Identification de Scripteurs Utilisant les Distributions d'Allographes

Writer Identification Using Allograph Distributions

Christian Viard-Gaudin [1], Guo Xian Tan [1,2], Alex C. Kot [2]

[1] IRCCyN, UMR CNRS 6597, Ecole Polytechnique de l'Université de Nantes, France
[2] Nanyang Technological University, Singapore
christian.viard-gaudin@univ-nantes.fr ; tanguoxian@pmail.ntu.edu.sg ; eackot@ntu.edu.sg

Manuscrit reçu le 15 décembre 2009

Résumé et mots clés

Ce papier propose une méthode permettant d'identifier le scripteur d'un texte quelconque de quelques lignes en le comparant à des écritures de références. La comparaison est basée sur une mesure de mise en correspondance des distributions des allographes de lettres représentatifs des styles d'écriture. Un système automatique segmente le texte en lettres, puis classe chaque lettre de manière probabiliste parmi les prototypes disponibles pour cette lettre. Deux bases de complexité différentes sont utilisées pour valuer ce système. Un taux d'identification de 99,2 % est obtenu sur une base de recherche de 120 textes écrits en français, tandis qu'il se situe à 87 % sur une base de recherche de 200 textes écrits en anglais. Cette méthode est développée sur de l'écriture en ligne.

Identification de scripteur, recherche d'information, écriture manuscrite en-ligne, k-plus-proches-voisins, allographe.

Abstract and key words

A method is proposed to allow the retrieval of the identity of the writer of a non-constraint handwritten text by matching it with some reference handwritten documents. The matching is based on a metric computed on the distributions of the allograph of the letters featuring a unique writing style. An automatic system segments the text into characters and assigns a partial membership to the different representative prototypes of the considered letter of the Roman alphabet. Two different datasets are used to assess this system. A writer identification rate of 99.2% is obtained when the reference dataset is composed of 120 French documents. On the other dataset with 200 English texts, the identification rate reaches 87%. Online handwriting is considered by this system.

Writer identification, information retrieval, online handwriting, k-nearest neighbor, allograph.

Remerciements

Cette recherche est soutenue conjointement par Nanyang Technological University (NTU, Singapour), le programme du Ministère des Affaires Etrangères Merlion PhD et le projet ANR Technologie Logicielle (CIEL 06-TLOG-009).

1. Introduction

La diversité des solutions de saisie ouvre la voie à l'acquisition et au stockage d'une quantité sans cesse croissante de documents de différentes natures contenant de l'écriture manuscrite en-ligne. Parmi les technologies en fort développement, nous pensons aux solutions de saisie de type papier/stylo digital qui ouvrent sur des applications à très large échelle et pour des documents de complexité quelconque [Jain 03]. L'ergonomie de ces solutions se rapproche maintenant de l'usage d'un stylo conventionnel. Deux technologies complémentaires sont à mentionner, celles dépendant d'un papier assurant le repérage spatial et celles déportant ce repérage dans un dispositif annexe intégrant des capteurs ultra-sonores. Face à ce flux de documents, il est nécessaire de proposer des fonctionnalités permettant un accès intelligent aux bases de données où ils sont entreposés. De nombreuses fonctionnalités sont souhaitables. On peut évoquer la catégorisation des documents, la recherche par mots-clés, mais également la capacité à retrouver les documents rédigés par une personne donnée. Cette problématique a jusqu'à présent été abordée essentiellement dans le domaine de l'écriture hors-ligne, nous proposons ici de la développer dans le cadre de signaux en-lignes.

L'accès à l'identité du scripteur ayant composé un document apporte de la valeur ajoutée au document. Dans ce cadre, deux fonctions différentes sont à envisager. La première est une fonction d'identification du scripteur, elle consiste à retrouver à partir d'un ensemble de référence contenant toutes les écritures des scripteurs incriminés celle d'un scripteur spécifique dont on dispose d'un échantillon supplémentaire. Elle nécessite une mise en correspondance de type 1 contre N. La seconde est une fonction de vérification, elle permet d'attester ou d'infirmer l'identité prétendue d'une personne. Dans ce cas, la confrontation est de type 1 contre 1. Les fonctions identification/vérification de scripteurs s'inscrivent dans une politique de gestion des droits numériques et de prévention de la fraude et du vol d'identité. Par exemple, l'une des applications pourrait être de traiter l'identité des étudiants composant à un examen à des fins de contrôle. Deuxièmement, dans des environnements où de grandes quantités de documents, formulaires, notes et procès-verbaux de réunions sont constamment en cours de traitement et de gestion, connaître l'identité du rédacteur donne une valeur supplémentaire pour distribuer de façon adaptée ces documents ou remonter à la source d'une information.

Le reste de ce document est organisé comme suit : la section 2 rappelle les principaux travaux sur ce sujet, ensuite la méthodologie proposée est détaillée en section 3, tandis que les résultats expérimentaux sont présentés dans la section 4. Enfin, les discussions et les pistes à explorer sont données dans la section 5.

2. Les méthodes existantes

À un premier niveau, on distingue deux types de systèmes d'identification/vérification de scripteurs : ceux qui sont indépendants du texte écrit et ceux qui reposent sur un texte imposé. Les signatures sont un cas particulier de cette seconde catégorie. À l'inverse, le système que nous proposons ici s'inscrit dans la première catégorie. Un second niveau de différenciation concerne la nature hors-ligne ou en-ligne du signal d'écriture. De nombreux systèmes ont été proposés pour traiter des images de documents [Hochberg 97], [Bensefia 05], [Busch 05], [Bulacu 07], [Niels 07], beaucoup moins de travaux concernent l'écriture en-ligne [Jain 03], [Pitak 04], [Niels 07], [Chan 08].

Enfin, les différents systèmes se distinguent par le type d'approches mises en oeuvre. On peut principalement en distinguer deux types. Certaines vont extraire des caractéristiques globales, significatives d'un point de vue macroscopique [Pitak 04], [Yasushi 03]. Il s'agit par exemple de la densité des lignes, de la fluctuation des lignes de bases, du respect de l'indentation. Des mesures d'entropie, de paramètres de textures [He 08] peuvent contribuer à ce type de description. À l'inverse, il peut être intéressant d'avoir une vue beaucoup plus locale et de baser l'identification sur des caractéristiques apparaissant à une échelle microscopique. C'est le cas des approches se basant sur les singularités des graphèmes composant l'écriture. Là encore, le niveau de granularité peut être variable. Il s'agit la plupart du temps de graphèmes extraits par des algorithmes de segmentation simples [Bensefia 05], où alors manuellement [Schomaker 04]. Le tableau 1 résume les principaux travaux consacrés à l'identification de scripteurs et en précise certaines spécificités. Le système proposé ici s'intéresse aux documents en-ligne, il s'agit d'une extension des travaux de [Tan 08]. Le calcul du vecteur caractéristique du style du scripteur utilise ici une méthode plus précise, de plus, nous avons étendu la base de référence et nous comparons les résultats sur une base française et sur une base de textes anglais. Enfin, nous proposons ici une analyse beaucoup plus poussée en nous intéressant à la sensibilité de la méthode vis-à-vis de la longueur du texte, du nombre de prototypes par lettre, et de la métrique utilisée dans la mise en correspondance des styles d'écriture. L'originalité de notre contribution réside dans le choix d'extraire des lettres grâce à un processus de segmentation/étiquetage automatique du texte en caractères. Nous pensons en effet que le niveau lettre porte une information biométrique de nature très stable, à l'instar de [Niels 07]. Cette démarche est également caractéristique des approches manuelles mises en oeuvre par les experts en criminologie [Stuart 05].

Tableau 1. Les principaux systèmes d'identification de scripteurs.

Auteurs	Année	Originalité	Taux d'identification	Langue
[Zois 00]	2000	Approche Morphologique	96.5 % avec 50 scripteurs	Anglais
[Said 00]	2000	Filtre de Gabor & méthode de co-occurrence	95 % avec 20 scripteurs	Anglais
[Srihari 01]	2002	Texte imposé, mise en oeuvre d'un perceptron	98 % avec 1000 scripteurs	Anglais
[Pitak 04]	2004	Transformée de Fourier	98.5 % avec 81 scripteurs	Thai
[Schlapbach 04]	2004	Un modèle de Markov par scripteur	96 % avec 100 scripteurs	Anglais
[Bensefia 05]	2005	Graphèmes multi-niveaux	95 % avec 88 scripteurs	Français
			86 % avec 150 scripteurs	Anglais
[Bulacu 07]	2007	Basée Texture et prototypes d'allographes	92 % avec 650 scripteurs	Anglais
[Niels 08]	2008	Mise en correspondance de prototypes d'allographes, segmentation manuelle	100 % avec 43 scripteurs	Anglais
[Chan 08]	2008	Assignation d'un prototype à chaque caractère, segmentation automatique	95 % avec 82 scripteurs	Français
Notre système	2008	Assignation floue de tous les prototypes à chaque caractère, segmentation automatique	99 % avec 120 scripteurs	Français
			87% avec 200 scripteurs	Anglais

3. La Méthode Proposée

Pour introduire la méthode que nous avons développée et analysée, nous proposons d'abord d'observer les quelques échantillons d'écriture représentés sur la figure 1.

Ces mots sont issus de la base IRONOFF [Viard-Gaudin 99] et proviennent de deux scripteurs. Comme on peut l'observer sur la figure 1, il y a une grande stabilité intra-scripteur sur le tracé de certaines lettres. C'est le cas en particulier dans cet exemple pour la lettre 'f'. Nous pouvons remarquer que ces deux scripteurs n'utilisent pas le même allographe pour la lettre 'f'. On rappelle ici qu'un allographe est une variante graphique pour représenter une même lettre. Dès lors, imaginons que l'on dispose de prototypes de référence pour représenter les différents allographes d'une lettre et que l'on assigne chaque instance d'une lettre au prototype le plus similaire, il en résulte la possibilité de calculer une distribution discrète sur la base des prototypes. Ainsi par exemple, si l'on considère les cinq prototypes définis dans le tableau 2, le scripteur *1* se verra assigner très fréquemment le prototype N° 5, tandis que le scripteur *2* aura une forte propension à utiliser le prototype N° 4. Il en résulte les distributions plausibles pour les scripteurs 1 et 2 telles que celles indiquées sur le tableau 2. Cette description vectorielle va permettre de comparer, avec une métrique restant à définir, les usages qui sont faits des allographes de cette lettre par différents scripteurs, il s'agit là d'une caractéristique déterminante du style d'écriture propre à un individu. En supposant que les scripteurs *1* et *2* constituent la base des scripteurs de référence, un scripteur inconnu *T* pourra être identifié par comparaison de sa distribution avec celles des scripteurs *1* et *2*. Au vu de l'exemple, il est clair que le scripteur 1 devrait être classé en première position.

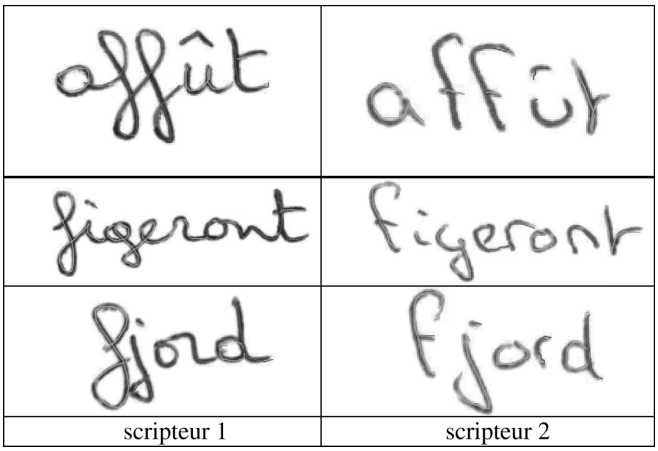


Figure 1. Echantillons de mots de la base IRONOFF.

À partir de la description simplifiée que nous venons d'introduire, nous pouvons résumer la méthode qui va être décrite. La méthode d'identification de scripteurs proposée repose sur trois étapes, *cf.* figure 2. La première consiste à sélectionner lettre par lettre les prototypes caractéristiques des allographes de cette lettre. La seconde étape représente le codage sur la base des allographes-prototypes des documents de référence et de test. Enfin, la troisième étape correspond à la classification du document de test vis-à-vis des documents de référence. Ces trois étapes sont décrites ci-après.

3.1. Construction des prototypes

Compte tenu de la nature en-ligne, donc temporelle, de l'écriture étudiée ici, plusieurs sources de variabilité sont attendues

Tableau 2. Distribution des caractères selon les prototypes d'allographe

k	1	2	3	4	5
Allographes de la lettre 'f' Distribution sur N = 5 prototypes					
scripteur 1	0.00	0.05	0.00	0.00	0.95
scripteur 2	0.00	0.05	0.15	0.80	0.00
scripteur *T*	0.10	0.00	0.05	0.00	0.85

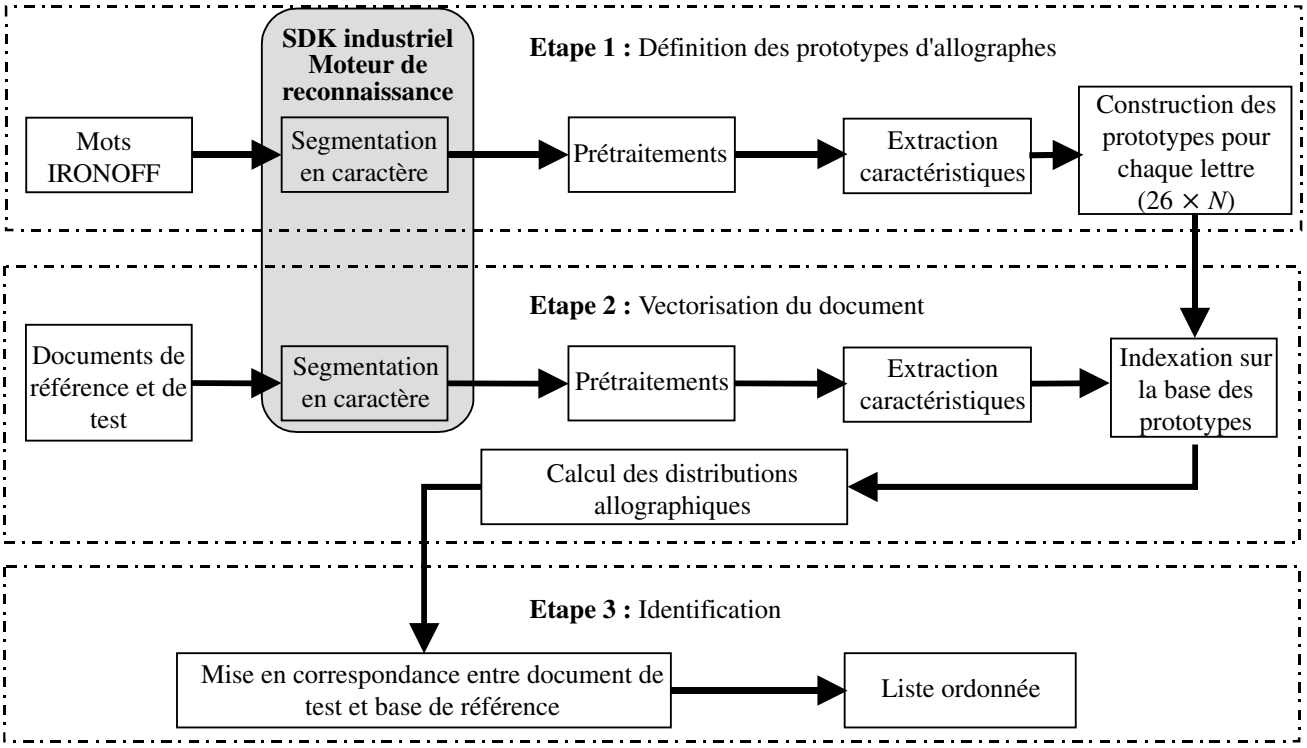
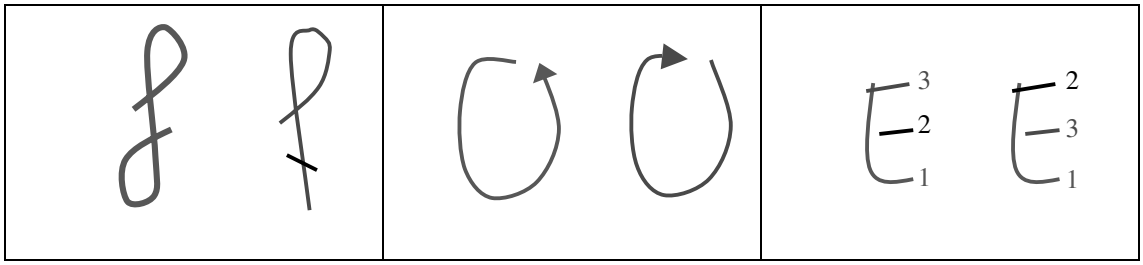


Figure 2. Schéma de la méthodologie proposée

dans les allographes de lettres. La figure 3 en résume les principales. La figure 3-a montre deux variantes morphologiques de la lettre 'f', tandis sur la figure 3-b, c'est le sens de parcours de la forme qui est modifiée d'une variante à l'autre, alors que sur la figure 3-c, il s'agit de l'ordre des tracés élémentaires composant la forme qui est différent. C'est cette variété des allographes d'une lettre qui va être modélisée grâce à la sélection de prototypes représentatifs des écritures rencontrées sur une large palette de scripteurs.

Pour construire les prototypes définissant les allographes de lettres, nous avons utilisé les mots de la base IRONOFF [Viard-Gaudin 99]. Cette base comporte 16 585 mots écrits par 373 sujets. Chaque mot dont on connaît le vrai label est ensuite segmenté automatiquement en lettres. Pour réaliser cette étape de segmentation automatique nous utilisons un moteur de reconnaissance d'écriture manuscrite en-ligne [Vision 09] auquel nous fournissons le label du mot à reconnaître. Le moteur nous retourne en sortie le meilleur alignement de la séquence de points de la trajectoire avec le label mot considéré.

Cela permet de disposer de 89 760 caractères, écrits en contexte de mots, répartis sur les différentes lettres de l'alphabet. La lettre la plus fréquente est bien entendue la lettre '*e*' avec 12 161



(a) (b) (c)

Figure 3. (a). Variante Morphologique. (b). Variante Directionnelle. (c). Variante Temporelle.

occurrences tandis que la moins fréquente est le '*w*' avec seulement 139 échantillons. Sur chacun des sous-ensembles correspondant aux 26 lettres minuscules de l'alphabet, un algorithme de classification non-supervisé (clustering) est mis en oeuvre. C'est lui qui permet de définir les N prototypes d'allographes qui permettront de représenter les styles d'écritures de chaque lettre. Ces prototypes sont communs à l'ensemble des scripteurs de cette base. Ces scripteurs sont différents de ceux qui seront utilisés en situation d'identification de l'écriture, la méthode proposée s'inscrit donc dans un cadre omni-scripteur. L'usage spécifique que fera chaque scripteur de ces prototypes génériques permettra l'identification de celui-ci.

Dans les expériences relatées dans cet article, nous avons utilisé un algorithme de type *K-moyennes* en faisant varier le nombre N de *clusters* (de prototypes d'allographes) entre 2 et 60 par lettre. La distance utilisée est la distance Euclidienne dans l'espace de description des caractères. Chaque caractère est représenté par un vecteur de 210 composantes, après avoir normalisée et ré-échantillonnée la trajectoire sur 30 points et extrait 7 caractéristiques par point. Les caractéristiques utilisées sont les coordonnées x et y, la direction $(\cos\theta \; et \; \sin\theta)$ de la tangente, la courbure $(\cos\Delta\theta \; et \; \sin\Delta\theta)$ et l'état posé/levé du stylo en ce point [Chan 08]. La figure 4 présente cinq des prototypes obtenus sur la sous-base des caractères reconnus comme étant des '*f*'. Les trajectoires en pointillé représentent l'état levé du stylo.

3.2. Vectorisation des documents

Concernant la seconde étape, chaque document, qu'il soit issu de la base des documents de référence ou qu'il soit un document de test dont on veut identifier le scripteur, va être projeté dans l'espace des prototypes d'allographes. En ne considérant que les lettres minuscules, cet espace est de dimension $26 \times N$. C'est dans cet espace que se fera la mise en correspondance entre un document de test et les documents de référence pour permettre d'ordonner ceux-ci vis-à-vis du document de test. Le système supporterait une extension aux majuscules, on pourrait aussi évoquer les signes de ponctuation et les digits qui font partie des symboles de l'alphabet supporté par le reconnaisseur. Toutefois dans un soucis de simplification, et notamment pour limiter la taille du vecteur d'indexation de chaque document, nous nous sommes limités aux lettres minuscules. Par ailleurs, nous ne sommes pas persuadés du bien fondé de l'apport de la prise en compte des majuscules et autres symboles additionnels. N'oublions pas que la méthode est fondée sur l'estimation des probabilités d'usage des différents prototypes d'une lettre. Si cette lettre est très peu fréquente, voire complètement absente, ce qui est souvent le cas avec les majuscules, alors cette estimation sera très peu fiable, voire invalide. Il en résulterait une augmentation du niveau de bruit de l'estimation, et donc une perte de précision dans le taux d'identification. Nous n'avons pas mené d'expérience sur les majuscules pour confirmer cette hypothèse, par contre les expériences relatées dans [Tan 09] qui consistent à supprimer progressivement les lettres minuscules les moins pertinentes au sens de leur capacité propre d'identification, appuient cette analyse.

Un des points importants de la méthode réside dans la capacité à reconnaître, segmenter et étiqueter automatiquement le texte au niveau caractère. Toutes ces tâches sont confiées à un moteur de reconnaissance industriel (MyScript Builder) [Vision 09].

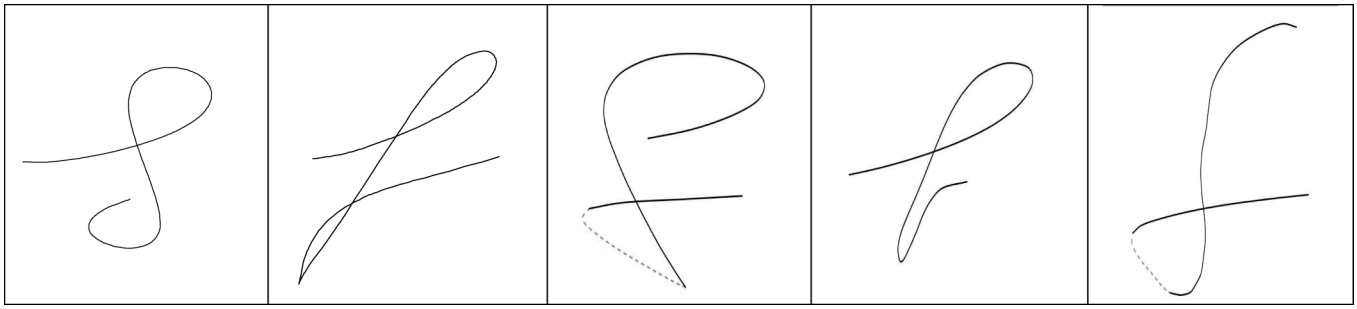


Figure 4. Exemples de prototypes de la lettre 'f' obtenus par clustering.

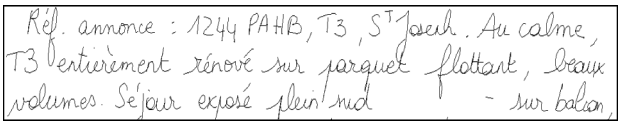


Figure 5. Exemple d'extrait de texte.

Bien entendu ici, à la différence de l'usage qui en avait été fait dans l'étape de construction des prototypes, le texte n'est pas supposé connu. Il s'agit donc d'une véritable situation de reconnaissance, puis de post-segmentation. Ce moteur est déjà entraîné, indépendamment de l'application traitée ici, et il supporte l'usage de modèles de langage afin d'améliorer ses capacités de reconnaissance [Perraud 05]. Nous avons ici associé le modèle de langage générique propre à la langue traitée : le Français ou l'Anglais.

La figure 5 montre un exemple de texte dont on veut identifier le scripteur. La figure 6 affiche le résultat de la reconnaissance d'un des mots, ainsi que deux des caractères issus de la segmentation automatique. Bien entendu, l'outil de reconnaissance ne réalise pas une segmentation et une reconnaissance parfaite. Il en résulte des erreurs dans l'affectation des séquences de points (segments) à la lettre qui aurait dû lui correspondre. Par exemple sur la figure 7, on montre l'effet d'une erreur de segmentation : l'accent de la lettre '*é*' du mot « *économie* » qui a été correctement reconnu est affecté à tort à la lettre suivante '*c*'. Les performances de l'outil de reconnaissance sont évaluées dans la section 4.

Une fois que tous les caractères d'un document sont extraits, nous projetons chacun de ces caractères dans la base des prototypes de la lettre qu'il représente, et nous mesurons une similitude entre ce caractère et chacun des prototypes de cette lettre. Cette mesure est normalisée pour être interprétée comme la probabilité que le prototype ait émis ce caractère [Han 06], [Hoppner 99].

$$P_\alpha(p_x|x) = \frac{\exp(-\beta \times dist(p_k,x))}{\sum_{k^*=1}^{N} \exp(-\beta \times dist(p_{k'},x))} \quad (1)$$

Dans cette formule x représente un échantillon d'une lettre de l'alphabet $\alpha \in$ ['a', 'b', ...'z'] dont on veut calculer la probabilité qu'il ait été généré par le prototype p_k. N est le nombre de prototypes d'allographes pour cette lettre. La distance $dist(p_k,x)$ est la distance de Mahalanobis entre le point x et le cluster de prototype p_k, elle est calculée dans l'espace de représentation des caractères, ici de dimension 210. Dans l'équation (1), $\beta > 0$ est un paramètre de réglage fixant la sélectivité des fonctions exponentielles. Ainsi, avec une faible valeur de β, la masse de probabilité sera distribuée sur des prototypes éloignés, tandis qu'avec une valeur de β plus élevée les prototypes les plus proches seront davantage concernés. Nous avons réglé expérimentalement $\beta = 0{,}01$, à la suite des expériences présentés à la figure 9.

En procédant ainsi, on ne prend pas une décision stricte en ne considérant que le plus proche voisin, comme cela avait été fait dans [Chan 08], mais tous les prototypes sont partiellement considérés grâce à la fonction d'appartenance floue définie par l'équation (1). On peut alors obtenir la fréquence du prototype p_k, soit $tf_{\alpha,k}$ (term frequency) sur l'ensemble du document en sommant sur tous les échantillons appartenant à la même lettre, *cf.* eq. (2). M_α est le nombre de caractères du document correspondant à la même lettre α de l'alphabet.

$$tf_{\alpha,k} = \frac{1}{M_\alpha} \sum_{m=1}^{M_\alpha} P_\alpha(p_k|x_m) \quad (2)$$

Figure 6. Reconnaissance et segmentation des caractères, cas du 'f' et du 'o'.

a) texte à reconnaître

b) mot correctement reconnu c) segmentation incorrecte

Figure 7. Erreur de segmentation.

$$idf_{\alpha,k} = \log \frac{\sum_{k'=1}^{N}\sum_{i=1}^{R} tf_{\alpha,k',i} + \varepsilon}{\sum_{i=1}^{R} tf_{\alpha,k,i} + \varepsilon} \quad (3)$$

Enfin, ces termes sont pondérés par un facteur idf_k (inverse document frequency, Eq. 3) indiquant le caractère fréquent ou non de ce terme dans l'ensemble de la base des documents de référence, en nombre R. Il en résulte un vecteur de dimension $26 \times N$: $[tf_{\alpha,k} \times idf_{\alpha,k}]$, avec $k \in [1,N]$, décrivant le style de l'écriture du document, en considérant les 26 lettres minuscules 'a' à 'z'. L'utilisation de descripteurs de type $tf \times idf$ est classique en recherche d'informations [Salton 88].

3.3. Classification

À partir de cette représentation vectorielle, il est possible de calculer la distance entre le document de test w_T, et chaque document de référence w_i. Le scripteur i dont le document minimise la distance $dist(w_i, w_T)$ sera considéré en première position. Nous avons envisagé trois métriques pour calculer cette distance. D'une part, classiquement la distance euclidienne, par ailleurs comme les composantes de ce vecteur proviennent de distribution, nous avons aussi utilisé la divergence de *Kullback-Leibler* (Eq. 4) [Cover 91], et enfin pour la même raison la distance du Chi^2 (Eq. 5). Le calcul de ces distances s'effectue en ne prenant en compte que les lettres de l'alphabet effectivement représentées à la fois dans le document de test, et le document de référence [Chan 08].

$$dist(w_i, w_T) = \sum_{\alpha='a'}^{'z'}\sum_{k=1}^{N} idf_{\alpha,k} \times tf_{\alpha,k,T} \log \frac{tf_{\alpha,k,T}}{tf_{\alpha,k,i}} \quad (4)$$

$$dist(w_i, w_T) = \sum_{\alpha='a'}^{'z'}\sum_{k=1}^{N} \frac{idf_{\alpha,k}(tf_{\alpha,k,i} - tf_{\alpha,k,T})^2}{tf_{\alpha,k,i} + tf_{\alpha,k,T}} \quad (5)$$

4. Résultats expérimentaux

4.1. Bases de données

Nous proposons d'évaluer cette méthode sur deux bases distinctes. La première, appelée FR_120, comporte des textes français, elle est issue de 120 scripteurs. Chacun de ces scripteurs a rédigé 2 documents, sur des sujets libres, totalement distincts et à 2 moments différents. La taille des documents est variable, elle s'échelonne entre 86 caractères pour le plus court à 972 caractères pour le plus long, la longueur moyenne étant de 465 caractères, soit une dizaine de lignes manuscrites. L'exemple présenté à la figure 5 est issu de ces documents. Deux sous-bases ont été constituées, chacune de 120 documents, contenant un document provenant de chaque scripteur. L'une de ces sous-bases correspond à la base de référence, l'autre servira de base de test pour évaluer la capacité du système à retrouver dans la base de référence le bon scripteur. Pour chaque document de la base de test, on ordonnera la base de référence pour permettre le calcul du taux d'identification jusqu'au TOP(n).

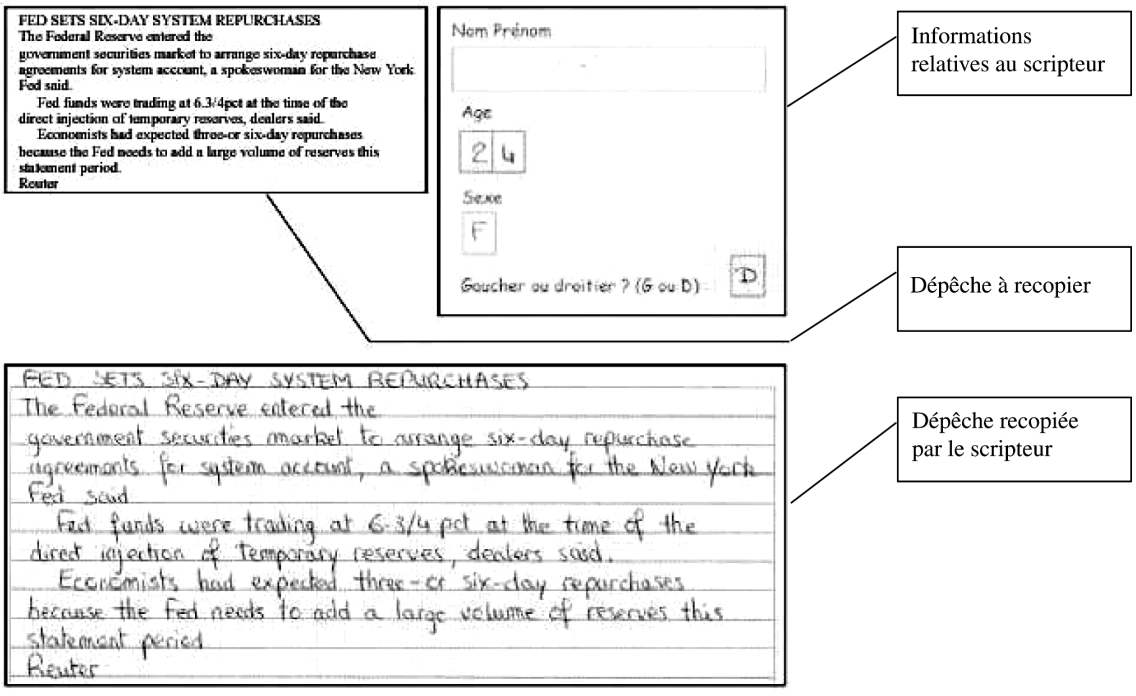


Figure 8. Collecte des dépêches REUTERS.

La seconde base considérée, REUTERS_200, comporte des textes écrits en Anglais. Elle a été obtenue par recopie manuscrite d'une partie de la base de dépêches de l'agence REUTERS, *cf.* figure. 8. Le corpus utilisé dans nos expériences est un sous-ensemble du corpus Reuters-21578 [Peña 08] reproduit sous forme manuscrite. Deux cents scripteurs ont chacun recopié deux dépêches, fournissant ainsi, deux cents documents de références et deux cents documents utilisés pour l'évaluation du taux d'identification du scripteur. Cette base est plus difficile que la précédente. En effet, d'une part la taille de la base de référence est plus grande. De plus, les textes sont plus courts, ils comptent en moyenne 304 caractères, soit approximativement six lignes. Enfin, ils sont constitués d'un mélange plus disparate de majuscules, minuscules, chiffres et noms hors-lexiques. Tout cela rend l'étape de reconnaissance plus ardue, ainsi que l'étape d'identification qui suit, puisque celle-ci se base uniquement sur les lettres minuscules.

Les documents de ces deux bases ont été collectés avec une technologie de type papier et stylos digitaux, essentiellement dans le même milieu culturel, celui des étudiants et des enseignants. On peut donc s'attendre à une certaine homogénéité si ce n'est dans le type d'écriture au moins dans l'usage qui est fait de l'écriture manuscrite, rendant ainsi le problème de l'identification non trivial.

Sur l'ensemble de ces deux bases, nous avons d'abord calculé le taux de reconnaissance de l'outil qui nous sert à faire la segmentation automatique en lettres. Nous l'utilisons pour cela dans sa version standard, déjà entraîné, avec les ressources linguistiques de base qui ne sont pas spécifiquement adaptées aux textes que nous traitons ici. Toutefois, bien entendu le modèle linguistique propre au Français est utilisé pour la base FR_120 et le modèle linguistique anglais pour la base REUTERS_200.

Le taux de reconnaissance au niveau lettre est respectivement de 91.0 %. pour la base française et de 89.0 % pour la base REUTERS_200. Ce calcul est restreint aux seules lettres minuscules sur lesquelles se base la méthode d'identification, en sont donc exclus tous les autres caractères. Les lettres mal reconnues (9 % ou bien 11 %) vont bruiter les assignations dans les bases lettres correspondantes.

4.2. Influence de la métrique de mise en correspondance (base FR_120)

Cette étude a été réalisée sur la base FR_120. Nous présentons dans le tableau 3 les résultats d'identification en première position obtenus avec trois métriques différentes pour calculer la mise en correspondance entre un document de test (w_T) et les documents de référence (w_i). Il s'agit de la distance euclidienne, de la divergence de Kullback-Leibler (Eq. 4) et d'une pseudo-distance du Chi2 (Eq. 5).

Nous présentons également dans la première colonne du tableau 3, les résultats obtenus dans une version antérieure [Chan 08] où en lieu et place de la formule (Eq. 1) permettant de distribuer l'origine d'un caractère sur l'ensemble des prototypes disponibles pour cette lettre, seul le plus proche prototype était retenu. Nous pouvons observer que cette version de base permettait d'obtenir un taux d'identification s'élevant à 96.7 %, ce qui signifie que sur la base de test, seul 4 scripteurs parmi les 120 n'ont pas été retrouvés en première position.

Tableau 3. Taux d'identification en première position, recherche dans un base de 120 scripteurs

1-Plus Proche Voisin Dist. euclidienne	Distribution sur l'ensemble des prototypes		
	Distance euclidienne	Divergence KL	Distance Chi^2
96.7 % 116/120	98.3 % 118/120	91.7% 110/120	99.2 % 119/120

Le taux d'erreur est divisé par deux avec la méthode proposée ici, puisque seuls deux scripteurs échappent à la première position lorsque l'on utilise la distance euclidienne. Avec la distance du Chi^2, on diminue encore l'erreur : un seul scripteur n'est pas retrouvé en première position. De plus, le vrai scripteur correspondant est positionné en seconde position de la liste des documents de référence. Par contre, les résultats sont sensiblement moins bons lorsque la divergence de Kullback-Leibler est utilisée. Il est possible que la non-symétrie de cette métrique soit un facteur défavorable, de plus son utilisation dans le cadre de distributions sur des données symboliques peut également poser problème.

L'amélioration que nous constatons par rapport à la méthode originale est à mettre au crédit d'une meilleure représentation dans l'espace des prototypes. Initialement cet espace était quantifié de façon discrète : un symbole discret représentait tout un sous-espace des caractéristiques, il en résultait un changement brutal de représentant lors de petits déplacements dans cet espace. Nous sommes passés avec la méthode proposée à une représentation continue dans cet espace en estimant la probabilité de chaque style vis-à-vis de l'échantillon de caractère étudié. De cette façon lorsqu'un style d'écriture se trouve en frontière de plusieurs prototypes, la stabilité de la description est bien meilleure.

4.3. Influence du paramètre de sélectivité des prototypes (base FR_120)

La formule (1) fait intervenir un paramètre β qui fixe la sélectivité radiale de chaque prototype. La masse de probabilité d'un échantillon x est distribuée plus fortement sur les prototypes les plus proches de x lorsque la valeur de β est élevée. À l'inverse, si la valeur de β est faible, on tend vers une distribution uniforme, ce qui voudrait dire que chaque échantillon ressemble pareillement à tous les prototypes d'allographes de cette lettre.

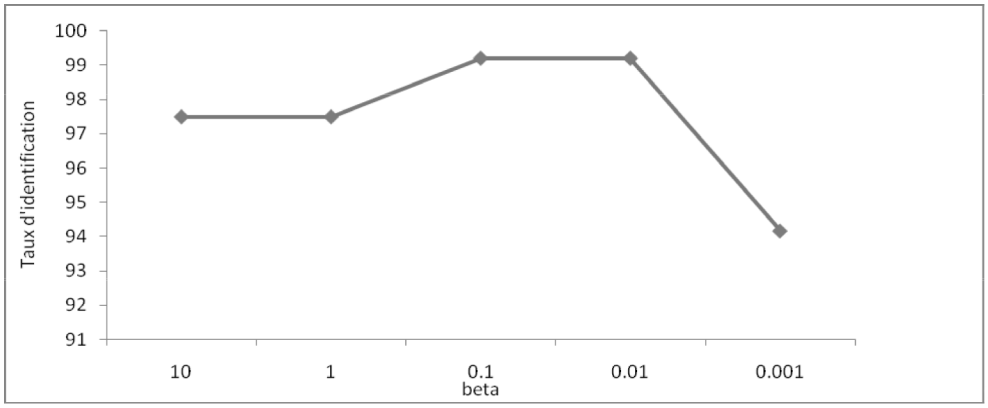


Figure 9. Influence de β sur le taux d'identification sur la base de 120 scripteurs (FR_120).

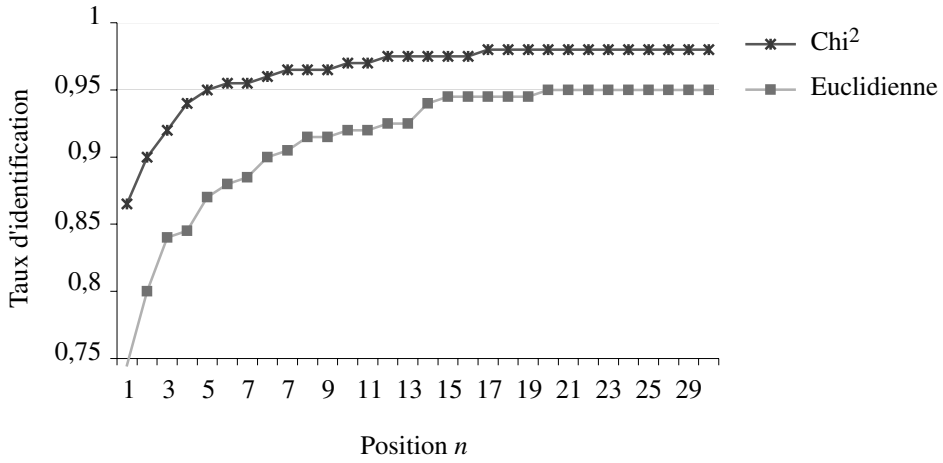


Figure 10. Taux d'identification en position n, recherche dans une base de 200 scripteurs (REUTERS _200).

Pour ajuster la valeur de ce paramètre nous l'avons fait varier sur une plage de quatre décades [0.001..10] et retenu la valeur correspondant au maximum du taux d'identification sur la base FR_120. Ce réglage a été conservé sur la base REUTERS_200. La figure 9 montre le résultat de cette étude.

4.4. Résultat sur la base REUTERS_200

Avec la meilleure configuration obtenue sur la base FR_120, à savoir la distance du Chi^2, nous obtenons un taux d'identification en première position sur la base REUTERS_200 de 86.5 %. On peut noter une baisse de performance de 12.7 % par rapport à la base FR_120 (99.2 %). Trois phénomènes complémentaires expliquent cette évolution. Premièrement, le système de reconnaissance est légèrement moins performant sur la base REUTERS_200 (89 % au lieu de 91 %), ensuite les textes de la base REUTERS_200 sont plus courts (304 caractères en moyenne contre 465) et comme le montre la figure 12 les performances se dégradent lorsque la longueur du texte diminue, enfin le nombre de classes augmentent : 200 au lieu de 120. Il est évidemment plus difficile de retrouver le bon scripteur parmi 200 que parmi 120, le risque de confusion inter-scripteur étant alors plus grand. Toutefois, comme le montre la figure 10, le taux d'identification s'élève rapidement lorsque l'on étend la taille de la liste de documents de référence considérés, ainsi par exemple, il passe à 95 % en considérant la bonne réponse parmi les 5 premiers candidats. Comme sur la base FR_120, la métrique du Chi^2 donne les meilleurs résultats ainsi que le montre la figure 10 qui permet une comparaison avec la distance Euclidienne.

4.5. Influence du nombre de prototypes

Dans toutes les expériences précédentes, à chaque lettre de l'alphabet est assigné un ensemble de $N = 10$ prototypes. Ce choix résulte d'une étude expérimentale dont les résultats sont reportés à la figure 11. Dans cette étude, le nombre de prototypes par lettre évolue de 2 à 60. Afin de vérifier la stabilité vis-à-vis de la base de test, nous avons subdivisé aléatoirement celle-ci en deux parties égales de 60 scripteurs tout en conservant les 120 documents dans la base de référence (FR_120). On observe sensiblement le même comportement sur les 2 sous-bases, à savoir

qu'avec moins de 10 prototypes, les performances sont dégradées, et de même au delà de 30. Ainsi, avec trop peu de prototypes la précision de la représentation est trop grossière, le biais de l'estimateur des fonctions de densité de probabilité (*ddp*) entraîne ces piètres performances. A l'inverse lorsque le nombre de prototypes est trop élevé, l'estimateur est trop sensible, le bruit présent dans les données se trouve trop pris en compte. Le nombre de 10 prototypes apparaît réaliser le meilleur compromis entre biais et variance dans l'estimation des *ddp*. On pourrait pousser plus loin l'analyse sur le nombre de prototypes en travaillant non plus globalement sur l'alphabet, mais lettre par lettre. Certaines lettres, plus complexes que d'autres, présentent davantage de variations allographiques. C'est sans doute le cas par exemple de la lettre '*f*' vis-à-vis de la lettre '*e*'.

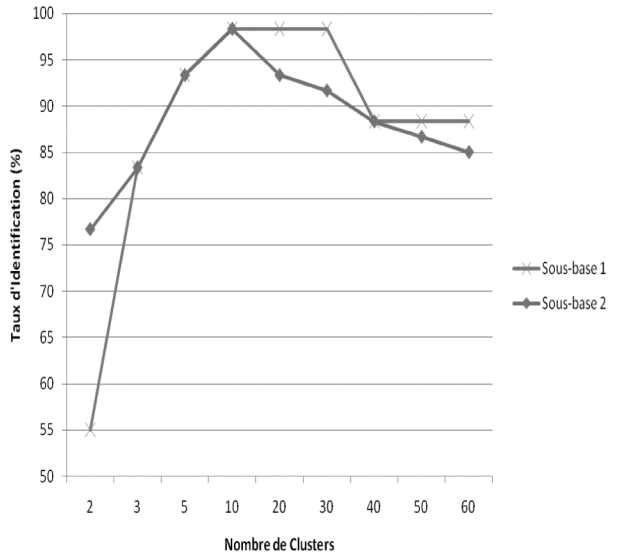


Figure 11. Sensibilité de la méthode au nombre de prototypes par lettre (base FR_120).

4.6. Influence de la longueur des textes

La mise en correspondance entre un texte de référence et un scripteur en test s'appuie sur l'estimation des distributions des prototypes d'allographes des différentes lettres de l'alphabet. Pour estimer de façon fiable ces distributions, il est nécessaire de disposer d'une quantité suffisante de caractères. C'est la sensibilité de la méthode vis-à-vis de ce nombre de caractères que nous allons explorer dans les résultats présentés ici. Dans cette expérience, nous choisissons de réduire de façon contrôlée la taille des documents de test, en ne conservant qu'un nombre fixé de caractères. Les caractères conservés sont choisis aléatoirement. La figure 12 montre en ordonnée le nombre moyen de scripteurs de la base de test qui ne sont pas retrouvés en première position en faisant décroître en abscisse le nombre de caractères de 350 à 60. On constate que ce nombre est relativement stable jusqu'à 160 caractères, ensuite il augmente significativement. Par exemple, avec seulement 60 caractères, soit moins de deux lignes de textes manuscrits, près de 16 scripteurs sur les 120 de la base de test ne sont pas positionnés en première position. Pour que les résultats soient indépendants des caractères sélectionnés, l'expérience est reproduite dix fois pour chaque taille de texte, d'où les valeurs moyennes indiquées.

5. Conclusion

Nous avons proposé une méthode d'identification du scripteur pour des documents contenant de l'écriture en-ligne. Cette méthode donne de très bons résultats sur les deux bases utilisées. Dans la première des situations évaluées, la taille de la base de référence est de 120 scripteurs parmi lesquels, il faut retrouver le scripteur à identifier. La base de test comporte également 120 documents ; parmi ceux-ci, un seul scripteur n'est pas retrouvé en première position, il est toutefois en seconde position. Ce nombre de 120 correspond déjà à un nombre significatif, par exemple celui d'un amphi d'étudiants composant à

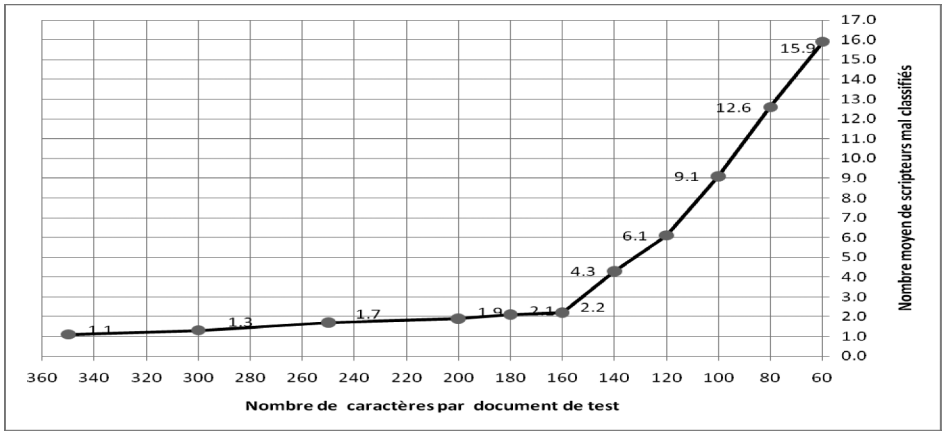


Figure 12. Sensibilité de la méthode à la taille du document de test (base FR_120).

un examen. Ces conditions d'évaluation correspondent donc à un usage réaliste, que l'on peut également retrouver en situation d'entreprise, où de nombreux collaborateurs (enquêteurs, journalistes, secrétaires, …) contribuent à alimenter le système d'information de la société.

Nous avons également cherché à évaluer cette méthode dans des conditions plus sévères, notamment en augmentant la taille de la base des scripteurs de référence. Bien entendu, on augmente alors les risques de confusion inter-scripteurs. Sur la base considérée, REUTERS_200, comportant 200 scripteurs, le taux d'identification du scripteur en première position atteint 86.5 %. Cette baisse de performance est due non seulement au changement de taille de la base de référence, mais aussi à la diminution de la longueur des textes utilisés pour faire l'identification.

La méthode proposée repose sur la spécificité de l'usage des styles allographiques entre les scripteurs. Ces allographes sont examinés au niveau caractère, pour cela la première étape de la méthode consiste à segmenter et reconnaître le texte à ce niveau caractère. À cet effet, un outil industriel automatique a été mis en oeuvre, il donne des performances très suffisantes pour cette tâche. Qui plus est, disposant d'une version électronique du texte, même partiellement dégradée, d'autres fonctions au delà de l'identification du scripteur sont envisagées. On peut évoquer des tâches de catégorisation [Peña 09] ou d'indexation de documents.

Les études expérimentales menées ont permis de répondre à plusieurs questions, concernant notamment le choix de la métrique dans l'espace de représentation des styles allographiques et aussi du nombre de prototypes nécessaires dans cet espace. Ainsi, nous avons obtenu les meilleurs résultats avec la distance du Chi^2 en utilisant un nombre de 10 prototypes par lettre de l'alphabet en ne considérant que les lettres minuscules. Le passage d'une quantification par un symbole discret correspondant au plus proche prototype d'un caractère à une représentation continue indiquant les probabilités *a posteriori* que ce caractère soit l'un des $N = 10$ prototypes disponibles a permis d'améliorer de façon notable les résultats d'identification. Cela a fait passer le taux d'identification du scripteur de 96,7 % à 99,2 %.

Pour augmenter les performances de ce système, il serait possible d'étendre l'alphabet considéré en s'appuyant sur davantage de symboles que les seules lettres minuscules, par exemple les majuscules et les digits. De plus, le nombre de prototypes par lettre est ici constant quelque soit la lettre, on pourrait adapter ce nombre en fonction de chaque lettre. Une autre extension possible serait de considérer plus d'un label en sortie du classifieur de caractères, et cela en fonction de la confiance associée à chaque classe, de façon à assigner chaque caractère segmenté à plus d'une lettre de l'alphabet. De cette façon, les lettres mal reconnues pourraient, si elles se trouvent dans une position pas trop éloignée, venir participer à l'identification du scripteur. Ainsi par exemple, supposons qu'un segment soit reconnu comme un 'h' en Top(1) avec un score normalisé de 0.8 et comme un 'f' en Top(2) avec un score de 0.2, alors la méthode actuelle ne met à jour que la distribution des allographes des 'h', l'extension proposée ici consisterait à s'intéresser également à la distributions des allographes des 'f' en pondérant les masses de probabilités par les scores normalisés.

Références

[Bensefia 05] A. BENSEFIA, T. PAQUET, L. HEUTTE, "Handwritten Document Analysis for Automatic Writer Recognition", *Electronic Letters on Computer Vision and Image Analysis*, 2005, vol. 5, no. 2, pp. 72-86.

[Bulacu 07] M. BULACU and L. SCHOMAKER, "Text-Independent Writer Identification and Verification sing Textural and Allographic Features", *IEEE Trans.Pattern Analysis and Machine Intelligence*, vol. 29, no.4, Apr 2007, pp. 701-717.

[Busch 05] A. BUSCH, W.W. BOLES and S. SRIDHARAN, "Texture for Script Identification", *IEEE Trans.Pattern Analysis and Machine Intelligence*, vol. 27, no.11 Nov 2005, pp. 1720-1732.

[Chan 08] S.K CHAN, C. VIARD-GAUDIN and Y.H TAY, "On line Text Independent Writer Identification Using Character Prototypes Distribution", *Proc.of SPIE IS&T Electronic Imaging: Document Recognition and Retrieval XV*, 2008, vol. 6815, pp.1-9

[Cover 91] T. COVER and J. THOMAS, "*Elements of Information Theory*" Wiley, 1991, pp.13-41.

[Han 06] J. HAN and M. KAMBER, "*Data Mining:Concepts and Techniques*", Elsevier, 2006, pp.383-460.

[He 08] Z. HE, X. YOU and Y. Y. TANG. "Writer identification of Chinese handwriting documents using hidden Markov tree model", *Pattern Recognition* 41, 2008, pp. 1295 – 1307.

[Hochberg 97] J. HOCHBERG, P. KELLY, T. THOMAS and L. KERNS, "Automatic script identification from document images using cluster-based templates", *IEEE Trans.Pattern Analysis and Machine Intelligence*, vol.19 no.2, Feb 1997, pp.176-181.

[Hoppner 99] F. HOPPNER, F. KLAWONN, R. KRUSE and T. RUNKLER, "*Fuzzy Cluster Analysis:Methods for Classification,Data Analysis and Image Recognition*", Wiley, 1999, pp. 5-31.

[Jain 03] A.K. JAIN and A. M. NAMBOODIRI, "Indexing and Retrieval of On-line Handwritten Documents", *Proceedings of the 7th International Conference on Document Analysis & Recognition*, 2003, pp.655-659.

[Niels 07] R. NIELS and L. VUURPIJL, Automatic Allograph Matching in Forensic Writer Identification, *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 21, No. 1 (2007), pp. 61–81.

[Niels 08] R. NIELS, F. GOOTJEN and L. VUURPIJL, "Writer Identification Through Information Retrieval: the Allograph Weight Vector", *International Conference on Frontiers in Handwriting Recognition*, 2008, pp. 481–486.

[Oviatt 00] S. OVIATT, P. COHEN, L. WU, L. DUNCAN, B. SUHM, J. BERS, T. HOLZMAN, T. WINOGRAD, J. LANDAY, J. LARSON and D. FERRO, "Designing the ser Interface for Multimodal Speech and Pen-Based esture Applications: State-of-the-Art Systems and Future Research Directions", *Human Computer Interaction*, 2000, Vol. 15, No. 4, pp. 263-322.

[Peña 08] S. PEÑA SALDARRIAGA, E. MORIN, and C. VIARD-GAUDIN, "Categorization of On-line Handwritten Documents", *Proceedings of the Eighth IAPR Workshop on Document Analysis Systems*, In Proc. DAS2008, Sep 2008, pp. 95-102.

[Peña 09] S. PEÑA SALDARRIAGA, C. VIARD-GAUDIN, and E. MORIN, "On-line Handwritten Text Categorization", *Proceedings of IS&T/SPIE Electronic Imaging, Document Recognition and Retrieval XVI*, vol. 7247, Jan 2009, pp. 727409-1 - 727409-11.

[Perraud 05] F. PERRAUD, C. VIARD-GAUDIN, E. MORIN, P.M. LALLICAN, "Statistical Language Models for On-Line Handwriting

Recognition", *IEICE Transactions on Information and Systems Image Understanding and Digital Document*, Vol.E88-D No.8, 2005, pp.1807-1814.

[Pitak 04] T. PITAK and T. MATSUURA, "On-line Writer Recognition for Thai Based on Velocity of Bary center of Pen-point Movement", *Proceedings of IEEE International Conference on Image Processing*, October 2004, pp.889-892.

[Said 00] H. E. S. SAID, T. N. TAN, and K. D. BAKER, "Personal identification based on handwriting," *Pattern Recognition*, vol. 33, 2000, pp. 149-160.

[Salton 88] G. SALTON and C. BUCKLEY, "Term-weighting approaches in automatic text retrieval", *Information Processing & Management* 24(5), 1988, pp. 513–523.

[Schlapbach 04] A. SCHLAPBACH and H. BUNKE, "Using HMM based recognizers for writer identification and verification," in *Proceedings International Workshop on Frontiers in Handwriting Recognition*, IWFHR, Tokyo, 2004, pp. 167-172.

[Schomaker 04] L. SCHOMAKER and M. BULACU, "Automatic Writer Identification sing Connected-Component Contours and Edge-Based Features of ppercase Western Script", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 26, No. 6, June 2004, pp. 787-798.

[Srihari 01] S. N. SRIHARI, S.-H. CHA, and S. LEE, "Establishing Handwriting Individuality sing Pattern Recognition Techniques," in *Proceedings of the Sixth International Conference on Document Analysis and Recognition*, 2001, pp. 1195-1204.

[Stuart 05] STUART H. JAMES, JON J. NORDBY, "Forensic science: an introduction to scientific and investigative techniques", *CRC Press*, 2005, ISBN0849327474, 9780849327476, 778 pages.

[Tan 08] GUO XIAN TAN, C. VIARD-GAUDIN, and A. KOT, « Identification de Scripteurs basée sur une Distribution Probabiliste de Prototypes d'Allographes », *Colloque International Francophone sur l'Ecrit et le Document*, Oct. 2008, pp. 139-144.

[Tan 09] GUO XIAN TAN, C. VIARD-GAUDIN, and A. KOT, "Automatic writer identification framework for online handwritten documents using character prototypes", in *Pattern Recognition 42 (2009)*, pp. 3313-3323.

[Viard-Gaudin 99] C. VIARD-GAUDIN, P-M LALLICAN, S. KNERR and P. BINTER, "The IRESTE On/Off (IRONOFF) Dual Handwriting Database", *Proceedings of the 5th International Conference on Document Analysis & Recognition*, Sep 1999, pp. 455-458.

[Vision 09] Vision Objects Industrial Text Recogniser SDK, "MyScript Builder Help", SDK documentation, http://www.visionobjects.com/about-us/download-center/_263/myscript-products-datasheets.html, 2009

[Yasushi 03] Y. YASUSHI, T. NAGAO and N. KOMATSU, "Text-indicated Writer Verification Using Hidden Markov Models", *Proceedings of the 7th International Conference on Document Analysis & Recognition*, 2003, pp.329-332.

[Zois 00] E. N. ZOIS and V. ANASTASSOPOULOS, "Morphological waveform coding for writer identification," *Pattern Recognition*, vol. 33, 2000, pp. 385-398.

Christian **Viard-Gaudin**

Christian VIARD-GAUDIN est professeur au département GEII de l'IUT de Nantes. Ses travaux de recherche sont axés sur le traitement de l'écriture manuscrite et des documents. Il coordonne le projet ANR Technologie Logicielle « Conversion et Indexation de l'Ecriture en-Ligne » (CIEL) qui comporte des axes sur la modélisation avancée de documents manuscrits en-ligne complexes, notamment les formules mathématiques et les schémas électriques, et sur la catégorisation et l'indexation de ces documents. Ses travaux se sont développés dans un cadre international avec des partenariats avec la Malaisie, Singapour et la Chine mais aussi avec la compagnie Vision Objects qu'il a contribué à faire naître sur la technopole nantaise.

Guo **Xian Tan**

Guoxian TAN est un étudiant en dernière année de doctorat. Il effectue sa thèse en cotutelle entre l'Université de Nantes à l'IRCCyN et l'Université de Technologie Nanyang à Singapour. Son sujet de thèse porte sur l'identification du scripteur mais aussi sur l'identification du type d'alphabet utilisé pour composer un document manuscrit en-ligne. Sur ces sujets, il est l'auteur de six articles en conférences internationales et d'un article dans la revue Pattern Recognition.

Alex **C. Kot**

Alex KOT est professeur et directeur adjoint d'une importante composante (college of Engineering) de la Nanyang Technological University à Singapour après avoir été directeur du département « Electrical and Electronic Engineering » pendant 8 ans. Ses travaux de recherche couvrent les domaines du traitement du signal pour les communications, de la biométrie et de l'authentification. Il est actuellement éditeur associé des revues EURASIP Journal on Applied Signal Processing, IEEE Transactions on Multimedia, IEEE Signal Processing Letter et IEEE Signal Processing Magazine.