

THESE
(NOUVEAU REGIME)

présentée au

CENTRE DE RECHERCHE EN INFORMATIQUE DE NANCY

de l'

INSTITUT NATIONAL POLYTECHNIQUE DE LORRAINE

par

Hervé Cuvelier

pour obtenir



LE TITRE DE DOCTEUR EN INFORMATIQUE

**UNE APPROCHE DE LA LOCALISATION DE L'INFORMATION
APPLIQUEE A LA RECONNAISSANCE DU COURRIER MANUSCRIT**

Soutenu le 21 novembre 1988 devant la commission d'examen

JURY

Monsieur

Messieurs



D 136 037520 7

K. Tombre
P. Mohr
J.C. Merlivat

Président

Rapporteur
Rapporteur

Examineur
Examineur
Examineur

1360375207

[H] 1988 CUVELIER H.

A l'issue d'un D.E.A. de Reconnaissance des Formes dans le laboratoire de M Berthod (I.N.R.I.A. Sophia-antipolis), et désireux de poursuivre mes investigations dans ce domaine, je me mis à rechercher une société française susceptible de m'accueillir pour effectuer un travail de thèse sous convention C.I.F.R.E. La Compagnie Générale d'Automatisme, en la personne de M Merlivat, accepta ma proposition et me laissa le choix de mon partenaire universitaire. Sur les bons conseils de M Berthod, je rencontrai au Centre de Recherche Informatique de Nancy M Mohr, qui me fit confiance et devint mon responsable de thèse. Il ne me restait plus qu'à obtenir la convention de l'Association Nationale de la Recherche Technique. En novembre 86, la commission agréa le contrat de collaboration industrie-université en établissant la convention N° 340/86.

Je tiens à ce que ces personnes citées trouvent ici l'expression de ma gratitude pour la confiance et l'aide qu'elles m'ont accordées.

Je remercie M Tombre, M Marchand et M Haton de me faire l'honneur d'accepter de juger ce travail.

Que toutes les personnes, de la société CGA-HBS et du laboratoire du CRIN trouvent dans ces lignes, le témoignage de ma plus vive reconnaissance pour leur aide et leurs conseils.

Mr Courjaret, Mr Gardin du Boisdulier, Mr Vanhulle, qui ont travaillé conjointement sur ce projet, m'ont permis d'appréhender les méthodes de travail des chercheurs de l'industrie et m'ont fait profiter de leur expérience. Qu'ils en soient remerciés très sincèrement.

Service Commun de la Documentation
INPL
Nancy-Brabois

TABLE DES MATIERES

1 - INTRODUCTION.	9
2 - CONSIDERATIONS SUR L'ORGANISATION "LECTEUR DE CODE".	15
2.1 - INTRODUCTION.	15
2.2 - CONDITIONS DE TESTS ET VALIDATION D'UNE APPROCHE.	16
2.3 - FLEXIBILITE D'UN SYSTEME.	17
2.4 - SELECTION D'UNE ARCHITECTURE.	17
2.4.1 - Investigations.	17
2.4.2 - Architecture envisagée.	20
3 - BINARISATION D'IMAGES.	26
3.1 - INTRODUCTION.	26
3.2 - RECHERCHE BIBLIOGRAPHIQUE.	29
3.2.1 - Seuillage global.	29
3.2.2 - Seuillage local.	30
3.2.3 - Seuillage dynamique.	33
3.3 - METHODES EXPERIMENTEES.	34
3.3.1 - Test d'une binarisation par seuillage d'histogrammes locaux.	34
3.3.2 - Test d'une binarisation par seuillage dynamique.	37
3.3.3 - Conclusion.	40
3.4 - ALGORITHME SELECTIONNE.	41
3.4.1 - Seuillage par "test de binarisation".	44
3.4.2 - "Lissage" de l'image binaire brute.	47
3.4.2.1 - Le nettoyage.	48
3.4.2.2 - Le bouchage.	49
3.4.2.3 - exemples de formes lissées et modèle de cablage des opérations.	49
3.5 - EVALUATION DES PERFORMANCES.	52

5.4 - ETIQUETAGE DES FORMES.	85
5.4.1 - Introduction.	85
5.4.2 - Critères d'identification.	86
5.4.3 - Détection des doublets.	87
5.4.4 - Détection de la virgule.	89
5.4.4.1 - Introduction.	89
5.4.4.2 - Détection.	89
5.4.4.3 - Etude du contexte.	89
5.4.4.4 - Validation.	90
5.5 - RESTAURATION DES FORMES BRISEES.	91
5.5.1 - Introduction.	91
5.5.2 - Investigations.	91
5.5.3 - Classification des formes brisées.	92
5.5.4 - Méthodologie de la restauration.	93
5.5.5 - Validation.	95
5.6. - SEPARATION DES CHIFFRES CONNECTES.	96
5.6.1 - Introduction.	97
5.6.2 - Investigations.	101
5.6.3 - Conclusion.	102
5.6.4 - Méthodologie de la déconnexion.	109
5.6.5 - Evaluation.	111
5.7 - ASSOCIATION DES FORMES.	111
5.7.1 - Introduction.	111
5.7.2 - Notion de "voisinage d'association".	114
5.7.3 - Principe de la procédure d'association.	114
5.7.4 - Règles d'association et contrôle de l'inclinaison des lignes et des mots.	114
5.7.5 - Evaluation.	118

5.8 - DETECTION DE CANDIDATS CODE-ETAT-VILLE.	129
5.8.1 - Introduction.	129
5.8.1 - Concepts de bases.	131
5.8.3 - Génération et pondération des hypothèses.	138
5.8.3.1 - Typologie et homogénéité du code.	138
5.8.3.2 - Classement des hypothèses formulées à partir de chaque bloc.	141
5.8.3.3 - Rejets de configurations et choix final.	148
5.8.4 - Evaluation.	149
6 - CONCLUSION GENERALE.	156
- ANNEXES.	159
- REFERENCES BIBLIOGRAPHIQUES.	173

TABLE DES FIGURES

Figure 1.1	: Schéma fonctionnel du système de lecture des codes postaux.	11
Figure 1.2	: Système de binarisation et de lissage d'image. Multi-processeur de reconnaissance d'adresse postale.	13
Figure 2.4.1	: Structure multi-niveaux adaptée à la reconnaissance des caractères.	18
Figure 2.4.2	: Architecture à données centrales.	22
Figure 3.1.1	: Structures géométriques pertinentes.	27
Figure 3.1.2	: Relations entre les pixels en fonction du type de connexité.	27
Figure 3.2.1.1	: Histogramme bimodal.	29
Figure 3.2.1.2	: Histogramme de l'image convolué par un opérateur Laplacien.	30
Figure 3.2.1.3	: Fonction de conversion de la moyenne.	32
Figure 3.3.1.1	: Seuillages d'histogrammes locaux.	35
Figure 3.2.2.1	: Seuillage haut.	38
Figure 3.3.2.2	: Seuillage bas.	38
Figure 3.3.2.3	: Image binaire résultant du seuillage dynamique.	38
Figure 3.3.2.4	: Connexion des caractères lors de la croissance de pixels.	39
Figure 3.3.2.5	: Détérioration des images de faible contraste contenant des défauts d'encrage.	39
Figure 3.4.1	: Résultat de la binarisation.	42
Figure 3.4.2	: Schéma d'une représentation 3-D d'une image multi-niveaux de gris.	43
Figure 3.4.3	: Creusement des caractères épais au cours de la binarisation.	43
Figure 3.4.1.1	: Voisinage de détection des minimums et voisinage de décision pour le test de binarisation.	45
Figure 3.4.1.2	: Courbe heuristique permettant de calculer le coefficient de contraste utilisé lors du test de binarisation.	46
Figure 3.4.2.1	: Synoptique de fonctionnement des opérations de lissage.	50
Figure 3.4.2.2	: Exemple de lissage de la forme par nettoyage et bouchage.	51
Figure 3.5.1	: Répercussion des trous du caractère sur le squelette.	53
Figure 3.5.2	: Caractères brisés par la binarisation.	53

Figure 4.1.1	: Différence entre 2 méthodes de squelettisation.	55
Figure 4.2.2.1	: Calcul de la distance du pixel de la forme au squelette.	60
Figure 4.2.2.2	: Configurations d'éliminations de points.	61
Figure 4.2.2.3	: Elimination finale de points.	61
Figure 4.3.1	: Conservation des caractéristiques géométriques de la forme.	64
Figure 4.4.1	: Exemples de caractères où l'information du squelette est insuffisante.	64
Figure 4.4.2	: Effet de duplication de noeuds.	66
Figure 4.4.3	: Formation d'aigrette parasite sur le caractère.	66
Figure 4.4.4	: Caractères à boucles pleines.	66
Figure 4.4.5	: Elimination des bulles parasites du squelette.	66
Figure 5.2.1.1	: Technique de lissage par croissance dirigée.	69
Figure 5.2.1.2	: Exemple de croissance de régions sur une lettre américaine.	69
Figure 5.2.2.1	: Détection des lignes d'information par histogramme.	71
Figure 5.2.2.2	: Détection des lignes en utilisant des histogrammes locaux.	70
Figure 5.3.2.1	: Cadres des formes.	73
Figure 5.3.2.2	: Codage de Freeman.	74
Figure 5.3.2.3	: Points extrêmes nord et sud.	74
Figure 5.3.3.1	: Description des formes par suivi de contour.	75
Figure 5.3.4.1	: Architecture "pipe-line" pour une acquisition, une binarisation, et une description en continue et parallèle.	77
Figure 5.3.4.2	: Voisinage de travail utilisé pour la description de l'image par balayage.	77
Figure 5.3.4.3	: Description de la forme et extraction des paramètres.	78
Figure 5.3.4.4	: Suivi des arborescences et configurations de segments.	78
Figure 5.3.4.5	: Intérêt de la détection des soulignements.	81
Figure 5.3.4.6	: Déconnexion des lignes d'écriture au niveau du soulignement.	81
Figure 5.3.4.7	: Détection des lignes et soulignements.	80
Figure 5.3.4.8	: Détection des formes connectées à une ligne.	82
Figure 5.3.4.9	: Calcul des points extrêmes.	83
Figure 5.4.2.1	: Critères d'étiquetage.	86
Figure 5.4.3.1	: Détection des doublets.	87
Figure 5.4.3.2	: Elimination des faux doublets.	88
Figure 5.4.4.1	: Détection de la virgule.	89
Figure 5.4.4.2	: Recherche des voisins de la virgule.	90
Figure 5.4.4.3	: Etude du contexte et critères de validation de la virgule.	90

Figure 5.5.2.1	: Restauration des formes brisées par recouvrement des profils.	91
Figure 5.5.2.2	: Association des formes par des critères de proximité.	92
Figure 5.5.3.1	: Catégories de chiffres brisés.	93
Figure 5.5.4.1	: Critères de restauration des chiffres brisés de manière accidentelle.	93
Figure 5.5.4.2	: Configurations particulières où le recouvrement des formes est très important.	94
Figure 5.5.4.3	: Critères d'association des morceaux de chiffres brisés symptomatiques.	94
Figure 5.5.5.1	: Exemples de chiffres brisés et de configurations "coupées-chaînées".	95
Figure 5.6.2.1	: Segmentation par recherche d'un minimum de l'histogramme.	98
Figure 5.6.2.2	: Coupures de chiffres connectés et générations d'hypothèses.	99
Figure 5.6.2.3	: Détection de pics de connexions dans le profil bas de la forme.	100
Figure 5.6.3.1	: Configurations de connexions complexes.	101
Figure 5.6.4.1	: Schéma fonctionnel de l'algorithme de déconnexion des doublets.	102
Figure 5.6.4.2	: Accrochage du doublet et tracé du chemin bas.	103
Figure 5.6.4.3	: Détection des points potentiels de coupure.	104
Figure 5.6.4.4	: Exemples de modélisations de zones de connexions.	106
Figure 5.6.4.5	: Ordre de priorités dans le choix des points de coupures.	107
Figure 5.6.4.6	: Mode de coupure par interpolation de 2 points ou de manière verticale.	107
Figure 5.6.4.7	: Exemples de résultats de coupures de doublets.	108
Figure 5.6.5.1	: Détections et coupures erronées des zones de connexions.	110
Figure 5.7.2.1	: Calcul de la zone d'influence d'un caractère commun.	111
Figure 5.7.2.2	: Calcul de la zone d'influence d'une barre verticale.	112
Figure 5.7.2.3	: Calcul de la zone d'influence d'un mot ou d'un graphique et élimination des jambages des mots.	113
Figure 5.7.4.1	: Documents où l'information est difficile à localiser: lignes interpénétrées et nombreuses connexions.	115
Figure 5.7.4.2	: Associations de deux caractères.	116
Figure 5.7.4.3	: Associations de deux mots.	116
Figure 5.7.4.4	: Association des formes horizontales au bloc étudié.	116
Figure 5.7.4.5	: Calcul dynamique de la base virtuelle d'un bloc.	117
Figure 5.7.4.6	: Calcul de la base virtuelle d'un mot.	117
Figure 5.7.5.1	: Résultats de la description de l'image et association des formes en blocs d'information.	119

Figure 5.8.1.1	: Syntaxe d'adressage au niveau des blocs.	130
Figure 5.8.1.2	: Syntaxe grammaticale de l'adresse visible au sein du bloc.	130
Figure 5.8.2.1	: Schéma fonctionnel du module de localisation du code, de l'état et de la ville.	132
Figure 5.8.2.2	: Procédure utilisant la syntaxe de la virgule.	133
Figure 5.8.2.3	: Procédure utilisant la syntaxe du code postal.	134
Figure 5.8.2.4.a	: Recherche des entités ville et état de manière syntaxique.	135
Figure 5.8.2.4.b	: Schéma fonctionnel du module de détection de l'état et de la ville.	136
Figure 5.8.2.5	: Détection des entités état et ville dans un bloc de formes.	137
Figure 5.8.3.1	: Typologie du code postal et méthode de détection.	139
Figure 5.8.3.2	: Rejet d'hypothèses dont le contenu du code est hétérogène.	140
Figure 5.8.3.3	: Classement des hypothèses pour un bloc donné.	142
Figure 5.8.3.4	: Code contenant un faux doublet.	143
Figure 5.8.3.5	: Concurrence entre 2 hypothèses: code, code + extension.	143
Figure 5.8.3.6	: Fiabilité des hypothèses code contenant un tiret séparateur.	144
Figure 5.8.3.7	: Foisonnement d'hypothèses à partir d'un bloc.	145
Figure 5.8.3.8	: Problèmes engendrés par de fausses détections de caractères doublets et les échecs de la restauration des chiffres brisés.	146
Figure 5.8.3.9	: Fiabilité des hypothèses "ville , état" lorsque la virgule existe.	147
Figure 5.8.3.10	: Rejets d'hypothèses: comparaisons des dimensions géométriques des 3 entités.	148
Figure 5.8.3.11	: Problèmes de choix final entre 2 hypothèses dont les taux de confiance sont suffisamment élevés pour être fiables	149

1 - INTRODUCTION.

- PRESENTATION GENERALE.

Dans la course à l'automatisation que nous connaissons actuellement, la robotique emploie très souvent la vision artificielle pour saisir des images de scènes. Ce type d'acquisition de données est généralement suivi d'un processus de reconnaissance de formes. Les contraintes de temps et les puissances de calculs limitées ne permettent pas à ces automates d'analyser une information brute souvent constituée d'images en teintes de gris. Pour réduire le volume d'information à traiter, les processus classiques de lecture de caractères débutent l'analyse d'image par une binarisation. Des pré-traitements permettent ensuite de localiser et d'extraire les caractéristiques pertinentes des formes. Le but de ces opérations est de ne soumettre à la reconnaissance que l'information utile.

Dans le domaine de la lecture optique, les recherches ont tendance à favoriser les phases de reconnaissance par rapport aux étapes de localisation de l'information. Les publications scientifiques relatent principalement les efforts fournis sur les tâches de reconnaissance au sens strict du terme. Peu d'ouvrages décrivent les traitements qui permettent, à partir d'une image, de fournir aux modules de lecture, des caractères ainsi que leur contexte. Une des raisons essentielles est que l'information est souvent structurée ou contrainte dans une zone de l'image, suivant qu'il s'agit de la lecture de documents dactylographiés ou d'imprimés à cases...

Si l'on désire analyser un document manuscrit non contraint, un effort doit être porté sur les phases de pré-traitement, car l'information y est moins accessible. Les techniques classiques de localisation ne sont plus suffisantes pour être utilisées de manière autonome. Elles doivent être combinées suivant un ensemble de règles issues d'une syntaxe, d'heuristiques, ou d'événements statistiques.

La reconnaissance d'adresse postale est un terrain privilégié pour élaborer de telles approches. Aussi, dans le cadre d'un projet sur la lecture des codes postaux américains manuscrits, les pré-traitements nous sont apparus aussi essentiels que la reconnaissance. En effet, les images de ce courrier hétérogène contiennent tous les problèmes courants du traitement d'images et de la reconnaissance des formes.

Les images de lettres sont texturées, bruitées, et n'ont jamais la même dynamique de teintes. Les lignes de l'adresse sont inclinées et souvent imbriquées. Les syntaxes d'écriture sont nombreuses. Les caractères et les chiffres sont chaînés ou coupés et parfois connectés à des traits de soulignement.

- PRESENTATION DU PROJET.

L' "United States Postal Service" (U.S.P.S.), homologue américain des P.T.T. françaises, a émis en juin 1986 un appel d'offre mondial dont l'objet était la reconnaissance du code postal américain manuscrit.

Les fonctions que devra remplir le système résument cette proposition :

- la localisation du **code postal**, de l'**état**, de la **ville**,
- l'identification des chiffres de ce **code postal** ("Z.I.P. code"),
- l'identification de l'**état**.

Sur les trente trois sociétés et universités qui ont été contactées pour répondre à cet appel d'offre, seules quatre ont été retenues :

- Compagnie Générale d'Automatisme (CGA-HBS)
St. Denis FRANCE.
- ElectroCom Automation, inc.
Arlington TX U.S.A.
- Environment Research Institute of Michigan (ERIM)
Ann Arbor, MI U.S.A.
- State University of New-York (SUNY)
Buffalo, NY U.S.A.

Le projet correspond à l'identification des problèmes de **réduction**, de **localisation**, de **reconnaissance** de l'information manuscrite et à la présentation des solutions envisagées pour les résoudre. La figure (1.1) représente la démarche qui a été retenue pour ce projet.

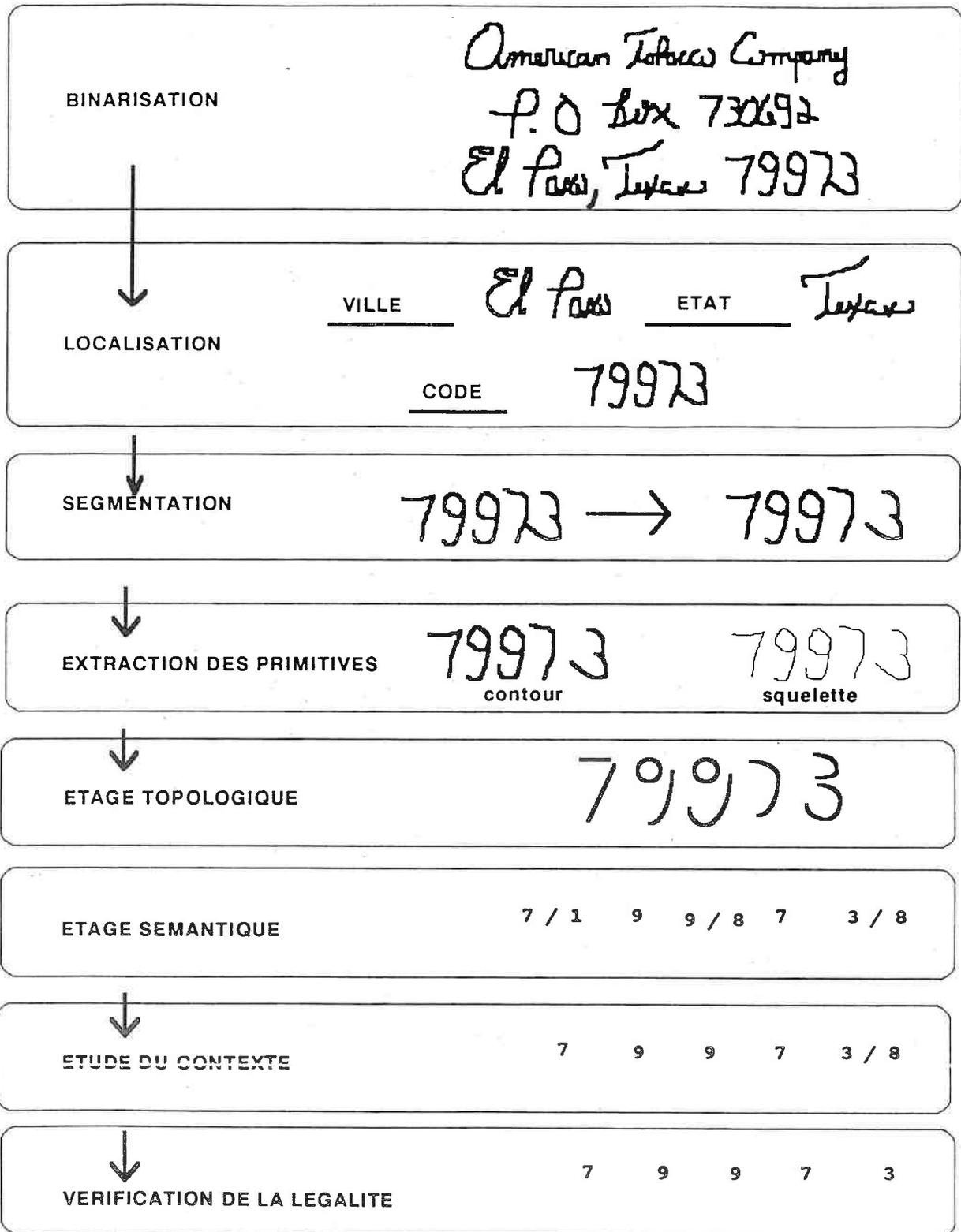


Figure 1.1 Schéma fonctionnel du système de lecture des codes postaux..

- UNE METHODOLOGIE DE TRAVAIL FONDEE SUR UNE EXPERIENCE INDUSTRIELLE.

Dans le cadre ce projet, une bibliothèque contenant 3.000 images de lettres a été fournie par l' "U.S.P.S.". Plusieurs banques de données, françaises et étrangères, ont été consultées et nous ont conduits à rassembler plus de 300 articles traitant des problèmes de **traitements opto-numériques des documents** et de la **reconnaissance des caractères**.

Les investigations et les réalisations algorithmiques furent conduites par un groupe de quatre personnes rattachées au service de Reconnaissance Optique des Caractères de la société CGA-HBS. Il y a cinq ans les Postes-Télécommunications-Télégraphes françaises (P.T.T.) ont fait appel à la Compagnie Générale d'Automatisme (CGA-HBS) pour la réalisation d'une machine capable de reconnaître le code postal du courrier manuscrit. Ce lecteur de code sera opérationnel fin 89.

Le système de lecture, représenté sur la figure (1.2), est intégré sur une machine de tri et se compose de deux ensembles bien distincts :

- Le SYBILIS permet l'**acquisition** d'une image "multi-niveaux" de gris lors du déplacement de la lettre. Il effectue ensuite la **binarisation** et le **lissage** de l'image. Toutes les opérations se font en continu, ce qui donne un temps de "pipe-line" de 70 ms.
- Le PRISME, système "multi-processeur" à base de 50 TRANSPUTERS, effectue la lecture du code dans l'adresse et vérifie la légalité de celui-ci dans un dictionnaire en moins de 500 ms.

Les acquis de la société en matière de lecture optique ont permis d'aborder cette étude avec une connaissance solide des problèmes de RF liés au traitement du courrier et de lui donner un certain nombre de contraintes nécessaires à la réalisation d'un projet industriel: contraintes de temps de calcul, flexibilité du système...

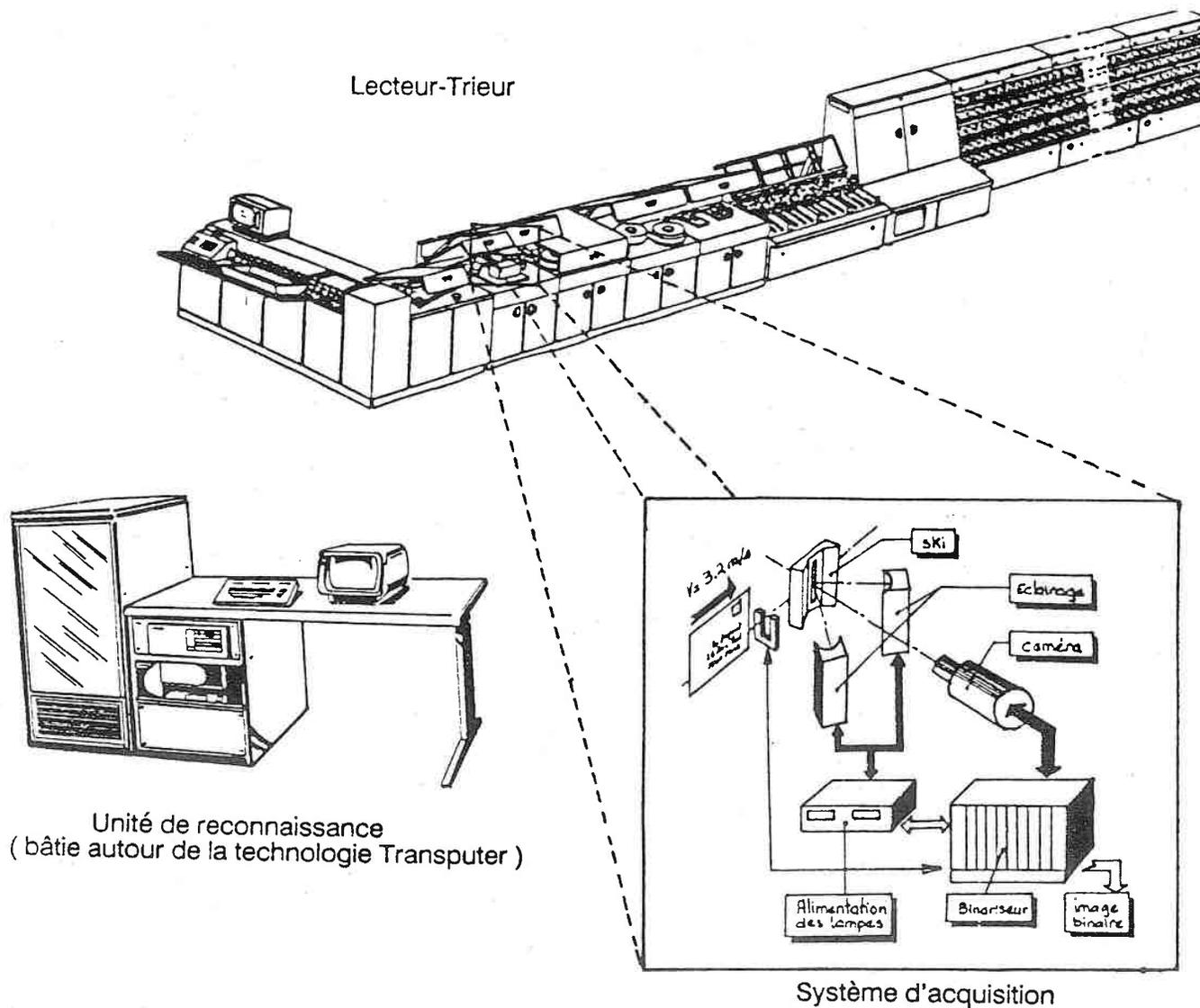


Figure 1.2 : Système de binarisation et de lissage d'images

: Multiprocesseur de reconnaissance d'adresse postale

- CONTENU DE CE MEMOIRE..

Chaque chapitre traite d'un problème spécifique à l'écriture manuscrite et présente les solutions envisagées pour le résoudre. Dans le cadre de cette thèse, l'objectif était de développer tous les outils qui permettent de fournir à l'étage de reconnaissance des caractères numériques, isolés et entiers, à partir de l'image multi-niveaux de gris.

Nous commencerons notre réflexion sur les possibilités de l'**architecture du système**. De manière générale, les procédures d'un lecteur de code et les simulations que nous avons effectuées, sont agencées selon une séquence d'opérations proche de celle décrite dans la figure(1.1). Nous aborderons ensuite la réduction de l'information au travers de la **binarisation** de l'image, de l'extraction du **squelette** et du **contour** des formes. Dans le chapitre traitant de la binarisation, nous étudierons les méthodes de seuillages d'histogrammes et les possibilités qu'offre l'étude du voisinage d'un point. Plusieurs méthodes de squelettisation seront exposées. Elles utiliseront des processus d'amincissements homogènes ou dirigés. Après la présentation de ces méthodes de traitement d'images, les chapitres suivants seront plus orientés vers la reconnaissance des formes. Pour résoudre les problèmes explicités auparavant, il nous est apparu intéressant de travailler sur les extrémités et les caractéristiques géométriques simples des caractères, car elles permettent de différencier certaines catégories de formes. Une pré-identification est très instructive quand celle-ci vous permet de détecter les caractères isolés, les mots et les soulignements... Ce type de **description**, comme nous le verrons, permet une **association** des formes en blocs. La notion de points caractéristiques sera reprise pour restaurer les chiffres brisés et séparer les chiffres connectés. A partir de ce mode de représentation, une série d'hypothèses fondées sur la syntaxe de l'adressage postal est envisageable, et permet de réduire le nombre des candidats **code+état+ville** à la reconnaissance. Cette tâche sera abordée dans le dernier chapitre de la **localisation**.

2 - CONSIDERATIONS SUR L'ORGANISATION DU SYSTEME " LECTEUR DE CODES " .

2.1 - INTRODUCTION .

Actuellement, le système développé au cours de ce projet se définit plutôt comme une simulation algorithmique d'un lecteur de code. Il nous est apparu préférable de réaliser les différents modules du système en nous appuyant sur un schéma fonctionnel idéal. Pour que l'architecture soit flexible et permettent d'évaluer ou de modifier chaque module de manière indépendante, nous avons recherché ce qui fait la qualité et la fiabilité d'un automate de reconnaissance des caractères.

De nombreux chercheurs et industriels ont abordé les problèmes de la reconnaissance de l'écriture manuscrite. En ce qui concerne l'écriture occidentale, l'intérêt s'est plutôt porté sur les caractères **bâtons**, imitation manuscrite des polices de caractères dactylographiés. Les expériences et les réalisations issues de l'industrie sont souvent confidentielles et n'apparaissent qu'à l'occasion de certains congrès privés où les sociétés sous contrat exposent leurs résultats (exemple : congrès de l'"U.S.P.S."). Les méthodologies défendues dont les résultats sont rarement comparables, tendent à prouver que la reconnaissance des caractères n'est pas le fruit d'une démarche unique . Les nombreuses réalisations issues de ces recherches apparaissent bien souvent comme étant des systèmes clos et inflexibles . L'intégration de nouveaux modules, ou d'étages complets, devient difficile, ce qui remet en cause la fiabilité et les performances du système . Il ne faut pas perdre de vue que l'objectif premier d'un fabricant de lecteurs de codes, c'est de pouvoir adapter son automate à tous les marchés, à toutes les contraintes. Son objectif second, c'est de toujours remettre en question les performances algorithmiques et électroniques de ce système.

2.2 - CONDITIONS DE TESTS ET VALIDATION D'UNE APPROCHE.

Dans le cadre de cette étude, l'expérience de CGA-HBS, et celle de ses partenaires dans le domaine des lecteurs de codes, furent des facteurs prépondérants pour le choix des méthodes applicables à la **reconnaissance optique des caractères**. Cette connaissance permet de porter un jugement critique sur les résultats publiés. Les conditions de tests, et la formulation des résultats de ces publications sont hétérogènes. Celles-ci rendent difficile, voire impossible, l'estimation des performances et les comparaisons. En effet, il est impossible de considérer ou de comparer deux méthodes, si les résultats sont formulés différemment, et si l'évaluation est effectuée sur des documents non représentatifs.

Les résultats susceptibles d'intéresser un industriel avisé doivent être issus d'évaluations réalisées selon certaines normes .

La normalisation des conditions de tests se traduit par ces quelques règles :

i) - Les performances des algorithmes de reconnaissance doivent être exprimées en :

- taux de reconnaissance ,
- taux de rejet ,
- taux de confusion (erreurs).

ii) - Pour chaque évaluation d'algorithme , quelle que soit son importance dans le système global , l'apprentissage et les améliorations doivent se faire à partir d'une base de données images, et l'évaluation sur une autre . Une base d'images se définit comme un échantillon d'images de lettres ou de formes représentatives du courrier étudié. Son contenu peut osciller entre quelques centaines et quelques milliers de documents. L'écriture appliquée de la population d'un laboratoire ne peut pas représenter un échantillon représentatif des styles négligés d'une nation . A titre d'exemple, une évaluation de performances d'un lecteur de code manuscrit se fait sur une base de 10.000 lettres, soit 50.000 caractères à analyser.

2.3 - FLEXIBILITE D'UN SYSTEME.

Pour être industrialisable , un système de reconnaissance doit pouvoir s'adapter aux différentes caractéristiques du courrier, de la syntaxe d'écriture et du tracé des caractères qui diffère d'un pays à l'autre . Le système ne devant pas changer d'architecture , sa flexibilité doit permettre l'intégration et l'évolution des modules à tous les niveaux . Ceci implique aussi la possibilité d'évaluer de manière indépendante ou globale les modules modifiés. En effet, l'augmentation des performances d'un système entraîne très souvent une complication des modules et une désorganisation de la gestion des données. L'évaluation devient alors difficile. La recherche des causes d'erreurs doit s'extraire de résultats cumulant les limites fonctionnelles de plusieurs algorithmes. Cette flexibilité architecturale, que l'intégration finale tendra d'approcher, est peut être apportée par une organisation dont le centre est une source de données. Chaque fonction puise ses sources dans cette base centrale de données et lui retransmet les résultats. Les modules peuvent être utilisés simultanément, les échanges d'informations se faisant par l'intermédiaire d'un noyau . Dans ce système idéal, les données peuvent guider les actions. Cette indépendance de toutes les tâches permet peut être de conserver un système ordonné et évaluable.

2.4 - SELECTION D'UNE ARCHITECTURE.

Afin de choisir une structure d'accueil pour nos modules traitant de problèmes spécifiques, nous nous sommes inspirés des réalisations décrites dans la littérature.

2.4.1 - INVESTIGATIONS .

La plupart des systèmes de reconnaissance issus de notre bibliographie se fondent sur des structures primaires où l'utilisation des données de sources différentes se fait uniquement de manière hiérarchique et séquentielle .

- Comme le montre la figure (2.4.1), les données de systèmes multi-niveaux se propagent d'étage en étage, au fur et à mesure qu'elles sont calculées. En résumant le processus d'identification, l'image digitalisée est d'abord binarisée , puis les

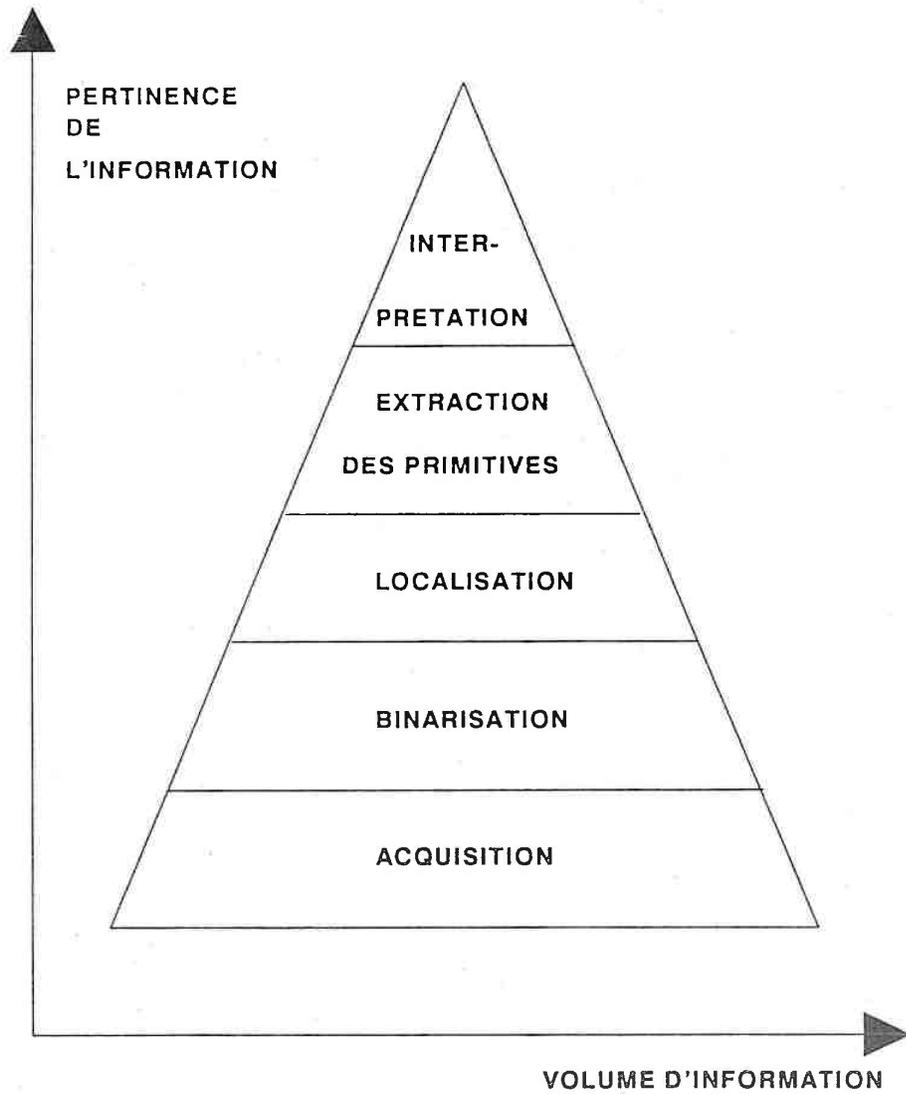


Figure 2.4.1 Structure Multi-niveaux adaptée à la
Reconnaissance des Caractères.

distributions de points ou les primitives géométriques des caractères en sont extraites . L'information ainsi décrite est soumise au **classificateur** qui identifie ou rejette celle-ci . L'activation de ces différents étages se faisant de manière séquentielle , l'intégration d'un nouvel étage à un niveau donné nécessite une connaissance parfaite de la gestion de l'information dans les niveaux inférieurs et supérieurs . Il apparait difficile d'harmoniser les données sans dégrader les performances initiales du système .

D'autres expériences ont été menées à partir d'une structure où plusieurs **classificateurs** travaillent en parallèle et communiquent entre eux . Par exemple , le système d'identification des chiffres manuscrits non contraints, proposé par B. Duerr (DUE 79), est fondé sur la combinaison d'une méthode structurelle et d'une méthode statistique. Chaque étage estime la fiabilité de sa classification. Le cumul des décisions sanctionne la forme comme étant définitivement reconnue ou devant subir d'autres tests . M. Shridhar et A. Badreldin (SHR 85) , J.C. Tou et R. Gonzalez (TOU 72) ont élaboré des systèmes similaires où la reconnaissance se fait à deux niveaux . Le premier niveau a la charge de classer définitivement les formes robustes caractérisées par des taux de confiance élevés et de rassembler les formes topologiquement proches dans des classes de confusion . Le second niveau lève l'ambiguïté sur les formes contenues dans les classes de confusion.

S.N. Srihari et R. Bozinovic dans les publications (SRI 83) , (BOZ 84), considèrent qu'un système bâti autour d'une base centrale de données , appelée "blackboard" en Intelligence Artificiel, permet la communication des différents sources de connaissance . Chaque résultat est obtenu par des procédures indépendantes assumant leurs propres limites et leurs erreurs . De nombreuses données de types différents peuvent coopérer à l'identification du code postal :

- les résultats issus de la détection des entités **code + état + ville** ,
- Les primitives topologiques des formes connexes issues de ces entités candidats ,
- Les hypothèses formulées au niveau de la reconnaissance du caractère.

Ce type d'architecture a été utilisé pour la première fois par l' "Information Sciences Institute" dans un système-expert de compréhension du langage , HEARSAY II , développé dans les années 70, et fut suivi de HEARSAY III , un langage de gestion de la connaissance permettant de contrôler et gérer de multiples sources d'informations .

S.N.Srihari et R.Bozinovic étudièrent aussi la manière de faire communiquer plusieurs démarches dans le but de reconnaître l'écriture cursive . Cette réflexion les a conduit à réaliser un système où une procédure de segmentation des mots génère des formes candidates ; un analyseur de formes identifie chaque forme et pondère des hypothèses "mot"; un lexique vérifie l'orthographe du mot ; parallèlement , une procédure calcule la probabilité de présence de préfixe, les combinaisons possibles de lettres et un analyseur grammatical vérifie la légalité du mot par rapport à sa position dans la phrase . Ce type de système parait facilement évaluable puisque chaque tâche évolue indépendamment . Cette indépendance permet de faire des améliorations et d'insérer d'autres applications sans pour cela dégrader les performances originelles .

2.4.2 - ARCHITECTURE ENVISAGEE .

L' "implémentation" d'une organisation utilisant plusieurs approches simultanées ou bien contenant des procédures capables de rebrousser chemin ("back-tracking") nous est apparue difficile et bien coûteuse à réaliser en considération de la durée du projet et du type d'information que le système aura à gérer. L'utilisation de plusieurs démarches entraîne une multiplication proportionnelle des tâches de codage, d'évaluation et d'amélioration. De plus, le choix du candidat final, décision qui n'est jamais triviale, doit être inévitablement précédé d'une confrontation des résultats obtenus par les différentes approches.

Dans le cadre de ce projet , après examen de ces publications et des investigations menées au sein de la société , nous nous sommes inspirés de l'architecture "**blackboard structure**" décrite dans les documents d'I.A., pour faire une esquisse de l'"architecture à données communes" de notre lecteur. De plus , comme cette étude est menée parallèlement par d'autres sociétés et universités dont certaines travaillent avec des langages symboliques, ce type d'organisation permet une grande flexibilité dans l'hypothèse d'insertions ou de remplacements de modules par d'autres plus performants . En effet, les modules que nous avons développés ont les caractéristiques des "Knowledge Sources" définies en I.A.. Sur le plan de la localisation de l'information et de la reconnaissance des caractères, les performances d'un système ne se jugent pas principalement au taux de sanctions justes mais au taux de confusions. Le taux de confusion se répercute fortement sur le succès d'un lecteur de caractères car dans certaines applications bancaires ou administratives, l'erreur est normalement proscrite.

Pour combattre les confusions, l'étage d'identification des caractères utilise deux **classificateurs** :

- Le premier, **de topologie globale**, étudie les primitives globales de la forme (segment, noeud, longueur, proportion, courbure...) et catalogue celles-ci dans une classe correspondant soit à une des variantes morphologiques d'un caractère, soit à une classe de confusion dont le modèle est la frontière de formes morphologiquement proches.
- Le deuxième, **de topologie fine**, étudie minutieusement chaque segment de la forme (épaisseur, point d'inflexion, orientation, angle...) et les relations spatiales existantes entre eux. Ce classificateur confirme de manière spécifique les modèles de confiance et lève l'ambiguïté sur les éléments des classes de confusions.

L'illustration (2.4.2) résume la structure du système choisie pour l'intégration des modules développés. La base centrale de données est divisée en deux niveaux . Le premier représenté par le **tableau de données** gère uniquement les images et les données numériques . Le second, réduit au **tableau des hypothèses**, contient l'information issue des raisonnements effectués à tous les niveaux.

Sur ces deux tableaux , formant le noyau central , sont greffés sept modules principaux :

- **Module de binarisation .**

L'image de la lettre acquise en multi-niveaux de gris est transformée en image binaire et stockée dans le tableau de données .

- **Modules de localisation de la zone adresse et détection des entités CODE-ETAT-VILLE .**

La localisation de ou des lignes de l'adresse entraîne une génération d'hypothèses code + état + ville . Ces listes de candidats sont pondérées, classées et stockées dans le tableau de données .

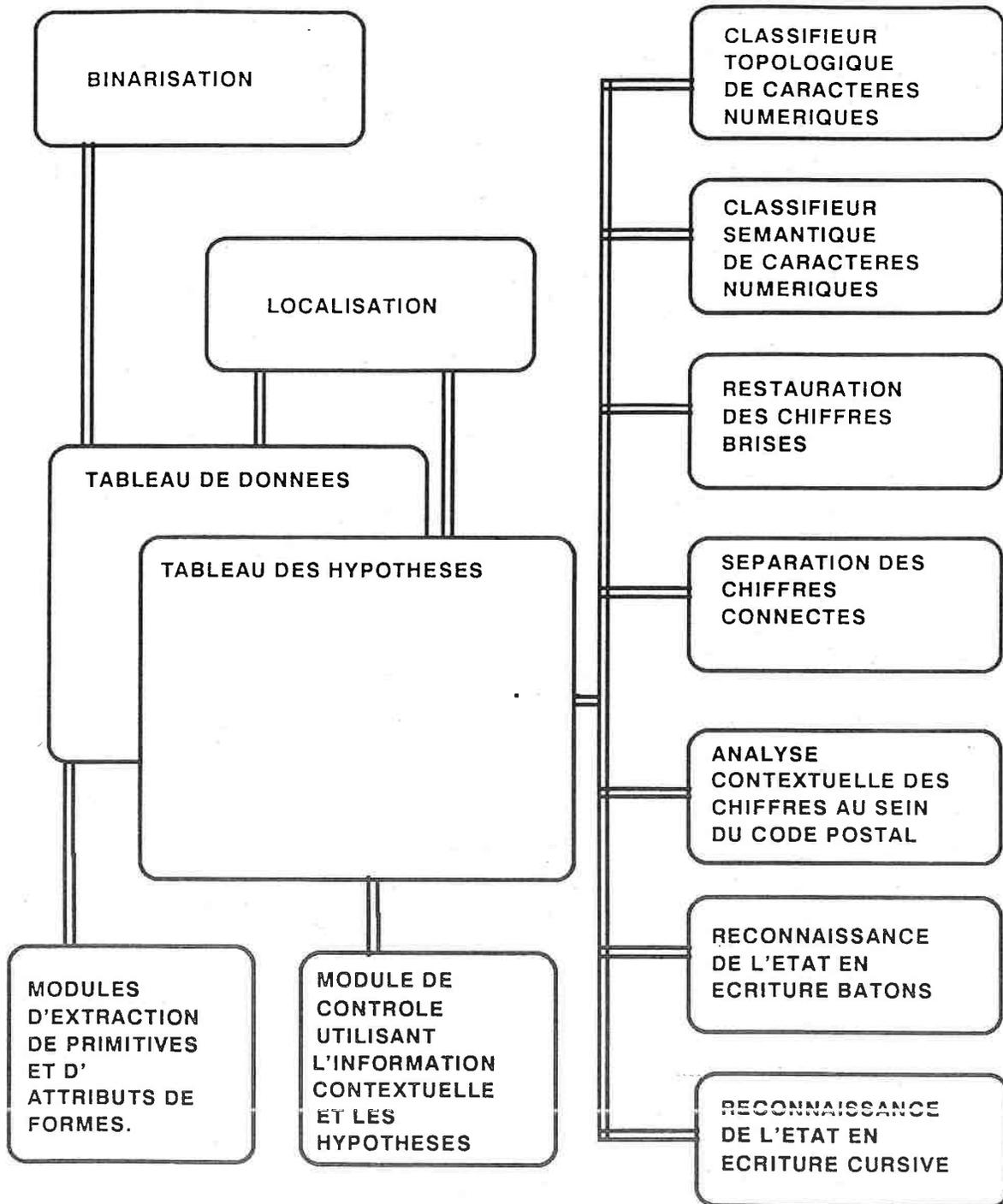


Figure 242 : Architecture à Données Centrales.

- **Module d'extraction des primitives .**

Ceci concerne l'extraction des primitives pertinentes des formes squelettisées qui seront abondamment utilisées dans les phases de reconnaissance . Des "**utilitaires de calcul**" y sont intégrés , fournissant sur la demande des classificateurs un lot de mesures et de résultats de tests effectués sur le squelette et le contour de l'image binaire. Les informations sont stockées dans le tableau des données .

- **Module d'identification des chiffres .**

Pour la ou les hypothèses code + état + ville les plus fiables , ce classificateur étudie toutes les formes du code postal candidat . Une identification est réalisée pour chaque caractère de ce code et le résultat est stocké dans le tableau d'hypothèses .

- **Modules d'analyse des chiffres brisés ou connectés .**

Ces modules sont capables, par des critères géométriques simples, de détecter les caractères connectés et les chiffres brisés lors de la localisation de la zone d'intérêt de l'image et de la déduction des codes postaux hypothétiques. Le but de ces modules est de fournir au classificateur des chiffres entiers non connectés. Les résultats de fusions et de coupures de formes sont stockés dans le tableau d'hypothèses .

- **Module d'identification de l'état .**

Comme le classificateur de chiffres , celui-ci est affecté à la reconnaissance de l'état . A partir du squelette du mot ou d'autres formes de représentation , ce module fournit une liste des noms d'états probables . Ces hypothèses sont stockées dans le tableau d'hypothèses .

- Modules de décision .

Toutes les données se trouvent désormais contenues dans ces deux tableaux . Ce module sanctionne la lettre, comme identifiée ou rejetée, en fonction du contexte géométrique de l'adresse et de la compatibilité des résultats obtenus par les deux classificateurs sur une ou plusieurs hypothèses "code-état-ville".

- Information contenue dans le tableau de données :

- images binaires sous format matriciel ,
- images des contours sous format matriciel ,
- images des squelettes étiquetés sous format matriciel ,
- descriptions des squelettes de formes ,
- mesures métriques et tests effectués sur le squelette ou ses primitives ,
- base d'informations sur la dernière ligne de l'adresse ,
- résultats de l'apprentissage bayésien (méthode statistique).

- Information contenue dans le "tableau des hypothèses" :

- liste de candidats code + état + ville proposée par la localisation ,
- liste de candidats code postal à 5 ou 9 chiffres, fournie par le classificateur de chiffres ,
- liste de candidats état fournie par l'identificateur de mots ,
- liste des différentes possibilités de déconnexion des doublets (et triplets) et de restauration de chiffres brisés .

Le fonctionnement de nombreux modules à objectifs différents conduit à effectuer des calculs similaires avec les mêmes données d'entrées . Pour éviter la duplication des opérations , les extractions de primitives et les résultats de parcours sont conservés dans le tableau central , et restent disponibles jusqu'à la fin du processus .

Dans les chapitres suivants , certains de ces sept étages seront présentés et commentés . Pour chaque module , des informations seront fournies sur le choix de la méthode et sur le contexte qui la définit . Les sujets traités seront :

- la binarisation,
- l'extraction des composantes connexes,
- la restauration des chiffres brisés,
- la séparation des chiffres connectés,
- la localisation de l'information recherchée.

3 - BINARISATION D'IMAGES .

3.1 - INTRODUCTION .

Les images d'enveloppes américaines ont été digitalisées avec une définition de 12 points au millimètre. Le stockage de la zone adresse se fait sur trois plans 512 x 512 pixels codés sur 8 bits (256 niveaux de gris). Chaque image américaine entière représente environ 2250 kbytes de données. Les contraintes industrielles, obligeant à tenir une cadence de lecture supérieure à **15 lettres par seconde**, rendent cet objectif irréalisable si chaque module du système doit se référer à l'image multi-niveaux de gris. En effet le volume de données à traiter (images), la complication des algorithmes qu'entraîne l'extraction des primitives sur une information ayant des spectres de gris variables, se répercute sur le temps d'exécution. La **binarisation** de l'image transforme l'image originale en une image simplifiée car binaire, diminuant ainsi le volume de données par huit. Les formes apparaissent noires sur fond blanc. La réduction de l'information doit se faire sans dégradation des parties essentielles de celle-ci. La qualité d'une binarisation se caractérise surtout par son insensibilité aux trames et textures du fond, et son adaptation aux variations du contraste et de l'éclairage (fenêtres réfléchissantes des enveloppes). Les matrices de points formées par les imprimantes à aiguilles doivent être connectées sans pour autant entraîner des agrégats de caractères successifs.

En reconnaissance des formes connexes, quelques caractéristiques doivent être préservées :

- les structures géométriques originales de la figure (3.1.1) qui seront utilisés pour l'élaboration d'une **classification topologique**,
- les dimensions géométriques des traits et des surfaces.

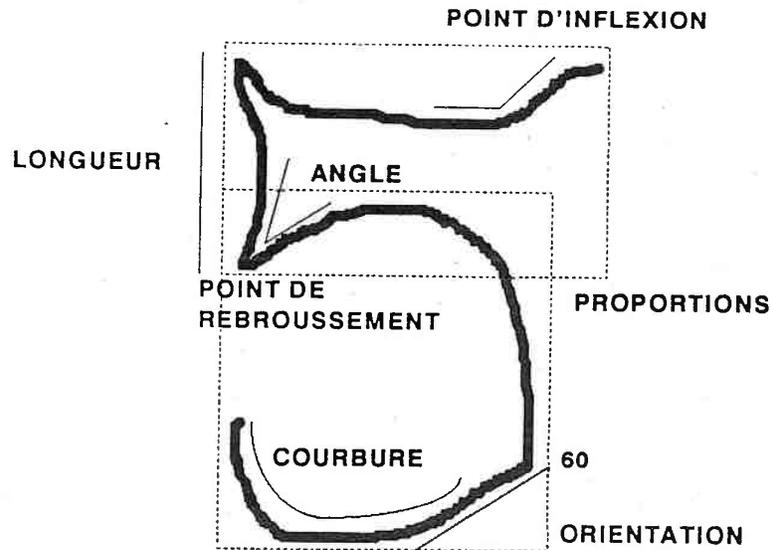
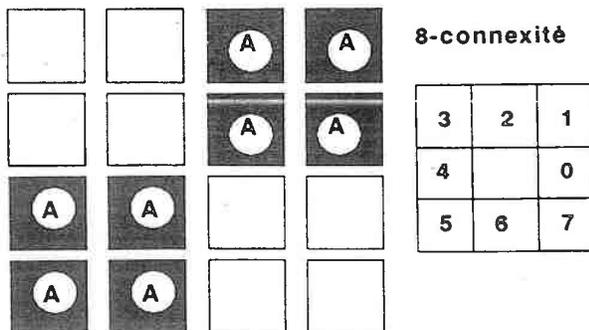


Figure 3.1.1 Structures géométriques pertinentes

Dans une image discrète, représentée par un pavage de pixels, chaque élément unité est entouré de 8 voisins. Si l'on considère dans ce plan discret, l'ensemble Z^2 des points de coordonnées entières, on peut introduire deux topologies correspondants à deux métriques différentes. Ces topologies définirons pour un point donné, un voisinage élémentaire de 4 ou de 8 voisins. Un ensemble E de points est dit connexe si pour tout couple de points A et B de E , il existe un chemin entre A et B contenu dans E , tel que tout point du chemin soit voisin du suivant. Selon le type de voisinage, on parlera de **connexité 4** pixels voisins ou de **connexité 8** pixels voisins. La figure (3.1.2) explique l'existence d'une ou de deux composantes connexes, selon le type de voisinage choisi.

les 2 amas de pixels sont connexes (1 forme A)



les 2 amas ne sont pas connexes (2 formes A, B)

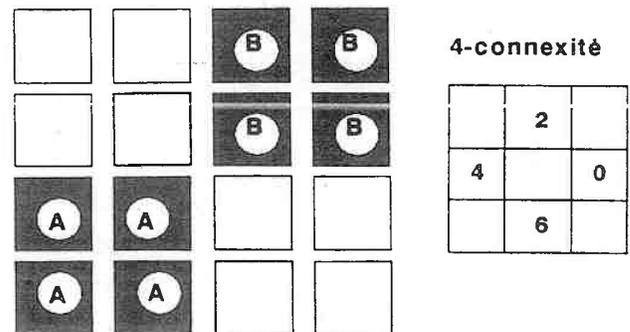


Figure 3.1.2 Relations entre les pixels en fonction du type de connexité.

3.2 - RECHERCHE BIBLIOGRAPHIQUE.

Les investigations ont été orientées vers deux sources d'informations : les références bibliographiques et les documents concernant des équipements existants. Les quelques méthodes présentées sont un reflet des techniques utilisées en reconnaissance des caractères.

Le **seuillage**, action de sanctionner de manière booléenne (noir ou blanc) un élément par rapport à un **seuil**, peut se traduire mathématiquement par la condition suivante :

Si $I(x,y) < T(x,y)$

alors $B(x,y) = 0$ (couleur noire des caractères)

sinon $B(x,y) = 1$ (couleur blanche du fond).

(x,y)	coordonnées du pixel,
$B(x,y)$	valeur 0 ou 1 du pixel dans l'image binaire calculée,
$I(x,y)$	valeur du niveau de gris du pixel dans l'image originale,
$T(x,y)$	valeur du seuil pour le pixel de coordonnées (x,y) .

L'opérateur de seuillage T peut s'exprimer d'une manière plus générale :

$$T [x, y, L(x,y), I(x,y)]$$

$L(x,y)$ représente les caractéristiques du voisinage de ce point : La valeur moyenne des teintes de gris en est un exemple.

J.S. Wezka (WES 77) divisa les seuillages en trois catégories en fonction du niveau de dépendance de l'opérateur de seuillage T .

Si T dépend seulement de $I(x,y)$ le seuillage est appelé **seuillage global**.

Si T dépend de $I(x,y)$ et de $L(x,y)$ le seuillage est appelé **seuillage local**.

SI T dépend de $I(x,y)$, de $L(x,y)$ et de la position spatiale du pixel en x et y , le seuillage est appelé **seuillage dynamique**.

3.2.1 - SEUILLAGE GLOBAL.

La méthode de seuillage statistique la plus connue et la plus employée est la méthode de **détection des modes** représentée sur la figure (3.2.1.1). Celle-ci s'effectue sur l'histogramme des distributions des pixels de l'image originale en fonction des teintes de gris. Elle consiste à déterminer la vallée qui sépare la population de pixels appartenant aux formes, de la population des pixels du fond. Cette vallée représente le niveau de gris utilisé pour le seuillage. L'histogramme n'est jamais parfait et contient de nombreuses vallées. Un lissage préliminaire de l'histogramme empêche les détections erronées. Malgré cela, plusieurs minimums significatifs peuvent persister. Le choix se porte sur la vallée la plus profonde ou sur celle qui s'étale le plus dans le spectre de teintes.

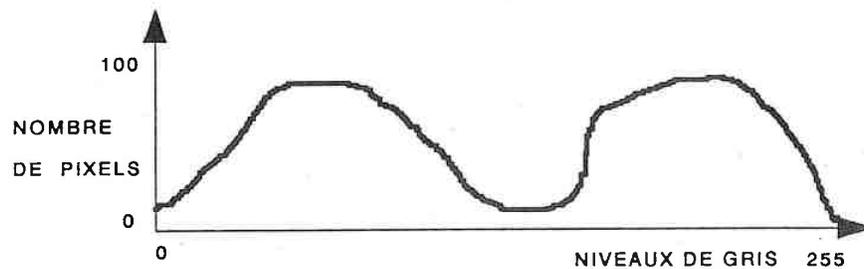


Figure 3.2.1.1 Histogramme Bimodal

Malheureusement, l'histogramme est rarement parfait et bimodal. De nombreuses méthodes permettent de faire ressortir les populations de pixels du **fond** ou des **objets** et de creuser les vallées de séparation.

Mason (MAS) propose une méthode de pondération de la distribution des pixels, en fonction d'une teinte, par des critères locaux. Les pixels, dans les zones contrastées de l'image, correspondent souvent à des transitions **fond-forme** ou **fond-texture**. A partir de la différence de niveau de gris calculée entre le pixel étudié et ses proches voisins, les distributions sont augmentées pour les pixels de zones homogènes et diminuées pour ceux de zones contrastées. Les zones homogènes du fond et des formes ressortent sous formes de pics et les textures et les transitions sont négligées.

R.N. Nagel et A. Rozenfeld (NAG 77) utilisent, pour la lecture de signatures de chèques bancaires, l'histogramme d'une image résultant d'une convolution de l'image originale par un opérateur Laplacien. Ce filtre rehausse les transitions existantes dans l'image. Ceci se traduit par une séparation plus franche des populations au niveau de l'histogramme comme le schématise la figure (3.2.1.2).

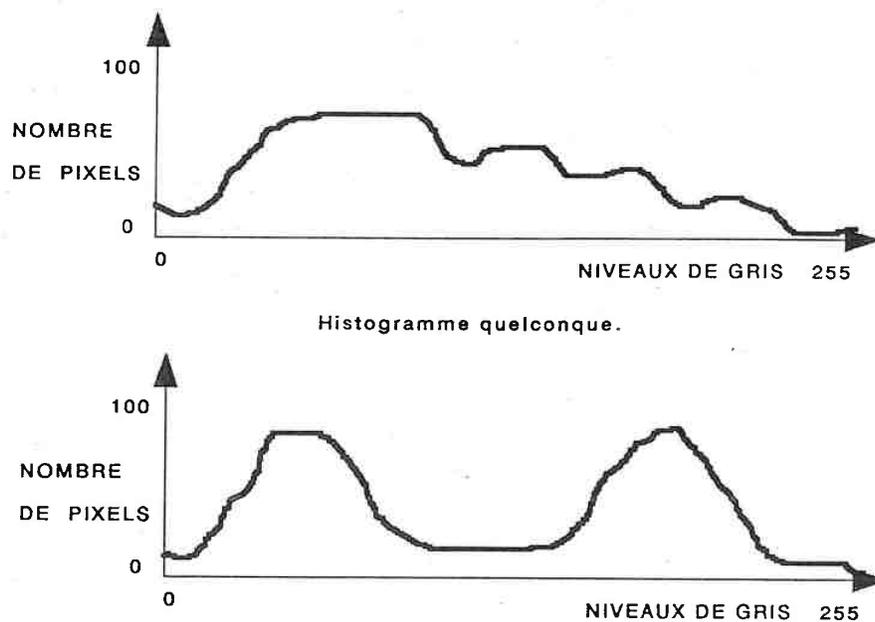


Figure 3.2.1.2 Histogramme de l'image convolué par un opérateur Laplacien.

Sur les images de courriers manuscrits, les transitions de teintes existantes au sein des formes et du fond ne sont jamais constantes et de même niveau dans l'échelle des gris. Il nous apparaît impossible d'utiliser ce type de seuillage.

3.2.2 - SEUILLAGE LOCAL.

Les opérations locales semblent plus adaptées aux types de problèmes auxquels tout système de lecture est confronté.

Premièrement, les documents originaux qui sont saisis par un **banc d'acquisition** sont souvent sales, texturés, de plusieurs couleurs, contenant des encrages et des modes d'impressions de natures différentes. De plus, les caractères sont souvent maculés et les traits apparaissent parfois peu clairement ou peu homogènes lorsqu'ils sont épais.

Deuxièmement, Le balayage d'acquisition effectue la digitalisation d'un document dont l'illumination n'est pas uniforme. A ces problèmes d'hétérogénéité d'éclairage et de réflectivité variable du support se greffent ceux de distorsion et d'absorption de l'objectif qui produit une réplique plus ou moins fidèle de l'image originale.

En utilisant le contexte dans lequel se trouve le pixel de l'image multi-niveaux de gris, il est possible de contourner ces problèmes d'hétérogénéité. L'approche la plus simple est la comparaison de la valeur du niveau de gris du pixel étudié avec celles de quelques voisins éloignés d'une certaine distance. Le pas est proportionnel à l'épaisseur moyenne du trait de l'objet recherché. Si le pixel est plus sombre que plusieurs de ces voisins, ce pixel est considéré comme "objet".

R.N. Wolfe (WOL 69) utilise deux opérateurs bidimensionnels de tailles différentes pour comparer les niveaux de gris des pixels. Le premier moyenne les niveaux de gris dans un voisinage 4 x 4. Sur cette image lissée, une comparaison est effectuée entre le pixel central $I(x,y)$ et l'ensemble de ces 8 pixels voisins séparés d'un pas de 4, $I'(x+4,y+4), I'(x-4,y-4), \dots$. Si le niveau de gris du pixel $I(x,y)$ est plus sombre que 2 voisins symétriques par rapport à lui, ce pixel est considéré comme "objet". Le second opérateur est similaire mais travaille sur un domaine plus vaste, ce qui permet de rattraper les formes épaisses.

J.M. White et G.D. Rohrer (WHI 83) soulignent les problèmes que génèrent ces méthodes de moyennes sur des voisinages de tailles variables. Ces méthodes sont coûteuses car elles nécessitent la conservation d'un grand nombre de pixels de l'image lors des calculs. Les performances montrent que ces démarches s'adaptent difficilement aux variations de contraste au sein du caractère. Une des solutions, pour éviter de travailler sur des zones trop vastes, est d'utiliser une **moyenne mobile** ("running average") lors d'un balayage des colonnes de l'image. Cette adaptation de la moyenne s'exprime mathématiquement par la formule :

$$A(y) = f * I(y) + (1-f) * A(y-1)$$

$A(y)$ représente la **moyenne mobile**,

$A(y-1)$ représente la **moyenne mobile** de la trame précédente,

$I(y)$ représente la valeur moyenne de la trame de pixels saisie,

f est une fraction.

Cette équation peut être représentée sous la forme :

$$A(y) = A(y-1) + f * [I(y) - A(y-1)]$$

Cette équation peut être rendue non linéaire et plus adaptative en remplaçant la fraction f par une fonction F

$$A(y) = A(y-1) + F [I(y) - A(y-1)]$$

La figure (3.2.2.1) représente la fonction F obtenue par l'expérimentation. Cette fonction permet d'adapter la moyenne aux brusques variations de la dynamique des teintes qui peuvent exister entre la colonne étudiée et la précédente.

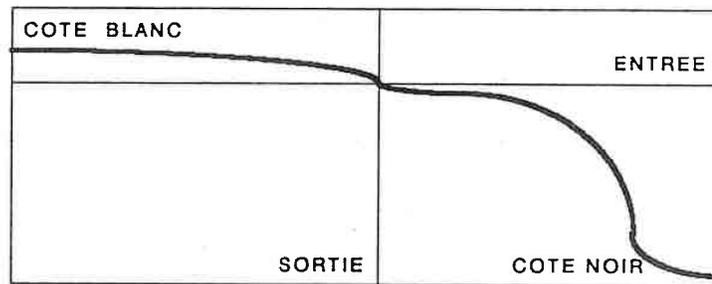


Figure 3221 Fonction de conversion de la Moyenne.

Ullman (ULL 69) propose un seuillage sélectif fondé sur les valeurs de gris de certains pixels se trouvant dans le voisinage 5×5 du pixel $I(x,y)$ étudié.

```

. X X X .
X X . X X
X . i . X
X X . X X
. X X X .

```

Dans ce voisinage de points, la recherche de la teinte la plus claire permet de décider lequel des deux tests suivants sera effectué sur le pixel étudié.

Si $\max(x,y) \leq \text{seuil } S$, le pixel étudié est **objet**,
si le test $I(x,y) - I(x',y') < \text{constante}$, T est positif pour tous ses voisins.

Si $\max(x,y) > \text{seuil } S$, le pixel étudié est **objet**,
si le test $I(x,y) < I(x',y') * \text{constante}$, M est positif pour tous ses voisins.

Des seuillages, effectués sur de petites zones de l'image, donnent des résultats satisfaisants. De nombreux équipements utilisent cette technique de **seuillages locaux**. P.W. Palumbo et S.N. Srihari (PAL 86) ont fait un exposé sur la fiabilité de ces techniques, insistant sur leur insensibilité au bruit.

3.2.3 - SEUILLAGE DYNAMIQUE.

Le seuillage dynamique a la particularité d'être fonction de la position spatiale du pixel étudié. Les démarches effectuées tendent de raccrocher le pixel à une zone de l'image par des notions de distances. Les algorithmes complexes qui en découlent nécessitent la conservation d'une grande partie de l'image. C'est pour cette raison que l'analyse du contexte spatial est plutôt utilisée après la binarisation et sur des voisinages réduits; par exemple, les amincissement et les expansions des formes (**érosion** et **dilatation**).

3.3 - METHODES EXPERIMENTEES.

Deux techniques facilement réalisables et estimées pertinentes ont été testées :

- le seuillage d'histogrammes locaux (C.K. Chow et T. Kaneko (CHO 72)),
- le seuillage dynamique.

Remarque : Les algorithmes développés qui s'inspirent de ces méthodes seront peu explicités ici afin de préserver le caractère confidentiel de leur réalisation.

3.3.1 - TEST D'UNE BINARISATION PAR SEUILLAGE D'HISTOGRAMMES LOCAUX.

Cette méthode consiste à découper l'image en petites aires égales se recouvrant et à calculer l'histogramme des distributions des pixels en fonction des teintes. Pour chacune de ces zones, un seuillage est calculé à partir de l'étude de l'histogramme. De manière classique, l'algorithme recherche la vallée séparant les deux populations **forme et fond**, où l'inflexion de la pente de l'histogramme cumulé. Ensuite chaque pixel subit un seuillage final dont le seuil correspond à l'interpolation des seuils locaux affectés aux zones entourant le pixel étudié. Cette méthode s'affranchit des problèmes de variations du contraste car celui-ci est assez stable sur des surfaces réduites. A une petite échelle, un seuillage global représente moins de danger. Des tests ont été effectués sur un échantillon de lettres particulièrement bruitées ou texturées.

Les exemples des figures (3.3.1.1) permettent de juger les résultats obtenus à partir d'histogrammes locaux calculés sur des aires de 16 x 16 pixels. L'aspect géométrique et topologique du caractère est conservé, mais les forts gradients, existant dans le tracé manuscrit et dans la texture, entraînent des seuillages trop bas dans les zones riches en informations. Ceci permet l'apparition des textures et du bruit.

Autre point défavorable: la réalisation difficile de ce schéma en temps réel. Pour que cette tâche de binarisation respecte les objectifs de temps d'exécution, il est nécessaire que les opérations faites sur les pixels de l'image soient mises en parallèle. Une carte électronique dédiée conservant une portion de l'image sous forme de zones se recouvrant, et calculant pour chacune d'elles des histogrammes locaux, demande une logique extrêmement complexe et difficile à réaliser en temps réel.

Boy
more
52736

856

01

1
6

Boy

856

more

01

52736

Figure 33.1.1 Seuillages d'histogrammes locaux.

farm
1750 Route
STATE
NJ 07470

farm
1750 Route
STATE
NJ 07470

Figure 3.3.1.1 Seuillages d'histogrammes locaux.

3.3.2 - TEST D'UNE BINARISATION PAR SEUILLAGE DYNAMIQUE.

Dans cette approche, le seuillage est calculé à partir des résultats d'un filtre passe-bande qui rehausse les variations spatiales des teintes dans l'image. Ce filtrage nécessite un opérateur de grande taille proportionnel à l'épaisseur du caractère. Les systèmes de traitement d'images classiques tel que le VICOM ne disposent que d'opérateurs 3 x 3 cablés. Les itérations de ces masques de petites tailles associés à une opération de bouchage de trous dans le caractère permettent une approximation de ce filtrage convolutionnel.

Dans l'exemple présenté, le filtre approché est du type "**chapeau mexicain**". A partir de cette image filtrée, deux images seuillées sont produites :

- Sur la figure (3.3.2.1), un **seuillage haut** fait apparaître sur l'image A des points issus de caractères, mais pas en totalité.
- Sur la figure (3.3.2.2), un **seuillage bas** fait au contraire apparaître sur l'image B tous les points issus de caractères mais aussi ceux de la texture ou du bruit.

A partir de l'image sous-informée A où théoriquement tout point noir appartient à un caractère, des croissances dirigées sont effectuées. Le cumul de ces opérations permet d'obtenir une image dilatée.

- Sur la figure (3.3.2.3), l'image binaire C finale ne contient que les pixels communs à l'image A dilatée et à l'image B sur-informée obtenue par seuillage bas.

Cette démarche est intéressante et facilement réalisable en temps réel, puisque les opérations sont effectuées directement sur l'image binaire et les voisinages de travail sont de quelques pixels de large.

- Sur la figure (3.3.2.4), si l'homogénéité est réalisée sur les caractères produits par les imprimantes à aiguilles, l'inconvénient de cette approche demeure dans l'apparition de connections indésirables entre les formes et les textures.
- Sur la figure (3.3.2.5) Certaines zones de caractères binaires sont rognées lorsque l'encrage des images originales est faible.

Box 856

Figure 3.3.2.1 Seuillage haut (A)

Box 856

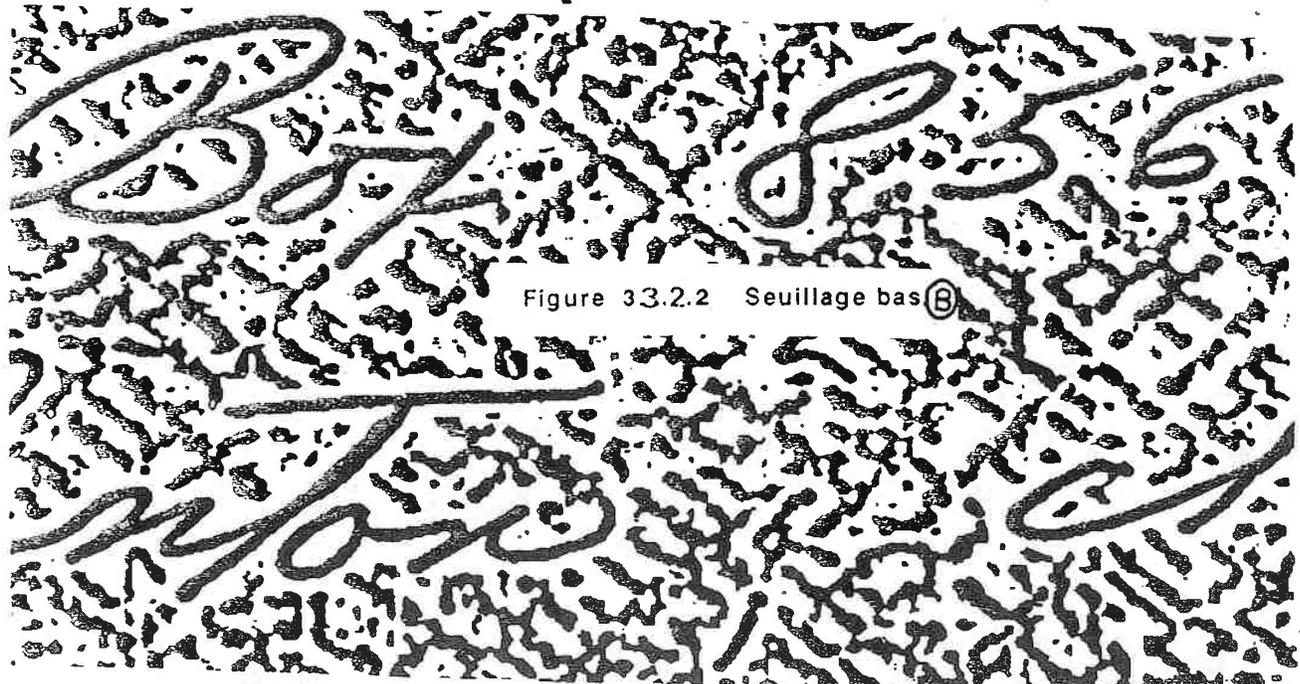


Figure 3.3.2.2 Seuillage bas (B)

Box 856

Figure 3.3.2.3 Image binaire resultant du seuillage dynamique. (C)

Box 856

MR CHARRAS
 H. B. S.
 64, RUE PLEYE
 93200 SAINT-DENIS
 REISCH Edouard
 Bat 13, esc 9
 Cité Les Francs Maitres
 93200 SAINT DENIS

Figure 3.3 2.4 Connexion des caracteres lors de la croissance des pixels.

YNBA WELK
 Shear Dr
 Sullivan Key
 Harold

Figure 33 2.5 Deterioration des images de faibles contrastes contenant des défauts d'encrage.

3.3.3 - CONCLUSIONS DE NOS INVESTIGATIONS.

En résumé, pour obtenir de bons résultats avec les techniques utilisant les histogrammes, il est nécessaire de travailler sur de petites zones de l'image et d'interpoler les résultats de chaque zone pour décider de l'appartenance du pixel au fond ou aux formes.

La puissance de calcul nécessaire pour respecter les contraintes d'un développement temps réel, sont des obstacles au choix d'une telle méthode.

Par contre, l'approche dynamique, économique en électronique et rapide en temps, est prometteuse pour les formes bien distantes.

La conclusion de ces essais tend à prouver que la démarche idéale doit être capable de déterminer si une faible variation de teinte est un défaut d'encrage ou une limite de forme, quel que soit le contraste et le spectre des niveaux de gris. De plus, ce processus doit différencier les caractères des textures qui modifient très souvent la topologie des formes.

3.4 - ALGORITHME SELECTIONNE.

Comme nous pouvons le constater sur la figure (3.4.1), l' algorithme de binarisation du prototype SYBILIS, réalisé pour le projet de lecture du code postal français, a des performances plus satisfaisantes que celles obtenues par les méthodes précédentes. Une adaptation de l' algorithme est néanmoins nécessaire car il s' agit de l' analyse de deux types de courriers différents. Outre les différences de graphismes, le courrier français est saisi à raison de 8 points/mm tandis que le courrier américain l' est à raison de 12 points/mm.

Le croquis de la figure (3.4.2), représentant une image multi-niveaux de gris, permet d' aborder les concepts de plateaux et vallées, utilisés en **traitement d' images**. Sur l' image multi-niveaux de gris les caractères sombres sont généralement représentés par des lignes de creux ou des vallées. On les appellera des **minimums régionaux**. Les talus sont les limites de formes et les plateaux le support clair. Le module de binarisation effectuée pour chaque point de l' image, un **seuillage local** présenté sous la forme d' un **test de binarisation**. Un pixel est considéré comme élément d' une forme connexe, si 2 de ces 8 voisins distants d' un pas de n-pixels sont plus clairs que lui.

Ce test est calculé à partir d' une image de gradients et de l' image originale. Pour calculer les valeurs de cette image, la détection des variations de teintes de gris produites par le passage (talus) du support (plateau) à une forme (vallée ou fosse) est effectué dans un voisinage 15 x 15 pixels. Ce noyau d' investigation, proportionnel à l' information recherchée, a la particularité de considérer la majorité des points du caractère comme points de transitions. Le test de binarisation se réfère à ces zones de transitions, appelés **écarts locaux**, pour étiqueter comme forme les pixels sombres de l' image originale placés dans ces zones. Ceci permet de réduire au maximum les graphiques épais.

Sur le plan visuel, une zone homogène, c' est à dire sans variation de teinte significative, est considérée comme faisant partie du fond. Un caractère écrit au marqueur, comme ceux présentés à la figure (3.4.3), se retrouve troué car l' épaisseur des traits est supérieure au voisinage d' étude. Ce problème est similaire au choix de la taille du noyau d' un filtre convolutionnel.



ecarts locaux
image de gradients

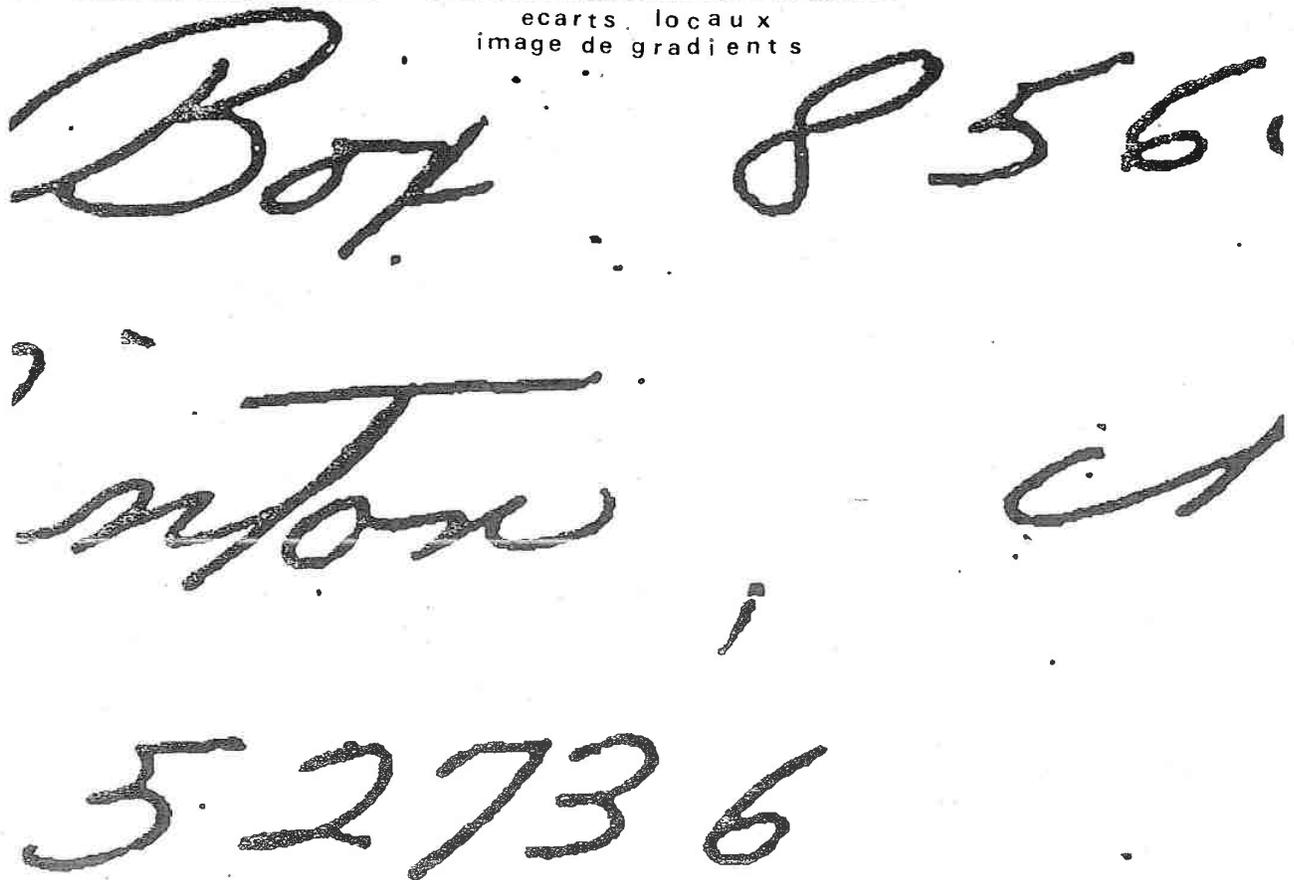


Figure 3.4.1 Resultat de la binarisation. image binaire.

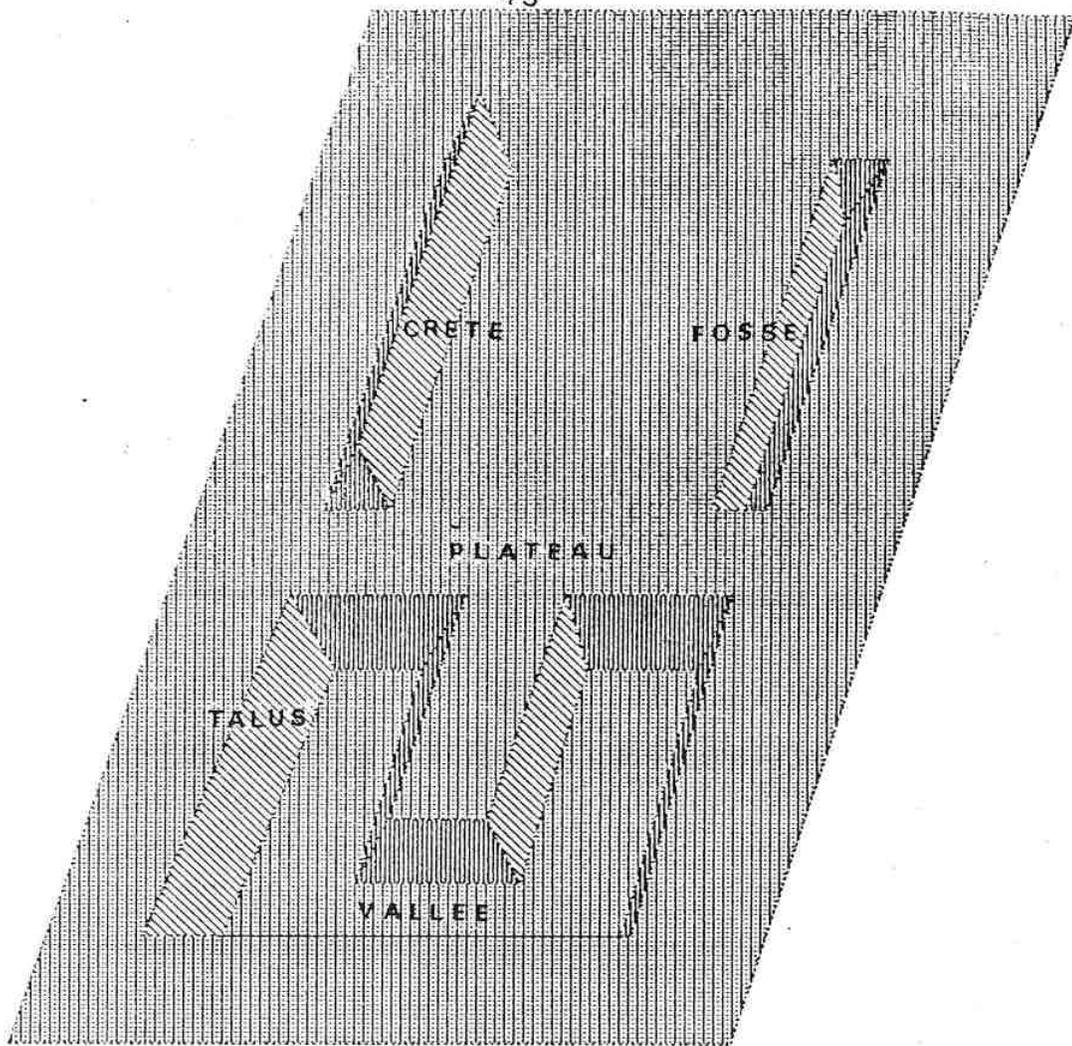


Figure 3.42 Schéma d'une représentation 3-D d'une image multi-niveaux de gris

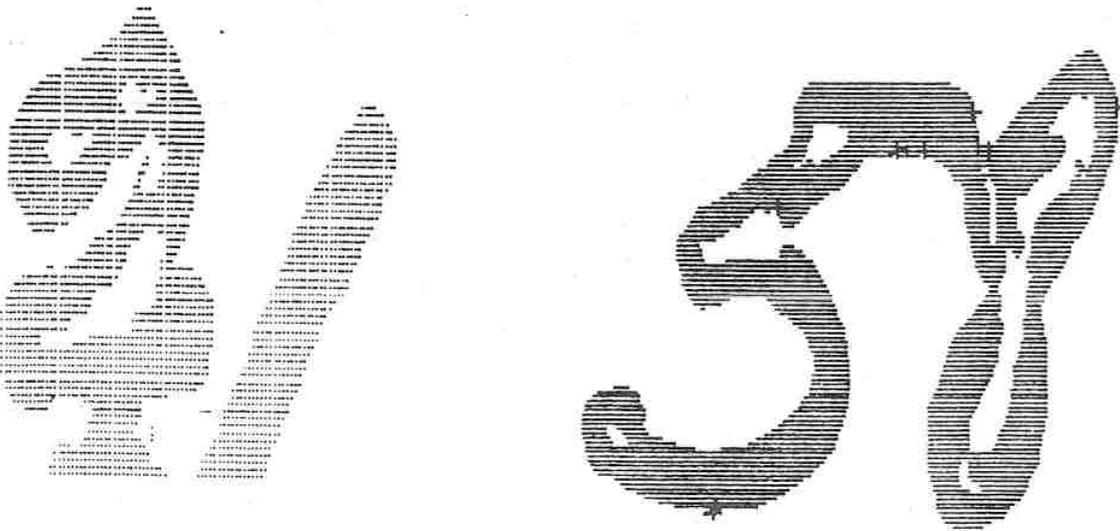


Figure 3.43 Creusement des caracteres epais au cours de la binarisation.

L'image binaire issue de ce test conserve des pixels parasites isolés ou collés sur le périmètre des formes. Certains caractères binaires présentent des trous ou des excroissances de quelques pixels. Pour rattraper ces défauts mineurs, le programme effectue un **lissage** de l'image. Ce type d'opération s'apparente aux opérations d'ouvertures et des fermetures de la morphologie mathématique.

3.4.1 - SEUILLAGE PAR TEST DE BINARISATION.

La fonction de test détermine si le pixel appartient à un minimum régional. L'examen a lieu dans huit directions (directions de la connexité-8). La zone d'exploration est à une distance précise du pixel central. Ce pas est fonction de l'épaisseur des traits que l'on désire faire ressortir. Pour le courrier français saisi à 8 points/mm, le pas est de 4 pixels. Pour le courrier américain saisi à 12 points/mm le pas est de 6 pixels.

```

x x x . . . x x x . . . x x x
x 4 x . . . x 3 x . . . x 2 x
x x x . . . x x x . . . x x x
. . . . .
. . . . .
. . . . .
x x x . . . x x x . . . x x x
x 5 x . . . x P x . . . x 1 x
x x x . . . x x x . . . x x x
. . . . .
. . . . .
. . . . .
x x x . . . x x x . . . x x x
x 6 x . . . x 7 x . . . x 8 x
x x x . . . x x x . . . x x x

```

Le pixel est déclaré objet si le test est positif pour au moins deux de ses voisins. Le test de binarisation est formulé de la manière suivante :

$$K(x,y,c) * L(x,y) < O(x_dir, y_dir)$$

c	valeur du gradient en (x,y)
x,y	coordonnées du pixel central.
x_dir, y_dir	coordonnées du pixel issues d'une des 8 directions et distant d'un pas.
K(x,y,c)	coefficient dynamique de contraste.
L(x,y)	valeur du niveau de gris du pixel central issu de l'image originale filtrée par un passe-bas.
O(x_dir,y_dir)	valeur du niveau de gris d'un pixel voisin issu de l'image originale

L'image originale est légèrement lissée afin d'atténuer les pics locaux et connecter les points des caractères provenant d'imprimantes à aiguilles.

Le processus de binarisation est schématisé sur la figure (3.4.1.1). La condition mise sur le nombre de tests positifs permet d'éviter le rattachement de pixels du fond à une forme ou la connexion de deux formes entre elles. Celle-ci évite partiellement le creusement des zones épaisses du caractère. Si le trait du caractère a une épaisseur nettement supérieure au voisinage d'étude, le contraste représenté par les écarts locaux est minime. Ceci implique aussi que les points utilisés pour le test de binarisation ont des niveaux de gris trop proches pour qu'une différenciation soit possible. Le pixel est alors considéré comme élément du fond.

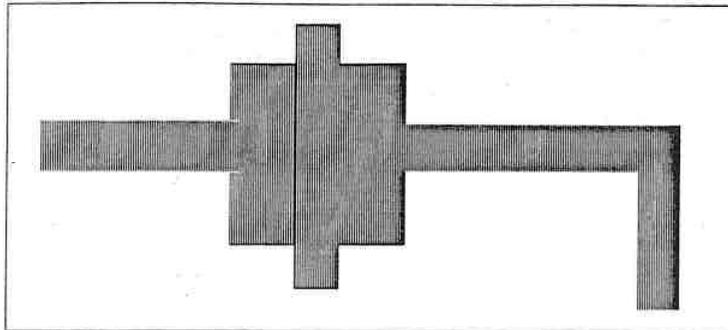
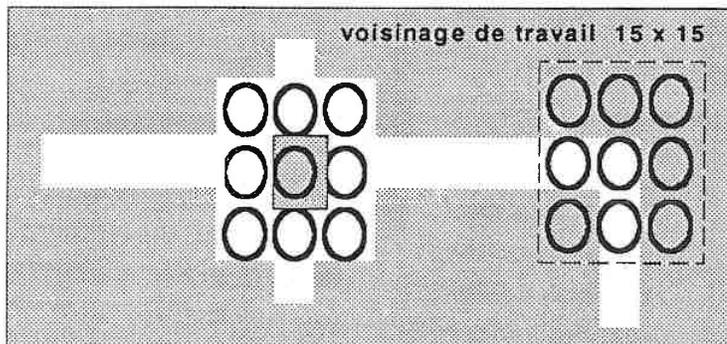


IMAGE ORIGINALE

IMAGE DES ECARTS
LOCAUX (GRADIENTS)

SI LA SURFACE DE LA FORME DEPASSE LE VOISINAGE D'ETUDE
LES PIXELS DE CETTE ZONE HOMOGENE NE SONT PAS CONSIDERES
COMME MINIMUMS REGIONAUX

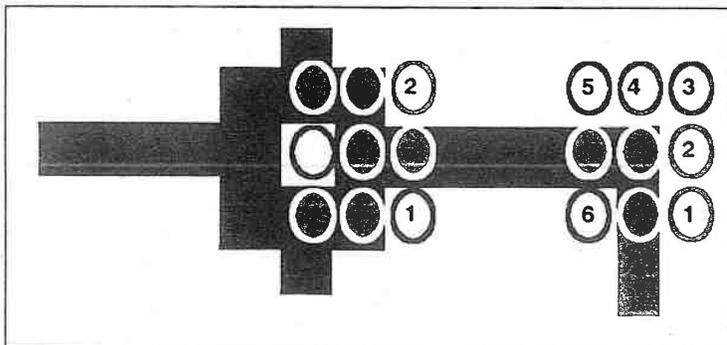


IMAGE BINAIRE

SI LE TEST DE BINARISATION EST POSITIF 2 FOIS, ALORS LE PIXEL EST "OBJET".

Figure 3.4.1.1 Voisinage de détection des minimumset
voisinage de décision pour le test de binarisation.

Les faibles variations de teintes de gris ne doivent pas être prises en compte dans une zone de contraste élevé. A l'inverse, celles-ci deviennent intéressantes dans les zones homogènes. C'est pour s'adapter au relief de l'image multi-niveaux de gris, que le **coefficient de contraste K** est une fonction de c , où c représente les valeurs résultantes de la convolution de l'image originale par un opérateur de type gradient. Cette adaptation atténue les problèmes de fond bruité et de mauvais éclairage. En effet un éclairage hétérogène produit des spectres réduits et variables en dynamique selon les zones de l'image.

Cette courbe heuristique de K tend vers une fonction logarithmique présentée sur la figure (3.4.1.2), où seules les valeurs comprises entre deux bornes produisent de fortes variations de la fonction. Les performances de la binarisation dépendent principalement des valeurs comprises entre c' et c'' , car cet intervalle est le reflet de l'instabilité des teintes dans les formes.

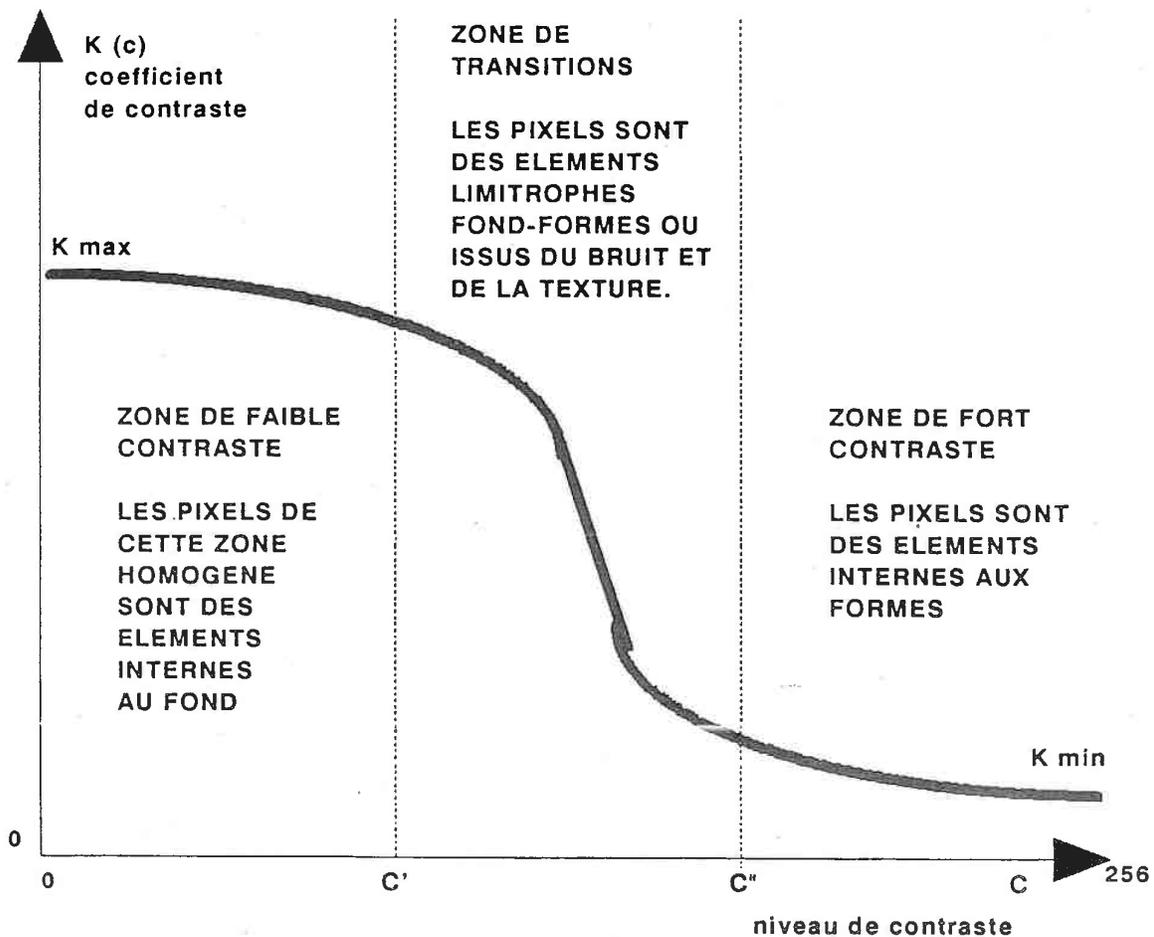


Figure 3. Courbe heuristique permettant de calculer le coefficient de contraste utilisé lors du test de binarisation.

3.4.2 - "LISSAGE" DE L'IMAGE BINAIRE BRUTE.

Pour être conforme à l'image originale, l'image binaire est codée de façon à faire apparaître les formes en noir sur un fond blanc. La binarisation produit une image contenant des amas de pixels connexes inutiles. Ceux-ci appartiennent au fond et aux formes. Pour épurer l'image et les contours de formes de ces éléments superflus, des opérations de morphologie mathématique sont effectuées sur l'image binaire. On peut formuler celles-ci de la manière suivante :

- Soit A l'ensemble des points de l'espace de l'image binaire.
- Soit B , un autre ensemble de l'espace, indépendant de A , que nous appellerons "élément structurant".
- Déplaçons B par translation pour le centrer au point x , et notons B_x , le nouvel ensemble translaté.
- Si B_x est inclus dans A , le point x est affecté de la valeur 1; la valeur est 0 dans le cas contraire.

Le lissage de cette image altérée, se fait par itérations de deux types de traitements séquentiels de ces opérations de "tout ou rien" que nous appellerons:

- le **nettoyage**,
- et le **bouchage**.

Les éléments structurants utilisés sont représentés par des masques de maille 3×3 qui correspondent aux configurations recherchées. Les masques présentés sont affectés à l'étude d'une image binaire où les formes sont codées 0 et le fond 1. Pour la description des masques nous utiliserons les conventions suivantes :

- 1 est un point du fond,
- 0 est un point de la forme,
- \bar{x} est un point quelconque.

3.4.2.1 - Le nettoyage.

Le **nettoyage** permet l'élimination des petites taches parasites et des excroissances du contour produites par les textures et le bruit. Cette suppression de points se fait en respectant la connexité 8 pixels voisins de la forme.

- i) Elimination des points **objets** isolés ou en extrémité de lignes.

Rotation du masque dans les 8 directions.

```

1 1 1
X 0 1
1 1 1

```

- ii) Elimination des points **objets** formant des angles droits ou des excroissances du contour. Rotation du masque dans les 8 directions.

```

X 1 1
0 0 1
X 1 1

```

- iii) Elimination des points **objets** formant des angles. La suppression de ce point superflus ne rompt pas la connexité. Rotation du masque dans les 4 directions (90°).

```

1 1 1
X 0 1
1 X 1

```

3.4.2.2 - Le Bouchage.

- Le **bouchage** égalise les contours et opture les trous internes aux caractères en rajoutant des pixels noirs. Cet ajout de points se fait en respectant la connexité 4 pixels voisins du fond.

i) Elimination des points du **fond** correspondant à un trou d'un pixel dans le caractère.

```

0 0 0
0 1 0
0 0 0

```

ii) Elimination des points du fond inclus dans le contour, inclus à l'intérieur de la forme, et les creux d'un pixel dans les zones de rebroussements. Rotation des masques dans les 8 directions.

```

X 0 0
1 1 0
X 0 0

```

3.4.2.3 - Exemples de formes lissées et cablage des opérations.

La réalisation pratique de ces opérations est obtenue par l'utilisation d'une table de correspondance adressable par 9 bits représentant les valeurs des pixels (0 ou 1) du voisinage 3 x 3. Le schéma de la figure (3.4.2.1) montre un exemple de plan synoptique de fonctionnement. Ce type de configuration existe sur les cartes de traitement d'image.

Sur la figure (3.4.2.2), nous pouvons constater l'intérêt de ces opérations. De nombreux tests doivent être effectués pour estimer le nombre d'itérations de **nettoyage** et de **bouchage** minimum et suffisant.

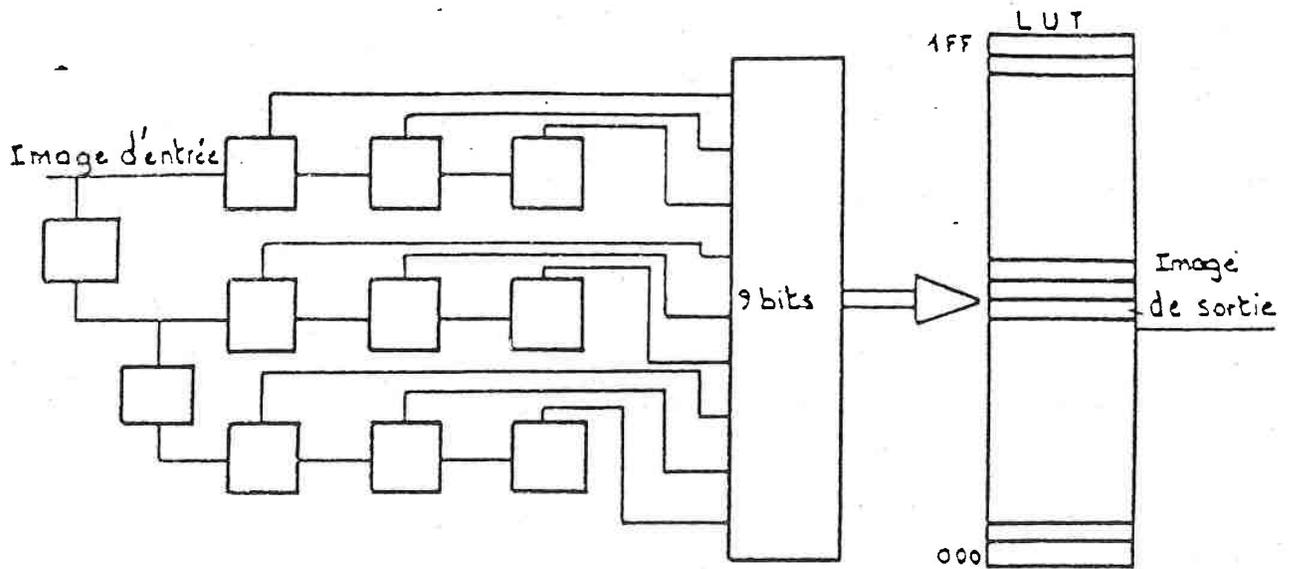


Schéma de principe d'un montage réalisant une opération "Tout ou Rien"

L'adressage de la LUT dépend directement des valeurs des pixels du noyau 3 x 3 constituant l'élément structurant. La mémoire de la LUT est chargée de la façon suivante :

- En entrée pixel central à 1 :
 0 en sortie pour un noyau correspondant aux configurations du nettoyage
 1 en sortie pour un noyau différent
- En entrée pixel central à 0 :
 1 en sortie pour un noyau correspondant aux configurations de bouchage
 0 en sortie pour un noyau différent

Les pixels du masque 3 x 3 sont numérotés de la façon suivante :

8	7	6
5	4	3
2	1	0

Figure 34.2.1 Synoptique de fonctionnement des opérations de lissage.

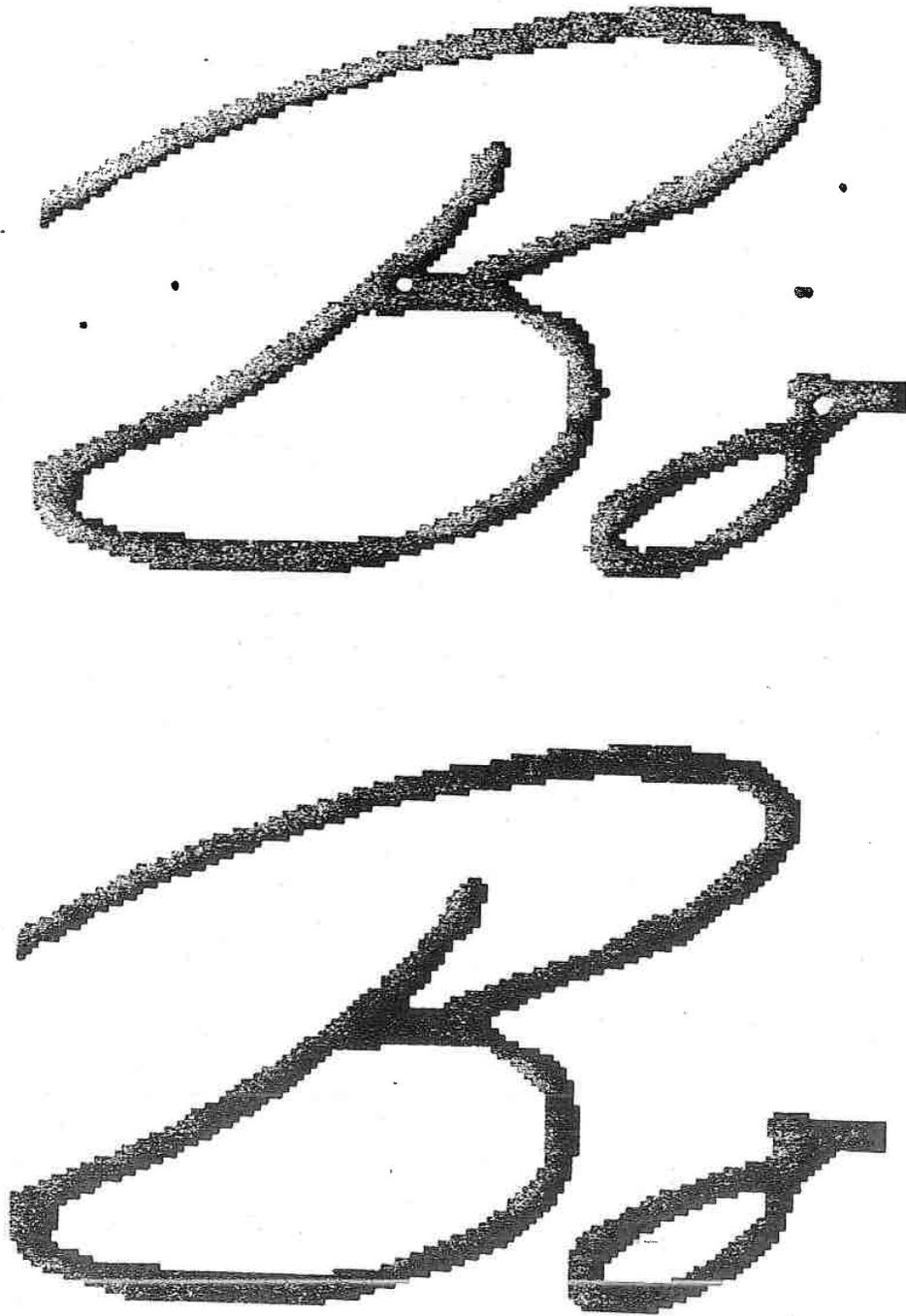


Figure 3.4.2.2 Exemple de lissage de la forme binaire par nettoyage et bouchage.

3.5 - EVALUATION DES PERFORMANCES.

Le codage de cet algorithme a été fait sur le VICOM, appareil de traitement d'image en temps réel. Le programme principal fait appel à la librairie VICOM pour les calculs. Ce système effectue des opérations élémentaires sur l'image et produit des convolutions en 40 ms avec des opérateurs de noyau 3 x 3 (pixels). La carte supplémentaire de morphologie mathématique permet de procéder à plusieurs opérations simultanément. La binarisation des trois plans 512 x 512 pixels de la lettre ne prend que quelques dizaines de secondes.

- POINTS POSITIFS.

- Récupération des images à faible contraste causé par :
 - les imprimantes à tissu,
 - les adressographes,
 - le mauvais éclairage.
- La topologie du caractère reste conforme à l'image originale.
- La connexion des points de caractères issus d'imprimantes à aiguilles est possible.
- Pas de connexion entre des informations d'origines différentes :
 - Ecriture manuscrite ou dactylographiée sur fond texturé ou bruité.
 - Ecriture manuscrite sur graphisme de fond à faible niveau de gris.
 - La séparation des caractères très proches est réalisée.
- L'élimination des trames et des parasites dans le fond permet au système de travailler sur une image propre ne contenant que des informations pertinentes.
- Les contours sont homogènes.

- POINTS NEGATIFS.

L'algorithme est étudié pour binariser correctement des formes d'une certaine épaisseur. Les traits au marqueur ou les zones d'épaississement du caractère génèrent des formes trouées. Ces trous se répercutent malheureusement sur le squelette en formant des bulles. Ces amas parasites détériorent les primitives de la forme et la rendent méconnaissable comme nous pouvons le voir sur la figure (3.5.1).

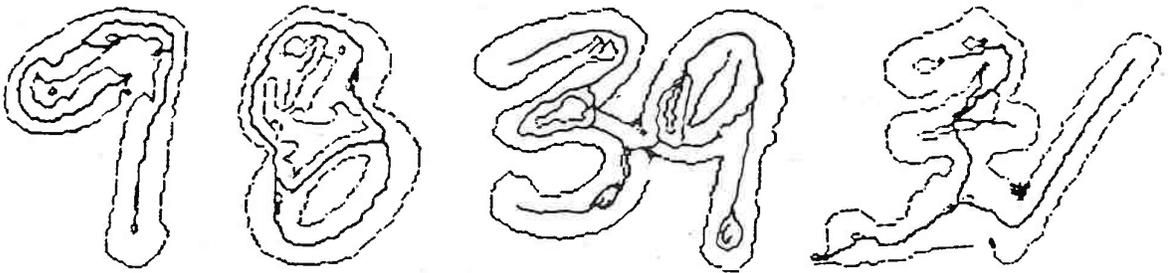


Figure 3.5.1 Repercution des trous du caractere sur le squelette.

Certains **levers de crayons** ou des défaillances de l'encrage produisent des zones de faibles teintes de gris au sein du caractère. La binarisation sectionne alors le caractère en plusieurs morceaux. Ce problème est symptomatique des zones anguleuses des chiffres 1, 4, 5 et 9. Quelques types de coupures sont représentés sur la figure (3.5.2).

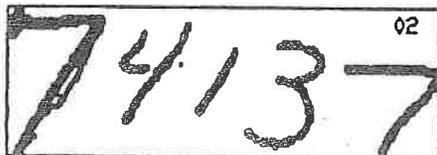


Figure 3.5.2 Caracteres brises par la binarisation.

4 - SQUELETTE ETIQUETE.

4.1 - INTRODUCTION.

Lors d'une identification de caractère, l'extraction des primitives pertinentes est souvent effectuée sur le **squelette** de la forme. Ce mode de représentation de la forme connexe est obtenu par un **amincissement** maximum du caractère. L'**amincissement** correspond à la contraction d'une forme par élimination des points du contour. Si la surface est diminuée au maximum, seuls des allongements de pixels subsistent. Ces lignes unitaires correspondent au squelette, représentation minimale et filiforme de l'objet.

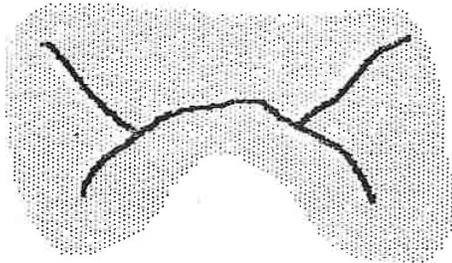
La "**squelettisation**" d'une forme, action d'amincir la forme jusqu'à l'obtention du squelette, peut être réalisée à partir de deux types de démarches.

Pour obtenir un processus réversible, certaines méthodes, comme l'**érosion topologique**, permettent de ronger la forme de manière homogène. Cette diminution se fait à vitesse constante et quelle que soit la direction. Une image souvent employée pour exprimer cette progression est le mouvement d'un feu encerclant une forêt ("grass-fire"). Deux fronts de flammes opposés se rejoignent en des points équidistants aux lisières. Cet amincissement isotrope donne à chaque point la propriété d'équidistance des bords. Le squelette issu est souvent complexe car chaque imperfection de la forme ou chaque défaut du contour se répercute fortement sur celui-ci.

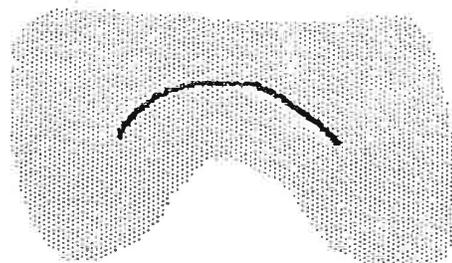
Les autres méthodes, que nous appellerons "**amincissements dirigés**" ("shrinking algorithms"), ne sont pas toujours réversibles mais permettent d'obtenir des squelettes plus représentatifs. Ces algorithmes sont insensibles aux détails et s'attachent plutôt à conserver les informations topologiques des formes, telles que les échelles, les courbes et les orientations.

Le principal défaut de l'amincissement est la perte de l'information "**épaisseur du caractère**". Pour combler cette lacune le squelette est étiqueté, ce qui avec l'image du contour permet d'apprécier la surface et l'étendue du caractère.

L'exemple de la figure (4.1.1) schématise le comportement de ces deux types de processus.



Squelette obtenu par un amincissement homogène et maximum.



Squelette obtenu par un amincissement dirigé et maximum.

4.2 - INVESTIGATIONS.

4.2.1 - AMINCISSEMENT HOMOGENE.

La première nécessité pour un algorithme d'**amincissement** est de ne pas briser la connexité des formes. Les algorithmes considèrent très souvent qu'un pixel issu d'une forme peut être connexe à 8 pixels voisins. Un squelette connexe-8 contient un minimum de points, ce qui facilite le stockage et la description de celui-ci sous forme de graphe.

L'**amincissement homogène** conduit à la formation d'un squelette où tous les points sont équidistants du bord. On définit en morphologie mathématique, le squelette comme l'ensemble des centres de disques maximaux (rayons maximums) inclus dans l'ensemble E des points de la forme connexe. Cette démarche d'amincissement permet de considérer tout détail comme instructif, quelle que soit sa taille. Ces détails peuvent être pertinents dans le cas de formes complexes. En ce qui concerne les caractères, les formes sont grossières et les barbelures qui apparaissent sur les squelettes sont des artéfacts ou du bruit.

La squelettisation d'une forme dans le plan euclidien a été mathématiquement formulée en premier par Th Motzkin (1935) puis H. Blum (1962), pour enfin être réactualisé par J. Serra (SER 82).

De nombreuses publications font état de la facilité d'utiliser, en traitement d'images, un **voisinage carré** autour du pixel central plutôt qu'un **voisinage hexagonal**. En effet, les huit pixels voisins ne sont pas équidistants du centre dans cette mosaïque orthogonale de pixels. Les algorithmes associent la connexité-8 à l'**objet** et définissent le **fond** selon la connexité-4.

L'algorithme de Serra élimine les points superflus qui correspondent aux configurations reproduites ci-dessous. Ces structures 3 x 3 subissent des rotations de 90°.

```

0 0 0   . 0 0
. 1 .   1 1 0
1 1 1   . 1 .

```

4.2.2 - AMINCISSEMENT DIRIGE.

Les différents algorithmes présentés dans cette partie tendent d'élaborer un squelette le plus instructif sans pour cela briser la connexité de la forme ou considérer les imperfections du contour.

- N. Naccache et Shinghal (NAC 84) en utilisant l'approche de C.J. Hilditch aboutissent à l'algorithme de squelettisation suivant :

Définitions :

```

P5 P4 P3
P6 P1 P2   notations des pixels dans le voisinage d'étude.
P7 P8 P9

```

- 1) $B(P1)$ est le nombre de pixels noirs parmi les 8 pixels voisins.
- 2) Un point de contour est un point noir avec au moins un point blanc dans le voisinage défini par la connexité-4.
- 3) Un point final est un point noir avec au moins un point noir parmi ses 8 pixels voisins.
- 4) Le nombre de passages est défini par :

$$X(P1) = \sum b_i, i \text{ variant de } 2 \text{ à } 5$$
 où $b_i = 1$ si $P(2i-2)$ est blanc et $P(2i-1)$ ou $P(2i)$ est noir
 sinon $b_i = 0$.

L'algorithme amincit par itérations. Au cours de chaque phase de traitement, les colonnes de la forme sont balayées de gauche à droite et de haut en bas. Un point central noir $P1$ est marqué comme superflu si les six tests suivants $H1$ à $H6$ sont positifs :

- | | |
|--|---|
| $H1 : P2 + P6 + P7 + P8 \leq 3$ | $P1$ est un point du contour. |
| $H2 : B(P1) \geq 2$ | $P1$ n'est pas un point final. |
| $H3 : N(P1) \geq 1$ | $N(P1)$ est le nombre de pixel noirs non marqués se trouvant dans le voisinage défini par la connexité-8.
Ce test empêche d'éliminer les pixels des extrémités de ligne. |
| $H4 : X(P1) = 1$ | $P1$ est un point de brisure.
son élimination entraîne une rupture de la connexité de la forme. |
| $H5 : \text{soit } P4 \text{ n'est pas marqué}$
ou $X4(P1) = 1$ | $X4(P1)$ est le nombre de passages sur $P1$ si l'on considère temporairement que le pixel $P4$ est blanc.
Ce test empêche une érosion excessive. |
| $H6 : \text{soit } P6 \text{ n'est pas marqué}$
ou $X6(P1) = 1$ | Ce test similaire à $H5$ empêche aussi une érosion excessive. |

À la fin de chaque passage, les pixels marqués sont éliminés et l'algorithme recommence ce processus jusqu'à ce qu'il ne puisse plus effectuer de marquage.

- Une autre approche intéressante est celle développée par TH. Pavlidis (PAV 81). Cette algorithm, lui aussi itératif, associe à chaque pixel du squelette un numéro de 1 à 4. Les points essentiels qui seront conservés sont appelés "**multiples**" et codés 3. Les points qui ne répondent pas aux critères des "multiples" sont codés 2. Si ces points codés 2 ne répondent pas aux critères de rattrapage des "**pro-multiples**", ils demeurent à la valeur 2 et sont éliminés. Les "pro-multiples" codés 4 sont aussi éliminés s'ils ne deviennent pas "multiples" en répondant à d'autres critères.

- Bases de l'algorithme d'amincissement.

R est l'image des points noirs codé 1 de l'image originale.

Itérations des phases 1 à 4.

- 1) Construction de l'ensemble B(R) contenant tous les points de R qui sont connexes-4 avec un point noir externe à l'ensemble R.
- 2) Identification du domaine M (R) contenant les pixels "**multiples**" de l'ensemble B(R).
- 3) Si $B(R) = M(R)$ le programme s'arrête.
- 4) Remplacement de R par $R - [B(R) - M(R)]$.

Pour comprendre cette approche et les notions utilisées, il est nécessaire de poser les définitions suivantes:

- Le codage de l'image binaire associe la valeur 1 aux pixels des formes connexes, et la valeur 0 aux pixels du fond.
- un pixel **objet** codé 1 qui ne répond pas aux quatre tests suivants (i),ii) puis iii),iiii)) est codé 2 et sera éliminé.
- un pixel est étiqueté définitivement comme "**multiple**" (codé 3) s'il satisfait l'une de ces deux conditions :
 - i) Au plus un de ses huit voisins est différent de 0.
 - ii) Son voisinage doit se conformer aux configurations décrites par les structures suivantes qui subissent des rotations de 90°.

A A A

0 P 0 Au moins un des trois pixels marqués de la lettre A doit être différent de 0.

B B B Au moins un des trois pixels marqués de la lettre B doit être différent de 0.

A A A

A P 0 Les pixels marqués de la lettre A peuvent avoir une valeur quelconque.

A 0 2

Le pixel codé 2 est rattrapé et étiqueté "**pro-multiple**" (codé 4) si l'une des deux conditions suivantes est satisfaite.

iii) Le pixel n'a pas de voisin étiqueté 1.

iii) Son voisinage doit se conformer aux configurations décrites par la structure suivante qui subit des rotations de 90°.

A A C

0 2 2 Soit l'un des pixels de chaque paire, marqués A, B, C, est différent de 0,

B B C soit les 2 pixels C sont non-nuls.

Ces diverses conditions peuvent être vérifiées de manière séquentielle ou parallèle. La préservation des lignes d'épaisseur 2 permet par la suite de choisir une élimination orientée de ces pixels.

0 2 2 0

0 2 2 0

0 2 2 0

Il est nécessaire ensuite d'assigner une valeur finale aux pixels qui ont été étiquetés "**pro-multiples**" pour les conserver (codé 3) ou les rejeter. La configuration suivante montre le besoin d'utiliser un label additionnel avant d'éliminer les "**pro-multiples**". Cette suppression se fait si l'un des deux tests est positif.

1 1 2 0 0 0

2 2 4 0 0 0

0 0 4 4 3 3

0 0 4 4 0 0

2 2 1 2 0 0

exemple de codage

- 3 2 1 - Le voisin 4 est nul et le voisin 0 est codé 3 ou 4.
 4 P 0
 5 6 7 - Le voisin 2 est nul, le voisin 6 est codé 3 ou 4
 et il n'y a pas d'autre voisin qui réponde à la condition précédent.

- Dans les années 70, H. Tamura (TAM 78) compara de nombreux algorithmes d'amincissement. Sa conclusion est qu'un algorithme d'amincissement doit considérer certaines configurations de pixels dits finaux afin de ne pas les éliminer. Ces points essentiels correspondent aux configurations suivantes :

X 1 0	0 1 X	X Y X	X Y X
Y 1 1	1 1 Y	1 1 Y	Y 1 1
X Y X	X Y X	0 1 X	X 1 0

les pixels X sont indifférents et les pixels Y ne doivent pas tous être nuls.

- La figure (4.2.2.1) représente le codage qu'effectue le squelettiseur de P. Marthon, A. Bruel, G. Biguet (MAR) au cours des passes de traitement. Cette algorithme analyse la sommation des vecteurs définis entre un point étudié et les points de ses huit voisins connexes. La norme de la longueur de ce vecteur final est représentée par un nombre entier compris entre 0 et 4.

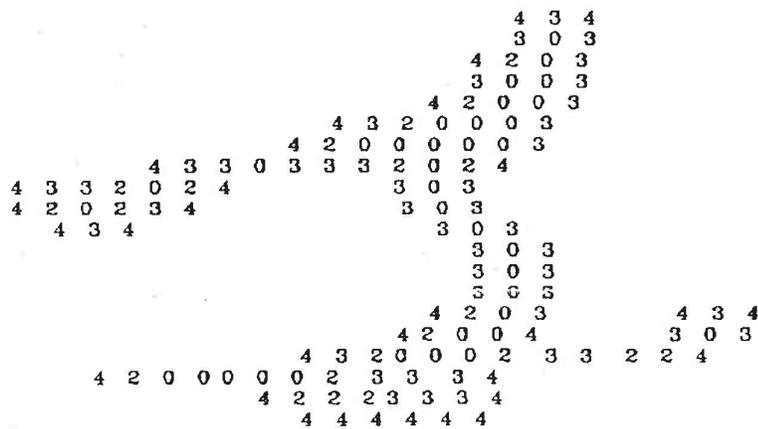


Figure 4.2.2.1 Calcul de la distance du pixel de la forme au squelette.

A priori les points se trouvant au sein de la forme doivent avoir un vecteur petit ou inexistant si la norme est approchée par la somme $|X| + |Y|$. Inversement, les points du contour associés à de fortes valeurs devraient être éliminés.

En fait, tous les pixels tels que $|X| + |Y| = 4$ sont éliminés et tous les points tels que $|X| + |Y| = 2$ sont conservés.

Pour les sommes égales à 3, problématiques, comme le montre la figure (4.2.2.2), il suffit de connaître le nombre de pixels voisins pour lever l'ambiguïté. Le point n'ayant que 3 voisins est éliminé.

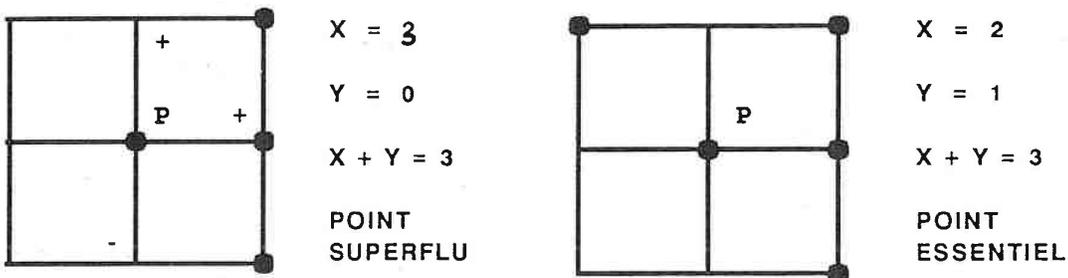


Figure 4.2.2.2 Configurations d'éliminations de points.

Malheureusement, Ce test n'est pas toujours suffisant. Il demeure deux exceptions représentées sur la figure (4.2.2.3), qui sont les lignes verticales ou horizontales d'épaisseur 2. Les familles de points A et B ont le même voisinage, mais il n'est possible d'éliminer dans le cas présent, qu'un des deux points A. Il est nécessaire pour trancher de rechercher le troisième élément.

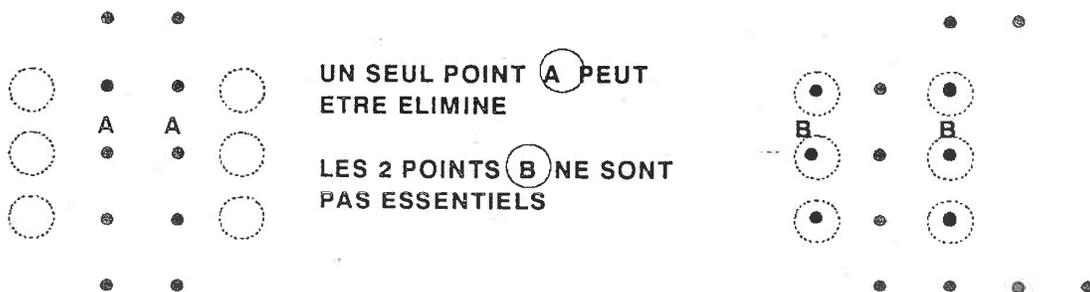


Figure 4.2.2.3 Elimination finale de points.

Cet algorithme conserve les propriétés de connexité de l'image :

- même nombre de composantes connexes entre la forme et son squelette pour une distance $D1(x,y) = \max [|x1-y1| - |x2-y2|]$
- même nombre de composantes connexes entre leurs formes complémentaires pour la distance $D2(x,y) = |x1-y1| + |x2-y2|$.

Cet algorithme est assez insensible au bruit. Les parasites du contour ne produisent pas d'artefacts sur le squelette.

Les techniques d'**amincissement** offrent, dans la plupart des cas, une bonne approximation du squelette de la forme, mais ont toujours des limites d'applications dans les zones de distorsions du caractère.

4.3 - METHODE D'AMINCISSEMENT SELECTIONNE.

A partir d'une forme complexe, un algorithme d'amincissement génère un squelette différent d'une méthode à l'autre. Certains algorithmes font ressortir des détails tels que les légères excroissances ou d'infimes brèches du contour en produisant des petits segments sur le squelette général. Dans le cadre d'analyse de structures ou d'objets complexes, l'étude de ces détails peut représenter l'unique but de la **squelettisation**.

Dans le cadre de la reconnaissance des caractères, ces barbelures sont très souvent parasites. De plus, elles sont dangereuses car tout segment de la forme est considéré comme significatif au niveau des tâches de **reconnaissance**.

L'expérience, acquise au cours du projet de lecture du code français, a permis de connaître toutes les configurations que produit l'algorithme de S. Thomé (THO 78). Notre choix s'est arrêté sur cette méthode, aux résultats très satisfaisants, qui amincit selon la **connexité 8 pixels voisins**.

L'algorithme développé érode la composante connexe en éliminant les points superflus, au cours de séquences de quatre étapes ordonnées.

La suppression des points commence par les configurations **nord** du contour, puis **sud**, **est** et **ouest**.

Les configurations recherchées sont détectées en utilisant quatre masques qui subissent des rotations de 90° en fonction de l'orientation d'élimination.

- Les conventions de codage des **masques** pour une image binaire contenant des formes noires sur un fond blanc sont :

- P indique le pixel central étudié ,
- 1 pixel blanc appartenant au fond ,
- 0 pixel noir appartenant à la forme ,
- X pixel sans condition.

```

X 1 X   1 1 1   X 1 1   1 1 1
0 P 0   1 P 1   0 P 1   1 P 1
X 0 X   X 0 0   X 0 X   0 0 X

```

Pour obtenir le résultat escompté, la recherche des configurations se fait selon un protocole précis. L'algorithme effectue la séquence des quatre amincissements suivants :

```

P4 P3 P2
P5 P0 P1
P6 P7 P8

```

- Etape 1 : domaine des pixels de bordure avec
P3 = 0.
- Etape 2 : domaine des pixels de bordure avec
P3 = 1, P7 = 0.
- Etape 3 : domaine des pixels de bordure avec
P1 = 0, P3 = 1, P7 = 1.
- Etape 4 : domaine des pixels de bordure avec
P1 = 1, P3 = 1, P5 = 0, P7 = 1.

Cette démarche d'amincissement dirigé diffère d'une érosion topologique homogène. Les extrémités de traits ne sont pas rognées et les distorsions du contour ne produisent pas des barbelures mais une courbure du squelette. Quelques exemples sont représentés sur la figure (4.3.1).

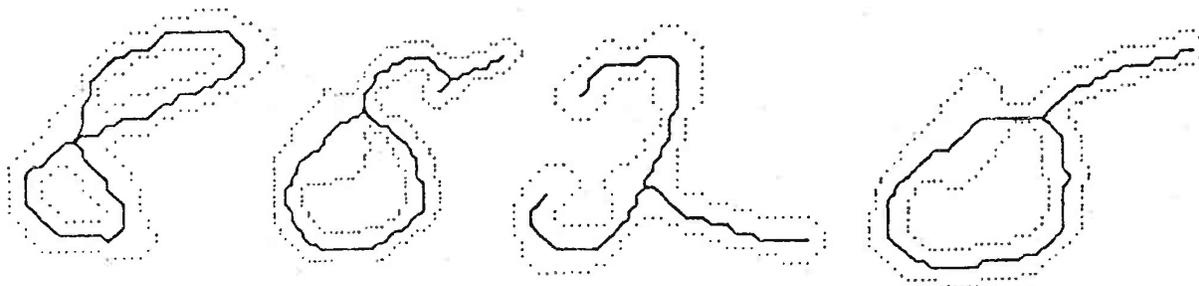


Figure 4.3.1 Conservation des caractéristiques géométriques de la forme.

4.4 - UTILISATION DE L'ÉTIQUETAGE ET DU CONTOUR.

Le problème majeur des représentations filiformes est la perte définitive de l'information concernant l'**épaisseur**, ce que nous pouvons remarquer sur les chiffres de la figure (4.4.1). Toute forme compacte voit son squelette différer du tout au tout. Dans les étapes de "**reconnaissance topologique**", une confrontation du squelette avec son contour est inévitable pour lever l'ambiguïté.

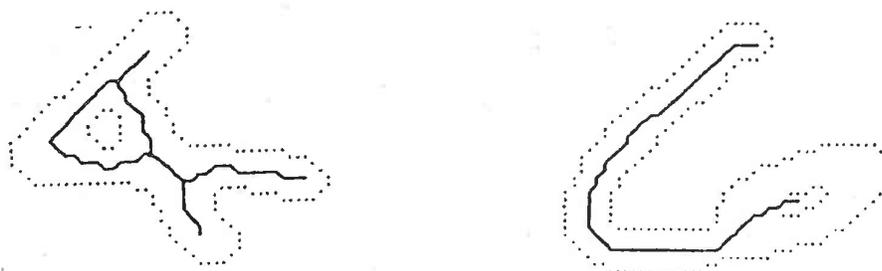


Figure 4.4.1 Exemples de caractères où l'information du squelette est insuffisante.

Pour améliorer la décision, l'étiquetage des pixels du squelette par le nombre d'itérations nécessaires pour l'obtenir, permet de caractériser certains noeuds complexes et les amas de boucles compactes.

Plusieurs causes sont à l'origine des aberrations du squelette :

- les artefacts de binarisation (trous dans le caractère, distorsions) ,
 - les artefacts de squelettisation (distorsion des croisements de traits ou les zones épaisses),
 - les dessins parasites,
 - les écritures riches en volutes (boucles abondantes).
- figure (4.4.2) : Effet de duplication de noeuds ("**necking effect**").

Cette distorsion représente la transformation d'un noeud à quatre branches en une configuration contenant un petit segment séparant deux noeuds à deux branches. La **binarisation** des lignes obliques produit un décalage du tracé au niveau des intersections de traits. Ce décalage est ensuite accentué par les itérations d'amincissement de la forme. L'**étiquetage** permet alors de savoir si certains segments courts sont authentiques ou parasites.

- figure (4.4.3) : Formation d'"**aigrette**" ("extra tail").

L'amincissement d'une zone de rebroussement du tracé produit un segment sur l'extrados de la courbure. Ce segment n'est pas essentiel et peut être caractérisé par la valeur élevée de son étiquetage.

- figure (4.4.4) : Problème des boucles pleines.

L'amincissement d'une boule finit par donner un point comme squelette. Un caractère contenant une boucle pleine voit sa boucle se transformer en un segment connecté au reste de la représentation filiforme. Le contour externe et l'étiquetage permet de reconnaître en ce segment une boucle fermée.

- Figure (4.4.5) : Elimination des boucles parasites.

Le problème majeur de la binarisation choisie, est la formation de trous dans les formes épaisses. Il suffit du trou d'un pixel pour créer une bulle non négligeable sur le squelette. Une démarche complexe permet de chasser du squelette ces amas de boucles parasites. Pour cela, il est nécessaire d'utiliser le contour, l'étiquetage au niveau des noeuds et les graphes de parcours qui peuvent être générés entre deux noeuds.

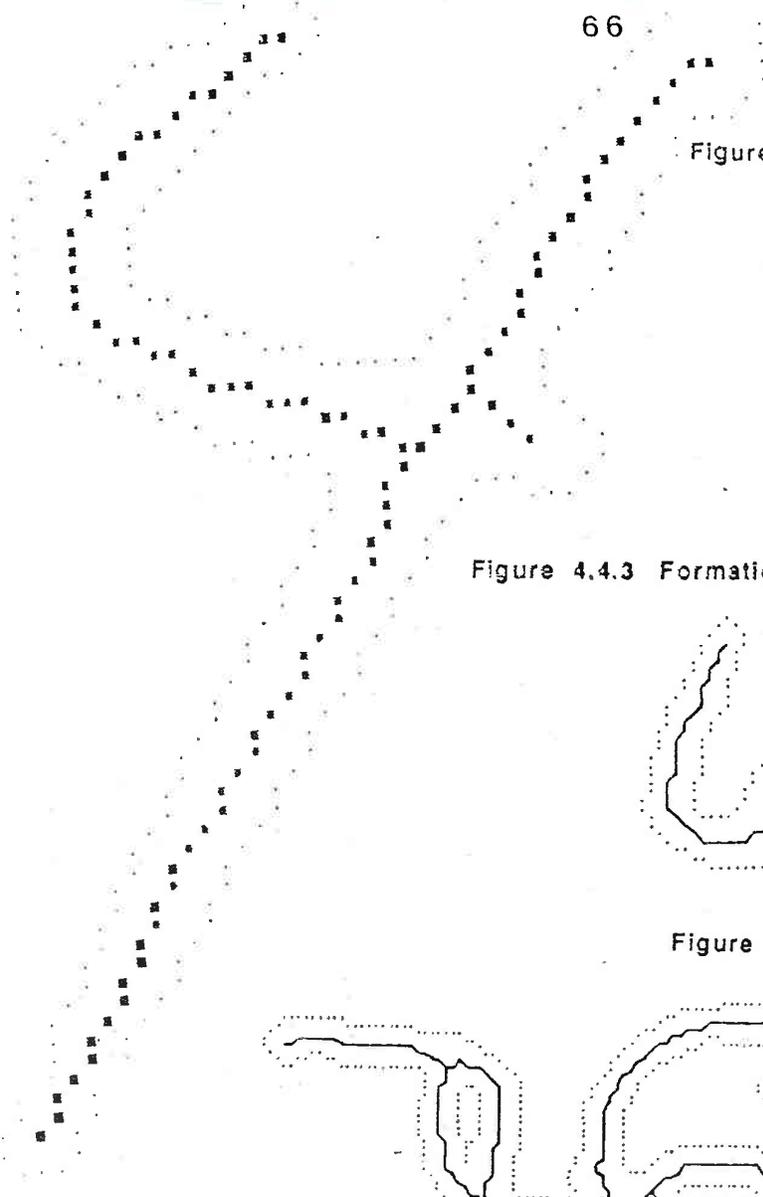


Figure 4.4.2 Effet de duplication de noeuds.

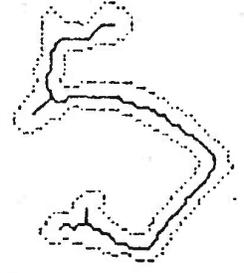


Figure 4.4.3 Formation d'aigrette parasite sur le caractere.

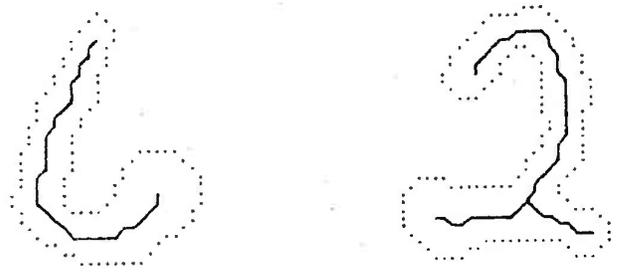


Figure 4.4.4 Caracteres a boucles pleines.

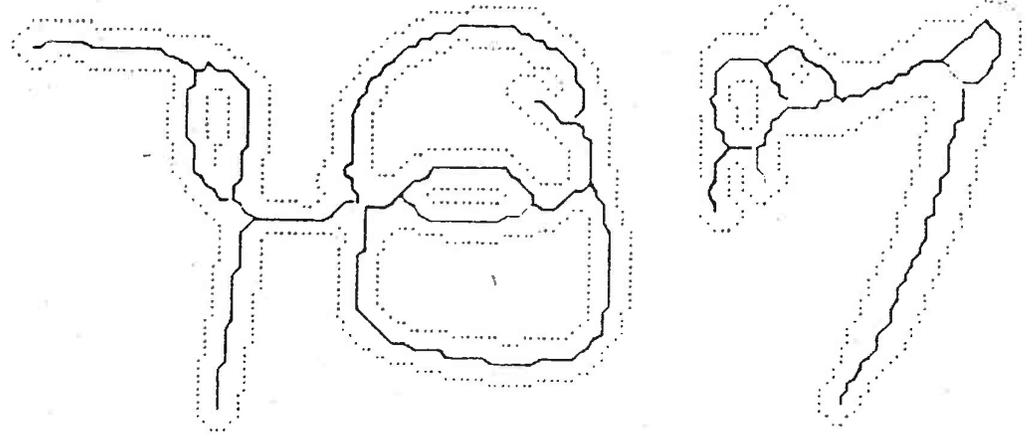
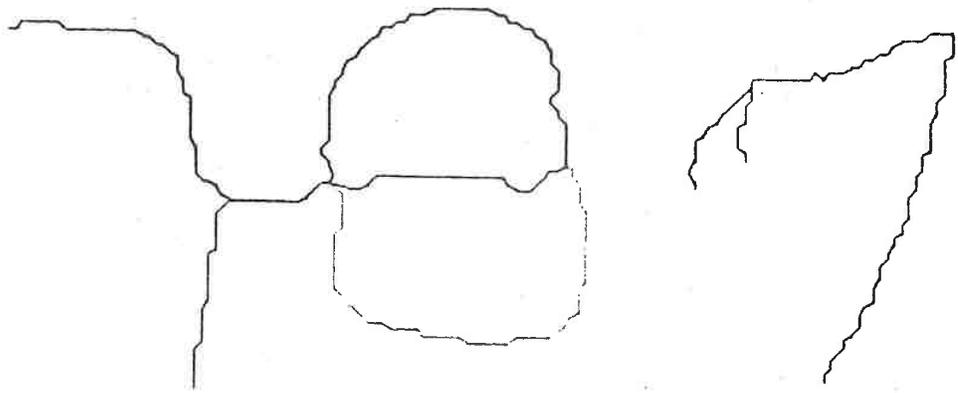


Figure 4.4.5 Elimination des bulles parasites du squelette.



5 - RECHERCHE DE L' INFORMATION CODE-ETAT-VILLE.

5.1 - INTRODUCTION.

Postulons qu'après la description des composantes connexes, le module d'identification des chiffres du code postal analyse la totalité de l'image. Ceci se traduit par une soumission obligatoire de l'ensemble des formes à l'**étage de reconnaissance**. Un certain nombre de sanctions vont étiqueter des caractères alphanumériques comme chiffres. Des candidats **codes** peuvent alors être générés à partir de ces formes. Au niveau du module de décision, le choix final provient de la comparaison des hypothèses codes formulées.

Actuellement les ordinateurs n'offrent pas assez de puissance de calcul à un coût raisonnable pour effectuer cette masse d'opérations. Le prototype "SYBILIS-PRISME" (CGA-HBS) effectue la lecture en temps réel car des opérations de pré-traitements diminuent considérablement le volume de données à traiter. Il est donc nécessaire de réduire au maximum la zone des recherches du code postal. Au niveau de la méthodologie choisie, nous avons préféré séparer les tâches de localisation et de reconnaissance, car nous savions qu'un classificateur de chiffres performant n'est pas "endurci" aux caractères alphabétiques. Explicitement, un module dédié uniquement à la reconnaissance des caractères numériques, ne possède pas les critères antagonistes lui permettant, d'une part de rattraper les formes marginales, et d'autre part de rejeter les lettres. Cette faiblesse permet à des successions de caractères alphabétiques, topologiquement proches de chiffres, d'être identifiées comme candidats **codes**. Si l'on pousse la démarche à vouloir détecter l'**état** et la **ville** (et à les identifier), les générations de candidats **état-ville** pour chaque candidat **code**, peuvent produire une explosion combinatoire. Il est donc préférable de suggérer à l'**étage de reconnaissance** les zones de l'image contenant uniquement le code postal, l'état et la ville. Pour arriver à cette **localisation**, de nombreuses tâches, dites de "bas niveaux", diminuent le volume d'information et tendent d'en extraire un ou plusieurs candidats **code-état-ville**.

5.2 - INVESTIGATIONS.

Dans les publications issues de nos recherches bibliographiques, la localisation des lignes et des blocs de caractères a été étudiée principalement sur des documents dactylographiés. Les démarches qui vont être présentées n'apparaissent pas très fiables sur le courrier américain pour les raisons suivantes :

- l'écriture manuscrite est rarement horizontale, surtout dans la zone de l'adresse,
- les hauteurs de lignes manuscrites sont variables, les hampes et les jambages des lettres pouvant provoquer l'entrelacement ou la connexion de plusieurs lignes,
- dans la ligne , l'espacement des lettres et des mots est variable, voire inexistant quand l'écriture est inclinée.

5.2.1 - TECHNIQUE DE LISSAGE PAR CROISSANCE DIRIGEE (" RUN LENGTH SMOOTHING ALGORITHM ").

Cette approche tend de former des lignes d'information en comparant des images binaires ayant subies des lissages directionnels. Ces lissages se font par dilatation des formes (croissance de pixels) dans le sens horizontal et vertical. La figure (5.2.1.1) montre un exemple de traitement d'un document imprimé.

La tentative faite sur les images américaines montre que cette méthode devient hasardeuse pour les caractères manuscrits car les blocs d'informations résultants connectent plusieurs lignes, comme nous pouvons le constater sur la figure (5.2.1.2).

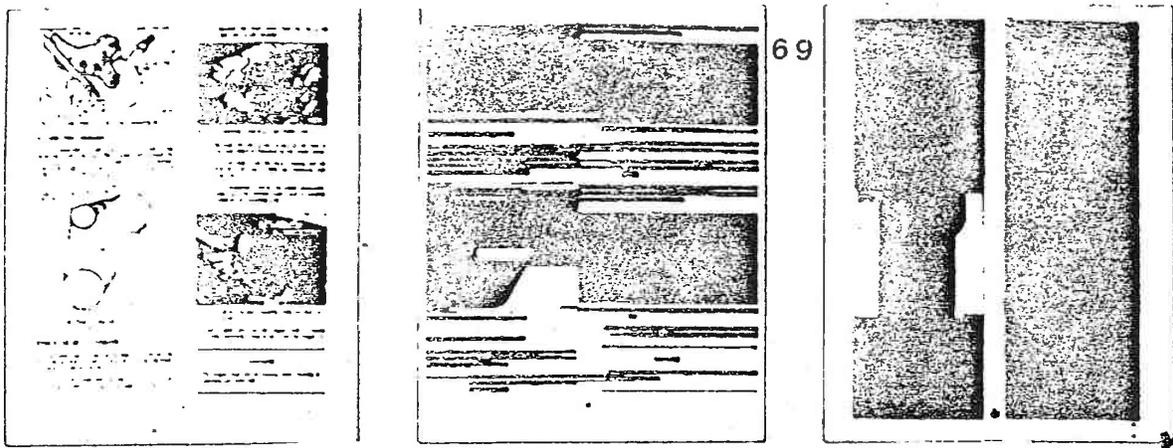
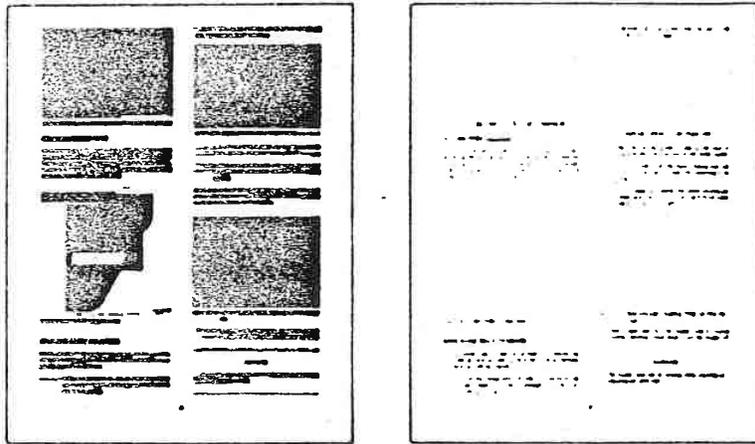


Figure 5.2.1.1 Technique de lissage par croissance dirigée.



Chang-Huei Wang,
 Dept. of Computer Science
 SUNY at Buffalo
 226 Bell Hall
 Buffalo, NY 14260

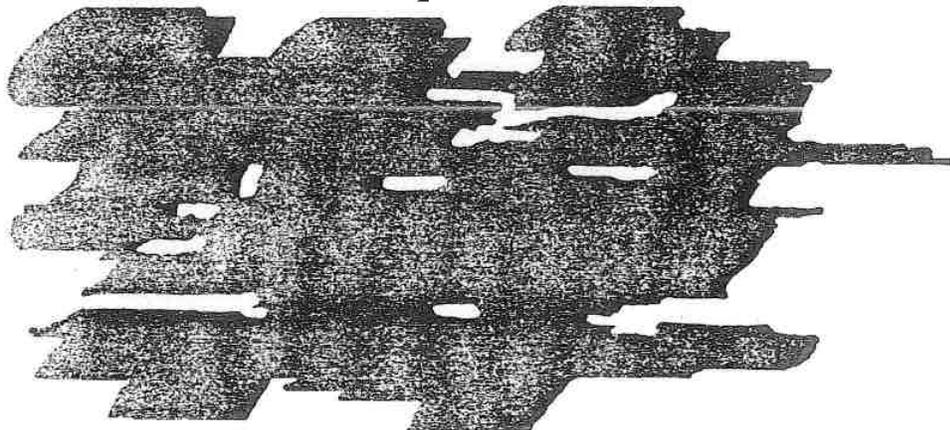
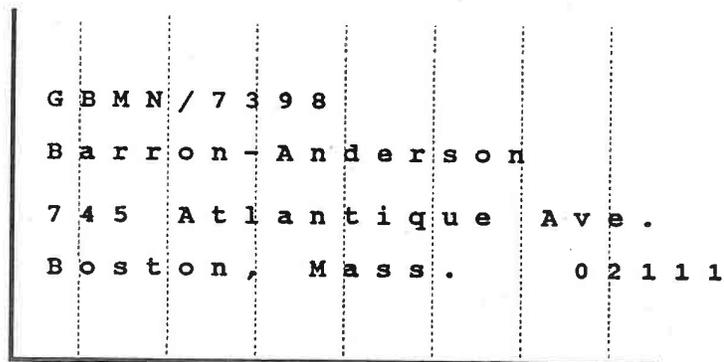


Figure 5.2.1.2 Exemple de croissance de régions sur une lettre américaine.

5.2.2 - TECHNIQUES UTILISANT LES HISTOGRAMMES.

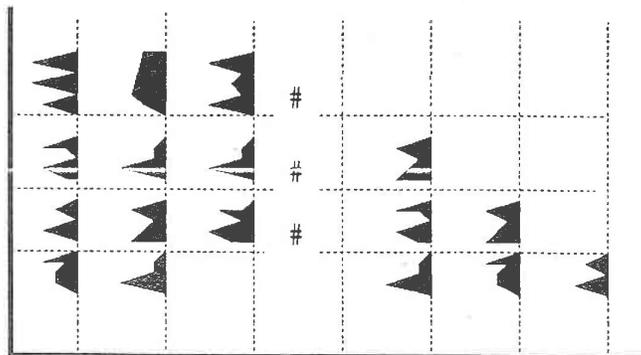
Ces techniques sont basées sur l'histogramme du nombre de pixels **objets** par ligne. Les zones basses de l'histogramme (faible densité) correspondent à des interlignes. le schéma de la figure (5.2.2.1) reprend l'approche qui avait été développée pour le projet de lecture du code postal français.

Pour les lignes inclinées, il est possible d'utiliser des histogrammes locaux en faisant les distributions de pixels **objets** par ligne dans des bandes verticales de l'image. Cette opération est schématisée sur l'exemple de la figure (5.2.2.2). L'algorithme recherche ensuite les vallées de chaque histogramme afin de reconstituer des lignes horizontales par corrélation. Le problème majeur reste la détection des vallées lorsque les lignes sont entrelacées et inclinées.



G B M N / 7 3 9 8
 B a r r o n - A n d e r s o n
 7 4 5 A t l a n t i q u e A v e .
 B o s t o n , M a s s . 0 2 1 1 1

IMAGE ORIGINALE



HISTOGRAMMES LOCAUX HORIZONTALS

Figure 5.2.2.2 Détection des lignes en utilisant des histogrammes locaux.

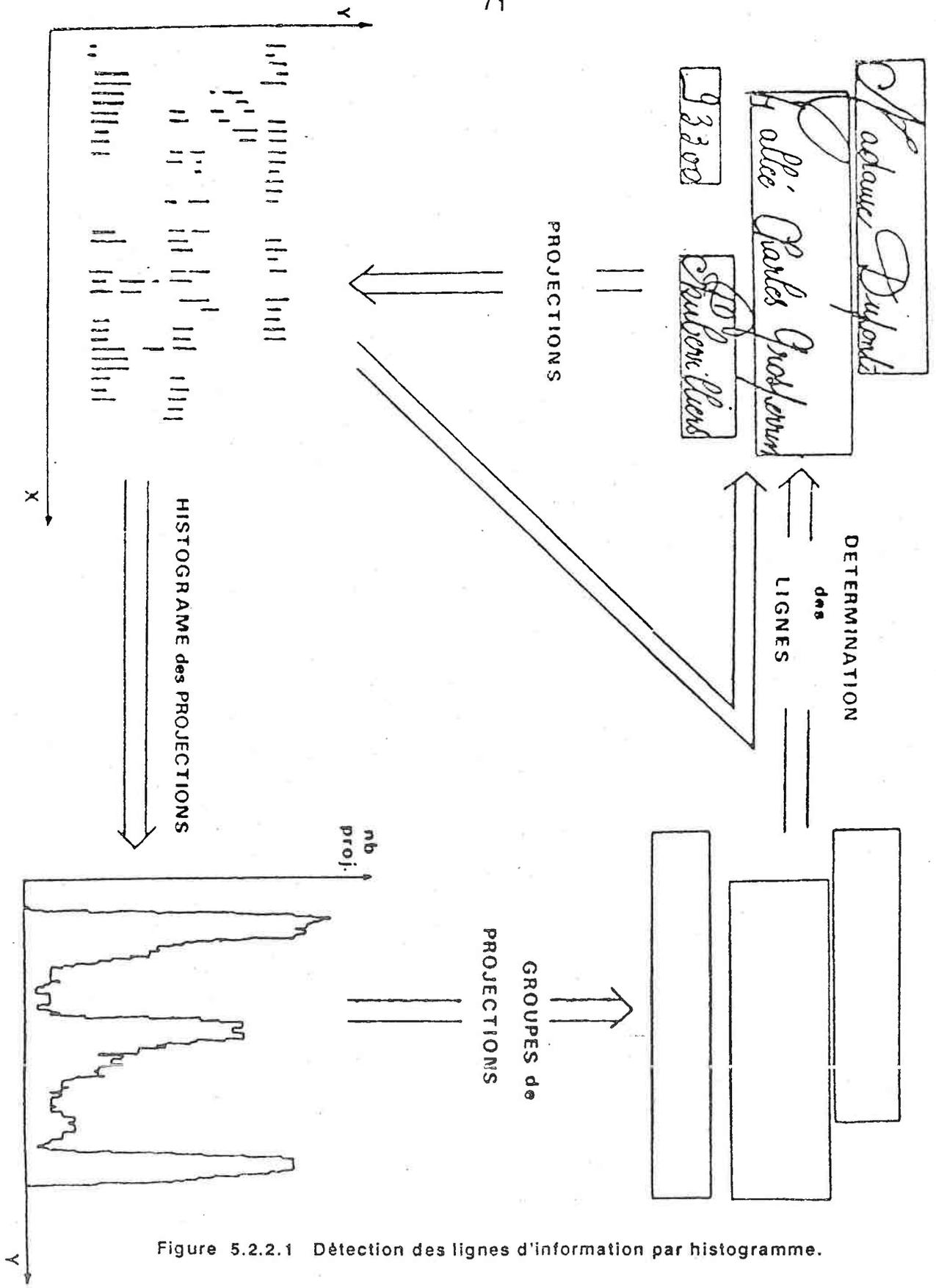


Figure 5.2.2.1 Détection des lignes d'information par histogramme.

5.2.3 - CONCLUSION.

Si les techniques d'histogrammes locaux apparaissent plus adaptées que les déterminations de lignes d'information par croissance de régions, les résultats obtenus ne donnent aucune indication sur le contenu. Les ensembles peu instructifs, générés par les méthodes précédentes, ne permettent pas d'éviter l'explosion combinatoire d'hypothèses "**code-état-ville**" et n'évitent pas non plus les confusions. En effet, les techniques présentées ont l'inconvénient de ne travailler qu'à l'échelle du pixel. La localisation aveugle qui s'en suit ne voit dans la composante connexe qu'un amas de points.

Pour générer un minimum de candidats, les lignes d'information doivent être tronquées en blocs pour ne contenir que des formes de même nature. A l'échelle du caractère, la signature de la forme peut être identifiée grâce à certains attributs facilement extraits. La procédure de **localisation** choisie a pour dessein de conserver les caractéristiques pertinentes des composantes connexes pour établir des catégories de formes. Il devient alors possible de comparer les positions relatives et les signatures de ces formes pour les regrouper en blocs qui peuvent contenir une ou plusieurs entités (nombre, mots). L'agencement spatial de ces blocs est aussi utilisable, car ceux-ci se rapprochent plus du mot que de la ligne. Il devient alors possible de différencier les blocs faits de caractères séparés (nombre, mots en écriture batons), et ceux contenant des mots cursifs (mots courts, mots longs). Ces informations sont suffisantes pour localiser le code, l'état et la ville. Cette localisation utilise bien sûr les caractéristiques des blocs et des formes, mais son originalité réside dans l'analyse du contexte qui s'extrait de ce type de représentation. Comme nous le verrons à mainte reprise, l'analyse du contexte permet de minimiser les causes d'erreurs, d'augmenter les sanctions positives, sans pour cela faire de la localisation un module de reconnaissance coûteux en temps.

5.3 - DESCRIPTION DE L'IMAGE.

5.3.1 - INTRODUCTION.

Après binarisation, l'image ne contient désormais que des formes géométriquement différenciables. Pour définir le mode de représentation que nous avons choisi, la première tâche doit extraire certains attributs pertinents des composantes connexes. Ce premier algorithme que nous nommerons description doit aussi étiqueter chaque unité d'image (pixel) par une référence caractéristique de l'objet décrit. Ces deux niveaux de description seront utiles pour la localisation et pour la reconnaissance.

Cette tâche de bas niveau se doit d'être rapide, sans gestion excessive de données et d'une algorithmique facilement réalisable en temps réel. Les caractéristiques recherchées doivent permettre l'obtention des hypothèses **code-état-ville** sans pour cela empiéter sur le domaine de la **description topologique** des formes (extraction des primitives).

5.3.2 - CARACTERISTIQUES PERTINENTES DES COMPOSANTES CONNEXES.

i) Cadre de la forme (figure (5.3.2.1)).

- Celui-ci définit la position et l'étendue de la forme dans l'image. Il est utile pour l'association des formes en lignes d'informations.

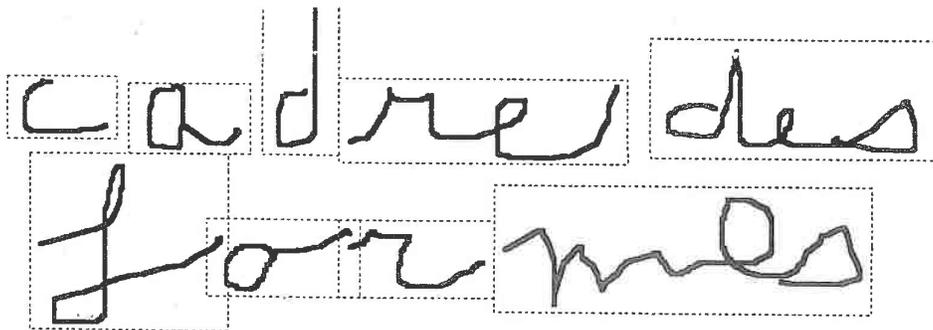


Figure 5.3.2.1 Cadres des formes.

ii) Surface.

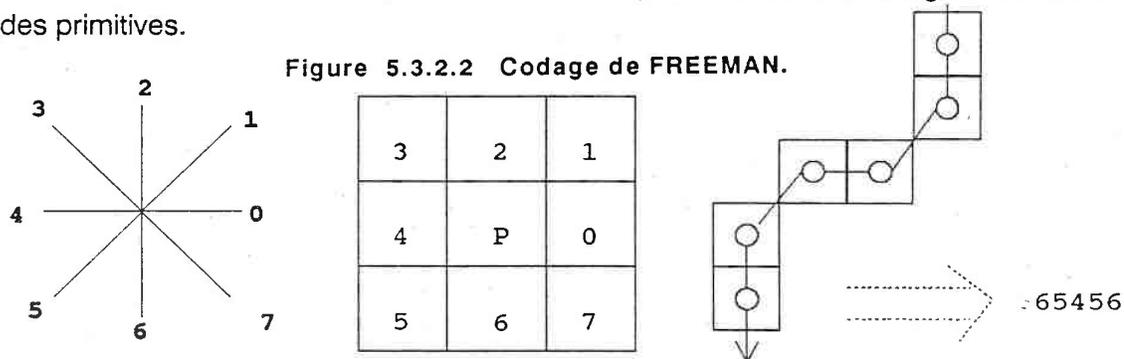
Celle-ci permet d'éliminer les formes aux dimensions extrêmes. L'homogénéité des surfaces des différentes catégories de formes est instructive pour l'association des éléments et la localisation du code postal.

iii) Périmètre.

Caractérisation des formes similaires à ce que l'on obtient avec la surface. Il y a possibilité de distinguer les périmètres internes et externes (boucles).

iii) Code de Freeman (figure (5.3.2.2)).

- Obtenu par suivi de contour en même temps que le périmètre. Le **code de Freeman** (graphe des directions) nécessite une étude minutieuse pour être instructif. Cette tâche coûteuse en temps de calcul est implicitement liée à l'étape d'extraction des primitives.



NOTATION UTILISEES POUR CODER LES CHAINES DE PIXELS DU CONTOUR.

iii) Points extrêmes (figure (5.3.2.3)).

- Ces points du contour externe correspondent à des positions extrêmes **nord** et **sud**. La stabilité du nombre et des positions de ces points est un attribut des plus pertinents pour une "pré-classification" des formes.

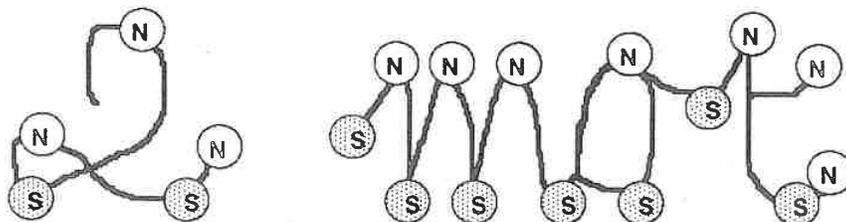


Figure 5.3.2.3 Points extrêmes Nord (haut) et Sud (bas).

5.3.3 - DESCRIPTION DE L'IMAGE PAR SUIVI DE CONTOUR.

Cette démarche classique en reconnaissance des formes utilise le contour binaire des formes pour en faire la description. Afin d'étiqueter les pixels et de déterminer si le contour est interne ou externe, les algorithmes de suivi les plus simples travaillent simultanément avec l'image du contour et l'image binaire. La figure (5.3.3.1) schématise un processus de suivi de contour simple à développer. Le cheminement aléatoire et les comparaisons d'images qui en résultent, sont coûteux en gestion des données. La totalité des deux images doit rester disponible tout au long de la procédure.

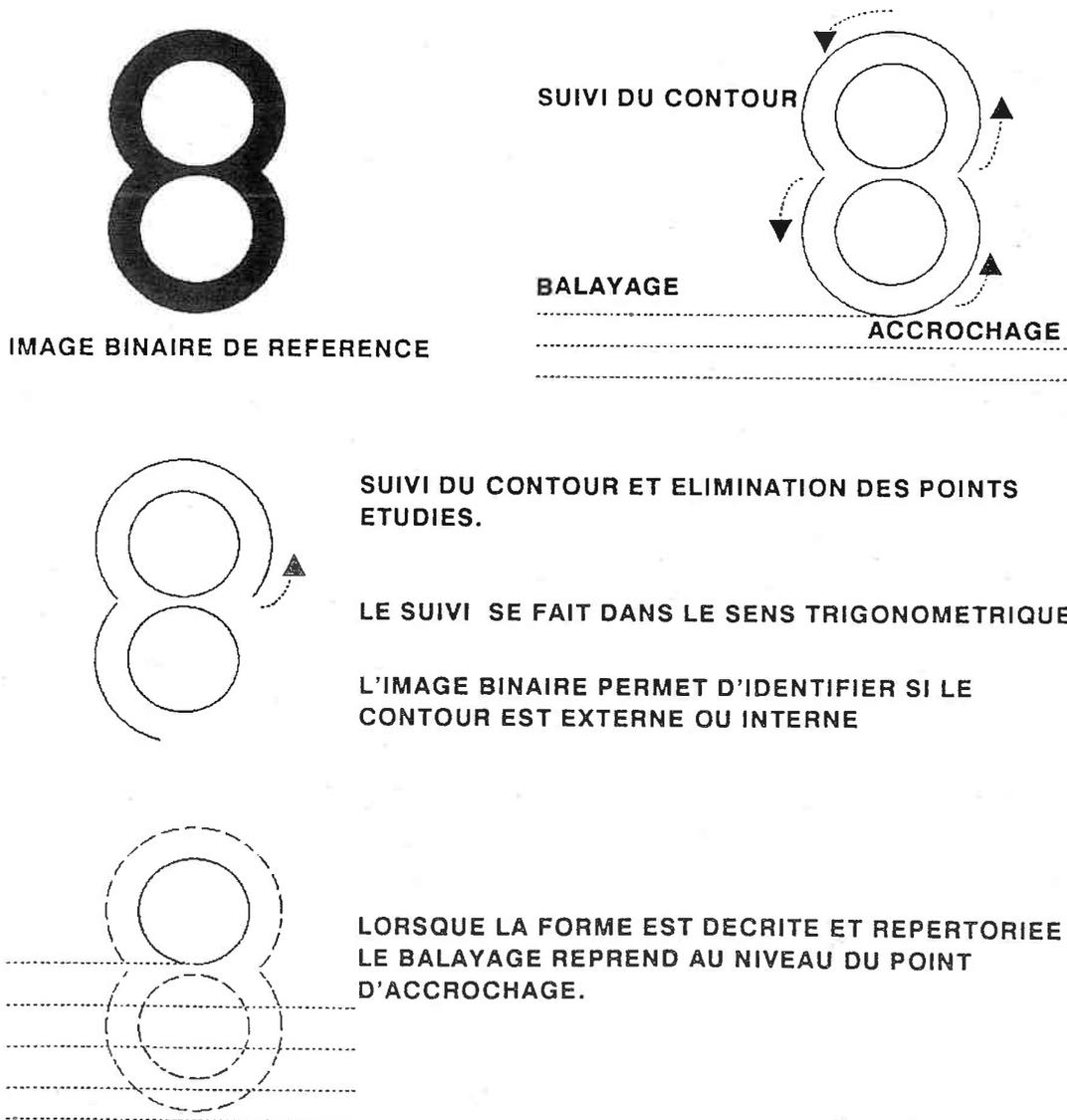


Figure 5.3.3.1 Description des formes par suivi de contour.

5.3.4 - DESCRIPTION PAR BALAYAGE DE L'IMAGE ET SUIVI DE GRAPHES.

En considération des contraintes de temps de calcul imposées aux systèmes industriels de lecture de code, la description des composantes connexes doit être faite très rapidement. Sur les conseils des électroniciens de CGA-HBS, nous avons recherché un algorithme dont les opérations sont réalisables en temps réel.

5.3.4.1 - Possibilités d'intégration et expérience industrielle.

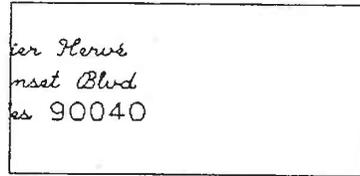
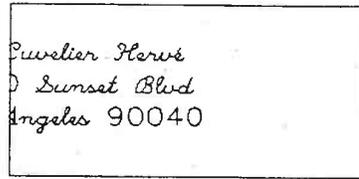
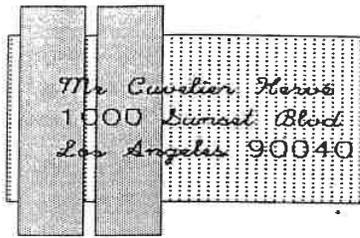
Actuellement, la mise en parallèle des opérations effectuées sur les pixels permet de travailler en temps réel sur une vingtaine de colonnes de l'image, lors d'un balayage horizontal.

L'image binaire que nous utilisons est représentée par 1456 colonnes de segments. Chaque colonne unitaire contient un certain nombre de segments de pixels appartenant à différentes formes.

Le choix d'une description de l'image par balayage apparaît inéluctable quand la fiabilité et la rapidité des calculs sont prouvées par un prototype opérationnel. C'est le cas de la binarisation par le système SYBILIS. Le balayage est économique en gestion des données et en temps de calcul.

Il est possible grâce aux microprocesseurs du type "transputer" et aux configurations temps réel de cartes électroniques (structure en "pipe-line") d'effectuer en parallèle l'**acquisition**, la **binarisation** et la **description** de l'image. Le décalage théorique entre les trois balayages n'est que de quelques colonnes. Le processus envisagé est représenté sur la figure (5.3.4.1).

ACQUISITION 12 POINTS AU MILLIMETRE PAR DEFILEMENT DE LA LETTRE DEVANT UNE BARRETTE CCD.



LA BINARISATION EST EFFECTUEE EN CONTINUE A PARTIR D'UNE BASE MINIMALE DE 15 COLONNES.

LA DESCRIPTION RECOIT LES PIXELS DE L'IMAGE BINAIRE ET DESCRIE EN CONTINU LES FORMES EN UTILISANT 2 COLONNES.

Figure 5.3.4.1 Architecture " Pipe - Line " pour une Acquisition, une Binarisation, et une Description en continue et parallèle.

5.3.4.2 - Schéma fonctionnel de l'algorithme.

Comme l'explique la figure (5.3.4.2), notre algorithme de suivi de graphes utilise simultanément deux colonnes: la colonne étudiée et la précédente qui sert de référence pour l'identification des formes et le marquage des nouveaux segments.

Le balayage de l'image s'effectue de gauche à droite dans l'image et de bas en haut dans la colonne.

Si le segment étudié n'est pas connexe avec l'un des segments de la colonne de référence, une nouvelle forme est créée.

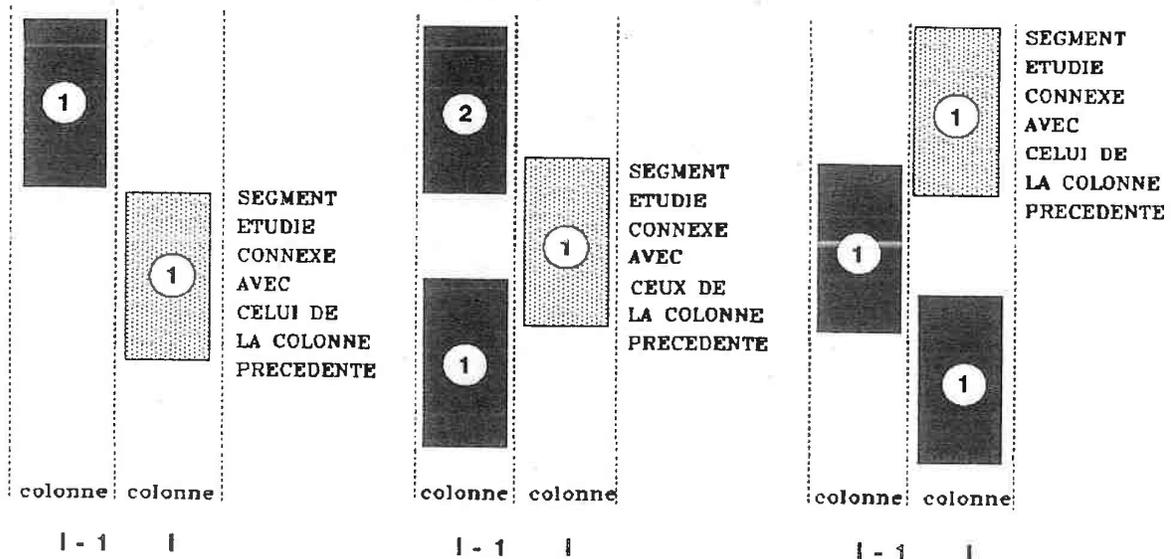


Figure 5.3.4.2 Voisinage de travail utilisé pour la description de l'image par balayage.

Sur l'exemple de la figure (5.3.4.3), tout segment connexe de la colonne étudié apporte ses caractéristiques géométriques à la forme et subit un marquage. Il est ainsi possible de définir une surface, un cadre, et de positionner cette composante connexe dans l'espace bidimensionnel de l'image.

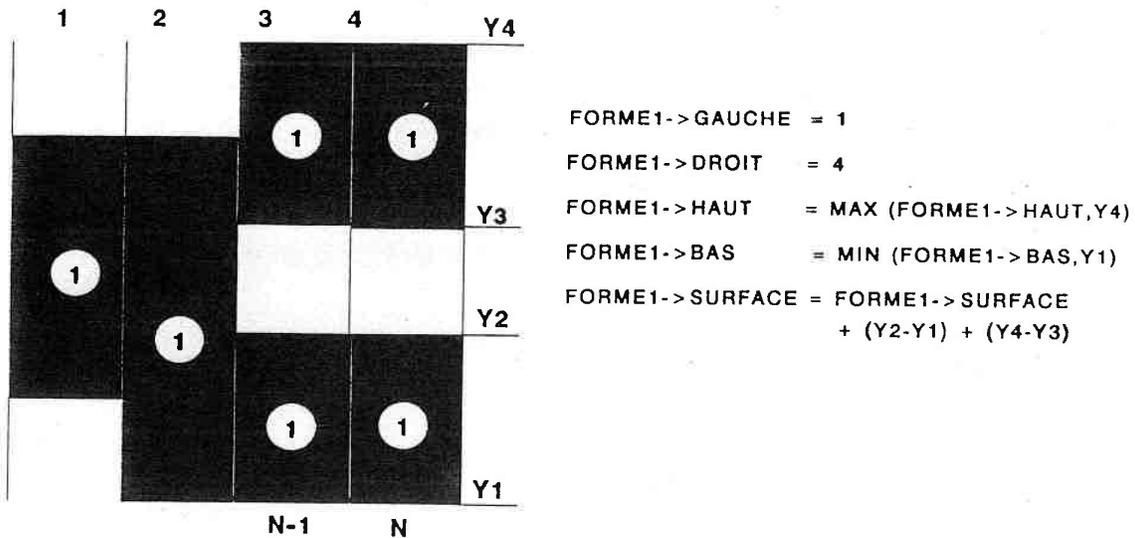


Figure 5.3.4.3 description de la forme et extraction des paramètres.

Au cours du balayage, les formes apparaissent comme des structures arborescentes (graphes) que le système parcourt soit dans le sens "père -> fils", soit dans le sens "fils -> père". La figure (5.3.4.4) représente les configurations possibles que peuvent prendre les segments d'une composante connexe.

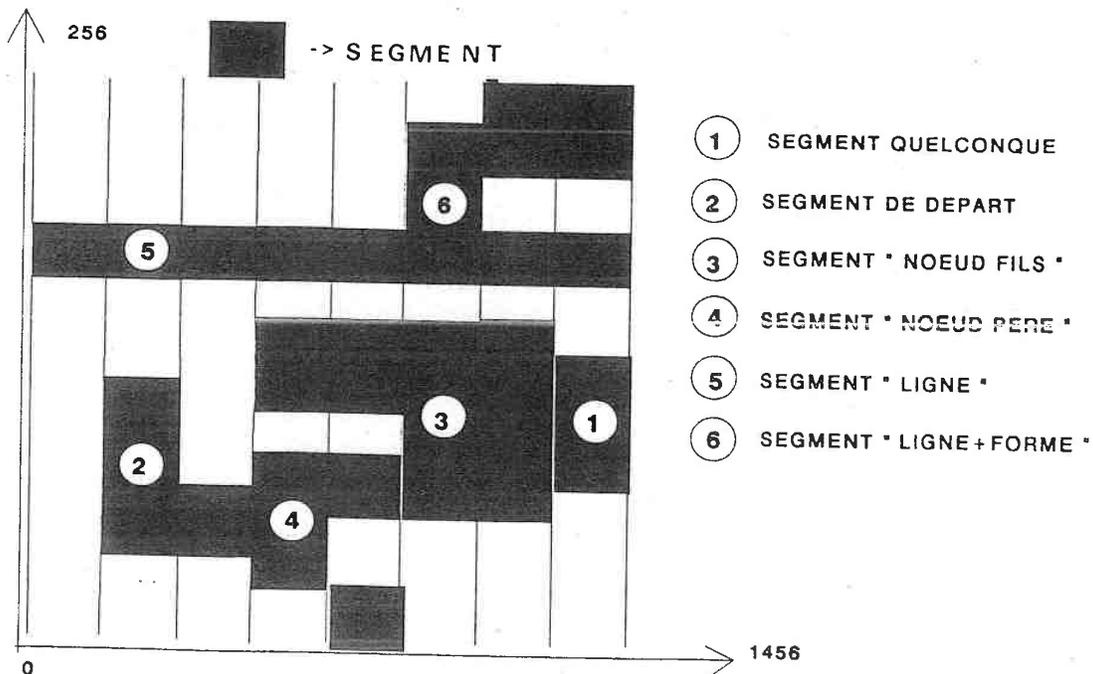


Figure 5.3.4.4 Suivi des arborescences et configurations de segments.

- 1 - **Segment quelconque** connexe à un seul segment non instructif pour l'étude de l'arborescence.
- 2 - **Segment de départ** sans connexion initialisant une nouvelle forme.
- 3 - **Segment "noeud fils"** correspondant à la connexion de plusieurs branches issues d'une ou de plusieurs étiquettes.
- 4 - **Segment "noeud père"** engendrant dans la colonne suivante de nouvelles branches "filles" sous la même étiquette.
- 5 - **Segment "ligne"** connexe à une structure linéaire stable détectée comme étant un soulignement.
- 6 - **Segment "ligne + forme"** correspondant à une forme soulignée.

Pour récupérer les formes soulignées, le processus tente de détecter si les variations verticales significatives de cette succession de segments de ligne ne cachent pas une connexion avec un caractère. Si c'est le cas, les segments sont coupés en deux au niveau supérieur estimé de la ligne. Les parties inférieures deviennent des segments **ligne** et les parties supérieures sont marquées comme **forme**.

5.3.4.3 - Détection des lignes horizontales et des formes connectées à celles-ci.

- Intérêt de la détection des soulignements (figure (5.3.4.5)).

La reconnaissance des caractères devient périlleuse voire impossible quand les caractères sont connectés les uns aux autres. En effet, dans cette apparence d'écriture cursive, le tracé inférieur des formes n'a pas les caractéristiques d'un mot car la base des lettres est noyée dans la ligne. Au cours de l'étape suivante d'association des caractères en bloc d'information, les lignes non détectées perturberont la démarche fondée sur des critères horizontaux.

- La déconnexion des lignes d'écriture est rendue très difficile par la multitude de liaisons possibles. Elle devient réalisable lorsque l'unique lien indésirable est une ligne. L'exemple de la figure (5.3.4.6) est assez représentatif de ce type de connexion.

- Caractéristiques des lignes et soulignements (figure (5.3.4.7)).

Quelle que soit la génération de la ligne, limite de fenêtre, soulignement manuel ou pré-imprimé, la ligne est une forme horizontale d'épaisseur assez constante. Une saisie défectueuse ou un tracé manuel peut la rendre légèrement oblique. Le système détecte des lignes dont l'inclinaison est inférieure à 10 %. La **binarisation** employée transforme les lignes obliques en successions de paliers d'une longueur supérieure à dix pixels. Le décalage vertical de ces marches est de l'ordre du pixel.

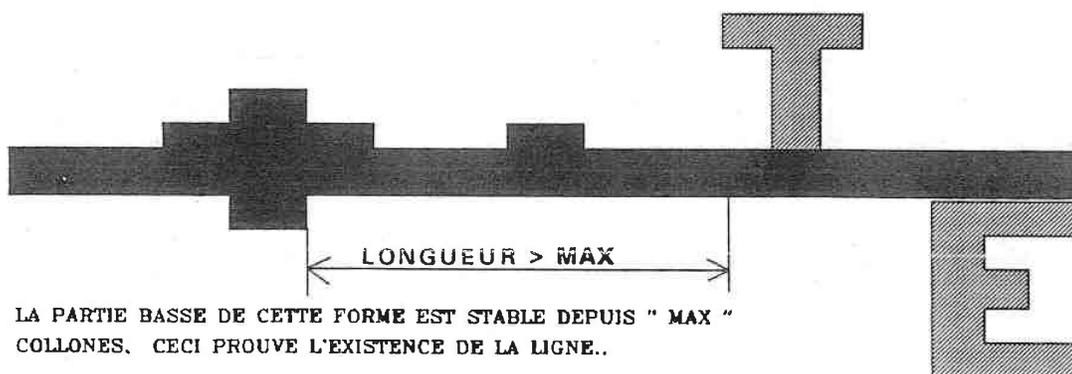


Figure 5.3.4.7 Détection des lignes et soulignements.

04 02
 Wright Walters Magazine
 PUBLICATION

05
 P.O. Box 56129

ADDRESS STATE ZIP CODE
 Boulder, Co. 80322-6129

02
 ELMIRA, NEW YORK 14901 TELEPHONE 7
 Keroll Furniture
 P.O. Drawer 189
 Hickory, N.C. 28601
 05

Figure 5.3.4.5 Intérêt de la détection des soulignements.

Figure 5.3.4.6 Déconnexion des lignes d'écriture au niveau du soulignement.

- figure (5.3.4.8) : La stabilité de la base de la ligne permet de déterminer sa présence. Sur son niveau supérieur, plus hétérogène, le système détecte les variations significatives qui s'y produisent. L'élimination des bruits est assez efficace et les excroissances persistantes sont considérées comme faisant parties d'un caractère. Ces zones sont alors isolées de la ligne.

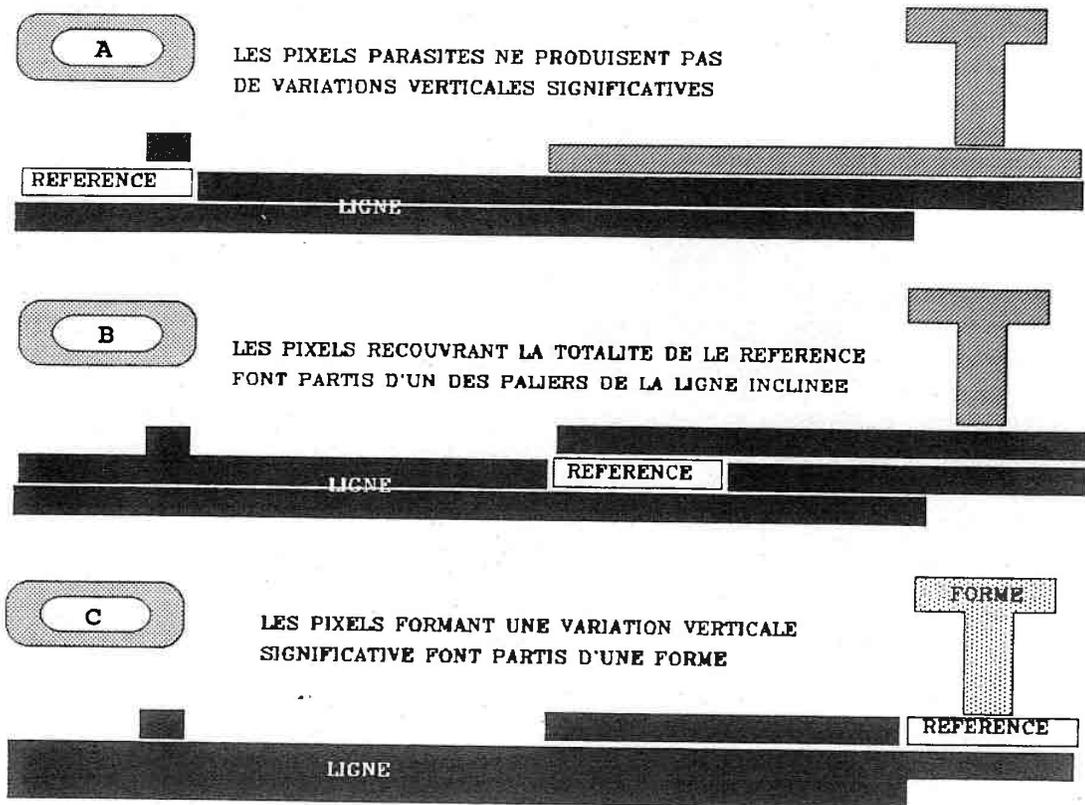
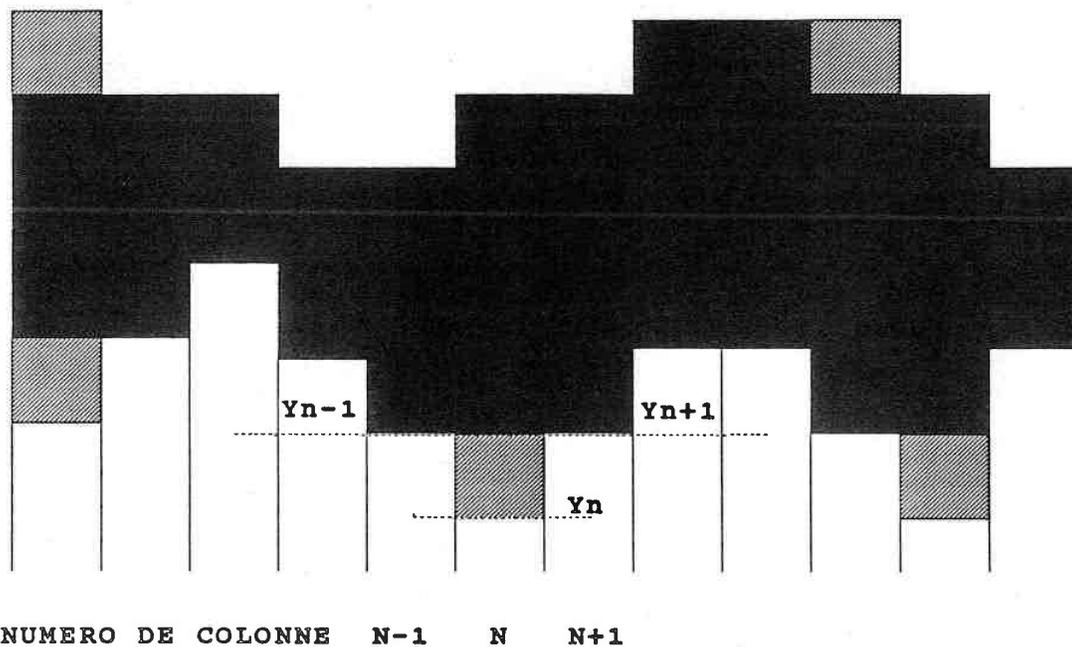


Figure 5.3.4.8 Détection des formes connectées à une ligne.

5.3.4.4 - Détection des points extrêmes.

- Calcul des points extrêmes (figure (5.3.4.9)).

Chaque segment est comparé à ses proches voisins de droite et de gauche. Les voisins qui le précèdent conservent les paramètres contextuels des colonnes étudiées, plus éloignées sur la gauche. A la fin de l'étude d'une colonne, les points extrêmes déterminés sont triés car les différentes branches d'une même forme peuvent générer des maximums et des minimums erronés.



```

if Yn # Yn-1 Prev_Y(n) = Yn-1
sinon      Prev_Y(n) = Prev_Y(n-1)

```

```

INFERIEUR[n] SI Yn < Prev_Y(n)
ET           Yn > Yn+1

```

```

SUPERIEUR[n] SI Yn > Prev_Y(n)
ET           Yn < YN+1

```

Figure 5.3.4.9 Calcul des points extrêmes.

5.3.4.5 - Evaluation.

L'évaluation se fait en considérant uniquement l'existence ou l'absence de connexions entre les chiffres du code postal et la ligne.

- Base d'apprentissage : 1000 images.

Sur 92 codes postaux soulignés traités, 25 séparations se sont avérées défectueuses.

- Base d'évaluation : 1000 images.

Sur 104 codes postaux soulignés traités, 21 séparations se sont avérées défectueuses.

Cet algorithme éprouve des difficultés à détecter les lignes quand elles sont altérées, trop inclinées, ou brisées. De plus, si une forme est connectée au début du soulignement, et si la présence de la ligne n'est pas prouvée, cette forme sera éliminée dans la suite du processus.

5.4 - ETIQUETAGE DES FORMES.

5.4.1 - INTRODUCTION.

Une image d'enveloppe est constituée de mots, de caractères isolés, de ponctuations et de graphiques. Sans un étiquetage sommaire de ces différentes catégories de formes, il est impossible de formuler rapidement et de manière fiable, des hypothèses contenant un **code postal** de 5 ou 9 chiffres, un **état** et une **ville**. De plus, il est primordial d'éviter l'explosion combinatoire de candidats que produit un processus ignorant le contenu des blocs d'information.

Pour palier à cela, la méthode d'étiquetage présentée s'appuie sur les attributs robustes des composantes connexes que fournit le module de description. Ces attributs permettent de faire la distinction entre :

- les virgules,
 - les tirets,
 - les 1 et les i,
 - les caractères isolés,
 - les doublets et triplets de caractères chaînés,
 - les mots courts.
 - les mots longs et les graphiques.
-
- Les **virgules** sont des ponctuations utilisées pour séparer le nom de l'état de celui de la ville dans l'adresse postale. Nous verrons par la suite l'intérêt de les détecter.
 - Les **tirets** aident à déterminer les codes postaux contenant une extension (12345 - 6789), car ils sont utilisés comme séparateurs. Les **tirets** représentent aussi les barres horizontales de chiffres (5) ou de lettres (E,F,L,T,Z) brisés.
 - Les **segments verticaux** proches du 1 et du i peuvent appartenir aux chiffres brisés 1, 4 et 9.

- La notion de **caractère unique** permet de donner des hypothèses fiables pour les codes postaux.
- Le **mot court** correspond à deux ou trois caractères connectés. Il représente souvent un doublet ou un triplet de chiffres connectés dans le **code postal** ou l'abréviation régulière de l'**état**. Dans la suite du processus, un module de déconnexion des chiffres chaînés coupe les **mots courts** du code postal (doublets et triplets de caractères).
- Le reste est considéré comme **mots longs** ou **graphiques**.

5.4.2 - CRITERES D'IDENTIFICATION.

Sur la figure (5.4.2.1) sont représentés les critères d'étiquetage qu'utilise notre système. Fidèles à l'approche que nous avons choisie pour les pré-traitements, cet étiquetage utilise les caractéristiques géométriques simples des formes, telles que les surfaces, les longueurs, le nombre et le positionnement des points extrêmes.

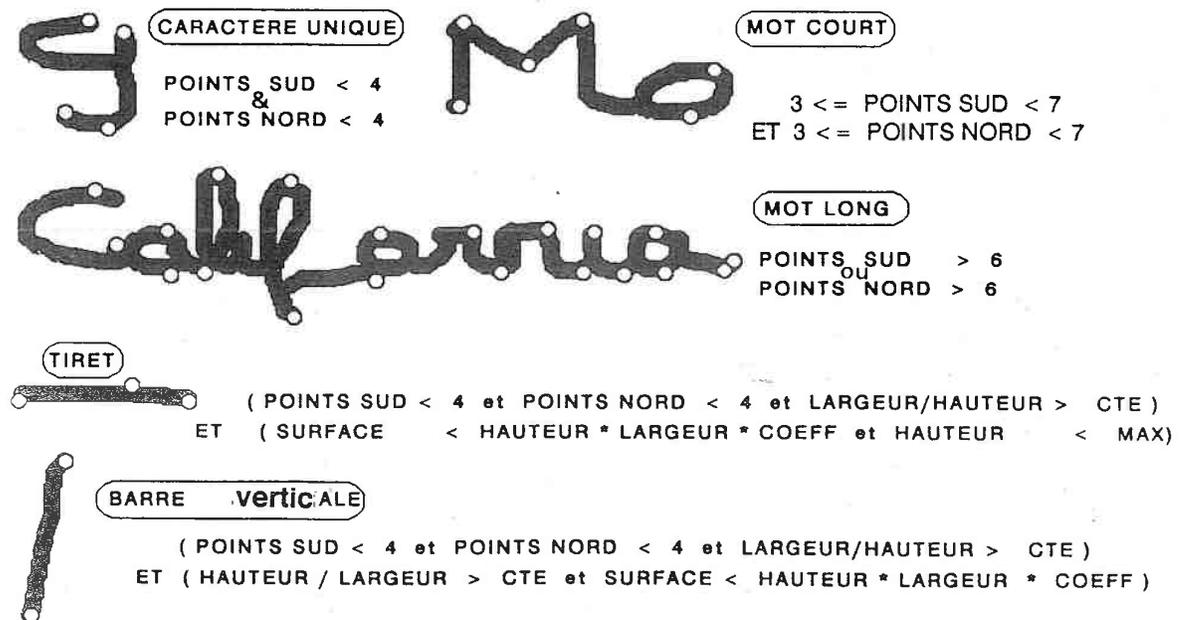


Figure 5.4.2.1. Critères d'étiquetage.

5.4.3 - DETECTION DES DOUBLETS ET TRIPLETS DE CHIFFRES CHAINES.

- Détection des **doublets** (figure (5.4.3.1)).

Si l'on postule que le doublet ou le triplet correspond à la fusion des caractéristiques des deux ou trois chiffres qui le composent, chaque moitié ou tiers du cadre de cette forme représente un chiffre unique. Les classificateurs du système ne pouvant pas identifier les caractères chaînés, une segmentation de ces formes est nécessaire. Dans une première étape, un module recherche seulement les ensembles de deux ou trois chiffres connectés. La détection se fait selon la même méthodologie que celle employée pour l'étiquetage.

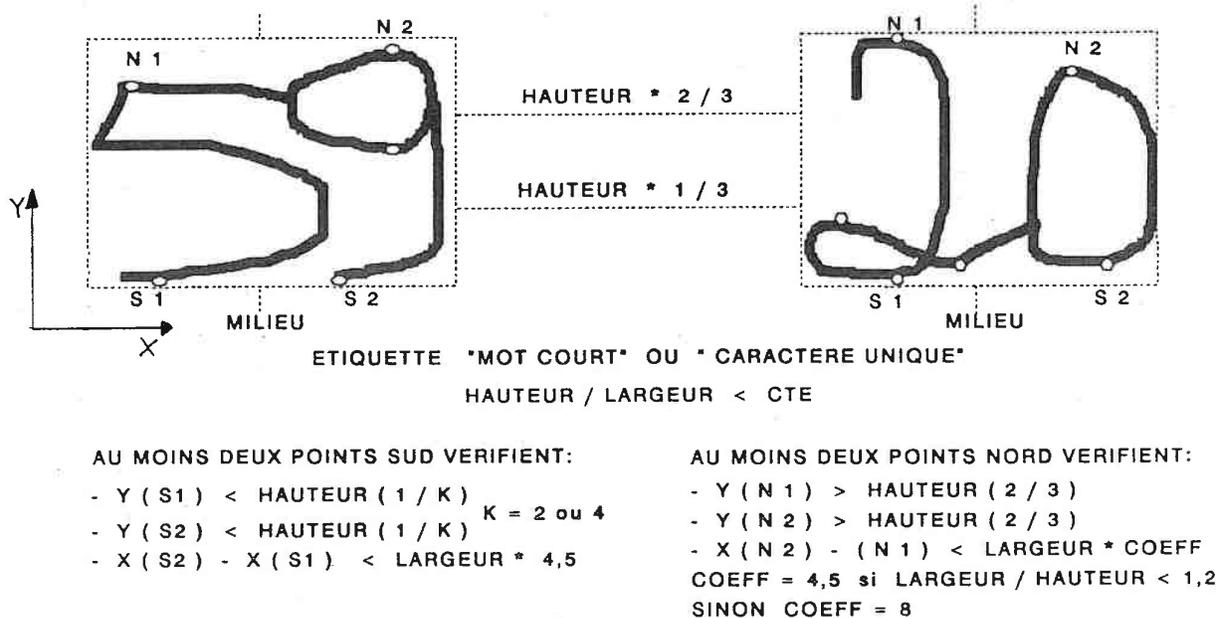
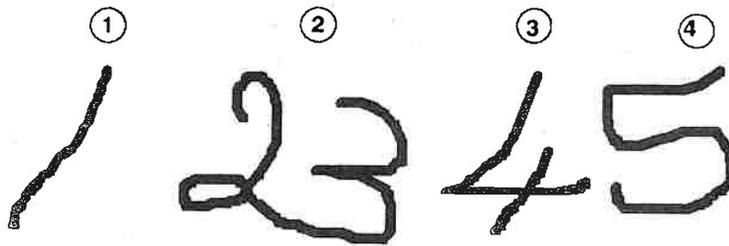


Figure 5.4.3.1 Détection des doublets.

Sur une base de 132 doublets, seulement 13 n'ont pas été détectés.

- Annulation ou validation du **doublet** en fonction du contexte (figure (5.4.3.2)).

Certains chiffres, souvent le 4 et le 5, peuvent avoir les attributs du doublet. L'élimination de ces faux doublets se fait par comparaison des dimensions géométriques (taille, surface, position...) du **mot court** avec les caractères uniques de droite et de gauche. Une distinction entre un chiffre quelconque et le 1 est indispensable, vu ses dimensions marginales.



LA FORME ② EST CONSIDERE COMME DOUBLET SI :

- CETTE FORME A ETE DETECTEE COMME DOUBLET
- SA SURFACE > MAX (SURFACE DES FORMES NORMALES ③ ET ④) * COEFF1
- SA SURFACE > SURFACE DE LA BARRE VERTICALE ① * COEFF2
- LE DOUBLET SE TROUVE DANS LES 4 DERNIERS CARACTERES QUAND LE BLOC CONTIENT PLUS DE 8 FORMES.

LA CONFIGURATION 8 FORMES PEUT CORRESPONDRE A UN CODE + EXTENSION

Figure 5.4.3.2 Elimination des faux doublets.

Sur une base de 1000 lettres, seulement 11 chiffres isolés ont été détectés comme doublets.

Lors de l'étape de formation des candidats **code-état-ville**, tout bloc intéressant et contenant cinq formes dont un **doublet** voit l'élimination de l'hypothèse d'une connexion entre chiffres. Cette condition permet de rattraper les faiblesses de l'algorithme de détection.

5.4.4 - DETECTION DE LA VIRGULE.

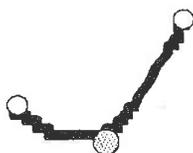
5.4.4.1 - Introduction.

La **syntaxe d'adressage** des lettres américaines veut que la **ville** soit séparée de l'**état** par une **virgule**. Dans 60 % des lettres, cette virgule est présente.

Cette ponctuation représente un élément discriminant dans la recherche d'une hypothèse **code-état-ville**. La ville se trouve placée à droite ou en dessous de la virgule. Il peut exister plusieurs virgules dans l'image, ou des caractères (i) détectés comme telles. Pour éviter les confusions, il est nécessaire d'utiliser le contexte que représente les formes de droite et de gauche.

5.4.4.2 - Détection (figure (5.4.4.1)).

La détection des virgules se fait sur des critères géométriques simples. Les tests sont suffisamment stricts pour ne pas trop détecter de fausses virgules.

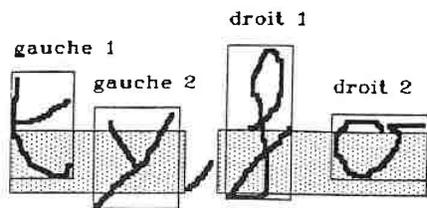


POINTS NORD < 3 ET POINTS SUD < 2
 & SURFACE < MAX_SURFACE
 & SURFACE < (HAUTEUR+LARGEUR) * EPAISSEUR

Figure 5.4.4.1 Détection de la virgule.

5.4.4.3 - Etude du contexte (figure (5.4.4.2)).

Recherche de formes voisines contenues dans le même bloc ou dans les blocs se trouvant sur la même ligne horizontale. L'existence d'un voisin sur la gauche est de manière syntaxique indispensable pour valider la virgule, car il représente un élément de la **ville**.



- 1) RECHERCHE DE VOISINS SUR LA DROITE ET SUR LA GAUCHE DE LA VIRGULE DANS LE BLOC.
- 2) SI LA VIRGULE EST ISOLE. RECHERCHE DE VOISINS DANS LES BLOCS PROCHES.

REGLE 1 : VOISIN [nord] > VIRGULE [sud]
 REGLE 2 : VOISIN [sud] < VIRGULE [nord] + SEUIL

SI GAUCHE 2 = " MOT" ALORS GAUCHE 1 ET GAUCHE 2 NE SONT PAS DES ELEMENTS DU CONTEXTE.

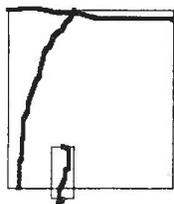
SI HAUTEUR [gauche 1] > 2 * HAUTEUR [gauche 2]
 ALORS GAUCHE 1 N'APPARTIENT PAS AU CONTEXTE..

Figure 5.4.4.2 Recherche des voisins de la virgule.

5.4.4.4 - Validation (figure (5.4.4.3)).

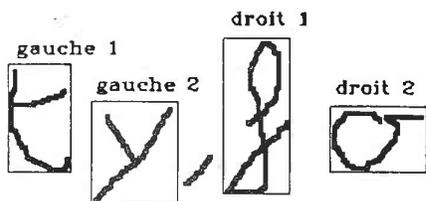
Selon les configurations, la virgule se trouve entourée de un à quatre caractères voisins. Ces formes vont permettre d'éliminer des hypothèses générées sur les dernières lettres de mots ou les barres de caractères coupés de la ligne inférieure.

Une fois validée et étiquetée, la virgule conserve la trace de ces proches voisins.



PREMIERE CONDITION :

TOUTE VIRGULE DONT LE CADRE EST FORTEMENT RECOUVERT PAR CELUI D'UN DE SES VOISINS EST CONSIDERE COMME CARACTERE ET NON COMME VIRGULE.



SECONDE CONDITION :

HAUTEUR = GAUCHE 1 [hauteur] + GAUCHE 2 [hauteur] + DROIT 1 [hauteur] + DROIT 2 [hauteur]
 SUD = GAUCHE 1 [sud] + GAUCHE 2 [sud] + DROIT 1 [sud] + DROIT 2 [sud]

REGLE 1 = VIRGULE [nord] < 1/4 (SUD + HAUTEUR)

REGLE 2 = VIRGULE [sud] < SUD / 4

minhauteur = MIN (GAUCHE 2 [hauteur] , DROIT 1 [hauteur]) * COEFF

REGLE 3 = GAUCHE 2 [nord] - DROIT 1 [sud] > minhauteur

REGLE 4 = DROIT 1 [nord] - GAUCHE [sud] > minhauteur

Figure 5.4.4.3 Etude du contexte et critères de validation de la virgule.

5.5 - RESTAURATION DES FORMES BRISEES.

5.5.1 - INTRODUCTION.

Outre les variations de l'encre qui peuvent entraîner de fortes différences de teintes, les brisures de chiffres sont essentiellement produites par le lever de crayon. Cette lacune de tracé n'est pas aléatoire et correspond aux rebroussements du contour du chiffre ou aux rajouts de tirets. Les brusques variations de teintes entraînent la formation d'une franche coupure à la **binarisation**. Les brisures de caractères dépendent du type d'écriture utilisé et des particularités de tracé.

Les chiffres français brisés sont en majorité des 1, 4, 5 et 8. Les barres supérieures de 5 sont détachées dans 20 % des cas. Le courrier américain contient lui aussi un très grand nombre de 5 brisés, et une forte population de 4 et de 9 coupés.

Notre objectif est de restaurer ces formes, c'est à dire pouvoir soumettre au module de reconnaissance une entité contenant tous les morceaux du chiffre brisé.

5.5.2 - INVESTIGATIONS.

- Dans l'article de M. Shridhar et A. Baldreldin (SHR 85) l'utilisation des profils droits et gauches des formes permet de fusionner les formes brisées par recouvrement des **profils**. La figure (5.5.2.1) montre un exemple de restauration. Malheureusement, dans le cas d'une succession de caractères, cette démarche n'explique pas l'appartenance d'une forme à l'élément de droite ou de gauche.

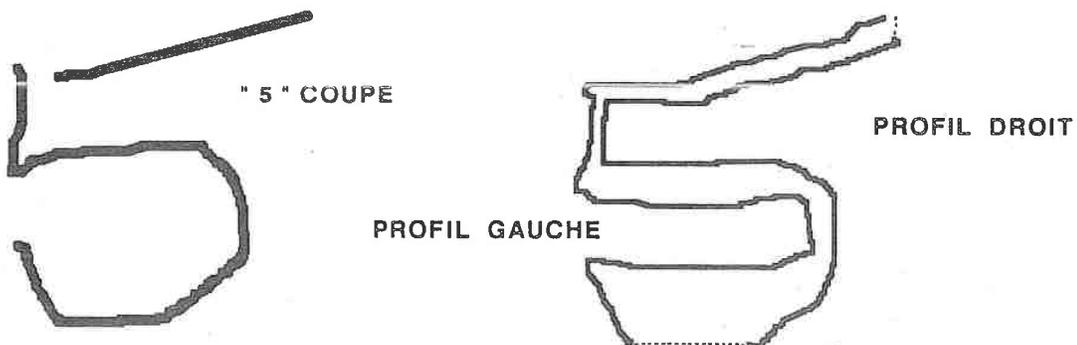
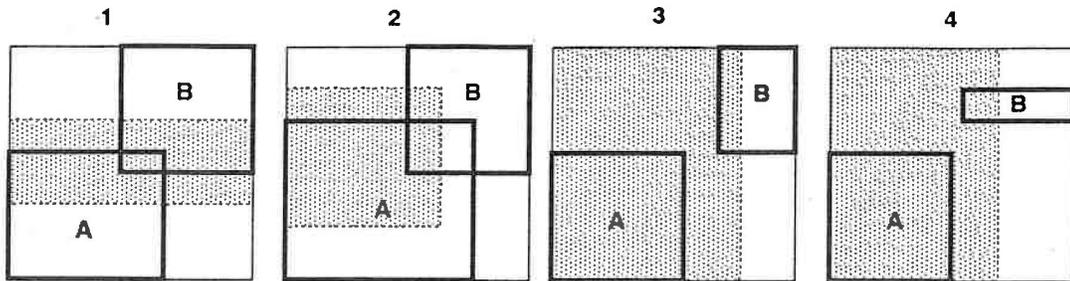


Figure 5.5.2.1 Restauration des formes brisées par recouvrement des profils.

- Dans le système fourni aux P.T.T. françaises, une tentative de connexion est envisagée pour les formes géométriquement proches lorsque le nombre de formes dépasse celui du code réglementaire de cinq chiffres. L'exemple de la figure (5.5.2.2) schématise les conditions indispensables pour que deux formes A et B puissent être associées. Ces formes sont représentées par les cadres et les conditions de proximité par le grisé.



LES 2 FORMES SONT ASSOCIEES SI LES CONDITIONS DE PROXIMITE (1 OU 2) ET (3 OU 4) SONT REMPLIES.

Figure 5.5.22 Association des formes par des critères de proximité.

5.5.3 - CLASSIFICATION DES FORMES BRISEES.

Deux catégories de chiffres brisés sont déduites de l'étude de la base de lettres américaines :

- Les **symptomatiques** du lever de crayon.

Le 5 avec sa barre supérieure détachée.

Le 4 et le 9 avec une ligne verticale et une courbe à sa gauche.

- Les **accidentels**.

Les 0, 3, 8 avec des boucles brisées.

Les 1, 7 avec des lignes brisées.

Quelques formes coupées appartenant à ces deux classes sont représentées sur la figure (5.5.3.1).



Figure 5.5.3.1 Catégories de chiffres brisés.

5.5.4 - METHODOLOGIE DE LA RESTAURATION.

Les coupures accidentelles ne peuvent être rattrapées que par des règles fondées sur le recouvrement latéral et vertical des formes. Sachant que l'écriture manuscrite peut être ramassée et fortement inclinée, les associations horizontales doivent être réalisées en fonction de critères très restrictifs. Le rapport des surfaces évite de fusionner deux chiffres inclinés. Les règles principales sont explicitées sur la figure (5.5.4.1).

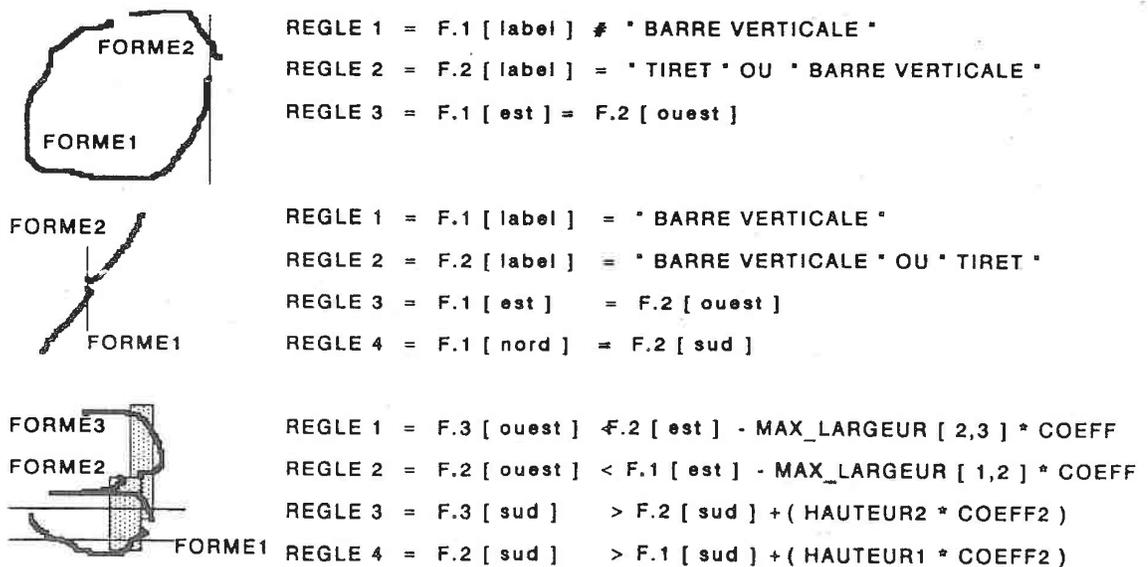


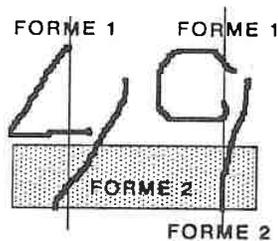
Figure 5.5.4.1 Critères de restauration des chiffres brisés de manière accidentelle.

Ces règles rigoureuses ne peuvent pas empêcher certaines associations erronées. Les configurations représentées par la figure (5.5.4.2) sont modélisables et la confusion peut alors être évitée.



Figure 5.5.4.2 Configurations particulières où le recouvrement des formes est très important.

La démarche de restauration diffère pour les formes symptomatiques car celles-ci ont des caractéristiques robustes. L'étude des brisures et des relations géométriques existant entre celles-ci, permet de définir l'appartenance de ces fragments à un modèle. Les critères principaux sont explicités sur la figure (5.5.4.3).

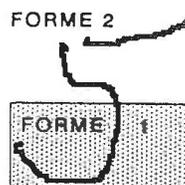


REGLE 1 = F.1 [label] ≠ " BARRE VERTICALE "

REGLE 2 = F.2 [label] = " BARRE VERTICALE "

REGLE 3 = F.1 [est] ≥ F.2 [ouest]

REGLE 4 = F.1 [sud] > F.2 [sud] + (HAUTEUR1 * COEFF2)



REGLE 1 = F.1 [label] ≠ " BARRE VERTICALE "

REGLE 2 = F.2 [label] = " BARRE VERTICALE " ou " TIRET "

REGLE 3 = F.2 [sud] > F.1 [sud] + (HAUTEUR1 * COEFF1)

Figure 5.5.4.3 critères d'association des morceaux de chiffres brisés symptomatiques

5.5.5 - VALIDATION.

Les résultats sont très satisfaisants dans les cas simples, mais si l'un des fragments du chiffre ne se trouve pas dans le même bloc, ou en est trop éloigné, l'association n'est pas faite.

Sur les 1000 images de la base d'évaluation :

- 175 chiffres ont été correctement restaurés,
- 19 sont restés brisés.

Une des lacunes de cette démarche est l'impossibilité de pouvoir détecter les **coupés-chainés** (figure (5.5.5.1)). L'élément le plus représentatif de cette catégorie est la configuration barre du 5 coupé liée par la droite à un autre chiffre.

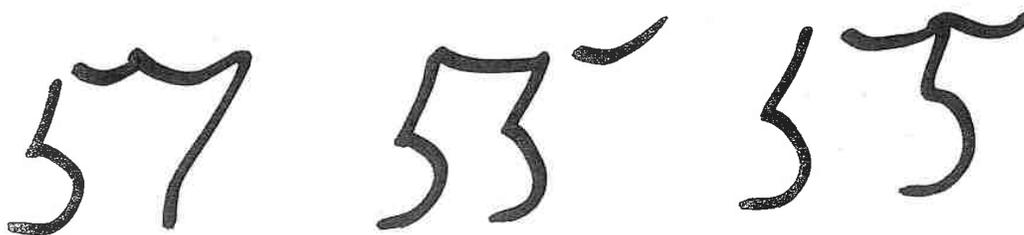


Figure 5.5.5.1 Exemples de chiffres brisés et de configurations "coupée-chainées".

5.6 - SEPARATION DES CHIFFRES CONNECTES

5.6.1 - INTRODUCTION.

L'approche difficile de l'écriture cursive pousse les industriels à ne travailler qu'au niveau du caractère isolé. Cette ignorance s'explique aussi par le fait que les caractères numériques recherchés, (code postal, chèque,...) ou les caractères alphabétiques (formulaires administratifs,...), sont très souvent séparés ou contraints dans des zones.

Au travers des documents issus de recherches bibliographiques, la reconnaissance de l'écriture cursive ou la séparation des caractères connectés demeure un écueil sur lequel se buttent les taux de lecture. Les publications relatent seulement certaines tentatives ponctuelles:

- une reconnaissance d'écritures cursives provenant d'un nombre restreint de personnes.
- Reconnaissance d'un nombre limité de mots.
- la reconnaissance de l'écriture sur tablette après apprentissage. Cette démarche est bien différente de celles utilisées lors de reconnaissances optiques. Cette antinomie provient de l'analyse des caractéristiques temporelles du tracé. Chaque point du mot est défini par des coordonnées spatiales et temporelles.

Le but de ce module est de déconnecter les ensembles de chiffres chaînés, détectés comme doublets ou triplets par l' algorithme présenté au chapitre 5.4.3

5.6.2 - INVESTIGATIONS.

Dans le cadre de cette étude sur le courrier américain, l'estimation des chiffres connectés dans le code postal approche les 20 %. Ce sont principalement des doublets et quelques triplets.

Sur les 2.000 lettres manuscrites américaines, une estimation relève la présence de:

- 80 % de codes détachés,
- 15 % de codes à doublets,
- 4 % de codes à triplets,
- 1 % de codes à connexions complexes.

Une étude entreprise par les P.T.T. françaises sur 4.830.000 lettres manuscrites relève la présence de :

- 76 % de codes détachés,
- 15 % de codes à doublets,
- 5 % de codes à triplets,
- 4 % de codes à connexions complexes.

L'identification globale de plusieurs chiffres chaînés par un classificateur ne peut être entreprise que sur certains doublets. Le nombre de combinaisons chiffre-chiffre, et les types de connexions, produisent trop de configurations pour que la modélisation soit possible.

L'approche la plus simple, pour identifier les chiffres connectés, est de les segmenter avant l'**étape de reconnaissance**. Cette **segmentation** peut être effectuée au niveau de l'image binaire ou sur le squelette. Nous avons préféré travailler au niveau de l'image binaire car l'information du contour de la forme est proche de la topologie de la forme. Le squelette est un mode de représentation trop difficile à analyser dans le cadre du cursif.

Une des techniques couramment employée est l'étude de l'histogramme des projections verticales cumulées des pixels d'une forme par rapport à une horizontale (figure (5.6.2.1)). La détection de la vallée la plus profonde au voisinage du milieu de la forme correspond à la position de la connexion. Une coupure est débutée à partir de cette abscisse. L'inconvénient de cette approche est la méthode de coupure, processus aveugle qui ne tient pas compte du contexte sémantique des formes qui côtoient la connexion.

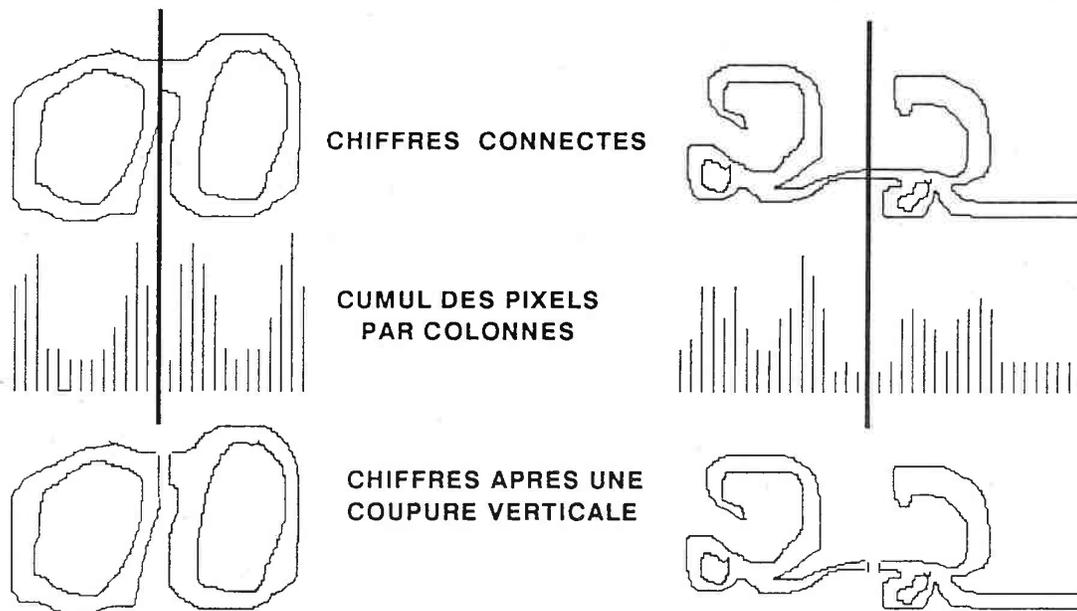


Figure 5.6.2.1 Segmentation par recherche d'un minimum de l'histogramme .

Pour venir à bout de ce problème, il est possible de générer plusieurs listes de candidats déconnectés à l'**étage de reconnaissance** en additionnant ou en supprimant la **zone de connexion**. Les hypothèses effectuées sur l'exemple de la figure (5.6.2.2) explique cette démarche.

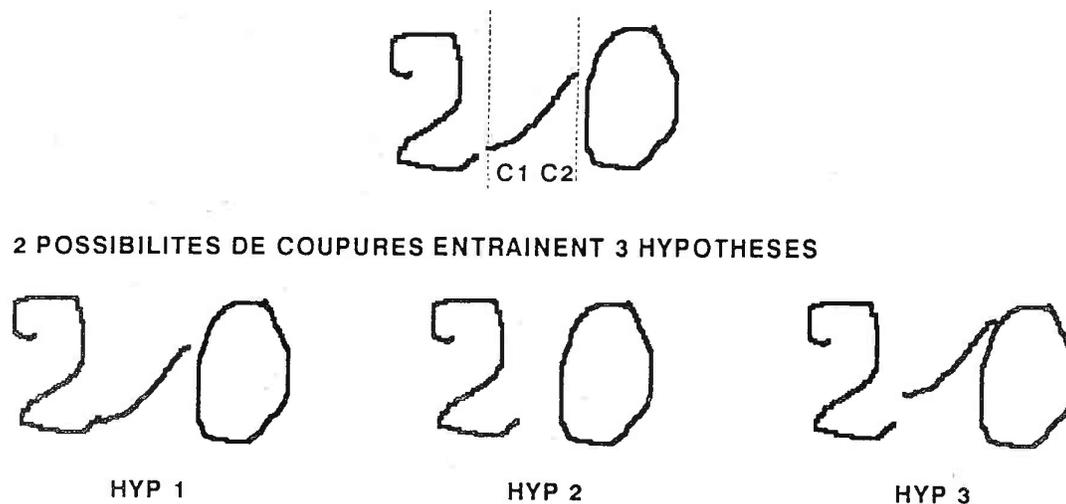


Figure 5.6.2.2 Coupures de chiffres connectés et générations d'hypothèses.

La technique utilisée par M. Shridhar et A. Badreldin (SHR 85), représentée sur la figure (5.6.2.3), est basée sur l'étude des contours externes des formes connexes. L'initialisation du traitement se fait par recherche du maximum du "profil bas" exécuté sur la partie inférieure du doublet. Le tracé de coupure part du bas de la forme à l'abscisse du pic et remonte en évitant de toucher le contour. Quand le contact est inévitable, la coupure progresse verticalement en évitant au maximum de rester au sein de la forme. Cette démarche très intéressante fait l'objet d'une réalisation d'un algorithme hiérarchisé qui procède par niveaux croissants de complexité. Cet algorithme s'adaptent aux configurations allant du caractère isolé à la connexion de deux chiffres inclinés.

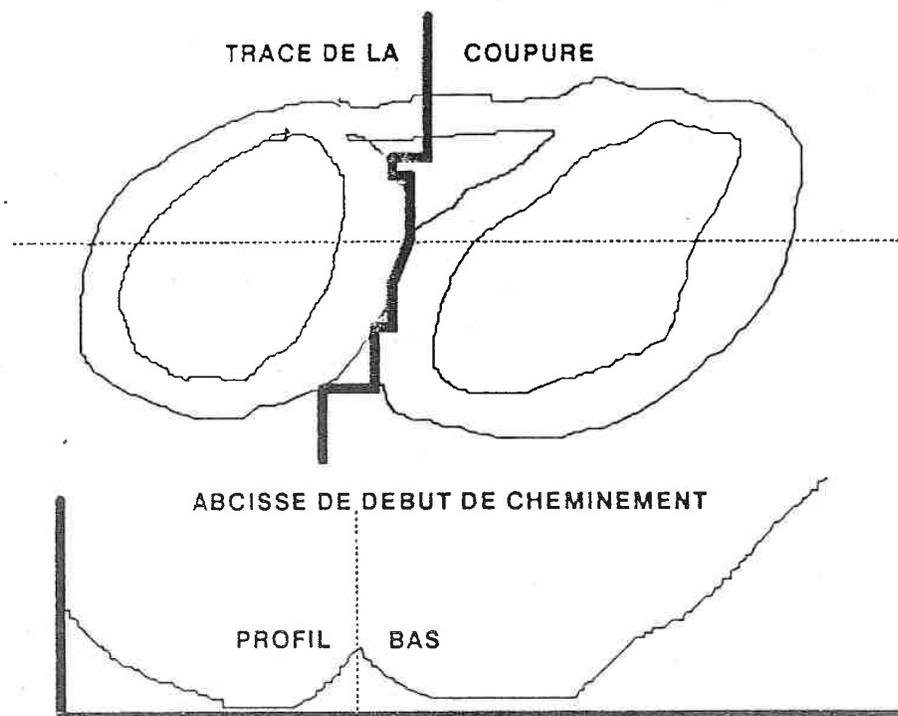


Figure 5 .6.2.3 Détection de pics de connexions dans le profil bas de la forme.

5.6.3 - CONCLUSION.

Les deux démarches présentées sont néanmoins dédiées à la coupure de connexions simples. Dans le cas d'un fort recouvrement des caractères, ce que nous pouvons voir sur la figure (5.6.3.1), la détection des vallées de l'histogramme ou le cheminement à travers la connexion ne permet pas de séparer correctement les caractères.

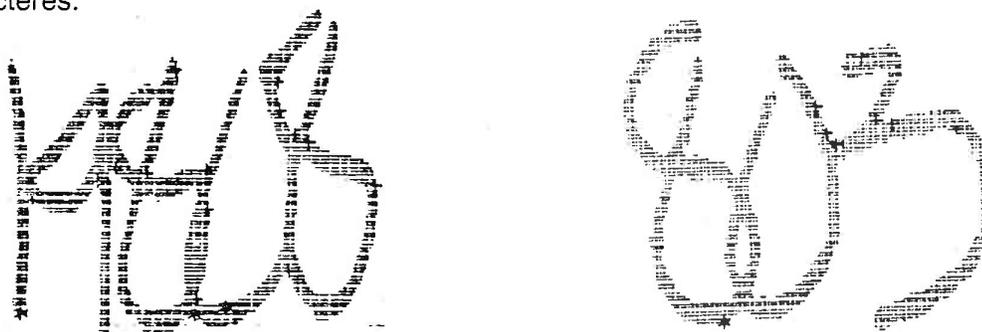


Figure 5.6.3.1 Configurations de connexions complexes.

Pour faire face aux connexions complexes, une segmentation est plus efficace si la forme ou le type de connexion est plus ou moins reconnu. Certains chercheurs ont appliqué ce raisonnement aux données cursives de tablettes graphiques et de bases manuscrites restreintes à quelques mots.

Un des schémas de reconnaissance nécessite un suivi de la description de la forme. Simultanément le système compare le cumul des primitives aux modèles de chiffres. Lorsqu'une partie de cette forme est reconnue comme proche d'un modèle, il y a **segmentation** de celle-ci et le processus continue sur le reste du mot. Dans l'étude des codes postaux, les chiffres sont le plus souvent déconnectés. Cette démarche devrait alors, pour des contraintes de temps, s'appliquer de manière parallèle aux cinq ou aux neuf chiffres. Une détection sommaire de la forme connexe apparaît nécessaire pour utiliser ce module de manière facultative.

Dans le cadre du projet de lecteur de code français, il est apparu plus fiable et plus rapide d'utiliser un **classificateur** pour certaines connexions fréquentes. Le **doublet** ou le **triplet** devient une **forme de base** équivalente à un chiffre.

Dans le courrier français, 27 % des chiffres connectés sont des doubles-zéros "00" et 5 % des triples-zéros "000".

5.6.4 - METHODOLOGIE DE LA DECONNEXION.

Les résultats prometteurs obtenus par la **description** de l'image nous ont amené à récidiver l'approche par une **détection de points caractéristiques**. Très souvent, une connexion produit sur le contour de la forme un accident anguleux (recouvrement de formes) ou un décalage net au niveau des zones horizontales de celui-ci (prolongation du tracé). Par conséquent, la détection des point anguleux du contour permet théoriquement de définir un couple de points disposés de part et d'autre de la connexion.

L'approche embryonnaire que nous avons développée, dont le schéma fonctionnel est représenté a la figure (5.6.4.1), s'appuie sur le postulat suivant : une zone de coupure peut être localisée, soit par un point pertinent du contour, soit par un couple de points issus des parties supérieures et inférieures de ce contour.

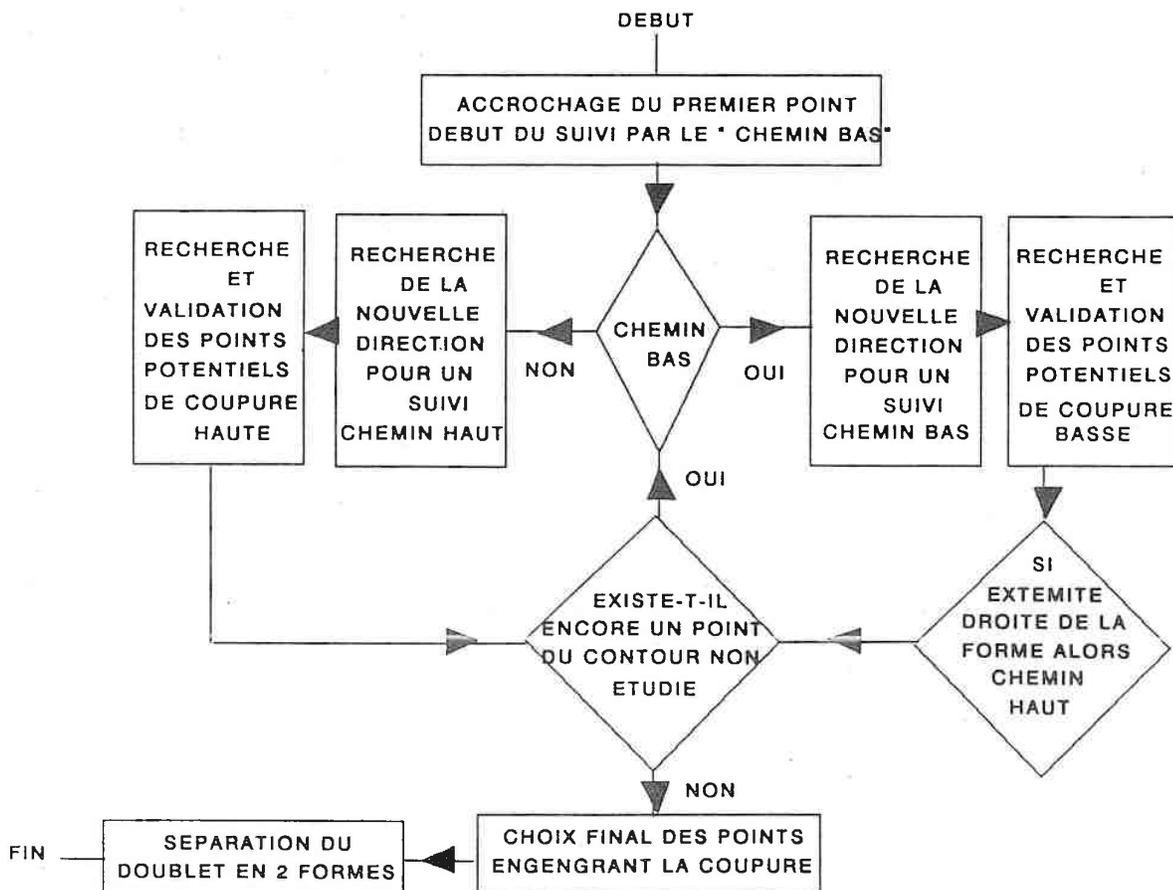
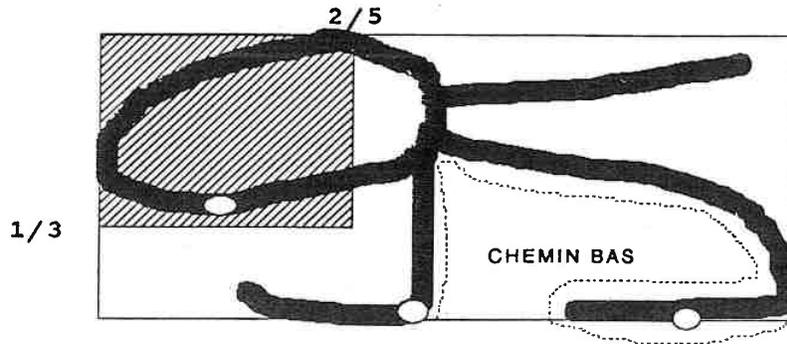


Figure 5.6.4.1 Schéma fonctionnel de l'algorithme de déconnexion des doublets.

Selon les règles de la figure (5.6.4.2), la forme est abordée par la partie basse, gauche (premier **point sud**), selon des critères d'intérêts de parcours.



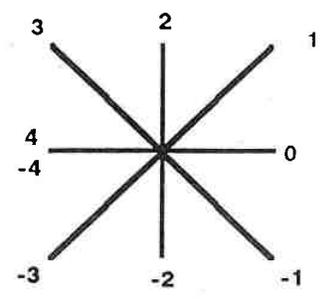
LA ZONE ZEBREE EST PROSCRITE POUR EVITER D'EFFECTUER DES COUPURES ERRONEES SUR CERTAINS DOUBLETS.
LE PROCESSUS DE SUIVI DEBUTE PAR LE " CHEMIN BAS " AU NIVEAU DU PREMIER POINT SUD DANS LA ZONE AUTORISEE.

Figure 5.6.4.2 Accrochage du doublet et tracé du chemin bas.

Le parcours de la forme se fait par l'extérieur, de manière trigonométrique et selon la connexité 8 pixels voisins, et génère un "chemin bas" et un "chemin haut". Le contexte spatial de tout point du contour permet de définir des points potentiels de coupure haute ou basse. Ces points sont classés en deux ordres en fonction de leur importance. La figure (5.6.4.3) présente pour chaque configurations recherchées un exemple typique.

Le cadre de la forme est ensuite découpé en espaces inégaux qui modélisent les zones de coupures idéales pour un certain nombre de doublets. Pour chaque espace, représentés sur la figure (5.6.4.4), des règles fondées sur la topologie de la forme et la position des autres points potentiels permettent de décider si un ou plusieurs points candidats sont des **points de coupure**.

code de FREEMAN



POINTS BAS POTENTIELS DE 1er ORDRE

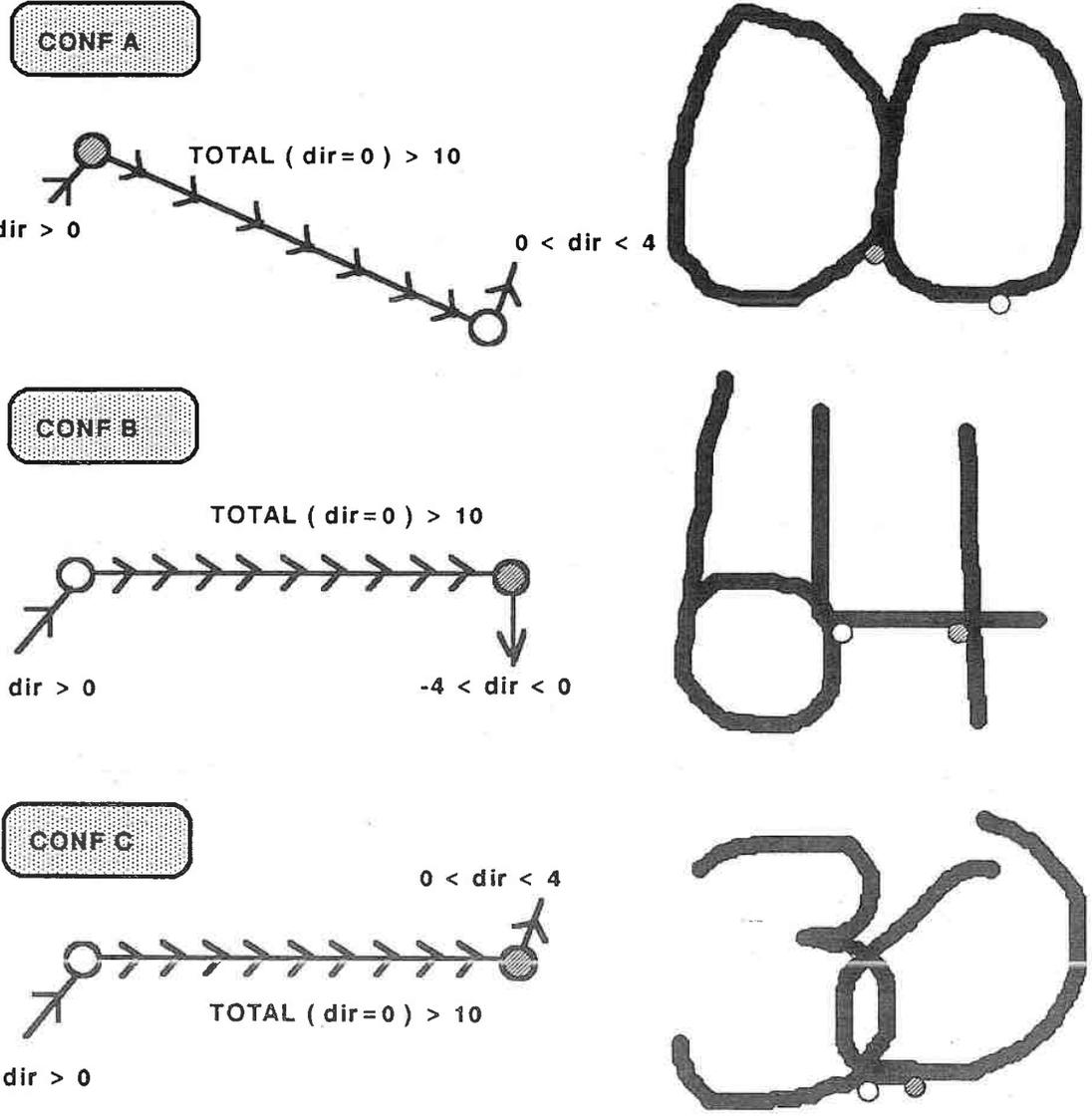


Figure 5.6.4.3 Détection des points bas potentiels de coupure basse (chemin bas)

POINTS HAUTS POTENTIELS DE 1er ORDRE

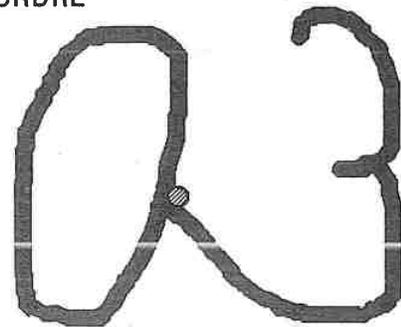
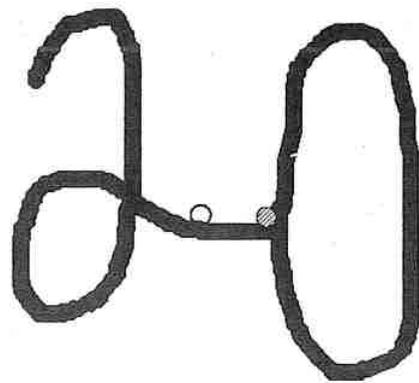
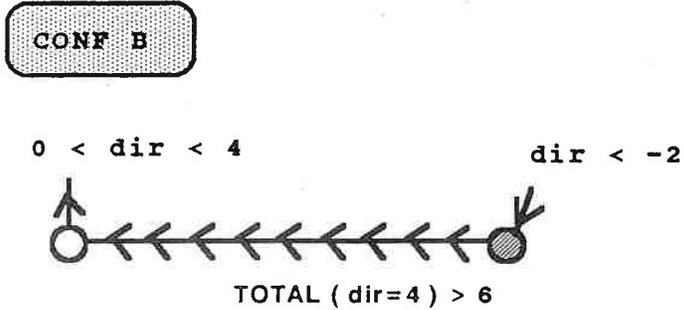
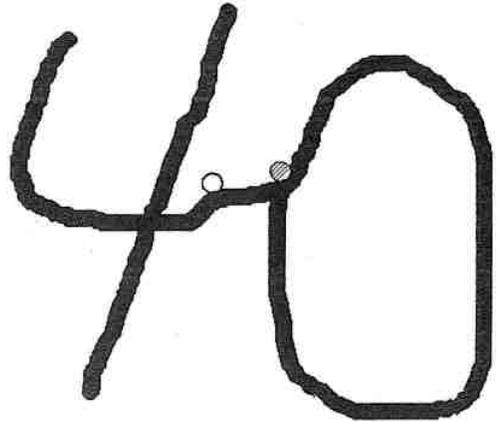
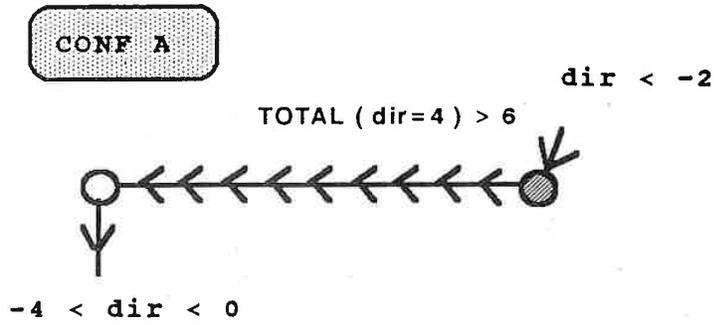


Figure 5.6.4.3 Détection des points hauts potentiels de coupure haute (chemin haut)

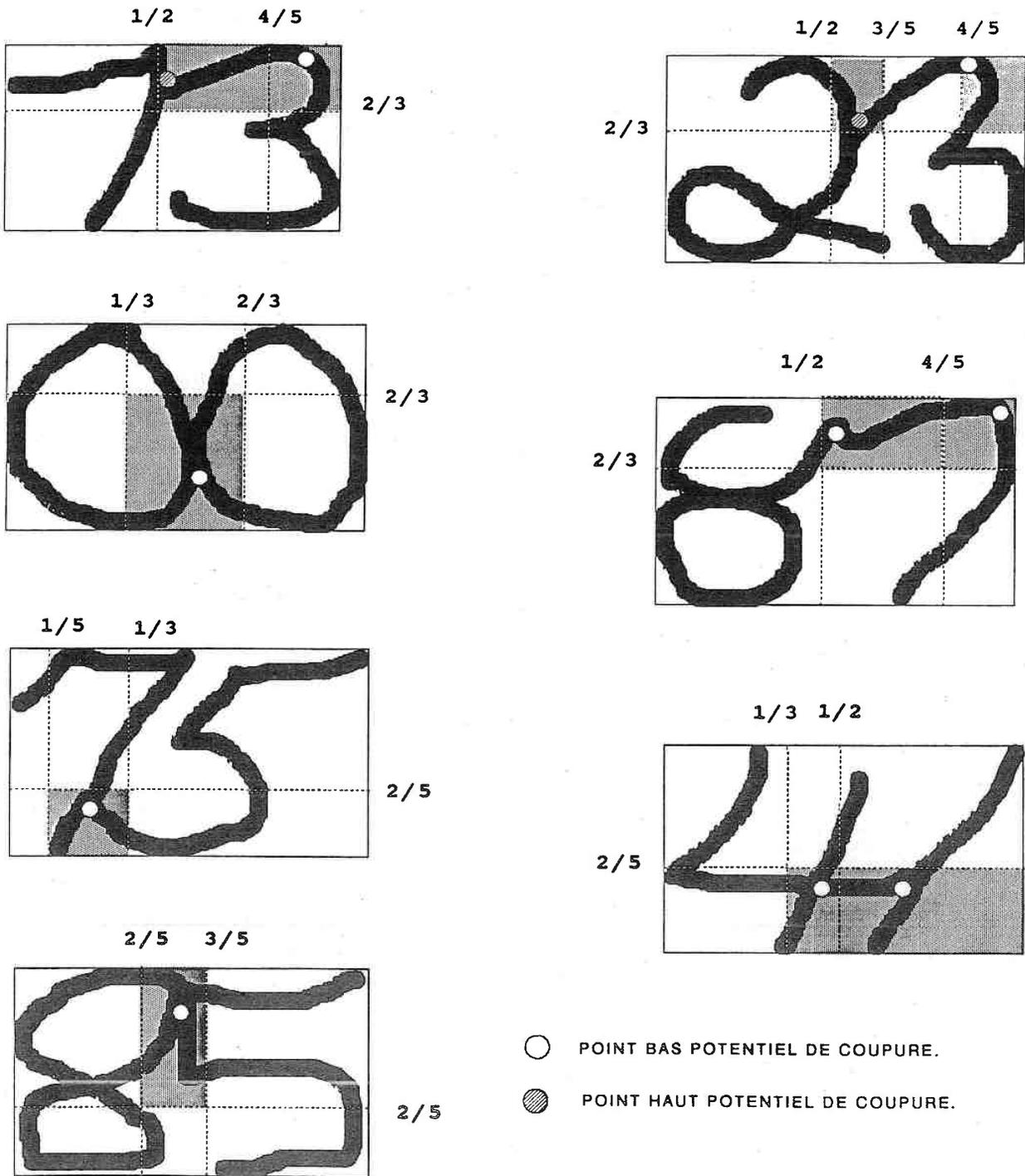


Figure 5.6.4.4 Exemples de modélisations de zones de connexions..

Le module travaille de préférence avec les **points inférieurs de premier ordre** et recherche ensuite un **point supérieur**, de premier ordre si possible. Cette préférence est expliquée par l'exemple de la figure (5.6.4.5)

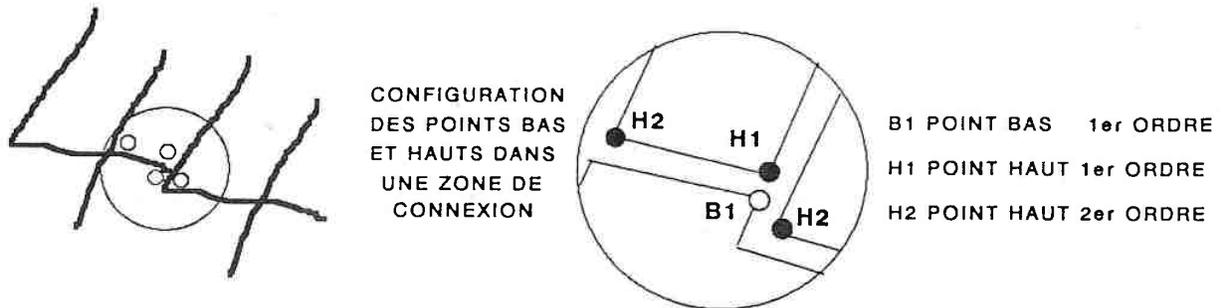
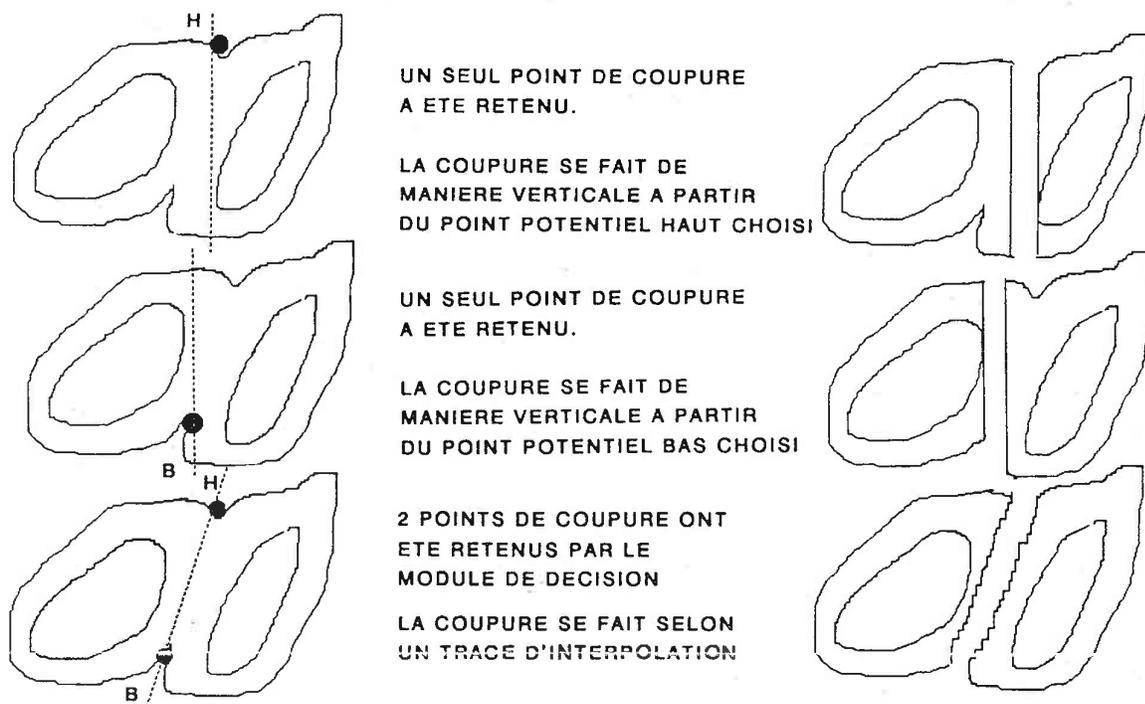


Figure 5.6.4.5 Ordre de priorités dans le choix des points de coupures.

Lorsque le module décision choisit le lieu de segmentation, la coupure se fait verticalement à partir d'un unique point ou par l'interpolation de deux points comme on peut le voir sur la figure (5.6.4.6).



- figure (5.6.4.7) : Lorsque la coupure est achevée, deux ou trois formes sont désormais identifiables. Le module recalcule les caractéristiques de chacune de ces parties.

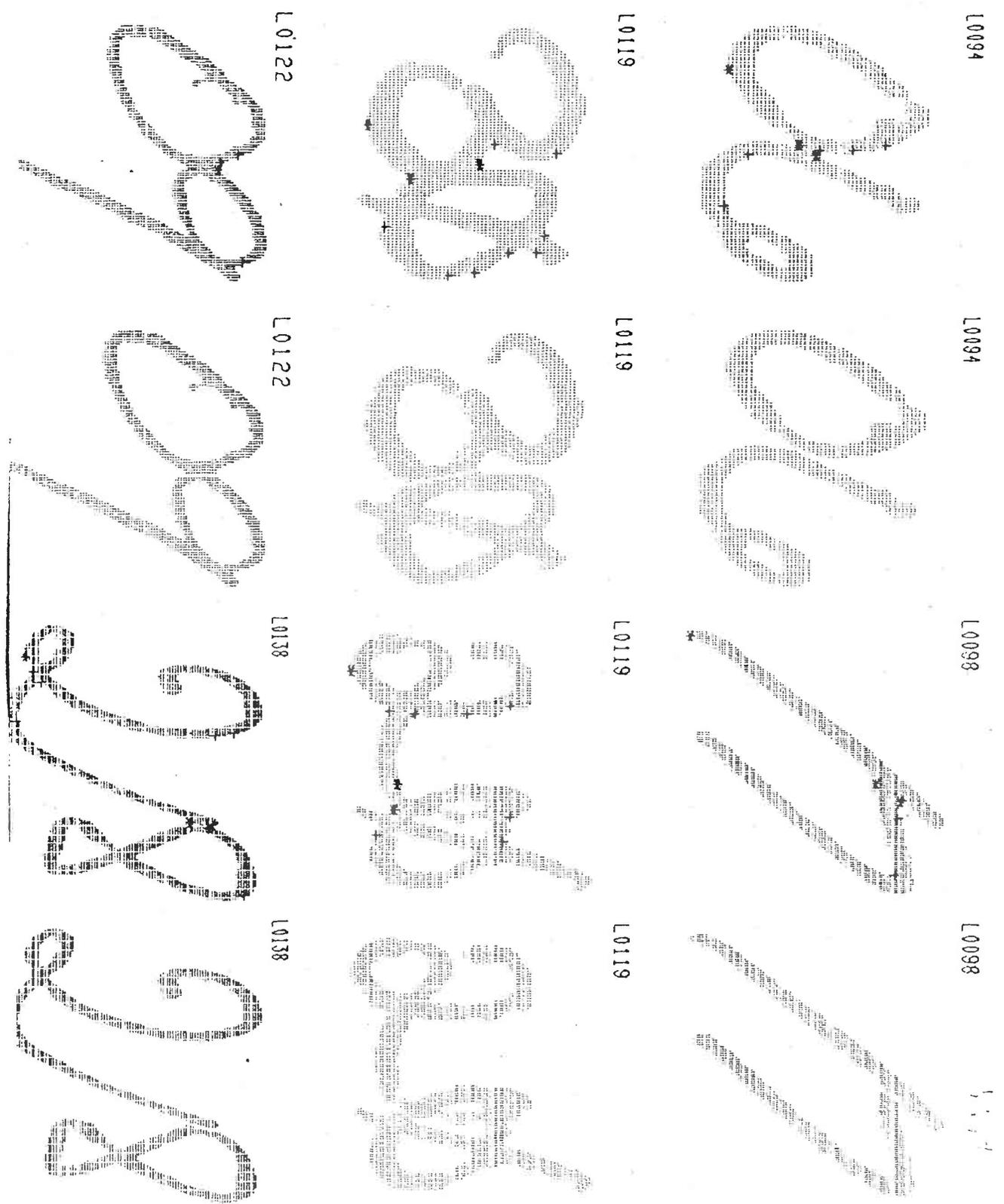


Figure 5.6.4.7 Exemples de résultats de coupures de doublets.

5.6.5 - EVALUATION.

Cette approche, élaborée à partir de cas d'espèces, fournit des résultats parfois surprenants quand le type de connexion est marginal. Certaines confusions proviennent des configurations antagonistes. Les points de coupures correctement trouvés sont éliminés par les règles de décision élaborées sur un certain modèle typique. Le faible nombre d'échantillons ne nous a pas permis de cerner correctement les classes de problèmes.

- Résultats de la "base d'apprentissage" (175 doublets) :

- 87 % des doublets sont correctement séparés ou contiennent des barbelures non dangereuses pour l'identification.

- 13 % des doublets ne sont pas séparés ou sont mal séparés.

- Résultats de la "base d'évaluation" (156 doublets) :

- 73 % des doublets sont correctement séparés ou contiennent des barbelures non dangereuses pour l'identification.

- 27 % des doublets ne sont pas séparés ou sont mal séparés.

L'exemples présenté sur la figure (5.6.5.1) montrent le besoin d'affiner les coupures réalisées trop brutalement. Une étude du contexte et une recherche plus poussée des relations entre points inférieurs et supérieurs devrait permettre de récupérer les connexions complexes (imbrication des formes et multi-connexions).

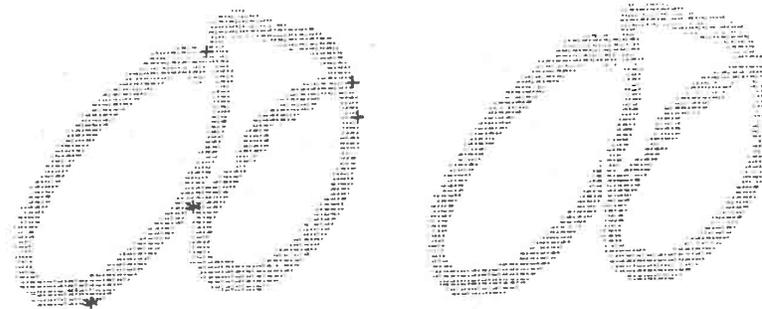


Figure 5.6.5.1 détection et coupures erronées des zones de connexions.

En ce qui concerne les confusions de points, il faut améliorer la démarche de décision qui étudie actuellement les zones du cadre de manière séquentielle et aveugle.

Les segments de liaison, souvent placés dans la partie basse, peuvent être éliminés. Les "**classificateurs topologiques et sémantiques**" sont capables d'étudier un chiffre incomplet même si la base du chiffre est assimilée comme segment de liaison. En effet, le module de reconnaissance doit parfois analyser des chiffres isolés qui sont naturellement incomplets. Si le segment de liaison n'a pas une forte signification pour l'identification, son absence aura peu d'influence dans la décision.

5.7 - ASSOCIATION DES FORMES.

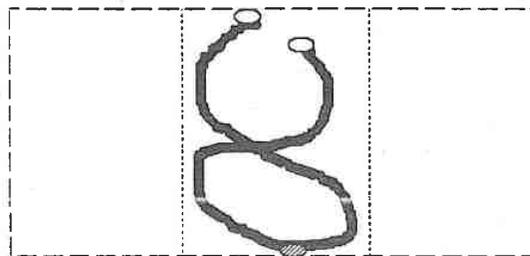
5.7.1 - INTRODUCTION.

A la sortie de la **description**, les formes de l'image sont décrites et représentées par des paramètres. L'identification du code postal ne pouvant s'effectuer sur la totalité des formes connexes, il est nécessaire de les regrouper en fonction de critères géométriques. Les ensembles ainsi créés, que nous appellerons **blocs** ou **pseudo-lignes d'information**, faciliteront la tâche de localisation du code postal, de l'état, et de la ville. Cet algorithme d'**association** des formes utilise les proximités géométriques pour réunir les lettres d'un même mot, les chiffres d'un même code postal, et sépare les lignes enchevêtrées dans le sens vertical.

5.7.2 - NOTION DE "VOISINAGE D'ASSOCIATION".

Pour éviter les associations erronées, chaque forme se voit attribuée une **zone d'influence**. Ce **voisinage d'association** est calculé en fonction du nombre de **points extrêmes hauts (nord)** et bas (**sud**) et des caractéristiques géométriques de la forme.

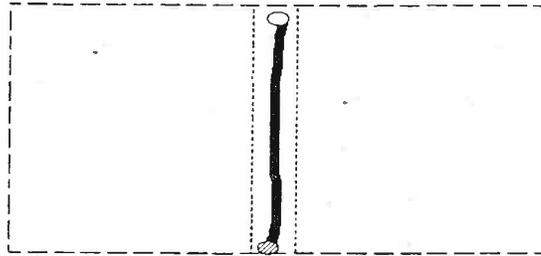
- Forme simple, type **caractère isolé** (figure (5.7.2.1)).



VOISINAGE D'ASSOCIATION D'UN CARACTERE :
 HAUTEUR = HAUTEUR DU CADRE DE LA FORME
 LARGEUR = LARGEUR DU CADRE * COEFF

Figure 5.7.2.1 Calcul de la zone d'influence d'un caractère commun.

- Forme allongée (1 ou i), type barre verticale (figure (5.7.2.2)).



VOISINAGE D'ASSOCIATION D'UNE VERTICALE :

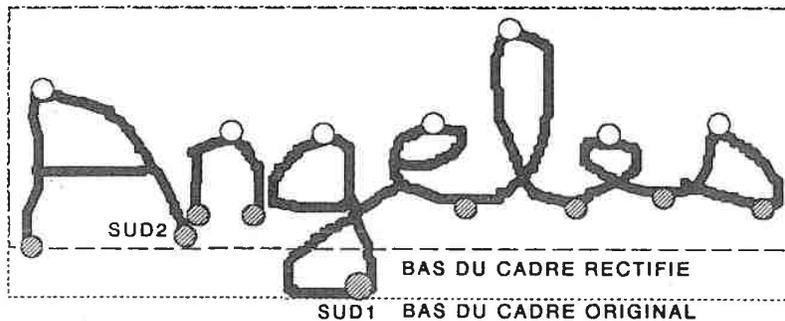
HAUTEUR = HAUTEUR DU CADRE DE LA FORME

LARGEUR = HAUTEUR DU CADRE * COEFF

Figure 5.7.2.2 Calcul de la zone d'influence d'une barre verticale.

- Forme complexe, type **mot long** ou **graphisme** figure (5.7.2.3).

i) Les jambages de consonnes f, g, j, p, q et y sont dangereux pour l'**association** et sans intérêt géométrique. Le cadre du mot est remonté au niveau de la base des autres caractères.

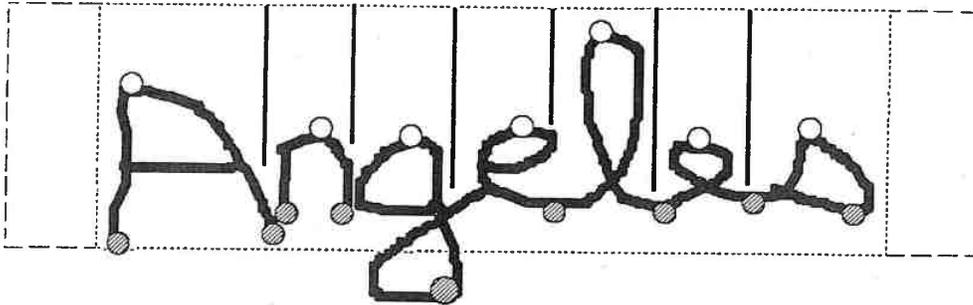


RECTIFICATION DE LA BASE DU CADRE POUR EVITER LES ASSOCIATIONS PRODUITES PAR LES JAMBAGES DES CONSONNES.

HAUTEUR = HAUTEUR - DELTA

SI IL EXISTE AU MOINS 3 POINTS SUD DANS LE 1/4 INFERIEUR DE LA FORME, DELTA = 0 DEVIENT SUD2 - SUD1.

ii) Les caractéristiques d'un mot peuvent être considérées comme étant le cumul des caractéristiques de chaque caractère. Ce postulat permet de faire l'approximation du nombre d'éléments chaînés et d'en déduire un **voisinage d'association moyen**.



VOISINAGE D'ASSOCIATION D'UN MOT OU D'UN GRAPHIQUE :
 HAUTEUR = HAUTEUR RECTIFIEE DU CADRE
 LARGEUR = LARGEUR * (1 + 2 / NOMBRE DE POINTS SUD)

Figure 5.7.2.3 Calcul de la zone d'influence d'un mot ou d'un graphique et élimination des jambages des mots.

5.7.3 - PRINCIPE DE LA PROCEDURE D'ASSOCIATION.

Les formes, en sortant de la **description**, sont numérotées de gauche à droite et de bas en haut.

Le processus d'**association** se fait de manière croissante en suivant la numérotation. Au départ, l'algorithme recherche une forme libre, non incluse dans un bloc. Quand cette forme est trouvée, elle devient le premier élément d'un nouveau bloc. Cet élément devient aussi maître du processus d'association. Ensuite, celui-ci cherche dans sa zone d'influence d'autres formes (libres), pour les intégrer et les rendre esclaves du bloc. Quand cet élément maître ne peut plus associer, c'est au tour des formes esclaves de devenir maître, et ainsi de suite. Lorsque l'élément maître n'a pas de voisin à associer et qu'il n'y a plus de formes esclaves dans le bloc, le bloc est clos. Le système redevient maître de l'association et recherche alors l'élément non inclus et de plus petit numéro susceptible de générer un nouveau bloc.

Chaque forme incluse apporte ses caractéristiques au bloc au cours de la croissance de celui-ci. Les paramètres dynamiques du bloc sont:

- positions géométriques dans l'image (cadre, base virtuelle),
- nombre d'éléments inclus,
- surface, hauteur, largeur,
- cumul des étiquettes des formes incluses (caractères, mot, ponctuation ...).

5.7.4 - REGLES D'ASSOCIATION ET CONTROLE DE L'INCLINAISON DES LIGNES ET DES MOTS.

Certaines lettres, comme celles représentées sur la figure (5.7.4.1), sont riches en dessins parasites, en caractères brisées, et en mots entrelacés. Une association erronée peut être fatale pour la reconnaissance de la zone d'intérêt **code-état-ville**. Pour aboutir à un algorithme fiable, il est nécessaire d'établir des règles d'associations adaptées au type de courrier et différentes pour chaque catégorie de formes. L'expérimentation sur une large base d'images nous a permis de définir les critères empiriques présentés dans les figures (5.7.2.2), (5.7.2.3), (5.7.2.4).

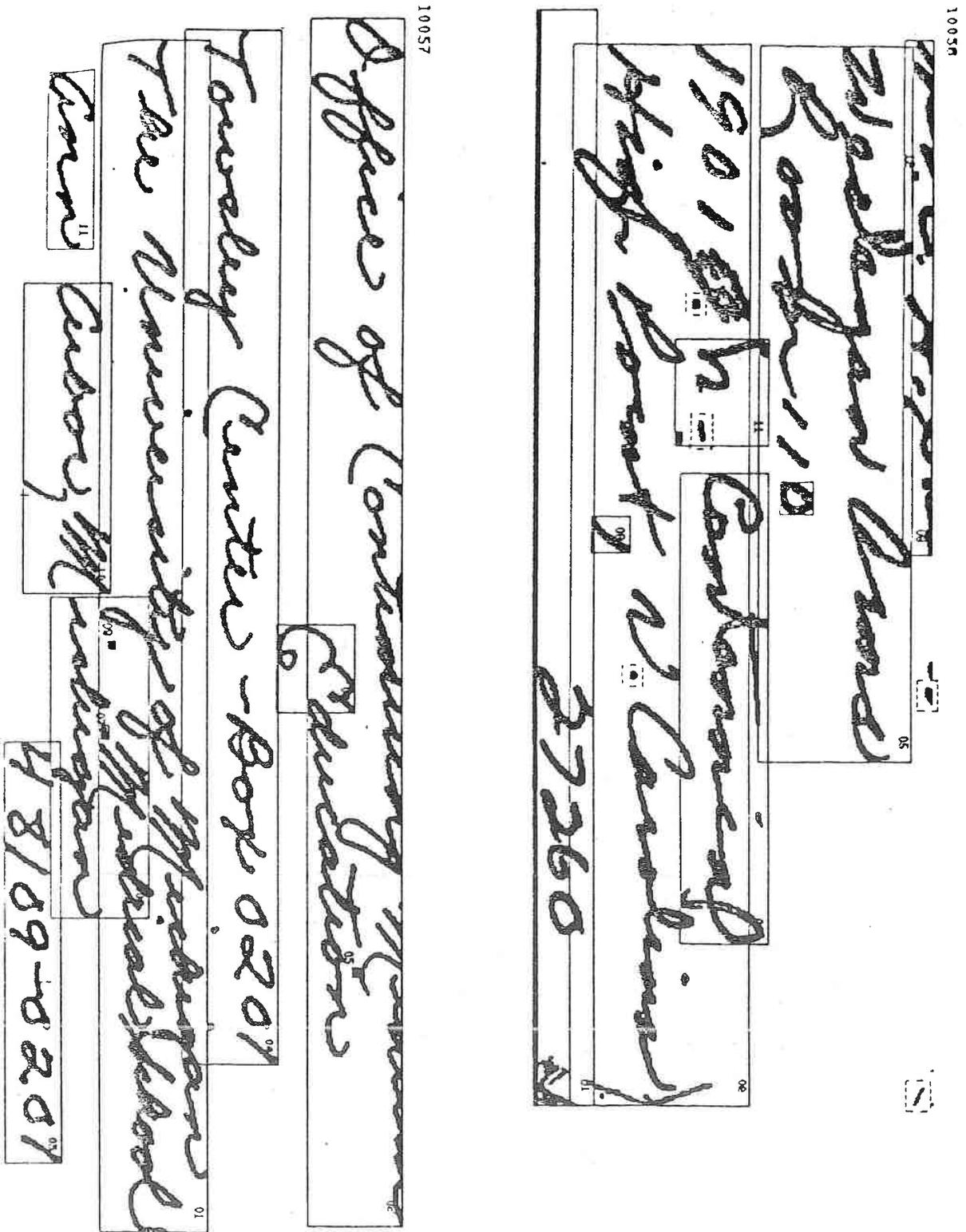


Figure 5.7.4.1 Documents où l'information est difficile à localiser. lignes interpénétrées et nombreuses connexions

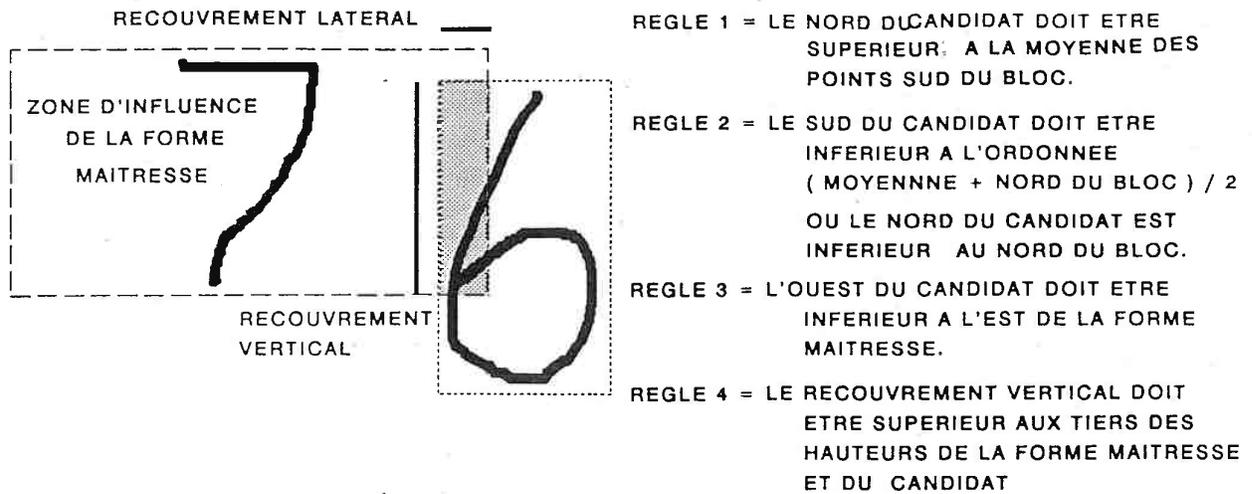


Figure 5.7.4.2 Associations de deux caractères.

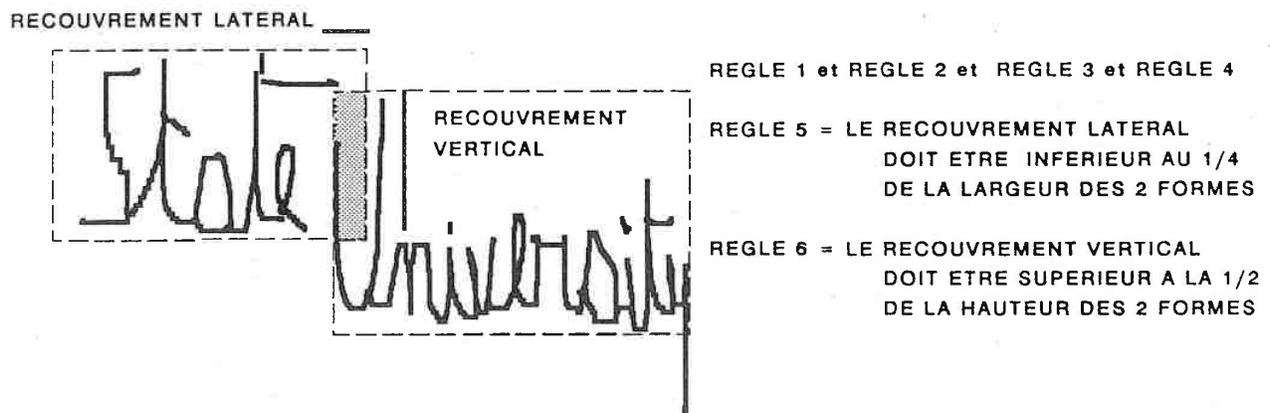


Figure 5.7.4.3 Association de deux mots.

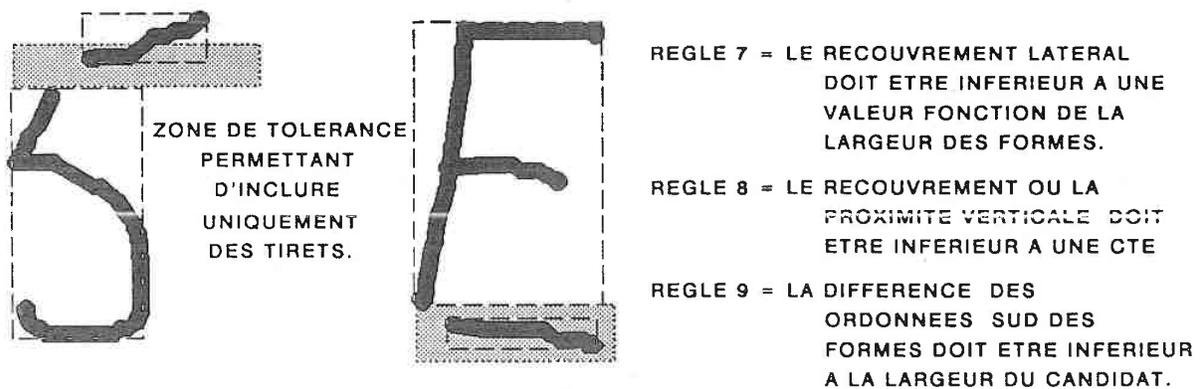


Figure 5.7.4.4 Association des formes horizontales au bloc étudié.

- Calcul dynamique de la **base virtuelle** du bloc (figure (5.7.4.5)).

Pour contrôler l'inclinaison de la ligne, à chaque inclusion de forme, le système recalcule la référence inférieure du bloc. Ce calcul tient compte de la pente en pondérant fortement les paramètres du dernier élément inclus.

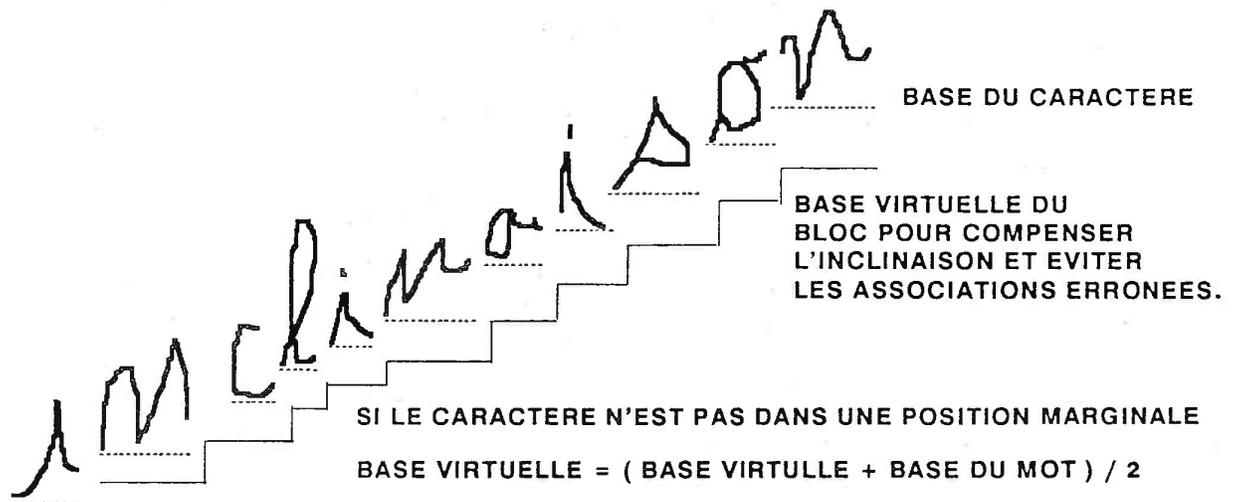
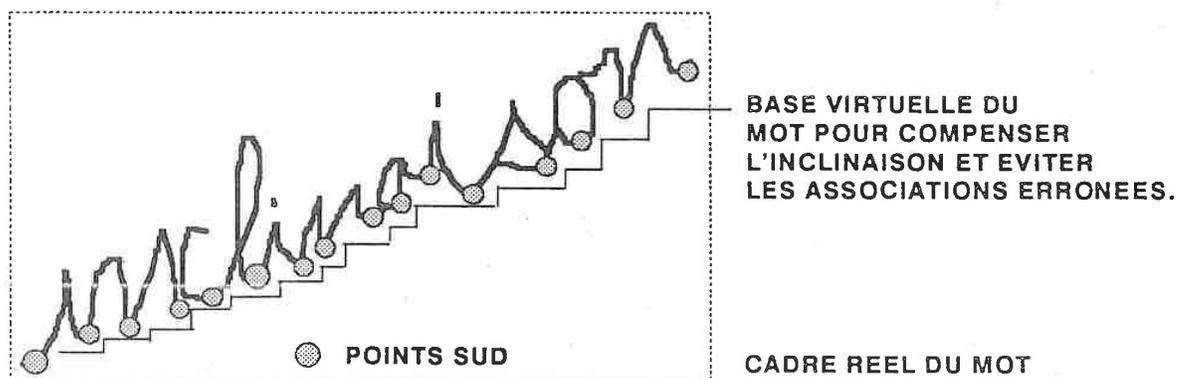


Figure 5.7.4.5 Calcul dynamique de la base virtuelle d'un bloc.

- Calcul de la base virtuelle d'un mot (figure (5.7.4.6)).

Une procédure identique à la précédente est appliquée au mot. Le calcul est effectué sur les points extrêmes bas (sud), choisis suffisamment espacés les uns des autres.



SI LA FORME EST UN MOT, LES BASES DE CARACTERES SONT LES POINTS SUD

$$BASE VIRTUELLE = (BASE VIRTUELLE + ORDONNEE DU POINT SUD DU MOT) / 2$$

Figure 5.7.4.6 Calcul de la base virtuelle d'un mot.

5.7.5 - EVALUATION.

Deux bases de 1000 lettres ont été formées. Une base dite d'**apprentissage** permet d'affiner les résultats des tests effectués. Lorsque ceux-ci sont satisfaisants, un test de validation est fait sur la **base d'évaluation**.

Le but de la procédure d'**association** est de conserver le code postal entier, l'état et la ville si possible (figures 5.7.5.1)). Tout autre ligne d'information, bien ou mal associée, n'est pas comptabilisée dans les résultats d'évaluation de l'algorithme.

- Performances obtenues sur les 1.000 lettres de la **base d'apprentissage** :

- 7 erreurs sont d'origine algorithmique :

- 4 erreurs sont des configurations marginales non étudiées,
- 3 images contiennent un 5 brisé dont la barre n'est pas incluse dans le même bloc.

- 24 erreurs proviennent des traitements antérieurs :

- 10 soulignements n'ont pas été détectés,
- 14 images sont de mauvaises qualités ou contiennent des lignes interconnectées.

- Total : 31 images sont incorrectement associées.

- Performances obtenues sur les 1000 lettres de la **base d'évaluation** :

- 7 erreurs sont d'origine algorithmique :

- 2 erreurs sont des configurations marginales non étudiées,
- 5 images contiennent un 5 brisé dont la barre n'est pas incluse dans le même bloc.

- 32 erreurs proviennent des traitements antérieurs :

- 17 soulignements n'ont pas été détectés,
- 15 images sont de mauvaises qualités ou contiennent des lignes interconnectées,

- Total : 39 images incorrectement associées.

Dept. of Microbiology and Immunology
 Medical School and
 V. H. Rackham School of Graduate Studies
 University of Michigan
 Ann Arbor, MI 48109
 U.S.A.

IMAGE NUMBER : bd_1321

Description stage

Dept. of Microbiology and Immunology
 Medical School and
 V. H. Rackham School of Graduate Studies
 University of Michigan
 Ann Arbor, MI 48109
 U.S.A.

IMAGE NUMBER : bd_1321

Association stage

✓ Norwich Road, Ipswich
 3823 Walsworth Rd
 Horton, Ohio
 44203
 Alto Road

IMAGE NUMBER : bd_1334

Description stage

✓ Norwich Road, Ipswich
 3823 Walsworth Rd
 Horton, Ohio
 44203
 Alto Road

IMAGE NUMBER : bd_1334

Association stage

Figure 5.7.5.1 Resultats de la description de l'image et association des formes en blocs d'information;

OPEN NETWORK
 DEPT SW
 BOX 18666
 DENVER, CO 80208
 U S A

OPEN NETWORK
 DEPT SW
 BOX 18666
 DENVER, CO 80208
 U S A

ZION yellowstone
 Grand Junction
 CO USA
 81503

ZION yellowstone
 Grand Junction,
 CO USA
 81503

Figure 5.7.5.1 Resultats de la description de l'image et association des formes en blocs d'information;

Susie Johnson
1365 Detroit St
Denver, Col
80206

IMAGE NUMBER : bd_1245

Description stage

Susie Johnson
1365 Detroit St
Denver, Col
80206

IMAGE NUMBER : bd_1245

Association stage

See Instructions on Back Before Opening
Thomas Hill
It is out of stock, V4
05091

IMAGE NUMBER : bd_1259

Description stage

See Instructions on Back Before Opening
Thomas Hill
It is out of stock, V4
05091

IMAGE NUMBER : bd_1259

Association stage

Figure 5.7.5.1 Results de la description de l'image et association des formes en blocs d'information;

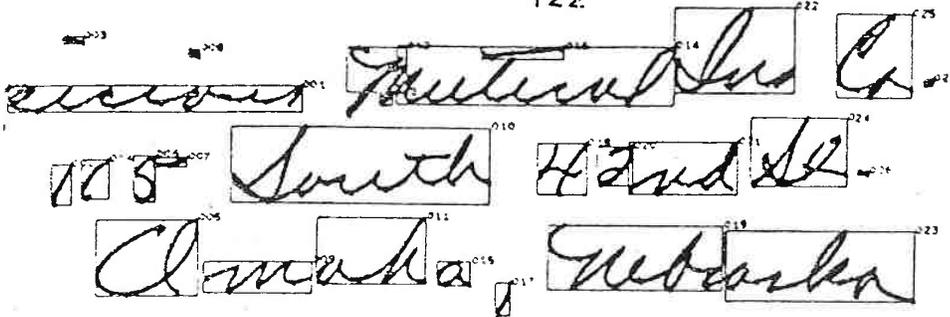


IMAGE NUMBER : bd_1075

Description stage

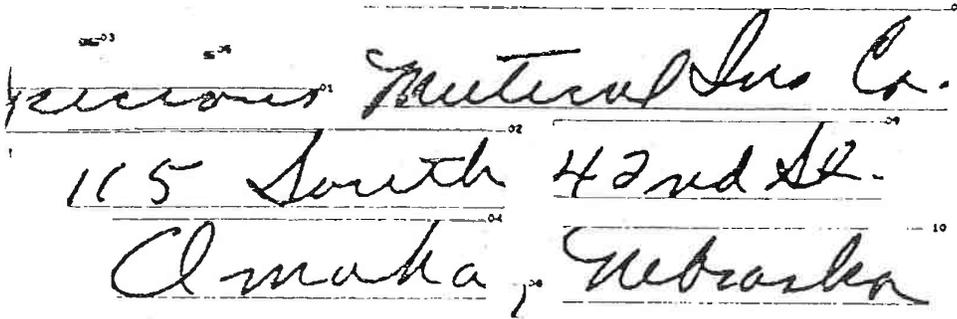


IMAGE NUMBER : bd_1075

Association stage

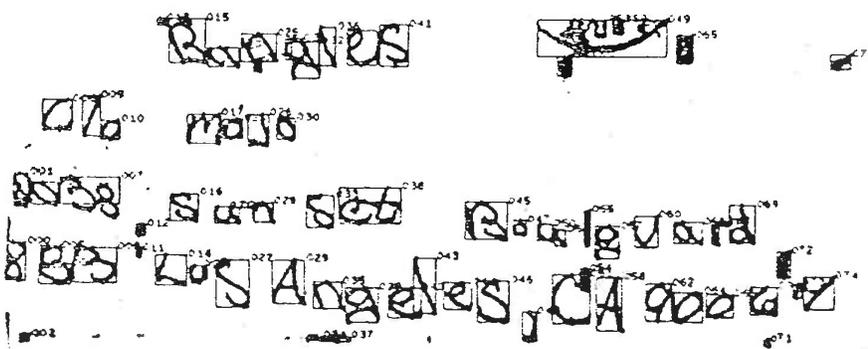


IMAGE NUMBER : bd_1080

Description stage

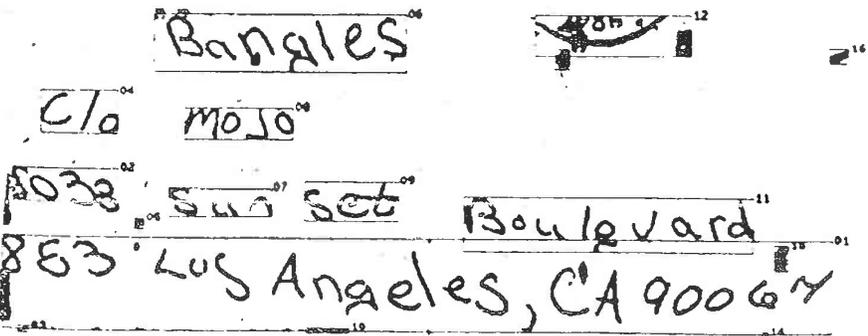


IMAGE NUMBER : bd_1080

Association stage

Figure 5.7.5.1 Resultats de la description de l'image et association des formes en blocs d'information;

IMAGE NUMBER : bd_1151

Description stage

10018 PIRATE'S TRAIL
 REMEMBERSVILLE, OHIO
 7502

IMAGE NUMBER : bd_1151

Association stage

10018 PIRATES TRAIL
 REMEMBERSVILLE, OHIO
 7502

IMAGE NUMBER : bd_1157

Description stage

Miss Mary Callahan, O.S.B.
 P.O. Box 488
 Callahan, Ala. 35056-0488

IMAGE NUMBER : bd_1157

Association stage

Miss Mary Callahan, O.S.B.
 P.O. Box 488
 Callahan, Ala. 35056-0488

Figure 5.7.5.1 Resultats de la description de l'image et association des formes en blocs d'information;

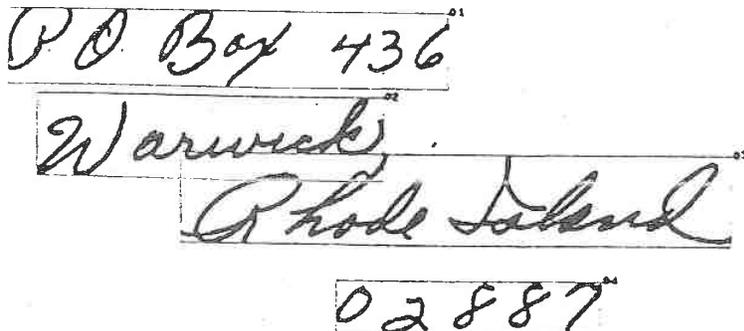
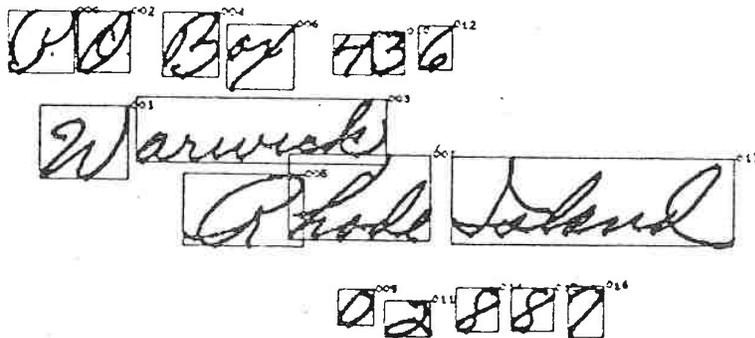
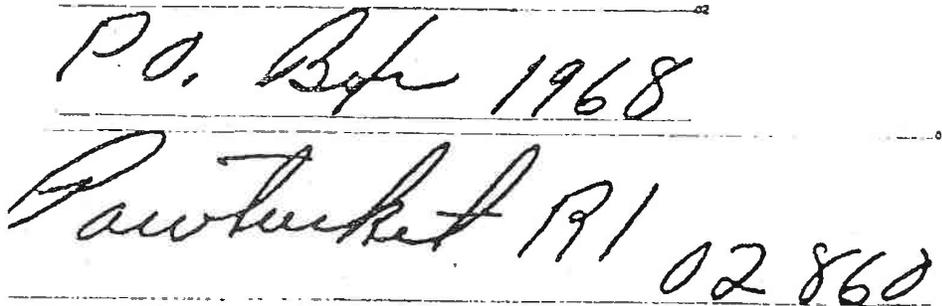
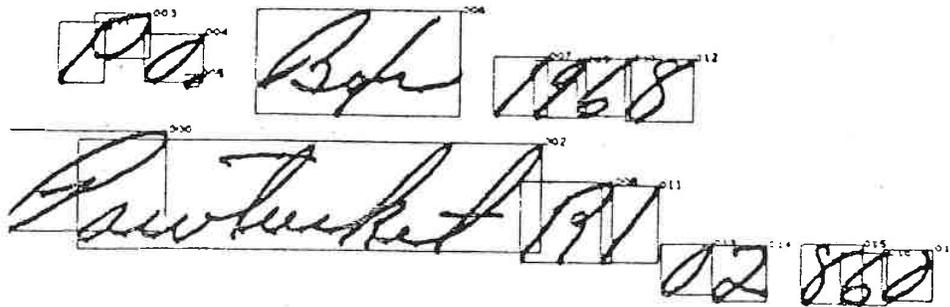


Figure 5.7.51 Resultats de la description de l'image et association des formes en blocs d'information;

Description stage

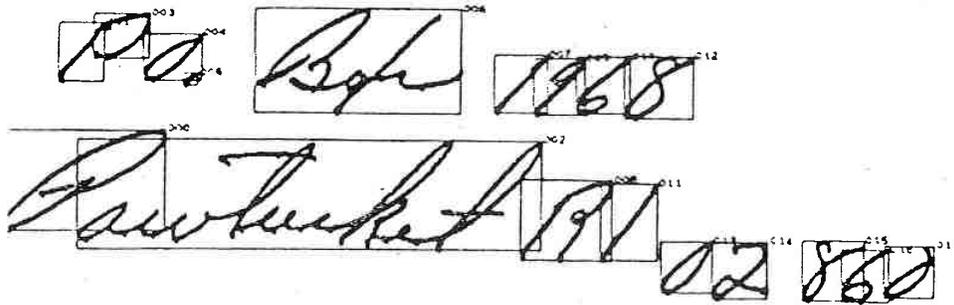


IMAGE NUMBER : bd_1387

Association stage

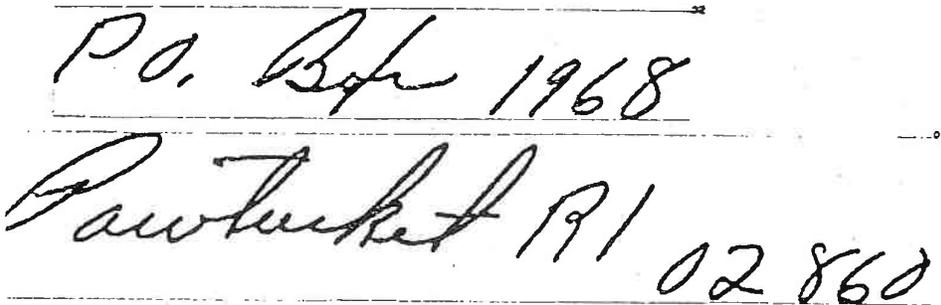


IMAGE NUMBER : bd_1388

Description stage

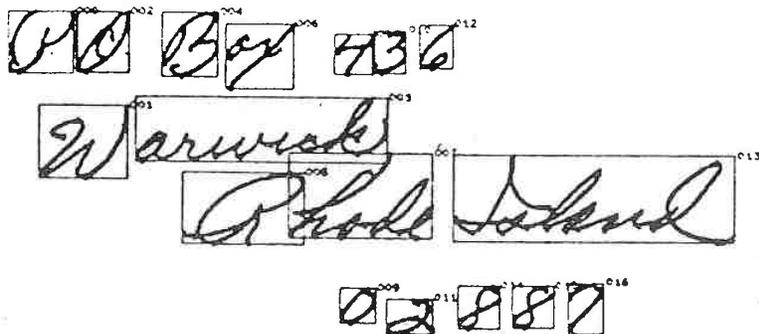


IMAGE NUMBER : bd_1388

Association stage

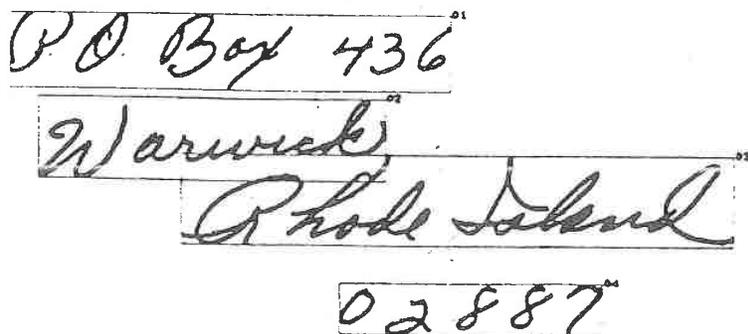


Figure 5.7.5.1 Resultats de la description de l'image et association des formes en blocs d'information;

Mm Margaret Bennett
 2720 E DuPont Ave
 Belle WU 25

IMAGE NUMBER : bd_1166

Description stage

Mm Margaret Bennett
 2720 E DuPont Ave
 Belle WU 25

IMAGE NUMBER : bd_1166

Association stage

English Dept
 American University
 4400 Massachusetts Ave
 Washington, D.C.
 20016

IMAGE NUMBER : bd_1175

Description stage

English Dept
 American University
 4400 Massachusetts Ave
 Washington, D.C.
 20016

IMAGE NUMBER : bd_1175

Association stage

Figure 5.7.5.1 Resultats de la description de l'image et association des formes en blocs d'information;

Equitable Life Assurance Society
 up Benefita Manager
 2 Box 11081
 Merrillville, Indiana 46411-1081

IMAGE NUMBER : bd_1184

Description stage

Equitable Life Assurance Society
 up Benefita Manager
 2 Box 11081
 Merrillville, Indiana 46411-1081

IMAGE NUMBER : bd_1184

Association stage

HARRY A BECKER
 125 GULF OF MEXICO DRIVE
 2PT # 405
 LONG BOAT KEY FL 33548

IMAGE NUMBER : bd_1202

Description stage

HARRY A BECKER
 125 GULF OF MEXICO DRIVE
 2PT # 405
 LONG BOAT KEY FL 33548

IMAGE NUMBER : bd_1202

Association stage

Figure 5.75.1 Resultats de la description de l'image et association des formes en blocs d'information;

OPEN NETWORK
 DEPT SW
 BOX 18666
 DENVER, CO 80218
 U S A

OPEN NETWORK
 DEPT SW
 BOX 18666
 DENVER, CO 80218
 U S A

ZION yellowstone
 Grand Junction
 CO USA
 81503

ZION yellowstone
 Grand Junction
 CO USA
 81503

Figure 5.7.5.1 Resultats de la description de l'image et association des formes en blocs d'information;

5.8 - DETECTION DE CANDIDATS "CODE-ETAT-VILLE".

5.8.1 - INTRODUCTION.

Après l'étape d'**association** des formes, l'image peut se représenter comme un ensemble de blocs d'information (pseudo-lignes). La ville, l'état et le code postal sont contenus dans un ou plusieurs blocs. Les noms composés de villes ou d'états et les codes postaux en deux parties (codes à 5 + 4 chiffres) peuvent être séparés dans plusieurs blocs se trouvant sur la même parallèle. L'objectif de ce module est de détecter dans cet ensemble, le bon candidat "**ville-état-code**". Si le véritable code est repéré, la soumission des caractères à l'étape de reconnaissance est possible.

Les statistiques effectuées sur les 2.000 images de notre base de données et les rapports de l'"U.S.P.S." ont permis de dégager les règles de **syntaxe de l'écriture** d'une adresse américaine. Pour classer cet ensemble de règles, nous définirons arbitrairement les deux syntaxes suivantes :

- La première syntaxe, appelée **spatiale**, existe au niveau de l'image, et régit tous les blocs d'informations. La figure (5.8.1.1) montre les relations géométriques qui existent entre les trois entités considérées: le code postal, l'état, la ville.

- La seconde syntaxe, que définirons comme **grammaticale**, assure l'appartenance des formes à une entité. Comme on peut le constater sur la figure (5.8.1.2), les formes de l'adresse sont agencées en phrases. Chaque ponctuation et chaque espace, positionné en fonction de règles grammaticales, délimite les entités de ces phrases.

En considération de la réelle structure de l'adresse, et du danger que représente l'investigation aléatoire, nous proposons que la formulation des hypothèses "**ville-état-code**" se calque sur les deux syntaxes de l'adressage.

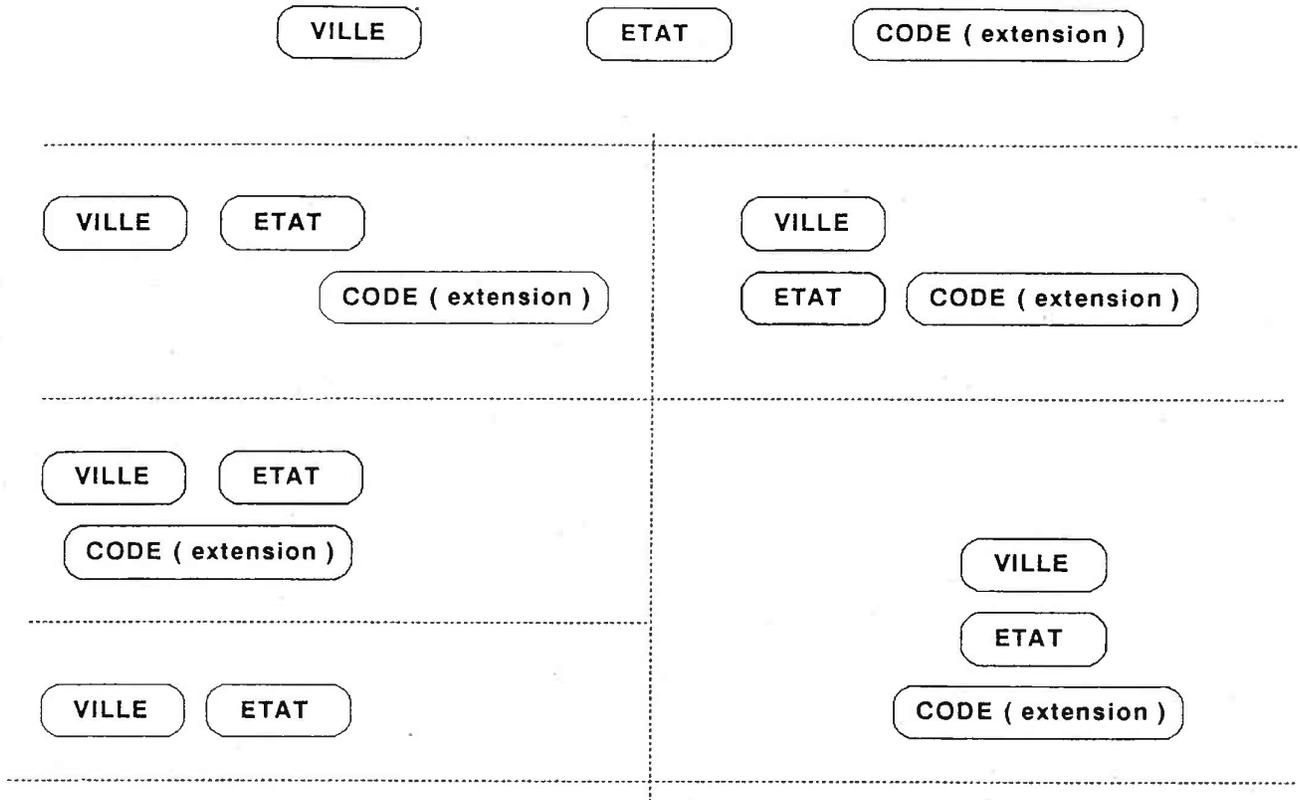


Figure 5.8.1.1 Syntaxe d'adressage au niveau des blocs.



Figure 5.8.1.2 Syntaxe grammaticale de l'adresse visible au sein du bloc.

5.8.2 - CONCEPTS DE BASES.

Compte tenu de l'existence de ces règles syntaxiques, mais aussi de l'abondance des configurations qui leur échappent, notre module de localisation étudie tous les blocs d'information. Une analyse de leur contenu permet de formuler des hypothèses. Pour détecter les bonnes solutions, configurées parfois de manière marginale, la génération des candidats "ville-état-code" se fait selon des critères peu restrictifs car cette syntaxe est malheureusement assez floue. Les hypothèses formulées et pondérées sont éliminées en fin de processus par le module de décision.

Il est possible de faire un parallèle avec le concept de **grammaire** d'un langage utilisé pour identifier des phrases écrites ou parlées et pour reconnaître les formes (caractères). En effet, dans le cas de l'adressage, il est possible de se donner une grammaire $G = (Vt, Vn, P, S)$ définie par ce quadruple :

- Vt vocabulaire terminal
Vt = [tiret, virgule, 4-caractères, 5-caractères, N-caractères, espace-intra-bloc, bloc-5-caractères, bloc-N-caractères]
- Vn vocabulaire auxiliaire (non terminal)
Vn = [ville, état, code, extension]
- S axiome de la grammaire, S élément de Vn.
S = adresse correspondant à la concaténation des éléments de Vn et Vt.
- P ensemble des règles de réécriture (ou production)
La forme générale d'une règle de production de la grammaire est :
a -> b où a, b sont des éléments de [Vt U Vn]*
c'est à dire [Vt U Vn]* = (a1,a2,...,an ai élément de Vt U Vn, n élément de N).

Contrairement aux réalisations les plus courantes qui recherchent la signification de la phrase ou de la forme en la décomposant en éléments du vocabulaire terminal, notre module de localisation connaît la signification mais doit concaténer des éléments de Vt pour construire une chaîne "code-état-ville" de Vn.

Le fonctionnement du module de **localisation** est représenté sur la figure (5.8.2.1). Ce schéma explique les différents chemins d'analyses des blocs d'informations. La division de cette recherche de candidats "ville-état-code" en deux phases

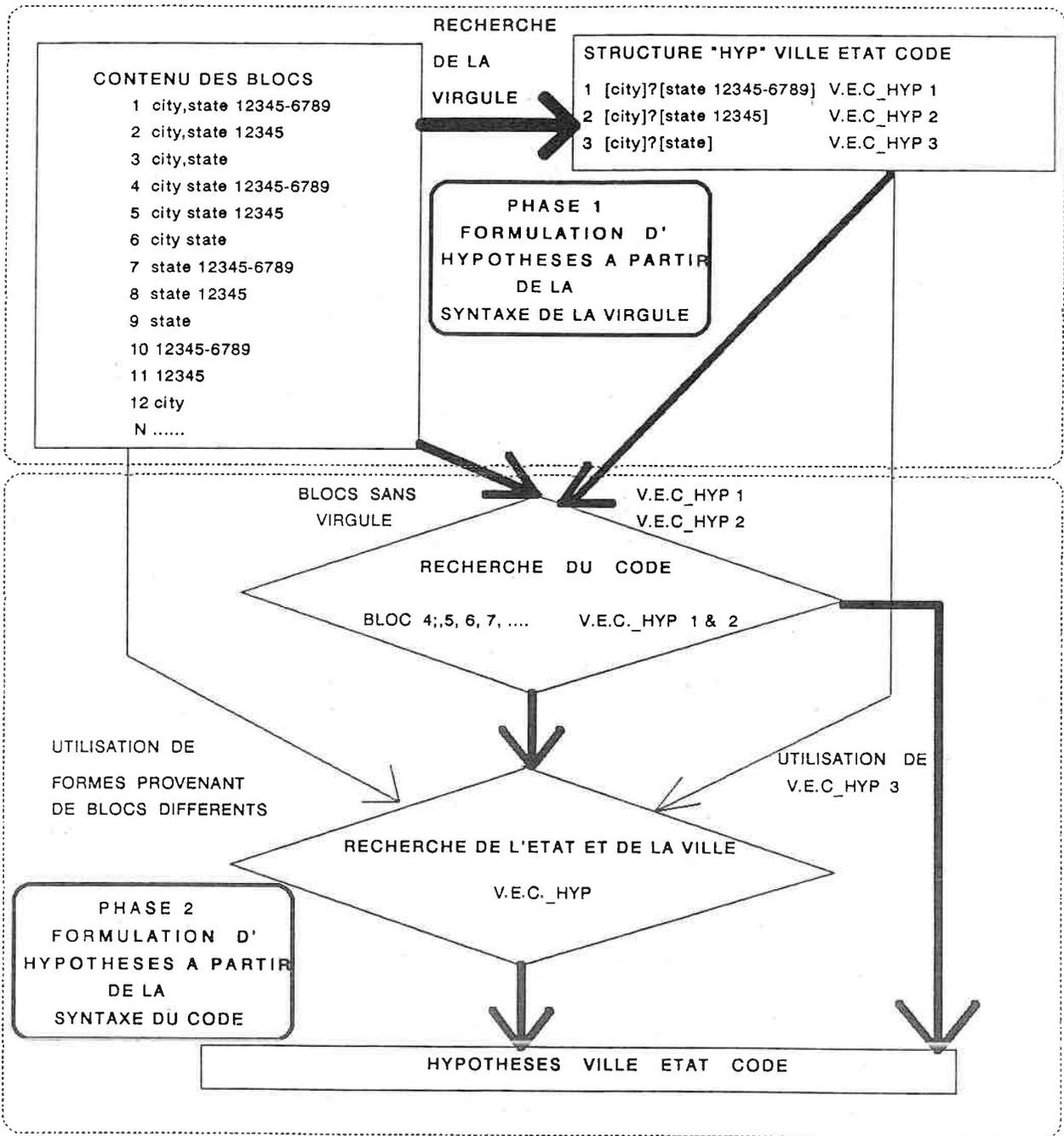


Figure 5.8.2.1 Schéma fonctionnel du module de localisation du code, de l'état, et de la ville.

séquentielles est déduite de la présence dans l'adresse d'éléments syntaxiques robustes et pertinents :

- La phase 1 est une conséquence de la présence dans 60 % des adresses d'une virgule séparant la ville de l'état. Cette considération se traduit au niveau du processus par une recherche des virgules dans l'image, et de formulations d'hypothèses "ville-virgule-état".

- La phase 2 attache de l'importance au tiret séparant le code de son extension facultative et à la stabilité du nombre d'éléments d'un code postal de 5 ou 5+4 formes. Au début de cette deuxième étape, le système est confronté à deux types de données, les blocs et les hypothèses "ville-virgule-état". Au cours de l'analyse, ce module de détection de codes postaux formule des candidats "code" à partir des blocs et des candidats "ville-virgule-état-code" à partir des hypothèses.

Pour chaque bloc de l'image, ces procédures tendent de créer une hypothèse en étudiant les éléments contenus. L'algorithme fait appel, si nécessaire, à d'autres blocs voisins pour compléter le candidat. C'est ce qui se passe pour les candidats codes isolés auxquels il faut associer une ville et un état. Ces entités sont extraites, soit dans le même bloc, soit dans un bloc différent, ou proviennent d'une hypothèse "ville-virgule-état". Un bloc peut générer plusieurs hypothèses **code** et devenir paradoxalement l'**état** ou la **ville** d'un autre candidat.

- Phase 1 : procédure utilisant la syntaxe de la virgule (figure (5.8.2.2)).



Figure 5.8.2.2 Procédure utilisant la syntaxe de la virgule.

Si la **virgule** est un élément discriminant pour séparer la ville de l'état, celle-ci ne nous permet pas de dire si les noms sont complets et si l'état est suivi d'un code postal inclus dans le même bloc. Des procédures annexes raccordent les blocs voisins, considérés comme faisant partie à priori des noms composés.

C'est la phase 2 qui est chargée de détecter l'existence possible du code postal à droite de l'état dans les hypothèses "ville-virgule-état".

Un état peut être considéré à tort comme "état-code" lors de la phase 2. Pour éviter cela, le processus de la phase 1 conserve des hypothèses "ville-virgule-état", considérées comme closes, où les limites de l'état et de la ville sont définitivement fixées.

- Phase 2 : procédure utilisant la syntaxe du code postal (figure (5.8.2.3)).

Statistiquement le code postal est formé de cinq chiffres rassemblés dans un même bloc, souvent disjoints. Ce bloc est généralement réduit à un ensemble de 5 formes. Tout bloc contenant cinq formes simples, du type **caractère isolé**, peut devenir **code**. Pour éviter la génération abusive de candidats dans les lettres dactylographiées ou à écritures non cursives, la procédure s'intéresse seulement aux blocs les plus à droite. Les **codes longs** ont la partie **code** à cinq chiffres séparée de l'**extension** de quatre chiffres par un espace ou un tiret.

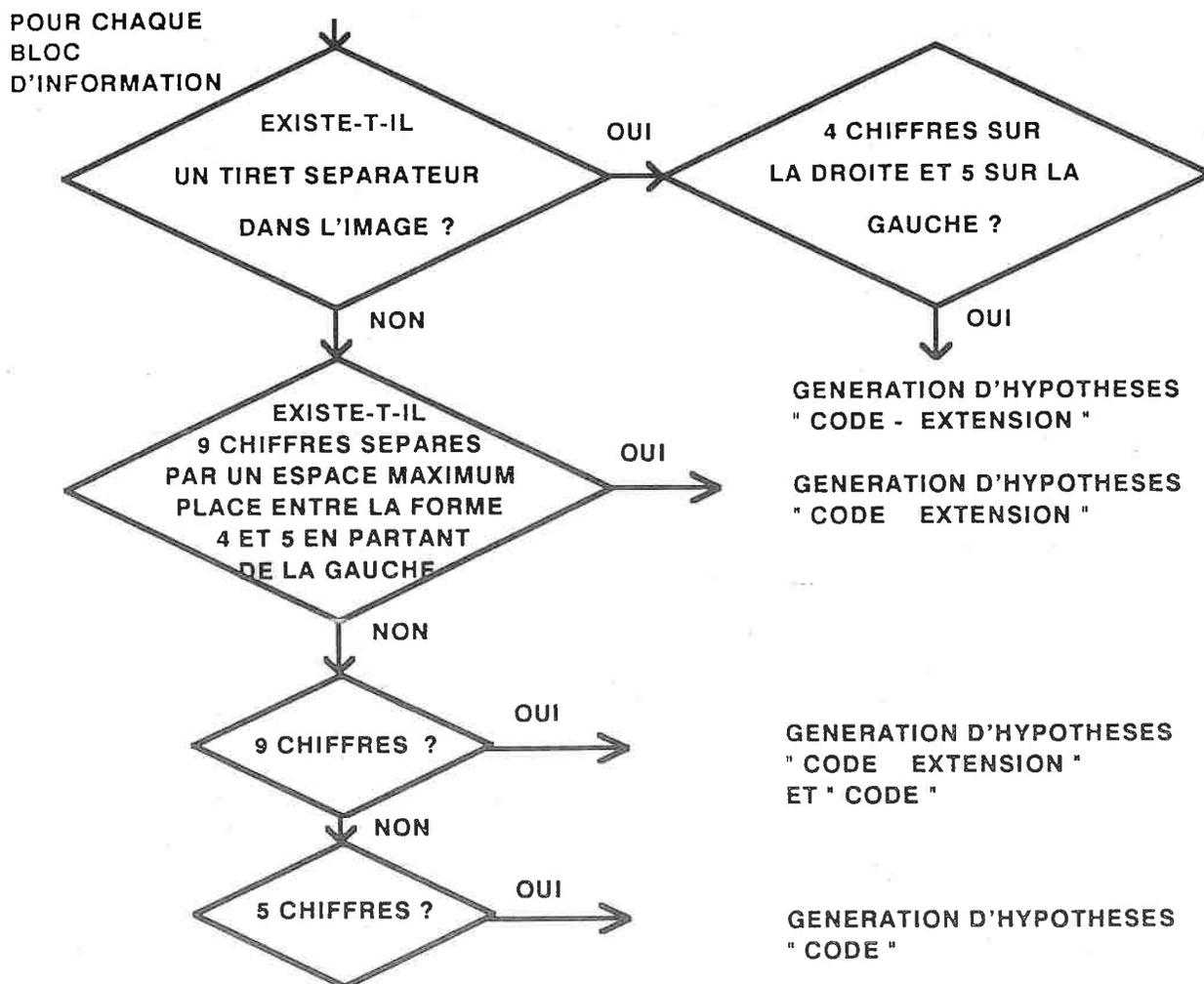


Figure 5.8.2.3 Procédure utilisant la syntaxe du code postal

- Phase 2 : recherche de la ville et de l'état pour les hypothèses **code**.
(figure (5.8.2.4)).

Le **code** présumé peut se trouver isolé ou inclus avec l'état ou l'état et la ville.

Isolé, le choix se porte sur la structure contenant une virgule, sinon sur les blocs de manière syntaxique les plus proches.

Inclus avec d'autres formes, le système reconnaît être en présence de l'état mais ne peut définir avec certitude si l'état est complet ou précédé de la ville.

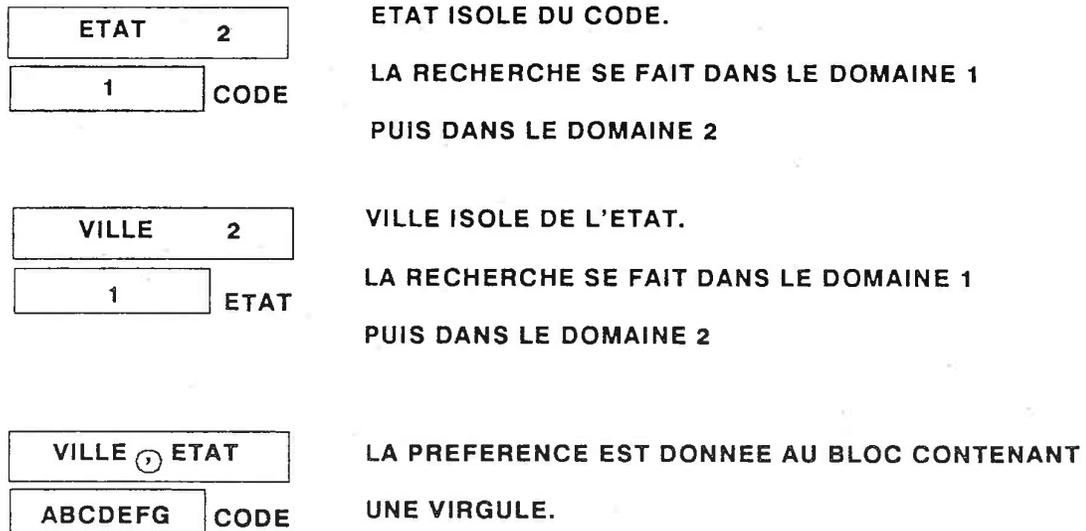


Figure 5.8.2.4 a) Recherche des entités ville et état de manière syntaxique.

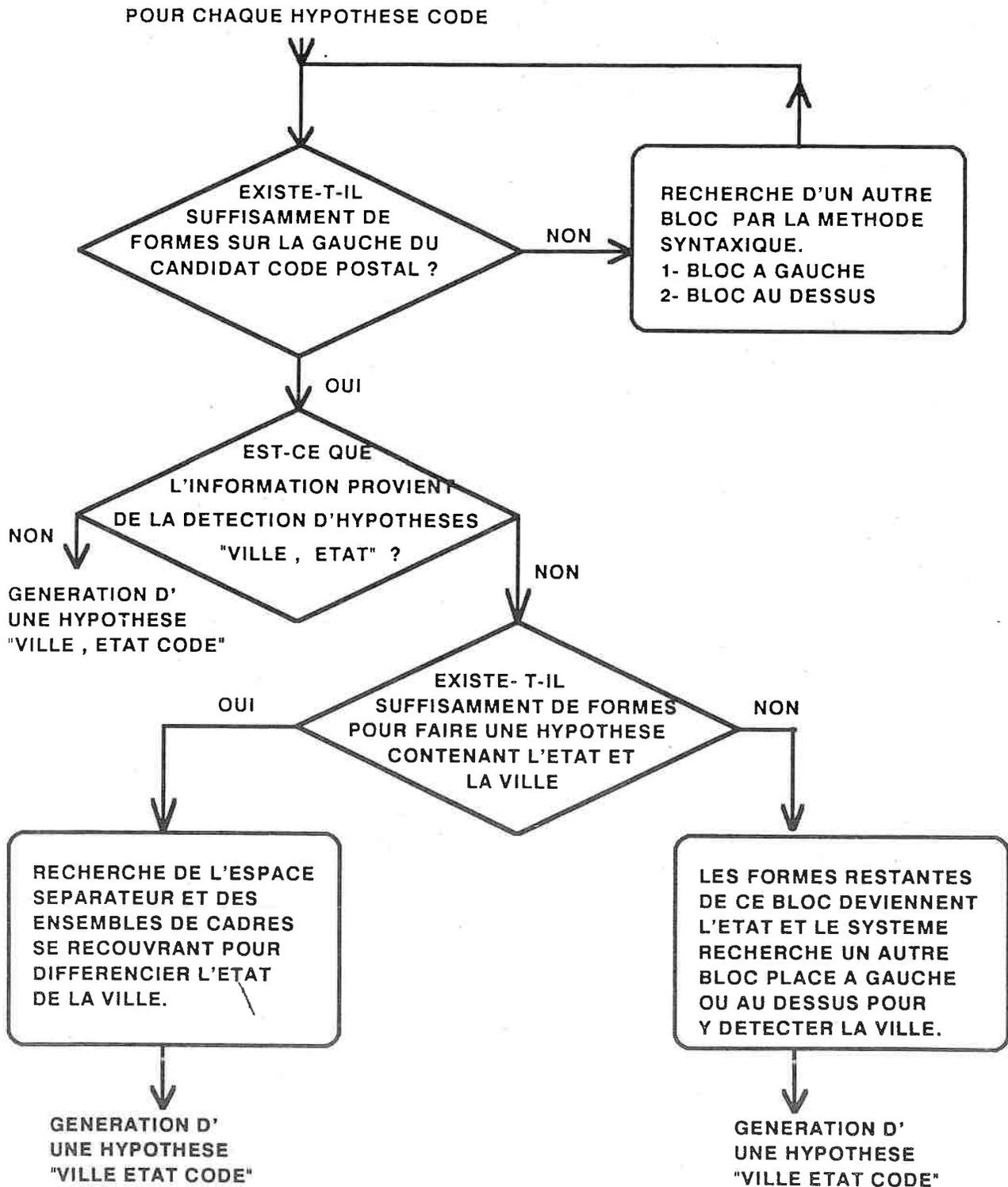
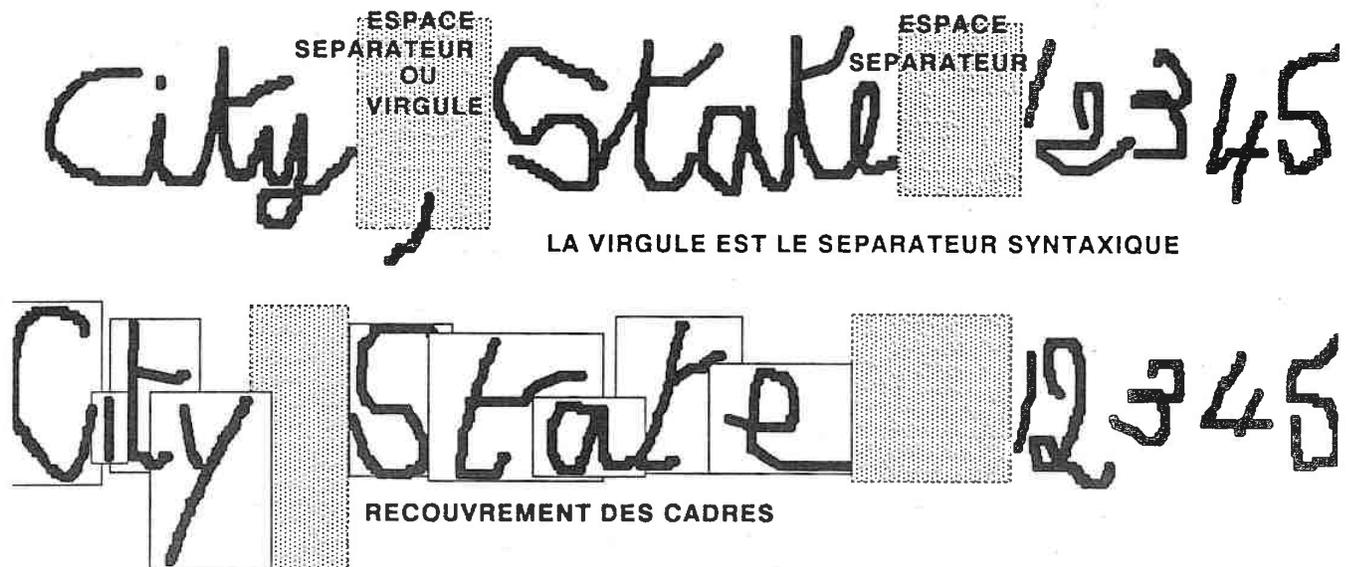


Figure 5.8.2.4.b) Schéma fonctionnel du module de détection de l'état et de la ville.

Pour délimiter les entités au sein du bloc, le système analyse les caractéristiques représentées sur la figure (5.8.2.5). La procédure utilise les espaces significatifs qui séparent souvent les groupes de caractères ou les **mots longs**. Le recouvrement mutuel des cadres de formes, la typologie des formes présentes (caractères, mots courts, mots longs), la surface des éléments et d'autres entités géométriques sont aussi instructifs pour cette recherche de limites.

A partir d'un **code** présumé, la recherche par plusieurs démarches de l'**état** puis de la **ville** entraîne une multiplication du nombre d'hypothèses "ville-état-code".



LES ESPACES MAXIMUMS DE LA LIGNE ET LES RECouvrements DES CADRES PERMETTENT DE SEPARER L'ETAT DU CODE ET L'ETAT DE LA VILLE QUAND CES ENTITES SONT INCLUSES DANS LE MEME BLOC.

Figure 5.8.2.5 Détection des entités état et ville dans un bloc de formes.

5.8.3 - GENERATION ET PONDERATION DES HYPOTHESES.

Les différentes possibilités de concaténation des caractères et des blocs en entités codes ou états ou villes sont utilisées de manière séquentielle et indépendante pour formuler une liste exhaustive de candidats "ville-état-code". Cette tolérance au niveau de cette "réécriture" de l'adresse permet de tenir compte des erreurs algorithmiques antérieures et de la multitude de présentations de l'adresse existant dans le courrier américain.

5.8.3.1 - Typologie et homogénéité du code.

Le code postal se trouve, soit isolé, soit placé dans la partie droite du bloc complexe, espacé des autres entités par un blanc. Le code peut s'obtenir soit par sommation des formes de la droite vers la gauche au sein du bloc étudié, soit par la recherche de l'espace significatif qui sépare le code de l'état.

- Typologie du code postal (figure (5.8.3.1)).

Le code à extension contient souvent un **tiret** ou un **espace** franc qui sépare les cinq chiffres du code des quatre autres. Le **tiret** est un élément aussi pertinent que la virgule par la stabilité de sa présence. Une hypothèse "**code+extension**" est alors générée. Quelles que soient les méthodes employées, le contenu du code est fixe, formé de chiffres uniques ou de mots courts détectés comme doublets. Cette démarche utilisée sur les lettres non cursives entraînent la formation de nombreux candidats. Une étude de l'homogénéité des éléments permet d'éliminer une partie des codes erronés (figure (5.8.3.2)).

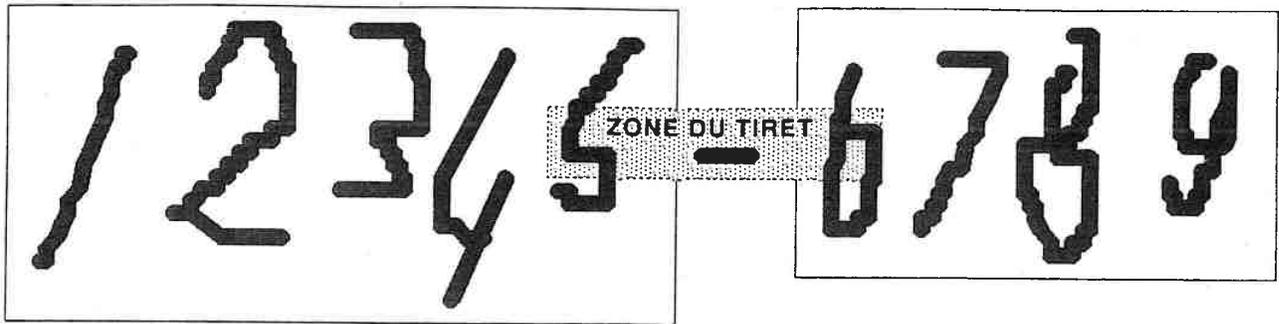
- Homogénéité des chiffres du candidat code.

Pour chaque hypothèse, le système analyse les relations géométriques des formes incluses. Les informations pertinentes utilisées pour les comparaisons sont :

- surface de la forme, de l'entité code,
- intervalle forme-forme,
- caractéristiques géométriques du cadre de la forme, de l'entité code,
- recouvrement des formes.

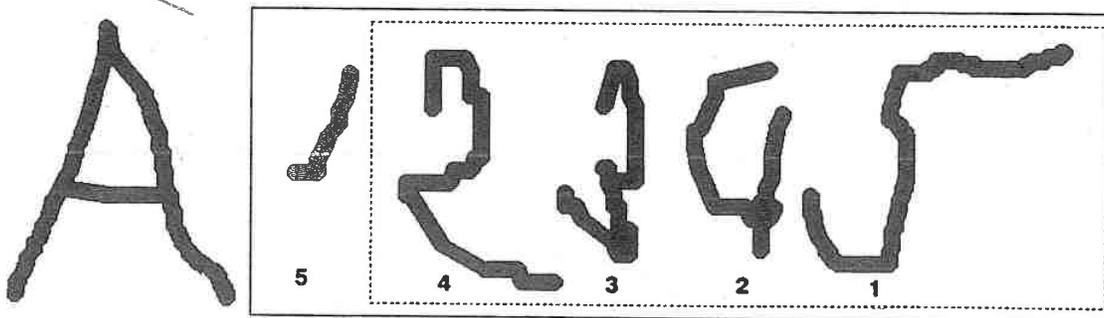
CODE POSTAL (5 CHIFFRES)

EXTENSION (4 CHIFFRES)



EXTENSION + 5 FORMES "UNIQUES"
=> CODE + EXTENSION

4 FORMES "UNIQUES" => EXTENSION



PAS D'ESPACE OU DE TIRET SEPARATEUR => PAS D'HYPOTHESES EXTENSION
HYPOTHESE PAR SOMMATION DES 5 FORMES "UNIQUES"

Figure 5.8.3.1 Typologie du code postal et méthode de détection.

English Dept
 American University
 4400 Massachusetts Ave
 Washington, DC
 20016

English Dept
 American University
 4400 Massachusetts Ave
 Washington, DC
 20016

English Dept
 American University
 4400 Massachusetts Ave
 Washington, DC
 20016

English Dept
 American University
 4400 Massachusetts Ave
 Washington, DC
 20016

English Dept
 American University
 4400 Massachusetts Ave
 Washington, DC
 20016

HYPOTHESE FINALE

Zbo = code isolé

Sbo = état isolé ou à gauche du code

Cco = ville avec virgule

HYPOTHESE REJETEE

Zsu = code par sommation

Sbo = état isolé ou à gauche du code

Cbo = ville isolée

HYPOTHESE REJETEE

HYP NUMBER : 3

Zsu = code par sommation

Sbo = état isolé ou à gauche du code

Cbo = ville isolée

HYPOTHESE REJETEE

Zsu = code par sommation

Sbo = état isolé ou à gauche du code

Cbo = ville isolée

HYPOTHESE REJETEE

Zsu = code par sommation

Sag = état par aggrégation

Cag = ville par aggrégation

Figure 5.8.3.2 Rejet d'hypothèses dont le contenu du code est hétérogène.

5.8.3.2 - Classement des hypothèses formulées à partir de chaque bloc.

Résumons le processus général: pour un bloc donné, chaque méthode d'investigation génère un candidats **code**; chaque **code** génère à son tour plusieurs hypothèses "**ville-état-code**"; les limites de l'**état** et de la **ville** sont obtenues, elles aussi, par des approches différentes (figure (5.8.3.3)).

Pour conserver uniquement les candidats les plus fiables, il est nécessaire de classer toutes ces hypothèses et de leur affecter un taux de confiance en fonction des méthodes employées. Sur les figures explicatives, les solutions finales sont en traits pleins, les rejets en pointillés. Les critères utilisés pour la pondération des hypothèses sont les suivants :

i) Le code.

- Un code postal isolé ou bien distinct de l'état est fortement pondéré.
- La pondération tient compte du coefficient d'homogénéité des formes du **code**.
- Le système ne tient compte des hypothèses obtenues par sommation que dans le cas de faible homogénéité des formes des autres candidats.
- figure (5.8.3.4) : Le système considère comme erronées les détections de formes brisées ou de doublets dans les hypothèses **code** formulées contenant cinq formes.
- figure (5.8.3.5) : Si dans un bloc, il y a concurrence entre une hypothèse **code** et une hypothèse **code+extension**, la préférence est donnée au code postal long.
- figure (5.8.3.6) : La présence d'un **tiret** est l'assurance de la présence d'un **code+extension**. Cette hypothèse est fortement pondérée. Sur les 1.000 images de la **base d'évaluation**, 70 des 72 codes postaux longs contiennent un tiret séparateur.
- figure (5.8.3.7) : Si un bloc est générateur de plusieurs hypothèses bien pondérées, il y a de fortes chances que ce bloc contienne le bon **code**.
- figure (5.8.3.8) : Plus le contenu du candidat est simple, plus le risque de confusion diminue (problèmes des faux doublets et des faux brisés).

Mr and Mrs Jack Targgart
2300 O Road

HYPOTHESE FINALE

Zsp = code par recherche de l'espace

Sre = état entre code et virgule

Cco = ville avec virgule

Columbus, Ohio 43221

Mr and Mrs Jack Targgart
2300 O Road

HYPOTHESE REJETEE

Z4b = extension par sommation

Sre = état entre code et virgule

Cco = ville avec virgule

Columbus, Ohio 43221

Mr and Mrs Jack Targgart
2300 O Road

HYPOTHESE REJETEE

Zsp = code par recherche de l'espace

Zsu = code par sommation

Sbo = état isolé ou à gauche du code

Cbo = ville isolée

Columbus, Ohio 43221

Mr and Mrs Jack Targgart
2300 Oxford Road

HYPOTHESE REJETEE

Zsu = code par sommation

Sbo = état isolé ou à gauche du code

Cbo = ville isolée

Columbus, Ohio 43221

Mr and Mrs Jack Targgart
2300 Oxford Road

HYPOTHESE REJETEE

Z4b = extension par sommation

Zbo = code isolé

Sbo = état isolé ou à gauche du code

Cbo = ville isolée

Columbus, Ohio 43221

Figure 5.8.3.3 Classement des hypothèses pour un bloc donné.

Dr P Moss
4706 Broadway
K C ^{Code}
84112

143

HYPOTHESE FINALE

Zbo = code isolé

Sbo = état isolé ou à gauche du code

Cbo = ville isolée

Figure 5.8.3.4 Code contenant un faux doublet.

Mrs Gary Callahan, O.S.B.

HYPOTHESE FINALE

Z4s = extension par recherche del'espa

Zbo = code isolé

Sbo = état isolé ou à gauche du coc

Cco = ville avec virgule

P.O. Box 488
Callahan, Ala 35056 0488

Mrs Gary Callahan, O.S.B.

HYPOTHESE REJETEE

Z4s = extension par recherche del'espa

Zbo = code isolé

Sbo = état isolé ou à gauche du coc

Cbo = ville isolée

P.O. Box 488
Callahan, Ala 35056 0488

Mrs Gary Callahan, O.S.B.

HYPOTHESE REJETEE

Zsp = code par recherche de l'espace

Sbo = état isolé ou à gauche du cor

Cbo = ville isolée

P.O. Box 488
Callahan, Ala 35056 0488

Mrs Gary Callahan, O.S.B.

HYPOTHESE REJETEE

Zsu = code par sommation

Sbo = état isolé ou à gauche du cc

Cbo = ville isolée

P.O. Box 488
Callahan, Ala 35056 0488

Figure 5.8.3.5 Concurrence entre 2 hypothèses: code, code+extension.

Mrs Gary Callahan, O.S.B.

HYPOTHESE REJETEE

Zsp = code par recherche de l'espace

Ssp = état séparé de la ville par un espace

Csp = ville séparé de l'état par un espace

P.O. Box 488
Callahan, Ala 35056 0488

Equitable Life Assurance Society
sup Benefita Manager

0 Box 11081

Terre Haute, Indiana 46411-1081

HYPOTHESE REJETEE

Zbo = code isolé

Sbo = état isolé ou à gauche du code

Cbo = ville isolée

Equitable Life Assurance Society
sup Benefita Manager

0 Box 11081

Terre Haute, Indiana 46411-1081

HYPOTHESE FINALE

Z4d = extension avec tiret

Zbo = code isolé

Sbo = état isolé ou à gauche du code

Cbo = ville isolée

Figure 5.8.3.6 Fiabilité des hypothèses code contenant un tiret séparateur.

PO BOX 12129
 Hamtramck BR
 Detroit, Michigan 48212

HYPOTHESE FINALE

Zsp = code par recherche de l'espace

Sre = état entre code et virgule

Cco = ville avec virgule

HYPOTHESE REJETEE

Zsu = code par sommation

Sre = état entre code et virgule

Cco = ville avec virgule

HYPOTHESE REJETEE

Zsp = code par recherche de l'espace

Sbo = état isolé ou à gauche du code

Cbo = ville isolée

HYPOTHESE REJETEE

Zsu = code par sommation

Sbo = état isolé ou à gauche du code

Cbo = ville isolée

HYPOTHESE REJETEE

Zsu = code par sommation

Sbo = état isolé ou à gauche du code

Cbo = ville isolée

Figure 5.8.3.7 Foisonnement d'hypothèses à partir d'un bloc.
 Les hypothèses proviennent de démarches différentes.

Mr. James F. Collier 146
26170 Dunning

Inkster Mich 48141

HYPOTHESE FINALE

Zsu = code par sommation

Sag = état par aggrégation

Cag = ville par aggrégation

Mr. James F. Collier
26170 Dunning

Inkster Mich 48141

HYPOTHESE FINALE

Zsu = code par sommation

Ssp = état séparé de la ville par un espace

Csp = ville séparé de l'état par un espace

Mr. James F. Collier

26170 Dunning

Inkster Mich 48141

HYPOTHESE FINALE

Zsu = code par sommation

Sbo = état isolé ou à gauche du code

Cbo = ville isolée

Mr. James F. Collier

26170 Dunning

Inkster Mich 48141

HYPOTHESE REJETEE

Zsu = code par sommation

Sag = état par aggrégation

Cag = ville par aggrégation

Mr. James F. Collier

26170 Dunning

Inkster Mich 48141

HYPOTHESE REJETEE

Zsu = code par sommation

Ssp = état séparé de la ville par un espace

Csp = ville séparé de l'état par un espace

Figure 5.8.3.8 Problèmes engendrés par les fausses détections de caractères doublets et les échecs de la restauration des chiffres brisés.

ii) L'état et la ville.

Comme le tiret pour le code postal à extension, la virgule est un atout dans la localisation des trois entités. Les hypothèses **code-état-ville** bâties autour de la virgule sont fortement pondérées. Les candidats dont la séparation ville-état a été obtenue par **recherche de l'espace séparateur** sont assez fiables dans le cas d'absence de virgule. Un exemple est présenté sur la figure (5.8.3.9).

iii) La syntaxe d'adressage.

Dans certaines lettres, il demeure un trop grand nombre d'hypothèses, même après pondération selon les critères précédemment exposés. Les causes les plus courantes de ce foisonnement de candidats sont les écritures non cursives, les virgules abondantes et certaines images très riches en formes. Une **pondération syntaxique** permet d'éliminer les hypothèses trop proches du sommet de l'image et la préférence est donnée aux **codes** se trouvant dans la partie basse.

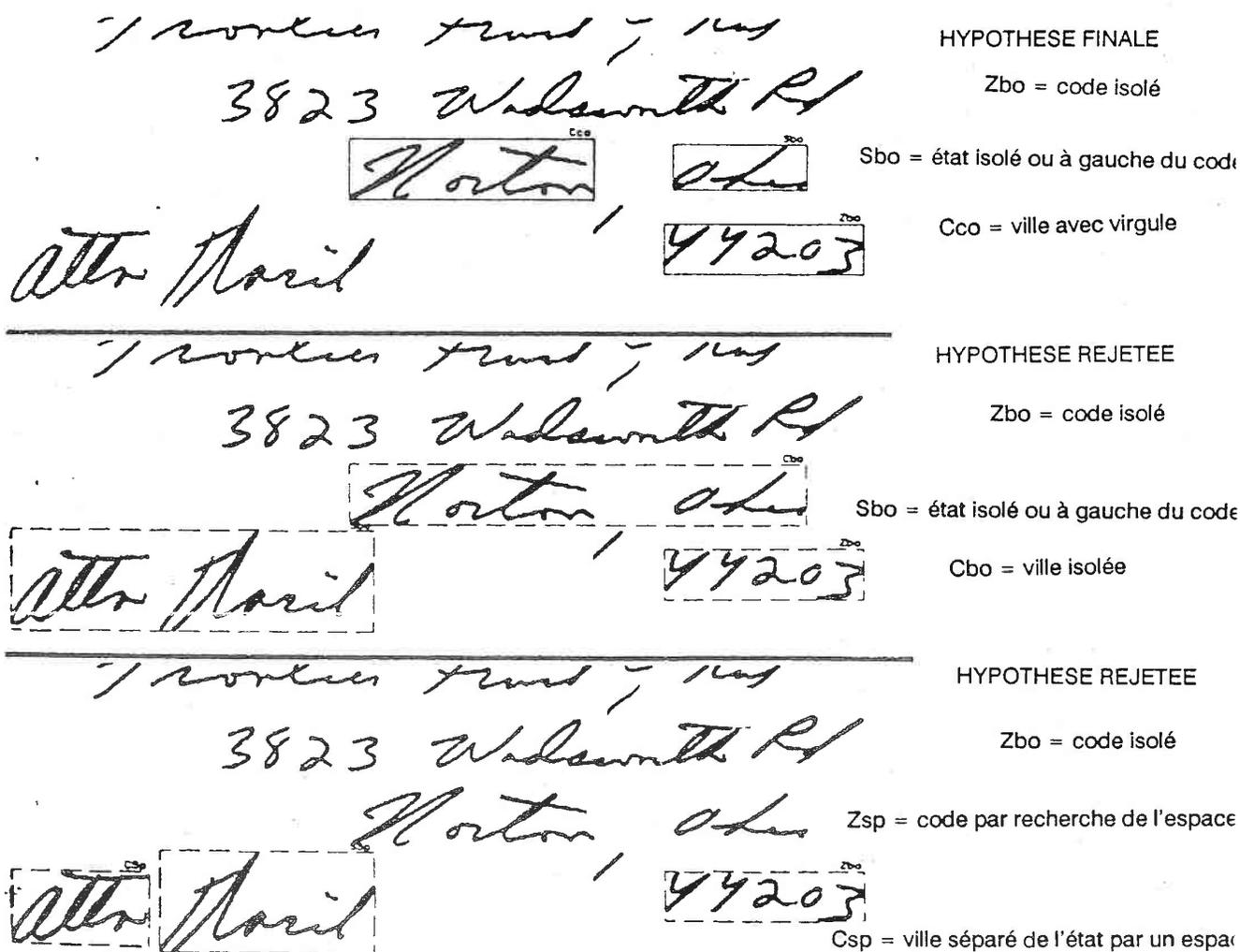


Figure 5.8.3.9 Fiabilité des hypothèses "ville,état" lorsque la virgule existe.

5.8.3.3 - Rejets de configurations et choix final.

Pour faire un choix final, seuls les candidats dont le taux de confiance dépasse un seuil critique sont conservés. Ce seuil est déduit de l'étude statistique effectuée sur l'ensemble des hypothèses générées sur les 2.000 images de référence.

La nécessité de déboucher sur une ou deux hypothèses finales oblige le système à éliminer certaines configurations.

La comparaison des positions, des dimensions, des surfaces des trois entités **code**, **état**, **ville** issues d'un candidat permet de détecter certaines anomalies. Les dimensions relatives aberrantes et les contenus excessifs ou hétérogènes de certains noms permettent de rattraper de nombreux cas. C'est ce que nous pouvons constater sur l'exemple de la figure (5.8.3.10).

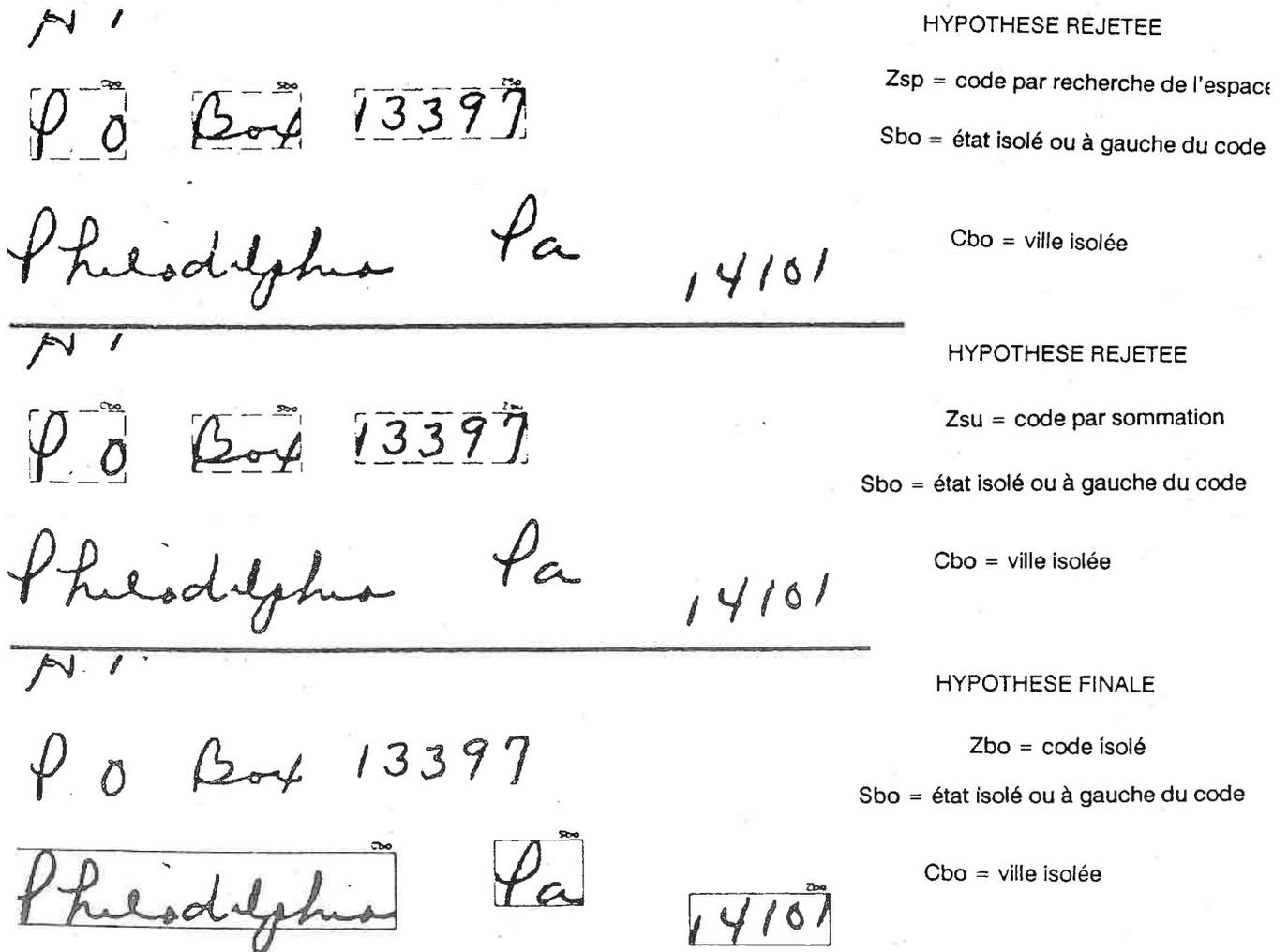


Figure 5.8.3.10 Rejets d'hypothèses: comparaisons des dimensions géométriques des 3 entités

Pour un ou plusieurs **codes** considérés comme justes, les configurations **état-ville** qui leur sont associés peuvent avoir des **taux de confiance** proches. A ce stade de la démarche, plus aucun critère ou paramètre ne peut assister la décision. Le choix final du code, de l'état et de la ville se porte sur les entités obtenues par la méthode qui donne statistiquement les meilleurs résultats. La figure (5.8.3.11) illustre bien le dilemme.

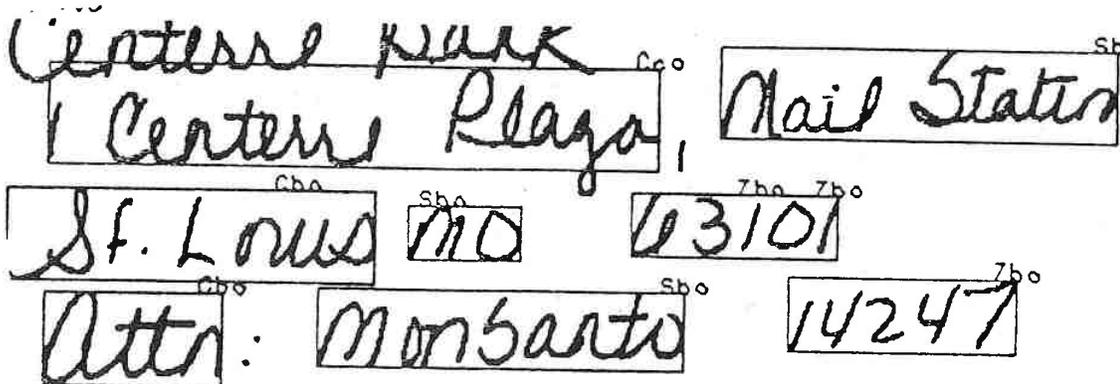


Figure 5.8.3.11

Problèmes de choix final entre deux hypothèses
dont les taux de confiance sont suffisamment élevés pour être fiables.

5.8.4 - EVALUATION.

Les résultats obtenus par le module de **localisation** peuvent s'interpréter en fonction des conditions d'évaluation. En considération de la variabilité des dimensions des images fournies par l'"U.S.P.S.", seule la zone de l'image contenant le code postal a été conservée. Cette surface de 512 x 1456 pixels est représentée par trois plans images 512 x 512 pixels se recouvrant légèrement.

En raison des erreurs de manipulation ou d'oublis de la part de la personne qui a rédigé l'adresse, certains codes postaux ne sont pas présents dans l'image ou impossible à identifier visuellement. Les raisons de l'illisibilité du code sont diverses. Le code postal peut être incomplet, barré par le rédacteur, oblitéré par un **code barre**, ou mal centré dans la fenêtre de l'enveloppe. Dans ces conditions extrêmes, il est logique de considérer ces codes postaux altérés comme inexistant.

Les chiffres qui sont avancés proviennent de l'évaluation de 1.000 images connues de la **base d'apprentissage** et de l'évaluation de 1.000 images inconnues de la **base d'évaluation**.

- Performances de détection du code postal sur les 1.000 lettres de la **base d'apprentissage** :

Total des bonnes sanctions : 91 % dont

- 84 % de codes postaux correctement détectés,
- 7 % de lettres sans code postal et sans hypothèse **code**,

Total des mauvaises sanctions : 9 % dont

- 7 % de lettres contenant un code postal non détecté.
- 1 % d'hypothèses erronées sur des lettres sans code postal.
- 1 % d'hypothèses erronées sur des lettres contenant un code postal.

- Performances de détection du **code** sur les 1.000 lettres de la **base d'évaluation** :

Total des bonnes sanctions : 86 % dont

- 77 % de codes postaux correctement détectés,
- 9 % de lettres sans code postal et sans hypothèse **code**,

Total des mauvaises sanctions : 14 % dont

- 12 % de lettres contenant un code postal non détecté.
- 1 % d'hypothèses erronées sur des lettres sans code postal.
- 1 % d'hypothèses erronées sur des lettres contenant un code postal.

Les causes de confusions ou d'erreurs de détections sont diverses et proviennent de la sommation des limites d'efficacité des modules appartenant aux étages de **binarisation** et de **localisation**. Un descriptif permet de juger l'impact de chaque module sur la localisation du **code**:

- 1) - Si les soulignements, ou les graphiques originaux connectés aux chiffres du code ne sont pas éliminés, le code postal est considéré comme perdu.
- 2) - Si plus de deux chiffres sont connectés ensemble, la procédure de **détection des doublets** n'est pas efficace et l'hypothèse n'est pas retenue.
- 3) - Si un ou plusieurs chiffres sont connectés à un mot ou à un caractère alphabétique, la détection ne trouvera pas le bon code postal.
- 4) - Si un doublet n'est pas détecté, le code est considéré comme une entité de quatre chiffres, ce qui correspond souvent à un numéro de boîte postal "PO BOX 1234". L'hypothèse n'est donc pas retenue.
- 5) - Si les chiffres sont trop fragmentés ou coupés de manière marginale, la procédure de **restauration des chiffres brisés** ne peut pas traiter correctement ces chiffres. Le code postal ne peut pas être considéré comme candidat **code**.
- 6) - Certaines images sont très bruitées et le code se trouve alors connecté à des graphiques parasites. La localisation du code postal est impossible.
- 7) - Certaines hypothèses **code** formulées sur le bon code postal sont rejetées par le module de décision de la **localisation** quand l'ensemble des caractéristiques géométriques des chiffres est considéré comme hétérogène.
- 8) - Les formes qui présentent des hauteurs excessives par rapport à l'image sont considérées comme des **graphiques** et sont éliminées. Parfois certains codes tombent dans cette catégorie.
- 9) - Certaines adresses rédigées par des rédacteurs indisciplinés ont leur code postal placé sur la gauche ou en haut de la lettre. Ces configurations ne suivent pas la syntaxe de l'adressage. Par conséquent, la localisation du code postal est impossible car le système ne considère que les candidats se trouvant sur la droite de l'image et de préférence dans la zone basse.

10) - La **binarisation** qui est effectuée sur les images américaines multi-niveaux de gris travaille sur des surfaces 16 x 16 pixels qui correspondent au voisinage d'étude du pixel. Ce voisinage a été choisi en fonction de l'épaisseur moyenne des caractères. Les adresses rédigées au marqueur produisent après binarisation des chiffres troués dont les contours sont dentelés. Ces formes génèrent de nombreux **points extrêmes** et sont souvent considérées comme **mots longs** ou **graphiques**. Le module de localisation des codes postaux ne travaille pas sur ce type d'information.

11) - Certains candidats **code** ont les apparences de codes postaux parfaits. Malheureusement, il s'agit parfois de mots de cinq lettres indépendantes formant des entités géométriquement homogènes. Seul l'étage d'**identification des chiffres** permettrait de lever l'ambiguïté. Ce cas est courant dans les lettres rédigées en **écriture bâtons** où la dernière ligne correspond à une ligne d'attention, message précisant le nom du destinataire.

12) - L'**association** des chiffres se fait selon des critères de proximité. Si certains chiffres sont anormalement espacés, certains codes postaux se trouvent séparés dans deux blocs différents. Dans ce type de configuration, le code postal est perdu.

Les objectifs du projet sont de fournir aussi des candidats **état** et **ville**. Cette information est manuscrite et souvent cursive. Si le code postal correspond à une succession de cinq ou neuf chiffres plus ou moins séparés, les noms d'états et de villes offrent de nombreuses configurations. Un état peut être représenté par un **code d'abréviation** formé de deux caractères ou peut être écrit de manière littérale. Lorsque les noms d'états ou de villes sont composés, les erreurs de délimitation sont fréquentes. Les performances avancées sont le résultat de la détection de l'**état** et de la **ville** pour les hypothèses dont le **code** a été correctement localisé .

- Performances de détection de l'**état** sur les 1.000 lettres de la **base d'apprentissage** :

- 91 % des candidats **état** ont été correctement localisés.

- 9 % des candidats **état** sont inutilisables. Ceci s'explique par :

- 4 % des candidats sont des détections erronés,

- 5 % des candidats sont incomplets ou contiennent des formes supplémentaires.

- Performances de détection de l'**état** sur les 1.000 lettres de la **base d'évaluation** :

- 86 % des candidats **état** ont été correctement localisés.

- 14 % des candidats **état** sont inutilisables. Ceci s'explique par :

- 12 % des candidats sont des détections erronés,

- 2 % des candidats sont incomplets ou contiennent des formes supplémentaires.

- Performances de détection de la **ville** sur les 1.000 lettres de la **base d'apprentissage** :

- 90 % des candidats **ville** ont été correctement localisés.

- 10 % des candidats **ville** sont inutilisables. Ceci s'explique par :

- 4 % des candidats sont des détections erronés,
- 6 % des candidats sont incomplets ou contiennent des formes supplémentaires.

- Performances de détection de la **ville** sur les 1.000 lettres de la **base d'évaluation** :

- 82 % des candidats **ville** ont été correctement localisés.

- 18 % des candidats **ville** sont inutilisables. Ceci s'explique par :

- 14 % des candidats sont des détections erronés,
- 4 % des candidats sont incomplets ou contiennent des formes supplémentaires.

Les erreurs de détection de la **ville** et de l'**état** proviennent principalement du manque d'information pertinente sur ces deux entités. De plus il arrive souvent que ces éléments soient incomplets ou connectés à d'autres lignes. Il est possible d'énumérer les causes d'échecs.

1) - Les soulignements non détectés peuvent empêcher une bonne localisation de la **ville** ou de l'**état**.

2) - Lorsque l'image 512 x 1456 pixels ne conserve qu'une petite partie d'un nom, ce morceau est parfois trop petit pour être détecté comme candidat.

3) - Un nom ou une partie de nom (majuscule) peut être éliminé si ses caractéristiques géométriques dépassent certains seuils. Le cas le plus courant est la connexion de mots appartenant à des lignes différentes.

4) - L'**association des formes** peut englober des graphismes ou des mots avec le nom de la ville ou de l'état. Il est difficile d'éliminer du bloc ces informations car le système ne sait pas si ces entités sont dépourvues d'intérêt ou si elles appartiennent à un nom composé.

5) - L'**association** sépare parfois les éléments d'une même entité en plusieurs blocs ou inversement rassemble trop d'entités dans le même bloc. Le système ne déduit pas facilement l'appartenance des formes aux entités correspondantes. Il suffit que les espaces séparateurs de mots ne soit pas significatifs ou que les noms soient composés pour entraîner la génération d'hypothèses erronées. La virgule qui est présente dans 60 % des lettres américaines est malheureusement le seul élément séparateur pertinent. Dans les lettres où la virgule est absente, le système est incapable de reconstituer les noms avec certitude si plusieurs **blocs** sont sur la même horizontale.

6) - Les graphismes isolés peuvent être pris pour des noms si leur position est en accord avec les règles syntaxiques. A partir de là, la détection est vouée à l'échec.

7) - Sur certaines images, les noms composés sont écrits sur deux lignes différentes ou bien l'information se trouvant à gauche du code postal n'est pas une des deux entités recherchées. Il est inéluctable que le système ne peut pas apprécier ces variations de la syntaxe d'écriture.

6 - CONCLUSION GENERALE.

Ce document, reflet partiel d'un **système de reconnaissance du code postal** manuscrit américain, montre l'importance de la manière dont l'information doit être caractérisée. Les performances d'une méthode d'identification sont dépendantes de la pertinence et de la fiabilité des primitives de l'image ou de la forme.

En considération des conclusions de l'"United States Postal Service", les résultats obtenus par l'ensemble des modules qui ont contribué à la formation d'hypothèses **code-état-ville** sont apparus globalement plus performants que ceux des trois sociétés et des universités concurrentes. En effet, sur la base d'évaluation, notre approche a permis de localiser le code postal dans **86 %** des cas, et n'a commis que **2 %** de confusions (erreurs). Cette méthodologie est la seule à pouvoir fournir des candidats ville et état.

Les approches des trois autres soumissionnaires sont assez semblables et s'effectuent en trois phases: extraction des lignes; segmentation de lignes en caractères; recherche du code postal. Ces localisations de lignes se font par analyse des projections horizontales de l'image binaire (histogramme,...). Ensuite, il y a isolement et déconnexion des caractères au sein de ces lignes. Ces segmentations utilisent la topologie des composantes connexes mais ne tirent aucune information de celles-ci. Leurs détections ne recherchent que le code postal et d'une manière malheureusement assez aveugle. Il en résulte des taux de bonnes sanctions proches des **66 %** avec des taux d'erreurs s'étalant entre **15 et 30 %**.

Reprenons l'ensemble des processus présentés pour en résumer les points forts et les défauts.

Le module de **binarisation** de l'image qui est semblable à celui du prototype SYBILIS est le fruit de nombreux tests effectués sur des courriers de nature différente. Les paramètres fixes du module sont les résultats du compromis entre la binarisation de lettres dactylographiées et celle de lettres manuscrites.

Le module de **description** de l'image fondé sur la détection des **points extrêmes** des formes et le module d'**association**, qui étudie les relations géométriques des formes pour les intégrer dans des **blocs d'information**, forment un système prometteur de **localisation** de mots. Par rapport à d'autres techniques plus classiques, cette démarche est instructive sur le contenu des lignes d'informations. D'une part, il est possible de séparer les caractères des traits de soulignement. D'autre part, les formes sont **pré-étiquetées** et peuvent appartenir aux catégories suivantes : caractères isolés, ponctuations, doublets, triplets, mots, lignes, graphiques. L'**association** des formes en blocs d'information permet ensuite de séparer les entités de nature différente.

Des améliorations sont néanmoins envisageables sur la détection des points extrêmes. Certains contours dentelés produisent un foisonnement de points extrêmes. Cette abondance de caractéristiques nuit à la description des formes.

Le module de **localisation** qui se repère à la syntaxe d'écriture pour générer une ou plusieurs hypothèses **code-état-ville** doit sa fiabilité à la qualité de l'information fournies par les deux modules précédents.

Le module de **restauration des formes brisées** par ses critères simples et robustes montre bien la fiabilité et la pertinence des points extrêmes.

Le module de **séparation des chiffres connectés** se fonde sur les points extrêmes pour détecter les doublets et les triplets. La recherche des zones de connexions sur le contour montrent aussi l'intérêt d'utiliser la topologie de la forme. Cette démarche embryonnaire qui identifie les zones de connexions, et produit la coupure, nécessite des améliorations au niveau de la décision finale. Malgré une détection fiable des zones de coupures potentielles, le système manque d'informations pour lever l'ambiguïté des configurations antagonistes. Il lui est impossible de procéder aux coupures de manière itérative. L'objectif final est de pouvoir séparer correctement les chiffres connectés, quel que soit leur nombre, en fonction d'une démarche contextuelle simple.

La méthodologie fondée sur les points pertinents des formes remporte un certain succès au sein de la société. Les algorithmes de **description** et d'**association** pourraient être repris pour d'autres projets.

Les exemples de projets dans le domaine postal sont :

- la localisation du bloc adresse sur les lettres non mécanisables et les colis postaux (vidéo-codage),
- la recherche d'une signature de la lettre pour la cataloguer dans le cadre d'une indexation différée dans le temps (post-indexation),

Les exemples de projets dans les domaines bancaires, administratifs et industriels sont :

- des lecteurs de sommes en chiffres, de lettres manuscrites ou des vérifications de la signature ,
- des lecteurs de plaques minéralogiques,
- des lecteurs de caractères sur pièce.(contrôle non destructif).

ANNEXES

PUBLICATION

REALISEE

POUR L'

L' ADVANCED TECHNOLOGY CONFERENCE

OF

THE " UNITED STATES POSTAL SERVICE "

3 - 5 mai 1988

**AN APPROACH FOR THE LOCATION OF
CITY/STATE/ZIP INFORMATION
ON HANDWRITTEN OR MACHINE-PRINTED
LETTERS**

François COURJARET & Hervé CUVELIER

CGA-HBS
3, rue de Tunis - Boite Postale 111
93203 SAINT-DENIS CEDEX 1 - FRANCE

ABSTRACT

CGA-HBS ALCATEL Company is presently under contract with the USPS to develop techniques for the automatic location and recognition of the address ZIP Code. The technique introduced here concerns the Location task which provides from a bi-level image a set of block candidates for the City/State names and the ZIP Code.

This paper provides first an understanding of City/State/ZIP Location problem and then presents the recommended approach as well as the performances obtained from a base of 3.000 images.

INTRODUCTION : POSITION OF THE PROBLEM

If one disposed of sufficient computing power and reliable recognition algorithms (especially in terms of low substitution rate), the extraction of that part of the address image corresponding to the ZIP Code block would not be necessary. As a matter of fact, the contents of the whole address image would be submitted to the recognition stage, and decision for the ZIP Code value would be obvious : it would be the unique information recognized by the numerical classifier as a set of five or nine digit values.

But today, the state of the art in computers does not offer enough power with respect to speeds of recognition higher than 10 letter pieces per second. Moreover, a classifier designed for numerical digit recognition does not reject absolutely any pattern different from a digit and is misleading especially for handwritten addresses. It is thus necessary to suggest to the recognition stage the most probable part of the image where the ZIP is located.

Line and block localization have been frequently investigated, but most of the techniques were developed for machine-printed text and do not perform on handwritten text for the following reasons :

- handwritten script lines are not always horizontal, making ineffective histogramming techniques
- handwritten lines are very variable in height and ascenders and descenders may connect two different lines
- handwritten letters are not regularly spaced as in type-printed text and no reliable use can be made of the letter pitch.

Moreover, the image base provided by the USPS (about 3.000 images) indicated that three other problems merit attention :

- digits are not always made of a unique topological set of pixels due to removal of the pen ball during tracing
- a lot of addresses contain ZIP Codes made-up of digits connected to each other during tracing (see Figure 1)
- a non neglectible part of the mail exhibits a printed form which guides the sender to fill the address. Most of the time some digits cross the printed text and produce a single topological set of pixels for several digits (see Figure 2).

Similar considerations are applicable for State and City information. We have intended in this research to focus on these different levels of difficulty, so that the locator algorithm will be in a position not only to suggest a ZIP Code block but to provide pattern candidates for five or nine digits of the Zip Code.

We have also decided to make an effort for the detection and use of information provided by commas and dashes since 60 % of the base exhibit a comma between the City and the State, 90 % of letters with a ZIP + 4 Code contain a dash.

TECHNICAL APPROACH

For a better comprehension we will use the term "pattern" to designate a set of pixels whose connectivity number is equal to one.

The approach relies on a clustering technique applied at the topological level of the image, and not at the pixel level.

The process is made-up of three stages which work first on the binary image data, then on a symbolic description of patterns to generate finally a list of ZIP, State and City block candidates weighted by a confidence ranking.

1 - DESCRIPTION STAGE

This pre-processing stage is in charge of providing from the raw pixel image a symbolic description of all the patterns contained in the address. For each pattern, a set of attributes is computed and is listed hereunder :

- order of appearance during the scanning of the image
- coordinates of the enclosing rectangular frame
- surface and perimeter
- coordinates of "Extremal Points".

The Extremal Points are all pixels belonging to the external contour of the pattern and where the derivative varies from a negative to a positive value in a clockwise following of the contour.

As illustrated by Figure 3, there are two kinds of such points : the North and South Extremal Points which belongs to the contour borders oriented respectively to the north or the south.

Use of these points, in conjunction with the other attributes, have been found to be very useful for classifying patterns according to the following scheme :

- a pattern will be supposed to be a digit if it exhibits less than a maximum number of extremal points
- a pattern will be supposed to be a string if it exhibits more than a minimum number of extremal points
- a pattern is removed from the image if its surface is too small, too high, or presents an abnormal height/width ratio (lines for example).

This labelling will be further reused during the clustering process.

Moreover, the DESCRIPTION STAGE is in charge of detecting line patterns and of disconnecting words and digits when they are linked to such lines, as illustrated by Figure 4.

The originality of this stage is to scan the image on a column by column basis and to manage in a dynamic mode the data processing only by two successive scans at a time.

2 - ASSOCIATION STAGE

Working on the data provided by the DESCRIPTION STAGE, this process will cluster patterns according to their attributes and according to geometrical and empirical criteria in order to provide line-oriented blocks as illustrated by Figures 5 and 6.

In this way, a simplification is obtained for processing the last stage called LABELLING STAGE which will be discussed later.

Conditions for clustering two patterns that may be a candidate numeric or string are sketched out in Figure 7.

A minimum of overlapping between the vertical projections of the patterns is required, as well as a maximum horizontal distance between the enclosing frames. Of course, these criteria are dependent on the class the patterns belong to, in particular they are more severe for strings rather than for isolated digits. Specific criteria have been set as well for two other types of patterns : dashes and vertical strokes.

Dashes occur between the first five and last four digits of the Zip + 4 Code or result in breaks in patterns like "5", "T" or "E". Vertical strokes occur as "1", "l" or part of a broken digit such as "9" or "4".

The creation of blocks of information is ensured through the clustering of patterns examined two by two. A vertical reference used in associating patterns is updated dynamically during the clustering process in order to adapt variations of position in slanted lines as illustrated by Figure 5.

The ASSOCIATION STAGE provides at the end of this process a list of horizontal blocks of information, in such a way that candidate numeric patterns belonging to a same ZIP Code cannot belong to different blocks. Examples are exhibited in Figures 5 and 6.

3 - LABELLING AND DECISION STAGE

First, a pre-processing is performed to detect connected digits and to combine fragmented patterns resulting from breaks in a digit, in such a way that the next stage will deal only with semantical significant entities : digits or non digits.

Moreover, as we established that the comma was written most of the time between the City and State names, the identification of commas is also carried out. Similarly, we operate an identification of dashes in ZIP + 4 Codes.

An analysis is then initiated on each block from right to left in order to generate sub-blocks corresponding first to the ZIP, then to the State and the City.

Search for the ZIP is made in a first step according to three criteria : the number of digits (5 or 9), the eventual presence of a space between the utmost left digit of the Code and the State, and the apparition of a string pattern on the left digit if the State is written in cursive form.

For each ZIP hypothesis, we then initiate a search for the State and the City. If there are no more patterns on the left of the ZIP, a new block is scanned in a way corresponding to the main syntax habits in US. If there are no more than two or three digits on the left of the ZIP, which traduces a State written in hand-printed form, then the City will be looked for in another block. In other cases, the segmentation between State and City information is performed through the detection of a comma, or if it does not exist, by the detection of the largest inter-pattern space.

The approach, sketched out by Figure 8, generates a list of ZIP, State, City hypotheses with a ranking directly dependent on the number of criteria satisfied, as well as the position in the image regarding to most common writing habits.

The output of this stage provides the recognition stage for each ZIP, City, State list, the raster image of each digit or string where connected digits have been segmented and broken characters associated. The order of submission is directly related to the ranking of the hypotheses.

Results of the process are illustrated in Figure 9.

CONCLUSIONS

The set of algorithms has been tested and refined on a learning base of 1.000 images and are under review using an evaluation base of 1.000 images.

The DESCRIPTION STAGE provides two outputs : the symbolic description of patterns (Isolated Digits, Strings, Other Data) and abstraction of these patterns from underlining or pre-printed frames. With the evaluation base, 200 images of the 2.000 contain underlinings of which 70 % are processed correctly and the patterns extracted.

The ASSOCIATION STAGE provides clustering of patterns constituting lines of information. Using the learning base, 97 % of the images exhibit correctly localized ZIP Codes, i.e. ZIP information is included in a single line block and slant and interlines recovering has been faced with success.

The LABELLING AND DECISION STAGE provides candidate City, State and Zip blocks. The output may include several candidates ranked according to their position in the image. We anticipate at this time using the learning base of 1.000 images to generate a first guess correctly 90 % and a second best guess 95 % of the residual images.

ACKNOWLEDGEMENTS

This work has been sponsored by the U.S. Postal Service and we thank the staff of this Organization and the members of Arthur D. Little at the Technology Resource Center for their attention and guidance.

REFERENCES

- TSUJI, Y. ; ASAI, K.
CIC Systems Research Laboratories, NEC Corporation
Sectioning Apparatus and Method for Optical Character Reader System
Office European des Brevets, 014 6147 (1985)
- MAEDA, Y. ; YODA, F. ; MATSUVRA, K. ; NAMBU, H.
Mitsubishi Electric Corporation
Character Segmentation in Japanese Handwritten Document Images
(IEEE, 1986)
- PALUMBO, P.W. ; SRIHARI, S.N.
State University of New-York
Address Recognition Techniques in Mail Sorting : Research Direction
(BUFFALO 1985)
- KROENERT, G. ; LAUBER, G. ; LOERSCHER, J. ; MEYNIEUX, E. ; POSTL, W. ;
SCHNEIDER, V. ; SEISEN, S. ; TOMBRE, K.
Document editing and Entry Based on the Standardized Office Document
Architecture
CNRS Document 86 - R - 100 (1986)
- HENRICHON, E.
A line Tracking Procedure for Optical Character Scanning
IEEE, 1982
- SRIHARI, S.N.
State University of New-York at BUFFALO
Advanced Character Recognition
Final Report to United States Postal Service, 1985
- TAPPERT, C.C.
IBM Corporation
Segmentation Function Enhancement to Cursive Script Recognition System
IBM Technical Disclosure Bulletin (1984)
- SHRIDHAR, M. ; BADRELDIN, A.
Department of Electrical Engineering, University of WINDSOR, Ontario
Recognition of Isolated and Simply Connected Handwritten Numerals
Pattern Recognition, 1986
- SAYRE, K.M.
University of Notre Dame, Indiana
Machine Recognition of Handwritten Words : A Project Report
Pattern Recognition, 1973

59715
20235
54205-9704
41071004

20852
73120-4199
00658 8200

Wright Watchers Magazine
PUBLICATION
P.O. Box 56129
ADDRESS
Boulder, Co. 80322-6129
CITY STATE ZIP CODE

College Scholarship Service
CN 6343
Princeton NJ. 08541
City State Zip Code

FIGURE 1 :

BROKEN AND CONNECTED CHARACTERS

FIGURE 2 :

IMAGES WITH UNDERLININGS

8477A
 BEARS ROEBUCK AND CO.
 200 MacArthur Dr.
 Wilmington, De. 19804

Angela

- UPPER EXTREMAL POINTS
- LOWER EXTREMAL POINTS

DIGIT CLASS : few extremal points
 STRIM? CLASS : lot of extremal points
 . more than 5 lowers
 . more than 4 uppers
 COMSP : 1 lower + 1 upper

Wings Weekly Magazine
 P.O. Box 56129
 Boulder Co. 80322-6129
 CITY STATE ZIP CODE

FIGURE 4 :

RESULT OF THE PROCESSING ON UNDERLININGS

FIGURE 3 :

EXTREMAL POINTS

Jewelers Co.
 P.O. Box 39
 Arlington, TN 38002

Prof. Ho Poong Kim
 Philosophy Dept.
 Eastern Illinois University
 Charleston, Illinois 61920

THE WOMENS
 HAZ CHENOBNE & DR
 MECHANICSburg Pa.
 17056

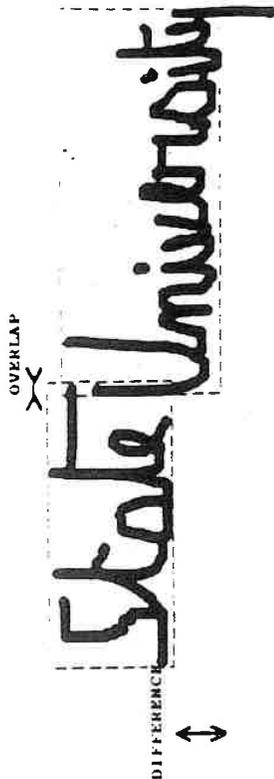
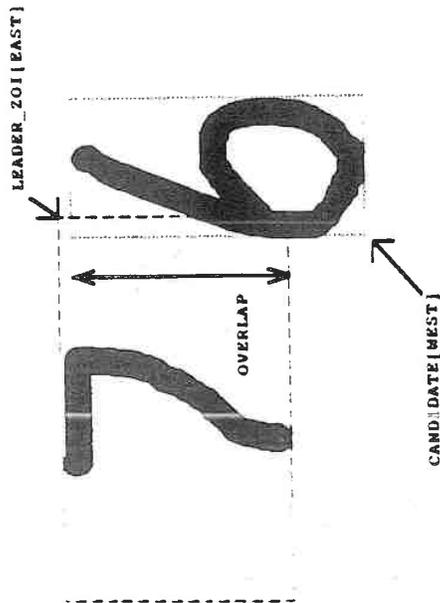
Country Handcraft
 Suite 391 B
 Milwaukee WI
 53201

FIGURE 5 :

BLOCK DETECTION FOR SKEMED LINES

FIGURE 6 :

OTHER EXAMPLES.



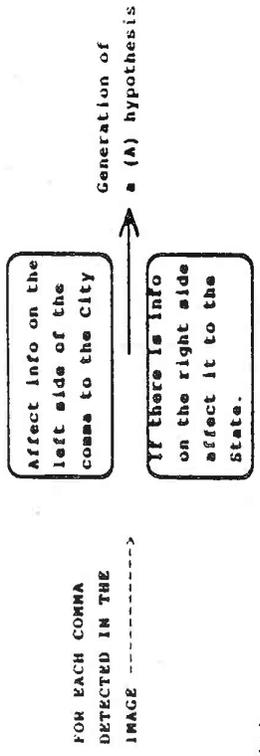
- RULE1) same than for digits
- RULE2) same than for digits
- RULE3) same than for digits
- RULE4) same than for digits
- RULE5) same than for digits
- RULE6) $OVERLAP < MIN (LEADER[WIDTH] , CANDIDATE[WIDTH]) / 4$
- DIFFERENCE < $MIN (LEADER[HEIGHT] / CANDIDATE[HEIGHT]) / 2$

- RULE1) $CANDIDATE[NORTH] > AVERAGE(lower marks of the block)$
- RULE2) $CANDIDATE[SOUTH] < (BLOCK[AVERAGE] + BLOCK[NORTH]) / 2$
- or
- RULE3) $CANDIDATE[NORTH] < BLOCK[NORTH]$
- RULE4) $CANDIDATE[WEST] < LEADER_ZOI[EAST]$
- and
- RULE5) $OVERLAP > LEADER[HEIGHT] / 3$
- and
- RULE6) $OVERLAP > CANDIDATE[HEIGHT] / 3$

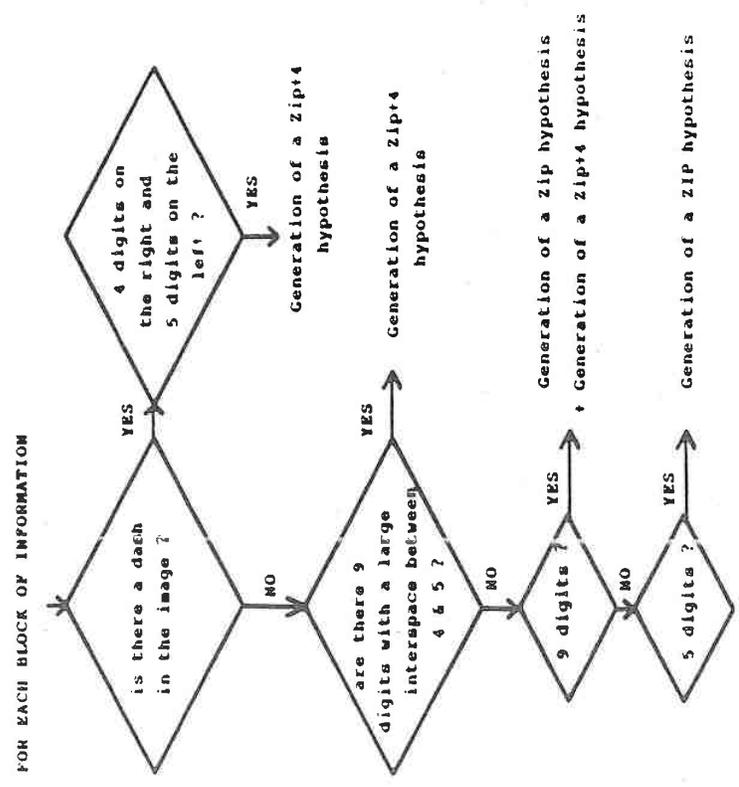
FIGURE 7 :

CRITERIA USED FOR THE BLOCK DETECTION

(1)



(2)



(3)

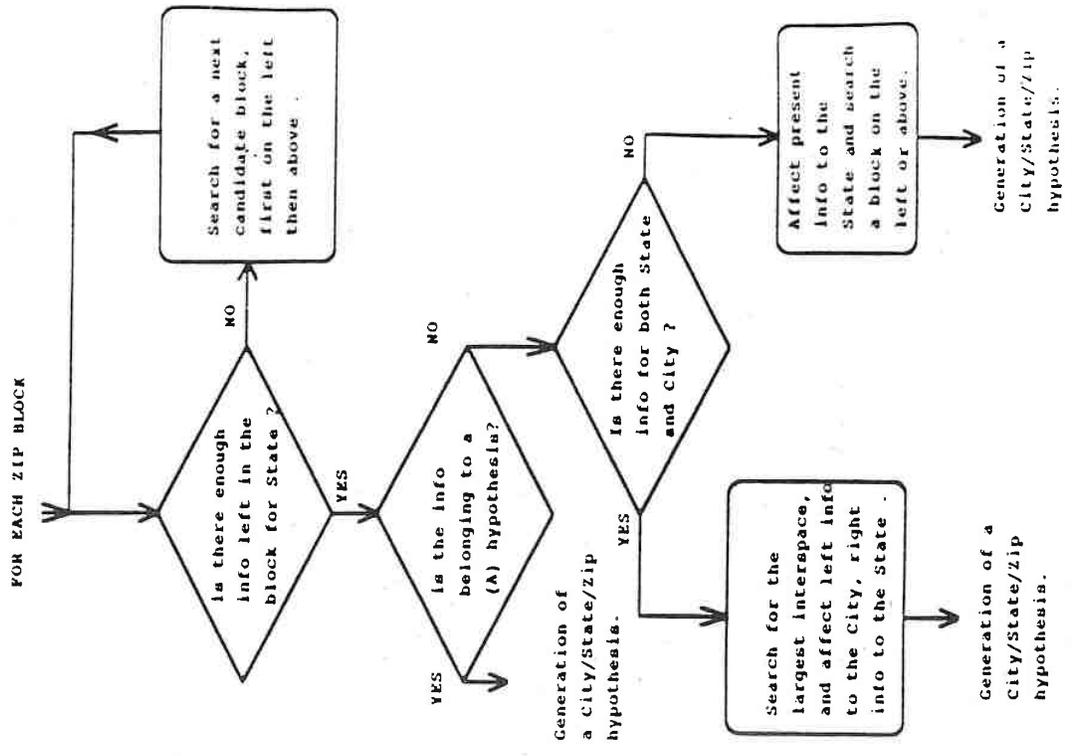


FIGURE 8 :

GENERATION OF CITY/STATE/ZIP HYPOTHESES

PUBLICATION *Wright Weekend Magazine*
 ADDRESS *P.O. Box 56129*
 CITY *Boulder* CO. *80322* STATE ZIP CODE *6129*

Department *ST*
 Box *3987*
 Springfield, *MO* *65808*

EMILY PARKER
 3110 CLAY ST. #3
 SAN FRANCISCO, CA
 94115

Fishn Shack
 PO Box 546-7C
 Camden, MO
 65020

FIGURE 9 : EXAMPLE OF CITY/STATE/ZIP HYPOTHESES PRODUCTION

REFERENCES BIBLIOGRAPHIQUES.

REFERENCE : BOZ 84

AUTEURS : R. Bozinovic et S.N. Shrihari

TITRE : Knowledge-based cursive script interpretation.

SOURCE : 1985, IEEE, T-PAMI, .

REFERENCE : CHO 72

AUTEURS : C.K. CHOW et T. KONEDO

TITRE : Boundary detection of radiographie images by a threshold method

SOURCE : Cargrss 71, booklet TA-7, pp 130-134 North holland, Amsterdam 1972.

REFERENCE : DUE 79

AUTEURS : B.Duerr, W. Haettich, H. Tropf, G. Winkler.

TITRE : A combination of statistical and syntactical pattern recognition applied to classification of unconstrained handwritten numerals.

SOURCE : 1979, Pattern recognition, vol 12, p 189-199

REFERENCE : NAC 84

AUTEURS : N.J. Naccache et R. Shinghal

TITRE : An investigation into the skeletonization approach of Hilditch

SOURCE : 1984, Pattern recognition, vol 17, n°3.

REFERENCE : NAG 77

AUTEURS : R.N. Nagel et A. Rosenfield

TITRE : Computer detection of freehand forgeries

SOURCE : 1977, IEEE, transactions on computers, vol C-26, n°9

REFERENCE : MAR 82

AUTEURS : P. Marthon, A. Bruel, G. Biguet

TITRE : Squelettisation par calcul d'une fonction discriminante sur un voisinage de 8 points

SOURCE : Groupe de traitement numérique des images ENSEEIHT

REFERENCE : PAL 86

AUTEURS : P.W. Palumbo et S.N. Srihari

TITRE : Document image binarisation: Evaluation of algorithms.

SOURCE : 1986, SPIE, vol 697, applications of digital image processing IX

REFERENCE : PAV 81

AUTEURS : T. Pavlidis

TITRE : A flexible parallel thinning algorithm

SOURCE : 1981, IEEE, T-PAMI, jul, p 301-312

REFERENCE : SER

AUTEURS : J.S. Serra

TITRE : Image analysis and mathematical morphology

SOURCE : Academic press 1982

REFERENCE : SHR 84

AUTEURS : M. Shridhar et A. Badreldin

TITRE : Handwritten numeral recognition by tree classification methods.

SOURCE : 1984, image and vision computing, vol 2, n°3.

REFERENCE : SHR 85

AUTEURS : M. Shridhar et A. Badreldin

TITRE : Recognition of isolated and simply connected handwritten numerals

SOURCE : 1986, Pattern recognition, vol 19, 1-12

REFERENCE : SRI 83

AUTEURS : S.N Srihari et R. Bozinovic

TITRE : Use of knowledge in the visual interpretation of cursive script.

SOURCE : 1983, IEEE, PAMI, N°6, pp 655-662

REFERENCE : TAM 78

AUTEURS : H. Tamura

TITRE : A comparison of line thinning algorithms from digital geometry viewpoint

SOURCE: Project PIPS of the agency of industrial and technologie, MITI, Japon.

REFERENCE : THO 78

AUTEURS : S. Thome

TITRE : Pré-traitement du chiffre manuscrit

SOURCE: Congrès AFCET-INRIA 1978 sur la reconnaissance des formes et le traitement d'image., p 736-739

REFERENCE : TOU 72

AUTEURS : J.T. Tou et R.C. Gonzalez

TITRE : Recognition of handwritten characters by topological feature extraction and multilevel categorization

SOURCE: IEEE, transactions on computers, july 1972

REFERENCE : ULL 69

AUTEURS : J.R. Ullman

TITRE : Experiments with the n_triple method of pattern recognition

SOURCE: 1969, IEEE, transactions on computers, vol 18, p 1135-1137

REFERENCE : WES 78

AUTEURS : J. Wezka

TITRE : A survey of threshold selection techniques

SOURCE: 1978, computer graphics and image processing 7, p 259-265

REFERENCE : WHI 83

AUTEURS : J.M. White et G.D. Rohrer

TITRE : Image Thresholding for optical character recognition and applications requiring character image extraction

SOURCE : IBM, Journal. Res develop., vol 27 N° 4, p 400-411

REFERENCE : WOL 69

AUTEURS : J.J. Wolf et W.A. Woods

TITRE : The HWIN speech understanding systeme

SOURCE : 1969, Bolt Beranek and Newman inc., Cambridge, Mass
IEEE; 1977; DA. 1977, pp 784-787

REFERENCE : ZAM 82

AUTEURS : P. Zamperoni

TITRE : Contour tracing of grey_scale images based on 2-D histograms

SOURCE : 1982, pattern recognition, vol 15, n°3, p 161-165



RESUME

Une approche de réduction et de localisation de l'information a été développée et intégrée dans un système de lecture du courrier manuscrit américain. Une étape préliminaire de binarisation de l'image diminue le volume d'information à traiter, atténue la présence d'éléments superflus (trames, bruits, graphiques) en évitant de détériorer les caractères. Une méthodologie originale, fondée sur des critères géométriques simples, montre l'intérêt d'utiliser les point extrêmes des formes pour décrire l'image et localiser l'information. Ces points pertinents permettent de détecter les caractères isolés, les mots et les soulignements..... Le type de description des formes qui en découle, facilite le regroupement des formes en blocs d'information. Ces points caractéristiques sont aussi utilisés pour restaurer les chiffres brisés et séparer les chiffres connectés. A partir de cette structuration de l'image en blocs d'information, une méthode de génération d'hypothèses "code postal + état + ville" choisit les formes à soumettre au système de lecture en fonction de la syntaxe de l'adressage postal.

SUMMARY

An approach of reduction and location of information has been implemented and integrated in an handwritten recognition system for the american mail processing. A preliminary task of image binarization reduces the volume of information processed, attenuates the presence of superfluous elements (texture, noise, graphics), avoiding to spoil characters. An original methodology, founded on simple geometrical criteria, shows the interest to use shapes extremal points for image description and for information location. These relevant points allows to detect isolated characters, strings and underlinings....This type of shapes description that follows, facilitates the clustering of the patterns in information blocks. these characteristic points are so used to restore broken digits and to separate connected digits. From this image structuration in information blocks, a method of "postal code + state + city" hypothesis generation chooses forms to submit through recognition stage, using postal address syntax.

Mots clefs : reconnaissance du manuscrit, binarisation, amincissement, contour, description de l'image, primitives de formes, segmentation, localisation de l'information.

keys words : handwritten recognition, binarization, thinning, contour, image description, shapes primitives, segmentation, information location.