celle du « o » et celle produite par le double contact du « c » et du « o »)), on peut s'attendre à ce que le taux d'erreur augmente par rapport au cas où l'on ne fait aucun travail sur les boîtes à lettres liées.

Les résultats du tableau 5 illustre bien la tendance selon laquelle le taux de reconnaissance (SL) augmente quand on effectue une séparation des lettres liées.

TABLEAU 5

Ratios de performance pour la fonte amss
Séparation des boîtes segmentées

FENETRE 1 (400 mots, 1 892 lettres)			
RATIOS	TAILLE DU DICTIONNAIRE RÉDUIT (nombre de mots)		
	1 000	3 000	5 000
SL	99,74	99,74	99,74
EL	0,00	0,00	0,00
AL	0,26	0,26	0,26
SLSA	100,00	100,00	100,00
SM	99,75	99,75	99,75
EM	0,00	0,00	0,00
AM	0,25	0,25	0,25
SMSA	100,00	100,00	100,00

5.6. RATIO DE RÉDUCTION GLOBALE

En observant les résultats de la figure 7, on peut relever 2 points intéressants :

— le RG est très faible (moins de 5 % dans la plupart des cas), ce qui est très intéressant en termes d'économie de temps-machine ; cela signifie que si un système de reconnaissance RCI fonctionnant sur une base de classification « Template Matching » fait 100 comparaisons pour reconnaître un mot donné (chaque gabarit des 26 lettres de l'alphabet est comparé à la lettre à reconnaître), notre système de reconnaissance de mots en fera *moins de 5* pour le même mot à reconnaître ;

— la croissance du RG en fonction de la taille du dictionnaire réduit n'est pas exponentielle, comme on

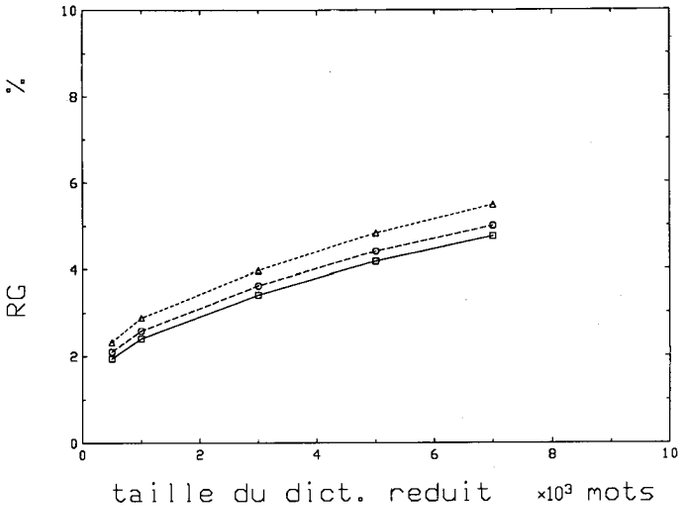

Fig. 7.

aurait pu le croire au premier abord : les courbes de la figure 7 montrent une évolution pratiquement linéaire. On peut comprendre pourquoi la courbe n'est pas exponentielle si on pense à la nature du dictionnaire réduit : ce sont les N mots (numérotés de 1 à N) les plus fréquemment utilisés dans la langue anglaise ; plus N est grand, plus les mots les moins fréquents (les tout derniers mots du dictionnaire réduit) sont longs et de formes inhabituelles par rapport aux premiers mots (ex. : « désoxyribonucléique » est sûrement très rare et est sûrement de forme inhabituelle (mot de 19 lettres)) ; on peut donc comprendre qu'à l'étape de génération d'hypothèses, le nombre d'hypothèses des mots moins fréquents sera plus petit que celui des mots plus fréquents. Le nombre réduit d'hypothèses va diminuer le NLW, le UL et, par conséquent, le RG. Il est même probable, en observant l'allure des courbes de la figure 7, qu'il y ait une « saturation » et que les courbes tendent vers une valeur maximum de RG. Si on se limite à l'hypothèse linéaire, un extrapolation du RG pour un dictionnaire réduit de 20 000 mots reste encore en deçà de 10 % (dans le pire des cas).

6. Conclusion

Le but de cette étude était de réaliser un système de reconnaissance de texte à partir des mots et de tester son comportement dans un environnement réel.

Le système est constitué de trois éléments de base suivants :

— un détecteur d'attributs qui a pour fonction d'extraire les caractéristiques fortes d'un mot (ascendantes, descendantes, points et boucles dans le mot) ;

— un dictionnaire où les mots ont été regroupés en différentes classes selon le nombre et la position des attributs dans les mots ;

— un système d'appoint de reconnaissance de caractères individuels.

Le processus de reconnaissance est le suivant : on extrait les attributs du mot à reconnaître ; on extrait du dictionnaire tous les mots qui ont la même « apparence graphique » (selon la position des attributs) que le mot à reconnaître ; un processus d'élimination progressive des mots extrait du dictionnaire (utilisation du système d'appoint) permet d'établir l'identité du mot à reconnaître.

Nous avons d'abord démontré que le concept de reconnaissance de texte à partir des mots est « viable ». Dans ce contexte, « viable » fait référence au nombre d'hypothèses générées en fonction de la taille du dictionnaire utilisé, pour un ensemble d'attributs donné. Les résultats ont été éloquents : pour une taille de dictionnaire réduit de 40 000 mots (soit près des deux tiers du dictionnaire Merriam-Webster), plus de 75 % des mots du dictionnaire génèrent 5 hypothèses ou moins et ce, avec 5 attributs facilement extractibles.

En plus de démontrer la viabilité du projet, nous avons dégagé les caractéristiques reliées à l'utilisation d'un tel concept : l'habileté à fonctionner dans un environnement omnifonte, la réduction du temps de traitement et le concept de « séparation » des caractères liés dans un environnement bruité. Considérons le résumé des résultats au tableau suivant pour illustrer les performances omnifontes du système ; on y donne les différents ratios de performance pour 3 fontes différentes (amss, amdh et ozub) pour un dictionnaire de 7 000 mots.

RATIOS	TYPE DE FONTE		
	amss	amdh	ozub
SL	99,10	96,02	96,78
EL	0,00	0,28	0,00
AL	0,90	3,69	3,22
SLSA	100,00	99,71	100,00
SM	99,25	95,24	96,27
EM	0,00	0,63	0,00
AM	0,75	4,13	3,73
SMSA	100,00	99,34	100,00

D'après le tableau, seule la fonte amdh a généré un faible pourcentage d'erreurs au niveau des lettres (EL = 0,28 %). En ne tenant pas compte des abstentions au niveau des lettres (le système s'abstient de prendre une décision), les pourcentages de reconnaissance des lettres (ratio SLSA) sont de 100,00 % pour les fontes amss et ozub, et de 99,71 % pour la fonte amdh.

De plus, nous avons montré que, une fois les attributs extraits et les hypothèses générées, seul un petit pourcentage de lettres d'un texte avaient à être reconnues pour en établir la reconnaissance totale : pour un dictionnaire réduit de 20 000 mots, environ 10 % de toutes les lettres d'un texte sont nécessaires pour reconnaître le texte en entier. Le concept de séparation est une nouvelle approche au problème des lettres liées : plutôt que d'essayer d'établir une coupure précise entre 2 lettres (segmentation), on ne fait que localiser les attributs sur chacune des lettres, ce qui est beaucoup plus simple que la segmentation ; à l'étape de réduction des hypothèses, on évitera de faire la reconnaissance de ces lettres liées. L'application de ce concept de séparation a permis de diminuer le pourcentage d'abstention au niveau des mots de plus de 2 % dans certains cas.

Pour prolonger les vues de ce projet, il serait intéressant d'analyser l'effet qu'aurait la variation de 2 paramètres du système de reconnaissance de textes à partir des mots :

— On sait que les attributs choisis (ascendantes, descendantes, points, boucles, nombre de caractères) ont donné de bons résultats quant au nombre d'hypothèses générées en fonction de la taille du dictionnaire réduit ; quel serait l'effet d'ajouter (ou de remplacer) des attributs judicieusement choisis ? Peut-on tendre vers une répartition plus « serrée » que celle obtenue dans ce projet ?

— La localisation des attributs se fait par la position de la lettre (dans le mot) à laquelle il appartient. Quel effet aurait sur le taux de reconnaissance l'utilisation d'une autre méthode de repérage des attributs ? Par exemple, la méthode de repérage normalisée : pour tenir compte de la grosseur du caractère, la longueur du mot est exprimée par un ratio (longueur du mot divisée par sa hauteur moyenne) ; la position des attributs dans le mot est exprimée en fraction de ce ratio (de façon continue ou discrète). Un autre exemple de méthode de repérage des attributs serait l'introduction d'une tolérance sur la position de l'attribut (ex. : l'attribut est situé à la troisième lettre + 1 lettre) pour tenir compte de l'effet des lettres liées ; quel effet aurait cette tolérance sur le nombre d'hypothèses générées en fonction de la taille du dictionnaire réduit ?

En dernier lieu, l'adjonction d'un système de reconnaissance sémantique et syntaxique à ce système de reconnaissance de mots serait l'ultime développement ; le travail de reconnaissance se ferait sur plusieurs niveaux : les caractères, le contexte entre les lettres (les mots) et le contexte entre les mots (le sens et la syntaxe de la phrase). Un tel système peut faire de la reconnaissance de textes dans un environnement très bruité car le système a la possibilité d'aller chercher l'information pertinente à la reconnaissance là où elle est réellement disponible.

Manuscrit reçu le 9 mai 1990, version révisée le 20 août 1991.

BIBLIOGRAPHIE

[1] C. Y. Suen, « Character Recognition by Computer and Applications », in *Handbook of Pattern Recognition and Image Processing*, Academic Press New York, 1986, pp. 569-586.

[2] C. Y. Suen, M. Berthod and S. Mori, « Automatic Recognition of Handprinted Characters. The state of Art », *Proceedings of the IEEE*, vol. 68, pp. 469-487, April 1980.

[3] C. Y. Suen, « Distinctive Features in Automatic Recognition in Handprinted Characters », *Signal Processing*, vol. 4, pp. 193-207, 1982.

[4] S. Kahan, T. Pavlidis and H. S. Baird, « On the Recognition of Handprinted Characters. The State of Art », *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 9, pp. 274-288, March 1987.

[5] J. J. Hull, « Hypothesis Generation in a Computational Model for Visual Word Recognition », *IEEE Expert*, vol. 4, pp. 63-70, Fall 1986.

[6] J. J. Hull and S. N. Srihari, « A Computational Approach to Visual Word Recognition : Hypothesis Generation and Testing », in *Proceedings of IEEE Computer Society Conf. Computer Vision and Pattern Recognition, Miami Beach, Fla.*, June 1986, pp. 156-161.

[7] J. J. Hull, « Inter-word Constraints in Visual Word Recognition », in *Proceedings of Sixth Canadian Conf. Artificial Intelligence, Montréal, Canada*, May 1986, pp. 134-138.

[8] J. J. Hull, « The use of Global Context in Text Recognition », in *Proceedings of IEEE Computer Society International Conf. on Pattern Recognition, Paris, France*, October 1986, pp. 1218-1220.

[9] M. J. Adams, « Models of Word Recognition », *Cognitive Psychology*, vol. 11, pp. 133-176, 1979.

[10] A. J. Marcel, « Conscious and Unconscious Perception : Experiments on Visual Masking and Word Recognition », *Cognitive Psychology*, vol. 15, pp. 197-237, 1983.

[11] J. C. Johnston, « A Test of the Sophisticated Guessing Theory of Word Perception », *Cognitive Psychology*, vol. 10, pp. 123-153, 1978.