LOCALISATION DE POINTS CARACTERISTIQUES DANS UN DESSIN : APPLICATION A LA RECONNAISSANCE DU CHINOIS MULTIFONTE

S. ZHANG, B. TACONET, A. FAURE

Université du Havre L.A.C.O.S. Place Robert Schuman 76610 LE HAVRE

RESUME

Nous présentons une méthode de reconnaissance statistique des caractères chinois multipolices, qui peut être transposée à tout dessin au graphisme complexe.

Les noeuds et des extrémités, extraits sur le squelette du caractère lissé, constituent les éléments caractéristiques de chaque classe . Une première classification est effectuée selon un indice de complexité, somme pondérée du nombre de noeuds et d'extrémités. Les coefficients de pondération sont choisis de telle sorte qu'une jonction intempestive entre une extrémité et le corps d'une branche ne modifie pas la valeur de l'indice. Une sous-classification est faite ensuite par comparaison entre les coordonnées des points caractéristiques de la forme et du modèle. Cependant, aucun élément retenu ne permet d'étiqueter ces points: il est donc nécessaire de faire une mise en correspondance entre les points du caractère inconnu et ceux du caractère modèle, en minimisant la distance. Le degré de ressemblance est calculé à l'aide d'un coefficient de similarité, qui décroît avec la distance. Des essais ont été effectués sur une base de données d'un millier de caractères environ, pour les plus faibles valeurs du coefficient de complexité. Le taux de reconnaissance est supérieur à 96 %.

SUMMARY

Here is presented a new statistical method of multifont Chinese character recognition, which can fit to any complex drawing.

Nodes and endings, extracted from the smoothed character skeleton, are the basic features of the classification. The first classification is made according to an index of complexity, defined as the weighted sum of nodes and endings. A junction between an ending and a stroke does not disturb the index value. The second level of classification is then made by comparing the coordinates of the characteristic points of the unknown form and the model. Note that it is absolutely necessary to match the points of the unknown form and the model. This is realized by minimising the Euclidian distance. A similarity index, decreasing as the distance, gives an evaluation of the likeness. Experiments on a database of one thousand of Chinese characters have been done, involving the lowest complexity characters. The recognition rate overheads 96%.

I. Introduction

En matière de reconnaissance de caractères, celle des caractères chinois se révèle la plus délicate, par la conjugaison de plusieurs sources de difficultés: une base de données volumineuse (plusieurs milliers de caractères courants), un graphisme très fréquemment complexe (nombre élevé de traits) et des formes très voisines pour certains groupes de caractères.

C'est à partir des années 60, qu'ont commencé les recherches en reconnaissance de caractères chinois imprimés. Le premier article a été publié par R. Casey et G. Nagy [1] en 1966, qui utilisent une méthode de masques. P.P.Wang [2] a étudié l'effet des transformations de Fourier discrète, rapide,et de la transformation d'Hadamard portant sur 63 radicaux de caractères chinois.Yamamoto et al.[3] ont développé une approche hiérarchique de la coïncidence avec les modèles . Sakai et Mori [4] ont défini et exploité la notion de similarité synthétique, mais ces méthodes semblent défaillantes pour le cas multifonte. On peut trouver aussi quelques articles traitant de la reconnaissance de caractères chinois multifontes, par exemple, Umeda [5] utilise des caractéristiques locales situées à la périphérie("mesh/peripheral pattern"); P. Chen et al. [6] définissent un code de relation entre segments, mais ces travaux méritent d'être approfondis. Nous présentons une nouvelle méthode de reconnaissance statistique des caractères chinois multifontes, fondée sur la stabilité de la position de points caractéristiques simples: noeuds, extrémités, pour une même catégorie de police de caractères. C'est une extension de l'approche développée au L.A.C.O.S. pour les caractères latins [7],[8].

II. Le principe de base et l'algorithme

II-1. L'extraction des points caractéristiques

Nous extrayons deux sortes de points caractéristiques:

1). les extrémités.

2). les noeuds.

A chaque noeud, on associe un coefficient de complexité qui est lié au nombre de ses branches(tableau 1.):

CN = nb - 2

où CN est le coefficient de noeud, et nb le nombre de branches

tableau 1

NB	CN	caractère
3	3-2=1	 (terre)
4	4-2=2	 (terre)
5	3	 (grand)
6	4	 (bois)
7	5	pas de caractère courant
8	6	 (riz)

avec : NB nombre de branches

Nous définissons le paramètre NPC (le nombre pondéré des points caratéristiques) exprimant la complexité d'un caractère:

NPC = NE + CN

où NE est le nombre des extrémités.

L'ensemble des caractères chinois peut se diviser en sous-classes suivant le paramètre de complexité NPC:

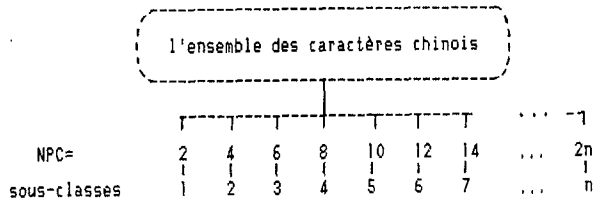

Figure 1. La classification de l'ensemble des caractères chinois

D'après la constitution des caractères chinois et définition de NPC,nous obtenons toujours des sous-classes avec un NPC pair.

A titre d'exemple, le caractère "  " est dans la sous-classe 3 (NPC=4+2=6).

L'extraction est effectuée sur la forme squelettisée du caractère. Pour s'adapter au cas multifonte, toutes les coordonnées des points caractéristiques sont normalisées, par multification d'un coéfficient d'échelle sur une fenêtre-type de dimension 60*60 pixels.

II-2. Importance du prétraitement

Le premier niveau de classification revet une importance capitale dans les performances de la reconnaissance; lorsque les caractères ont un tracé d'épaisseur constante, exempt de coupure, l'emploi d'une méthode classique de squelettisation, précédée d'un lissage, suffit et le taux d'erreur sur le calcul du nombre de complexité est très faible. Cependant, le cas d'un tracé épais, et d'épaisseur variable, doit être étudié, car certains styles de fontes imprimées,très populaires en Chine, possèdent de telles fioritures (style de Fangsong).

En effet, une squelettisation de type classique produit des branches parasites à cause des ornements situés aux extrémités (figure 2).

草文见学
论会政计

figure 2

Une opération d'ébarbulage avec seuil fixe risque d'amputer le squelette de branches utiles. C'est pourquoi nous avons cherché à rendre l'ébarbulage plus sélectif, en affectant à chaque point du squelette son degré de profondeur au sens du 8-voisinage (figure 3).Le seuil d'ébarbulage est fonction de la longueur de la branche et de la profondeur au noeud de liaison.

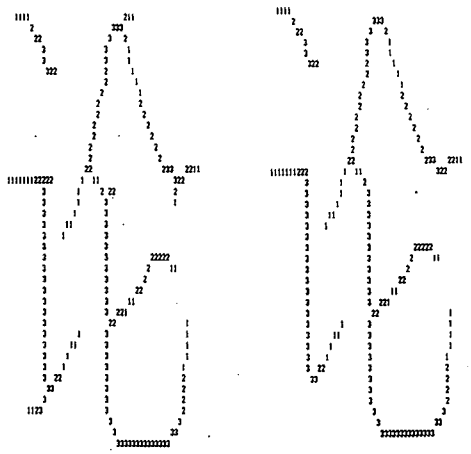

Figure 3

II-3. La création du dictionnaire des caractères-modèles

Il est nécessaire de créer un dictionnaire des points caractéristiques des caractères modèles afin de faire ,dans chaque sous-classe, le calcul de la similarité avec le caractère inconnu. Les caractères modèles sont divisés en sous-classes selon le paramètre de complexité NPC et les coordonées de position des points caractéristiques sont mémorisées. Les coordonnées ont été normalisées dans la fenêtre-type.

Le cas multifonte fait apparaître plusieurs modèles par caractère, et il est naturel de composer un modèle unique en effectuant une moyenne sur toutes les polices. Pour ce faire, nous avons choisi une méthode qui impose aux modèles d'avoir une même structure, c'est-à-dire même nombre caractéristique de complexité, et même nombre d'extrémités. Parmi les polices à reconnaître, on retient celle qui paraît la plus proche de la moyenne, comme référence de départ.

On cherche d'abord à mettre en correspondance les extrémités. Autour d'une extrémité du modèle de la police de référence, on délimite une aire d'agrément (12x12 pixels).Une extrémité d'un modèle d'une autre police est appariée si les directions des branches qui en partent sont voisines (distance d'une unité sur le codage de Freeman), puis si la distance euclidienne est minimale, en cas de candidats multiples. La figure 4 illustre un cas très rare pour lequel l'appariement se fait de façon erronnée.

Les noeuds sont de deux sortes : dans la première catégorie, une branche relie le noeud à une extrémité. La mise en correspondance se calque sur celle des extémités, en tenant compte des coefficients de noeuds . Reste l'autre catégorie, celle des noeuds qui sont reliés exclusivement à des noeuds. En général cette catégorie comporte peu de points, et par conséquent on peut envisager d'employer une méthode optimale rigoureuse (méthode arborescente, ou méthode hongroise).

Les coordonnées d'un point caractéristique du caractère modèle multifonte se calculent selon la formule :

$$x_M = 0.5[\mathrm{Min}(x_k) + \mathrm{Max}(x_k)]$$

II-4. L'algorithme de similarité

Contrairement aux autres algorithmes qui traitent de la similarité, on ne peut pas

représenter les points caractéristiques d'un caracatère sous forme vectorielle, puisque l'ordre des points caractéristiques est incertain. Pour estimer la similarité entre 2 points, nous définissons une zone de recherche (12*12 pixels) centrée en un point caractéristique du caractère inconnu. Après examen de tous les points caratéristiques d'un caractère modèle dans cette zone, on retient comme point correspondant celui qui est le plus proche du point du caractère inconnu (Figure 4), au sens de la distance euclidienne.

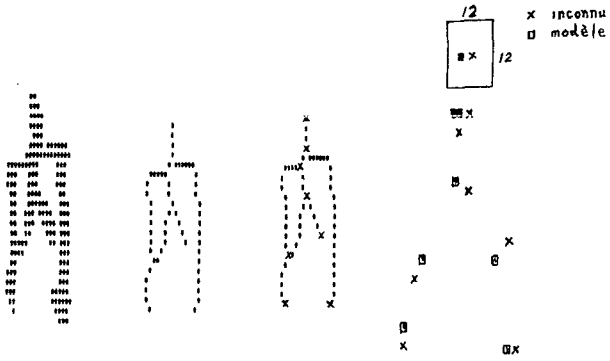

Figure 4

Nous construisons alors une fonction de similarité qui décroît avec la distance (Figure 5):

$$S_i = \begin{cases} 1 - \frac{D_i}{16}; & D_i \leq 12 \\ 0; & D_i > 12 \end{cases}$$

où: S_i désigne la similarité, pour un couple de points caractéristiques, entre le caractère modèle et le caractère inconnu.

D_i désigne la distance euclidienne minimale entre le caractère modèle et le caractère inconnu, pour un couple de points caractéristiques.

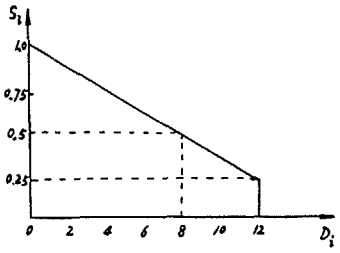

Figure 5. Fonction de similarité

La similarité globale S entre le caractère inconnu et le caractère modèle est la somme des similarités partielles:

$$S = \sum_{i=1}^{NPC} S_i$$

Nous marquons les points caractéristiques qui sont déjà utilisés au cours du calcul de la similarité globale et ils ne sont plus considérés dans la suite.

Les fonctions de similarité entre un caractère inconnu et chaque caractère modèle sont calculées de façon syntématique et celui dont la valeur de la similarité est la plus élevée est retenu dans la décision finale.

III. Les résultats d'expériences

En appliquant la méthode mentionnée précédemment, nous avons mené des expériences sur une base de données d'envivon mille caractères chinois imprimés de la sous-classe 3 (NPC=6) à la sous-classe 10 (NPC=20), chaque sous-classe comportant en moyenne une centaine de caractères. La structure de la reconnaissance est indiquée dans la figure 6. Pour une même police de caractères(avec épaisseur du tracé constant), avec apprentissage, le taux de reconnaissance dépasse 98%.Pour plusieurs polices voisines(cas multifonte), il dépasse 96%.

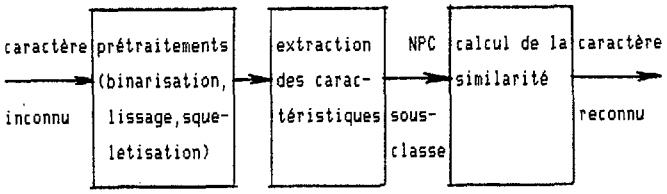

Figure 6. La structure de la reconnaissance

IV. Conclusion

Comparée avec d'autres méthodes de reconnaissance des caractères chinois imprimés, notre méthode a des avantages suivants:

1. Elle s'adapte à la reconnaissance des caractères chinois multifontes. Evidemment, cette méthode dispose d'une bonne tolérance aux variations de position des points caractéristiques. La figure 7. donne quelques groupes des caractères bien reconnus qui proviennent de fontes différentes.

2. L'approche de la classification et l'algorithme de similarité sont tous assez simples. Cela permet d'obtenir une vitesse de traitement assez élevée, ce qui est capital pour

la reconnaissance d'une volumineuse base de caractères(typiquement pour les caratères chinois).

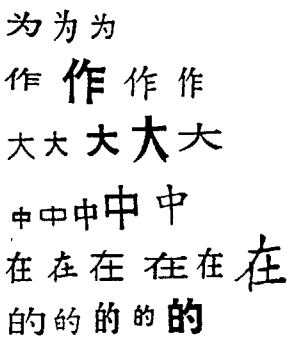

Figure 7

3. Cette méthode a été conçue pour permettre la reconnaissance en dépit de certaines dégradations causées par la qualité de l'impression. Dans l' exemple donné figure 8, le caractère "据" devient "据" dont les extrémités 1 et 2 deviennent des noeuds à cause des contacts et le noeud 3 devient une extremité à cause de la coupure. Mais grâce à la définition du paramètre de complexité NPC, celui-ci reste inchangé et le caractère est reconnu malgré ces modification de structure. Mentionnons aussi le cas qui aboutit à la diminution du nombre de points caractéristiques NPC, par exemple, "研" devient "研" où manquent deux extrémités, ce probleme peut être résolu par l'ajout d'un modèle supplémentaire dans une autre sous-classe (de nombre de complexité diminué de 2 unités).

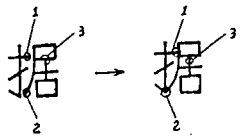

Figure 8

Les deux principales sources de rejet, et de fausse reconnaissance, sont les coupures dans le tracé,et la présence de branches parasites surgies au cours de l'opération de squelettisation: elles conduisent à une valeur erronnée du nombre de complexité. Pour y remédier, nous envisageons d'agrandir la notion de sous-classe: elle comportera non seulement les modèles de même nombre de complexité, mais aussi ceux de complexité voisine. Des résultats partiels très encourageants nous invitent à poursuivre la recherche dans cette voie.

Bibliographie :

[1] R.Casey and G.Nagy, "Recognition of printed chinese characters," IEEE Transaction on Electronic Computer. Vol. EC-15. No.1 Feb. 1966. pp.91-101.

[2] P.Wang and R.Shian," Machine recognition of printed chinese characters via transformation algorithms ", Pattern Recognition. Vol.5. pp. 303-321 (1973).

[3]S.Yamamoto,A.Nakajima,K.Nakata, " Chinese characters recognition by hierarchical pattern matching ", Proc. 1st. Int. Conf. on Pattern Recognition. pp. 187-196(1973).

[4] Kenichi Mori and Isao Masuda " Advances in recognition of chinese characters ", Proc. 5th. Int. Conf. on Pattern Recognition. Miami. pp. 692-702 (1980).

[5] Michio Umeda " Recognition of multi-font printed chinese characters ", 6th. Int. Conf. on Pattern Recognition . pp. 793-796 (1982).

[6] P.Chen,Y.Chen,W.Hsu " Stroke relation coding -- a new approach to the recognition of multi-font printed chinese characters ", International Journal of Pattern Recognition and Artificial Intelligence. Vol.2. No.1(1988). pp.149-160.

[7] F.Yatim,B.Taconet " Système d'acquisition et de reconnaissance de caractères alphanumériques multifontes." 6°Congrès R.F.I.A., Antibes,1987, pp. 419 - 428 .

[8] F.Yatim,"Reconnaissance de caractères multifontes par une structure pluri-procédure" Thèse de troisième cycle, Lille, France, 1988.