Identification du scripteur utilisant les fréquences de graphèmes

Application à des données en-ligne

Jinpeng Li[*,**]

[*] *IVC (IRCCyN-CNRS-UMR 6597)*
Site de la Chantrerie -rue Christian Pauc
44306 Nantes France
[**] *Computer Science Faculty*
Guangdong University of Technology
51000 Chine
jinpeng.li@etu.univ-nantes.fr

RÉSUMÉ. *Nous proposons une méthode permettant d'identifier le scripteur d'un texte de quelques lignes manuscrites en-ligne à partir d'une base de documents manuscrits de références. La méthode est basée sur l'apparition de motifs récurrents dans le document requête et les documents de la base de référence. Ces motifs correspondent à des graphèmes extraits automatiquement par un algorithme de segmentation. Cette méthode est évaluée sur deux bases de données, l'une de 120 scripteurs en langue française pour l'apprentissage, l'autre de 200 scripteurs en langue anglaise pour le test. scripteurs. Les résultats obtenus (87,6%) sont comparables avec ceux obtenus avec une méthode plus complexe nécessitant la reconnaissance explicite du texte.*

ABSTRACT. *We propose a method for writer identification which is capable of retrieving the identities from a reference database according to a piece of online handwritten script. The identification is based on the graphemes extracted automatically by a simple segmentation algorithm, and this method is evaluated in two databases, French database (120 writers) for learning the best parameters and English database (200 writers) for testing. The top1 accuracy (87.6%) is comparable with a more complex approache.*

MOTS-CLÉS : *Identification de scripteur, Graphèmes, Écriture manuscrite en-ligne, Recherche d'information.*

KEYWORDS: *Writer identification, Graphemes, Online handwriting, Information retrieval.*

1. Introduction

L'identification du scripteur est une technique de reconnaissance biométrique de type comportementale. A partir d'un dispositif de saisie de l'information manuscrite en-ligne[1], notre approche permet de chercher automatiquement l'identité du scripteur parmi un ensemble d'utilisateurs. Nous utilisons pour l'identification le style graphique de l'écriture qui est une caractéristique propre à chaque personne. Tan (Tan *et al.,* 2009a, Tan *et al.,* 2009b) a proposé une méthode d'identification basée sur des styles de caractères. Lors de l'identification du scripteur d'un nouveau texte, il s'agit d'abord de reconnaître les caractères par un système de segmentation/reconnaissance du texte, puis de calculer la fréquence d'utilisation de chaque prototype, l'utilisateur avec le même profil de fréquence est désigné comme le scritpeur. Pour comparer les fréquences d'apparition des prototypes, la distance du Chi^2 est une meilleure mesure. Dans (Tan *et al.,* 2009a), ils utilisent le coefficient d'information alphabétique pour améliorer la performance.

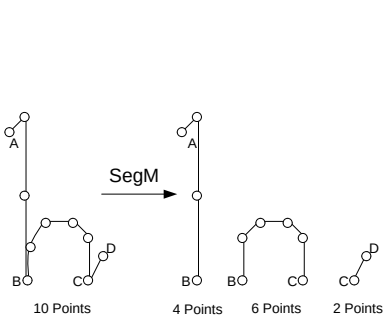
Dans cet article nous nous plaçons dans la continuité des travaux de Tan. L'apport de notre approche réside dans l'échelle d'analyse sur des données en-ligne qui évite l'étape coûteuse de reconnaissance tout en obtenant d'aussi bons résultats. En effet au lieu de découper l'écriture en caractères, ce qui nécessite une reconnaissance complète du texte, nous segmentons l'écriture en unités graphique d'écriture appelées graphèmes qui correspondent généralement à des morceaux de caractère. Par ailleurs des prototypes de graphèmes sont définis par k-means pour identifier scripteurs. La section 2 décrit l'algorithme de la segmentation en petits graphèmes. Les résultats expérimentaux sont montrés dans la dernière section .

2. Segmentations en graphèmes

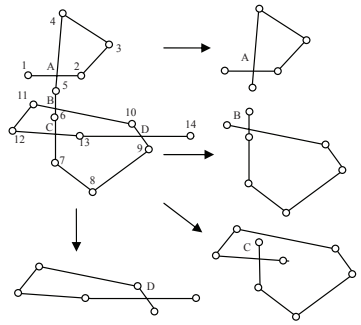
Nous proposons deux solutions de segmentation de l'écriture en petits graphèmes. La première segmentation utilise les minimums verticaux de l'écriture manuscrite et est appelée SegM (Segmentation sur Minimum). Cette méthode a été appliquée également dans Schomaker et al.(Schomaker *et al.,* 2004, Schlapbach *et al.,* 2008) sur de l'écriture hors-ligne. Concernant l'écriture en-ligne, les minimums locaux du signal sont considérés comme des points de segmentation. La figure 1a montre un exemple de segmentation pour un tracé, et il en résulte 2 points minimums (B et C), ainsi trois segments seront générés ($A \rightarrow B$, $B \rightarrow C$ et $C \rightarrow D$). La seconde segmentation proposée,appelée SegL (Segmentation Loop), recherche les boucles dans l'écriture manuscrite. En effet, les experts considèrent la boucle comme une caractéristique importante pour l'identification (Huber *et al.,* 1999). L'extraction des boucles est basée sur la recherche d'intersection de deux paires de points sur la trajectoire. La Fig. 1b montre un exemple de segmentation pour un caractère f, il existe 4 points d'intersection (A, B, C et D) correspondant à 4 boucles.

1. L'écriture en-ligne est définie par la séquence de points du tracé. L'information temporelle est donc conservée contrairement à l'écriture dite hors-ligne.

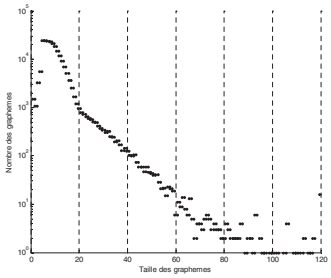
La Fig. 1a montre un phénomène intéressant : le troisième segment est court (2 points) alors que le second est long. Ce phénomène n'est pas rare comme le montre l'histogramme de la Fig. 1c représentant le fréquence des segments FR_120 segmentée par la méthode SegM en fonction de leur taille en nombre de points. La plupart des graphèmes ont moins de 20 points. Au delà de 20 points, leur fréquence diminue nettement. On note également qu'il y a beaucoup de petits graphèmes de 2 ou 3 points qui de part leur taille sont assez instables. Ces nombreux petits graphèmes risquent de perturber l'identification et donc peuvent être considérés comme du bruit. Toutefois leur élimination pure et simple serait préjudiciable. Nous proposons une stratégie qui fusionne les segments courts avec leur voisin gauche (2 voisins normalement) dans une stroke sauf le premier segment qui est fusionné avec celui suivant. La figure 1d montre un exemple de fusion des segments de moins de 3 points (seuil de fusion).



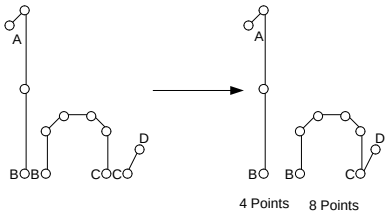
(a) Exemple de segmentation sur les minimums verticaux (SegM)



(b) Exemple de boucles de segmentation (SegL)



(c) Histogramme des tailles de graphèmes (SegM sur FR_120)



(d) Exemple de fusion des segments trop courts (seuil = 3)

Figure 1 – Des stratégies de segmentation

3. Résultats expérimentaux

Bases des données : La base de mots IRONOFF (Viard-Gaudin *et al.,* 1999) est utilisée pour construire les prototypes de graphèmes. Cette base de données comporte

16 585 mots écrits par 373 sujets. Pour l'identification, nous utilisons deux bases de textes FR_120 et Reuters200 (Said *et al.,* 1998, Tan *et al.,* 2009b). Chaque scripteur a saisi deux textes, un texte de référence et un pour le test d'identification. FR_120 comporte 240 textes français sur des sujets libres issus de 120 scripteurs et dont la longueur varie entre 86 caractères et 972 caractères avec une longueur moyenne de 465 caractères. La base Reuters200 comporte 400 textes écrits en Anglais : 200 scripteurs ont recopié chacun deux documents de l'agence de presse Reuter. La taille de la base de référence de Reuters200 est plus grande que celle de FR_120 et ses textes sont plus courts, ils comptent en moyenne 304 caractères, soit approximativement six lignes. Par conséquent, le corpus Reuters200 est plus difficile que celui de FR_120.

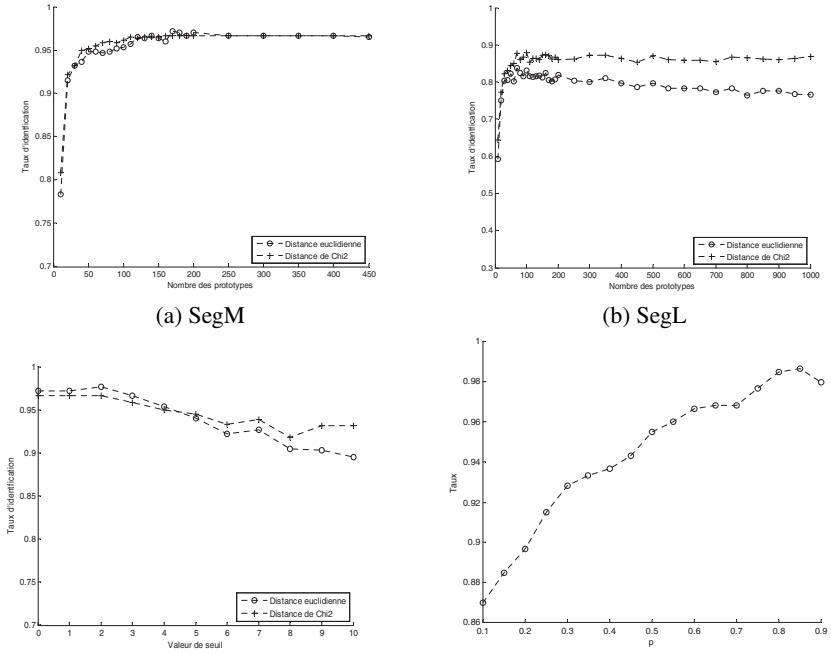
Protocole de test : L'apprentissage de notre approche consiste à deux méta-paramètres : le nombre de prototypes d'allographe (appris sur IRONOFF) et le cas échéant le seuil de fusion de des petits graphèmes. Pour évaluer cet apprentissage nous utilisons la base FR_120 (apprentissage) puis les méta-paramètres optimums sont testés sur Reuters200 (test). À cause de l'initialisation aléatoire du clustering, tous les taux présentés correspondent à la valeur moyenne des expériences répétées cinq fois (on observe une variance moyenne de 0.006%).

SegM et SegL : Les figures 2a et 2b montrent respectivement les taux d'identification en première position par SegM et SegL sur FR_120. Dans la première étude (SegM), le nombre de prototypes évolue de 20 à 450. On observe qu'avec moins de 100 prototypes, les performances sont dégradées. Le taux est 97,17% pour 170 prototypes sur la base du FR_120, soit 3 à 4 scripteurs non-identifiés en première position sur les 120 scripteurs. Dans l'étude avec SegL (Figure 2b), le nombre de prototypes évolue de 20 à 1000. Nous obtenons meilleur taux de 87,83% sur la base FR_120 avec 100 prototypes. Apparemment les performances avec les graphèmes de type SegL sont inférieures à celles avec SegM. En effet SegL utilise seulement une partie du tracé de l'écriture, seules les boucles sont extraites et le reste du tracé n'est pas utilisé. Le reste de l'écriture contient probablement également des informations importantes.

Traitement des petits graphèmes : Comme mentionné auparavant, la méthode avec SegM produit de nombreux petits graphèmes. La Figure 2c montre le taux d'identification sur FR_120 selon la valeur du seuil de fusion. La valeur optimale du seuil est une longueur de 2 points et le taux obtenu est de 97,67%.

Combinaison : Une combinaison linéaire de deux méthodes, $dist(w_i, w_j, p) = p \times distSegM(w_i, w_j) + (1 - p) \times distSegL(w_i, w_j)$, est utilisée pour retrouver le document référence. Les fonctions $distSegM$ et $distSegL$ sont les fonctions de distances entre deux documents basées sur SegM et SegL. Nous avons balayé le poids (p) de 0,1 à 0,9 avec un pas de 0,05. La Fig. 2d montre les taux d'identification sur la base FR_120 avec cette combinaison. Le taux de 98,67% est obtenu avec la distance euclidienne en choisissant $p = 0,85$, soit 1,6 scripteurs en moyenne mal reconnus.

Bilan et comparaison : Le Tableau 1 résume les taux en première position avec toutes les méthodes étudiées et toutes leurs combinaisons. La première ligne indique les taux obtenu par Tan et al. sur les deux mêmes bases. Leur meilleur taux est de



(a) SegM

(b) SegL

(c) Fusion des petits graphèmes par SegM

(d) Combinaison de SegM+Fusion (170 prototypes) et SegL (100 prototypes) sur FR_120

Figure 2 – Taux d'identification sur FR_120

99,16% sur la base FR_120, soit 1 scripteur sur 120 non retrouvé en première position, et de 87% sur la base Reuters200. Dans nos travaux, tout d'abord nous faisons l'apprentissage sur FR_120 : 98,67% d'identification pour optimiser des paramètres. Puis le test sur Reuters200, le taux de 82,7% d'identification est obtenu. En effect ce nombre de prototypes peut-être n'est pas suffisant à la base Reuters200 plus compliquée. Ainsi dans la dernière colonne de tableau 1, un expérimental ayant des parametres optimisées a été étudié sur Reuters200. Nous avons obtenu un taux de 87,6%.

4. Conclusion

Deux méthodes de segmentations en graphèmes pour l'identification du scripteur ont été étudiées. Son taux optimisé est comparable avec ceux obtenus avec une méthode plus compliquée nécessitant la reconnaissance explicite du texte et sa segmentation en lettres. Concernant des perspectives, nous pouvons continuer à quantifier des graphèmes de différentes tailles influençant l'identification. Par ailleurs il existe plusieurs définitions de boucles selon ses directions et ses croisées.

Tableau 1 – Taux d'identification

	FR_120 (apprentissage)	Reuters200 (test)		Reuters200 (Optimisé)
Tan et al.	99,17%	87%		87%
SegM	97,17%	77%	(NP = 170, Eu)	84,8%(NP=400, Chi^2)
SegL	87,83%	64,9%	(NP = 100, Chi^2)	68%(NP=600, Chi^2)
SegM + F	97,17%	79,6%	(NP = 170,T = 2, Eu)	85,8%(T=2, NP=400, Chi^2)
SegM + SegL	98,33%	81%	(NP_{SegM} = 170, NP_{SegL} = 100, p=0,8, Eu)	86,7%(NP_{SegM}=400, NP_{SegL}=600, p=0,8, Chi^2)
(SegM+F) + SegL	98,67%	82,7%	(NP_{SegM} = 170, NP_{SegL} = 100, p=0,85 Eu)	87,6%(NP_{SegM}=400, NP_{SegL}=600, p=0,8, Chi^2)

NP : Nombre des prototypes Eu : La distance euclidienne T : Seuil de fusion
F : Fusionner Chi^2 : La distance du Chi^2

Remerciements : Nous remercions en particulier mes directeurs, Harold MOUCHERE, Guo Xian TAN et Christian VIARD-GAUDIN pour les soutiens.

5. Bibliographie

Huber R., Headrick A. M., *Handwriting Identification : Facts and Fundamentals*, CRC Press, LCC,, 1999.

Said H., Baker K., Tan T., « Personal Identification Based on Handwriting », *ICPR '98 -Volume 2*, IEEE Computer Society, Washington, DC, USA, p. 1761, 1998.

Schlapbach A., Liwicki M., Bunke H., « A writer identification system for on-line whiteboard data », *Pattern Recogn.*, vol. 41, n° 7, p. 2381-2397, 2008.

Schomaker L., Bulacu M., « Automatic Writer Identification Using Connected-Component Contours and Edge-Based Features of Uppercase Western Script », vol. 26, IEEE Computer Society, Washington, DC, USA, p. 787-798, 2004.

Tan G. X., Viard-Gaudin C., Kot A., « Online Writer Identification Using Alphabetic Information Clustering », *SPIE-IS&T Electronic Imaging : Document Recognition and Retrieval XVI*, vol. 7247, p. 72470F-1 - 72470F-8, 2009a.

Tan G. X., Viard-Gaudin C., Kot A. C., « Automatic writer identification framework for online handwritten documents using character prototypes », *Pattern Recogn.*, vol. 42, n° 12, p. 3313-3323, 2009b.

Viard-Gaudin C., Lallican P. M., Binter P., Knerr S., « The IRESTE On/Off (IRONOFF) Dual Handwriting Database », *ICDAR '99*, IEEE Computer Society, Washington, DC, USA, p. 455, 1999.

ANNEXE POUR LE SERVICE FABRICATION
A FOURNIR PAR LES AUTEURS AVEC UN EXEMPLAIRE PAPIER
DE LEUR ARTICLE ET LE COPYRIGHT SIGNE PAR COURRIER
LE FICHIER PDF CORRESPONDANT SERA ENVOYE PAR E-MAIL

1. ARTICLE POUR LA REVUE :

L'objet. Volume ? – n°0/2000

2. AUTEUR :

Jinpeng Li[*,**]

3. TITRE DE L'ARTICLE :

Identification du scripteur utilisant les fréquences de graphèmes

4. TITRE ABRÉGÉ POUR LE HAUT DE PAGE MOINS DE 40 SIGNES :

Identification du scripteur de graphèmes

5. DATE DE CETTE VERSION :

11 février 2010

6. COORDONNÉES DES AUTEURS :

– adresse postale :
[*] IVC (IRCCyN-CNRS-UMR 6597)
Site de la Chantrerie -rue Christian Pauc
44306 Nantes France
[**] Computer Science Faculty
Guangdong University of Technology
51000 Chine
jinpeng.li@etu.univ-nantes.fr
– téléphone : 33 2 40 68 30 48
– télécopie : 33 2 40 68 32 32
– e-mail : jinpeng.li@etu.univ-nantes.fr

7. LOGICIEL UTILISÉ POUR LA PRÉPARATION DE CET ARTICLE :

LaTeX, avec le fichier de style `article-hermes.cls`,
version 1.2 du 03/03/2005.

8. FORMULAIRE DE COPYRIGHT :

Retourner le formulaire de copyright signé par les auteurs, téléchargé sur :
`http://www.revuesonline.com`

SERVICE ÉDITORIAL – HERMES-LAVOISIER
14 rue de Provigny, F-94236 Cachan cedex
Tél : 01-47-40-67-67
E-mail : revues@lavoisier.fr
Serveur web : http://www.revuesonline.com